Exploring the Additive Effects of Religious Participation on Multivariate, Demographics Based Machine Learning Models

by

Ian Michael Jacobs

Bachelor of Science in Biochemistry, Beloit College, 2022

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

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Ian Michael Jacobs

It was defended on

April 3, 2024

and approved by

Dr. Haley Grant, Assistant Professor, Department of Biostatistics

Dr. Nilesh Shah, Assistant Professor, Department of Dental Public Health

Dr. Lu Tang, Assistant Professor, Department of Biostatistics

Thesis Advisor: Dr. Ada Youk, Associate Professor, Department of Biostatistics

Exploring the Additive Effects of Religious Participation on Multivariate, Demographics Based Machine Learning Models

Ian Michael Jacobs, MS

University of Pittsburgh, 2024

Through the 21st century, vaccine hesitancy has had a significant effect on the implementation of vaccine development and rollout in the United States. A known and well documented factor that contributes to this kind of structural hesitancy is regular participation in a religious congregation or community whose doctrine or teachings condemn vaccination and/or modern medicine in some form. The public health contribution of this thesis is to support the use of machine learning in the prediction of public health outcomes, as well as promote the contribution of socially anchored metrics within demographics-based models.

Data for this project was sourced from The Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation, The U.S. Department of Agriculture's Economic Research Survey, and The Association of Statisticians of American Religious Bodies' U.S. Religion Census. These data were cleaned at the U.S. county level and the remaining variables were categorized into six major demographic categories: education, population, poverty, unemployment, vaccine hesitancy, and religious participation. This cleaning process resulted in 54 usable demographic variables and one outcome variable.

After data cleaning was performed, four machine learning techniques were implemented on the variable set to compare their prediction ability: elastic net, multivariate adaptive regression splines, random forest, and gradient boosted trees. Using the root mean square error and R-squared

of each of these models, it was determined that the gradient boosted trees method had the greatest prediction ability with this particular dataset.

Variable selection was performed, and it was determined through importance testing that 26 of the 54 variables had a significant contribution to the model and provided the most substantial prediction ability. Of those 26 variables, two originated from the religion category. Results from the gradient boosted tree analysis indicated a decrease in prediction ability when the selected religion variables were removed from the model, which supports a data-based linkage between vaccine hesitancy and religious participation. Post-hoc hierarchical clustering was performed at a county level to give a visual representation of the demographically constructed clusters and to provide a geographically based comparison between the selected demographics and vaccine hesitancy.

Table of Contents

List of Tables

List of Figures

Acknowledgments

The author would like to thank Dr. Ada Youk for her constant guidance and wisdom throughout the development of this project, Dr. Lu Tang for his machine learning advice and expertise, Dr. Nilesh Shah and Dr. Haley Grant for their efforts as members of the committee, and the late Dr. Debra Majeed for fostering a lifelong curiosity for religious studies in all of the diverse ways it shapes our world. It is a privilege to be able to commemorate her life and legacy within this thesis. Additionally, the author would like to thank his wonderful grandmother, Karen Malmisur. Her spectacular proofreading abilities were invaluable in the final stages of the editing process, and the weekend visits to her home throughout the course of this thesis were similarly invaluable for the authors mental health and wellbeing. Finally, the author would like to thank Pitt Public Health, the Carnegie Library in Squirrel Hill, and the 61C Cafe for their wonderful spaces in which this thesis was developed.

1.0 Introduction

This chapter contains three sections: Background for this research, overall Research Objectives, and the potential Research Contributions. In the Background, religious participation and vaccine hesitancy are discussed and linked through previously published research. In Research Objective, the main research question is detailed. In Research Contribution, potential contributions to the field of biostatistics, social epidemiology, and machine learning are expanded upon.

1.1 Background

Religious participation and vaccine hesitancy have been linked since the inception of vaccination. This phenomenon has negatively impacted the effectiveness of vaccination campaigns and population-protection from vaccine-preventable disease, and has deep roots in many different religious traditions including Protestantism, Catholicism, Judaism, Islam, Christianity, Amish faiths, Hinduism, and Sikhism (Kibongani Volet et al., 2022). Individuals with strong religious beliefs will choose to follow the teachings of their religion, and those teachings often encourage alternative approaches such as through the use of holy water, religious ceremonies, or different forms of prayer in an attempt to combat illness or disease (Garcia & Yap, 2021). This sort of faith-based hesitancy came to the forefront of discourse surrounding COVID-19 and the rollout of its multiple vaccines and is a crucial area of study for the field of social epidemiology. Vaccinations at large have been confirmed to be effective and safe in the treatment of disease, particularly in the case of COVID-19. Receiving vaccination has been demonstrated to significantly decrease

mortality during hospitalization for COVID-19, correlated with an increase in likelihood of being discharged home, and resulted in an overall decreased length of hospital (Lee et al., 2023).

The overarching goal of this thesis is to demonstrate the merit in incorporating sociologically based metrics, in this case religious participation, to demographics-based models with an outcome variable anchored in personal choice and community engagement like vaccine hesitancy. Machine learning provides useful, flexible, and innovative techniques that allow practitioners to perform multivariate analysis with high dimensional data. These qualities are the reason that this area of analysis was selected for the prediction modeling involved in this thesis over more traditional regression models. Additionally, different machine learning models are compared to determine the ideal methodological approach while modulating between variable sets. The machine learning techniques explored include elastic net regression, multivariate adaptive regression splines, random forest models, and gradient boosted tree models. This varied and multi-step approach was inspired by Nicholson et al, whose 2022 paper on a machine learning and clustering approach to COVID-19 data research at the county level had a significant influence over the way that this thesis was structured and how the analysis was performed.

1.2 Research Objective

The goal of this thesis is to analyze and promote the influence of religious participation on multivariate, demographics-based machine learning models that share a community influenced outcome variable of vaccine hesitancy. Promoting the use of these machine learning methods in the public health space is incredibly important for the future of data-based prediction modeling for public health outcomes at the national, state, as well as local level. A combination of county-level

data sources have been assembled, including data on education, population metrics, poverty, unemployment, vaccine hesitancy, and religious participation. These disparate data sources were cleaned in order to derive the most effective machine learning approach for prediction, to determine which data dimensions are the most critical, and to then determine the significance of the critical religion-based variables on the overall predictive ability of a final, multivariate model. A secondary goal is to present a cluster-based analysis of the significant dimensions determined during the main analysis of this thesis in order to provide geographically grounded, county level groups. The aim of this grouping analysis is to provide a visual representation of how the final demographic variable set is distributed at the national and regional levels and how these clusters compare visually to a similar map vaccine hesitancy.

1.3 Research Contribution

The work performed as part of this thesis will contribute to the ongoing work in the field of machine learning as it relates to public health at large. The methods presented are being utilized in a number of scientific fields and will add to the body of knowledge related to county-level, multivariate analysis in a novel and innovative way. This work will also contribute to the field of social epidemiology through computational, machine learning techniques that are relatively new to that area of research as they are applied to socially informed public health outcomes.

2.0 Data Processing

This chapter consists of two sections: Data Sourcing and Data Cleaning. The Data Sourcing section details the source of the six datasheets that were utilized for this analysis, and the Data Cleaning section contains details related to the cleaning process, missing data, and software used.

2.1 Data Sourcing

Each source utilized for this analysis are free to use and publicly available. Specific selected variable information for each of these datasets is contained within Appendix A. Each set of variables contained differing variable naming syntax, so the naming convention was standardized for the final analysis. Both the original variable name and the generated analysis variable name are included within the tables of Appendix A.

2.1.1 Economic Research Survey (U.S.D.A.)

The United States Department of Agriculture's Economic Research Survey (U.S. Department of Agriculture, 2018) contained four out of six of the county level datasheets that were incorporated into the final data file: education, population, poverty, and unemployment. These sheets contain county level data, and each row represents a United States county or county equivalent. Each of these datasets contained a number of variables that were unnecessary for this analysis, and those variables were removed.

The education dataset contained educational level attainment variables since 1970, but this project used the most recent data (2017-2021). This dataset also contained Rural-urban continuum codes and urban influence codes for 2003 and 2013. Rural-urban continuum codes, as well as urban influence codes, form a classification scheme that distinguishes metropolitan (metro) counties by the population size of their metro area, and nonmetropolitan (nonmetro) counties by degree of urbanization and adjacency to a metro area or areas (National Institutes of Health, 2014). Table 6 of Appendix A lists the education variables that were selected. This table also contains county name, state abbreviation, and the five-digit FIPS (Federal Information Processing Standards) code of the county. These variables were shared between all sheets before data merging, and are listed in Table 12 of Appendix A.

The population dataset was similar in content to the education data set and contained data for years 2020, 2021, and 2022. The year 2021 was selected based on other year related entries across the other five data sources. The final selected variables can be viewed in Table 7 of Appendix A. Finally, this dataset also contains the same shared variables contained in Table 12 of Appendix A.

The poverty dataset only contained data from the year 2021. Each listed variable for poverty had an additional upper- and lower-ninety percent confidence interval variable, which were dropped for ease of use with the multivariate models. The final selected variables can be viewed in Table 8 of Appendix A. Finally, this dataset also contains the same shared variables contained in Table 12 of Appendix A.

The employment dataset contained one hundred total variables, but only six employment specific variables were retained for the year 2021. The final selected variables can be viewed in

Table 9 of Appendix A. Finally, this dataset also contains the same shared variables contained in Table 12 of Appendix A.

2.1.2 Vaccine Hesitancy for COVID-19 (HHS ASPE)

COVID-19 vaccine hesitancy data was provided in a report prepared by the U.S. Department of Health and Human Services Office of the Assistant Secretary for Planning and Evaluation (U.S. Department of Health and Human Services & Office of the Assistant Secretary for Planning and Evaluation, 2021). These data were pulled from the Centers for Disease Control and Prevention's public website (Centers for Disease Control and Prevention, 2021). The ASPE estimated hesitancy rates using the U.S. Census Bureau's Household Pulse Survey (U.S. Census Bureau, 2024) data and utilized the estimated values to predict hesitancy rates at the Public Use Microdata Areas (PUMA) level using the Census Bureau's 2019 American Community Survey (ACS) 1-year Public Use Microdata Sample (PUMS). To create county-level estimates, they used a PUMA-to-county crosswalk from the Missouri Census Data Center. PUMAs spanning multiple counties had their estimates apportioned across those counties based on overall 2010 Census populations. This description of their methods is directly from the CDC's website (Centers for Disease Control and Prevention, 2021)

All variables that were provided were included for analysis except for Percent adults fully vaccinated against COVID-19 (as of 6/10/21). This variable was not used due to a large number of missing counties, including the entire state of Texas. Three different HPS vaccine hesitancy dimensions are included: estimated hesitant, estimated strongly hesitant, and estimated hesitant or unsure. For the bulk of analysis, estimated hesitant was used as the main outcome of interest. This variable represented the percentage of individuals who were either 'probably not' or 'definitely

not' going to receive a COVID-19 vaccine by county. This covered all individuals who reported a tendency to forgo the vaccine. Estimated strongly hesitant only included 'definitely not' individuals, while estimated hesitant or unsure included 'probably not' and 'definitely not' individuals, as well as individuals who were simply 'unsure' whether they would receive the vaccine.

This dataset also included variables like the Social Vulnerability Index (SVI) and the COVID-19 Vaccine Coverage Index (CVAC). SVI uses 16 U.S. census variables to help local officials identify communities that may need support before, during, or after disasters. (Agency for Toxic Substances and Disease Registry, 2020). CVAC captures supply- and demand-related challenges that may hinder rapid, widespread COVID-19 vaccine coverage in U.S. counties, through five specific themes: historic under-vaccination, sociodemographic barriers, resource-constrained healthcare system, healthcare accessibility barriers, and irregular care-seeking behaviors (Surgo Ventures, 2021). These data also included county level percentages of race, which provided a productive addition to the population table. All included values from this dataset can be viewed in Table 10 and Table 11 of Appendix A. This dataset also contains the same shared variables contained in Table 12 of Appendix A.

2.1.3 2020 U.S. Religion Census (ASARB)

Religious participation data was sourced from the 2020 United States Religion Census performed by The Association of Statisticians of American Religious Bodies (ASARB). These data contained count and rank variables for congregations and their adherents from 372 different faith groups within the United States at the national, state, county, as well as metropolitan area level (The Association of Statisticians of American Religious Bodies, 2023). For the scope of this

project, county level data was utilized. Additionally, for the purposes of this thesis religious participants were not divided by faith. Vaccine hesitancy related to religious doctrine is not specific to one faith, thus religion remain generalized to congregations of faith as well as adherents to a faith in a nonspecific way.

All variables included in the ASARB 2020 Summary data were included. These variables can be viewed in Table 13 of Appendix A. This data source included the FIPS code and county name that are contain in the shared variable table (Table 12) in Appendix A.

2.2 Data Cleaning

RStudio 2022.07.2 Build 576 was utilized in order to merge these six datasets. Packages utilized included *tidyverse* 1.3.2, *readxl* 1.4.1, *dplyr* 1.1.3, *rvest* 1.0.3, *htmlTable* 2.4.2, and *data.table* 1.14.1. FIPS codes and county/county equivalent names were harmonized between datasets, and the six sets were joined on the variables FIPS_Code, Area_Name, and State (state abbreviation). As previously addressed, custom, homogenous variable names were generated at the authors discretion for ease of use with later analysis and figure creation. Census Region and Census Division were also added to the dataset based on the State Name variables by the author for potential use with potential figure creation (U.S. Census Bureau et al., 2010).

There are two states that were uniquely difficult to clean. In 2021, Connecticut successfully voted to adopt a new form of county equivalents for the coming year of 2023, census planning regions (Federal Register, 2020). The state went from eight total counties to nine total census planning regions. This became a problem while cleaning the 2021 ERS data. All datasheets in that set recorded by county, except for the population dataset which preemptively recorded their

population metrics at the census planning region. There was no easy way to revert the data from the nine planning regions into the previous county format, and county level population data was not easily accessible for Connecticut in 2021. This created missing population variables cells. Supervised learning requires complete data, so these eight counties in Connecticut from before 2023 are not included in the later supervised learning procedures.

Alaska had somewhat similar difficulties during cleaning. Alaska contains a combination of municipalities, boroughs, and census areas for their county equivalents (U.S. Census Bureau, 2021). Every set had slightly differing areas listed, and some defunct county equivalents still listed. During cleaning, the thirty county equivalents listed from the census in 2020 were used as a guideline. Only two of the thirty areas had missing variables; the Chugach census area and the Copper River census area both had missing variables from the hesitancy dataset. These counties were also not included during later supervised learning procedures.

Two other counties contained missing values. Kalawao county in Hawaii did not contain data from the poverty dataset, but was not missing any other data from any other source. Rio Arriba county in New Mexico was missing a single variable, social vulnerability index (SVI). Both of these counties were also dropped for later supervised learning procedures. Overall, 12 out of 3143 counties/county equivalents contained missing values and were not included in the later supervised learning. This only represents 0.38% of all counties in the United States, and 3131 counties/county equivalents are used during supervised learning method selection, variable selection, and post-hoc clustering.

3.0 Methods

This chapter consists of four sections: Supervised Machine Learning, Variable Selection, Post-Hoc Clustering, and Computational Tools and Software. The Supervised Machine Learning section details the four supervised learning methods that were compared on the cleaned dataset. The Variable Selection section details the method by which the final significant variables were selected after a supervised learning method was chosen. The Post-Hoc Clustering Section provides an explanation of the methodology used to cluster counties based on the selected variables from the final model. The Computational Software and Tools section provides notes on the programs and statistical packages utilized as part of the analysis process.

3.1 Supervised Machine Learning

Four different supervised machine learning methods were compared to determine which technique yielded the most significant prediction. Statistical descriptions of these methods were summarized from *The Elements of Statistical Learning* (Hastie et al., 2004) and *A Machine Learning and Clustering-Based Approach for County-Level COVID-19 Analysis* (Nicholson et al., 2022). Root mean square error (RMSE) and R^2 were utilized as measures of prediction for all models. A randomized training testing split was utilized with both the random forest and gradient boosted tree approach. These techniques both require a randomized calibration set before being applied to a testing set, and in this case 70% of the data was used to train and 30% to test.

3.1.1 Elastic Net Regression (ENET)

Elastic net regression is a regression technique that reduces overfitting and performs automatic feature selection within a linear regression model (Nicholson et al., 2022). It accomplishes this goal by using both the L1 penalization method of the lasso (least absolute shrinkage and selection operator) method and the L2 penalization method from ridge regression. The hyperparameters that are tuned as part of this method are the penalty weight and the mixing parameter associated with balancing the L1 and L2 elements in the cost function (λ_1 and λ_2). Input variables are ranked in order of overall model importance using the t-values associated with the $\hat{\beta}_{ENET}$ coefficients in the equation below.

$$\hat{\beta}_{ENET} = \left(1 + \frac{\lambda_2}{n}\right)\left\{\underset{\beta}{\arg min} \left\| \boldsymbol{y} - \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{x_j}\beta_j \right\|^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2\right\} \quad (1)$$

where $x_1, \ldots, x_m$ are m predictors and $y = (y_1, \ldots, y_m)^T$ is the response variable for n observations.

3.1.2 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regressions Splines (MARS) provides a regression procedure that is well suited for high dimension problems. The basic equation that is utilized for this method is provided below (Hastie et al., 2004).

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \quad (2)$$

Where each $h_m(X)$ is a new function from the permitted set C, or the product of two or more such functions. New features that isolate the ranges of values from the original input data are created through the use of hinge functions (Nicholson et al., 2022). After this hinging process, variables

and their interactions are added sequentially to this piecewise linear regression model. In addition to this regression process, this technique utilizes a stepwise, backwards elimination procedure in order to reduce the number of features. This process also optimizes the generalized cross validation performance statistic (GCV) in order to affect the number of parameters based on the size for the given number of terms. The importance of variables is thus derived from this GCV metric; the greater the effect of the presence of each variable has on the GCV, the more important that given variables is overall. GCV in this case takes on the form of the following equation (Hastie et al., 2004).

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}\left(y_i - \hat{f}_\lambda(x_i)\right)^2}{\left(1 - \frac{M(\lambda)}{N}\right)^2} \quad (3)$$

Where $M(\lambda)$ is the effective number of parameters in the model and $\hat{f}_\lambda$ is the estimated best model of each size λ. This occurs during the backwards elimination procedure.

3.1.3 Random Forest (RF)

Random forest models are a popular choice for classification and regression that are relatively simple to tune and train (Hastie et al., 2004). This technique relies on utilizing an ensemble of weak learners (Nicholson et al., 2022), which are a collection of models with weak predictive ability that evolve over time to create a final idealized predictive model. In a regression setting, a regression tree is fitted to many bootstrapped samples, a method of sampling where new datasets are drawn with replacement from a designated training subset of data, and those results are averaged. These samples are the same size as the original training dataset. This trained model can then be applied to a testing subset, and the fit of the model and the importance of individual

variables can be assessed. The importance of variables in a random forest is determined by which variables improve the mean square error. The greater the increase in mean square error for a given variable, the more important that variable is to the overall model.

Random forest models for regression take on the form below (Hastie et al., 2004). After B trees are grown, this equation is informed by the total number of trees as defined by $\{T(x;\Theta_b)\}_1^B$.

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b) \tag{4}$$

Where Θ_b characterizes the bth random forest tree in terms of split variables, cut points at each node of the forest, and terminal-node values.

3.1.4 Gradient Boosted Trees (GBT)

Gradient boosted tree models share many of the traits that make random forest models so effective. This model also utilizes weak learners for regression. However, GBT differs by construction. This model builds a predefined number of relatively simple decision trees where each subsequent tree is constructed based on the results of the previous tree's predictive error. This provides an iterative approach that random forests do not have. The alterable hyperparameter values, which are simply alterable facets of the model, for GBT models include the number of trees to fit, the maximum depth of each tree, the learning rate, and the minimum number of observations in the terminal nodes of the trees (Nicholson et al., 2022). Gradient boosted tree model also differs in regard to variable importance. GBT models measure importance by how often a feature is selected in the construction of underlying trees.

This process begins with an equation of the optimal constant model, which is part of a greater algorithm GBT modeling for regression and is a single terminal node tree. This initialized first form of a generic GBT model is provided below (Hastie et al., 2004).

$$f_0(x) = \arg min_\gamma \sum_{i=1}^{N} L(y_i, \gamma) \quad (5)$$

Where $L(y_i, \gamma)$ is the inserted loss criteria, which serves a similar role to the hyperparameter.

3.2 Method Selection

These four machine learning techniques were compared using both R^2 and RMSE. Gradient Boosted Trees was selected as the most optimal model based on these two metrics. Given the full, cleaned variable set, the GBT technique minimized root mean square error while maximizing R^2.

3.3 Variable Selection

Variable importance was determined for the GBT model with all variables included, and a number of models were run with increasing numbers of variables included while comparing their RMSE values and R^2 values. GBT models measure importance by how often a feature is selected in the construction of underlying trees. Variables were added into the model by decreasing order of overall importance. When those values reached a global maximum/minimum respectively, the dimensions that were included within the model in which those values were idealized was selected and included. This represented the overall most effective model at predicting vaccine hesitancy.

3.4 Post-Hoc Clustering

Hierarchical clustering (HC) was chosen in order to visualize and contextualize the 26 chosen dimensions geographically. This technique produces hierarchical representations of cluster profiles in a tree structure where each cluster at every level is created by merging clusters at the subsequent level where the highest level is one cluster containing all datapoints (Hastie et al., 2004). This type of bottom to top clustering is referred to as agglomerative and is the most practical and popular approach to hierarchical clustering. This technique also utilized Ward's method to obtain the distance between clusters instead of single linkage or complete linkage. Ward's method minimizes the within sum of squares error at every iteration while combining clusters.

Let C_i and C_j denote two mutually exclusive clusters consisting of n_i and n_j points, respectively. Let $d(C_i, C_j)$ denote the dissimilarity between C_i and C_j. Ward's method computes dissimilarity as the increase in the sum of squares if C_i and C_j are merged. Mathematically, this is equivalent to:

$$d_{Ward}(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \left\| \mu_{C_i} - \mu_{C_j} \right\|^2 \tag{6}$$

Where μ_{C_i} and μ_{C_j} are the mean clusters of C_i and C_j, respectively (Nicholson et al., 2022).

The gap statistic method was utilized to select the number of clusters. This method leverages Monte Carlo simulation in order to help determine the optimal number of clusters and is applicable to the gradient boosted tree method (Tibshirani et al., 2001). This method largely outperforms more traditional methods of cluster selection, and provides a figure to visually inform the selection of an appropriate number of clusters.

3.5 Computational Software and Tools

RStudio 2022.07.2 Build 576 was utilized in order to perform these analyses. Below is a bulleted list of the packages that were used by the analysis procedure they were used in.

- **ENET:** *glmnet* 4.1-8, *caret* 6.0-94, *dplyr* 1.1.3, *ggplot2* 3.4.0
- **MARS:** *earth* 5.3.2, *caret* 6.0-94, *dplyr* 1.1.3, *ggplot2* 3.4.0
- **RF:** *randomForest* 4.7-1.1, *caret* 6.0-94
- **GBT/Variable Selection:** *xgboost* 1.7.5.1, *caret* 6.0-94, *dplyr* 1.1.3
- **Hierarchical Clustering:** *factoextra* 1.0.7, *cluster* 2.1.6

4.0 Results

This chapter contains five sections: Supervised Machine Learning Results, Variable Selection Results, Model Comparison, Post-Hoc Clustering results, and Graphical Representation of Hesitancy. The Supervised Machine Learning Results section compares the outcomes of the machine learning methods. The Variable Selection Results section details which variables were selected for the final model. The Model Comparison section compares the final model with a model with religion removed. The Post-Hoc Clustering Results section contains details about clustering analysis performed. Graphical Representation of Hesitancy shows a map of vaccine hesitancy.

4.1 Supervised Machine Learning Results

Three total outcome variables were considered (Table 1). As detailed in section 2.1.2, estimated hesitant was used as the main outcome of interest. This variable represented the percentage of individuals of in a county who were either ‘probably not’ or ‘definitely not’ going to receive a COVID-19 vaccine when it becomes available to them. This particular category represents all individuals who reported a tendency to forgo the vaccine, which promotes conclusions related to vaccine hesitancy as a public health outcome. Estimated strongly hesitant only included ‘definitely not’ individuals, while estimated hesitant or unsure included ‘probably not’ and ‘definitely not’ individuals, as well as individuals who were simply ‘unsure’ whether they would receive the vaccine. These estimates were based on the Household Pulse Survey, and Table 1 provides a formal description of the national survey questions that were described above.

Table 1: Available Outcome Variables

Outcome	Description
ESTHES (Estimated Hesitant)	Estimate of percentage of adults who describe themselves as "probably not" or "definitely not" going to get a COVID-19 vaccine once one is available to them, based on national survey data
ESTHESoUNS (Estimated Hesitant or Unsure)	Estimate of percentage of adults who describe themselves as "unsure", "probably not", or "definitely not" going to get a COVID-19 vaccine once one is available to them, based on national survey data
ESTSTRHES (Estimated Strongly Hesitant)	Estimate of percentage of adults who describe themselves as "definitely not" going to get a COVID-19 vaccine once one is available to them, based on national survey data

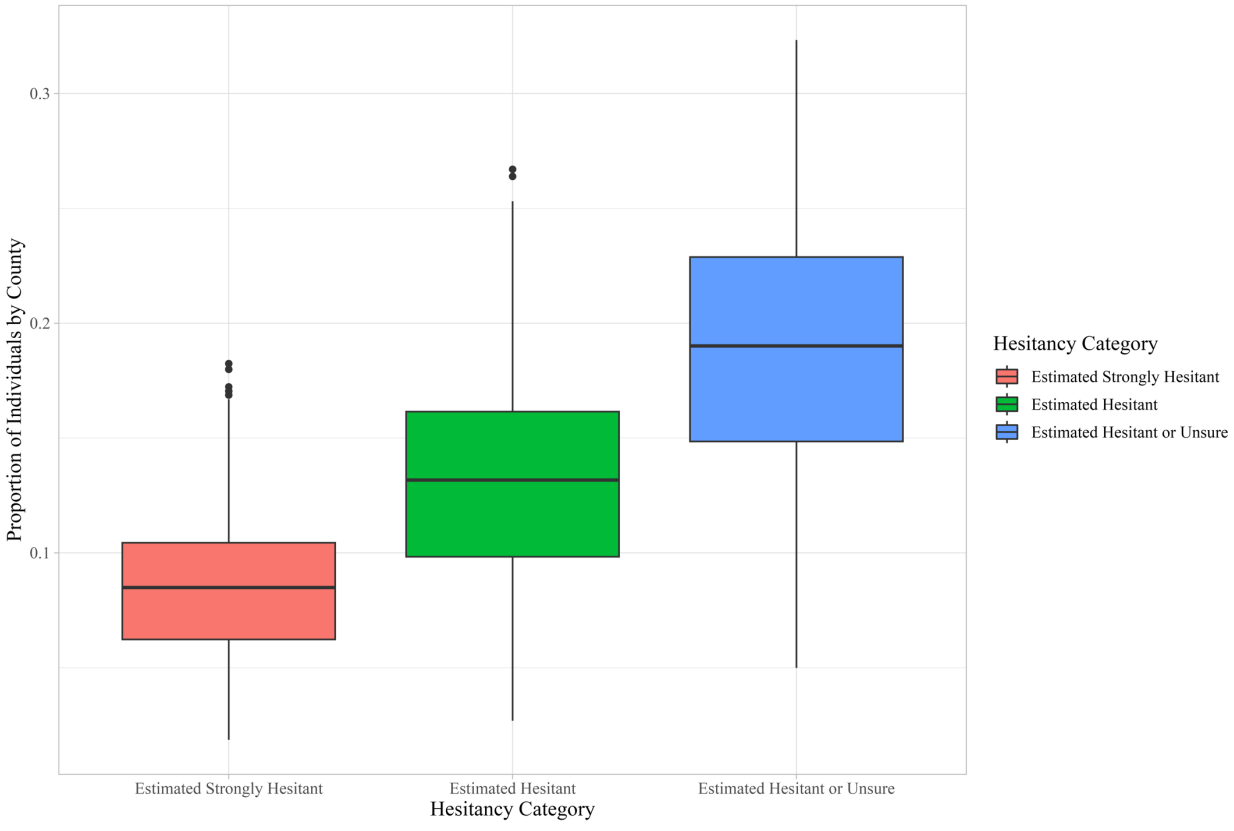

Figure 1: Boxplot of Outcome Variables (Hesitancy Category)

Using estimated hesitant as the chosen outcome variable, four machine learning methods detailed in the previous section were performed on the cleaned data. During the process of preparing the data for the machine learning analysis, the nonnumeric variables included within the data were excluded. This included the variables State Name, State (state abbreviation), County Name, Census Region, Census Division, Geographical Point, County Boundary, State Boundary, SVI Category, and CVAC Level of Concern. These variables are denoted in the tables in Appendix A with an asterisk. Additionally, the five-digit FIPS Code variable was converted to the row names of the dataframe. This dimension is denoted in the Selected Shared Variables table (Table 12) in Appendix A with a double asterisk. This left 1 outcome variable and 54 variables for the supervised machine learning model analysis. The result of this analysis is contained in Table 2.

Table 2: Supervised Learning Results

Outcome	**Metric**	**Supervised Learning Method**			
		ENET	MARS	RF	GBT
Estimated Hesitant	RMSE	0.0328	0.0292	0.0303	**0.0282**
	R^2	0.5122	0.6016	0.5842	**0.6217**

The supervised learning model that minimized the RMSE and maximized R^2 was the gradient boosted tree model, and was thus determined to be the most productive model at predicting estimated vaccine hesitancy based on the 54 variables included in the dataset. This model was utilized in the subsequent variable selection section.

4.2 Variable Selection Results

Of the 54 included variables in the gradient boosted tree analysis, 26 were found to be critical to overall predictive ability. Tables 3 and 4 provide the variable name and category, along with a variable description and its gain value, an indication of variable importance. Variables within their demographic categories are ranked in descending order by gain value, a measure of overall variable importance.

Table 3: GBT Critical Dimensions (1)

Var. Cat.	Var. Name	Description	Gain
Hesitancy	CVACLOCVR	Surgo Covid-19 Vaccine Coverage Index (CVAC) level of concern for vaccination rollout	0.0797
	PctAMINAN	Percent of county population that is non-Hispanic American Indian/Alaska Native	0.0577
	PctHISP	Percent of county population that is Hispanic	0.0540
	PctBLK	Percent of county population that is non-Hispanic Black	0.0310
	PctWHI	Percent of county population that is non-Hispanic White	0.0238
	PctASN	Percent of county population that is non-Hispanic Asian	0.0164
	SVI	2018 CDC Social Vulnerability Index (SVI)	0.0117
Religion	CONGpc	Number of religious congregations per capita	0.0722
	PctADH	Religious practitioners as a percentage of entire population	0.0170
Education	PctHSDO	Percent of adults with only a high school diploma	0.0319
	PctLtHSD	Percent of adults with less than a high school diploma	0.0236
	PctSCoAD	Percent of adults with some college or an associate's degree	0.0188
	HSDO	Count of adults with only a high school diploma	0.0064

Table 4: GBT Critical Dimensions (2)

Var. Cat.	Var. Name	Description	Gain
Population	RtNETMIG	Net migration rate in period 7/1/2020 to 6/30/2021	0.0266
	RtBRTH	Birth rate in period 7/1/2020 to 6/30/2021	0.0261
	RtDTH	Death rate in period 7/1/2020 to 6/30/2021	0.0155
	POPCHNG	Numeric change in resident total population 7/1/2020 to 7/1/2021	0.0126
	GQEST	7/1/2021 Group Quarters total population estimate	0.0095
	NCHNG	Natural change in period 7/1/2020 to 6/30/2021	0.0068
Poverty	PctPOVALL	Estimated percent of people of all ages in poverty 2021	0.0277
	PctPOV517	Estimated percent of related children aged 5-17 in families in poverty 2021	0.0157
	POVALL	Estimate of people of all ages in poverty 2021	0.0107
Unemployment	MEDHHINC	Estimate of median household income, 2021	0.2054
	PctMEDHHINC	County household median income as a percent of State total median household income, 2021	0.1364
	RtUNEMP	Unemployment rate, 2021	0.0549
	UNEMP	Number unemployed annual average, 2021	0.0083

Figure 2 contains a plot of variable importance. The two most critical variables to this model were county median household income and county median household income as a percent of state total median household income, which both come from the unemployment dataset. Also note that CONGpc, a religion metric, has the fourth highest importance to the model when applied to the training subset.

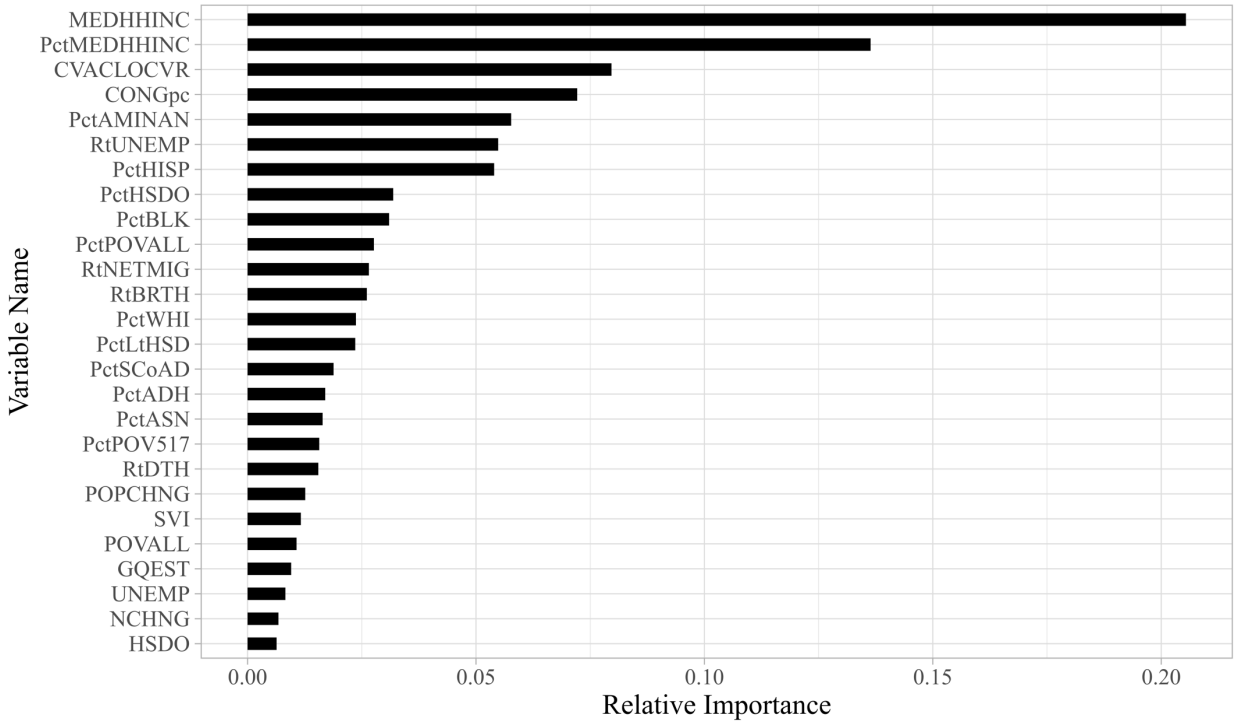

Figure 2: Relative Importance Plot for Final GBT Model

The model with these 26 critical variables generated an RMSE of 0.0261 and an R^2 of 0.6549. Both of these metrics represent an increase in predictive ability from the full model with all dimensions included. To note CONGpc, or the number of religious congregations per capita, and PctADH, of religious practitioners as a percentage of entire population, are both critical dimensions for this multivariate, demographically oriented model.

4.3 Model Comparison Results

This section is focused on comparing a model with the 26 critical variables with a model with the 24 non-religious critical variables. This was performed to verify that the critical religion

metrics in combination with the 24 additional variables promoted the prediction ability of the final model. Table 5 displays the significant decrease in RMSE and increase R^2 demonstrated by the model with the critical religious variables include as compared to the model with the critical religious variables excluded.

Table 5: Critical Religious Dimension Model Comparison

Outcome	Metric	Dataset	
		24 Dimensions (No Religion)	26 Dimensions (With Religion)
Estimated Hesitant	RMSE	0.0277	**0.0261**
	R^2	0.6123	**0.6549**

4.4 Post-Hoc Clustering Results

The goal of the clustering analysis is to identify how the critical factors utilized within the final model relate to one another and to geographically represent similarly grouped factors for the United States. As a preliminary step, a correlation plot of the 26 variables was generated in order to review preliminary relationships (Figure 3). The most significant correlations occur between variables that originated from the same dataset, but there were a few relationships that occurred between variables in the different datasets. These relationships are denoted in the chart by darker shades of either red or blue depending on the directionality of the relationship.

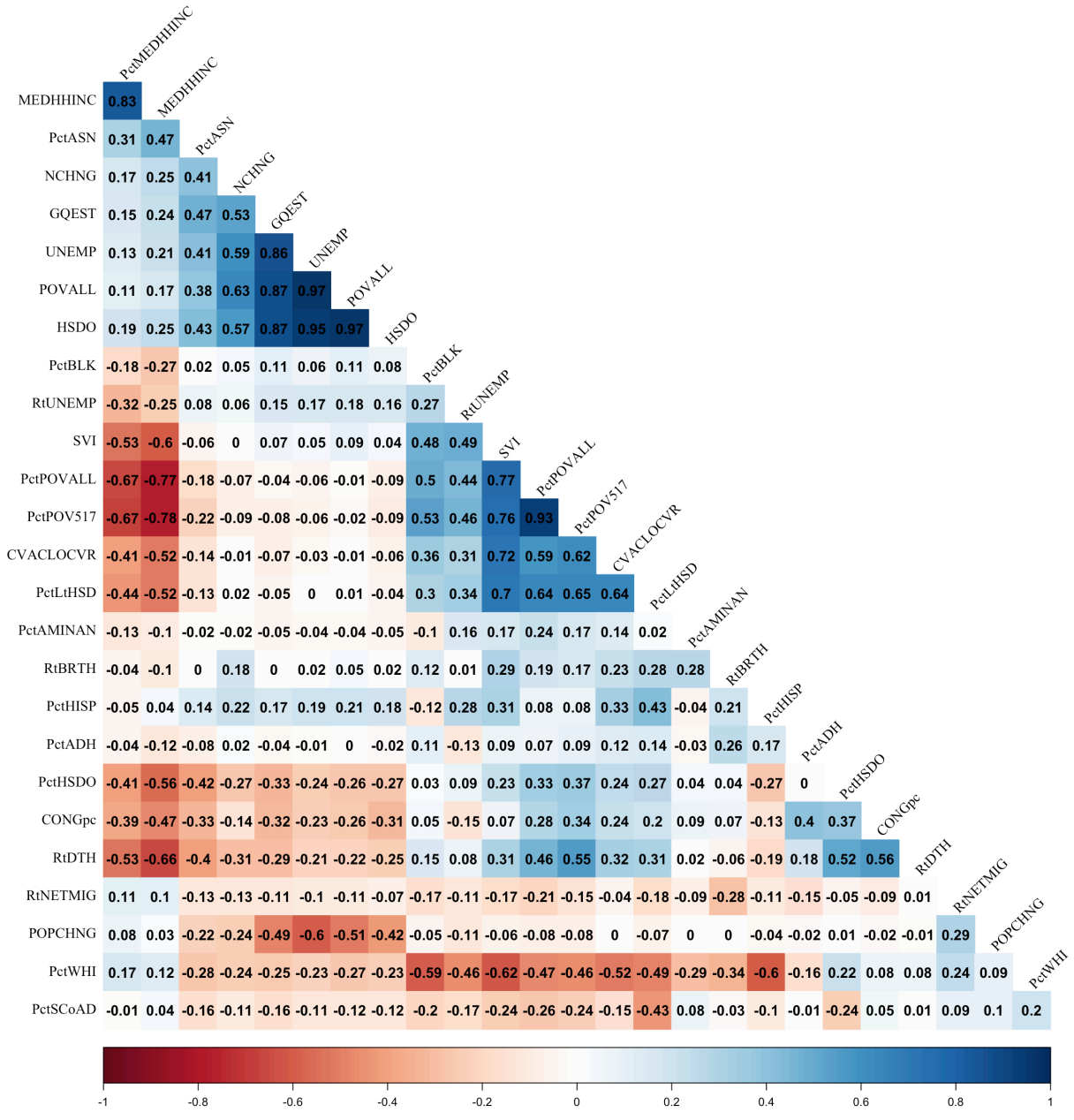

Figure 3: Correlation Plot of Critical Dimensions

To highlight a few interesting correlations, POVALL (Estimate of all people in poverty), HSDO (Count of adults with only a high school diploma), UNEMP (annual average of the number of unemployed people), and GQEST(Estimate of the population living in group quarters) were

highly correlated with one another. Every correlation between these four variables had a correlation coefficient greater than 0.85, and represented the strongest correlations between variables that did not originate from the same data source.

The first goal of the cluster analysis was to select the correct number of clusters. To assess this, the gap statistic (Nicholson et al., 2022) was utilized, and Figure 4 provides a graphical representation of the gap statistic with the number of k clusters.

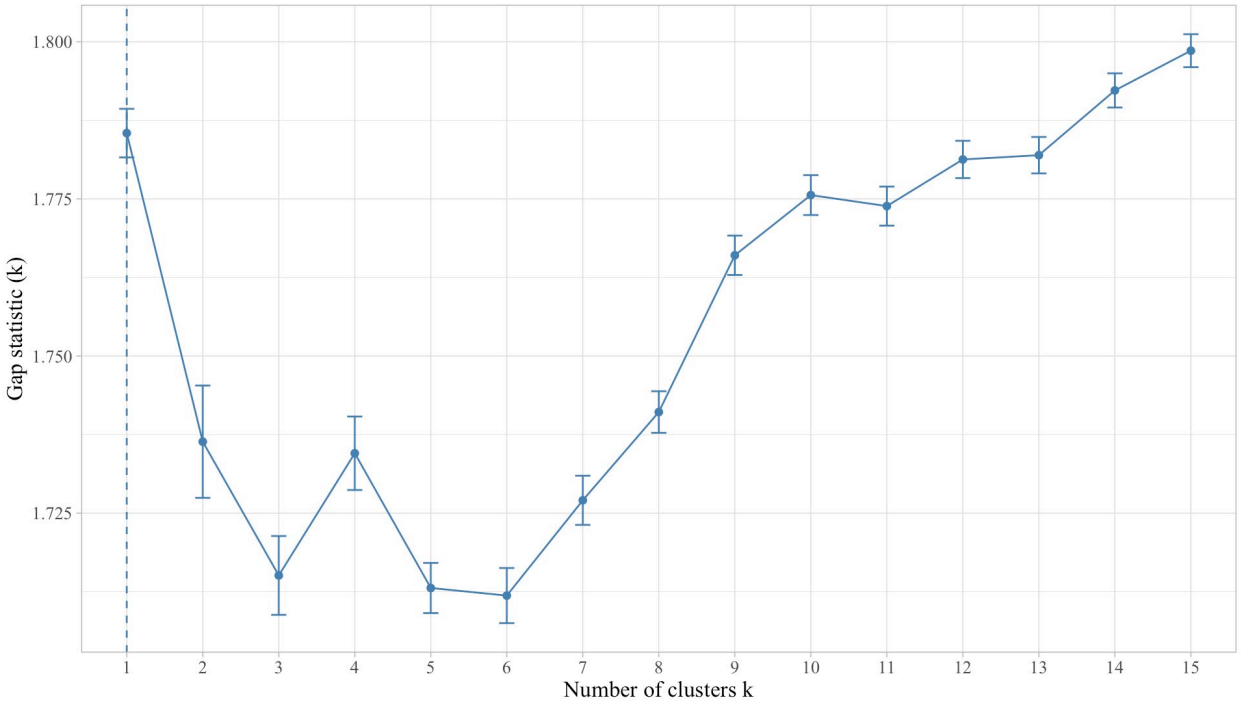

Figure 4: Number of Clusters Plotted by Gap Statistic

Based on Figure 4, nine clusters were selected for hierarchical clustering on this data. Typically, the ideal number of clusters occurs when the gap statistic is maximized. In this case, this chart does not have a useful global maximum (first local maxima are at 1 and 15 clusters) and thus cluster selection by way of the gap statistic became more discretionary. After visually

inspecting Figure 4 and analyzing a number of cluster profiles with differing numbers of groups ranging from 6 to 10, 9 was determined to be an appropriate number of clusters for the scope of this clustering analysis and to support further research in this machine learning niche of public health. Nine clusters generated a number of county level groupings that allowed for some demographic patterns to appear on rural and urban divides as well as regional and divisional boundaries that were not reflected in other cluster profiles. Figure 5 provides a geographical representation of the nine clusters for the 26 critical dimensions.

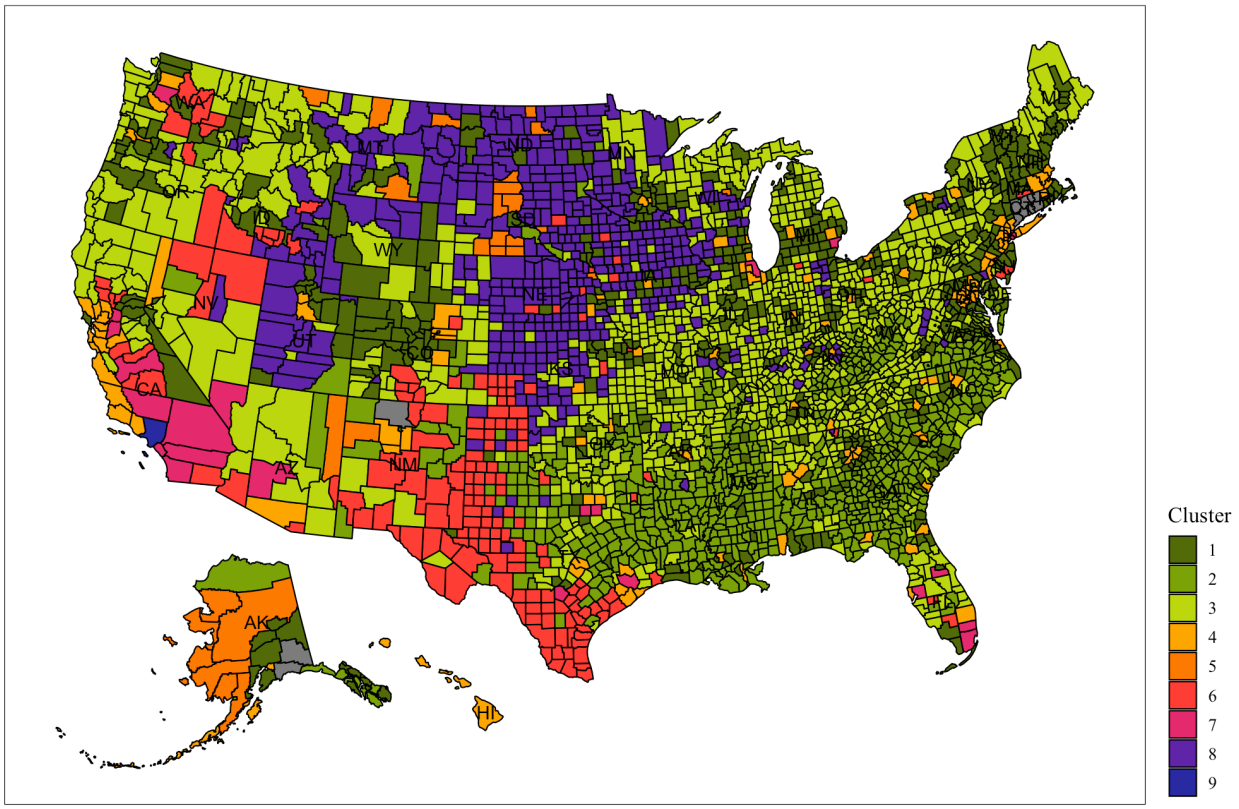

Figure 5: Graphical Representation of Hierarchical Clusters

Clusters 1 and 3 seem to exist throughout much of the United States. Both of these clusters can be seen next to one another throughout the Northeast, Midwest, Lower Appalachia, the Pacific

Northwest, and parts of Colorado, Wyoming, Arizona, and Nevada. Cluster 2 appears to take up predominantly the South, stretching from Virginia to Texas. Cluster 4 is relatively rare, representing all of Hawaii as well as a number of highly populous metropolitan areas such as the areas surrounding New York City, Washington D.C., Denver, Albuquerque, and Pittsburgh. Cluster 5 is even more sparse, and represents a large portion of Alaska as well as parts of the Dakotas. Cluster 6 is situated largely within the Southwest, taking up large portions of Texas, New Mexico, and Oklahoma. Cluster 7, another small cluster, largely represents parts of central and southern California, a county in Arizona, and a handful of counties in southern Florida. Cluster 8, the third large cluster is almost entirely located within the Northern United States with counties in Wisconsin, Minnesota, the Dakotas, Montana, Nebraska, Kansas, and as far west as Utah and Idaho. Cluster 9 is an outlier. That cluster represents Los Angeles County, which is understandably an area demographically different enough from all other counties to warrant its own cluster.

Each cluster also has a number of standout demographic characteristics that differentiate it from other groupings. The full breakdown of cluster profiles can be reviewed in Table 14 of Appendix B. Below is a bulleted list describing some unique cluster features.

- **Cluster 1** has a particularly low poverty rate as well as a particularly low proportion of individual with less than a high school diploma. This cluster also has the second lowest unemployment rate and highest level of concern for vaccine rollout of all groupings.
- **Cluster 2** has a higher poverty rate than most clusters as well a low median household income and high proportion of religious adherents. This cluster also has by far the highest proportion of black individuals of all clusters, a high social vulnerability index, and the highest death rate.
- **Cluster 3** is a predominantly white grouping. This cluster has a low birth rate, high death rate, a particularly high rate of net migration, and a very low proportion of all other racial groupings.

- **Cluster 4** has a very high number of individuals estimated to be living in group quarters. this cluster also has the highest median household income of all groupings, and has a low overall poverty rate compared to other clusters.
- **Cluster 5** has the highest proportion of American Indian/Alaska Native individuals of all groupings and the lowest proportion of both black and white individuals. This cluster also has the highest birth rate, the lowest group quarters estimate and the highest overall poverty rate.
- **Cluster 6** has the highest proportion of Hispanic individuals of all groupings and the highest proportion of individuals with less than a high school diploma. This cluster also has the second highest social vulnerability score and the second highest proportion of religious adherents.
- **Cluster 7** has the lowest congregations per capita than all other groupings. This cluster has the second highest proportion of both Asian and Black individuals, and a relatively low death rate.
- **Cluster 8** This cluster has the highest proportion of religious adherents, the highest congregations per capita, and the lowest social vulnerability score. This cluster also has the highest proportion of white individuals, the lowest unemployment rate, the largest proportion of individuals with some college or an associate's degree, and second lowest poverty rate.
- **Cluster 9** has the highest unemployment rate, the lowest birth and death rates, and the lowest congregations per capita of all groupings. This cluster also has by far the lowest rate of net migration, the lowest proportion of individuals with some college or an associate's degree, and a particularly high social vulnerability index.

4.5 Graphical Representation of Hesitancy

This section includes a graphical representation of estimated vaccine hesitancy. This figure is to provides a comparison between the county map plot of the generated clusters and the geographical distribution of the estimated hesitant variable.

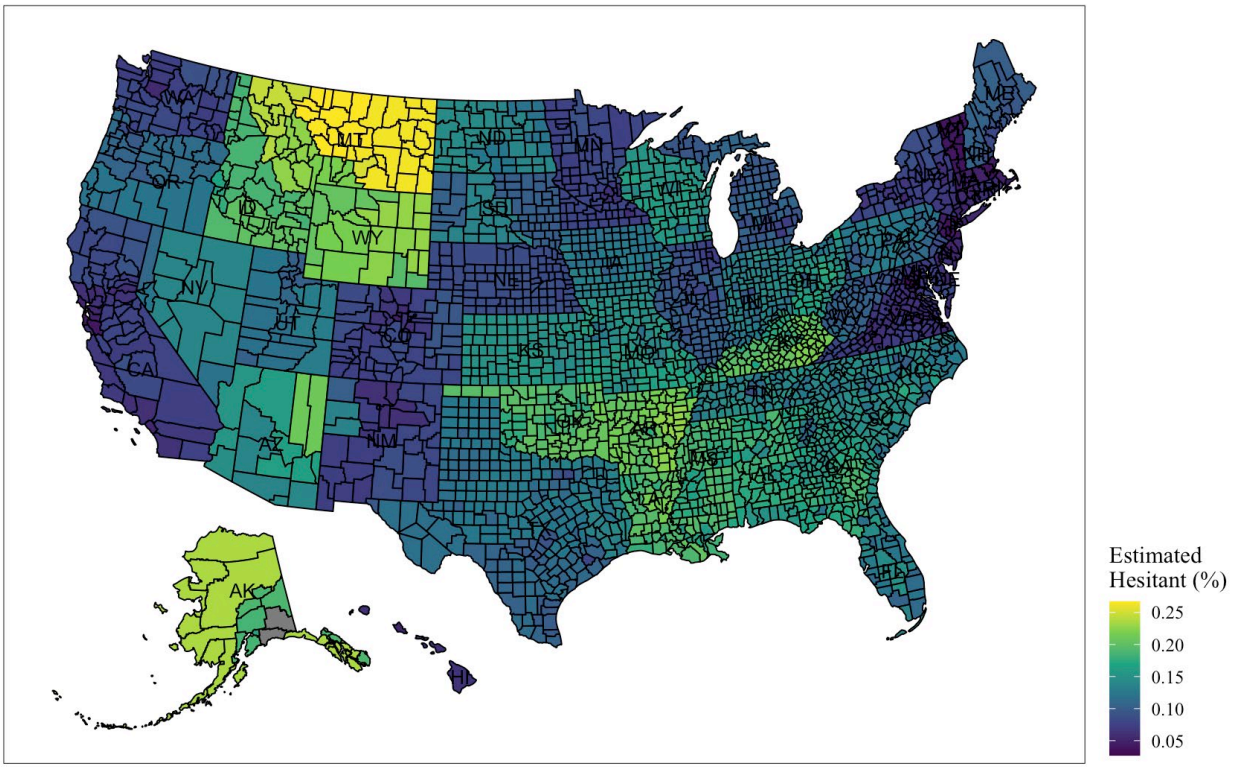

Figure 6: Graphical Representation of Estimated Hesitant Variable

In contrast to Figure 5, Figure 6 provides less of a commentary about rural/urban and regional differences as much as it displays a state-by-state contrast in hesitancy. In a number of instances on this map, rigid gradient differences can be seen on state lines. States like Texas and Wisconsin, for example, are a visually distinct color than the states that surround them. This is to

be expected given that this gradient scale is built on one variable, but it is true that on the whole this estimated hesitancy variable appears to be more homogenous within states than between states.

5.0 Discussion

This chapter contains two sections: Research Implications and Future Work. The Research Implications section details the significant outcomes from this project, and the Future Work section describes the ways in which this work could be extended or expounded upon.

5.1 Research Implications

This project has demonstrated the merit of incorporating religious participation into multivariate machine learning models regarding vaccine hesitancy. Despite the relatively moderate predictive ability of this model ($R^2 = 0.6549$), a similar model with religious variables removed performed substantially less optimally ($R^2 = 0.6123$). Vaccine hesitancy is a difficult public health phenomenon to attempt to predict and this work promotes the idea of incorporating estimated social factors such as religious participation into predicative hesitancy models. Additionally, this project supports the specific use of the gradient boosted trees as an ideal method with a mixed data source model such as the one generated here. This, however, may not be entirely generalizable to all mixed data projects. It is also interesting for the sake of this project that both a measure of adherence as well as a measure congregation are critical to the model's success. This promotes the idea that religion may be a productive metric to include in public health data related supervised machine learning research.

The implications of this work are potentially wide-reaching. Public health outcomes are intimately tied to social behavior, and machine learning utilization in the field of public health is

in its infancy. In future work with similar statistical learning methodology, this thesis supports incorporation of varied forms of data related to social behavior. Public health prediction is a complex field that is deeply tied to human experience and the communities that individuals inhabit. It is necessary to view problems in the field through a multi-faceted lens, and religion will never cease to influence the ways in which people and congregations approach their personal and collective health. Another goal of this thesis was to provide clustering analysis for future machine learning work, as well as to provide a geographic comparison between demographic groupings from this data and vaccine hesitancy estimates. This sort of profiling analysis is not necessarily unique, but potentially useful to describe future vaccine hesitancy.

5.2 Future Work

Future work could include that application of this model onto regional and divisional subgroups to be able to compare where these 26 dimensions apply in the most productive ways. The final model was trained on a national dataset, but it would be interesting to analyze the predictive ability on more granular, compartmentalized county level samples. It might also be productive to train regional models to be applied to regional data sets as well. The included clustering analysis could also be utilized to create training and testing subgroups for more specific machine learning models.

The final model of this thesis is not widely applicable to hesitancy of all vaccines, and the scope of this work was intentionally narrow in that way providing a model-based snapshot into a world grappling with the rollout of a new vaccine for COVID-19. All demographic data for this project was for the years of 2020 or 2021, a time where the world and its attitudes towards

vaccination, as well as the number of available vaccines and vaccine preventable diseases, have since changed in a number of ways. It is important to note, however, that the methodology presented was utilized because of its potential adaptability to other public health projects, and this work affirms the idea that the process outlined within Nicholson et al. (2022) provides a flexible and productive framework for analyzing high dimensional mixed demographics data. This methodology could easily be applied to other vaccine related outcomes, particularly in the field of COVID-19 research as it was originally intended.

It would be productive to add more literature based social metrics to statistical learning models in the area of vaccines. The aim of this project was to analyze the contribution of religious metrics but there are a litany of different factors that influence vaccine hesitancy that may enhance the predictive ability of this model. It is entirely possible that the inclusion of other dimensions in fields such as of voting records or political representations would sharpen this model and provide a more holistic representation of hesitancy on a county-by-county basis.

6.0 Conclusion

In summation, after assembling an array of demographics-based variables including religious participation, vaccine hesitancy was assessed with a number of machine learning methods to determine an optimized approach. Gradient boosted trees was chosen as an ideal method of prediction for this data. After determining variable importance and selecting variables, the included analysis supports the contribution of religious participation variables on the overall predictive ability of machine learning models, particularly that of GBT models. The formally selected religion metrics improved the prediction ability of the final model, which is significant to public health at large. Public health outcomes are, and always will be tied to social behavior. This thesis contributes to the growing body of work that promotes the incorporation of social metrics and data within large data work and county level analysis with public health outcomes. In addition, clustering analysis of the selected variables yielded a number of productive trends and observations about the demographics of the United States as compared to vaccine hesitancy.

Appendix A Variable Tables for Data Sources

Table 6: Selected Education Data Variables

Var. Name	Variable Description	Analysis Var. Name
Less than a high school diploma, 2017-21	Count of adults with less than a high school diploma	LtHSD
High school diploma only, 2017-21	Count of adults with only a high school diploma	HSDO
Some college or associate's degree, 2017-21	Count of adults with some college or an associate's degree	SCoAD
Bachelor's degree or higher, 2017-21	Count of adults with a bachelor's degree or higher	BDOH
Percent of adults with less than a high school diploma, 2017-21	Percent of adults with less than a high school diploma	PctLtHSD
Percent of adults with a high school diploma only, 2017-21	Percent of adults with only a high school diploma	PctHSDO
Percent of adults completing some college or associate's degree, 2017-21	Percent of adults with some college or an associate's degree	PctSCoAD
Percent of adults with a bachelor's degree or higher, 2017-21	Percent of adults with a bachelor's degree or higher	PctBDoH

Table 7: Selected Population Data Variables

Var. Name	Variable Description	Analysis Var. Name
CENSUS_2020_POP	4/1/2020 resident Census 2020 population	POP_2020
POP_ESTIMATE_2021	7/1/2021 resident total population estimate	POPEST_2021
N_POP_CHG_2021	Numeric change in resident total population 7/1/2020 to 7/1/2021	POPCHNG
BIRTHS_2021	Births in period 7/1/2020 to 6/30/2021	BRTH
DEATHS_2021	Deaths in period 7/1/2020 to 6/30/2021	DTH
NATURAL_CHG_2021	Natural change in period 7/1/2020 to 6/30/2021	NCHNG
INTERNATIONAL_MIG_2021	Net international migration in period 7/1/2020 to 6/30/2021	INTLMIG
DOMESTIC_MIG_2021	Net domestic migration in period 7/1/2020 to 6/30/2021	DOMMIG
NET_MIG_2021	Net migration in period 7/1/2020 to 6/30/2021	NETMIG
RESIDUAL_2021	Residual for period 7/1/2020 to 6/30/2021	RES
GQ_ESTIMATES_2021	7/1/2021 Group Quarters total population estimate	GQEST
R_BIRTH_2021	Birth rate in period 7/1/2020 to 6/30/2021	RtBRTH
R_DEATH_2021	Death rate in period 7/1/2020 to 6/30/2021	RtDTH
R_NATURAL_CHG_2021	Natural increase rate in period 7/1/2020 to 6/30/2021	RtNCHNG
R_INTERNATIONAL_MIG_2021	Net international migration rate in period 7/1/2020 to 6/30/2021	RtINTLMIG
R_DOMESTIC_MIG_2021	Net domestic migration rate in period 7/1/2020 to 6/30/2021	RtDOMMIG
R_NET_MIG_2021	Net migration rate in period 7/1/2020 to 6/30/2021	RtNETMIG

Table 8: Selected Poverty Data Variables

Var. Name	Variable Description	Analysis Var. Name
POVALL_2021	Estimate of people of all ages in poverty 2021	POVALL
PCTPOVALL_2021	Estimated percentage of people of all ages in poverty 2021	PctPOVALL
POV017_2021	Estimate of people aged 0-17 in poverty 2021	POV017
PCTPOV017_2021	Estimated percentage of people aged 0-17 in poverty 2021	PctPOV017
POV517_2021	Estimate of related children aged 5-17 in families in poverty 2021	POV517
PCTPOV517_2021	Estimated percentage of related children aged 5-17 in families in poverty 2021	PctPOV517

Table 9: Selected Unemployment Data Variables

Var. Name	Variable Description	Analysis Var. Name
Civilian_labor_force_2021	Civilian labor force annual average, 2021	CLF
Employed_2021	Number employed annual average, 2021	EMP
Unemployed_2021	Number unemployed annual average, 2021	UNEMP
Unemployment_rate_2021	Unemployment rate, 2021	RtUNEMP
Median_Household_Income_2021	Estimate of median household income, 2021	MEDHHINC
Med_HH_Income_Percent_of_State_Total_2021	County household median income as a percent of State total median household income, 2021	PctMEDHHINC

Table 10: Selected Hesitancy Data Variables (1)

Var. Name	Variable Description	Analysis Var. Name
Estimated hesitant	Estimate of percentage of adults who describe themselves as "probably not" or "definitely not" going to get a COVID-19 vaccine once one is available to them, based on national survey data	ESTHES
Estimated hesitant or unsure	Estimate of percentage of adults who describe themselves as "unsure", "probably not", or "definitely not" going to get a COVID-19 vaccine once one is available to them, based on national survey data	ESTHESoUNS
Estimated strongly hesitant	Estimate of percentage of adults who describe themselves as "definitely not" going to get a COVID-19 vaccine once one is available to them, based on national survey data	ESTSTRHES
Social Vulnerability Index (SVI)	SVI values range from 0 (least vulnerable) to 1 (most vulnerable)	SVI
*SVI Category	SVI categorized as follows: Very Low (0.0-0.19), Low (0.20-0.39); Moderate (0.40-0.59); High (0.60-0.79); Very High (0.80-1.0)	SVICAT
CVAC level of concern for vaccination rollout	CVAC Index values range from 0 (lowest concern) to 1 (highest concern)	CVACLOCVR
*CVAC Level of Concern	CVAC categorized as follows: Very Low (0.0-0.19), Low (0.20-0.39); Moderate (0.40-0.59); High (0.60-0.79); Very High (0.80-1.0)	CVACLOC

***Nonnumeric variables**

Table 11: Selected Hesitancy Data Variables (2)

Var. Name	Variable Description	Analysis Var. Name
Percent Hispanic	Percent of county population that is Hispanic	PctHISP
Percent non-Hispanic American Indian/Alaska Native	Percent of county population that is non-Hispanic American Indian/Alaska Native	PctAMINAN
Percent non-Hispanic Asian	Percent of county population that is non-Hispanic Asian	PctASN
Percent non-Hispanic Black	Percent of county population that is non-Hispanic Black	PctBLK
Percent non-Hispanic Native Hawaiian/Pacific Islander	Percent of county population that is non-Hispanic Native Hawaiian/Pacific Islander	PctNHPI
Percent non-Hispanic White	Percent of county population that is non-Hispanic White	PctWHI
*Geographical Point	Geographical center point of the county	GP
*County Boundary	Multipolygon county boundaries	CB
*State Boundary	Multipolygon state boundaries	SB

***Nonnumeric variables**

Table 12: Selected Shared Variables

Var. Name	Variable Description	Analysis Var. Name
**FIPS Code	Five Digit County Level Federal Information Processing Standards Code	FIPS_Code
*County Name	Name of County/equivalent	Area_Name
*State	Abbrev. of State	State
*Census Region	Census defined Region	Census.Region
*Census Division	Census defined Division	Census.Division

***Nonnumeric variables **Later converted to rowname during analysis**

Table 13: Selected Religious Participation Variables

Var. Name	Variable Description	Analysis Var. Name
*State Name	State full Name	State_Name
Congregations	Groups that gather for religious worship	CONG
Adherents	Followers of religion	ADH
Congregations Per 100,000 Population	Congregations per capita	CONGpc
Adherents as % of Population	Religious practitioners as a percentage of entire population	PctADH
Population Rank	Population count ranking	POPRNK
Congregations Rank	Congregation count ranking	CNGRNK
Adherents Rank	Adherents count ranking	ADHRNK
Congregations Per 100,000 Pop. Rank	Rank of congregations per capita	CONGpcRNK
Adherents as % of Population Rank	Rank of religious practitioners as a percentage of entire population	PctADHRNK

***Nonnumeric variables**

Appendix B Hierarchical Cluster Profiles

Table 14: Dimensional Profiles of Each Cluster by Dimension Category

Cluster	1	2	3	4	5	6	7	8	9
Counties	**554**	**704**	**1015**	**141**	**27**	**180**	**26**	**483**	**1**
Hesitancy									
CVACLOCVR	0.28	0.77	0.51	0.34	0.80	0.79	0.62	0.25	0.71
PctHISP	0.07	0.08	0.05	0.17	0.03	0.50	0.35	0.05	0.48
PctAMINAN	0.01	0.02	0.01	0.01	0.71	0.01	0.00	0.01	0.00
PctWHI	0.82	0.62	0.88	0.55	0.22	0.43	0.37	0.90	0.26
PctBLK	0.06	0.25	0.03	0.14	0.00	0.03	0.16	0.01	0.08
PctASN	0.02	0.01	0.01	0.08	0.01	0.01	0.09	0.01	0.14
SVI	0.27	0.80	0.49	0.46	0.90	0.78	0.73	0.21	0.77
Religion									
CONGpc	121.17	307.90	243.59	84.93	343.56	219.57	63.95	343.61	56.37
PctADH	0.41	0.55	0.44	0.50	0.44	0.59	0.48	0.60	0.51
PctLtHSD	7.33	17.16	11.46	9.42	14.55	22.48	15.12	7.95	19.96
PctHSDO	27.90	37.23	37.54	22.49	37.11	31.06	24.15	33.49	20.39
PctSCoAD	31.05	28.97	31.22	27.18	32.49	28.66	27.81	35.43	25.61
HSDO	24511	8502	10561	98179	2471	10233	365701	4163	1411475
Population									
RtBRTH	9.81	11.07	9.91	11.03	18.06	12.90	11.48	11.12	9.60
RtDTH	9.99	16.06	14.73	8.73	14.81	11.81	8.93	13.69	8.70
RtNETMIG	8.92	2.57	10.03	-2.94	-11.31	-4.19	-7.77	2.53	-18.50
POPCHNG	1190	-21	421	-649	-108	47	-12840	45	-180394
NCHNG	-4.51	-79.95	-179.68	1443.04	1.52	158.96	5858.08	-14.50	8581.00
GQEST	3287	1458	1148	14579	205	1432	40722	486	180236

Poverty									
PctPOV517	11.51	28.27	18.92	13.76	33.13	21.58	19.47	13.50	18.50
PctPOVALL	10.05	20.96	14.17	11.35	28.39	16.54	15.04	11.04	14.10
POVALL	13476.4	7067.1	5669.03	71272.1	3284.2	10442.1	326645	1980.85	1365808
Unemployment									
RtUNEMP	4.32	5.42	4.62	5.12	7.14	5.91	7.13	3.05	8.90
UNEMP	3014.6	867.6	898.8	17280.3	340.26	1888.3	76139.7	337.4	445871
MEDHHINC	75321.7	45862.2	55008.7	82934.1	44907.3	55733.9	70602.1	61337.6	77356.0
PctMEDHHINC	109.22	75.84	85.65	112.45	64.97	82.42	95.68	90.80	91.20

Appendix C Code Appendix

This appendix contains 5 sections: Data Cleaning Code, Machine Learning Comparison Code, Gradient Boosted Tree Code, Hierarchical Clustering Code, and Figure Generation Code

Appendix C.1 Data Cleaning Code

```
# Ship of Thesis-eus (2023-12-27)
# Package Library
setwd("/Users/ianjacobs/Desktop/Thesis/Data Items/Cleaning")
library(tidyverse)
library(readxl)
library(dplyr)
library(rvest)
library(htmlTable)
library(data.table)

# Raw Data Loading
educ0 <- read_excel('Education.xlsx', range = 'A4:BC3289')
## ERS Dept. of Ag. Education Estimates
popu0 <- read_excel('PopulationEstimates.xlsx', range = 'A5:BA3209')
## ERS Dept. of Ag. Population Estimates
pove0 <- read_excel('PovertyEstimates.xlsx', range = 'A5:AH3200')
## ERS Dept. of Ag. Poverty Estimates
unem0 <- read_excel('Unemployment.xlsx', range = 'A5:CV3282')
## ERS Dept. of Ag. Unemployment Estimates
hesi0 <- read.csv('Vaccine_Hesitancy_for_COVID-19__County_and_local_estimates
_20240111.csv')
## ASPE COVID-19 Vaccine Hesitancy
reli0 <- read_excel('2020_USRC_Summaries.xlsx', sheet = '2020 County Summary'
)
## USRC 2020 U.S. Religion Census

# Column dropping and FIPS standardizing, also doing some nomenclature standa
rdizing

educ <- educ0 %>% rename('FIPS_Code' = 'Federal Information Processing Standa
rd (FIPS) Code', Area_Name = 'Area name') %>%
                  select(-("2003 Rural-urban Continuum Code":"Percent of adul
```

```
ts with a bachelor's degree or higher, 2008-12"))
educ$Area_Name <- gsub(pattern = ",.*", replacement = "", x=educ$Area_Name)
setDT(educ)[FIPS_Code=="02020", Area_Name:="Anchorage Borough"]
setDT(educ)[FIPS_Code=="02110", Area_Name:="Juneau Borough"]
setDT(educ)[FIPS_Code=="02220", Area_Name:="Sitka Borough"]
setDT(educ)[FIPS_Code=="02195", Area_Name:="Petersburg Borough"]
setDT(educ)[FIPS_Code=="02275", Area_Name:="Wrangell Borough"]
setDT(educ)[FIPS_Code=="02282", Area_Name:="Yakutat Borough"]
setDT(educ)[FIPS_Code=="02230", Area_Name:="Skagway Borough"]
setDT(educ)[FIPS_Code=="17099", Area_Name:="LaSalle County"]
setDT(educ)[FIPS_Code=="22000", Area_Name:="Louisiana"]
setDT(educ)[FIPS_Code=="22059", Area_Name:="LaSalle Parish"]
setDT(educ)[FIPS_Code=="35013", Area_Name:="Doña Ana County"]

##########

popu <- popu0 %>% rename('FIPS_Code' = 'FIPStxt', Area_Name = 'Area_Name') %>
%
                  select(FIPS_Code, State, Area_Name, CENSUS_2020_POP, POP_ES
TIMATE_2021, N_POP_CHG_2021, BIRTHS_2021, DEATHS_2021, NATURAL_CHG_2021, INTE
RNATIONAL_MIG_2021, DOMESTIC_MIG_2021,   NET_MIG_2021, RESIDUAL_2021, GQ_ESTI
MATES_2021, R_BIRTH_2021, R_DEATH_2021, R_NATURAL_CHG_2021, R_INTERNATIONAL_M
IG_2021, R_DOMESTIC_MIG_2021, R_NET_MIG_2021)
setDT(popu)[FIPS_Code=="02020", Area_Name:="Anchorage Borough"]
setDT(popu)[FIPS_Code=="02110", Area_Name:="Juneau Borough"]
setDT(popu)[FIPS_Code=="02195", Area_Name:="Petersburg Borough"]
setDT(popu)[FIPS_Code=="02220", Area_Name:="Sitka Borough"]
setDT(popu)[FIPS_Code=="02230", Area_Name:="Skagway Borough"]
setDT(popu)[FIPS_Code=="02275", Area_Name:="Wrangell Borough"]
setDT(popu)[FIPS_Code=="02282", Area_Name:="Yakutat Borough"]
setDT(popu)[FIPS_Code=="35013", Area_Name:="Doña Ana County"]

##########

pove <- pove0 %>% rename('FIPS_Code' = 'FIPS_Code', 'State'='Stabr', Area_Nam
e = 'Area_name') %>%
                  select(!c("Rural-urban_Continuum_Code_2003","Rural-urban_Co
ntinuum_Code_2013","Urban_Influence_Code_2003","Urban_Influence_Code_ 2013","
POV04_2021","CI90LB04_2021","CI90UB04_2021","PCTPOV04_2021","CI90LB04P_2021",
"CI90UB04P_2021","CI90LBALL_2021","CI90UBALL_2021","CI90LBALLP_2021","CI90UBA
LLP_2021","CI90LB017_2021","CI90UB017_2021","CI90LB017P_2021","CI90UB017P_202
1","CI90LB517_2021","CI90UB517_2021","CI90LB517P_2021","CI90UB517P_2021","CI9
0LBINC_2021","CI90UBINC_2021","MEDHHINC_2021"))
setDT(pove)[FIPS_Code=="02275", Area_Name:="Wrangell Borough"]
setDT(pove)[FIPS_Code=="02230", Area_Name:="Skagway Borough"]
setDT(pove)[FIPS_Code=="17099", Area_Name:="LaSalle County"]
setDT(pove)[FIPS_Code=="18033", Area_Name:="DeKalb County"]
setDT(pove)[FIPS_Code=="18087", Area_Name:="LaGrange County"]
setDT(pove)[FIPS_Code=="18091", Area_Name:="LaPorte County"]
setDT(pove)[FIPS_Code=="22059", Area_Name:="LaSalle Parish"]
```

```
setDT(pove)[FIPS_Code=="35011", Area_Name:="De Baca County"]
setDT(pove)[FIPS_Code=="42083", Area_Name:="McKean County"]
setDT(pove)[FIPS_Code=="35013", Area_Name:="Doña Ana County"]

##########

unem0$Area_Name <- gsub(pattern = ",.*", replacement = "", x=unem0$Area_Name)
unem <- unem0 %>% rename('FIPS_Code' = 'FIPS_Code', Area_Name = 'Area_Name')
%>%
                  select(-(Rural_Urban_Continuum_Code_2013:Unemployment_rate_
2020)) %>%
                  select(-(Civilian_labor_force_2022:Unemployment_rate_2022))
setDT(unem)[FIPS_Code=="02020", Area_Name:="Anchorage Borough"]
setDT(unem)[FIPS_Code=="02110", Area_Name:="Juneau Borough"]
setDT(unem)[FIPS_Code=="02220", Area_Name:="Sitka Borough"]
setDT(unem)[FIPS_Code=="02275", Area_Name:="Wrangell Borough"]
setDT(unem)[FIPS_Code=="02282", Area_Name:="Yakutat Borough"]
setDT(unem)[FIPS_Code=="02230", Area_Name:="Skagway Borough"]
setDT(unem)[FIPS_Code=="06075", Area_Name:="San Francisco County"]
setDT(unem)[FIPS_Code=="08014", Area_Name:="Broomfield County"]
setDT(unem)[FIPS_Code=="08031", Area_Name:="Denver County"]
setDT(unem)[FIPS_Code=="15003", Area_Name:="Honolulu County"]
setDT(unem)[FIPS_Code=="42101", Area_Name:="Philadelphia County"]
setDT(unem)[FIPS_Code=="17099", Area_Name:="LaSalle County"]
setDT(unem)[FIPS_Code=="25019", Area_Name:="Nantucket County"]
setDT(unem)[FIPS_Code=="35013", Area_Name:="Doña Ana County"]

##########

hesi0$County.Name <- gsub(pattern = ",.*", replacement = "", x=hesi0$County.N
ame)
hesi0$FIPS.Code <- sprintf("%05s", hesi0$FIPS.Code)
hesi <- hesi0 %>% rename('FIPS_Code' = 'FIPS.Code', Area_Name = 'County.Name'
, State_Name = 'State') %>% filter(FIPS_Code != c("02261")) # Valdez-Cordova
Census Area (AK)
setDT(hesi)[FIPS_Code=="02020", Area_Name:="Anchorage Borough"]
setDT(hesi)[FIPS_Code=="02110", Area_Name:="Juneau Borough"]
setDT(hesi)[FIPS_Code=="02220", Area_Name:="Sitka Borough"]
setDT(hesi)[FIPS_Code=="02230", Area_Name:="Skagway Borough"]
setDT(hesi)[FIPS_Code=="02275", Area_Name:="Wrangell Borough"]
setDT(hesi)[FIPS_Code=="02282", Area_Name:="Yakutat Borough"]
setDT(hesi)[FIPS_Code=="35013", Area_Name:="Doña Ana County"]

##########

reli <- reli0 %>% rename('FIPS_Code' = 'FIPS', Area_Name = 'County Name') %>%
                  select(!c("2020 Population"))
setDT(reli)[FIPS_Code=="02020", Area_Name:="Anchorage Borough"]
setDT(reli)[FIPS_Code=="02110", Area_Name:="Juneau Borough"]
setDT(reli)[FIPS_Code=="02220", Area_Name:="Sitka Borough"]
```

```
setDT(reli)[FIPS_Code=="02230", Area_Name:="Skagway Borough"]
setDT(reli)[FIPS_Code=="02275", Area_Name:="Wrangell Borough"]
setDT(reli)[FIPS_Code=="02282", Area_Name:="Yakutat Borough"]
setDT(reli)[FIPS_Code=="35013", Area_Name:="Doña Ana County"]

# All of these are going to be joined on FIPS_Code, State, Area_Name

ERSmerge1 <- merge(educ, popu, by.x= c('FIPS_Code','State','Area_Name'), by.y
= c('FIPS_Code','State','Area_Name'), all=TRUE)

ERSmerge2 = merge(ERSmerge1, pove, by.x= c('FIPS_Code','State','Area_Name'),
by.y = c('FIPS_Code','State','Area_Name'), all=TRUE)

ERS1 = merge(ERSmerge2, unem, by.x= c('FIPS_Code','State','Area_Name'), by.y
= c('FIPS_Code','State','Area_Name'), all=TRUE) %>% subset(State!="PR")

# notes dealing with typos/differing classifications
# ERS1 %>% group_by(FIPS_Code) %>% filter(n()>1)
## Alaska
# AK    Anchorage Borough (POPU, UNEM, EDUC)
# AK  Juneau Borough (POPU, UNEM, EDUC)
# AK    Petersburg Borough (EDUC)
# AK    Sitka Borough (POPU, UNEM, EDUC)
# AK    Wrangell Borough (POVE, POPU, UNEM) Funky stuff
# AK  Yakutat Borough (POPU, UNEM, EDUC)
## Not Alaska
# CA    San Francisco County
# CO    Broomfield County
# CO    Denver County
# HI    Honolulu County
# IL    LaSalle County
# IN    DeKalb County
# IN    LaGrange County
# IN    LaPorte County
# LA    Louisiana
# LA    LaSalle Parish
# MA    Nantucket County
# NM    De Baca County
# NM    Dona Ana County
# PA    McKean County
# PA    Philadelphia County

# dropping the ', ##' suffix from county names

ERS0 <- ERS1 %>% group_by(FIPS_Code, State) %>% summarise(Area_Name = paste(A
rea_Name, collapse = ", "))
ERS = merge(ERS1, ERS0, by.x= c('FIPS_Code','State','Area_Name'), by.y = c('F
IPS_Code','State','Area_Name'), all=TRUE) %>% subset(State!="PR")
ERS
```

```
# dropping state values, the entire US, and other included areas that arent o
fficially counties or county equivalents. Cross referenced with wikipedia

ERSfinal <- ERS %>% filter(FIPS_Code != c("30113")) %>% # Yellowstone NTL. Pa
rk (MT)
                   filter(FIPS_Code != c("51560")) %>% # Cliftion Forge(VA)
                   filter(FIPS_Code != c("51515")) %>% # Bedford(VA)
                   filter(FIPS_Code != c("02010")) %>% # Aleutian Islands(AK
)
                   filter(FIPS_Code != c("02160")) %>% # Kuskokwim Division
(AK)
                   filter(FIPS_Code != c("02201")) %>% # Prince of Wales-Out
er Ketchikan Census Area (AK)
                   filter(FIPS_Code != c("02231")) %>% # Skagway-Yakutat-Ang
oon Census Area (AK)
                   filter(FIPS_Code != c("02232")) %>% # Skagway-Hoonah-Ango
on Census Area
                   filter(FIPS_Code != c("02250")) %>% # Upper Yukon Divisio
n (AK)
                   filter(FIPS_Code != c("02261")) %>% # Valdez-Cordova Cens
us Area (AK)
                   filter(FIPS_Code != c("02280")) %>% # Wrangell-Petersburg
Census Area
                   filter(FIPS_Code != c("09110")) %>% # Capitol Planning Re
gion
                   filter(FIPS_Code != c("09120")) %>% # Greater Bridgeport
Planning Region
                   filter(FIPS_Code != c("09130")) %>% # Lower Connecticut R
iver Valley Planning Region
                   filter(FIPS_Code != c("09140")) %>% # Naugatuck Valley Pl
anning Region
                   filter(FIPS_Code != c("09150")) %>% # Northeastern Connec
ticut Planning Region
                   filter(FIPS_Code != c("09160")) %>% # Northwest Hills Pla
nning Region
                   filter(FIPS_Code != c("09170")) %>% # South Central Conne
cticut Planning Region
                   filter(FIPS_Code != c("09180")) %>% # Southeastern Connec
ticut Planning Region
                   filter(FIPS_Code != c("09190")) %>% # Western Connecticut
Planning Region
                   filter(FIPS_Code != c("00000")) %>% # USA
                   filter(FIPS_Code != c("01000")) %>% # Alabama
                   filter(FIPS_Code != c("02000")) %>% # Alaska
                   filter(FIPS_Code != c("04000")) %>% # Arizona
                   filter(FIPS_Code != c("05000")) %>% # Arkansas
                   filter(FIPS_Code != c("06000")) %>% # California
                   filter(FIPS_Code != c("08000")) %>% # Colorado
                   filter(FIPS_Code != c("09000")) %>% # Connecticut
                   filter(FIPS_Code != c("10000")) %>% # Delaware
```

```
                    filter(FIPS_Code != c("11000")) %>% # DC
                    filter(FIPS_Code != c("12000")) %>% # Florida
                    filter(FIPS_Code != c("13000")) %>% # Goergia
                    filter(FIPS_Code != c("15000")) %>% # Hawaii
                    filter(FIPS_Code != c("16000")) %>% # Idaho
                    filter(FIPS_Code != c("17000")) %>% # Illinois
                    filter(FIPS_Code != c("18000")) %>% # Indiana
                    filter(FIPS_Code != c("19000")) %>% # Iowa
                    filter(FIPS_Code != c("20000")) %>% # Kansas
                    filter(FIPS_Code != c("21000")) %>% # Kentucky
                    filter(FIPS_Code != c("22000")) %>% # Louisiana
                    filter(FIPS_Code != c("23000")) %>% # Maine
                    filter(FIPS_Code != c("24000")) %>% # Maryland
                    filter(FIPS_Code != c("25000")) %>% # Massachusetts
                    filter(FIPS_Code != c("26000")) %>% # Michigan
                    filter(FIPS_Code != c("27000")) %>% # Minnesota
                    filter(FIPS_Code != c("28000")) %>% # Mississippi
                    filter(FIPS_Code != c("29000")) %>% # Missouri
                    filter(FIPS_Code != c("30000")) %>% # Montana
                    filter(FIPS_Code != c("31000")) %>% # Nebraska
                    filter(FIPS_Code != c("32000")) %>% # Nevada
                    filter(FIPS_Code != c("33000")) %>% # New Hampshire
                    filter(FIPS_Code != c("34000")) %>% # New Jersey
                    filter(FIPS_Code != c("35000")) %>% # New Mexico
                    filter(FIPS_Code != c("36000")) %>% # New York
                    filter(FIPS_Code != c("37000")) %>% # North Carolina
                    filter(FIPS_Code != c("38000")) %>% # North Dakota
                    filter(FIPS_Code != c("39000")) %>% # Ohio
                    filter(FIPS_Code != c("40000")) %>% # Oklahoma
                    filter(FIPS_Code != c("41000")) %>% # Oregon
                    filter(FIPS_Code != c("42000")) %>% # Pennsylvania
                    filter(FIPS_Code != c("44000")) %>% # Rhode Island
                    filter(FIPS_Code != c("45000")) %>% # South Carolina
                    filter(FIPS_Code != c("46000")) %>% # South Dakota
                    filter(FIPS_Code != c("47000")) %>% # Tennessee
                    filter(FIPS_Code != c("48000")) %>% # Texas
                    filter(FIPS_Code != c("49000")) %>% # Utah
                    filter(FIPS_Code != c("50000")) %>% # Vermont
                    filter(FIPS_Code != c("51000")) %>% # Virginia
                    filter(FIPS_Code != c("53000")) %>% # Washington
                    filter(FIPS_Code != c("54000")) %>% # West Virginia
                    filter(FIPS_Code != c("55000")) %>% # Wisconsin
                    filter(FIPS_Code != c("56000")) # Wyoming

table1 <- table(ERSfinal$State)
table1
sum(table1)

url <- "https://en.wikipedia.org/wiki/List_of_United_States_FIPS_codes_by_cou
nty"
```

```
html <- read_html(url)
county_table <- html %>%
  html_element("table.wikitable.sortable") %>%
  html_table()


table2 <- table(county_table$`State or equivalent`)
table2

table3 <- table(hesi$'State Name')
table3
sum(table3)

hesiAK <- subset(hesi, State_Name=="ALASKA")
hesiAK

# checking identified states with issues

CT <- subset(ERSfinal, State == 'CT')
CT
MT <- subset(ERSfinal, State=="MT")
MT
VA <- subset(ERSfinal, State=="VA")
VA
AK <- subset(ERSfinal, State=="AK")
AK

Alaska <- merge(hesiAK, AK, by.x=c('FIPS_Code'), by.y=c('FIPS_Code'), all=TRU
E)
Alaska

ind0 <- duplicated(Alaska[,1])
Alaska[ind0,]
Alaska[!complete.cases(Alaska), ]
Alaska %>% summarise(across(everything(), ~ sum(is.na(.))))

# checking for missing values before final merge

na_rows <- ERSfinal[!complete.cases(ERSfinal), ]
na_rows

ERSfinal %>% summarise(across(everything(), ~ sum(is.na(.))))
ERSfinal[!complete.cases(ERSfinal), ]

# no poverty for Kalawao county HI

# merging hesitation with ERS data
FinalMerge1 <- merge(ERSfinal, hesi, by.x= c('FIPS_Code', 'Area_Name'), by.y
= c('FIPS_Code','Area_Name'), all=TRUE)  %>% filter(FIPS_Code != c("02261"))
ind1 <- duplicated(FinalMerge1[,1])
```

```r
FinalMerge1[ind1,]
FinalMerge1[!complete.cases(FinalMerge1), ]
FinalMerge1 %>% summarise(across(everything(), ~ sum(is.na(.))))

# merging religion data with the 5 other sheets
# final, usable sheet generation

Final0 <- merge(FinalMerge1, reli, by.x= c('FIPS_Code', 'Area_Name'), by.y =
c('FIPS_Code','Area_Name'), all=TRUE)
Final1 <- Final0[-c(1,3154)]

FinalSheetON <- subset(Final1, select = -c(State_Name, State.Code))

# variable renaming and nomenclature standardizing for ease of use during ana
lysis, also added census region and division data for potential figure creati
on

Data <- FinalSheetON %>% rename(LtHSD = "Less than a high school diploma, 201
7-21",
                                HSDO = "High school diploma only, 2017-21",
                                SCoAD = "Some college or associate's degree,
2017-21",
                                BDoH = "Bachelor's degree or higher, 2017-21"
,
                                PctLtHSD = "Percent of adults with less than
a high school diploma, 2017-21",
                                PctHSDO = "Percent of adults with a high scho
ol diploma only, 2017-21",
                                PctSCoAD = "Percent of adults completing some
college or associate's degree, 2017-21",
                                PctBDoH = "Percent of adults with a bachelor'
s degree or higher, 2017-21") %>%
                          rename(POP_2020 = CENSUS_2020_POP,
                                POPEST_2021 = POP_ESTIMATE_2021,
                                POPCHNG = N_POP_CHG_2021,
                                BRTH = BIRTHS_2021,
                                DTH = DEATHS_2021,
                                NCHNG = NATURAL_CHG_2021,
                                INTLMIG = INTERNATIONAL_MIG_2021,
                                DOMMIG = DOMESTIC_MIG_2021,
                                NETMIG = NET_MIG_2021,
                                RES = RESIDUAL_2021,
                                GQEST = GQ_ESTIMATES_2021,
                                RtBRTH = R_BIRTH_2021,
                                RtDTH = R_DEATH_2021,
                                RtNCHNG = R_NATURAL_CHG_2021,
                                RtINTLMIG = R_INTERNATIONAL_MIG_2021,
                                RtDOMMIG = R_DOMESTIC_MIG_2021,
                                RtNETMIG = R_NET_MIG_2021) %>%
                          rename(POVALL = POVALL_2021,
```

```
                                PctPOVALL = PCTPOVALL_2021,
                                POV017 = POV017_2021,
                                PctPOV017 = PCTPOV017_2021,
                                POV517 = POV517_2021,
                                PctPOV517 = PCTPOV517_2021) %>%
                         rename(CLF = Civilian_labor_force_2021,
                                EMP = Employed_2021,
                                UNEMP = Unemployed_2021,
                                RtUNEMP = Unemployment_rate_2021,
                                MEDHHINC = Median_Household_Income_2021,
                                PctMEDHHINC = Med_HH_Income_Percent_of_State_
Total_2021) %>%
                         rename(ESTHES = 'Estimated.hesitant',
                                ESTHESoUNS = 'Estimated.hesitant.or.unsure',
                                ESTSTRHES = 'Estimated.strongly.hesitant',
                                SVI = 'Social.Vulnerability.Index..SVI.',
                                SVICAT = 'SVI.Category',
                                CVACLOCVR = 'CVAC.level.of.concern.for.vaccin
ation.rollout',
                                CVACLOC = 'CVAC.Level.Of.Concern',
                                PctADFV = 'Percent.adults.fully.vaccinated.ag
ainst.COVID.19..as.of.6.10.21.',
                                PctHISP = 'Percent.Hispanic',
                                PctAMINAN = 'Percent.non.Hispanic.American.In
dian.Alaska.Native',
                                PctASN = 'Percent.non.Hispanic.Asian',
                                PctBLK = 'Percent.non.Hispanic.Black',
                                PctNHPI = 'Percent.non.Hispanic.Native.Hawaii
an.Pacific.Islander',
                                PctWHI = 'Percent.non.Hispanic.White',
                                GP = 'Geographical.Point',
                                CB = 'County.Boundary',
                                SB = 'State.Boundary') %>%
                         rename(State_Name = "State Name",
                                CONG = "Congregations",
                                ADH = "Adherents",
                                CONGpc = "Congregations Per 100,000 Populatio
n",
                                PctADH = "Adherents as % of Population",
                                POPRNK = "Population Rank",
                                CNGRNK = "Congregations Rank",
                                ADHRNK = "Adherents Rank",
                                CONGpcRNK = "Congregations Per 100,000 Pop. R
ank",
                                PctADHRNK = "Adherents as % of Population Ran
k") %>%
                         mutate(Census.Region = factor(State_Name, levels= c(
'Alabama', 'Alaska',  'Arizona', 'Arkansas', 'California', 'Colorado', 'Conne
cticut', 'Delaware', 'District of Columbia', 'Florida', 'Georgia', 'Hawaii',
'Idaho', 'Illinois', 'Indiana', 'Iowa', 'Kansas', 'Kentucky', 'Louisiana', 'M
```

```
aine', 'Maryland', 'Massachusetts', 'Michigan', 'Minnesota', 'Mississippi', '
Missouri', 'Montana', 'Nebraska', 'Nevada', 'New Hampshire', 'New Jersey', 'N
ew Mexico', 'New York', 'North Carolina', 'North Dakota', 'Ohio', 'Oklahoma',
'Oregon', 'Pennsylvania', 'Puerto Rico', 'Rhode Island', 'South Carolina', 'S
outh Dakota', 'Tennessee', 'Texas', 'Utah', 'Vermont', 'Virginia', 'Washingto
n', 'West Virginia', 'Wisconsin', 'Wyoming'), labels= c("South", "West", "Wes
t","South", "West", "West", "East", "South", "South", "South", "South", "West
", "West", "Midwest", "Midwest", "Midwest", "Midwest", "South", "South", "Eas
t", "South", "East", "Midwest", "Midwest", "South", "Midwest", "West", "Midwe
st", "West", "East", "East", "West", "East", "South", "Midwest", "Midwest", "
South", "West", "East", "South", "East", "South", "Midwest", "South", "South"
, "West", "East", "South", "West", "South", "Midwest", "West"))) %>%
                        mutate(Census.Division = factor(State_Name, levels=
c('Alabama', 'Alaska',  'Arizona', 'Arkansas', 'California', 'Colorado', 'Con
necticut', 'Delaware', 'District of Columbia', 'Florida', 'Georgia', 'Hawaii'
, 'Idaho', 'Illinois', 'Indiana', 'Iowa', 'Kansas', 'Kentucky', 'Louisiana',
'Maine', 'Maryland', 'Massachusetts', 'Michigan', 'Minnesota', 'Mississippi',
'Missouri', 'Montana', 'Nebraska', 'Nevada', 'New Hampshire', 'New Jersey', '
New Mexico', 'New York', 'North Carolina', 'North Dakota', 'Ohio', 'Oklahoma'
, 'Oregon', 'Pennsylvania', 'Puerto Rico', 'Rhode Island', 'South Carolina',
'South Dakota', 'Tennessee', 'Texas', 'Utah', 'Vermont', 'Virginia', 'Washing
ton', 'West Virginia', 'Wisconsin', 'Wyoming'), labels= c("East South Central
", "Pacific", "Mountain","West South Central", "Pacific", "Mountain", "New En
gland", "South Atlantic", "South Atlantic", "South Atlantic", "South Atlantic
", "Pacific", "Mountain", "East North Central", "East North Central", "West N
orth Central", "West North Central", "East South Central", "West South Centra
l", "New England", "South Atlantic", "New England", "East North Central", "We
st North Central", "East South Central", "West North Central", "Mountain", "W
est North Central", "Mountain", "New England", "Middle Atlantic", "Mountain",
"Middle Atlantic", "South Atlantic", "West North Central", "East North Centra
l", "West South Central", "Pacific", "Middle Atlantic", "South Atlantic", "Ne
w England", "South Atlantic", "West North Central", "East South Central", "We
st South Central", "Mountain", "New England", "South Atlantic", "Pacific", "S
outh Atlantic", "East North Central", "Mountain"))) %>% filter(FIPS_Code != c
("Totals")) # totals

ind3 <- duplicated(Data[,1])
Data[ind3,]
Data[!complete.cases(Data), ]
Data %>% summarise(across(everything(), ~ sum(is.na(.))))

Data %>% count(State)
Data000 <- Data %>% filter(is.na(State))
Data000

# Ouputting 'Data' to CSV for analysis

write.csv(Data, "/Users/ianjacobs/Desktop/Thesis/Analysis/Thesis_Data.csv", r
ow.names=TRUE)
```

Appendix C.2 Machine Learning Comparison Code

```
# Thesis-eus and the Minotaur (2024-02-05)
# Package Library
setwd("/Users/ianjacobs/Desktop/Thesis/Analysis")
library(tidyverse)
library(readxl)
library(devtools)
library(dplyr)
library(rvest)
library(htmlTable)
library(data.table)
library(Metrics)
library(ggplot2)
library(ggridges)
library(ggpubr)
library(usmap)
library(usmapdata)
library(corrplot)
library(psych)
library(glmnet)
library(caret)
library(ISLR)
library(earth)
library(ggbiplot)
library(caTools)
library(xgboost)
library(randomForest)
library(partykit)

# Ingest County Level Data
data <- read.csv('Thesis_Data.csv')
data$FIPS_Code <- sprintf("%05s", data$FIPS_Code)

# Dropping Alaska counties with no hesitancy values
dataNONA <- data %>% filter(FIPS_Code != c("02063")) %>% # Chugach Census Are
a
                     filter(FIPS_Code != c("02066")) %>%    # Copper River Ce
nsus Area
                     rename('fips' = 'FIPS_Code')


# Dropping categorical variables, retaining all numerical values, dropping al
l columns with na values (10 total rows)
sapply(dataNONA, class)
remove_cols <- c('X','Area_Name','State','GP','CB','SB','Census.Region','Cens
us.Division', 'CVACLOC','SVICAT', 'State_Name', 'PctADFV','ESTHESoUNS','ESTST
RHES')
```

```
dataNUM0 = subset(dataNONA, select = !(names(dataNONA) %in% remove_cols))
dataNUM <- dataNUM0 %>% mutate_if(is.integer, as.numeric) %>% column_to_rowna
mes(., var='fips')
dataNUMnona <- na.omit(dataNUM)

ind666 <- duplicated(dataNUM0[,1])

dataNUM0[ind666,]
dataNUM0[!complete.cases(dataNUM0), ]
dataNUM0 %>% summarise(across(everything(), ~ sum(is.na(.))))

# Machine learning Analysis(1/4)
## elastic net

X <- dataNUMnona %>% select(ESTHES) %>% scale(center = TRUE,
                                              scale = FALSE) %>% as.matrix()
Y <- dataNUMnona %>% select(-ESTHES) %>% scale(center = TRUE,
                                               scale = FALSE) %>% as.matrix()
set.seed(1234)
custom <- trainControl(method = "repeatedcv",
                       number = 5,
                       repeats = 5,
                       search = "random",
                       verboseIter = TRUE)

NetALL <- train(ESTHES~.,
                data = cbind(X, Y),
                method='glmnet',
                tuneLength=25,
                preProcess = c("center","scale"),
                trControl=custom)

valALL <- mean(NetALL$resample$RMSE)
valALL
plot(varImp(NetALL,scale=TRUE))

NetALL_pre <- predict(NetALL, Y)
NetALL_pre

rsq <- cor(X, NetALL_pre)^2
rsq

NetALL

# Machine learning Analysis (2/4)
## multivariate adaptive regression splines

hyper_grid <- expand.grid(degree = 1:3,
                          nprune = seq(2, 50, length.out = 10) %>%
                          floor())
set.seed(1234)
```

```
MARSALL <- train(
  x = subset(dataNUMnona, select = -c(ESTHES)),
  y = dataNUMnona$ESTHES,
  method = "earth",
  metric = "RMSE",
  trControl = trainControl(method = "cv", number = 10),
  tuneGrid = hyper_grid)
MARSALL$results %>% filter(nprune==MARSALL$bestTune$nprune, degree==MARSALL$b
estTune$degree)

# Discover Important Features (3/4)
## random forest

split <- sample.split(dataNUMnona, SplitRatio = 0.7)
train <- subset(dataNUMnona, split == "TRUE")
test <- subset(dataNUMnona, split == "FALSE")
set.seed(1234)
RFALL = randomForest(x = train[-38],
                             y = train$ESTHES,
                             ntree = 50)
print(RFALL)
y_pred = predict(RFALL, newdata = test[-38])
plot(RFALL)

importance(RFALL)
varImpPlot(RFALL)

predicted <- unname(predict(RFALL, test))

which.min(RFALL$mse)
sqrt(RFALL$mse[which.min(RFALL$mse)])
1 - (sum((test$ESTHES-predicted)^2)/sum((test$ESTHES-mean(test$ESTHES))^2))

# Discover Important Features (4/4)
## gradient boosted

set.seed(1234)
parts = createDataPartition(dataNUMnona$ESTHES, p = .7, list = F)
train2 = dataNUMnona[parts, ]
test2 = dataNUMnona[-parts, ]

train_x = data.matrix(train2[-38])
train_y = train2$ESTHES
test_x = data.matrix(test2[-38])
test_y = test2$ESTHES

xgb_train = xgb.DMatrix(data = train_x, label = train_y)
xgb_test = xgb.DMatrix(data = test_x, label = test_y)
watchlist = list(train=xgb_train, test=xgb_test)

model = xgb.train(data = xgb_train, max.depth = 3, watchlist=watchlist, nroun
```

```
ds = 173)
pred_y <- predict(model, test_x)

caret::RMSE(test_y, pred_y)
1 - (sum((test_y-pred_y)^2)/sum((test_y-mean(test_y))^2))
xgb.importance(model=model)
```

Appendix C.3 Gradient Boosted Tree Code

```
# Thesis-eus in the Gradient Boosted Forest(2024-02-19)
# Package Library
setwd("/Users/ianjacobs/Desktop/Thesis/Analysis")
library(tidyverse)
library(readxl)
library(devtools)
library(dplyr)
library(rvest)
library(htmlTable)
library(data.table)
library(Metrics)
library(ggplot2)
library(ggridges)
library(ggpubr)
library(usmap)
library(usmapdata)
library(corrplot)
library(psych)
library(glmnet)
library(caret)
library(ISLR)
library(earth)
library(ggbiplot)
library(caTools)
library(xgboost)
library(randomForest)
library(partykit)
library(ROCR)
library(Ckmeans.1d.dp)

# Ingest County Level Data
data <- read.csv('Thesis_Data.csv')
data$FIPS_Code <- sprintf("%05s", data$FIPS_Code)

# Dropping Alaska counties with no hesitancy values
dataNONA1 <- data %>% filter(FIPS_Code != c("02063")) %>% # Chugach Census Ar
```

```
ea
                      filter(FIPS_Code != c("02066")) %>%    # Copper River Ce
nsus Area
                      rename('fips' = 'FIPS_Code')

# Dropping categorical variables, retaining all numerical values, dropping al
l columns with na values (10 total rows)
sapply(dataNONA1, class)
remove_cols <- c('X','Area_Name','State','GP','CB','SB','Census.Region','Cens
us.Division', 'CVACLOC','SVICAT', 'State_Name', 'PctADFV')
dataNUM01 = subset(dataNONA1, select = !(names(dataNONA1) %in% remove_cols))
dataNUM1 <- dataNUM01 %>% mutate_if(is.integer, as.numeric) %>% column_to_row
names(., var='fips')
dataNUMnona1 <- na.omit(dataNUM1)

# Dividing into Hesitance categories
fullHESoUNS <- dataNUMnona1 %>% select(-c('ESTHES','ESTSTRHES'))
fullHES <- dataNUMnona1 %>% select(-c('ESTHESoUNS','ESTSTRHES'))
fullSTRHES <- dataNUMnona1 %>% select(-c('ESTHES','ESTHESoUNS'))

# creating sets with no religion metrics
noreliHESoUNS <- fullHESoUNS %>% select(-c('CONG':'PctADHRNK'))
noreliHES <- fullHES %>% select(-c('CONG':'PctADHRNK'))
noreliSTRHES <- fullSTRHES %>% select(-c('CONG':'PctADHRNK'))


## fullHES 2

set.seed(1234)
parts2 = createDataPartition(fullHES$ESTHES, p = .7, list = F)
train2 = fullHES[parts2, ]
test2 = fullHES[-parts2, ]

train_x2 = data.matrix(train2[-38])
train_y2 = train2$ESTHES
test_x2 = data.matrix(test2[-38])
test_y2 = test2$ESTHES

xgb_train2 = xgb.DMatrix(data = train_x2, label = train_y2)
xgb_test2 = xgb.DMatrix(data = test_x2, label = test_y2)
watchlist2 = list(train=xgb_train2, test=xgb_test2)

model2 = xgb.train(data = xgb_train2, max.depth = 3, watchlist=watchlist2, nr
ounds = 400)
pred_y2 <- predict(model2, test_x2)

caret::RMSE(test_y2, pred_y2)
1 - (sum((test_y2-pred_y2)^2)/sum((test_y2-mean(test_y2))^2))
```

```r
## noreliHES 5

set.seed(1234)
parts5 = createDataPartition(noreliHES$ESTHES, p = .7, list = F)
train5 = noreliHES[parts5, ]
test5 = noreliHES[-parts5, ]

train_x5 = data.matrix(train5[-38])
train_y5 = train5$ESTHES
test_x5 = data.matrix(test5[-38])
test_y5 = test5$ESTHES

xgb_train5 = xgb.DMatrix(data = train_x5, label = train_y5)
xgb_test5 = xgb.DMatrix(data = test_x5, label = test_y5)
watchlist5 = list(train=xgb_train5, test=xgb_test5)

model5 = xgb.train(data = xgb_train5, max.depth = 3, watchlist=watchlist5, nr
ounds = 400)
pred_y5 <- predict(model5, test_x5)

caret::RMSE(test_y5, pred_y5)
1 - (sum((test_y5-pred_y5)^2)/sum((test_y5-mean(test_y5))^2))

IMPORTANT <- fullHES %>% select(c("ESTHES", "PctMEDHHINC", "MEDHHINC", "CVACL
OCVR", "PctHISP", "PctAMINAN", "CONGpc", "RtUNEMP","PctWHI", "PctPOVALL", "Pc
tLtHSD","PctBLK", "PctHSDO", "SVI","RtBRTH","RtDTH","RtNETMIG","PctSCoAD", "P
ctADH","UNEMP", "POVALL","POPCHNG","PctASN","NCHNG","HSDO","PctPOV517","GQEST
"))


set.seed(123)
parts666 = createDataPartition(IMPORTANT$ESTHES, p = .7, list = F)
train666 = IMPORTANT[parts666, ]
test666 = IMPORTANT[-parts666, ]

train_x666 = data.matrix(train666[-1])
train_y666 = train666$ESTHES
test_x666 = data.matrix(test666[-1])
test_y666 = test666$ESTHES

xgb_train666 = xgb.DMatrix(data = train_x666, label = train_y666)
xgb_test666 = xgb.DMatrix(data = test_x666, label = test_y666)
watchlist666 = list(train=xgb_train666, test=xgb_test666)

model666 = xgb.train(data = xgb_train666, max.depth = 3, watchlist=watchlist6
66, nrounds = 232, method = "xgbTree", trControl = trainControl("cv", number
= 10))
pred_y666 <- predict(model666, test_x666)

min(IMPORTANT$ESTHES)
```

```
max(IMPORTANT$ESTHES)
caret::RMSE(test_y666, pred_y666)
1 - (sum((test_y666-pred_y666)^2)/sum((test_y666-mean(test_y666))^2))
model666$bestTune

trees = xgb.importance(model=model666)

IMPORTANT2 <- fullHES %>% select(c("ESTHES", "PctMEDHHINC", "MEDHHINC", "CVAC
LOCVR", "PctHISP", "PctAMINAN", "RtUNEMP","PctWHI", "PctPOVALL", "PctLtHSD","
PctBLK", "PctHSDO", "SVI","RtBRTH","RtDTH","RtNETMIG","PctSCoAD", "UNEMP", "P
OVALL","POPCHNG","PctASN","NCHNG","HSDO","PctPOV517","GQEST"))


set.seed(123)
parts6666 = createDataPartition(IMPORTANT2$ESTHES, p = .7, list = F)
train6666 = IMPORTANT2[parts6666, ]
test6666 = IMPORTANT2[-parts6666, ]

train_x6666 = data.matrix(train6666[-1])
train_y6666 = train6666$ESTHES
test_x6666 = data.matrix(test6666[-1])
test_y6666 = test6666$ESTHES

xgb_train6666 = xgb.DMatrix(data = train_x6666, label = train_y6666)
xgb_test6666 = xgb.DMatrix(data = test_x6666, label = test_y6666)
watchlist6666 = list(train=xgb_train6666, test=xgb_test6666)

model6666 = xgb.train(data = xgb_train6666, max.depth = 3, watchlist=watchlis
t6666, nrounds = 232, method = "xgbTree", trControl = trainControl("cv", numb
er = 10))
pred_y6666 <- predict(model6666, test_x6666)

min(IMPORTANT$ESTHES)
max(IMPORTANT$ESTHES)
caret::RMSE(test_y6666, pred_y6666)
1 - (sum((test_y6666-pred_y6666)^2)/sum((test_y6666-mean(test_y6666))^2))
model6666$bestTune

varlist <- as.list(dataNUMnona1)
```

Appendix C.4 Hierarchical Clustering Code

```
# Thesis-eus and Hier-polyta(2024-02-19)
# Package Library
setwd("/Users/ianjacobs/Desktop/Thesis/Analysis")
library(tidyverse)
```

```
library(readxl)
library(devtools)
library(dplyr)
library(rvest)
library(htmlTable)
library(data.table)
library(Metrics)
library(ggplot2)
library(ggridges)
library(ggpubr)
library(usmap)
library(usmapdata)
library(corrplot)
library(psych)
library(glmnet)
library(caret)
library(ISLR)
library(earth)
library(ggbiplot)
library(caTools)
library(xgboost)
library(randomForest)
library(partykit)
library(ROCR)
library(factoextra)
library(cluster)

# Ingest County Level Data
data <- read.csv('Thesis_Data.csv')
data$FIPS_Code <- sprintf("%05s", data$FIPS_Code)

# Dropping Alaska counties with no hesitancy values
dataNONA1 <- data %>% filter(FIPS_Code != c("02063")) %>% # Chugach Census Ar
ea
                     filter(FIPS_Code != c("02066")) %>%  # Copper River Cens
us Area
                     rename('fips' = 'FIPS_Code')

# Dropping categorical variables, retaining all numerical values, dropping al
l columns with na values (10 total rows)
sapply(dataNONA1, class)
remove_cols <- c('X','Area_Name','State','GP','CB','SB','Census.Region','Cens
us.Division', 'CVACLOC','SVICAT', 'State_Name', 'PctADFV')
dataNUM01 = subset(dataNONA1, select = !(names(dataNONA1) %in% remove_cols))
dataNUM1 <- dataNUM01 %>% mutate_if(is.integer, as.numeric) %>% column_to_row
names(., var='fips')
dataNUMnona1 <- na.omit(dataNUM1)
fullHES <- dataNUMnona1 %>% select(-c('ESTHESoUNS','ESTSTRHES'))

# Final Set
```

```
IMPORTANT <- fullHES %>% select(c("PctMEDHHINC", "MEDHHINC", "CVACLOCVR", "PctHISP", "PctAMINAN", "CONGpc", "RtUNEMP","PctWHI", "PctPOVALL", "PctLtHSD","PctBLK", "PctHSDO", "SVI","RtBRTH","RtDTH","RtNETMIG","PctSCoAD", "PctADH","UNEMP", "POVALL","POPCHNG","PctASN","NCHNG","HSDO","PctPOV517","GQEST"))

important <- scale(IMPORTANT)

m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

ac <- function(x) {
  agnes(important, method = x)$ac
}

sapply(m, ac)

clust1 <- agnes(important, method = "ward")

pltree(clust1, cex = 0.6, hang = -1, main = "Dendrogram")

gap_stat2 <- clusGap(important, FUN = hcut, K.max = 15, B = 25)

d <- dist(important, method = "euclidean")

final_clust <- hclust(d, method = "ward.D2" )

groups <- cutree(final_clust, k=9)

table(groups)

important2 <- cbind(IMPORTANT, cluster=groups)

head(important2)

important0 <- as.data.frame(important2)

important00 <- rownames_to_column(important0, "fips")

important00$cluster <- as.character(important00$cluster)

tb <- aggregate(IMPORTANT, by=list(cluster=important00$cluster), mean)

write.csv(tb, "my_tb.csv")
```

Appendix C.5 Figure Generation Code

```
# Thesis-eus, Hero of Graph-ens(2024-01-22)

ClusPlot <- plot_usmap(data = important00, values = "cluster", labels=TRUE) +
            theme(panel.background = element_rect(colour = "black")) +
            scale_fill_manual(values = c('1' = "#536e0a", '2' = "#7da50f",'3'
='#bdda0f','4'='#ffa800','5'='#ff7a00','6'='#ff3d35','7'='#e52b6f','8'='#6226
a9','9'='#2c29a2'), name = "treatment") +
            theme(legend.position = "right")+
       theme(text=element_text(size=13,  family="Times New Roman")) + ggtitle
("") + guides(fill=guide_legend(title="Cluster"))


ESTPlot <- plot_usmap(data = dataNONA1, values = "ESTHES", labels=TRUE) +
            theme(panel.background = element_rect(colour = "black")) +
            theme(legend.position = "right") +
            theme(text=element_text(size=13,  family="Times New Roman")) +
            ggtitle("") +
            scale_fill_viridis_c( name = "Estimated
Hesitant (%)")
ESTPlot

sportk <- fviz_gap_stat(gap_stat2) + theme_light() +
       theme(text=element_text(size=13,  family="Times New Roman")) + ggtitle
("")
sportk

ggsave('sportk.png',dpi=3000)

FULLdataCORR = cor(important, use="pairwise.complete.obs")

clunk = corrplot(FULLdataCORR, method="color",  type="lower", order="hclust",
addCoef.col = "black", tl.col="black", tl.srt=45, sig.level = 0.01, insig = "
blank", diag=FALSE, family="Times New Roman")

clunk

ggsave('clunk.png',dpi=3000)

figg <- dataNONA %>%  select(ESTHESoUNS, ESTHES, ESTSTRHES) %>%
             pivot_longer(cols = everything(), names_to = "Hesitancy_Categor
y", values_to = "Value") %>%
             ggplot(aes(x = reorder(Hesitancy_Category,Value), y = Value, fi
ll = reorder(Hesitancy_Category,Value))) +
             geom_boxplot()


figg + scale_x_discrete(labels = c("Estimated Strongly Hesitant","Estimated H
```

```
esitant","Estimated Hesitant or Unsure")) +
       xlab("Hesitancy Category") + ylab("Proportion of Individuals by County
") +
       scale_fill_discrete(labels = c("Estimated Strongly Hesitant","Estimate
d Hesitant","Estimated Hesitant or Unsure")) +
       labs(fill = "Hesitancy Category") +
       theme_light() +
       theme(text=element_text(size=13,  family="Times New Roman"))

ggsave('figg.png',dpi=1000)

corn <- xgb.ggplot.importance(importance_matrix = trees[1:26], n_clusters = 1
) +  xlab("Variable Name") + ylab("Relative Importance") +
        theme_light() + theme(text=element_text(size=13,  family="Times New R
oman")) +  scale_fill_manual(values=c("black"), guide="none") + ggtitle("")
corn

# ggsave('corn.png',dpi=3000)
```

Bibliography

Agency for Toxic Substances and Disease Registry. (2020, October 15). CDC's Social Vulnerability Index (SVI). Www.atsdr.cdc.gov. https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

Centers for Disease Control and Prevention. (2021, June 17). Vaccine Hesitancy for COVID-19: County and Local Estimates | Data | Centers for Disease Control and Prevention. Data.cdc.gov. https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw/about_data

Federal Register. (2020, December 14). Federal Register :: Request Access. Unblock.federalregister.gov. https://www.federalregister.gov/documents/2020/12/14/2020-27459/change-to-county-equivalents-in-the-state-of-connecticut

Garcia, L. L., & Yap, J. F. C. (2021). The role of religiosity in COVID-19 vaccine hesitancy. Journal of Public Health, 43(3). https://doi.org/10.1093/pubmed/fdab192

Hastie, T., Tibshirani, R., & Friedman, J. H. (2004). The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. Springer.

Kibongani Volet, A., Scavone, C., Catalán-Matamoros, D., & Capuano, A. (2022). Vaccine Hesitancy Among Religious Groups: Reasons Underlying This Phenomenon and Communication Strategies to Rebuild Trust. Frontiers in Public Health, 10(10). https://doi.org/10.3389/fpubh.2022.824560

Lee, S. W., Ma, D., Davoodian, A., Ayutyanont, N., & Werner, B. (2023). COVID-19 vaccination decreased COVID-19 hospital length of stay, in-hospital death, and increased home discharge. Preventive Medicine Reports, 32(102152). https://doi.org/10.1016/j.pmedr.2023.102152

Murtagh, F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms. The Computer Journal, 26(4), 354–359. https://doi.org/10.1093/comjnl/26.4.354

National Institutes of Health. (2014, April 15). Rural-Urban Continuum Code - SEER Datasets. SEER. https://seer.cancer.gov/seerstat/variables/countyattribs/ruralurban.html

Nicholson, C., Beattie, L., Beattie, M., Razzaghi, T., & Chen, S. (2022). A machine learning and clustering-based approach for county-level COVID-19 analysis. PLOS ONE, 17(4), e0267558. https://doi.org/10.1371/journal.pone.0267558

Surgo Ventures. (2021). Surgo U.S. COVID-19 Vaccine Coverage Index. Vaccine.precisionforcovid.org. https://vaccine.precisionforcovid.org/

The Association of Statisticians of American Religious Bodies. (2023, June 23). Maps and data files for 2020 | U.S. Religion Census | Religious Statistics & Demographics. Www.usreligioncensus.org. https://www.usreligioncensus.org/node/1639

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411–423. https://doi.org/10.1111/1467-9868.00293

U.S. Census Bureau. (2021, August 25). Alaska, Least Densely Populated State, Had Population of 733,391 in 2020. Census.gov. https://www.census.gov/library/stories/state-by-state/alaska-population-change-between-census-decade.html

U.S. Census Bureau. (2022, December 15). Microdata. The United States Census Bureau. https://www.census.gov/programs-surveys/acs/microdata.html

U.S. Census Bureau. (2024, February 8). Household Pulse Survey. The United States Census Bureau. https://www.census.gov/programs-surveys/household-pulse-survey.html

U.S. Census Bureau, U.S. Department of Commerce, & Economics and Statistics Administration. (2010). Census Regions and Divisions of the United States. In US Census Bureau.

U.S. Department of Agriculture. (2018). USDA ERS - County-level Data Sets. Usda.gov. https://www.ers.usda.gov/data-products/county-level-data-sets/

U.S. Department of Health and Human Services, & Office of the Assistant Secretary for Planning and Evaluation. (2021, June 16). Vaccine Hesitancy for COVID-19: State, County, and Local Estimates. ASPE. https://aspe.hhs.gov/reports/vaccine-hesitancy-covid-19-state-county-local-estimates

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x