Assessing the Effects of Weather on Bike Share Usage in Philadelphia

by

Alex Christopher Watts

Bachelor of Science, University of Alabama, 2018

Submitted to the Graduate Faculty of the

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Alex Christopher Watts

It was defended on

April 24, 2023

and approved by

Jeanine M. Buchanich, M.Ed., M.P.H., Ph.D. Associate Professor, Department of Biostatistics

Jenna C. Carlson, Ph.D., Assistant Professor, Department of Biostatistics

Aleksandar Stevanovic, Ph.D., Associate Professor, Department of Civil & Environmental Engineering

Thesis Advisor: Jeanine M. Buchanich, M.Ed., M.P.H., Ph.D. Associate Professor, Department of Biostatistics

Assessing the Effects of Weather on Bike Share Usage in Philadelphia

Alex Christopher Watts, MS

University of Pittsburgh, 2023

Physical inactivity and pollution from motor vehicle emissions are major public health risks in the 21st century. One potential way of addressing both of these issues is shifting transportation needs from cars to bicycles. Bike shares offer cities and other organizations an opportunity to provide convenient, affordable access to bicycles. While these systems can improve access to active transportation, there is a large body of evidence that suggests weather may pose a barrier to people using bikes for more trips. Past work looking at the bike share system in Washington, D.C. found that riders reduce the number and duration of bicycle trips in colder, wetter conditions. However, this does not necessarily apply to Philadelphia, as rider responses to weather can vary a great deal between cities. In this thesis, I used negative binomial regression and linear regression with a log-transformed outcome to estimate the effects of various weather conditions on the number and duration of rides in Philadelphia's Indego bike share system. To fit these models, I used data from January 1, 2016 to December 31, 2022 for 5,280,976 rides obtained from Indego's website and historical weather data for 61,362 hours obtained from Visual Crossing. As a secondary analysis, I partitioned the data into the group of rides taken by riders with a monthly or yearly pass and the group of rides taken by the rest of the riders. Both ride counts and mean duration generally increased with temperature and decreased with the introduction of precipitation. The group of monthly and yearly passholders showed a diminished response to both of these compared to riders overall. This suggests that weather may pose a barrier to the use of bike share as a form of transportation, and public health officials and transportation planners may be able to alleviate this by expanding long-term memberships. Shifting trips from personal vehicles to bicycles and bike share is crucial to reducing pollution from motor vehicle emissions and increasing physical activity.

Table of Contents

List of Tables

List of Figures

Preface

I would like to thank my fiancée, Taylor, who has had to listen to me talk about bikes more than any sane person should. I would also like to thank my committee, who have been extremely patient through a lot of mind-changing and rabbit hole-diving.

1.0 Introduction

Physical inactivity and motor vehicle emissions both pose serious health risks (CDC, 2023; Gentner et al., 2017a, 2017b; World Health Organization, 2018). Transitioning trips from motor vehicles to bicycles is associated with benefits from decreases in pollution and increased levels of exercise (Lindsay et al., 2011). Additionally, bicycle use introduces a previously underused form of exercise in commuters and other travelers' lives. Taking this into account, there is good reason to believe increasing the modal share using bikes for everyday transportation will provide ample health benefits to the users, as well as the overall community. Bike shares—systems that provide a fleet of bicycles available for short-term rentals on demand—can offer a convenient option for planners and public health officials to encourage this shift; however, weather conditions may pose a barrier to such efforts.

1.1 Bike Shares v. Private Bicycles

Bike shares offer several advantages over personal bikes. They offer affordable, readily available access to active transportation to populations that otherwise might travel by more sedentary means. Bike shares present a lower up-front cost to travelers than personal bicycles, which could appeal to those hesitant about the full cost of a personal bike. Additionally, bike shares pose a lower stress alternative to personal bikes because the user does not take on any of the maintenance costs or the risk of theft. Personal bikes need to be locked up and looked after. A personal bike has to be brought with the user from place to place, whereas a shared bike can be dropped off with no concern as to what happens after the ride ends.

In addition to riders without access to a personal bike, bike shares can be useful to those with their own bike. A recent review found that bike share users were more likely to have a personal bike available to them, as well. However, a 2012 survey of 10,661 users of bike share in Montreal, Toronto, Washington, D.C., and Minneapolis/St. Paul found ~30% of respondents reported using alternatives to bicycles less often due to bike share, including taxi and private automobiles (Shaheen et al., 2012). Thus, while bike share users may own personal bikes, a substantial portion of travelers who may have otherwise driven or taken another form of transportation are doing so less often as a result of bike shares.

In many cases, bike shares attract travelers who otherwise would have used public transportation, walked, or driven a car. One analysis of the bike shares in Melbourne, Brisbane, Washington, D.C., London, and Minneapolis/St. Paul found that less than 10% of bike share users would otherwise have ridden a personal bicycle (Fishman et al., 2015).

Bike shares attract more women—and possibly more riders of ethnic minorities—than personal bicycles. These results have been somewhat mixed and depend on the city in question. In New York, one review of bike share riders compared to riders as a whole found more women ride bike share bikes, but—compared to the city overall—all bike riders are disproportionately white, educated, and wealthy (Crossa et al., 2022). In other words, people who ride bikes tend to be from less marginalized communities, but more bike shares attract a larger proportion of women than ride personal bicycles. A similar study in Washington, D.C. found that, while bike share riders were still more likely to be white and male, this disparity was far less than riders of personal bicycles (Buck et al., 2013).

1.2 Weather and Bike Rider Behavior

Riders have previously been found to respond to weather conditions in a variety of contexts. However, this response appears to vary by region. In the Netherlands, cycling comprises 25% of all trips, and between 50 and 80% of all daily fluctuations in cycle trips can be attributed to weather, with more rides being taken in warmer weather with less precipitation (Thomas et al., 2008, 2013). In 2021, Goldmann and Wessel found significant variation in sensitivity to weather among bike users when comparing 30 German cities (Goldmann & Wessel, 2021). Goldmann and Wessel compare the difference between the average count of riders in the 25th percentile of weather severity and the average count of riders in the 75th percentile. The city of Oldenburg was reported to have no difference, while Würzberg observed a 20% reduction in ridership. Even within the same province as Würzberg, Erlangen saw a reduction of just 10%. Goldmann and Wessel attribute the bulk of these differences to the share of the city's population between 18 and 25 and the connectivity of the city's bicycle network. These variations within one country, and even within a single province of a country, suggest that it is important to consider the conditions separately.

Previous work in Washington, D.C. looking at the Capital Bikeshare (CaBi) found significant relationships between temperature and precipitation and both the number of and the mean duration of rides (Gebhart & Noland, 2014). Thus Philadelphia, although close to Washington, D.C., may have distinct rider behavior in response to the weather. Gebhart and Noland do, however, provide a good template for analyzing the counts and mean durations of ride-share trips within Philadelphia. They employ negative binomial regression for the counts and linear regression for the mean durations. They provide a secondary result analyzing the proximity of starting ride-share stations to transit stops to determine whether there is a difference in response to weather based on the availability of alternative transit methods within 0.25 miles. They do this by way of partitioning their overall dataset into rides that began within 0.25 miles of a transit stop and those that begin farther away. They conclude that riders who start trips close to transit are more likely to shift trips to alternative modes in inclement weather than those for whom transit is less convenient.

1.3 Bike Share in Philadelphia

Philadelphia has an ever-growing bicycle culture. The 2020 American Community Survey estimated Philadelphia's bike commuting rate to be 2.1%, which is higher than any other of America's 10 largest cities (The Bicycle Coalition of Greater Philadelphia, 2022). In 2021, bike traffic increased by 29% from 2020 (The Bicycle Coalition of Greater Philadelphia, 2022). The city ranks above average for all US cities in bicycle safety and accessibility (PeopleForBikes, 2022). The city's bike share has no doubt bolstered these numbers.

The Indego bike-share program was introduced as a component of Philadelphia's transit network in April 2015. It is owned by the city itself, under the Office of Transportation, Infrastructure, and Sustainability. Although privately sponsored by Independence Blue Cross and operated by Bicycle Transit Systems, it functions as part of the public transportation system, offering an affordable, healthy, and environmentally conscious mode of transportation for residents throughout the city. Residents are free to request a station anywhere within the city, and infill stations are added regularly. The program added 42 stations between January 1, 2023 and March 16, 2023. As part of the Better Bike Share Partnership, this program maintains a commitment to understanding and addressing the barriers "to the use of bike share in low-income and communities of color," which are generally overlooked by planners when implementing similar programs in US cities (Better Bike Share Partnership, 2023; Ursaki & Aultman-Hall, 2015).

Indego has operated under a few different fee structures. Initially, riders could purchase individual rides or a monthly pass. This was meant to reflect common fee structures for transit. The individual ride would cost a fee plus $0.20 for each minute. Monthly passes entitled members to unlimited 60-minute rides with a fee of $0.20 per minute over the hour. Beginning in April 2018, this system was replaced with a new fee structure that allowed users to purchase passes for a day, month, or year. All these memberships allowed for unlimited 60-minute rides within the set period, with a fee of $0.20 for each minute after. Beginning in November 2018, e-bikes were available for an additional $0.20 per minute (Foursquare ITP, 2018). In April 2022, the cost of each pass type was increased (Hooven, 2022). Passes and rides are available at a reduced rate for Pennsylvania residents who receive SNAP benefits through the state's ACCESS program (*Buy a Pass*, 2020). While the available passes have changed over time, the overall fee structure has remained largely consistent, with some users maintaining long-term memberships and others only engaging with the system for short periods. This raises the question of whether behavior differs between these two groups.

In 2021, there were 1,500 bikes in the system, with 8% of all rides in the city taking place on Indego bikes (The Bicycle Coalition of Greater Philadelphia, 2022). While this is a relatively small portion of the overall bicycle usage for the city, it represents an improvement of two percentage points from 2018 (Bicycle Coalition of Greater Philadelphia, 2018). Continued increases in the usage of Indego would mean greater access to micromobility for more citizens around the city, conferring the benefits outlined above.

1.4 Current Thesis

With evidence of increasing usage of Philadelphia's Indego bike share established, further understanding of what impacts riders' decisions to use bike shares is imperative. While previous work has found that rider behavior is influenced by weather patterns, with one analysis finding that the CaBi bike share in Washington, D.C. experienced a reduction in rides in poor weather, there is reason to believe rider behavior varies from region to region, and there has not been a recent analysis of rider responses to weather in Philadelphia, PA.

The purpose of this thesis is to replicate the findings of Gebhart and Noland in the similar but distinct city of Philadelphia using a larger, more recent dataset. I hypothesize that these results will follow closely to the Washington, D.C. population; however, Philadelphia's public transportation system is not quite as robust as Washington's, as riders in Philadelphia have fewer alternatives to bike sharing available and may be less inclined to forego a ride in poor conditions. Additionally, more recent trends suggest that bicycle commuting may be more popular in recent years than during the CaBi study's timeframe. Additionally, rather than evaluating results based on proximity to public transportation—as Gebhart and Noland have done—my secondary analyses examine riders who had monthly/yearly passes and no pass/daily passes. I have partitioned my dataset into two subsets. The first contains rides made by long-term passholders, and the other has rides from those who have a day pass or less. There is reason to believe that committed members of bike share programs respond differently to weather conditions than casual riders. As described below, in the case of estimating the impact of weather on ride counts, I employed a negative binomial model; for the mean ride duration, I employed a log-linear approach.

2.0 Methods

2.1 Statistical Analysis

2.1.1 Negative Binomial Model

When modeling count data, a Poisson model is often preferred. Sometimes, data are not Poisson distributed, and another approach is necessary. When data show a variance larger than their mean, they are considered overdispersed. In such cases, a correction is necessary. One good alternative when dealing with such overdispersion is the Negative Binomial model (Engel, 1984).

For a collection of counts $\boldsymbol{Y},$ Poisson regression generally relies on three assumptions:

(i) The counts, $\boldsymbol{Y}$, are independent.

(ii) Each count $Y_i \sim Poisson(\lambda)$. That is, each individual count follows a Poisson distribution with rate λ.

(iii) $\log(\lambda)$ has a linear relationship with the specified set of covariates $\boldsymbol{x}$. Put another way, $\lambda = e^{\boldsymbol{x}^T\boldsymbol{\beta}}$, where $\boldsymbol{x}$ is a set of covariates and $\boldsymbol{\beta}$ is a set of unknown parameters.

The key problem with overdispersed data is assumption (ii). For a count to follow a Poisson distribution, its mean μ must be equal to its variance σ^2. In many cases, this is not met. In cases where $\sigma^2 \gg \mu$, an alternative model is required.

According to Engel, we can consider the rate, λ, of a Poisson variable, X, to be a random variable itself, M, which is distributed Gamma with shape parameter α and scale parameter θ (Engel, 1984). Under this new assumption, X is distributed Negative Binomial with dispersion parameter α and probability $p = \frac{\theta}{1+\theta}$. This allows us to account for the excessive variation by replacing assumption (ii) with the following assumption:

(ii*) $X_i \sim NB\left(\alpha, p = \frac{\theta}{(1+\theta)}\right)$, or, each individual count follows a negative binomial distribution with probability p and dispersion parameter α.

Referring back to assumption (iii), we define $\mu(\boldsymbol{x_i}) = \lambda = e^{\boldsymbol{x_i^T \beta}} = \alpha\theta$ where $\boldsymbol{x_i}$ is the collection of covariates associated with Y_i, the ith count and $\boldsymbol{\beta}$ is a vector of unknown parameters. Then we can characterize the mean and variance of Y_i as the following:

$$E(Y_i|\boldsymbol{x_i}) = \mu(\boldsymbol{x_i}), Var(Y_i|\boldsymbol{x_i}) = \mu(\boldsymbol{x_i}) + \alpha\mu(\boldsymbol{x_i})^2$$

Considered in this lens, then we can consider the distribution of the counts $Y \sim NB(\mu(\boldsymbol{x}), \alpha)$ (Lawless, 1987). As $\alpha \to 0$, this approaches the Poisson model with rate $\mu(\boldsymbol{x})$.

For the ith count, this yields the following model:

$$\ln\big(\mu(\boldsymbol{x_i})\big) = \beta_0 + \ \beta_1\, x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

Here, $\boldsymbol{x_i} = [x_{1i} \ldots x_{ki}]^{\boldsymbol{T}}$ is the k-dimensional vector of covariates associated with the i-th count.

Results from such models are best interpreted using an Incidence Rate Ratio (IRR), computed by exponentiating the estimated coefficients. An IRR of 1 suggests an incidence rate equivalent to the reference group. Greater than 1 suggests a greater incidence, and less than 1 suggests a lower incidence.

2.1.2 Linear Regression with Log-Transformed Outcome

A standard linear regression is often the most appropriate starting point for constructing a model. However, linear models are inappropriate for cases where outcome values are exclusively positive and skewed. A linear model would allow for—and assume the possibility of—negative outcome values (Gelman & Hill, 2006). This is not possible for time-valued outcomes. In a case like this, it is more appropriate to fit the following model:

$$\ln(Y_i) = \beta_0 + \ \beta_1 X_{1i} + \cdots \ \beta_K X_{Ki} + E_i$$

Which is equivalent to

$$Y_i = e^{\beta_0} e^{\beta_1 X_{1i}} \dots e^{\beta_K X_{Ki}} e^{E_i}$$

This equation models the data as a multiplicative process rather than an additive one. This ensures the values estimated for the outcome will be positive, which fits the underlying data much better (Gelman & Hill, 2006).

Moving back to the linearized interpretation, we can use ordinary least squares (OLS) under the usual assumptions to estimate the $\boldsymbol{\beta}$ for $\ln(Y_i) = \boldsymbol{X_i^T \beta}$, where Y_i is the mean duration for the ith hour and $\boldsymbol{X_i}$ is the vector of associated covariates.

OLS is an algorithm for estimating the coefficients of a linear regression model of the form $Z = \boldsymbol{X^T} \boldsymbol{\beta} + \epsilon$, where Z is the outcome of interest, $\boldsymbol{X}$ is the vector of covariates associated with that outcome, $\boldsymbol{\beta}$ is the vector of unknown coefficients, and ϵ is a random error term (Wooldridge, 2013). Under the following assumptions, OLS provides the best, least unbiased estimators:

(i) Z is linear with respect to $\boldsymbol{X}$.

(ii) Each sample is independent.

(iii) No perfect collinearity within $\boldsymbol{X}$.

(iv) The error term, ϵ is conditionally centered on 0. In other words, $E(\epsilon|\boldsymbol{X}) = 0$.

In this particular case, if we take $Z = \ln(Y)$, it is reasonable to expect that these properties still hold. Thus, OLS is an appropriate tool for estimating this model.

A small consideration is necessary when interpreting this particular setup, though. While traditional multiple linear regression can be interpreted directly as a 1-unit change in X_k being associated with a β_k-unit change in Z, here we interpret the results as a 1-unit change in X_k being associated with a β_k% change in Y (Gelman & Hill, 2006; Wooldridge, 2013).

2.2 Data Source

I obtained data on Indego bike share usage from Indego's website (Indego, 2022). I obtained hourly historical weather data from Visual Crossing (Visual Crossing, 2023). I restricted data for this project to the period spanning January 2016 – December 2022, which includes the first full year and last full year of available data.

The variables of interest in the Indego data were the trip start date and time, the user's passholder type, and the ride duration.

The variables of interest in the weather history data were temperature, wind speed, humidity, solar energy, and precipitation. I removed visibility as a potential covariate due to low variability.

2.3 Data Processing

Once I acquired the data, I selected the variables of interest from the datasets. I then joined the weather to the bike share data by the hour in which the trip start time fell. From there, I aggregated the rides for each hour in two ways: counts and mean duration. I created three datasets for each of these. The first was the overall data. This provides the total number of rides from all users summed for each hour as the variable count and the arithmetic mean duration of rides from all users for each hour as the mean duration. The other two datasets provide the same measures for users with a monthly or yearly pass and a day pass or no pass.

Rides were only included in the data if they lasted longer than one minute and fewer than 1,440 minutes. Rides shorter than one minute were likely an immediate return, and anything longer than 1,440 minutes (24 hours) was likely the result of an issue returning the bike.

Additionally, I created a new categorical variable for temperature, splitting it over 10°F intervals. This allows for a non-linear effect relationship between temperature and the outcome variables without sacrificing interpretation. I used the mean temperature category (50-59°F) as the reference group. As a proxy for dark conditions, I created a variable that took all hours where solar energy was either not available or 0 as dark. I created the categorical variables for weekend and peak hours based on the day of the week or hour during which the start time fell. A ride is flagged as "weekend" if it fell on Saturday or Sunday (days 6 and 7 based on the default day coding used by the week() function in the R package Lubridate). A ride is flagged as "peak" if it began during the intervals from 7 a.m. – 9 a.m. and 4 p.m. – 7 p.m.

All other variables were left as they were initially reported by the data sources.

2.4 Analysis Plan

After data collection and processing, I had the following variables for each hour:

- Ride count – The number of trips started during that hour
- Mean duration – The mean duration of trips started during that hour

- Temperature – The mean temperature in °F for that hour
- Temperature (discrete) – A set of ten indicator variables denoting the range of the mean temperature for that hour binned as
  - < 10°F
  - 10-19°F
  - 20-29°F
  - 30-39°F
  - 40-49°F
  - 50-59°F
  - 60-69°F
  - 70-79°F
  - 80-89°F
  - >= 90°F
- Humidity – The relative humidity in % for that hour
- Windspeed – The average wind speed in MPH for that hour
- Dark – An indicator for whether conditions were dark (i.e., solar energy was missing or 0) during that hour
- Precipitation – The precipitation type that occurred during that hour
  - None
  - Rain
  - Snow
  - Other (hail, sleet, etc.)
- Weekend – An indicator of whether that hour fell on a Saturday or Sunday

- Peak Hours – An indicator of whether that hour fell on in the range of 7 a.m. – 9 a.m. or 4 p.m. – 7 p.m.

I have three total datasets, each with the same variables:

- Overall – The total number of rides lasting longer than 1 minute and less than 1440 minutes taken by any rider using any type of pass between 00:00:00am EST January 1, 2016 and 11:59:59pm EST December 31, 2022
- Passholders – The total number of rides lasting longer than 1 minute and less than 1440 minutes taken by riders using a monthly or yearly pass between 00:00:00am EST January 1, 2016 and 11:59:59pm EST December 31, 2022
- Non-passholders/casual riders – The total number of rides lasting longer than 1 minute and less than 1440 minutes taken using no pass or a day pass between 00:00:00am EST January 1, 2016 and 11:59:59pm EST December 31, 2022

Based on past research, there was reason to suspect that temperature would have a non-linear relationship with rider behavior (Ahmed et al., n.d.; Gebhart & Noland, 2014; Wessel, 2020). To prioritize the interpretability of results, I chose to include the discretized temperature rather than a more advanced non-linear treatment of continuous temperature.

The reference groups for the categorical variables in all models were: weekday, off-peak hours, temperature 50-59°F, conditions were not dark, and there was no precipitation.

For hourly ride counts, I used a negative binomial model with the following log-linear formulation:

$$\begin{aligned}\ln(count_i) = {} & \beta_0 + \beta_1 I(< 10°F_i = 1) + \beta_2 I(10 - 19°F_i = 1) + \beta_3 I(20 - 29°F_i = 1) \\ & + \beta_4 I(30 - 39°F_i = 1) + \beta_5 I(40 - 49°F_i = 1) + \beta_6 I(60 - 69°F_i = 1) \\ & + \beta_7 I(70 - 79°F_i = 1) + \beta_8 I(80 - 89°F_i = 1) + \beta_9 I(\geq 90°F_i = 1) \\ & + \beta_{10} \frac{Windspeed_i}{5} + \beta_{11} I(Dark_i = 1) + \beta_{12} I(Weekend_i = 1) \\ & + \beta_{13} I(Peak\ Hours_i = 1) + \beta_{14} I(Rain_i = 1) + \beta_{15} I(Snow_i = 1) \\ & + \beta_{16} I(Other_i = 1) + \beta_{17} \frac{Humidity_i}{5} + u_i\end{aligned}$$

The duration model, which is a linear regression with a log-transformed outcome, has this formulation:

$$\begin{aligned}&\ln(Mean\ Duration_i) \\ & = \beta_0 + \beta_1 I(< 10°F_i = 1) + \beta_2 I(10 - 19°F_i = 1) + \beta_3 I(20 - 29°F_i = 1) \\ & + \beta_4 I(30 - 39°F_i = 1) + \beta_5 I(40 - 49°F_i = 1) + \beta_6 I(60 - 69°F_i = 1) \\ & + \beta_7 I(70 - 79°F_i = 1) + \beta_8 I(80 - 89°F_i = 1) + \beta_9 I(\geq 90°F_i = 1) \\ & + \beta_{10} \frac{Windspeed_i}{5} + \beta_{11} I(Dark_i = 1) + \beta_{12} I(Weekend_i = 1) \\ & + \beta_{13} I(Peak\ Hours_i = 1) + \beta_{14} I(Rain_i = 1) + \beta_{15} I(Snow_i = 1) \\ & + \beta_{16} I(Other_i = 1) + \beta_{17} \frac{Humidity_i}{5} + \epsilon_i\end{aligned}$$

I divide humidity and windspeed by 5 here to estimate the effect for a 5% or 5 MPH increase, respectively. If left unadjusted, these estimates may be too small to warrant notice. Using five unit increments provides a more intuitive interpretation without biasing the results.

In both cases here, the model was assumed to use a linear combination of the covariates. However, the assumptions listed in sections 2.1.1 and 2.1.2 lead to different interpretations and results for each.

I employed several packages in R to conduct this analysis. For data management and processing, I used the *tidyverse* ecosystem of packages, which includes *dplyr*, *tidyr*, *tibble*, *forcats*, *readr*, *stringr*, *purrr*, and *ggplot2* (Müller & Wickham, 2023; Wickham, 2016, 2022a, 2022b; Wickham et al., 2019, 2022; Wickham, François, et al., 2023; Wickham, Vaughan, et al., 2023; Wickham & Henry, 2023). I used *skimr* to perform an initial inspection of the data, and I checked for missing values using the *md.pattern()* function from the *mice* package (van Buuren & Groothuis-Oudshoorn, 2022; Waring et al., 2022). To handle the date and time components of the analysis, I processed data using the *lubridate* package (Grolemund & Wickham, 2011).

To produce and format summary tables, I employed a combination of functions from the *table1*, *gtsummary*, and *flextable* packages (Gohel & Skintzos, 2022; Rich, 2023; Sjoberg et al., 2023). In addition to *ggplot2*, I used *ggsci* and *ggpubr* for visualizations (Kassambara, 2022; Xiao, 2018).

For the negative binomial model, I used the *glm.nb()* function from the *MASS* package in R (Venables & Ripley, 2002). I reported the results here as incidence rate ratios, which allow for direct interpretation against the reference group used. The three partitions of the dataset (overall, pass, and no pass) were structured the same way, so I fit the same model for all three. This allowed for a direct interpretation of the results for the two subsets against the overall model. This was also true for the duration model, in which the base R *lm()* function was used (R Core Team, 2020). I report these results as the estimated coefficients, which can be interpreted as the percent change in the outcome for a one-unit change in a given covariate.

To produce p-values below the default R threshold of 2e^{-16}, I used the *Rmpfr* package to set that threshold lower (Maechler, 2021). In addition to the tables of results, I produced forest plots to visualize effects using the *ggforestplot* package (Scheinin et al., 2023). For the negative binomial results, I report all values as IRRs to allow for direct interpretation. I do this in the tables, as well as the forest plot. I report linear model results for the duration with the directly estimated coefficients.

3.0 Results

3.1 Preliminary Analysis

Conditions for the time period from January 1, 2016 to December 31, 2022 are shown in Table 1. The study period contained a total of 61,361 hours. During this period, the mean temperature was 57.2°F, and the median was 57.6°F. The standard deviation for the temperature was 18°F. The lowest temperature reported during this period was 4.7°F, and the highest was 98°F. The plurality of hours fell in the range between 70-79°F; however, the measures of centrality, mean and mode, both fell within the 50-59°F range. Because of this, 50-59°F served as the reference group.

The mean wind speed was 8.08 MPH, and the median was 7.40 MPH. Windspeed ranged from 0 to 32.6 MPH, and the standard deviation of windspeed was 4.88 MPH.

Relative humidity ranged from 12.53% to 100%. The mean humidity was 64.414%, with a standard deviation of 19.456%. The median humidity was 64.52%.

Of the total hours in this period, 50% (n = 30,659) were flagged as “dark”. Weekend hours comprised 28.6% of the data (n = 17,568), which is roughly 2/7 of the total time. A total of 20.8% (n = 12,785) of the total time fell during peak hours, which is consistent with 5 of 24 hours being defined as peak hours.

For most of this period, there was no precipitation, with none being reported 87.2% (n = 53,519) of the time. Rain was reported for 11.6% (n = 7,101) of the hours, and snow was reported for 1.2% (n = 723) of the hours. Other forms of precipitation occurred less than 0.1% of the time (n = 18).

Table 1: Summary Statistics of Hourly Conditions in Study Period (N = 61,361)

Characteristic	Mean/Count (%)[1]	SD	Min	Median	Max
Temperature	57.2	18.0	4.7	57.6	98
< 10°F	45 (0.1%)				
10-19°F	547 (0.9%)				
20-29°F	2,653 (4.3%)				
30-39°F	8,506 (13.9%)				
40-49°F	10,686 (17.4%)				
50-59°F	9,472 (15.4%)				
60-69°F	10,017 (16.3%)				
70-79°F	12,099 (19.7%)				
80-89°F	6,166 (10.0%)				
>= 90°F	1,170 (1.9%)				
Windspeed (MPH)	8.08	4.88	0	7.40	32.6
Dark	30,659 (50.0%)				
Weekend (Sat, Sun)	17,568 (28.6%)				
Peak Hours (7am-9am, 4pm-7pm)	12,785 (20.8%)				
Precipitation					
None	53,519 (87.2%)				
Rain	7,101 (11.6%)				
Snow	723 (1.2%)				
Other	18 (<0.1%[2])				
Humidity (%)	64.414	19.456	12.53	64.52	100

[1] Means reported for continuous variables. Counts with proportions reported for categorical variables.

[2] 0.0003%

Detailed counts of rides by categorical variable and passholder type, as well as an overall summary, are reported in Table 2. The overall mean hourly count of rides was 86.1, with a standard deviation of 81.2. Squaring that to obtain the variance, we get 6,593.44, which is much larger than the mean of 86.1, confirming that Poisson regression is inappropriate. The hour with the most rides had a total of 570, which took place on a weekday, with temperature in the 60-69°F range with no precipitation. There were many hours during which no rides took place.

Among monthly and yearly passholders, the mean ride count was 75.8, with a standard deviation of 72. The maximum number of rides among this group was 532. Among the daily or no passholders, the mean ride count was 10.3, with a standard deviation of 16.5, a median of 4, and a maximum of 236.

Table 3 contains the mean hourly ride duration by categorical variable and passholder type. The mean hourly mean duration of rides for all riders was 21.5 minutes, with a standard deviation of 23.5, a median of 17.0, and a maximum of 1,210. Among users with a monthly or yearly pass, the mean was 18.9 minutes. The standard deviation was 20.2, the median was 15.2, and the maximum was 1,150. Riders with a day pass or less had a mean duration of 39.1 minutes, a standard deviation of 56.6, a median of 27.1, and a maximum of 1,410.

The plurality of rides for all groups fell in the 70-79°F range, and 91.9% (n = 4,853,719) of rides occurred when there was no precipitation. Most occurred when it was not dark, and only 28.3% (n = 1,496,133) occurred on weekends. This percentage is higher among casual riders (38.9%; n = 245,968).

Table 2: Ride Count by Passholder Type and Conditions

	Overall Dataset					Passes Only Dataset					No Pass Dataset				
Characteristic	Mean	SD	Min	Median	Max	Mean	SD	Min	Median	Max	Mean	SD	Min	Median	Max
Ride Count	86.1	81.2	0	63.0	570	75.8	72.0	0	57.0	532	10.3	16.5	0	4.00	236
Temperature															
< 10°F	7.09	12.0	0	3.00	69	6.91	12.0	0	3.00	69	0.178	0.576	0	0	3
10-19°F	14.9	19.7	0	8.00	132	14.5	19.5	0	8.00	132	0.404	0.793	0	0	5
20-29°F	25.6	30.0	0	15.0	183	24.8	29.3	0	14.0	179	0.796	2.44	0	0	75
30-39°F	37.8	39.4	0	26.0	337	36.0	38.0	0	24.0	283	1.78	5.51	0	0	188
40-49°F	54.6	50.5	0	44.0	497	51.0	48.4	0	40.0	431	3.60	5.73	0	2.00	102
50-59°F	71.5	63.3	0	59.0	534	64.6	58.8	0	52.0	441	6.89	10.1	0	3.00	123
60-69°F	95.4	80.4	0	80.0	570	83.8	72.9	0	70.0	444	11.6	16.5	0	6.00	183
70-79°F	114	89.6	0	103.0	545	98.6	79.7	0	88.0	470	15.9	19.9	0	9.00	190
80-89°F	171	81.9	0	162.0	554	143	74.6	0	131	532	27.9	22.7	0	22.0	236
>= 90°F	192	72.7	68	179.0	465	166	72.5	40	151	431	25.9	16.3	1	22.0	118
Dark	39.4	45.7	0	22.0	463	35.4	42.0	0	19.0	415	3.96	7.32	0	1.00	129
Weekend (Sat, Sun)	85.2	79.4	0	61.0	554	71.2	65.4	0	53.0	425	14.0	20.9	0	5.00	236
Peak Hours (7am-9am, 4pm-7pm)	155	99.4	0	145.0	570	141	91.6	0	129	532	14.1	19.1	0	6.00	188
Precipitation															
None	90.7	82.6	0	70.0	570	79.5	73.2	0	62.0	532	11.1	17.1	0	4.00	236
Rain	58.4	63.4	0	36.0	464	53.4	57.8	0	33.0	448	4.98	9.92	0	1.00	174
Snow	16.1	20.0	0	9.00	154	15.6	19.6	0	9.00	151	0.451	0.971	0	0	7
Other	29.3	24.3	1	25.0	78.0	28.5	23.3	1	23.5	70	0.833	1.89	0	0	8

Table 3: Mean Duration by Passholder Type and Conditions

	Overall Dataset					Passes Only Dataset					No Pass Dataset				
Characteristic	Mean	SD	Min	Median	Max	Mean	SD	Min	Median	Max	Mean	SD	Min	Median	Max
Mean Duration	21.5	23.5	1	17.0	1210	18.9	20.2	1	15.2	1150	39.1	56.6	1	27.1	1410
Temperature															
< 10°F	50.1	186	4.25	10.9	1150	51.3	191	4.25	10.8	1150	19.5	12.3	8	20.0	39
10-19°F	23.4	66.5	3	12.2	1210	20.1	37.7	3	12.0	574	38.7	127	1	14.0	1210
20-29°F	19.3	30.6	1	12.7	581	18.0	24.1	1	12.3	352	32.4	86.6	1	14.0	1390
30-39°F	19.1	25.7	2	13.8	661	17.9	24.4	2	13.2	613	34.0	76.0	1	17.0	1410
40-49°F	19.5	25.2	2	14.8	917	17.6	19.5	2	13.8	702	36.4	69.3	1	20.7	1310
50-59°F	20.7	22.4	2	16.3	1020	18.5	21.6	2	14.7	1020	37.3	52.8	1	24.7	1390
60-69°F	22.7	21.0	3	17.8	625	19.5	17.9	2	15.6	455	41.7	58.1	1	28.5	1280
70-79°F	24.0	21.3	1	19.6	1070	20.5	18.1	1	16.7	841	42.7	49.1	1	31.4	1070
80-89°F	23.1	10.9	3	21.0	408	19.5	9.89	3	17.6	408	39.9	26.2	1	33.9	545
>= 90°F	21.2	6.47	11.4	19.5	57.1	18.4	5.53	11.1	16.8	55.2	37.2	20.5	6	32	196
Dark	23.3	32.1	1	16.0	1210	20.7	27.7	1	14.5	1150	40.9	73.8	1	22.4	1410
Weekend (Sat, Sun)	22.2	21.4	2	18.2	917	19.0	16.8	2	15.8	513	39.8	54.3	1	28.5	1280
Peak Hours (7am-9am, 4pm-7pm)	18.0	8.14	4.5	14.9	309	16.3	6.72	3.67	14.7	135	33.9	40.3	1	25.6	1160
Precipitation															
None	21.6	22.0	1	17.3	1210	18.9	18.8	1	15.3	1150	39.2	53.9	1	27.7	1410
Rain	21.2	31.4	2	15.6	1020	19.3	27.7	2	14.6	1020	38.2	70.6	1	22.2	1280
Snow	22.2	39.7	2	13.8	581	19.7	27.3	2	13..5	308	52.2	166	1	15.0	1390
Other	11.5	5.88	4	11.2	31.2	11.3	5.83	4	11.1	31.2	21.0	9.93	5	20.0	32

Table 4 provides a breakdown of the proportion of rides occurring across conditions. Most occurred in warmer temperatures, with the plurality of rides occurring during periods where mean hourly temperatures were between 70°F and 79°F. A majority of rides took place on weekdays, and a slight majority fell outside of peak hours.

Table 4: Counts of Rides in Each Condition by Pass Type

Characteristic	Overall (N = 5,280,976)	Pass (N = 4,648,352)	No Pass (N = 631,835)
Temperature			
< 10°F	326 (<0.1%[1])	311 (<0.1%[3])	8 (<0.1%[5])
10-19°F	8,220 (0.2%)	7,946 (0.2%)	221 (<0.1%[6])
20-29°F	68,153 (1.3%)	65,889 (1.4%)	2,113 (0.3%)
30-39°F	321,358 (6.1%)	306,003 (6.6%)	15,116 (2.4%)
40-49°F	583,854 (11.1%)	545,167 (11.7%)	38,494 (6.1%)
50-59°F	677,203 (12.8%)	611,859 (13.2%)	65,269 (10.3%)
60-69°F	955,852 (18.1%)	839,824 (18.1%)	115,983 (18.4%)
70-79°F	1,385,014 (26.2%)	1,192,953 (25.7%)	192,038 (30.4%)
80-89°F	1,056,089 (20.0%)	883,822 (19.0%)	172,264 (27.3%)
>= 90°F	224,907 (4.3%)	194,578 (4.2%)	30,329 (4.8%)
Dark	1,208,538 (22.9%)	1,086,461 (23.4%)	121,333 (19.2%)
Weekend (Sat, Sun)	1,496,133 (28.3%)	1,249,994 (26.9%)	245,968 (38.9%)
Peak Hours (7am-9am, 4pm-7pm)	1,982,719 (37.5%)	1,802,050 (38.7%)	180,645 (28.6%)
Precipitation			
None	4,853,719 (91.9%)	4,257,003 (91.6%)	596,112 (94.3%)
Rain	415,028 (7.9%)	379,524 (8.2%)	35,382 (5.6%)
Snow	11,701 (0.2%)	11,312 (0.2%)	326 (0.1%)
Other	528 (<0.1%[2])	513 (<0.1%[4])	15 (<0.1%[7])

[1] 0.00617%

[2] 0.00100%

[3] 0.00669%

[4] 0.01104%

[5] 0.00127%

[6] 0.03498%

[7] 0.00237%

Visual examination of the number of trips started each hour by membership status and weather status (Figure 1) suggests passholders were more likely to ride during the peak hours of

7am-9am and 4pm-7pm. Non-passholders tended to ride more during the early afternoon. Passholders also appear to be less sensitive to inclement weather.

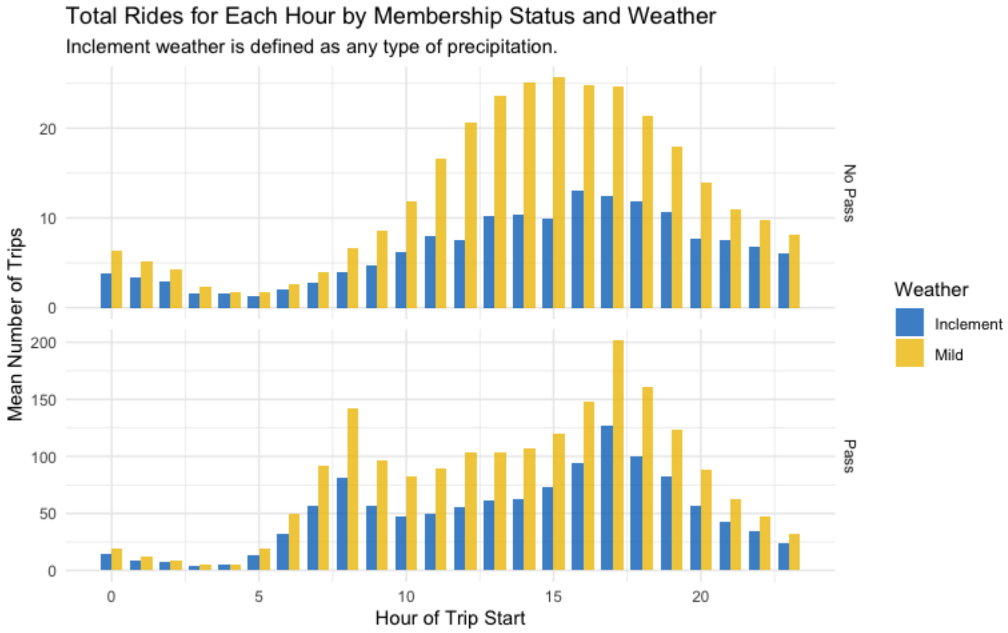

Figure 1: Mean Number of Trips by Start Time

Visual examination of ride duration (Figure 2) suggests similar responses to the time of trip regardless of membership. Passholders tended to take shorter trips than non-passholders, with the mean duration of trips being 18.9 minutes for passholders and 39.1 minutes for non-passholders.

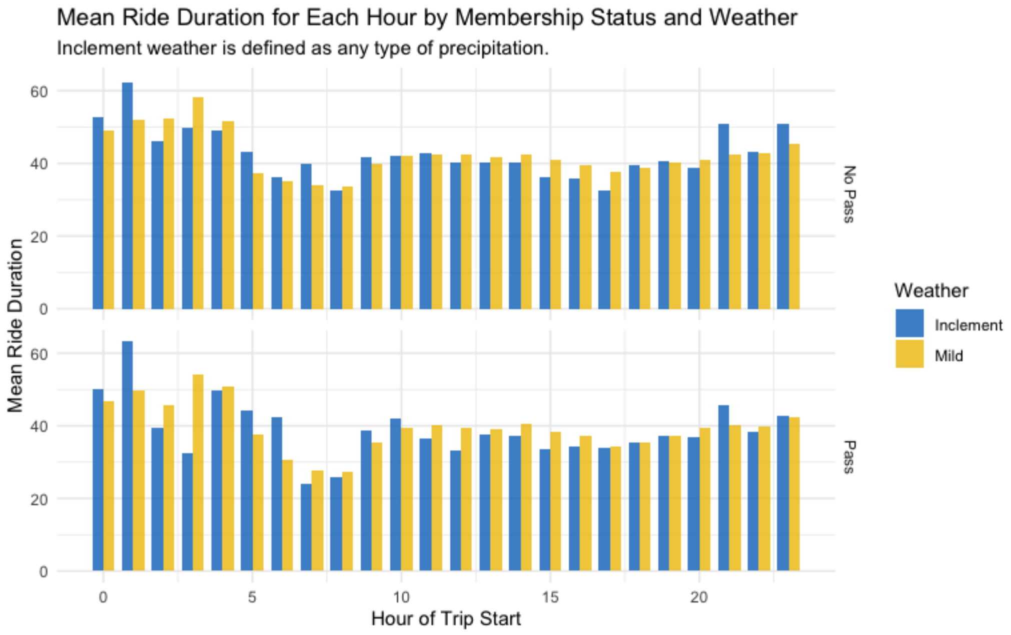

Figure 2: Mean Ride Duration by Start Time

Temperature and ride count are highly correlated on a daily scale. Figure 3 shows the unadjusted association between mean daily temperature and total ride count. These two variables have a correlation coefficient of 0.91, which suggests a strong direct relationship. The ellipse here, centered on the mean of each variable, represents the overall association between these variables. It assumes a Student's T distribution and encapsulates 95% of the values under that assumption. While it cannot be used for any formal arguments about the relationship between counts and temperature, it is a useful tool for gaining an intuition for what might be occurring here (Friendly et al., 2013).

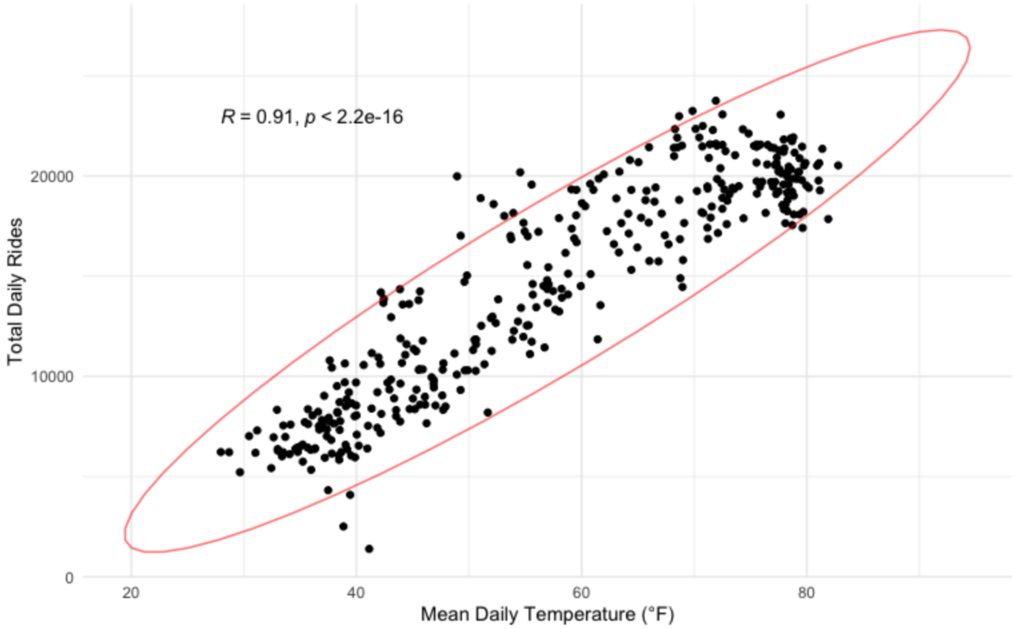

Figure 3: Scatterplot of Mean Daily Temperature Against Total Daily Rides with Accompanying Data Ellipse Representing the Association in Variance Between the Wwo

Figure 4 explores this further, showing each against the day of the year. The two curves are closely aligned, with both peaking around day 200, or July 19.

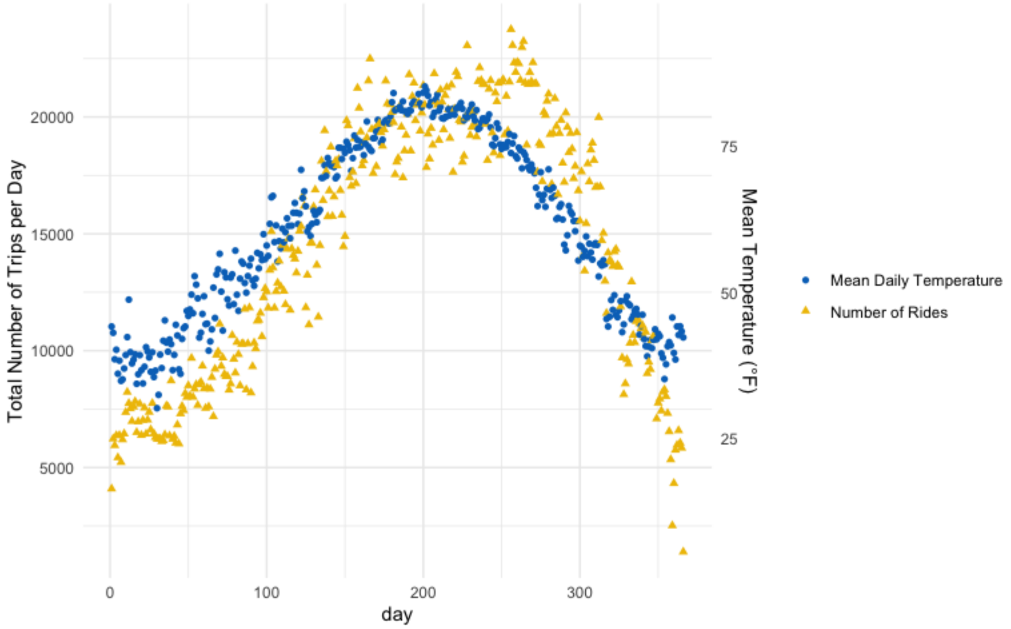

Figure 4: Mean Daily Temperature and Total Daily Number of Rides by Day of the Year

Figure 5 demonstrates the same comparison but looking at mean daily ride duration. The relationship there is weaker, with the dip in rid duration being less severe than the dip in temperature during the beginning and end of the year.

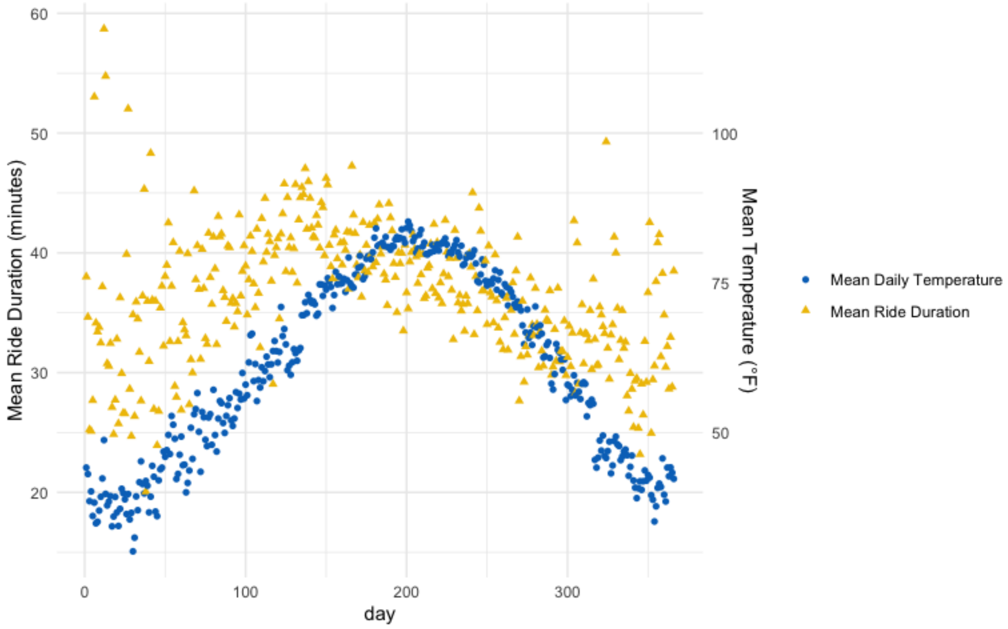

Figure 5: Mean Daily Temperature and Mean Daily Ride Duration by Day of the Year

3.2 Ride Counts

The results from the negative binomial model fit on the full dataset are reported with Incidence Rate Ratios in Table 5. Temperatures below the reference group of the 50°F-59°F range were associated with a decline in the rate of rides (IRR 0.08-0.76), while temperatures above this range were associated with an increase in the rate of rides (IRR 1.31-1.70). However, this increase dropped off slightly at and above 90°F (IRR 1.39). Windspeed was associated with a 6% decrease in ride incidence for a windspeed increase of 5 MPH. Rides fell by 54% at night. Weekends were not associated with any change in ride counts; however, peak hours were associated with a 98% increase in rides. Rain and snow were associated with decreases in ridership—19% and 41%, respectively. Other types of precipitation did not have a significant association with ridership. Humidity was associated with a 6% decrease in rides per 5 percentage-point increase in humidity.

Table 5: Negative Binomial Model with Full Data

Characteristic	IRR[1]	95% CI[1]	p-value
Temperature			<0.001[i]
50-59°F	—	—	
< 10°F	0.08	0.06, 0.11	
10-19°F	0.18	0.17, 0.19	
20-29°F	0.33	0.32, 0.34	
30-39°F	0.54	0.52, 0.55	
40-49°F	0.76	0.74, 0.77	
60-69°F	1.31	1.28, 1.33	
70-79°F	1.54	1.51, 1.57	
80-89°F	1.70	1.66, 1.75	
>= 90°F	1.39	1.32, 1.45	
Windspeed (5 MPH increments)	0.94	0.94, 0.95	<0.001[ii]
Dark	0.46	0.45, 0.46	<0.001[iii]
Weekend (Sat, Sun)	0.99	0.98, 1.00	0.121
Peak Hours (7am-9am, 4pm-7pm)	1.98	1.95, 2.02	<0.001[iv]
Precipitation			<0.001[v]
None	—	—	
Rain	0.81	0.79, 0.83	
Snow	0.59	0.55, 0.63	
Other	1.06	0.75, 1.55	
Humidity (5% increments)	0.94	0.94, 0.94	<0.001[vi]

[1] IRR = Incidence Rate Ratio, CI = Confidence Interval

[i] 1.122e-3997

[ii] 8.849e-59

[iii] 6.260e-2795

[iv] 3.033e-1859

[v] 1.370e-127

[vi] 2.643e-870

Results for the truncated models of passholder type and e-bike usage are reported in Tables 6 and 7, respectively.

Among passholders (Table 6), a similar trend with respect to temperature was seen, although it was slightly weaker. Windspeed was also similar, with a 5% decrease in ride incidence associated with a 5 MPH increase in windspeed. Darkness was associated with a 54% decrease in rides among passholders. This group had an estimated 6% decrease in ride incidence on weekends, as well as a 107% increase in rides during peak hours. The effects of precipitation among passholders were in line with the effects of precipitation overall, with rain corresponding to a 17% decrease, snow corresponding to a 41% decrease, and other precipitation associated with no change. Passholders also had an edtimated 6% reduction in rides per 5 percentage-point increase in humidity.

Table 6: Negative Binomial Model with Passholders Only

Characteristic	IRR[1]	95% CI[1]	p-value
Temperature			<0.001[i]
50-59°F	—	—	
< 10°F	0.09	0.07, 0.12	
10-19°F	0.19	0.18, 0.21	
20-29°F	0.36	0.34, 0.37	
30-39°F	0.56	0.55, 0.58	
40-49°F	0.78	0.76, 0.80	
60-69°F	1.27	1.24, 1.29	
70-79°F	1.46	1.43, 1.49	
80-89°F	1.58	1.54, 1.62	
>= 90°F	1.33	1.27, 1.39	
Windspeed (5 MPH increments)	0.95	0.94, 0.95	<0.001[ii]
Dark	0.46	0.45, 0.47	<0.001[iii]
Weekend (Sat, Sun)	0.94	0.92, 0.95	<0.001[iv]
Peak Hours (7am-9am, 4pm-7pm)	2.07	2.04, 2.10	<0.001[v]
Precipitation			<0.001[vi]
None	—	—	
Rain	0.83	0.81, 0.84	
Snow	0.59	0.55, 0.63	
Other	1.07	0.76, 1.59	
Humidity (5% increments)	0.94	0.94, 0.94	<0.001[vii]

[1] IRR = Incidence Rate Ratio, CI = Confidence Interval

[i] 8.570e-3130

[ii] 1.530e-48

[iii] 1.751e-2619

[iv] 3.493e-20

[v] 2.177e-2015

[vi] 9.796e-109

[vii] 2.019e-701

Table 7: Negative Binomial Model with No Passholders

Characteristic	IRR[1]	95% CI[1]	p-value
Temperature			<0.001[i]
50-59°F	—	—	
< 10°F	0.02	0.01, 0.05	
10-19°F	0.05	0.04, 0.06	
20-29°F	0.11	0.10, 0.12	
30-39°F	0.28	0.27, 0.29	
40-49°F	0.53	0.52, 0.55	
60-69°F	1.69	1.64, 1.74	
70-79°F	2.39	2.32, 2.46	
80-89°F	2.85	2.76, 2.95	
>= 90°F	2.00	1.88, 2.13	
Windspeed (5 MPH increments)	0.92	0.91, 0.93	<0.001[ii]
Dark	0.42	0.41, 0.43	<0.001[iii]
Weekend (Sat, Sun)	1.55	1.52, 1.58	<0.001[iv]
Peak Hours (7am-9am, 4pm-7pm)	1.10	1.08, 1.13	<0.001[v]
Precipitation			<0.001[vi]
None	—	—	
Rain	0.64	0.62, 0.66	
Snow	0.47	0.41, 0.54	
Other	0.46	0.21, 0.98	
Humidity (5% increments)	0.90	0.89, 0.90	<0.001[vii]

[1] IRR = Incidence Rate Ratio, CI = Confidence Interval

[I] 2.646e-7120

[ii] 3.726e-58

[iii] 3.676e-1663

[iv] 7.157e-470

[v] 4.767e-16

[vi] 1.607e-182

[vii] 3.198e-1278

Figure 6 shows a clear increasing trend in all 3 models for temperatures below the reference group of 50-59°F, which was more pronounced among non-passholders. Once temperatures pass the reference group, there was an upward trend until temperatures advance beyond 90°F. At that point, ridership dipped again. For these hotter temperatures, the non-passholder group was more likely to ride . In all cases, the different groups display similar responses to the given conditions, but the passholders had an attenuated effect.

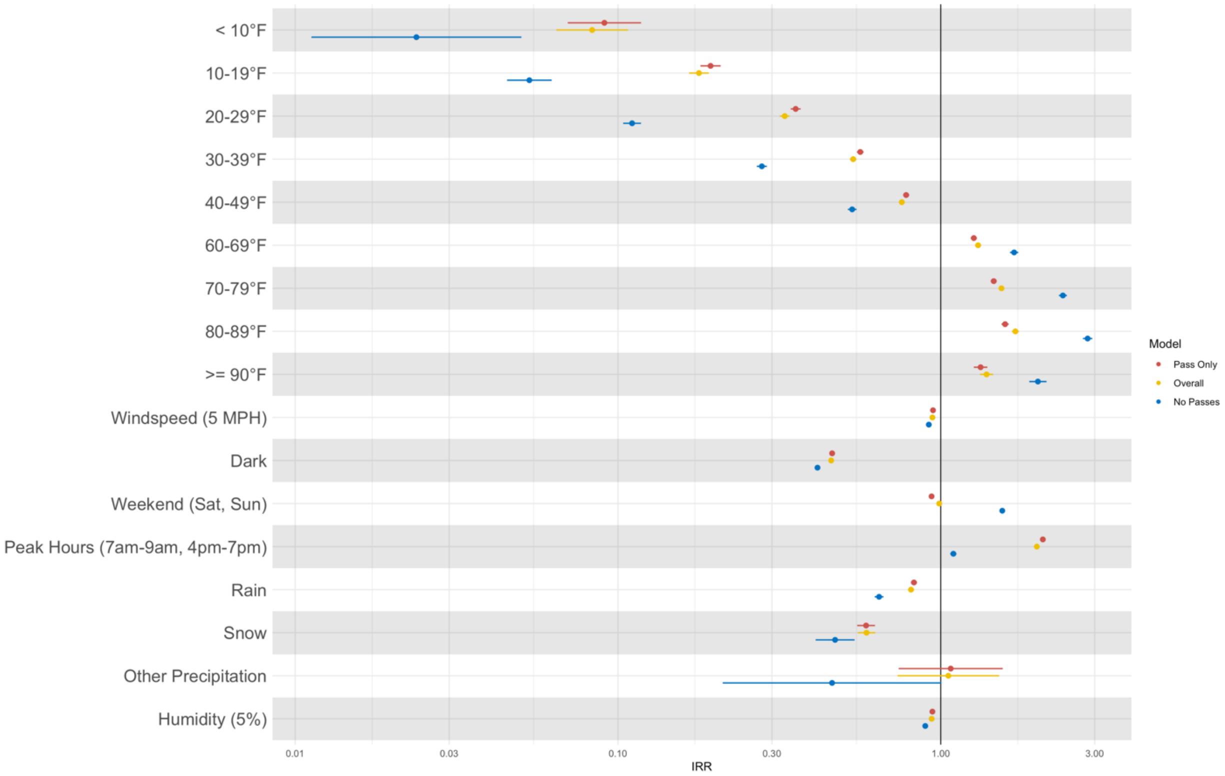

Figure 6: : Incidence Rate Ratios for Ride Count by Dataset Used

3.3 Ride Duration

Tables 9-11 display the results of the linear model with the log-transformed mean duration as the outcome. Just as with the ride counts, the first result is for the overall dataset.

Among all riders (Table 9), temperatures below the reference group of 50-59°F were associated with a decrease in ride duration (Beta -0.21,-0.09), while temperatures above were associated with an increase until temperatures reach above 90°F (Beta 0.1,0.19). At this point, the estimated ride duration was still greater than the reference group but not as much as with the 70-79°F or 80-89°F groups. Windspeed was associated with a 1% reduction in ride duration for a 5 MPH increase. Rides were 2% longer in dark conditions. Weekend rides were 7% longer, while peak-hour rides were 10% shorter. Rainy conditions were associated with 3% shorter rides. Perhaps counterintuitively, snowy conditions were associated with 12% longer rides. All other forms of precipitation were associated with 35% shorter rides. Rides were 1% shorter per 5 percentage-point increase in relative humidity.

Table 8: Linear Model with Log-Transformed Outcome with Full Data

Characteristic	Beta	95% CI[1]	p-value
Temperature			<0.001[i]
50-59°F	—	—	
< 10°F	-0.21	-0.37, -0.06	
10-19°F	-0.27	-0.32, -0.23	
20-29°F	-0.26	-0.28, -0.24	
30-39°F	-0.17	-0.19, -0.16	
40-49°F	-0.09	-0.10, -0.08	
60-69°F	0.10	0.09, 0.11	
70-79°F	0.18	0.16, 0.19	
80-89°F	0.19	0.18, 0.21	
>= 90°F	0.13	0.10, 0.16	
Windspeed (5 MPH increments)	-0.01	-0.02, -0.01	0.003
Dark	0.02	0.02, 0.03	0.008
Weekend (Sat, Sun)	0.07	0.06, 0.07	<0.001[ii]
Peak Hours (7am-9am, 4pm-7pm)	-0.10	-0.11, -0.09	<0.001[iii]
Precipitation			<0.001[iv]
None	—	—	
Rain	-0.03	-0.05, -0.02	
Snow	0.12	0.08, 0.16	
Other	-0.35	-0.58, -0.13	
Humidity (5% increments)	-0.01	-0.02, -0.01	<0.001[v]

[1] CI = Confidence Interval
[i] 2.875e-230
[ii] 2.358e-13
[iii] 3.674e-22
[iv] 7.189e-4
[vi] 3.049e-28

Among passholders, the temperature trend was similar to the overall trend. Windspeed was also the same, with a 1% decrease in duration per 5 MPH increase in windspeed. Dark conditions were associated with a 3% increase in ride length. On weekends, passholders took rides 3% longer than weekdays. Peak hours were associated with 6% shorter rides. Rain did not appear to have an effect on passholder ride duration. Snowy conditions were associated with a 10% increase in duration, while all other precipitation was associated with a 32% reduction. Again, ride duration decreased by 1% per 5 percentage-point increase in relative humidity.

Table 9: Linear Model with Log-Transformed Outcome with Passholders Only

Characteristic	Beta	95% CI[1]	p-value
Temperature			<0.001[i]
50-59°F	—	—	
< 10°F	-0.12	-0.27, 0.03	
10-19°F	-0.18	-0.22, -0.14	
20-29°F	-0.17	-0.19, -0.15	
30-39°F	-0.11	-0.12, -0.10	
40-49°F	-0.06	-0.07, -0.05	
60-69°F	0.07	0.06, 0.08	
70-79°F	0.14	0.12, 0.15	
80-89°F	0.15	0.13, 0.16	
>= 90°F	0.11	0.08, 0.14	
Windspeed (5 MPH increments)	-0.01	-0.01, -0.01	0.049
Dark	0.03	0.02, 0.04	<0.001[ii]
Weekend (Sat, Sun)	0.03	0.03, 0.04	<0.001[iii]
Peak Hours (7am-9am, 4pm-7pm)	-0.06	-0.07, -0.05	<0.001[iv]
Precipitation			0.045
None	—	—	
Rain	0.00	-0.01, 0.01	
Snow	0.10	0.06, 0.14	
Other	-0.32	-0.53, -0.11	
Humidity (5% increments)	-0.01	-0.01, -0.01	<0.001[v]

[1] CI = Confidence Interval
[i] 1.219e-112
[ii] 7.940e-4
[iii] 1.695e-4
[iv] 8.946e-9
[v] 1.485e-12

The same trends were observed for non-passholders with respect to temperature, albeit a more pronounced one. In this group, windspeed was associated with a 2% reduction in ride duration per 5 MPH increase. Dark conditions were associated with a 13% reduction in ride duration. On weekends, non-passholders took 7% longer rides. Peak conditions were associated with a 13% reduction in ride duration among non-passholders. Rainy conditions were associated with a 10% reduction in ride duration. Snowy conditions did not appear to have a definitive effect on ride duration among non-passholders, nor did other types of precipitation. A 5 percentage-point increase in relative humidity was associated with a 3% decrease in ride duration among non-passholders.

Table 10: Linear Model with Log-Transformed Outcome for Non Passholders

Characteristic	Beta	95% CI[1]	p-value
Temperature			<0.001[i]
50-59°F	—	—	
< 10°F	-0.48	-1.1, 0.16	
10-19°F	-0.50	-0.62, -0.38	
20-29°F	-0.54	-0.59, -0.49	
30-39°F	-0.33	-0.36, -0.30	
40-49°F	-0.16	-0.18, -0.13	
60-69°F	0.13	0.11, 0.16	
70-79°F	0.23	0.21, 0.25	
80-89°F	0.22	0.19, 0.24	
>= 90°F	0.13	0.08, 0.18	
Windspeed (5 MPH increments)	-0.02	-0.03, -0.01	<0.001[ii]
Dark	-0.13	-0.14, -0.11	<0.001[iii]
Weekend (Sat, Sun)	0.07	0.06, 0.09	<0.001[iv]
Peak Hours (7am-9am, 4pm-7pm)	-0.13	-0.14, -0.11	<0.001[v]
Precipitation			<0.001[vi]
None	—	—	
Rain	-0.10	-0.12, -0.07	
Snow	0.04	-0.07, 0.15	
Other	-0.03	-0.57, 0.52	
Humidity (5% increments)	-0.03	-0.03, -0.02	<0.001[vii]

[1] CI = Confidence Interval
[i] 5.917e-371
[ii] 2.565e-4
[iii] 7.426e-33
[iv] 5.759e-13
[v] 3.121e-28
[vi] 1.563e-7
[vii] 2.048e-63

Figure 7 directly compares regression coefficients across all three models. This shows that riders tended to take shorter trips in lower temperatures, although not significantly so when temperatures were below 10°F. All groups showed an increase in duration with warmer temperatures until temperatures reach 80°F or above, where duration plateaus. Duration was slightly positively associated with darkness in the overall dataset and among passholders but negatively associated with darkness among casual users. This was the one point of directional disagreement among the groups.

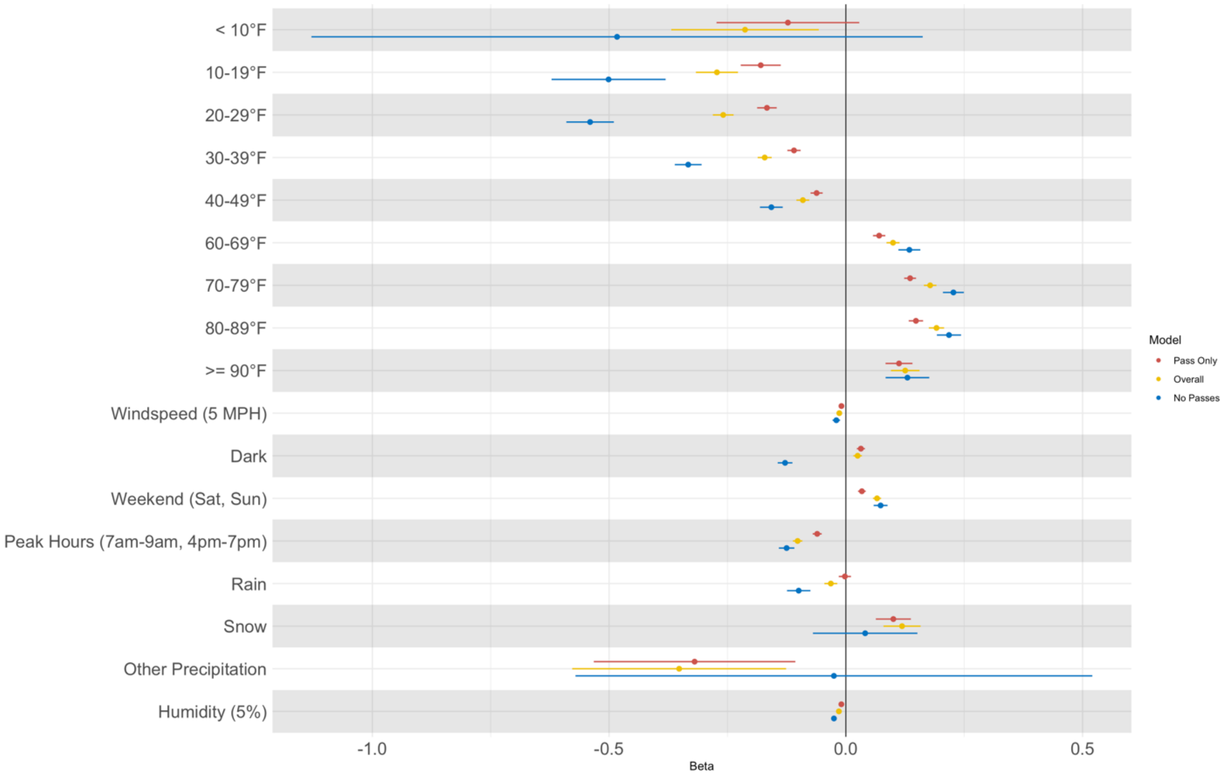

Figure 7: Estimated Coefficients from Duration Model Results by Dataset

4.0 Discussion

Rider behavior responds to weather conditions, as expected from previous studies (Bean et al., 2021; Gebhart & Noland, 2014; Miranda-Moreno & Nosal, 2011; Thomas et al., 2013). Gebhart and Noland found an association in the same direction as I've found here. However, there was reason to suspect these two cities' populations would have different responses. This analysis has found that the response from the population of Philadelphia falls roughly in line with previous studies.

Riders with passes exhibited less drastic responses to weather than riders overall, while riding behavior among non-passholders varied more with weather differences. This was especially true for ride counts. In all groups, fewer rides occurred in colder conditions and conditions with precipitation. In passholders, though, the effects were not as strong compared to casual riders.

Durations were less affected by conditions, suggesting the rides that occur in poor weather may be more utilitarian (i.e., with a fixed route). If rides in all weather types were equal parts recreational and utilitarian, the effects of weather on duration would be as pronounced and consistent as the effects on counts. The smaller effect of weather on duration suggests a greater portion of rides in inclement weather are for utilitarian purposes than in nice weather. A recreational ride may meander or be taken at a leisurely pace. A utilitarian ride has a specific destination in mind, and there is only so much variation possible in duration. The consistency of passholder ride duration suggests that riders with monthly or yearly memberships may be engaging with bike share as a form of primary transportation and not recreation in ideal conditions.

This analysis offers a preliminary result that confirms previous findings. Future work could consider geospatial aspects of Indego's bike share. Additionally, future analyses could examine the effects of e-bikes on rider behavior. There could be good reason to suspect riders would engage with bike share differently if an e-bike is available than otherwise. Electric pedal assist could help riders feel more comfortable by allowing them to reach speeds closer to surrounding traffic and reducing the required exertion for a given trip. It can also offer as simple a benefit as reducing the amount of time the trip takes. If we suppose a given rider has predetermined thresholds for convenience and comfort required to use bike share for a trip, any of these stated benefits could help meet these thresholds. Even if the current weather may reduce the convenience or comfort of taking a bike for a particular trip, electric pedal assist may still make that trip worth taking.

Overall, it appears that passholders were more likely to engage with the bike share in poor conditions. This could suggest a selection bias among the users who have monthly and yearly passes; however, more research is needed to determine whether that is the case. It is also possible that riders with a monthly or yearly pass were simply responding to the fact that they can take unlimited free rides of up to 60 minutes at any point without having to update or renew an already active pass. The cost of activating a day pass may lead potential users to reconsider whether they would want to use the bike share on a given day based on the current or expected weather conditions. There is also reason to suspect that rider response to rain has more to do with the available infrastructure (Goldmann & Wessel, 2021). This may suggest that these results are less an indication of eagerness on the part of passholders to ride regardless of weather and more a manifestation of hesitancy on the part of casual users.

Future analyses could attempt to determine the exact nature and cause of these behavioral responses to provide a better understanding of barriers to modal shifts for cycling.

5.0 Conclusion

This analysis supports previous findings from Washington, D.C. that bike share users are sensitive to current weather conditions, with increases in the number of and duration of rides in more moderate conditions. Policymakers and public health officials should take into account the barrier weather conditions pose to the widespread adoption of bicycles as a mode of transportation in the United States.

In the city of Philadelphia, bike share users were sensitive to inclement weather. Riders who maintain a consistent membership with the system were less likely to alter their behavior, though. Instituting programs that encourage riders to sign up for regular memberships could play a pivotal role in increasing the modal share of bikes in the city. Shifting trips from personal cars to bicycles is crucial to addressing several public health concerns faced by the city.

Appendix A : R Code

In this analysis, I had a script written to build all the datasets, and I used three R Markdown files to run the separate analysis of each dataset.

Building the dataset:

```
## code to prepare `DATASET` dataset goes here
library(tidyverse)


# read in csvs containing trip data
trip_files <- paste0("data-raw/",
                     stringr::str_subset(list.files(
                       "data-raw"
                       ),
                                         "indego-trips"
                                         )
                     )


trips_list <- lapply(trip_files, read_csv)


for( i in c(1,2,3,4,5,15,17,18,19,20,21,22,23,24,25,26,27,28)) {
  trips_list[[i]]$start_time <- trips_list[[i]]$start_time %>% lubridate::mdy_hm()
  trips_list[[i]]$end_time <- trips_list[[i]]$end_time %>% lubridate::mdy_hm()
}
```

```
for( i in 1:length(trips_list)){
  trips_list[[i]]$bike_id <- trips_list[[i]]$bike_id %>% as.character()
  trips_list[[i]]$start_lat <- trips_list[[i]]$start_lat %>% as.double()
  trips_list[[i]]$start_lon <- trips_list[[i]]$start_lon %>% as.double()
  trips_list[[i]]$end_lat <- trips_list[[i]]$end_lat %>% as.double()
  trips_list[[i]]$end_lon <- trips_list[[i]]$end_lon %>% as.double()
  }

rm(i)

trips <- bind_rows(trips_list)



trips <- trips %>%
  # recompute duration to ensure common unites (minutes)
  mutate(duration = as.double(end_time - start_time)/60,
         trip_date = as.Date(start_time)) %>%
  # filter out rides where the bike was returned or the trip was greater than 24 hrs (1440 minutes)
  dplyr::filter(
    !is.na(end_lat),
    !is.na(end_lon),
    !is.na(end_time),
    duration <= 1440,
    duration >= 1,
    !is.na(start_time)) %>%
  #remove any duplicate trips, such as the month of september 2021
  distinct() %>%
  mutate(
    start_station = ifelse(is.na(start_station), start_station_id, start_station),
    end_station = ifelse(is.na(end_station), end_station_id, end_station),
    bike_type = ifelse(is.na(bike_type), "standard", bike_type)) %>%
  dplyr::select(
    -start_station_id,
    -end_station_id
    )


weather_files <- paste0("data-raw/",
                      stringr::str_subset(list.files(
                        "data-raw"
                      ),
                      "Philadelphia"
                      )
)
```

```
weather_list <- lapply(weather_files, read_csv)

weather <- bind_rows(weather_list)

rm(weather_list)
rm(weather_files)

weather <-weather %>% transmute(
  datetime = lubridate::with_tz(datetime, Sys.timezone()),
  temp,
  humidity,
  preciptype,
  windspeed,
  visibility,
  solarenergy
)

trips <- trips %>% transmute(
  trip_id,
  duration,
  start_time = lubridate::with_tz(start_time, Sys.timezone()),
  plan_duration,
  passholder_type,
  bike_type,
  trip_route_category,
  trip_date = lubridate::floor_date(start_time, unit = "hour"),
  pass_ind = ifelse(passholder_type %in% c("Indego30", "IndegoFlex", "Indego365"), 1, 0) %>%
    factor(levels = c("0", "1"), labels = c("No Pass", "Pass"))
)

trip_data <- full_join(trips, weather, by = c("trip_date" = "datetime"))

full_data <-
  trip_data %>%
  group_by(trip_date) %>%
  summarise(count = n(), mean_duration = mean(duration)) %>%
    mutate(count = ifelse(is.na(mean_duration), 0, count)) %>%
  full_join(trip_data, by = "trip_date")

full_data <- full_data %>%
  group_by(trip_date, bike_type) %>%
  summarise(electric_prop = n()) %>%
  pivot_wider(names_from = "bike_type", values_from = "electric_prop") %>%
  mutate(electric = ifelse(is.na(electric), 0, electric)) %>%
  dplyr::select(trip_date, electric) %>%
  full_join(full_data, by = "trip_date") %>%
  mutate(electric = electric/count)

full_data <- full_data %>%
  group_by(trip_date, pass_ind) %>%
  summarise(pass = n()) %>%
  pivot_wider(names_from = "pass_ind", values_from = "pass") %>%
  mutate(pass = ifelse(is.na(Pass), 0, Pass)) %>%
  dplyr::select(trip_date, pass) %>%
  full_join(full_data, by = "trip_date") %>%
  mutate(pass = pass/count)

pass_data <-
  trip_data %>% dplyr::filter(pass_ind == "Pass") %>%
  group_by(trip_date) %>%
  summarise(count = n(), mean_duration = mean(duration)) %>%
  mutate(count = ifelse(is.na(mean_duration), 0, count)) %>%
  full_join(trip_data, by = "trip_date")

pass_data <- pass_data %>% dplyr::filter(pass_ind == "Pass") %>%
    group_by(trip_date, bike_type) %>%
    summarise(electric_prop = n()) %>%
    pivot_wider(names_from = "bike_type", values_from = "electric_prop") %>%
    mutate(electric = ifelse(is.na(electric), 0, electric)) %>%
    dplyr::select(trip_date, electric) %>%
    full_join(pass_data, by = "trip_date") %>%
    mutate(electric = electric/count)

no_pass_data <-
    trip_data %>% dplyr::filter(pass_ind == "No Pass") %>%
  group_by(trip_date) %>%
    summarise(count = n(), mean_duration = mean(duration)) %>%
    mutate(count = ifelse(is.na(mean_duration), 0, count)) %>%
    full_join(trip_data, by = "trip_date")

no_pass_data <- no_pass_data %>% dplyr::filter(pass_ind == "No Pass") %>%
    group_by(trip_date, bike_type) %>%
    summarise(electric_prop = n()) %>%
    pivot_wider(names_from = "bike_type", values_from = "electric_prop") %>%
    mutate(electric = ifelse(is.na(electric), 0, electric)) %>%
    dplyr::select(trip_date, electric) %>%
    full_join(no_pass_data, by = "trip_date") %>%
    mutate(electric = electric/count)

electric_data <-
    trip_data %>% dplyr::filter(bike_type == "electric") %>%
  group_by(trip_date) %>%

    summarise(count = n(), mean_duration = mean(duration)) %>%

    mutate(count = ifelse(is.na(mean_duration), 0, count)) %>%

    full_join(trip_data, by = "trip_date")


  electric_data <- electric_data %>% dplyr::filter(bike_type == "electric") %>%

    group_by(trip_date, pass_ind) %>%

    summarise(pass = n()) %>%

    pivot_wider(names_from = "pass_ind", values_from = "pass") %>%

    mutate(pass = ifelse(is.na(Pass), 0, Pass)) %>%

    dplyr::select(trip_date, pass) %>%

    full_join(electric_data, by = "trip_date") %>%

    mutate(pass = pass/count)


standard_data <-

    trip_data %>% dplyr::filter(bike_type == "standard") %>%

  group_by(trip_date) %>%

    summarise(count = n(), mean_duration = mean(duration)) %>%

    mutate(count = ifelse(is.na(mean_duration), 0, count)) %>%

    full_join(trip_data, by = "trip_date")


  standard_data <- standard_data %>% dplyr::filter(bike_type == "standard") %>%

    group_by(trip_date, pass_ind) %>%
  summarise(pass = n()) %>%
  pivot_wider(names_from = "pass_ind", values_from = "pass") %>%
  mutate(pass = ifelse(is.na(Pass), 0, Pass)) %>%
  dplyr::select(trip_date, pass) %>%
  full_join(standard_data, by = "trip_date") %>%
  mutate(pass = pass/count)


write_csv(full_data, "data/trip_data.csv")
write_csv(pass_data, "data/pass_data.csv")
write_csv(no_pass_data, "data/no_pass_data.csv")
write_csv(electric_data, "data/electric_data.csv")
write_csv(standard_data, "data/standard_data.csv")


rm(list = ls())
```

Analysing the overall dataset:

~~~
        ```{r setup, include=FALSE}
        knitr::opts_chunk$set(echo = TRUE, fig.width = 8, fig.height = 5)
        options(tidyverse.quiet = TRUE, tidymodels.quiet = TRUE)
        ~~~

        ```{r packages}
        pacman::p_load(here,         # file locator
                  tidyverse,    # data management + ggplot2 graphics
```

~~~
        MASS,        # for negative binomial modeling
        skimr,       # for initial data inspection
        mice,        # for missing data inspection
        lubridate,   # for date handling
        table1,      # for constructing summary table
        flextable,   # for formatting tables to export to word
        gtsummary,   # for constructing model summary tables
        Rmpfr,       # for adjusting the smallest possible reported values
        ggsci,       # for visualization color palette
        ggpubr,      # for correlations in one of the visualizations
        ggforestplot  # for creating forest plots of
        )
```

```{r external functions}
source("R/indego_palette.R")
source("R/functions.R")

palette <- pal_jco()(5)
```

# Dataset
~~~

In this analysis, I am using data for rides taken from January 1, 2016 (the first full year the program was available)and December 31, 2022 (the last full year before this analysis) collected from the Indego bikeshare platform in Philadelphia. I've connected this data with hourly historical weather data obtained from visual crossing.

The completed dataset includes the following variables:

- trip_id: A unique identifier of the ride

- duration: the number of minutes the ride took. Values larger than 1440 (24 hrs) are removed.

- start_time: the date, hour, and minute the ride began

- plan_duration: the number of days the plan used to book the ride was active for

- passholder_type: the name of the plan the rider used. Rides without a plan are coded "walkup"

- bike_type: either standard or electric, the electric bikes were introduced to the system in 2018

- trip_date: the hour of the ride. Used to match to weather data. Also used to aggregate individual ride data into ride counts

- temp: the temperature during that hour in Philadelphia

- humidity: the humidity during that hour in Philadelphia

- preciptype: the form of precipitation during that hour in Philadelphia. NA indicates no precipitation. *Need to recode missing values to reflect that.*

- windspeed: the windspeed during that hour in Philadelphia

- solarenergy: the sunlight available during that hour in Philadelphia. NA values provide a good indication of night time. That last sentence isn't actually true. Need to work out the best way to recode here.

For this initial analysis, I'll randomly sample 10% of the data.

~~~
```{r read data}
set.seed(19103)

data_sample <- read_csv(here("data", "trip_data.csv"))
~~~

```
data <- read_csv(here("data", "trip_data.csv")) %>% select(trip_date, count,
mean_duration, electric, pass, temp, humidity, preciptype, windspeed,visibility, solarenergy)
%>% distinct(trip_date, .keep_all = TRUE)
```

```
data <- data %>% mutate(
  count = ifelse(is.na(mean_duration), 0, count),
  preciptype = ifelse(is.na(preciptype), "None",
                      ifelse(preciptype == "rain", preciptype,
                             ifelse(str_detect(preciptype, "snow"), "snow", "Other"))) %>%
    factor(
      levels = c("None", "rain", "snow", "Other"),
      labels = c("None", "Rain", "Snow", "Other")) %>% relevel(ref = "None"),
  solarenergy = ifelse(is.na(solarenergy), 0, solarenergy),
  daylight = ifelse(solarenergy == 0, "Night", "Day"),
  electric = ifelse(is.na(electric), 0, electric),
  year = year(trip_date),
  quarter = quarter(trip_date),
  month = month(trip_date),
  week = week(trip_date),
  day = yday(trip_date),
  hour = hour(trip_date),
  weekend = ifelse(wday(trip_date) > 5, "Weekend", "Weekday") %>% factor(),
  peak = ifelse(hour < 7, "Off-Peak",
                ifelse(hour < 9, "Peak",
                       ifelse(hour < 16, "Off-Peak",
                              ifelse(hour < 19, "Peak", "Off-Peak")))) %>%
    factor(),
  temp_discrete =
    ifelse(temp <= 9, "Temperature 10",
          ifelse(temp <= 19, "Temperature 20",
                 ifelse(temp <= 29, "Temperature 30",
                        ifelse(temp <= 39, "Temperature 40",
                               ifelse(temp <= 49, "Temperature 50",
                                      ifelse(temp <= 59, "Temperature 60",
                                             ifelse(temp <= 69, "Temperature 70",
                                                    ifelse(temp <= 79, "Temperature 80",
                                                           ifelse(temp <= 89, "Temperature 90",
                                                                  "Temperature 100")
                                                           )
                                                    )
                                             )
                                      )
                               )
                      )
                  )
              ) %>%
        factor(
          levels = c(
            "Temperature 10",
            "Temperature 20",
            "Temperature 30",
            "Temperature 40",
            "Temperature 50",
            "Temperature 60",
            "Temperature 70",
            "Temperature 80",
            "Temperature 90",
            "Temperature 100"),
          labels = c(
            "< 10°F",
            "10-19°F",
            "20-29°F",
            "30-39°F",
            "40-49°F",
            "50-59°F",
            "60-69°F",
        "70-79°F",
        "80-89°F",
        ">= 90°F"
        )
      ) %>%
    relevel(ref = "50-59°F"),
   covid = ifelse(trip_date > as.Date("2020-03-22"), 1, 0),
   windy = ifelse(windspeed > quantile(windspeed, 0.1), 1, 0) %>% factor(levels = c("0", "1"), labels = c("calm", "windy"))
   )

summary_data <- data_sample %>% select(
  temp,
  trip_date,
  solarenergy,
  preciptype,
  duration
) %>%
  mutate(
  preciptype = ifelse(is.na(preciptype), "None",
                      ifelse(preciptype == "rain", preciptype,
                             ifelse(str_detect(preciptype, "snow"), "snow", "Other"))) %>%
    factor(
    levels = c("None", "rain", "snow", "Other"),

    labels = c("None", "Rain", "Snow", "Other")) %>% relevel(ref = "None"),

solarenergy = ifelse(is.na(solarenergy), 0, solarenergy),

daylight = ifelse(solarenergy == 0, "Night", "Day"),

year = year(trip_date),

quarter = quarter(trip_date),

month = month(trip_date),

week = week(trip_date),

day = yday(trip_date),

hour = hour(trip_date),

weekend = ifelse(wday(trip_date) > 5, "Weekend", "Weekday") %>% factor(),

peak = ifelse(hour < 7, "Off-Peak",

              ifelse(hour < 9, "Peak",

                     ifelse(hour < 16, "Off-Peak",

                            ifelse(hour < 19, "Peak", "Off-Peak")))) %>%

  factor(),

temp_discrete =

  ifelse(temp <= 9, "Temperature 10",

        ifelse(temp <= 19, "Temperature 20",

               ifelse(temp <= 29, "Temperature 30",

                      ifelse(temp <= 39, "Temperature 40",

                             ifelse(temp <= 49, "Temperature 50",

                                    ifelse(temp <= 59, "Temperature 60",
                                         ifelse(temp <= 69, "Temperature 70",
                                                ifelse(temp <= 79, "Temperature 80",
                                                       ifelse(temp <= 89, "Temperature 90",
                                                              "Temperature 100")
                                                       )
                                                )
                                         )
                                  )
                           )
                    )
             )
      ) %>%
factor(
  levels = c(
    "Temperature 10",
    "Temperature 20",
    "Temperature 30",
    "Temperature 40",
    "Temperature 50",
    "Temperature 60",
    "Temperature 70",
    "Temperature 80",
    "Temperature 90",
    "Temperature 100"),
   labels = c(
    "< 10°F",
    "10-19°F",
    "20-29°F",
    "30-39°F",
    "40-49°F",
    "50-59°F",
    "60-69°F",
    "70-79°F",
    "80-89°F",
    ">= 90°F"
    )
   ) %>%
  relevel(ref = "50-59°F"))

glimpse(data)

skim(data)

glimpse(data)

skim(data)
```

~~~~
```

## Checking for Missing Data

```{r missing plot, eval = F}
md.pattern(data, rotate.names = TRUE)
```

For these missing values, the missingness is important. It either
suggests there were favorable weather conditions, such as in preciptype,
or perhaps a nighttime duration as in solarenergy. It also suggests there were no rides in a time period, thus no e-bikes used.

# Descriptive Statistics

## Univariate

I'll begin by looking at the distributions of the outcome variables.
This analysis looks at two models, one for each outcome variable.
~~~~

~~~
### Ride Count

```{r count graphical}
data %>%
  ggplot(aes(x = count)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = pal_jco()(1),
                 alpha = 0.8) +
  labs(title = "Histogram of Rides per Hour",
       x = "Number of Rides Taken that Hour") +
  theme_minimal()

data_sample %>%
  mutate(
    Weather = ifelse(is.na(preciptype),
                     "Mild",
                     "Inclement"),
    pass_ind = ifelse(is.na(pass_ind),
                      "No Pass",
                      pass_ind),
    weekend = ifelse(wday(trip_date) > 5, "Weekend", "Weekday"),
~~~

~~~
             ) %>%
 group_by(trip_date, pass_ind, Weather, weekend) %>%
 summarize(count = n()) %>%
 mutate(Hour = hour(trip_date),
        peak = ifelse(Hour < 7, "Off-Peak",
         ifelse(Hour <= 9, "Peak",
             ifelse(Hour < 16, "Off-Peak",
                 ifelse(Hour <= 19, "Peak", "Off-Peak")))) %>%
  factor()) %>%
 group_by(Hour, pass_ind, Weather, weekend, peak) %>%
 summarize(mean = mean(count, na.rm = T)) %>%
 ggplot(aes(x = Hour, y = mean, fill = Weather)) +
 geom_col(position = "dodge", alpha = 0.8, width = 0.75) +
 facet_grid(pass_ind ~ weekend, scales = "free_y") +
 scale_fill_jco() +
 theme_minimal() +
 labs(title = "Total Rides for Each Hour by Membership Status and Weather",
    subtitle = "Inclement weather is defined as any type of precipitation.",
    x = "Hour of Trip Start",
    y = "Mean Number of Trips")
```
~~~

~~~
```{r}

data_sample %>%
  mutate(
    Weather = ifelse(is.na(preciptype),
                     "Mild",
                     "Inclement"),
    pass_ind = ifelse(is.na(pass_ind),
                      "No Pass",
                      pass_ind),
    Hour = hour(trip_date),
     weekend = ifelse(wday(trip_date) > 5, 1, 0)) %>%
  group_by(pass_ind, Weather, weekend) %>%
  summarize(mean = mean(mean_duration, na.rm = T))

data_sample %>%
  mutate(
    Weather = ifelse(is.na(preciptype),
                     "Mild",
                     "Inclement"),
    pass_ind = ifelse(is.na(pass_ind),
                      "No Pass",
                      pass_ind),
~~~

```
    Hour = hour(trip_date),
     weekend = ifelse(wday(trip_date) > 5, 1, 0),
    peak = ifelse(Hour < 7, "Off-Peak",
                 ifelse(Hour < 9, "Peak",
                        ifelse(Hour < 16, "Off-Peak",
                               ifelse(Hour < 19, "Peak", "Off-Peak"))))
    ) %>%
  group_by(Hour, pass_ind, Weather, weekend, peak) %>%
  summarize(mean = mean(mean_duration, na.rm = T)) %>%
  ggplot(aes(x = Hour, y = mean, fill = Weather)) +
  geom_col(position = "dodge", alpha = 0.8, width = 0.75) +
  facet_grid(pass_ind ~ weekend) +
  scale_fill_jco() +
  theme_minimal() +
  labs(title = "Mean Ride Duration for Each Hour by Membership Status and Weather",
       subtitle = "Inclement weather is defined as any type of precipitation.",
       x = "Hour of Trip Start",
       y = "Mean Ride Duration")
```

```{r}
data %>% ggplot(aes(x = windspeed, y = count)) + geom_point()
```

```
data_sample %>%
  mutate(day = yday(trip_date)) %>%
  group_by(day) %>%
  summarize("Number of Rides" = n(),
            "Mean Daily Temperature" = mean(temp, na.rm = T)/0.004) %>%
  pivot_longer(!day,
               names_to = "measure",
               values_to = "value") %>%
  ggplot(aes(x = day, y = value, color = measure, shape = measure)) +
  geom_point() +
  scale_y_continuous(
    name = "Total Number of Trips per Day",
    sec.axis = sec_axis(~.*.004, name = "Mean Temperature (°F)")) +
  scale_color_jco() +
  theme_minimal() +
  theme(legend.title = element_blank())

data_sample %>%
  mutate(day = yday(trip_date)) %>%
  group_by(day) %>%
  summarize("Mean Ride Duration" = mean(mean_duration, na.rm = T),
            "Mean Daily Temperature" = mean(temp, na.rm = T)*0.5) %>%
  pivot_longer(!day,
               names_to = "measure",
               values_to = "value") %>%
  ggplot(aes(x = day, y = value, color = measure, shape = measure)) +
  geom_point() +
  scale_y_continuous(
    name = "Mean Ride Duration (minutes)",
    sec.axis = sec_axis(~./0.5, name = "Mean Temperature (°F)")) +
  scale_color_jco() +
  theme_minimal() +
  theme(legend.title = element_blank())

data %>%
  group_by(day) %>%
  summarize(count = sum(count), temp = mean(windspeed, na.rm = T)) %>%
  ggplot(aes(x = temp, y = count)) + geom_point() + stat_ellipse(color = "red", alpha = 0.6) + stat_cor() + theme_minimal() +
  labs(y = "Total Daily Rides", x = "Mean Daily Temperature (°F)")
```

### Ride Duration

```{r duration numerical}
label(data$trip_date) <- "Time"
```

```
    label(data$count) <- "Hourly Count of Rides"

    label(data$mean_duration) <- "Hourly Mean Duration of Rides"

    label(data$temp) <- "Temperature"

    label(data$humidity) <- "Humidity"

    label(data$preciptype) <- "Precipitation"

    label(data$windspeed) <- "Windspeed"

    label(data$daylight) <- "Dark"

    label(data$pass) <- "Rides Taken with Pass"

    label(data$temp_discrete) <- "Temperature"

    label(data$weekend) <- "Weekend"

    label(data$peak) <- "Peak Hours"


    units(data$trip_date) <- "hours"

    units(data$mean_duration) <- "minutes"

    units(data$temp) <- "°F"

    units(data$humidity) <- "%"

    units(data$windspeed) <- "MPH"


    table1(~ temp + temp_discrete + humidity + windspeed + daylight + preciptype + weekend

+ peak, data, digits = 4) %>%

     t1flex()
```

~~~~
table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + daylight + preciptype, data = data) %>% t1flex()


data %>% summarise(mean(humidity), sd(humidity), min(humidity), max(humidity))
data %>% summarise(mean(temp), sd(temp), min(temp), max(temp))
data %>% summarise(mean(windspeed), sd(windspeed), min(windspeed), median(windspeed), max(windspeed))
data %>% summarize(mean(count), sd(count), min(count), max(count))
data %>% summarize(mean(mean_duration, na.rm = T), sd(mean_duration, na.rm = T), min(mean_duration, na.rm = T), max(mean_duration, na.rm = T))


summary_data %>% group_by(temp_discrete) %>% summarize(n())
summary_data %>% group_by(daylight) %>% summarize(n())
summary_data %>% group_by(weekend) %>% summarize(n = n(), n/nrow(summary_data))
summary_data %>% group_by(peak) %>% summarize(n = n(), n/nrow(summary_data))
summary_data %>% group_by(preciptype) %>% summarize(n())


summary_data %>% table1(~ temp_discrete + daylight + weekend + peak + preciptype, .) %>% t1flex()
```
~~~~

Precipitation types are too small outside of none, rain, and snow. I'll collapse the rest into "other" and code the cases of "rain, snow" as just snow. This will come in a bit.

~~~
```{r}
data %>%
  ggplot(aes(x = mean_duration)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  labs(title = "Histogram of Hourly Mean Ride Duration",
       x = "Mean Duration (minutes)") +
  theme_minimal()

data %>%
  ggplot(aes(x = log(mean_duration)))+
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  labs(title = "Histogram of Hourly Mean Ride Duration",
       x = "Mean Duration (minutes)") +
~~~

~~~
  theme_minimal()



data %>% summarise(mean = mean(log(mean_duration), na.rm = T),
                   sd = sd(log(mean_duration), na.rm = T))
```
~~~

The duration could be modeled as a log-normal linear regression, or it could be treated as just a linear regression.

~~~
```{r}
```
~~~

### Independent Variables

~~~
```{r passholder numeric}

```
~~~

There's a lot of variation in the passholder type, largely because the pass structure system has changed over the years. There are some threads that have always been here, though. Day passes (one or two day) have always been offered, even if they appear differently depending on the year. Walkups and NULL are the same type of pass. The monthly , yearly, and flex passes have been largely unchanged. So I'll group passholder_type under a new variable, pass_type, based on these categories.

~~~
```{r}
data %>% ggplot(aes(x = pass)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  theme_minimal()
```
~~~

~~~
```{r bike numeric}
data %>% ggplot(aes(x = electric)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  theme_minimal()
```
~~~

~~~
```{r precip numeric}



data %>% ggplot(aes(x = preciptype)) +
  geom_bar(color = "black",
           fill = "navy",
           alpha = 0.6) +
  theme_minimal()
~~~

```{r continuous numeric}
data_cont_long <- data %>%
  pivot_longer(
    cols = c("temp", "humidity", "windspeed", "solarenergy"),
    names_to = "variable",
    values_to = "measure"
    )

data_cont_long %>%
  ggplot(aes(x = measure)) +
  geom_histogram(bins = 100,
                 color = "black",
```

~~~
                 fill = "navy",
                 alpha = 0.6) +
  facet_grid(cols = vars(variable), scales = "free") +
  theme_minimal()
```


```{r temp graphic}
data %>% ggplot(aes(y = temp, x = month, group = month)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Hourly Temperatures by Month")
```



```{r humidity graphic}
data %>% ggplot(aes(y = humidity, x = hour, group = hour)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
~~~

~~~
  theme_minimal() +
  labs(title = "Plot of Hourly Humidity by Time of Day")
```


```{r windspeed graphic}
data %>% ggplot(aes(y = windspeed, x = hour, group = hour)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Hourly Windspeed by Time of Day")
```


```{r solar graphic}
data %>% ggplot(aes(y = solarenergy, x = month, group = month)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Solar Energy by Month")
~~~

~~~
data %>% ggplot(aes(y = daylight, x = hour, group = hour)) +
  geom_jitter(color = "black", alpha = 0.3) +
  theme_minimal() +
  labs(title = "Plot of Daylight Classification by Hour")
```
~~~

Clearly using solar energy as a proxy for day/night classification is imperfect, but the overwhelming majority of cases appear to match based on the time of day.

## Bivariate Descriptions/Preliminary Analysis

I want to do a descriptive for each of my outcomes of interest against
each independent variable.

Let's start with ride count.

~~~
```{r counttables}
table1(~ count, data = data) %>% t1flex()
table1(~ count | temp_discrete, data = data) %>% t1flex()
table1(~ count | daylight, data = data) %>% t1flex()
table1(~ count | weekend, data = data) %>% t1flex()
table1(~ count | peak, data = data) %>% t1flex()
table1(~ count | preciptype, data = data) %>% t1flex()
```
~~~

~~~~
```{r meandurationtables}
table1(~ mean_duration, data = data) %>% t1flex()
table1(~ mean_duration | temp_discrete, data = data) %>% t1flex()
table1(~ mean_duration | daylight, data = data) %>% t1flex()
table1(~ mean_duration | weekend, data = data) %>% t1flex()
table1(~ mean_duration | peak, data = data) %>% t1flex()
table1(~ mean_duration | preciptype, data = data) %>% t1flex()
```
~~~~

~~~~
```{r}
data %>% ggplot(aes(x = temp_discrete, y = mean_duration)) + geom_col()
```
~~~~

There's a degree of seasonality in this data that is important to keep
in mind.

~~~~
```{r}

data %>% ggplot(aes(x = day, y = temp)) + geom_point(alpha = 0.25) + theme_minimal()

data %>% filter(daylight == "Day") %>%
~~~~

~~~
  group_by(trip_date) %>%
  summarize("Hourly Rides" = count,
            "Hourly Temperature" = temp) %>%
  pivot_longer(cols = c("Hourly Rides", "Hourly Temperature"),
               names_to = "measure",
               values_to = "value") %>%
  ggplot(aes(x = trip_date, y = value, color = measure)) +
  geom_point(alpha = 0.8) +
  theme_minimal()

data %>% filter(daylight == "Day") %>%
  group_by(trip_date) %>%
  summarize("Hourly Mean Duration" = mean_duration,
            "Hourly Temperature" = temp) %>%
  pivot_longer(cols = c("Hourly Mean Duration", "Hourly Temperature"),
               names_to = "measure",
               values_to = "value") %>%
  ggplot(aes(x = trip_date, y = value, color = measure)) +
  geom_point(alpha = 0.8) +
  theme_minimal()

```
~~~

~~~
# Model

## Negative Binomial

```{r}
nb_model <- glm.nb(count ~ temp_discrete + I(windspeed/5) + daylight + weekend + peak + preciptype + I(humidity/5), data = data)

pois_model <- glm(count ~ temp_discrete + I(windspeed/5) + daylight + weekend + peak + preciptype + I(humidity/5), data = data, family = poisson(link = "log"))

duration_model <- lm(log(mean_duration) ~ temp_discrete + I(windspeed/5) + daylight + weekend + peak + preciptype + I(humidity/5), data = data)
~~~

```{r}
plot(nb_model)

plot(duration_model)
```
```{r}
```

```
nb_model %>% summary()
duration_model %>% summary()


full_nb_table <- nb_model %>%
  tbl_regression(exponentiate = T,
                 intercept = F, show_single_row = c('weekend', 'daylight', 'peak'),
                 label = c(
                   temp_discrete ~ "Temperature",
                   daylight ~ "Dark",
                   peak ~ "Peak Hours (7am-9am, 4pm-7pm)",
                   weekend ~ "Weekend (Sat, Sun)",
                   preciptype ~ "Precipitation",
                   `I(windspeed/5)` ~ "Windspeed (5 MPH increments)",
                   `I(humidity/5)` ~ "Humidity (5% increments)"
                 )) %>%
  add_global_p() %>%
  as_flex_table()


full_duration_table <- duration_model %>%
  tbl_regression(intercept = F, show_single_row = c('weekend', 'daylight', 'peak'),
                 label = c(
                   temp_discrete ~ "Temperature",
                   daylight ~ "Dark",
```

~~~~
                  peak ~ "Peak Hours (7am-9am, 4pm-7pm)",
                  weekend ~ "Weekend (Sat, Sun)",
                  preciptype ~ "Precipitation",
                  `I(windspeed/5)` ~ "Windspeed (5 MPH increments)",
                  `I(humidity/5)` ~ "Humidity (5% increments)"
                )) %>%
  add_global_p() %>%
  as_flex_table()

full_nb_table %>% save_as_docx(path = here("files", "full_nb_results.docx"))
full_duration_table %>% save_as_docx(path = here("files", "full_duration_results.docx"))

full_nb_table
full_duration_table
```



```{r}
source(here("R", "global_p.R"))

global_p(nb_model)
# 2 'mpfr' numbers of precision  200   bits
# [1] 2.0298465195466694380461698263104467695985404891415345411446557e-217
~~~~

~~~
# [2] 6.7359508146380783828098958784721518979443306021620032358869696e-315
```


```{r}
global_p(duration_model)
```


```{r}
nb_summary <-  summary(nb_model)


nb_results <- nb_summary$coefficients %>% data.frame()


rownames(nb_results) <-
  c("(Intercept)",
    "< 10°F",
    "10-19°F",
    "20-29°F",
    "30-39°F",
    "40-49°F",
    "60-69°F",
    "70-79°F",
    "80-89°F",
    ">= 90°F",
~~~

```
    "Windspeed (5 MPH)",
    "Dark",
    "Weekend (Sat, Sun)",
    "Peak Hours (7am-9am, 4pm-7pm)",
    "Rain",
    "Snow",
    "Other Precipitation",
    "Humidity (5%)")

nb_df <-  nb_results %>% transmute(
  beta = Estimate, se = Std..Error, names = rownames(nb_results), dataset = "Overall"
) %>% bind_rows(
  summary(pass_nb_model)$coefficients %>% data.frame() %>%
    transmute(beta = Estimate,
              se = Std..Error,
              names = rownames(nb_results),
              dataset = "Pass Only")
) %>% bind_rows(
  summary(no_pass_nb_model)$coefficients %>% data.frame() %>%
    transmute(beta = Estimate,
              se = Std..Error,
              names = rownames(nb_results),
              dataset = "No Passes")
```

```
)


nb_df %>% filter(names != "(Intercept)") %>%
  forestplot(
    df = .,
    name = names,
    estimate = beta,
    se = se,
  logodds = TRUE,
  colour = dataset,
  alpha = 0.6) + theme_minimal() + scale_color_manual(values =
    c(palette[1], palette[2], palette[4])
  ) + labs(x = "IRR", colour = "Model") + theme(axis.text.y = element_text(size = 16))
```

~~~
```{r}
duration_summary <-  summary(duration_model)


duration_results <- duration_summary$coefficients %>% data.frame()


rownames(duration_results) <-
  c("(Intercept)",
    "< 10°F",
~~~

```
    "10-19°F",

    "20-29°F",

    "30-39°F",

    "40-49°F",

    "60-69°F",

    "70-79°F",

    "80-89°F",

    ">= 90°F",

    "Windspeed (5 MPH)",

    "Dark",

    "Weekend (Sat, Sun)",

    "Peak Hours (7am-9am, 4pm-7pm)",

    "Rain",

    "Snow",

    "Other Precipitation",

    "Humidity (5%)")


duration_df <-  duration_results %>% transmute(

  beta = Estimate, se = Std..Error, names = rownames(duration_results), dataset = "Overall"

) %>% bind_rows(

  summary(pass_duration_model)$coefficients %>% data.frame() %>%

    transmute(beta = Estimate,

              se = Std..Error,
```

~~~
              names = rownames(duration_results),
              dataset = "Pass Only")
) %>% bind_rows(
  summary(no_pass_duration_model)$coefficients %>% data.frame() %>%
    transmute(beta = Estimate,
              se = Std..Error,
              names = rownames(duration_results),
              dataset = "No Passes")
)



duration_df %>% filter(names != "(Intercept)") %>%
  forestplot(
    df = .,
    name = names,
    estimate = beta,
    se = se,
  logodds = FALSE,
  colour = dataset) + theme_minimal() + scale_color_manual(values =
    c(palette[1], palette[2], palette[4])
  ) + theme(axis.text = element_text(size = 16)) + labs(x = "Beta", colour = "Model")
```
~~~

Analyzing the rides from passholders only:

~~~
```{r}
pacman::p_load(here,         # file locator
               tidyverse,    # pass_datamanagement + ggplot2 graphics
               MASS,         # for negative binomial modeling
               lubridate,    # for date handling
               table1,       # for constructing summary table
               gtsummary     # for constructing model summary tables
               )
```
~~~

This document focuses on the duration component with passholders.

# Question:

What is the effect of inclement weather on the number and duration of rides from both members and non-members of Philadelphia's Indego Bikeshare?

# Hypothesis:

Riders take fewer and shorter rides during inclement weather, but demand for rides from passholders is less elastic with respect to weather than non-passholders.

~~~
#Read Data

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, fig.width = 8, fig.height = 5)
options(tidyverse.quiet = TRUE, tidymodels.quiet = TRUE)
~~~

```{r read data}
pass_data_sample <- read_csv(here::here("data", "pass_data.csv")) %>% filter(pass_ind == "Pass")

pass_data <- read_csv(here::here("data", "pass_data.csv")) %>% select(trip_date, count, mean_duration, electric, temp, humidity, preciptype, windspeed,visibility, solarenergy) %>% distinct(trip_date, .keep_all = TRUE)


pass_data <- pass_data %>% mutate(
  count = ifelse(is.na(mean_duration), 0, count),
  preciptype = ifelse(is.na(preciptype), "None",
                      ifelse(preciptype == "rain", preciptype,
                             ifelse(str_detect(preciptype, "snow"), "snow", "Other"))) %>%
```

```
  factor(
    levels = c("None", "rain", "snow", "Other"),
    labels = c("None", "Rain", "Snow", "Other")) %>% relevel(ref = "None"),
solarenergy = ifelse(is.na(solarenergy), 0, solarenergy),
daylight = ifelse(solarenergy == 0, "Night", "Day"),
electric = ifelse(is.na(electric), 0, electric),
year = lubridate::year(trip_date),
quarter = lubridate::quarter(trip_date),
month = lubridate::month(trip_date),
week = lubridate::week(trip_date),
day = lubridate::yday(trip_date),
hour = lubridate::hour(trip_date),
weekend = ifelse(wday(trip_date) > 5, "Weekend", "Weekday") %>% factor(),
peak = ifelse(hour < 7, "Off-Peak",
              ifelse(hour < 9, "Peak",
                     ifelse(hour < 16, "Off-Peak",
                            ifelse(hour < 19, "Peak", "Off-Peak")))) %>%
  factor(),
temp_discrete =
  ifelse(temp <= 9, "Temperature 10",
        ifelse(temp <= 19, "Temperature 20",
               ifelse(temp <= 29, "Temperature 30",
                      ifelse(temp <= 39, "Temperature 40",
                            ifelse(temp <= 49, "Temperature 50",
                                   ifelse(temp <= 59, "Temperature 60",
                                          ifelse(temp <= 69, "Temperature 70",
                                                 ifelse(temp <= 79, "Temperature 80",
                                                        ifelse(temp <= 89, "Temperature 90",
                                                               "Temperature 100")
                                                        )
                                                 )
                                          )
                                   )
                            )
                     )
              )
      ) %>%
factor(
  levels = c(
    "Temperature 10",
    "Temperature 20",
    "Temperature 30",
    "Temperature 40",
    "Temperature 50",
    "Temperature 60",
    "Temperature 70",
      "Temperature 80",
      "Temperature 90",
      "Temperature 100"),
    labels = c(
      "< 10°F",
      "10-19°F",
      "20-29°F",
      "30-39°F",
      "40-49°F",
      "50-59°F",
      "60-69°F",
      "70-79°F",
      "80-89°F",
      ">= 90°F"
      )
    ) %>%
   relevel(ref = "50-59°F"),
  covid = ifelse(trip_date > as.Date("2020-03-22"), 1, 0),
  windy = ifelse(windspeed > quantile(windspeed, 0.1), 1, 0) %>% factor(levels = c("0", "1"), labels = c("calm", "windy"))
  )

pass_summary_data <- pass_data_sample %>% select(
  temp,
  trip_date,
  solarenergy,
  preciptype,
  duration
) %>%
  mutate(
  preciptype = ifelse(is.na(preciptype), "None",
                     ifelse(preciptype == "rain", preciptype,
                            ifelse(str_detect(preciptype, "snow"), "snow", "Other"))) %>%
    factor(
      levels = c("None", "rain", "snow", "Other"),
      labels = c("None", "Rain", "Snow", "Other")) %>% relevel(ref = "None"),
  solarenergy = ifelse(is.na(solarenergy), 0, solarenergy),
  daylight = ifelse(solarenergy == 0, "Night", "Day"),
  year = lubridate::year(trip_date),
  quarter = lubridate::quarter(trip_date),
  month = lubridate::month(trip_date),
  week = lubridate::week(trip_date),
  day = lubridate::yday(trip_date),
  hour = lubridate::hour(trip_date),
  weekend = ifelse(wday(trip_date) > 5, "Weekend", "Weekday") %>% factor(),
  peak = ifelse(hour < 7, "Off-Peak",
                  ifelse(hour < 9, "Peak",
                        ifelse(hour < 16, "Off-Peak",
                              ifelse(hour < 19, "Peak", "Off-Peak")))) %>%
      factor(),
    temp_discrete =
      ifelse(temp <= 9, "Temperature 10",
            ifelse(temp <= 19, "Temperature 20",
                  ifelse(temp <= 29, "Temperature 30",
                        ifelse(temp <= 39, "Temperature 40",
                              ifelse(temp <= 49, "Temperature 50",
                                    ifelse(temp <= 59, "Temperature 60",
                                          ifelse(temp <= 69, "Temperature 70",
                                                ifelse(temp <= 79, "Temperature 80",
                                                      ifelse(temp <= 89, "Temperature 90",
                                                            "Temperature 100")
                                                      )
                                                )
                                          )
                                    )
                              )
                        )
                  )
            ) %>%
factor(
  levels = c(
    "Temperature 10",
    "Temperature 20",
    "Temperature 30",
    "Temperature 40",
    "Temperature 50",
    "Temperature 60",
    "Temperature 70",
    "Temperature 80",
    "Temperature 90",
    "Temperature 100"),
  labels = c(
    "< 10°F",
    "10-19°F",
    "20-29°F",
    "30-39°F",
    "40-49°F",
    "50-59°F",
    "60-69°F",
    "70-79°F",
    "80-89°F",
    ">= 90°F"
```

~~~~
        )
      ) %>%
    relevel(ref = "50-59°F"))

glimpse(data)

skimr::skim(data)
```
~~~~

## Checking for Missing Data

~~~~
```{r missing plot, eval = F}
mice::md.pattern(data, rotate.names = TRUE)
```
~~~~

For these missing values, the missingness is important. It either suggests there were favorable weather conditions, such as in preciptype, or perhaps a nighttime duration as in solarenergy. It also suggests there were no rides in a time period, thus no e-bikes used.

~~~
# Descriptive Statistics

## Univariate

I'll begin by looking at the distributions of the outcome variables.
This analysis looks at two models, one for each outcome variable.

### Ride Count

```{r count graphical}
pass_data%>%
  ggplot(aes(x = count)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = ggsci::pal_jco()(1),
                 alpha = 0.8) +
  labs(title = "Histogram of Rides per Hour",
       x = "Number of Rides Taken that Hour") +
  theme_minimal()

data_sample %>%
  mutate(
~~~

~~~
    Weather = ifelse(is.na(preciptype),
                  "Mild",
                  "Inclement"),
    pass_ind = ifelse(is.na(pass_ind),
                   "No Pass",
                   pass_ind)
                    ) %>%
  group_by(trip_date, pass_ind, Weather) %>%
  summarize(count = n()) %>%
  mutate(Hour = lubridate::hour(trip_date)) %>%
  group_by(Hour, pass_ind, Weather) %>%
  summarize(mean = mean(count, na.rm = T)) %>%
  ggplot(aes(x = Hour, y = mean, fill = Weather)) +
  geom_col(position = "dodge", alpha = 0.8, width = 0.75) +
  facet_grid(pass_ind ~ ., scales = "free_y") +
  ggsci::scale_fill_jco() +
  theme_minimal() +
  labs(title = "Total Rides for Each Hour by Membership Status and Weather",
       subtitle = "Inclement weather is defined as any type of precipitation.",
       x = "Hour of Trip Start",
       y = "Mean Number of Trips")
```
~~~

~~~
```{r}
data_sample %>%
  mutate(
    Weather = ifelse(is.na(preciptype),
                     "Mild",
                     "Inclement"),
    pass_ind = ifelse(is.na(pass_ind),
                      "No Pass",
                      pass_ind),
    Hour = lubridate::hour(trip_date)) %>%
  group_by(Hour, pass_ind, Weather) %>%
  summarize(mean = mean(mean_duration, na.rm = T)) %>%
  ggplot(aes(x = Hour, y = mean, fill = Weather)) +
  geom_col(position = "dodge", alpha = 0.8, width = 0.75) +
  facet_grid(pass_ind ~ .) +
  ggsci::scale_fill_jco() +
  theme_minimal() +
  labs(title = "Mean Ride Duration for Each Hour by Membership Status and Weather",
       subtitle = "Inclement weather is defined as any type of precipitation.",
       x = "Hour of Trip Start",
       y = "Mean Ride Duration")
```
~~~

~~~
```{r}
data_sample %>%
  mutate(day = yday(trip_date)) %>%
  group_by(day) %>%
  summarize("Number of Rides" = n(),
            "Mean Daily Temperature" = mean(temp, na.rm = T)/0.004) %>%
  pivot_longer(!day,
               names_to = "measure",
               values_to = "value") %>%
  ggplot(aes(x = day, y = value, color = measure, shape = measure)) +
  geom_point() +
  scale_y_continuous(
    name = "Total Number of Trips per Day",
    sec.axis = sec_axis(~.*.004, name = "Mean Temperature (°F)")) +
  ggsci::scale_color_jco() +
  theme_minimal() +
  theme(legend.title = element_blank())

data_sample %>%
  mutate(day = yday(trip_date)) %>%
  group_by(day) %>%
  summarize("Mean Ride Duration" = mean(mean_duration, na.rm = T),
            "Mean Daily Temperature" = mean(temp, na.rm = T)*0.25) %>%
~~~

~~~
  pivot_longer(!day,
              names_to = "measure",
              values_to = "value") %>%
  ggplot(aes(x = day, y = value, color = measure, shape = measure)) +
  geom_point() +
  scale_y_continuous(
    name = "Mean Ride Duration (minutes)",
    sec.axis = sec_axis(~./0.25, name = "Mean Temperature (°F)")) +
  ggsci::scale_color_jco() +
  theme_minimal() +
  theme(legend.title = element_blank())

pass_data%>%
  group_by(day) %>%
  summarize(count = sum(count), temp = mean(temp, na.rm = T)) %>%
  ggplot(aes(x = temp, y = count)) + geom_point() + stat_ellipse(color = "red", alpha = 0.6) + ggpubr::stat_cor() + theme_minimal() +
  labs(y = "Total Daily Rides", x = "Mean Daily Temperature (°F)")
```
~~~

### Ride Duration

~~~
```{r duration numerical}
library(table1)
table1::label(data$count) <- "Hourly Count of Rides"
table1::label(data$mean_duration) <- "Hourly Mean Duration of Rides"
table1::label(data$temp) <- "Temperature"
table1::label(data$humidity) <- "Humidity"
table1::label(data$preciptype) <- "Type of Precipitation"
table1::label(data$windspeed) <- "Windspeed"
table1::label(data$windy) <- "Count of Rides Taken in Windy Conditions"
table1::label(data$daylight) <- "Indication of Night/Day"
table1::label(data$electric) <- "Percentage of Rides Taken on E-bikes"
table1::label(data$temp_discrete) <- "Count of Rides in 10°F Intervals"
table1::label(data$weekend) <- "Count of Rides Taken on Saturday or Sunday"
table1::label(data$peak) <- "Count of Rides Taken during Peak Hours"


table1::units(data$trip_date) <- "hours"
table1::units(data$mean_duration) <- "minutes"
table1::units(data$temp) <- "°F"
table1::units(data$humidity) <- "%"
table1::units(data$windspeed) <- "MPH"
~~~

~~~
table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + windy + daylight + weekend + peak + preciptype + electric, data = pass_data) %>% table1::t1flex() %>% flextable::save_as_docx(path = here("files", "table1.docx"))

pass_data %>% summarise(mean(humidity), sd(humidity), min(humidity), max(humidity))
pass_data %>% summarise(mean(temp), sd(temp), min(temp), max(temp))
pass_data %>% summarise(mean(windspeed), sd(windspeed), min(windspeed), max(windspeed))
pass_data %>% summarize(mean(count), sd(count), min(count), max(count))
pass_data %>% summarize(mean(mean_duration, na.rm = T), sd(mean_duration, na.rm = T), min(mean_duration, na.rm = T), max(mean_duration, na.rm = T))

pass_summary_data %>% group_by(temp_discrete) %>% summarize(n())
pass_summary_data %>% group_by(daylight) %>% summarize(n())
pass_summary_data %>% group_by(weekend) %>% summarize(n())
pass_summary_data %>% group_by(peak) %>% summarize(n = n(), n/nrow(pass_summary_data))
pass_summary_data %>% group_by(preciptype) %>% summarize(n())

pass_summary_data %>% table1(~ temp_discrete + daylight + weekend + peak + preciptype, .) %>% t1flex()
```
~~~

Precipitation types are too small outside of none, rain, and snow. I'll collapse the rest into "other" and code the cases of "rain, snow" as just snow. This will come in a bit.

~~~
```{r}
pass_data%>%
  ggplot(aes(x = mean_duration)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  labs(title = "Histogram of Hourly Mean Ride Duration",
       x = "Mean Duration (minutes)") +
  theme_minimal()

# pass_data%>%
#   ggplot(aes(x = log(mean_duration))) +
#   geom_histogram(bins = 100,
#                  color = "black",
#                  fill = "navy",
#                  alpha = 0.6) +
#     geom_histogram(data = data,
~~~

~~~
#                    aes(x = log(mean_duration)),
#                    bins = 100,
#                  color = "black",
#                  fill = "gold",
#                  alpha = 0.45) +
#     geom_histogram(data = no_pass_data,
#                    aes(x = log(mean_duration)),bins = 100,
#                  color = "black",
#                  fill = "grey",
#                  alpha = 0.3) +
#   labs(title = "Histogram of Hourly Mean Ride Duration",
#        x = "Mean Duration (minutes)") +
#   theme_minimal()
```
~~~

The duration could be modeled as a log-normal linear regression, or it could be treated as just a linear regression.

~~~
```{r}
```
~~~

### Independent Variables

~~~
```{r bike numeric}
pass_data%>% ggplot(aes(x = electric)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  theme_minimal()
```


```{r precip numeric}


pass_data%>% ggplot(aes(x = preciptype)) +
  geom_bar(color = "black",
           fill = "navy",
           alpha = 0.6) +
  theme_minimal()
```

```{r continuous numeric}
data_cont_long <- pass_data%>%
  pivot_longer(
~~~

~~~
    cols = c("temp", "humidity", "windspeed", "solarenergy"),
    names_to = "variable",
    values_to = "measure"
    )

data_cont_long %>%
  ggplot(aes(x = measure)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  facet_grid(cols = vars(variable), scales = "free") +
  theme_minimal()
```

```{r temp graphic}
pass_data%>% ggplot(aes(y = temp, x = month, group = month)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Hourly Temperatures by Month")
~~~

~~~
```



```{r humidity graphic}
pass_data%>% ggplot(aes(y = humidity, x = hour, group = hour)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Hourly Humidity by Time of Day")
```


```{r windspeed graphic}
pass_data%>% ggplot(aes(y = windspeed, x = hour, group = hour)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Hourly Windspeed by Time of Day")
```
~~~

~~~
```{r solar graphic}
pass_data%>% ggplot(aes(y = solarenergy, x = month, group = month)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Solar Energy by Month")


pass_data%>% ggplot(aes(y = daylight, x = hour, group = hour)) +
  geom_jitter(color = "black", alpha = 0.3) +
  theme_minimal() +
  labs(title = "Plot of Daylight Classification by Hour")
```
~~~

Clearly using solar energy as a proxy for day/night classification is imperfect, but the overwhelming majority of cases appear to match based on the time of day.

## Bivariate Descriptions/Preliminary Analysis

I want to do a descriptive for each of my outcomes of interest against
each independent variable.

Let's start with ride count.

~~~
```{r outcomes-pass}
table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + daylight + electric + peak + weekend + windy | preciptype, data = pass_data)

table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + preciptype + electric | daylight, data = pass_data)

table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + electric | daylight*preciptype, data = pass_data)

table1(~ count + mean_duration + temp + humidity + windspeed + daylight + electric + preciptype | temp_discrete, data = pass_data)
```

```{r}
pass_data%>% ggplot(aes(x = temp_discrete, y = mean_duration)) + geom_col()
```
~~~

There's a degree of seasonality in this pass_datathat is important to keep
in mind.

~~~
```{r}
~~~

```
    pass_data%>% ggplot(aes(x = day, y = temp)) + geom_point(alpha = 0.25) +
theme_minimal()

    pass_data%>% dplyr::filter(daylight == "Day") %>%
     group_by(trip_date) %>%
     summarize("Hourly Rides" = count,
               "Hourly Temperature" = temp) %>%
     pivot_longer(cols = c("Hourly Rides", "Hourly Temperature"),
                  names_to = "measure",
                  values_to = "value") %>%
     ggplot(aes(x = trip_date, y = value, color = measure)) +
     geom_point(alpha = 0.8) +
     theme_minimal()

    pass_data%>% dplyr::filter(daylight == "Day") %>%
     group_by(trip_date) %>%
     summarize("Hourly Mean Duration" = mean_duration,
               "Hourly Temperature" = temp) %>%
     pivot_longer(cols = c("Hourly Mean Duration", "Hourly Temperature"),
                  names_to = "measure",
                  values_to = "value") %>%
     ggplot(aes(x = trip_date, y = value, color = measure)) +
```

~~~
  geom_point(alpha = 0.8) +
  theme_minimal()

```


```{r counttables}
table1(~ count, data = pass_data) %>% t1flex()
table1(~ count | temp_discrete, data = pass_data) %>% t1flex()
table1(~ count | daylight, data = pass_data) %>% t1flex()
table1(~ count | weekend, data = pass_data) %>% t1flex()
table1(~ count | peak, data = pass_data) %>% t1flex()
table1(~ count | preciptype, data = pass_data) %>% t1flex()
```



# Model


## Negative Binomial


```{r}
pass_nb_model <- glm.nb(count ~ temp_discrete + I(windspeed/5) + daylight + weekend + peak + preciptype + I(humidity/5), data = pass_data)
~~~

~~~~
pass_duration_model <- lm(log(mean_duration) ~ temp_discrete + I(windspeed/5) + daylight + weekend + peak + preciptype + I(humidity/5), data = pass_data)
```

```{r}
plot(nb_model)



plot(duration_model)
```
```{r}
pass_nb_model %>% summary()
pass_duration_model %>% summary()

pass_nb_table <- pass_nb_model %>%
  tbl_regression(exponentiate = T, intercept = F, show_single_row = c('weekend', 'daylight', 'peak'),
                label = c(
                  temp_discrete ~ "Temperature",
                  daylight ~ "Dark",
                  peak ~ "Peak Hours (7am-9am, 4pm-7pm)",
                  weekend ~ "Weekend (Sat, Sun)",
                  preciptype ~ "Precipitation",
~~~~

```
                  `I(windspeed/5)` ~ "Windspeed (5 MPH increments)",
                  `I(humidity/5)` ~ "Humidity (5% increments)"
                )) %>%
  add_global_p() %>%
  as_flex_table()

pass_nb_table %>%
  flextable::save_as_docx(path = here("files", "pass_nb_table.docx"))

pass_duration_table <- pass_duration_model %>%
  tbl_regression(intercept = F, show_single_row = c('weekend', 'daylight', 'peak'),
                label = c(
                  temp_discrete ~ "Temperature",
                  daylight ~ "Dark",
                  peak ~ "Peak Hours (7am-9am, 4pm-7pm)",
                  weekend ~ "Weekend (Sat, Sun)",
                  preciptype ~ "Precipitation",
                  `I(windspeed/5)` ~ "Windspeed (5 MPH increments)",
                  `I(humidity/5)` ~ "Humidity (5% increments)"
                )) %>%
  add_global_p() %>%
  as_flex_table()
```

```
pass_duration_table %>%
  flextable::save_as_docx(path = here("files", "pass_duration_table.docx"))

pass_nb_table
pass_duration_table
```

~~~
```{r}
global_p(pass_nb_model)
```

```{r}
global_p(pass_duration_model)
```

```{r counttables}
table1(~ count, data = pass_data) %>% t1flex()
table1(~ count | temp_discrete, data = pass_data) %>% t1flex()
table1(~ count | daylight, data = pass_data) %>% t1flex()
table1(~ count | weekend, data = pass_data) %>% t1flex()
table1(~ count | peak, data = pass_data) %>% t1flex()
table1(~ count | preciptype, data = pass_data) %>% t1flex()
~~~

~~~
```

```{r meandurationtables}
table1(~ mean_duration, data = pass_data) %>% t1flex()
table1(~ mean_duration | temp_discrete, data = pass_data) %>% t1flex()
table1(~ mean_duration | daylight, data = pass_data) %>% t1flex()
table1(~ mean_duration | weekend, data = pass_data) %>% t1flex()
table1(~ mean_duration | peak, data = pass_data) %>% t1flex()
table1(~ mean_duration | preciptype, data = pass_data) %>% t1flex()
```
~~~

Analyzing the rides from casual riders only:

~~~
```{r}
pacman::p_load(here,         # file locator
          tidyverse,    # data management + ggplot2 graphics
          MASS,         # for negative binomial modeling
          lubridate,    # for date handling
          table1,       # for constructing summary table
          gtsummary     # for constructing model summary tables
          )
```
~~~

This document focuses on the duration component with passholders.

# Question:

What is the effect of inclement weather on the number and duration of rides from both members and non-members of Philadelphia's Indego Bikeshare?

# Hypothesis:

Riders take fewer and shorter rides during inclement weather, but demand for rides from passholders is less elastic with respect to weather than non-passholders.

#Read Data

~~~~
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, fig.width = 8, fig.height = 5)
options(tidyverse.quiet = TRUE, tidymodels.quiet = TRUE)
```
~~~~

~~~~
```{r read data}
no_pass_data_sample <- read_csv(here::here("data", "no_pass_data.csv")) %>% filter(pass_ind == "No Pass")
~~~~

```
no_pass_data <- read_csv(here::here("data", "no_pass_data.csv")) %>% select(trip_date,
count, mean_duration, electric, temp, humidity, preciptype, windspeed,visibility, solarenergy)
%>% distinct(trip_date, .keep_all = TRUE)
```

```
no_pass_data <- no_pass_data %>% mutate(
  count = ifelse(is.na(mean_duration), 0, count),
  preciptype = ifelse(is.na(preciptype), "None",
                     ifelse(preciptype == "rain", preciptype,
                            ifelse(str_detect(preciptype, "snow"), "snow", "Other"))) %>%
    factor(
      levels = c("None", "rain", "snow", "Other"),
      labels = c("None", "Rain", "Snow", "Other")) %>% relevel(ref = "None"),
  solarenergy = ifelse(is.na(solarenergy), 0, solarenergy),
  daylight = ifelse(solarenergy == 0, "Night", "Day"),
  electric = ifelse(is.na(electric), 0, electric),
  year = lubridate::year(trip_date),
  quarter = lubridate::quarter(trip_date),
  month = lubridate::month(trip_date),
  week = lubridate::week(trip_date),
  day = lubridate::yday(trip_date),
  hour = lubridate::hour(trip_date),
  weekend = ifelse(wday(trip_date) > 5, "Weekend", "Weekday") %>% factor(),
  peak = ifelse(hour < 7, "Off-Peak",
          ifelse(hour < 9, "Peak",
              ifelse(hour < 16, "Off-Peak",
                  ifelse(hour < 19, "Peak", "Off-Peak")))) %>%
    factor(),
  temp_discrete =
    ifelse(temp <= 9, "Temperature 10",
        ifelse(temp <= 19, "Temperature 20",
            ifelse(temp <= 29, "Temperature 30",
                ifelse(temp <= 39, "Temperature 40",
                    ifelse(temp <= 49, "Temperature 50",
                        ifelse(temp <= 59, "Temperature 60",
                            ifelse(temp <= 69, "Temperature 70",
                                ifelse(temp <= 79, "Temperature 80",
                                    ifelse(temp <= 89, "Temperature 90",
                                        "Temperature 100")
                                    )
                                )
                            )
                        )
                    )
```

```
                   )
             )
      ) %>%
factor(
  levels = c(
    "Temperature 10",
    "Temperature 20",
    "Temperature 30",
    "Temperature 40",
    "Temperature 50",
    "Temperature 60",
    "Temperature 70",
    "Temperature 80",
    "Temperature 90",
    "Temperature 100"),
  labels = c(
    "< 10°F",
    "10-19°F",
    "20-29°F",
    "30-39°F",
    "40-49°F",
    "50-59°F",
    "60-69°F",
                    "70-79°F",
                    "80-89°F",
                    ">= 90°F"
                    )
                  ) %>%
                relevel(ref = "50-59°F"),
               covid = ifelse(trip_date > as.Date("2020-03-22"), 1, 0),
               windy = ifelse(windspeed > quantile(windspeed, 0.1), 1, 0) %>% factor(levels = c("0",
"1"), labels = c("calm", "windy"))
               )

            no_pass_summary_data <- no_pass_data_sample %>% select(
              temp,
              trip_date,
              solarenergy,
              preciptype,
              duration
            ) %>%
              mutate(
              preciptype = ifelse(is.na(preciptype), "None",
                                   ifelse(preciptype == "rain", preciptype,
                                          ifelse(str_detect(preciptype, "snow"), "snow", "Other"))) %>%
                factor(
    levels = c("None", "rain", "snow", "Other"),
    labels = c("None", "Rain", "Snow", "Other")) %>% relevel(ref = "None"),
solarenergy = ifelse(is.na(solarenergy), 0, solarenergy),
daylight = ifelse(solarenergy == 0, "Night", "Day"),
year = lubridate::year(trip_date),
quarter = lubridate::quarter(trip_date),
month = lubridate::month(trip_date),
week = lubridate::week(trip_date),
day = lubridate::yday(trip_date),
hour = lubridate::hour(trip_date),
weekend = ifelse(wday(trip_date) > 5, "Weekend", "Weekday") %>% factor(),
peak = ifelse(hour < 7, "Off-Peak",
               ifelse(hour < 9, "Peak",
                       ifelse(hour < 16, "Off-Peak",
                               ifelse(hour < 19, "Peak", "Off-Peak")))) %>%
  factor(),
temp_discrete =
  ifelse(temp <= 9, "Temperature 10",
        ifelse(temp <= 19, "Temperature 20",
                ifelse(temp <= 29, "Temperature 30",
                        ifelse(temp <= 39, "Temperature 40",
                                ifelse(temp <= 49, "Temperature 50",
                                        ifelse(temp <= 59, "Temperature 60",
                                         ifelse(temp <= 69, "Temperature 70",
                                                ifelse(temp <= 79, "Temperature 80",
                                                       ifelse(temp <= 89, "Temperature 90",
                                                              "Temperature 100")
                                                       )
                                                )
                                         )
                                  )
                           )
                    )
             )
      ) %>%
factor(
  levels = c(
    "Temperature 10",
    "Temperature 20",
    "Temperature 30",
    "Temperature 40",
    "Temperature 50",
    "Temperature 60",
    "Temperature 70",
    "Temperature 80",
    "Temperature 90",
```

~~~
        "Temperature 100"),
      labels = c(
        "< 10°F",
        "10-19°F",
        "20-29°F",
        "30-39°F",
        "40-49°F",
        "50-59°F",
        "60-69°F",
        "70-79°F",
        "80-89°F",
        ">= 90°F"
        )
      ) %>%
    relevel(ref = "50-59°F"))

glimpse(data)


skimr::skim(data)
```
~~~

## Checking for Missing Data

```{r missing plot, eval = F}
mice::md.pattern(data, rotate.names = TRUE)
```

For these missing values, the missingness is important. It either
suggests there were favorable weather conditions, such as in preciptype,
or perhaps a nighttime duration as in solarenergy. It also suggests there were no rides in a time period, thus no e-bikes used.

# Descriptive Statistics

## Univariate

I'll begin by looking at the distributions of the outcome variables.
This analysis looks at two models, one for each outcome variable.

### Ride Count

```{r count graphical}
```

~~~
no_pass_data %>%
  ggplot(aes(x = count)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  labs(title = "Histogram of Rides per Hour",
       x = "Number of Rides Taken that Hour") +
  theme_minimal()
```


### Ride Duration


```{r duration numerical}
library(table1)
table1::label(data$count) <- "Hourly Count of Rides"
table1::label(data$mean_duration) <- "Hourly Mean Duration of Rides"
table1::label(data$temp) <- "Temperature"
table1::label(data$humidity) <- "Humidity"
table1::label(data$preciptype) <- "Type of Precipitation"
table1::label(data$windspeed) <- "Windspeed"
table1::label(data$windy) <- "Count of Rides Taken in Windy Conditions"
table1::label(data$daylight) <- "Indication of Night/Day"
~~~

~~~
table1::label(data$electric) <- "Percentage of Rides Taken on E-bikes"
table1::label(data$temp_discrete) <- "Count of Rides in 10°F Intervals"
table1::label(data$weekend) <- "Count of Rides Taken on Saturday or Sunday"
table1::label(data$peak) <- "Count of Rides Taken during Peak Hours"


table1::units(data$trip_date) <- "hours"
table1::units(data$mean_duration) <- "minutes"
table1::units(data$temp) <- "°F"
table1::units(data$humidity) <- "%"
table1::units(data$windspeed) <- "MPH"


table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + windy + daylight + weekend + peak + preciptype + electric, data = no_pass_data)

no_pass_summary_data %>% table1(~ temp_discrete + daylight + weekend + peak + preciptype, .) %>% t1flex()
```
~~~

Precipitation types are too small outside of none, rain, and snow. I'll collapse the rest into "other" and code the cases of "rain, snow" as just snow. This will come in a bit.

~~~~
```{r}
no_pass_data %>%
  ggplot(aes(x = mean_duration)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  labs(title = "Histogram of Hourly Mean Ride Duration",
       x = "Mean Duration (minutes)") +
  theme_minimal()

no_pass_data %>%
  ggplot(aes(x = log(mean_duration))) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  labs(title = "Histogram of Hourly Mean Ride Duration",
       x = "Mean Duration (minutes)") +
  theme_minimal()
```
~~~~

The duration could be modeled as a log-normal linear regression, or it could be treated as just a linear regression.

~~~~
```{r}
```

### Independent Variables

```{r bike numeric}
no_pass_data %>% ggplot(aes(x = electric)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  theme_minimal()
```


```{r precip numeric}


no_pass_data %>% ggplot(aes(x = preciptype)) +
  geom_bar(color = "black",
~~~~

```
           fill = "navy",
           alpha = 0.6) +
  theme_minimal()
```
~~~

```{r continuous numeric}
data_cont_long <- no_pass_data %>%
  pivot_longer(
    cols = c("temp", "humidity", "windspeed", "solarenergy"),
    names_to = "variable",
    values_to = "measure"
    )

data_cont_long %>%
  ggplot(aes(x = measure)) +
  geom_histogram(bins = 100,
                 color = "black",
                 fill = "navy",
                 alpha = 0.6) +
  facet_grid(cols = vars(variable), scales = "free") +
  theme_minimal()
```
~~~

~~~
```{r temp graphic}
no_pass_data %>% ggplot(aes(y = temp, x = month, group = month)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  geom_point(aes(y = count, x = month),
              color = "black",
              fill = "gold",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Hourly Temperatures by Month")
```
```{r duration numerical}
library(table1)
table1::label(data$count) <- "Hourly Count of Rides"
table1::label(data$mean_duration) <- "Hourly Mean Duration of Rides"
table1::label(data$temp) <- "Temperature"
table1::label(data$humidity) <- "Humidity"
table1::label(data$preciptype) <- "Type of Precipitation"
table1::label(data$windspeed) <- "Windspeed"
table1::label(data$windy) <- "Count of Rides Taken in Windy Conditions"
table1::label(data$daylight) <- "Indication of Night/Day"
~~~

```
table1::label(data$electric) <- "Percentage of Rides Taken on E-bikes"
table1::label(data$temp_discrete) <- "Count of Rides in 10°F Intervals"
table1::label(data$weekend) <- "Count of Rides Taken on Saturday or Sunday"
table1::label(data$peak) <- "Count of Rides Taken during Peak Hours"
```

```
table1::units(data$trip_date) <- "hours"
table1::units(data$mean_duration) <- "minutes"
table1::units(data$temp) <- "°F"
table1::units(data$humidity) <- "%"
table1::units(data$windspeed) <- "MPH"
```

```
table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + windy + daylight + weekend + peak + preciptype + electric, data = pass_data) %>% table1::t1flex() %>% flextable::save_as_docx(path = here("files", "table1.docx"))
```

```
no_pass_data %>% summarise(mean(humidity), sd(humidity), min(humidity), max(humidity))
no_pass_data %>% summarise(mean(temp), sd(temp), min(temp), max(temp))
no_pass_data %>% summarise(mean(windspeed), sd(windspeed), min(windspeed), max(windspeed))
no_pass_data %>% summarize(mean(count), sd(count), min(count), max(count))
```

~~~
no_pass_data %>% summarize(mean(mean_duration, na.rm = T), sd(mean_duration, na.rm = T), min(mean_duration, na.rm = T), max(mean_duration, na.rm = T))

no_pass_summary_data %>% group_by(temp_discrete) %>% summarize(n())
no_pass_summary_data %>% group_by(daylight) %>% summarize(n())
no_pass_summary_data %>% group_by(weekend) %>% summarize(n())
no_pass_summary_data %>% group_by(peak) %>% summarize(n = n(), n/nrow(no_pass_summary_data))
no_pass_summary_data %>% group_by(preciptype) %>% summarize(n())

no_pass_summary_data %>% table1(~ temp_discrete + daylight + weekend + peak + preciptype, .) %>% t1flex()
```

```{r humidity graphic}
no_pass_data %>% ggplot(aes(y = humidity, x = hour, group = hour)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Hourly Humidity by Time of Day")
```
~~~

~~~~
```{r windspeed graphic}
no_pass_data %>% ggplot(aes(y = windspeed, x = hour, group = hour)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Hourly Windspeed by Time of Day")
```

```{r solar graphic}
no_pass_data %>% ggplot(aes(y = solarenergy, x = month, group = month)) +
  geom_violin(draw_quantiles = c(0.25, 0.5, 0.75),
              color = "black",
              fill = "navy",
              alpha = 0.6) +
  theme_minimal() +
  labs(title = "Plot of Solar Energy by Month")

no_pass_data %>% ggplot(aes(y = daylight, x = hour, group = hour)) +
  geom_jitter(color = "black", alpha = 0.3) +
  theme_minimal() +
~~~~

~~~
  labs(title = "Plot of Daylight Classification by Hour")
```
~~~

Clearly using solar energy as a proxy for day/night classification is imperfect, but the overwhelming majority of cases appear to match based on the time of day.

## Bivariate Descriptions/Preliminary Analysis

I want to do a descriptive for each of my outcomes of interest against
each independent variable.

Let's start with ride count.

~~~
```{r outcomes-pass}
table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + daylight + electric + peak + weekend + windy | preciptype, data = no_pass_data)

table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + preciptype + electric | daylight, data = no_pass_data)

table1(~ count + mean_duration + temp + temp_discrete + humidity + windspeed + electric | daylight*preciptype, data = no_pass_data)
~~~

~~~~
table1(~ count + mean_duration + temp + humidity + windspeed + daylight + electric + preciptype | temp_discrete, data = no_pass_data)
```

```{r}
no_pass_data %>% ggplot(aes(x = temp_discrete, y = mean_duration)) + geom_col()
```

There's a degree of seasonality in this data that is important to keep
in mind.

```{r}

no_pass_data %>% ggplot(aes(x = day, y = temp)) + geom_point(alpha = 0.25) + theme_minimal()

no_pass_data %>% dplyr::filter(daylight == "Day") %>%
  group_by(trip_date) %>%
  summarize("Hourly Rides" = count,
            "Hourly Temperature" = temp) %>%
  pivot_longer(cols = c("Hourly Rides", "Hourly Temperature"),
               names_to = "measure",
               values_to = "value") %>%
~~~~

```
  ggplot(aes(x = trip_date, y = value, color = measure)) +
  geom_point(alpha = 0.8) +
  theme_minimal()

no_pass_data %>% dplyr::filter(daylight == "Day") %>%
  group_by(trip_date) %>%
  summarize("Hourly Mean Duration" = mean_duration,
            "Hourly Temperature" = temp) %>%
  pivot_longer(cols = c("Hourly Mean Duration", "Hourly Temperature"),
               names_to = "measure",
               values_to = "value") %>%
  ggplot(aes(x = trip_date, y = value, color = measure)) +
  geom_point(alpha = 0.8) +
  theme_minimal()

```

# Model

Create a train/test split and a cv split.

~~~
## Negative Binomial

```{r}
no_pass_nb_model <- glm.nb(count ~ temp_discrete + I(windspeed/5) + daylight + weekend + peak + preciptype + I(humidity/5), data = no_pass_data)

no_pass_duration_model <- lm(log(mean_duration) ~ temp_discrete + I(windspeed/5) + daylight + weekend + peak + preciptype + I(humidity/5), data = no_pass_data)
```

```{r}
plot(no_pass_nb_model)

plot(no_pass_duration_model)
```
```{r}
no_pass_nb_model %>% summary()
no_pass_duration_model %>% summary()
~~~

```
no_pass_nb_table <- no_pass_nb_model %>%
  tbl_regression(intercept = F, show_single_row = c('weekend', 'daylight', 'peak'),
                label = c(
                  temp_discrete ~ "Temperature",
                  daylight ~ "Dark",
                  peak ~ "Peak Hours (7am-9am, 4pm-7pm)",
                  weekend ~ "Weekend (Sat, Sun)",
                  preciptype ~ "Precipitation",
                  `I(windspeed/5)` ~ "Windspeed (5 MPH increments)",
                  `I(humidity/5)` ~ "Humidity (5% increments)"
                )) %>%
  add_global_p() %>%
  as_flex_table()

no_pass_duration_table <- no_pass_duration_model %>%

  tbl_regression(intercept = F, show_single_row = c('weekend', 'daylight', 'peak'),
                label = c(
                  temp_discrete ~ "Temperature",
                  daylight ~ "Dark",
                  peak ~ "Peak Hours (7am-9am, 4pm-7pm)",
                  weekend ~ "Weekend (Sat, Sun)",
                  preciptype ~ "Precipitation",
                    `I(windspeed/5)` ~ "Windspeed (5 MPH increments)",
                    `I(humidity/5)` ~ "Humidity (5% increments)"
                  )) %>%
    add_global_p() %>%
    as_flex_table()

no_pass_nb_table %>% flextable::save_as_docx(path = here("files", "no_pass_nb_results.docx"))
no_pass_duration_table %>% flextable::save_as_docx(path = here("files", "no_pass_duration_results.docx"))

no_pass_nb_table
no_pass_duration_table
```

```{r}
global_p(no_pass_nb_model)
```
```{r}
global_p(no_pass_duration_model)
```

```{r counttables}
```

~~~
table1(~ count, data = no_pass_data) %>% t1flex()
table1(~ count | temp_discrete, data = no_pass_data) %>% t1flex()
table1(~ count | daylight, data = no_pass_data) %>% t1flex()
table1(~ count | weekend, data = no_pass_data) %>% t1flex()
table1(~ count | peak, data = no_pass_data) %>% t1flex()
table1(~ count | preciptype, data = no_pass_data) %>% t1flex()
```


```{r meandurationtables}
table1(~ mean_duration, data = no_pass_data) %>% t1flex()
table1(~ mean_duration | temp_discrete, data = no_pass_data) %>% t1flex()
table1(~ mean_duration | daylight, data = no_pass_data) %>% t1flex()
table1(~ mean_duration | weekend, data = no_pass_data) %>% t1flex()
table1(~ mean_duration | peak, data = no_pass_data) %>% t1flex()
table1(~ mean_duration | preciptype, data = no_pass_data) %>% t1flex()
```
~~~

Bibliography

Ahmed, F., Rose, A. G., & Jacob, C. (n.d.). *Impact of weather on commuter cyclist behaviour and implications for climate change adaptation*.

Bean, R., Pojani, D., & Corcoran, J. (2021). How does weather affect bikeshare use? A comparative analysis of forty cities across climate zones. *Journal of Transport Geography*, *95*, 103155. https://doi.org/10.1016/j.jtrangeo.2021.103155

Better Bike Share Partnership. (2023). *Better Bike Share—The Better Bike Share Partnership is a JPB Foundation-funded collaboration between The City of Philadelphia, Bicycle Coalition of Greater Philadelphia, the National Association of City Transportation Officials (NACTO) and the PeopleForBikes Foundation to build equitable and replicable bike share systems.* https://betterbikeshare.org/

Bicycle Coalition of Greater Philadelphia. (2018). *2018 Bike PHL Facts*. https://phlbikecoalition.maps.arcgis.com/apps/Cascade/index.html?appid=5101f5b7a3094258af8fe02ef01a2f78&folderid=cb3a38c8a7f641288c23e6a31085f672

Buck, D., Buehler, R., Happ, P., Rawls, B., Chung, P., & Borecki, N. (2013). Are Bikeshare Users Different from Regular Cyclists?: A First Look at Short-Term Users, Annual Members, and Area Cyclists in the Washington, D.C., Region. *Transportation Research Record*, *2387*(1), 112–119. https://doi.org/10.3141/2387-13

Buy a Pass. (2020, August 12). Indego. https://www.rideindego.com/buy-a-pass/

CDC. (2023, March 23). *Adult Physical Inactivity*. Centers for Disease Control and Prevention. https://www.cdc.gov/physicalactivity/data/inactivity-prevalence-maps/index.html

Crossa, A., Reilly, K. H., Wang, S. M., Lim, S., & Noyes, P. (2022). If We Build It, Who Will Come? Comparing Sociodemographic Characteristics of Bike Share Subscribers, Cyclists, and Residents of New York City. *Transportation Research Record*, *2676*(3), 634–642. https://doi.org/10.1177/03611981211055664

Engel, J. (1984). MODELS FOR RESPONSE DATA SHOWING EXTRA-POISSON VARIATION. *Statistica Neerlandica*, *38*(3), 159–167. https://doi.org/10.1111/j.1467-9574.1984.tb01107.x

Fishman, E., Washington, S., & Haworth, N. (2015). Bikeshare's impact on active travel: Evidence from the United States, Great Britain, and Australia. *Journal of Transport & Health*, *2*(2), 135–142. https://doi.org/10.1016/j.jth.2015.03.004

Foursquare ITP. (2018). *Indego 2018 Business Plan Update*. https://www.phila.gov/media/20220831123354/Indego-business-plan-2018.pdf

Friendly, M., Monette, G., & Fox, J. (2013). Elliptical insights: Understanding statistical methods through elliptical geometry. *Statistical Science*, *28*(1), 1–39. https://doi.org/10.1214/12-STS402

Gebhart, K., & Noland, R. B. (2014). The impact of weather conditions on bikeshare trips in Washington, DC. *Transportation*, *41*(6), 1205–1225. https://doi.org/10.1007/s11116-014-9540-7

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Gentner, D. R., Jathar, S. H., Gordon, T. D., Bahreini, R., Day, D. A., El Haddad, I., Hayes, P. L., Pieber, S. M., Platt, S. M., de Gouw, J., Goldstein, A. H., Harley, R. A., Jimenez, J. L., Prévôt, A. S. H., & Robinson, A. L. (2017a). Review of Urban Secondary Organic Aerosol Formation from Gasoline and Diesel Motor Vehicle Emissions. *Environmental Science & Technology*, *51*(3), 1074–1093. https://doi.org/10.1021/acs.est.6b04509

Gentner, D. R., Jathar, S. H., Gordon, T. D., Bahreini, R., Day, D. A., El Haddad, I., Hayes, P. L., Pieber, S. M., Platt, S. M., de Gouw, J., Goldstein, A. H., Harley, R. A., Jimenez, J. L., Prévôt, A. S. H., & Robinson, A. L. (2017b). Review of Urban Secondary Organic Aerosol Formation from Gasoline and Diesel Motor Vehicle Emissions. *Environmental Science & Technology*, *51*(3), 1074–1093. https://doi.org/10.1021/acs.est.6b04509

Gohel, D., & Skintzos, P. (2022). *flextable: Functions for tabular reporting* [Manual]. https://CRAN.R-project.org/package=flextable

Goldmann, K., & Wessel, J. (2021). Some people feel the rain, others just get wet: An analysis of regional differences in the effects of weather on cycling. *Research in Transportation Business & Management*, *40*, 100541. https://doi.org/10.1016/j.rtbm.2020.100541

Grolemund, G., & Wickham, H. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, *40*, 1–25. https://doi.org/10.18637/jss.v040.i03

Indego. (2022). *Data – Indego* [Data set]. Indego. https://www.rideindego.com/about/data/

Kassambara, A. (2022). *ggpubr: Ggplot2 based publication ready plots* [Manual]. https://rpkgs.datanovia.com/ggpubr/

Lawless, J. F. (1987). Negative Binomial and Mixed Poisson Regression. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *15*(3), 209–225. https://doi.org/10.2307/3314912

Lindsay, G., Macmillan, A., & Woodward, A. (2011). Moving urban trips from cars to bicycles: Impact on health and emissions. *Australian and New Zealand Journal of Public Health*, *35*(1), 54–60. https://doi.org/10.1111/j.1753-6405.2010.00621.x

Maechler, M. (2021). *Rmpfr: R MPFR - multiple precision floating-point reliable* [Manual]. https://rmpfr.r-forge.r-project.org/

Miranda-Moreno, L. F., & Nosal, T. (2011). Weather or Not to Cycle: Temporal Trends and Impact of Weather on Cycling in an Urban Environment. *Transportation Research Record*, *2247*(1), 42–52. https://doi.org/10.3141/2247-06

Müller, K., & Wickham, H. (2023). *tibble: Simple data frames* [Manual]. https://CRAN.R-project.org/package=tibble

PeopleForBikes. (2022). *Philadelphia Pennsylvania City Rating Page | PeopleForBikes 2022 City Ratings*. PeopleForBikes. https://cityratings.peopleforbikes.org/philadelphia-pa

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. https://www.R-project.org/

Rich, B. (2023). *table1: Tables of Descriptive Statistics in HTML* (1.4.3). https://CRAN.R-project.org/package=table1

Scheinin, I., Kalimeri, M., Jagerroos, V., Parkkinen, J., Tikkanen, E., Würtz, P., & Kangas, A. (2023). *ggforestplot: Forestplots of measures of effects and their confidence intervals* [Manual].

Shaheen, S. A., Martin, E. W., Cohen, A. P., & Finson, R. S. (2012). *PUBLIC BIKESHARING IN NORTH AMERICA: EARLY OPERATOR AND USER UNDERSTANDING* (No. 11–26). Mineta Transportation Institute.

Sjoberg, D. D., Larmarange, J., Curry, M., Lavery, J., Whiting, K., Zabor, E. C., Bai, X., Drill, E., Flynn, J., Hannum, M., Lobaugh, S., Pileggi, S., Tin, A., & Wainberg, G. Z. (2023). *gtsummary: Presentation-Ready Data Summary and Analytic Result Tables* (1.7.0). https://CRAN.R-project.org/package=gtsummary

The Bicycle Coalition of Greater Philadelphia. (2022, May 17). *2020-2021 Bike Counts*. ArcGIS StoryMaps. https://storymaps.arcgis.com/stories/b70bff61ab8a4fd0ab8135257e3ed183

Thomas, T., Jaarsma, R., & Tutert, B. (2008). *Temporal variations of bicycle demand in the Netherlands: The influence of weather on cycling*.

Thomas, T., Jaarsma, R., & Tutert, B. (2013). Exploring temporal fluctuations of daily cycling demand on Dutch cycle paths: The influence of weather on cycling. *Transportation*, *40*(1), 1–22. https://doi.org/10.1007/s11116-012-9398-5

Ursaki, J., & Aultman-Hall, L. (2015). *QUANTIFYING THE EQUITY OF BIKESHARE ACCESS IN US CITIES* (No. 15–011). University of Vermont Transportation Research Center.

van Buuren, S., & Groothuis-Oudshoorn, K. (2022). *mice: Multivariate imputation by chained equations* [Manual]. https://CRAN.R-project.org/package=mice

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Visual Crossing. (2023). *Historical Weather Data & Weather Forecast Data | Visual Crossing* [Data set]. Visual Crossing. https://www.visualcrossing.com/weather-data

Waring, E., Quinn, M., McNamara, A., de la Rubia, E. A., Zhu, H., & Ellis, S. (2022). *skimr: Compact and flexible summaries of data* [Manual]. https://CRAN.R-project.org/package=skimr

Wessel, J. (2020). Using weather forecasts to forecast whether bikes are used. *Transportation Research Part A: Policy and Practice*, *138*, 537–559. https://doi.org/10.1016/j.tra.2020.06.006

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2022a). *forcats: Tools for working with categorical variables (factors)* [Manual]. https://CRAN.R-project.org/package=forcats

Wickham, H. (2022b). *stringr: Simple, consistent wrappers for common string operations* [Manual]. https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* [Manual]. https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2023). *purrr: Functional programming tools* [Manual]. https://CRAN.R-project.org/package=purrr

Wickham, H., Hester, J., & Bryan, J. (2022). *readr: Read rectangular text data* [Manual]. https://CRAN.R-project.org/package=readr

Wickham, H., Vaughan, D., & Girlich, M. (2023). *tidyr: Tidy messy data* [Manual]. https://CRAN.R-project.org/package=tidyr

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). South-Western.

World Health Organization. (2018). *Global action plan on physical activity 2018–2030: More active people for a healthier world*. World Health Organization. https://apps.who.int/iris/handle/10665/272722

Xiao, N. (2018). *ggsci: Scientific journal and sci-fi themed color palettes for ggplot2* [Manual]. https://CRAN.R-project.org/package=ggsci