COVID-19 Impact: Classification

Data Mining Project 3

PREPARED BY

Trevor Dohm, Eileen Garcia, Blake Gebhardt

Executive Summary

The COVID-19 pandemic has presented a significant challenge for policymakers, health professionals, and the public. To tackle this challenge, data mining techniques have been employed to analyze COVID-19 data from various regions of the United States. The primary focus of this report is to utilize classification techniques to answer key research questions related to the pandemic. Our approach involved collecting data from various sources, including the CDC, state and local health departments, and publicly available data. In order to prevent a “fifth wave”, we would like to classify counties or states in high/low or low/medium/high risk in terms of how affected they would be by a fifth wave. These results can be used to prepare the infrastructure and plan possible interventions (e.g., mask mandates, temporarily closing businesses and schools, etc.). Early interventions based on data might dampen a severe outbreak and therefore save lives and shorten the length of necessary closings.

Our comprehensive analysis sheds light on the varying infection rates and response strategies to COVID-19 across the United States. By employing classification techniques, we were able to group counties into high-risk, medium-risk, and low-risk categories, providing valuable insights for policymakers and health professionals. Understanding the unique characteristics and risk profiles of these regions enables the targeted allocation of resources and interventions, ultimately aiding in controlling the virus's spread. The report encompasses the entire US, offering a holistic view of the COVID-19 pandemic. Using machine learning models such as K-Nearest Neighbors, Random Forest, and Artificial Neural Networks, we were able to identify relationships between various demographic, socioeconomic, and COVID-19-related factors. These findings can guide the CDC and other stakeholders in tailoring their strategies to areas with the greatest need, maximizing the effectiveness of their efforts. First, we evaluated trends at both the state and county levels, revealing significant variations in infection rates and responses across regions. We identified areas that demonstrated particular success in managing the pandemic, as well as those that struggled. Classification techniques facilitated the grouping of regions with similar attributes, such as population density, age distribution, and economic indicators. Next, we developed predictive models that could estimate the risk of future COVID-19 waves in different areas. This information is invaluable for decision-makers, as it allows them to anticipate challenges, allocate resources efficiently, and implement targeted public health measures.

Table of Contents

Business Understanding

COVID-19, also known as the novel coronavirus, is a highly infectious respiratory illness that was declared the cause of a global pandemic since its initial outbreak in Wuhan, China in December 2019. It is caused by a virus known as SARS-CoV-2 and is primarily spread through respiratory droplets from an infected individual when they speak, cough, or sneeze. Since it was first discovered, nearly 800 million people across the globe have been infected, and nearly 7 million have died from the disease [1]. Social distancing and the term "flattening the curve" are efforts aimed at slowing the spread of the virus. Social distancing involves staying at least 2 meters away from other people, avoiding large gatherings, and working from home if possible [2]. Social distancing aims to reduce the spread of COVID-19 by limiting contact between others and preventing prolonged contact with respiratory droplets. By staying physically separate, the likelihood of infection by inhaling infected droplets decreases. "Flattening the curve" refers to slowing the rate of new cases so that the healthcare system can withstand the volume of patients. The "curve" refers to the graph of new infections. According to the New York Times, the curve peaked in winter 2020-2021 and winter 2021-2022 [3]. By practicing social distancing and other disease-fighting measures, the rate of new infections will slow, flattening the curve.

Classification methods have been utilized in combating COVID-19 to categorize the transmission patterns, patient admissions, and available resources such as hospital capacity, breathing support devices, and medical personnel. The insights gained from classification analysis are essential for healthcare authorities and policymakers in their decision-making. By evaluating case numbers, hospitalizations, and fatalities, health officials can assess the influence of COVID-19 on communities and pinpoint the regions most impacted. Classification approaches also facilitate the recognition of hotspots where the virus is propagating swiftly, enabling targeted allocation of resources. By tracking admission data, healthcare professionals can estimate the required number of hospital beds, ventilators, and medical staff to provide adequate care for patients. This information assists decision-makers in evaluating the efficacy of various strategies, such as social distancing and immunization programs, and making informed choices regarding the best course of action during the pandemic. By comprehending these aspects, we can strive to manage the pandemic and safeguard the health and welfare of people worldwide, with the goal of avoiding widespread infections in the future.

Data Understanding

U.S. COVID-19 Cases and Census Dataset

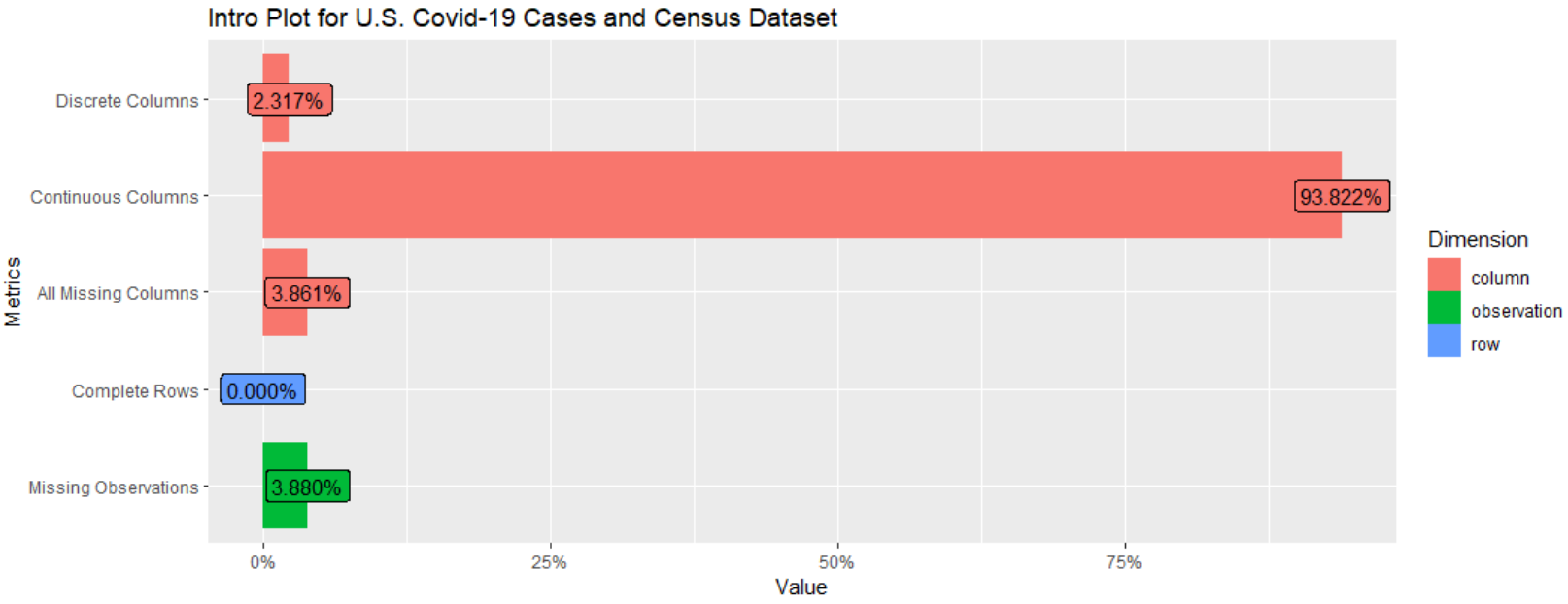

Figure 1. Intro Plot for U.S. Covid-19 Cases and Census Dataset.

The U.S. COVID-19 Cases and Census Dataset is a comprehensive collection of data related to the ongoing COVID-19 pandemic in the United States. This dataset includes critical information on confirmed COVID-19 cases and deaths, as well as a vast range of demographic data such as age, gender, race, and ethnicity obtained from the U.S. Census. This data provides a valuable tool for researchers and policymakers to identify and address disparities and inequalities in COVID-19 outcomes among various populations. To facilitate our analysis, we have chosen to focus exclusively on data points from the state of Texas within the U.S. COVID-19 Cases and Census Dataset. However, as we began to explore the raw dataset visualized in Figure 1, we quickly encountered significant discrepancies that required cleaning and refinement, as displayed by the percentage of missing columns and observations in the data. Our team recognized the importance of thorough cleaning and refinement of the U.S. COVID-19 Cases and Census Dataset to ensure accurate and reliable analysis. As such, we embarked on a meticulous process to identify and eliminate any inconsistencies or errors in the data, which ultimately led to the exclusion of features that were missing entirely from our analysis. However, we encountered some unexpected issues in this process as well, which we will discuss in more detail in the following sections.

United States County Coordinates Dataset

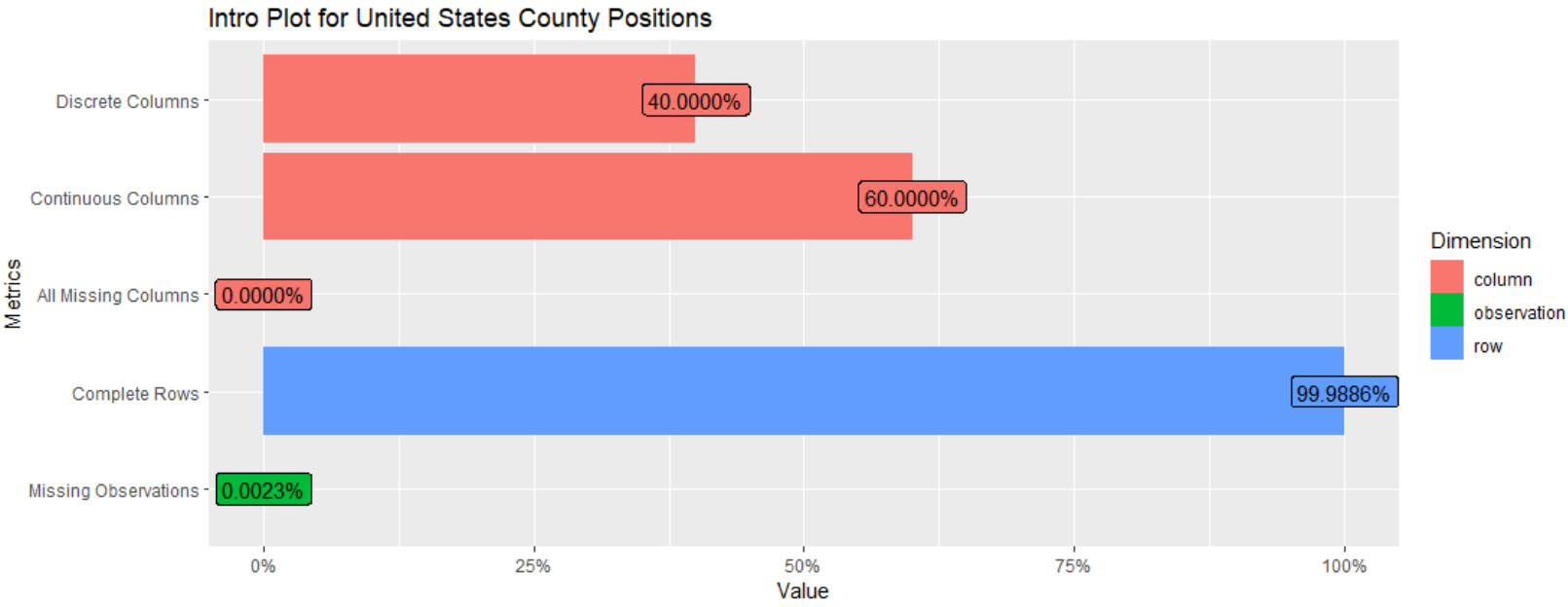

Figure 2. Intro Plot for U.S. County Positions Dataset.

The United States County Coordinates Dataset, visualized in Figure 3, is a vital resource containing latitude and longitude coordinates for all counties in the country. We leveraged this dataset to visualize our classification data on a United States map accurately. By combining the classification results with the county coordinates, we created a visually informative representation of the data that demonstrated how various counties in the country were grouped based on their characteristics. This approach provided us with valuable insights into the geographical distribution of the classifications and their correlation with different regions of the country. In this context, the United States County Coordinates Dataset was a critical resource that enabled us to create comprehensive and visually appealing data visualizations. It is worth noting that this dataset contained nearly complete rows, providing us with a reliable and accurate dataset for our analysis. We noted that Hawaiian and Alaskan counties were not given, nor District of Columbia, Washington DC. This noninclusion of these counties here will be visible later on when we visualize results using maps.

Data Preparation

Data Cleaning

As we had run into issues with combining the census and vaccine datasets on past projects, our team decided to perform dimensionality reduction on the Census data alone, transforming the data until we found some optimal class variables that would allow us to perform the classification task effectively and optimally. Similarly to past projects, our solution to the issues we faced in cleaning was to take the final feature dataset and then subset the original census data upon these features. Here, we found thirty important features and then added some extra from both calculations of new columns as well as other variables in the original dataset that we found interesting. These extra variables not given from the feature importance were "employed_pop", "unemployed_pop", "in_school", "in_undergrad_college", “cases_per_10000”, “deaths_per_10000”, and finally “death_per_case”, the three latter of which being the calculated variables. By completing these steps that we will see outlined below, we were able to obtain a complete mapping of the United States, and with the cleaned, important features, along with some other variables, that we had hoped for.

The first step of our process involved reading the Census data and performing some initial data cleaning. We then removed columns with missing values and computed the correlation matrix between the remaining numeric features. To simplify our analysis, we removed highly correlated variables using a threshold of 0.95. This resulted in a much more manageable set of important features to work with, which allowed us to analyze the features we wanted to cluster with. The reduced dimensionality helped us to better understand the data we were working with. Figure 3 illustrates the correlation matrix after performing this reduction. Despite this dimensionality reduction, the matrix was still quite complex. In the next section, we will discuss how we addressed this issue.

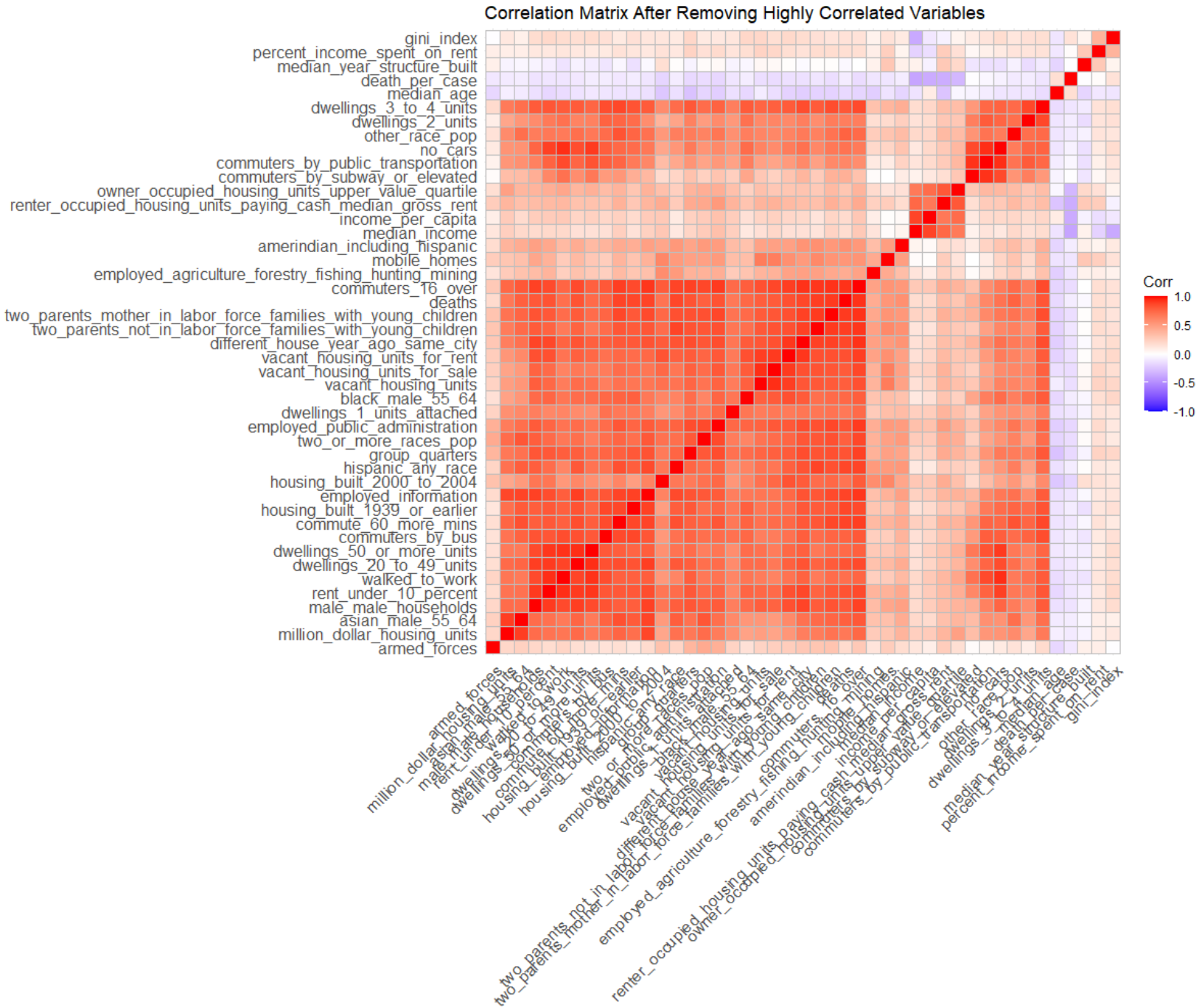

Figure 3. Correlation Matrix After Removing Highly Correlated Variables.

After removing highly correlated variables with a threshold of 0.95, we still had a large number of variables that made classification in our project difficult. To avoid overfitting and improve model performance, we decided to use variable importance (VarImp) as an approach for feature extraction. By using VarImp, we were able to identify the most important features in our dataset and use them for classification. To implement this approach, we first trained a controlled model using the repeated cross-validation (repeated-cv) technique with 10 steps and 3 repeats. This helped ensure that our model was generalizable and not overfitting on our dataset. We then trained a model to predict the "death_per_case" target variable using linear regression with scaled variables and the previously constructed control model. This allowed us to identify which variables were most important in predicting our target variable and thus useful for the classification task.

With these two components, we applied the variable importance method to extract the important features that the algorithm identified. These features were crucial for our classification analysis and allowed us to focus on the most significant variables in our dataset, while also removing any irrelevant features. The important variables that the algorithm identified are shown in Figure 4 and many of them are what we expected for pandemic-related variables. Overall, this approach allowed us to improve the accuracy of our model by using only the most important features, and we were able to gain valuable insights into the factors that contribute to COVID-19 mortality rates.

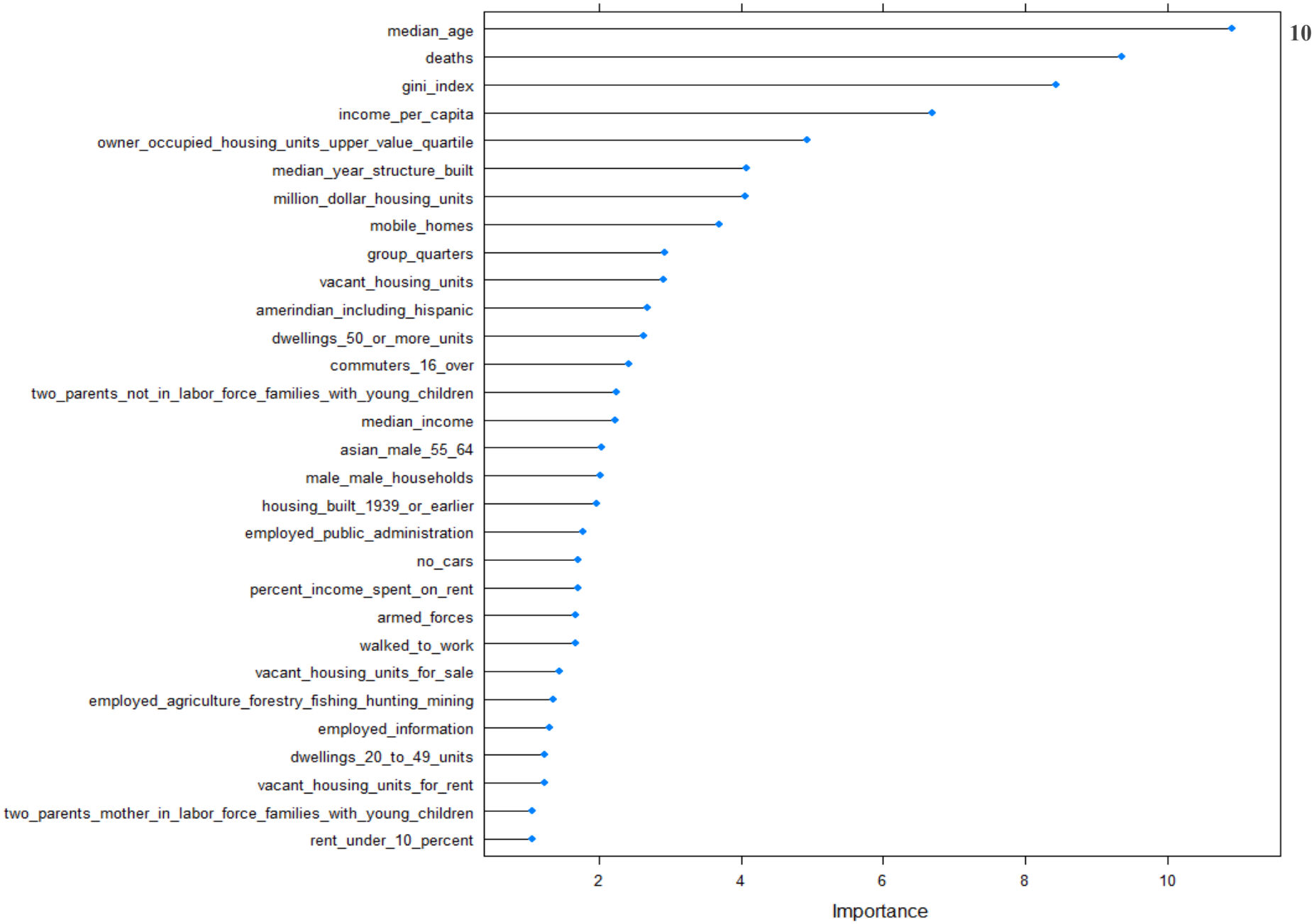

Figure 4. Top 30 Important Features.

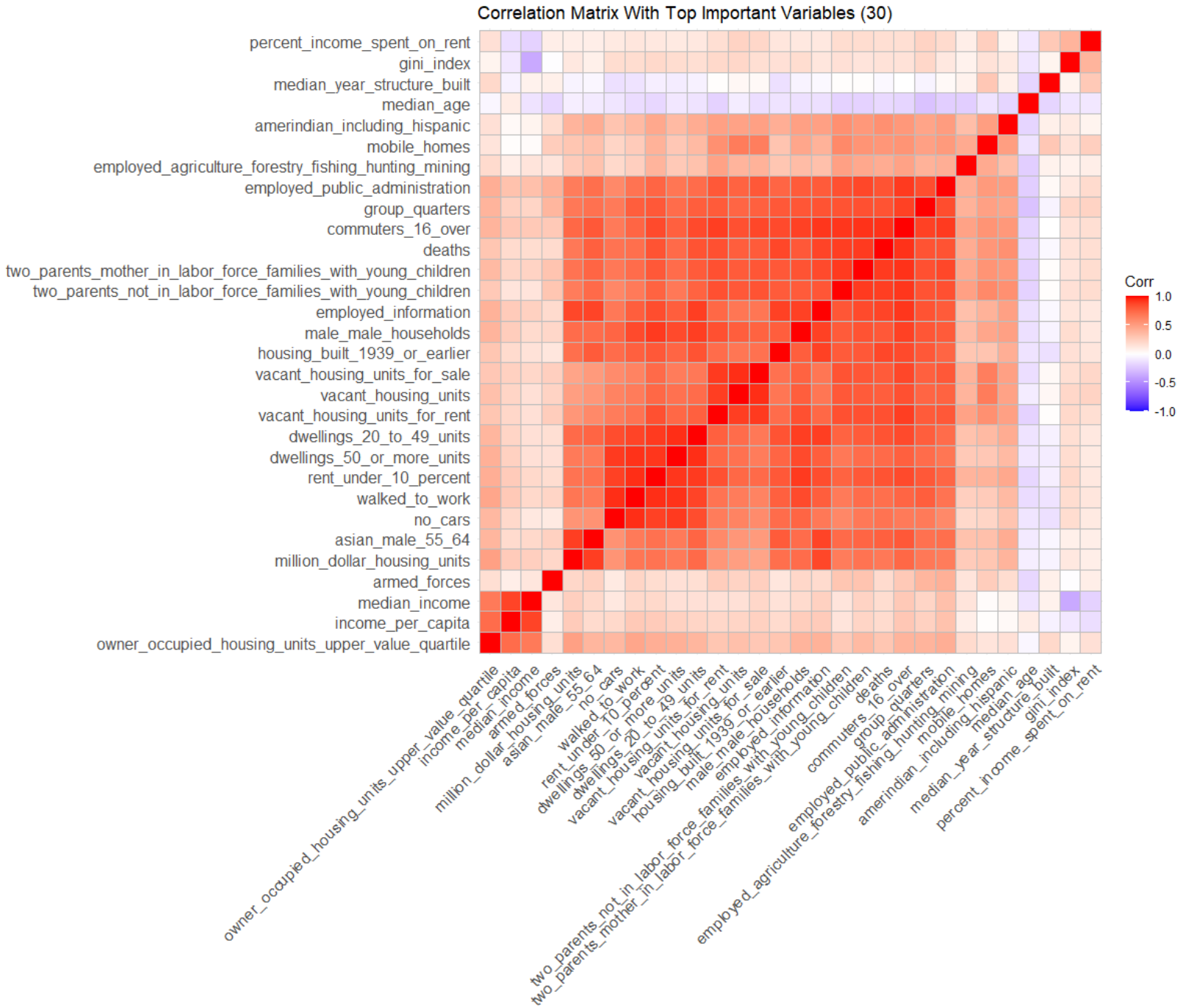

Figure 5. Correlation Matrix With Important Features.

After identifying the important features, we attempted to perform outlier removal on the remaining subset of the initial data. However, we found that the outlier removal process was actually counterproductive to the classification task. Therefore, we decided to take the same approach as our previous project, which involved subsetting the original data based on the important variables we identified, along with the additional class predictors we mentioned earlier. We will discuss the outlier removal process in more detail in a later section, although we did not use it to create our finalized dataset. Instead, we chose to subset the original data and add any anomalies back into the data. This decision was based on the fact that anomalies, such as Los Angeles County's high population, do exist in data about United States counties. Removing such counties from the data would not be ideal for the classification task, as we want to have as much information as possible to accurately classify states into their respective risk levels for an upcoming fifth wave if there is one. Therefore, we decided to skip outlier removal for now and move directly to the subsetting stage, which presented its own set of complications, including the disappearance of entire states in our mapped data and more.

Subsetting And Complications

Following a similar approach to our previous project, we began by subsetting the original census dataset to include the most important features. We also added four extra variables, namely "employed_pop", "unemployed_pop", "in_school", and "in_undergrad_college". To obtain more valuable information, we calculated three additional columns: “cases_per_10000”, “deaths_per_10000”, and “death_per_case”. To ensure the quality of the data, we removed counties with zero or fewer confirmed cases and omitted any rows with missing values. We noted that six total counties had been removed - five from the missing value omission and one from the filtering. After subsetting and adding the features, we normalized and combined the data with county positions to create a visual map. However, we noticed that several counties, including all counties in Louisiana, were missing from the map. This was unexpected since we had only removed six counties during the subsetting process. We knew that counties in Hawaii, Alaska, and Washington DC would not appear on the map, as discussed in our data exploration for the county position data, but many of these counties not displayed were simply unaccounted for. To identify the issues, we delved into the data and made extensive corrections to our subsetting and combining steps. This process required careful attention to detail and significant effort.

The Louisiana Issue

The issue that brought our attention to this major problem in our subsetting and combining section was the complete absence of the state of Louisiana from our graphs. Despite Louisiana counties being present in both our census and county position data, we could not see them on the map. We initially suspected that we had accidentally dropped a table of rows

containing Louisiana counties, but we could not find any problems in these related steps. After struggling to find errors in our processes, we decided to look at the provided starter code. However, even here, we noticed that Louisiana was missing from all maps, which we found to be quite suspicious. This discovery prompted one of the authors, who is a Political Science major, to realize that Louisiana is an exception when it comes to political subdivisions. Unlike other states in the US, Louisiana is divided into parishes instead of counties. This is because when Louisiana became part of the United States, its legal system was based on French and Spanish law, which recognized parishes as a division of government. The Louisiana Territorial Council formally adopted the parish system of government in 1807, which was later enshrined in the state's constitution in 1812. We checked our datasets and found that this was indeed the cause of our issue. The dataset containing COVID-19 cases and census had Louisiana counties listed with the word "parish" appended at the end of each name, while the county position data did not. Since we were merging these datasets with the county names, the parishes of Louisiana were left out, leading to their absence from our visualization. As such, we knew that if we could remove the appended “parish” in the census data, we would be able to merge these datasets and maintain the counties in Louisiana. To resolve the issue, we utilized regular expressions to remove the word "parish" from the Louisiana county names in the COVID-19 cases and census dataset, which allowed them to match with the county position data used for creating maps. After merging the datasets, Louisiana was successfully mapped onto the counties in the visualization, meaning we had successfully removed the appended “parish” in the census data. However, this fix led us to notice other counties that were also left out of the map visualization. To address this issue, we performed additional regex and mutation on both datasets to ensure that all county names were consistent and matched with the county position data. To provide a comprehensive overview, we list each of the remaining discrepancies between the datasets and their respective solutions below. Since we already mention the Louisiana Issue, we leave it out of the list.

- Some county names in the census data contained special characters, such as apostrophes and periods, while the county position data did not. We removed these special characters from the census data to ensure consistency in the county names across datasets. This fixed many of the issues, including the difference between “St.” and “St”, “Ste.” and “Ste”, “O’Brien” and “OBrien”, and “Queen Anne's” and “Queen Annes”, just to name a few. This added back important counties such as St. Louis, which is a large city in Missouri.

- We also found that some county names were split into two words sometimes and combined into a single word otherwise. For instance, "Du Page County" and "DuPage County" were both present in both datasets as well. We decided that it would be best to merge all instances of such names in both datasets, rather than adding spaces to some words, as the addition of spaces in county names could prove to be a difficult task. We focused on specific cases, such as counties with "La" and "De" in their names, and

performed the necessary corrections. After these changes were made, we observed a significant improvement in the visualization.

- The code provided modifies the county names in the dataset to standardize their format. This is done by removing certain phrases and fixing a spelling difference, such as changing "consolidated municipality of " to an empty string, effectively removing it from the county name. Similar changes are made for other phrases like "city and county of ", "city of st louis", and "town and county of ". After these changes, the code also standardizes the county names that begin with "city of " by removing this prefix. This is done to maintain a consistent format for all county names in the dataset. We briefly go over each of these changes and why they are listed differently between the datasets:
    - Consolidated Municipality Of Carson City: In 1969, the county merged into Carson City to consolidate government services and is now a consolidated municipality, making for a much more efficient government. We needed to remove this beginning part from the census data so as to be able to perform the merge.
    - City Of St. Louis: On August 22, 1876, the city of St. Louis voted to secede from St. Louis County and become an independent city, and, following a recount of the votes in November, officially did so in March 1877. Industrial production continued to increase during the late 19th century. This difference was extremely important since there were two counties differently named “St. Louis” and “St. Louis City”. Making sure we did not overlap the names was important.
    - Town And County Of Nantucket: It constitutes the Town and County of Nantucket, a combined county/town government in Massachusetts, a U.S. state. As such, it is named in this way, causing issues with the merge.
    - The remaining two are simply different naming conventions and don’t have any explicit reasoning behind their differing names. For example, in the census data, we have “City and County of San Francisco”, while in the county position data, we simply have “San Francisco”. We fixed these issues as well and merged.
- The county position data used to create maps also had county names in full lowercase, while the census data had county names in title case. To make the county names consistent across datasets, we simply converted the names in the census data to lowercase. At this step, we also removed the appended “county” from the census data.
- The county position data used to create maps did not include Alaska counties, Hawaii counties, or Washington DC, which we had previously noted in our data exploration. Since we did not want to manually add the longitude and latitude values to correctly

display these areas, we decided to leave them out. Note that with more time, we would be able to add these counties back and obtain the fullest possible picture for classification.

In the analysis, attention is given to the counties that, even after applying regular expressions for data cleaning, were not included in the final dataset due to missing values or other discrepancies. The removed counties consist of Daggett County in Utah, where the percent income spent on rent was not available; King and Kennedy Counties in Texas, both lacking owner-occupied housing units upper-value quartile data; Kalawao County in Hawaii, with essentially garbage data and no confirmed cases; Lake and Peninsula Borough in Alaska, missing median year structure built data; and Hoonah-Angoon Census Area in Alaska, which was filtered out due to zero confirmed cases. It is important to note that the owner-occupied housing units upper-value quartile is considered a significant variable, and thus it is retained in the dataset. Additionally, the District of Columbia in Washington DC and Hawaii and Alaska have missing county location information. The standardized county names allow for consistent identification and comparison of the data across different counties. This is important when making decisions and drawing conclusions from the dataset. Furthermore, removing counties with missing or unreliable data is necessary to maintain the integrity of the analysis. Missing values can lead to biased or inaccurate results, which could affect the conclusions drawn and the decisions made based on those conclusions. By filtering out counties with missing data, such as Daggett County in Utah, King County and Kenedy County in Texas, Kalawao County in Hawaii, Lake and Peninsula Borough in Alaska, and Hoonah-Angoon Census Area in Alaska, the analysis can focus on the most reliable and complete information available.

As an extra bit, we decided to look at the state of Alaska. Alaska is divided into boroughs, not counties. Because of the state's vast size and sparse population, traditional county government structures are not practical. Unlike counties, boroughs in Alaska are not based on population density or geographic area. Instead, they are formed by local petition and voter approval, allowing communities to define their own boundaries and services. We looked at the COVID-19 cases/census dataset and found that nearly all Alaska counties ended with the word “borough” or were the “Consolidated City and Borough of”. Although we are not using Alaska in our analysis, it is still an interesting challenge posed by politics and culture. It is worth noting that data preprocessing is just the first step in the analysis process. Once the data has been cleaned and standardized, it can be used for various purposes, such as identifying trends, making predictions, and informing policy decisions. In the context of the COVID-19 pandemic, having accurate and reliable data is essential for health officials and policymakers to make informed decisions about public health measures and resource allocation. By ensuring that the dataset is clean and consistent, these decision-makers can have a better understanding of the situation on the ground and can tailor their responses accordingly.

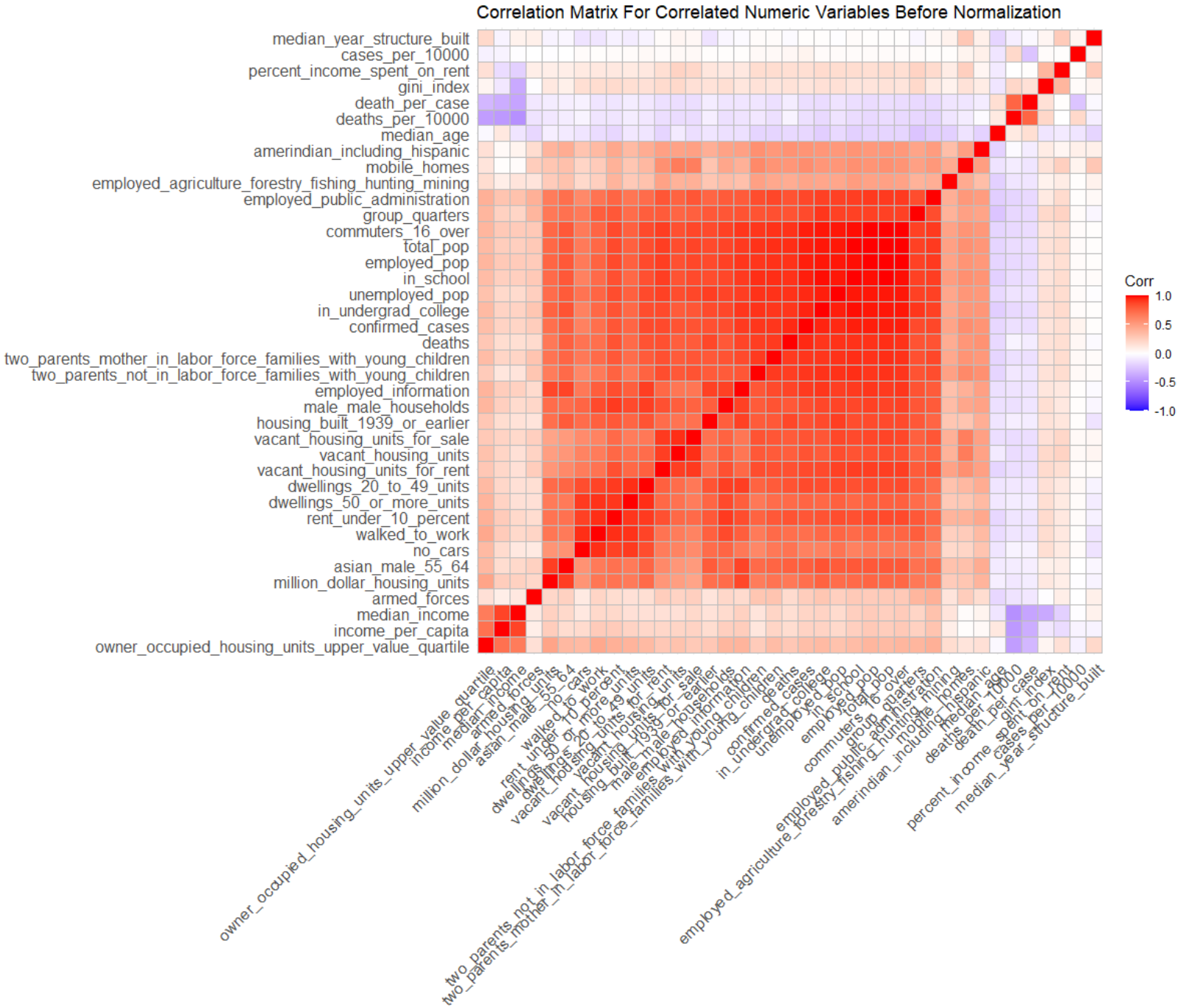

Figure 6. Final Correlation Matrix Before Normalizing Features.

Figure 6. Final Correlation Matrix After Normalizing Features.

Normalization And Finalized Data

Normalization is a data preprocessing technique that aims to scale numeric variables to a common range, usually between 0 and 1 or -1 and 1. This is particularly useful when dealing with variables that have different units or scales, as it helps ensure that each variable is treated equally during the analysis. In the case of a correlation matrix, normalization can impact the matrix because it eliminates the influence of different scales on the correlations between variables. Without normalization, variables with larger ranges or higher magnitudes might dominate the correlations, leading to potentially misleading or inaccurate results. For example, consider two variables: one representing income in dollars (ranging from thousands to tens of thousands) and another representing the percentage of the population with a college degree (ranging from 0 to 100). If these variables are not normalized, the income variable's higher magnitude might overshadow the relationship between the two variables in the correlation matrix, making it difficult to identify any meaningful patterns or relationships. By normalizing the variables, each variable is brought to the same scale, allowing for a more accurate and meaningful comparison of their relationships. This ensures that the correlation matrix reflects the true relationships between variables, rather than being influenced by the variables' magnitudes or scales. In this way, normalization contributes to a more reliable and informative correlation matrix, which can be used to better understand the underlying patterns and relationships between variables in the context of the analysis. Above, we can see the effect of normalization and how the correlation matrix changes depending on normalization alone. Note that in our case, we normalized by the total population for each given county, and as such, values are greater than zero with no capacity specifically defined as is common for normalization. Further, we left some unnormalized features in the dataset but did not perform classification on them. Rather, we used them to create our classes, as we will discuss later on. In other words, we may expect some interesting results for the statistics summary, but these have all been accounted for. Note that for the statistics summary, since the numbers for each column vary drastically among different important features shown in the important variables table and having two rows with the same number for some columns provides no value for the purpose of this report, we decided against displaying mode. Further, we opted for displaying standard deviation rather than variance as the interpretability of the numbers for variance was difficult since the values were quite large. Finally, we included minimum and maximum rather than range since we can garner more information from the data with these statistics than we could from range. Additionally, the range for most of the features is a trivial calculation since the minimum was commonly 0.00. This feature selection allowed us to analyze the data more accurately.

In the context of COVID-19, the correlation matrix presents correlations between various socioeconomic and demographic factors, as well as COVID-19 case and death data. The matrix has been constructed after normalizing the data to ensure accurate comparisons and conclusions.

The large red square in the correlation matrix signifies that there is a strong positive correlation between income per capita, median age, median year structure built, median income, percent income spent on rent, Gini index, and owner-occupied housing units upper-value quartile. This implies that these variables tend to increase or decrease together. For example, an area with higher income per capita might also have a higher median age and a more recent median year structure built. The strong positive correlation could indicate that these factors are interconnected, suggesting that the economic well-being of a region might be closely related to the age of its population and the quality of its housing. The smaller red square shows a positive correlation between confirmed cases, total population, and deaths. This is not surprising, as regions with larger populations would be expected to have more confirmed cases and deaths due to the higher number of individuals potentially exposed to the virus. It is essential to consider the population size when examining COVID-19 data to account for differences in scale and to better understand the impact of the virus on various communities. The darkest blue square between mobile homes and commuters over the employed population signifies a strong negative correlation. This means that as the number of mobile homes in an area increases, the proportion of commuters within the employed population decreases, and vice versa. This negative correlation could be attributed to the different lifestyles and living situations associated with mobile homes and commuters. Mobile home residents may be less likely to commute for work due to factors such as job availability, income, and transportation options. Conversely, areas with a higher proportion of commuters might have fewer mobile homes as residents choose other types of housing that are closer to their workplaces or offer better access to transportation networks. Understanding these correlations is crucial in the context of the COVID-19 pandemic, as it provides insights into how the virus spreads and impacts various communities. For instance, the strong positive correlation between socioeconomic factors and housing quality could suggest that wealthier areas might be better equipped to handle the pandemic due to their access to resources and more robust infrastructure. Policymakers and health officials can use this information to target interventions and allocate resources more effectively, ensuring that vulnerable communities receive the support they need.

Similarly, the correlation between confirmed cases, total population, and deaths highlights the importance of considering population size when evaluating the impact of the virus. This knowledge can help health officials make informed decisions about public health measures, such as lockdowns, social distancing guidelines, and vaccination campaigns. Moreover, understanding the relationship between mobile homes, commuters, and the employed population can provide insights into the unique challenges faced by different communities during the pandemic. For example, mobile home residents might face specific barriers to accessing healthcare or adhering to social distancing guidelines due to their living situations. With that, we finally had the dataset on which we wanted to perform the classification problem on. Below is the intro plot for this finalized dataset.

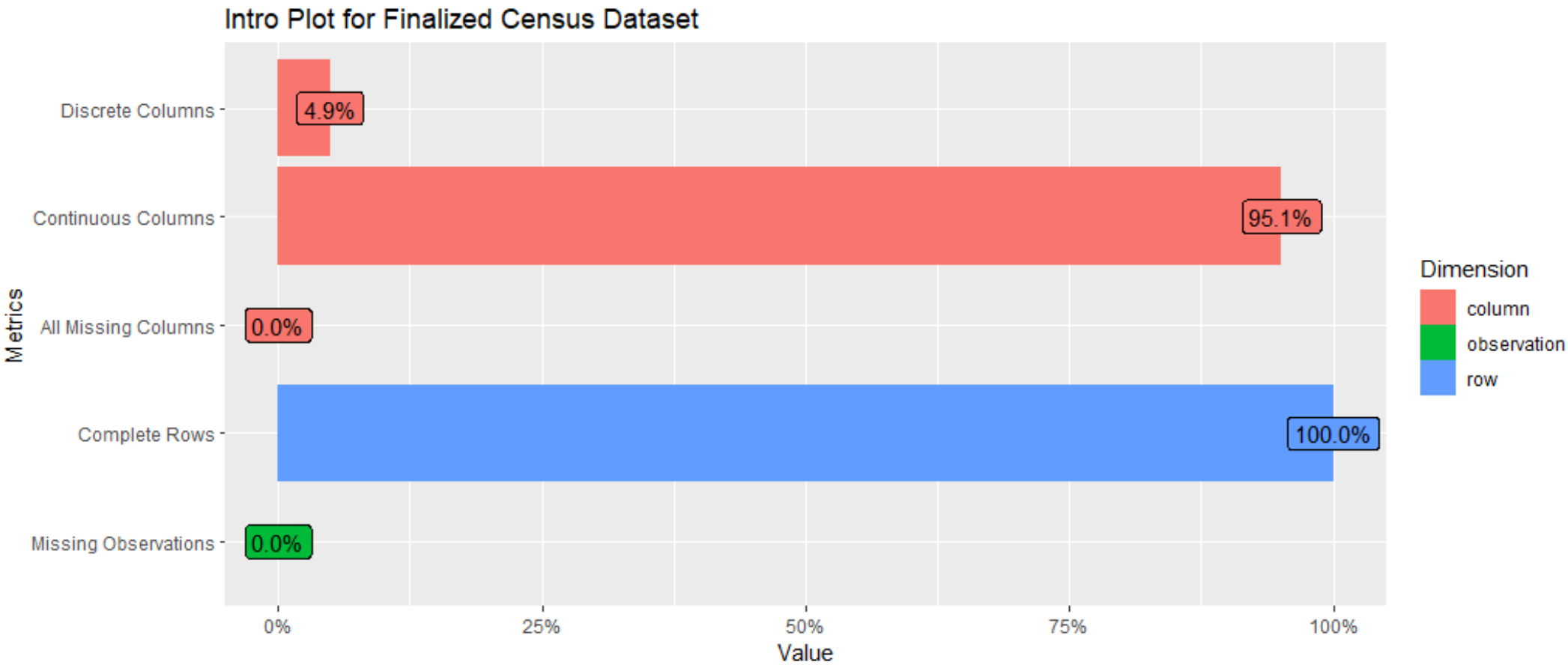

Figure 7. Intro Plot for Finalized Census Dataset.

Outlier Removal

As previously mentioned, we included the methodology for outlier removal below the creation of the finalized dataset since it did not have any impact on our final product. Outliers are data points that are significantly different from the rest of the dataset. In classification, outliers can have a significant impact on the accuracy of the model. When outliers are present in the training set, they can affect the model's ability to learn the underlying patterns in the data. If the outlier is labeled incorrectly, it can mislead the classification algorithm and affect its ability to accurately classify new data. For example, if an outlier in the training set is labeled as belonging to a particular class, but is not representative of that class, the model may overfit the outlier and perform poorly on new data. This can result in a high error rate and reduced accuracy. Additionally, if the outlier is a significant distance from other data points in the training set, it can affect the decision boundaries of the classification algorithm. The decision boundaries define the regions in which the different classes are separated. An outlier that is far from the other data points can affect the location of the decision boundaries, leading to the misclassification of new data. When outliers are present in the testing set, they can also affect the accuracy of the classification model. Outliers in the testing set can be difficult for the model to classify accurately, as they are not representative of the underlying patterns in the data. If the outlier is very different from the other data points in the testing set, it can lead to a misclassification of the entire sample. This can be particularly problematic if the outlier is a rare event or a new pattern that the model has not seen before. In this case, the model may not be able to generalize to new data, leading to poor performance on future classification tasks. However, in some cases, the removal of outliers can be detrimental to our classification model, causing issues with the results and not learning the variables in an effective way. Below, we see the box plots corresponding to the normalized census data, both before and after removing the outliers from the dataset.

Since our dataset creation is completed and we have looked at the methodologies used to clean, merge, and finalize our dataset, we now can move on to the class creation within the dataset, which we need to perform the classification task. To summarize what we have done up to this point, we used our census data for the classification problem and created a singular finalized dataset fit for the classification problem - a single table with some class attribute, which we will define in the next step. We also have identified the predictive features in the data using variable importance, created additional predictive features and another which will work as the class attribute as we will discuss later on, and dealt with missing data or problematic data by mutating, filtering, and constructing our dataset such that the classification model can handle the whole dataset. We define these important features and their individual statistics, including mean, median, standard deviation, minimum, and maximum, in the tables below. We used these statistics for similar reasons to previous projects - we felt these most effectively captured the meaning and expected values for each individual variable.

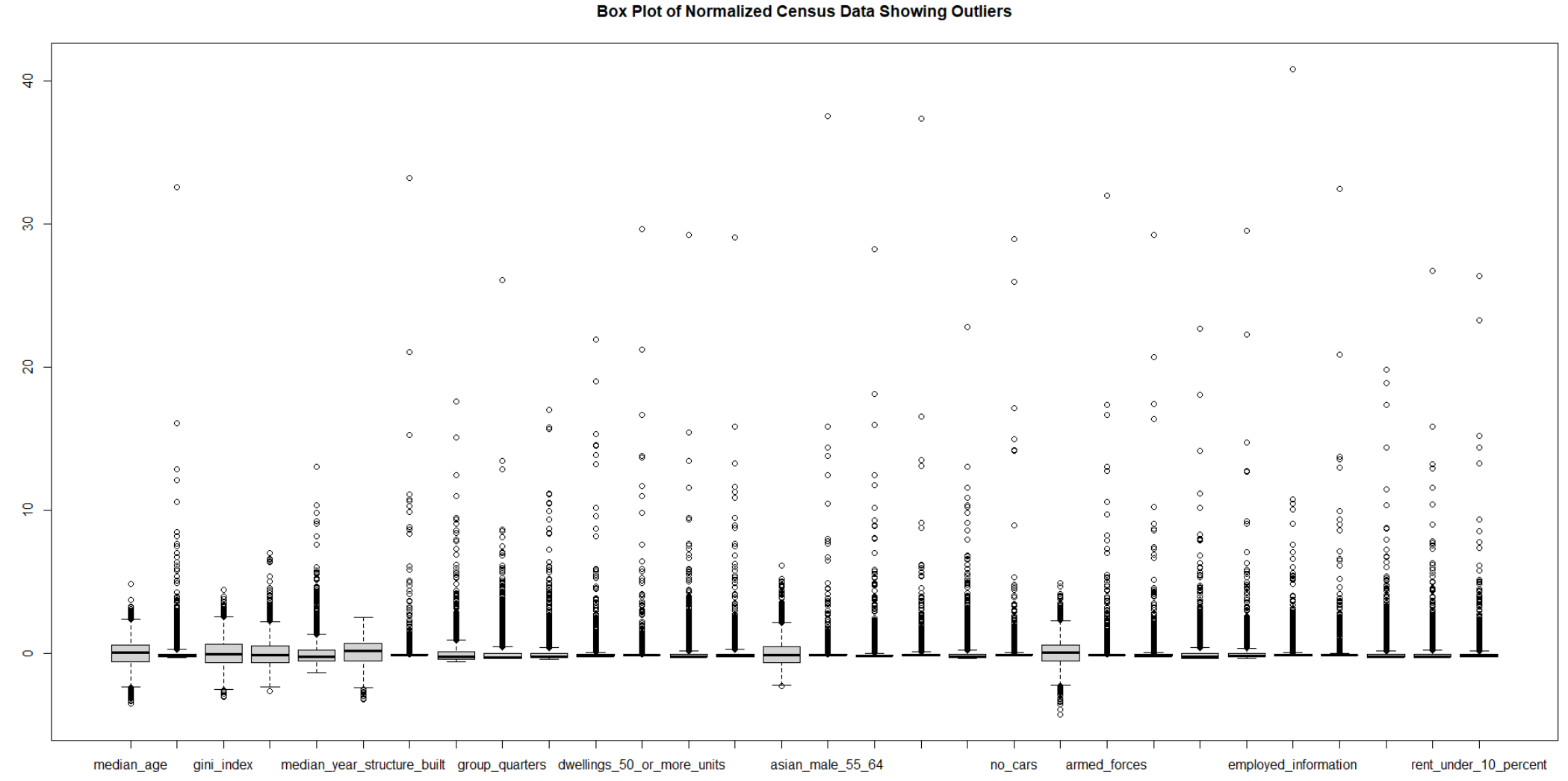

Figure 8. Box Plot of Normalized Census Data Showing Outliers.

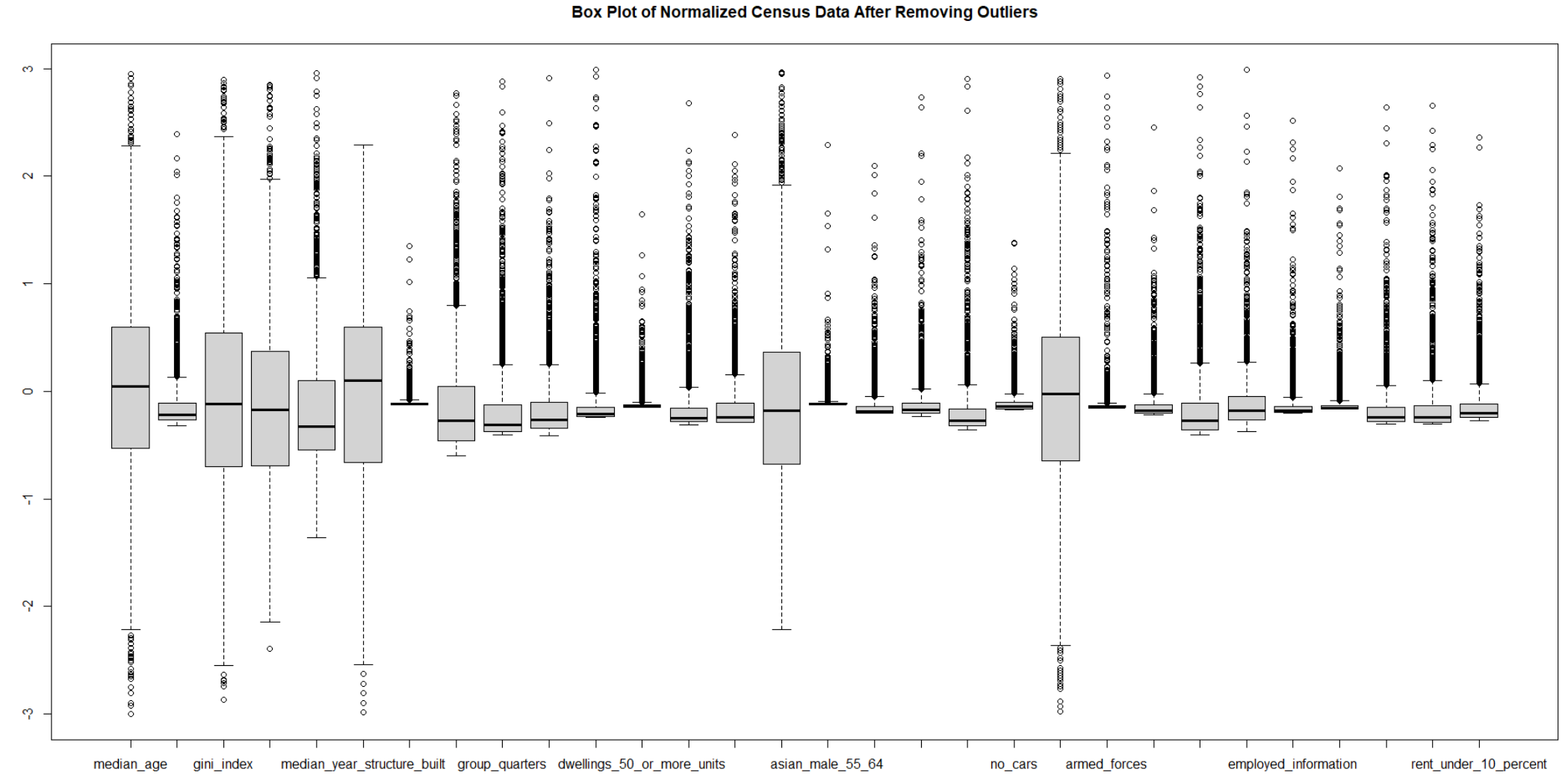

Figure 9. Box Plot of Normalized Census Data After Removing Outliers.

Important Features

Feature	Scale	Dataset	Description
state	Nominal	Census Data	State name for a given county.
county	Nominal	Census Data	County name denoting an observation.
total_pop	Ratio	Census Data	The total population in the given county.
confirmed_cases	Ratio	Census Data	The number of confirmed cases in a given county.
deaths	Ratio	Census Data	The number of deaths in a given county.
cases_per_10000	Ratio	Census Data	The number of cases per 10,000 people in a given county.
deaths_per_10000	Ratio	Census Data	The number of deaths per 10,000 people in a given county.
death_per_case	Ratio	Census Data	A ratio of deaths to COVID-19 cases.
median_age	Ratio	Census Data	The median age of individuals in a county.
gini_index	Interval	Census Data	A count of non-rented housing units that fall within the upper-value quartile within a given county.
income_per_capita	Ratio	Census Data	The mean income is computed for every individual in a given county.
owner_occupied_housing_units_upper_value_quartile	Ratio	Census Data	A count of non-rented housing units that fall within the upper-value quartile within a given county.
median_year_structure_built	Interval	Census Data	The median construction year in a county.
million_dollar_housing_units	Ratio	Census Data	The number of housing units valued at a million dollars or more in a given county.
mobile_homes	Ratio	Census Data	The number of mobile homes registered in a county.
group_quarters	Ratio	Census Data	A count of individuals living in group quarters, like college dorms or military barracks.
vacant_housing_units	Ratio	Census Data	The number of vacant housing units in a given county.
amerindian_including_hispanic	Ratio	Census Data	The number of individuals that identify as Amerindian, including Hispanic, in a given county.
dwellings_50_or_more_units	Ratio	Census Data	The number of dwellings with 50 or more housing units in a given county.
commuters_16_over	Ratio	Census Data	The number of commuters that are aged 16 and over in a given county.
median_income	Ratio	Census Data	The median income for individuals in a county.

two_parents_not_in_labor_force_families_with_young_children	Ratio	Census Data	The number of family units where two parents are not in the labor force in a given county.
asian_male_55_64	Ratio	Census Data	The number of individuals who identify as Asian and male, and are between the ages of 55 and 64.
male_male_households	Ratio	Census Data	The number of households led by same-sex male couples in a given county.
housing_built_1939_or_earlier	Ratio	Census Data	The number of housing units built in 1939 or earlier in a given county.
employed_public_administration	Ratio	Census Data	The number of individuals employed in the field of public administration within a given county.
no_cars	Ratio	Census Data	The number of households without a car in a given county.
percent_income_spent_on_rent	Ratio	Census Data	The average percentage of income spent by households in a given county.
armed_forces	Ratio	Census Data	A count of individuals in the armed forces.
walked_to_work	Ratio	Census Data	The number of individuals who walk to work in a given county.
vacant_housing_units_for_sale	Ratio	Census Data	The number of vacant housing units that are also listed for sale in a given county.
employed_agriculture_forestry_fishing_hunting_mining	Ratio	Census Data	The number of people employed in the agriculture, forestry, fishing, hunting, or mining industries.
employed_information	Ratio	Census Data	The number of individuals employed in the information industry in a given county.
dwellings_20_to_49_units	Ratio	Census Data	The number of dwellings with 20 to 49 housing units in a given county.
vacant_housing_units_for_rent	Ratio	Census Data	The number of vacant housing units that are also listed for rent in a given county.
two_parents_mother_in_labor_force_families_with_young_children	Ratio	Census Data	The number of households where the parental units are two mothers working in the labor force who also have young children in a given county.
rent_under_10_percent	Ratio	Census Data	The number of households where rent is under 10 percent of their income in a given county.
employed_pop	Ratio	Census Data	The number of individuals who are employed in a given county.
unemployed_pop	Ratio	Census Data	The number of individuals who are unemployed in a given county.
in_school	Ratio	Census Data	The number of individuals who are in school in a given county.
in_undergrad_college	Ratio	Census Data	The number of individuals who attend an undergraduate college in a given county.

Table 1. Important Features Selected For Classification. All Features are in Euclidean Distance Measures.

Summary Statistics

Feature	Mean	Std. Dev	Min	Pctl. 25	Pctl. 75	Max
state	-	-	-	-	-	-
county	-	-	-	-	-	-
total_pop	102359	3.285762e+05	74	11004	67583	10105722
confirmed_cases	29803	1.018584e+05	30	2940	19629	3416156
deaths	315.74	1.023181e+03	0.00	44.75	247.25	34351.00
cases_per_10000	2833.0	1.065771e+03	286.9	2399.5	3235.8	47702.7
deaths_per_10000	38.93	1.708500e+01	0.00	27.18	49.63	150.00
death_per_case	0.014505	8.106367e-03	0.000000	0.009718	0.017368	0.090278
median_age	0.0042548	1.635170e-02	0.0000036	0.0005804	0.0037816	0.7770270
gini_index	4.134e-05	1.143098e-04	5.000e-08	6.603e-06	3.990e-05	4.600e-03
income_per_capita	2.5265	1.007141e+01	0.0030	0.3754	2.0935	480.1351
owner_occupied_housing_units_upper_value_quartile	17.7571	5.242809e+01	0.0558	3.1029	15.9109	1858.1081
median_year_structure_built	0.186985	5.949310e-01	0.000194	0.029125	0.179490	26.675676
million_dollar_housing_units	0.001449	3.042191e-03	0.0000000	0.0001265	0.0014667	0.0509904
mobile_homes	0.06435	5.125402e-02	0.00000	0.02463	0.09218	0.56355
group_quarters	0.03533	4.623558e-02	0.00000	0.01242	0.03839	0.60734
vacant_housing_units	0.101557	1.053930e-01	0.007259	0.046948	0.115674	1.676711
amerindian_including_hispanic	0.019161	7.557999e-02	0.000000	0.001645	0.008197	0.928276
dwellings_50_or_more_units	0.0057986	1.241413e-02	0.0000000	0.0002496	0.0070117	0.2925121
commuters_16_over	0.41007	5.906652e-02	0.09394	0.37288	0.45260	0.67981
two_parents_not_in_labor_force_families_with_young_children	0.0007717	1.347308e-03	0.0000000	0.0000000	0.0009581	0.0270308
median_income	4.6908	2.170406e+01	0.0060	0.7294	3.8979	1093.7568

asian_male_55_64	0.0005988	1.759604e-03	0.0000000	0.0000000	0.0006308	0.0338866
male_male_households	0.0003091	5.275121e-04	0.0000000	0.0000000	0.0004612	0.0110375
housing_built_1939_or_earlier	0.02619	1.613861e-02	0.00000	0.01543	0.03340	0.16035
employed_public_administration	0.02379	1.264171e-02	0.00000	0.01595	0.02807	0.14370
no_cars	0.02460	1.536697e-02	0.00000	0.01656	0.02964	0.35284
percent_income_spent_on_rent	2.347e-03	6.878231e-03	3.430e-06	4.269e-04	2.400e-03	3.041e-01
armed_forces	0.0022489	1.220716e-02	0.0000000	0.0000000	0.0008645	0.4021269
walked_to_work	0.014223	1.886777e-02	0.000000	0.005557	0.016796	0.423607
vacant_housing_units_for_sale	0.005950	4.732624e-03	0.000000	0.003190	0.007552	0.065421
employed_agriculture_forestry_fishing_hunting_mining	0.029687	3.466819e-02	0.000000	0.007734	0.037010	0.289183
employed_information	0.006267	4.159455e-03	0.000000	0.003694	0.008048	0.089485
dwellings_20_to_49_units	0.006784	9.329965e-03	0.000000	0.001836	0.008613	0.153758
vacant_housing_units_for_rent	0.008874	7.385541e-03	0.000000	0.004731	0.010981	0.123023
two_parents_mother_in_labor_force_families_with_young_children	0.0013727	1.536740e-03	0.0000000	0.000344	0.0018261	0.0179286
rent_under_10_percent	0.005820	4.309161e-03	0.000000	0.003385	0.007049	0.109308
employed_pop	0.4382	6.509182e-02	0.1017	0.3959	0.4861	0.7205
unemployed_pop	0.02881	1.212808e-02	0.00000	0.02152	0.03494	0.12619
in_school	0.23140	4.277665e-02	0.07762	0.20648	0.25192	0.63040
in_undergrad_college	0.04320	3.441545e-02	0.00000	0.02599	0.04859	0.53901

Table 2. Statistical Summary of Important Features. Normalized By Total Population.

Class Definition

After creating the finalized dataset, we needed to determine which class (or classes) to predict. After much consideration, we decided to use the "death_per_case" variable as our class attribute. We believe that this variable provides the most accurate indication of a county's future risk since counties with high death per case rates are more likely to experience significant mortality rates compared to other counties. This variable is preferred over just using deaths or cases because a county like Los Angeles may have many cases and deaths, but this is largely due to its large population size compared to most other counties in the United States.

To prepare for the classification task, we used the “death_per_case” variable to create a class variable that would predict the risk of a county facing mass death in the future. We opted to split the data into three classes to provide more accurate planning for any potential fifth wave. We thought that, by giving more classes, experts could more accurately prepare infrastructure and possible interventions in the correct locations, therefore saving lives and shortening the length of necessary closings. We then needed to split the “death_per_case” variable into three classes, each of which contained similar amounts of data. The reason why we wanted each class to have a similar amount of data was to avoid class imbalance, which could skew our final results and cause incorrect planning. For example, if there is a class imbalance towards “low risk”, the majority of predicted classes would show as “low risk”, meaning that we would not accurately find the counties most at risk for the fifth wave. We called these classes “low risk,” “medium risk,” and “high risk,” and divided the “death_per_case” variable into each of these classes to create the new “deaths_class” variable. We assigned the “high risk” class to counties with “death_per_case” values greater than 0.016, “medium risk” to counties with values between 0.011 and 0.016, and “low risk” to counties with values below 0.011. We confirmed that each class had roughly 1000 instances to ensure balanced data for our models.

Type	High	Medium	Low
Full Dataset Instance Count	998	1032	1106

Table 3. Balanced Instance Counts for Each Class in Combined Dataset.

We checked the class distribution and found that there was no class imbalance. Additionally, we examined the counties that were most strongly associated with each risk category. These results allowed us to choose the states that would best help in the classification task. We took these strongly associated counties as our training data, and a subset of the remaining data as our testing, as we will see in the next section regarding modeling and the performing of the classification task.

Modeling

Preparing Data for Training, Testing, and Tuning

After assigning classes to our entire cleaned dataset, we split the datasets into a training dataset and a testing dataset. We selected Alabama, Maine, California, Michigan, Colorado, South Carolina, New Hampshire, Virginia, North Dakota, Wisconsin, Kansas, and Idaho as our test states, and selected every other state to be in our training data. Allowing our models to train from more states increases their accuracy. We similarly verified the balance of the training and test datasets. We spent several hours selecting particular states to be in each dataset based on keeping each dataset balanced. The final balanced results are in the table below.

Type	High	Medium	Low
Training Instance Count	729	750	854
Testing Instance Count	269	282	252

Table 4. Balanced Instance Counts for Each Class in Training/Testing Data. Ensure Models Can Learn Reasonably Well From This Data.

To further illustrate this balance in our training dataset, we plotted the risk level of each county in our training dataset onto a map of United States counties. As Figure 10 below shows, there is a roughly even divide between low-risk, medium-risk, and high-risk counties in our training data.

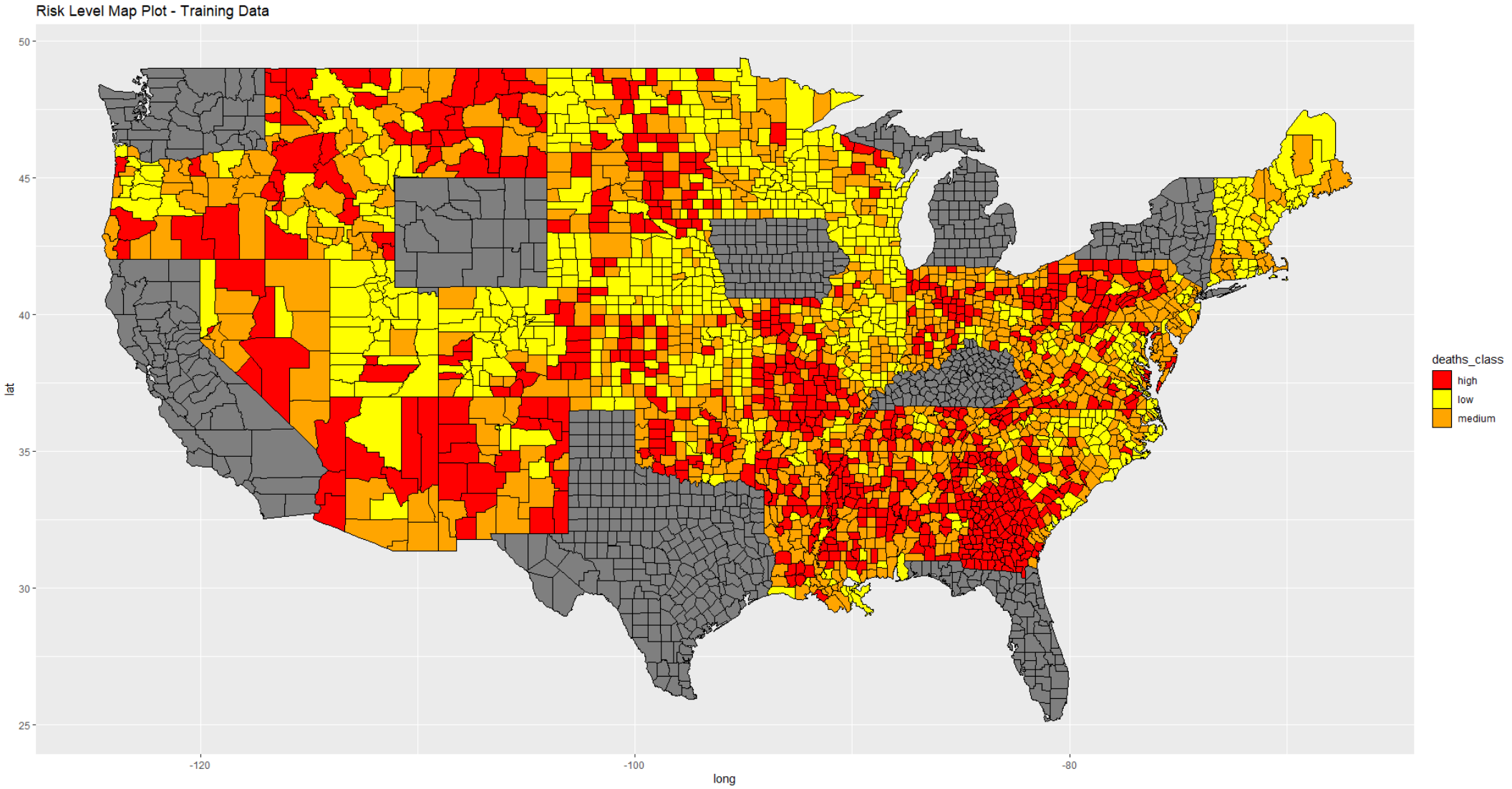

Figure 10. Risk Level Plot of Training Data.

At this point, we also checked for variable importance by running multiple chi-squared tests. The chi-squared test for attribute importance is a statistical test that is helpful for determining whether there is a significant relationship between a categorical attribute and a target variable in a dataset. In our case, we tested for significant relationships between all other variables and our target variable, the county's risk level. We use this test to rank the attributes in our dataset according to their importance for predicting the county's risk level. Initially, this gave poor results and stated that our class variable deaths per case had an attribute importance of 1. As shown in the table below, the attributes that ranked highest in terms of importance were some of the most obvious, such as deaths_per_10000.

Variable	Attribute Importance
death_per_case	1.0000000
county	0.7809546
deaths_per_10000	0.7084264
state	0.4607516
commuters_16_over	0.3652409
employed_pop	0.3611994

Table 5. Initial Chi-Squared Test for Attribute Importance.

To address the issue of unhelpful results, one solution we considered was to remove the class variable deaths_per_case. By doing so, we conducted another Chi-Squared test to determine the attribute importance, and the results are presented below.

Variable	Attribute Importance
county	0.7809546
deaths_per_10000	0.7084264
state	0.4607516
commuters_16_over	0.3652409
employed_pop	0.3611994
mobile homes	0.2862992

Table 6. Second Chi-Squared Test for Attribute Importance. Removed Class Variable.

In order to yield more significant and interesting attribute importance measurements, we also removed COVID-19-related variables such as deaths_per_10000, cases_per_10000, confirmed_cases, and deaths. Removing those variables yielded the results in the table below.

Variable	Attribute Importance
county	0.7809546
state	0.4607516
commuters_16_over	0.3652409
employed_pop	0.3611994
mobile homes	0.2862992
dwellings_20_to_49_units	0.2438369
vacant_housing_units	0.2411704
in_school	0.2238543
median_age	0.2171870
male_male_households	0.2137835

Table 7. Third Chi-Squared Test for Attribute Importance. Removed COVID-Related Variables.

Ignoring the first two attributes county and state since they are nominal, these tests yield interesting results. From these tests, for example, we know that the number of commuters ages 16 and over or the number of employed individuals are important attributes to consider when predicting the county's risk level. After removing obviously related to our deaths per case class variable, we created 3 different classification models and trained them with the training dataset.

K-Nearest Neighbors Model

First, we built a K-Nearest Neighbors model and trained it on our training dataset. It is important to note that to avoid confusing the model, we subset the training dataset to select all columns except for the columns "county" and "state". Our chi-squared tests showed those variables being important to determine the county's risk level, however, we want to be able to determine a county's risk level by their makeup, not by their geographic location. To achieve this,

we did not allow the model to train with the columns “county” and “state.” Furthermore, we centered and scaled the training data before fitting the model and specified a tune length of 5 to perform the tuning process for a maximum of 5 values of the hyperparameter “k.” We specified this “k” value to be a number between 1 and 10, inclusive. The tuning process is performed to find the optimal value of the hyperparameter “k” in the K-Nearest Neighbors algorithm. The “k” value controls the number of nearest neighbors to consider when making a prediction for a new observation. The optimal value of “k” will balance the bias-variance trade-off and usually results in the best predictive performance on test data. Our tuning process splits the training data into 10 equally-sized folds and evaluates the model on each fold, using the remaining folds for training. We used accuracy to select the optimal model. The final value used for the model was k = 1. In the table below, only the first 3 of the 10 k values are shown to illustrate the significant drop in accuracy after k = 1. For this reason, we used k = 1 in our final KNN model.

K value	Accuracy	Kappa
1	0.9502804	0.9251947
2	0.7299832	0.5935146
3	0.7008095	0.5491339

Table 8. Resampling Results Across Tuning Parameters For K-Nearest Neighbors (KNN).

Using k = 1, our model is a 1-nearest neighbor model. The training set outcome distribution for this model is shown in the table below.

Type	High	Medium	Low
Training Set Outcome Distribution	729	750	854

Table 9. Training Set outcome distribution for the 1-Nearest Neighbors Model.

Next, we gave the model our test dataset and we generated a confusion matrix for the 1-Nearest Neighbor model. Overall, the model had an accuracy of 0.3861, or 38.61%. Below is the map of the KNN Classifier’s predictions.

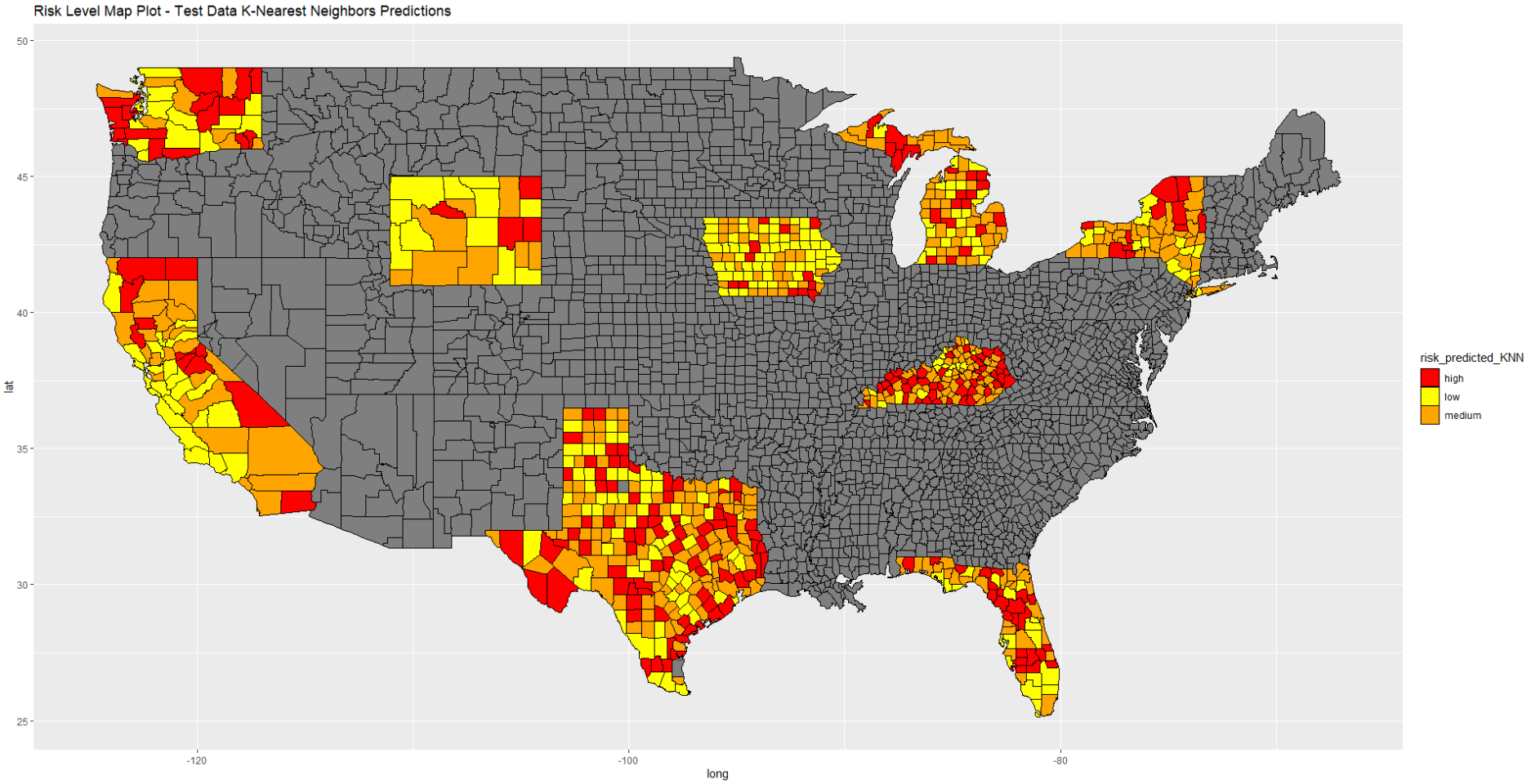

Figure 11. KNN Test Classifications.

	Actual		
Predicted	**High**	**Medium**	**Low**
High	87	56	85
Medium	109	114	111
Low	73	112	56

Table 10. Confusion Matrix for 1-Nearest Neighbors Model.

The 3x3 confusion matrix can, most importantly, tell us the number of correctly classified observations for each category. Our KNN model correctly classified 87 counties as high risk, 114 as medium risk, and 56 as low risk. These entries along the main diagonal are referred to as the diagonal elements. The confusion matrix can also tell us more about incorrect predictions. For example, our KNN model misclassified 56 counties that are actually classified as high-risk. Similarly, our model misclassified 73 counties that should have been classified as low-risk.

Random Forest Model

Next, we built a Random Forest model and trained it on our training dataset. Again, we subset the training dataset to select all columns except for the columns “county” and “state”.

We trained our RF model on the training dataset with 10-fold cross-validation and tuned the “mtry” parameter. We split the data into 10 folds, and the model is trained on 9 of the folds and tested on the remaining fold, with this process repeated 10 times. The “mtry” parameter, which controls the number of variables considered at each split in the decision tree, is specified to be tuned over a grid of 5 values. The optimal “mtry” value is selected based on the performance of the resulting models on the validation sets during cross-validation.

We used accuracy to select the optimal model. The final value used for the model was mtry = 18. In the table below, 5 mtry values are shown to illustrate that mtry = 18 results in the greatest accuracy. For this reason, we used mtry = 18 in our final RF model.

mtry value	Accuracy	Kappa
2	0.5649374	0.3464211
10	0.5653868	0.3468997
18	0.5701042	0.3545048
26	0.5658141	0.3476268
34	0.5653941	0.3468891

Table 11. Resampling Results Across Tuning Parameters for Random Forest (RF).

We also ran a variable importance function over our random forest model. We found some overlap with the initial variable importances gathered from our chi-squared tests, but we also found some interesting attributes that had not been listed previously. Below is a table that lists the top 10 most important variables shown out of the 34 found. An asterisk marks variables that were not listed in the chi-squared test.

	Random Forest Variable Importance
commuters_16_over	100.00
employed_pop	97.08
in_school	45.04
amerindian_including_hispanic*	41.11
mobile homes	38.25
dwellings_20_to_49_units	32.72
no_cars*	30.90
housing_built_1939_or_earlier*	30.50
group_quarters*	27.88
employed_public_administration*	26.73

Table 12. Random Forest Variable Importance. Only 10 Most Important Variables Were Shown.

Running this variable importance test on our RF model helped us gain insights into the relationship between predictor variables and a county's risk level. For example, the RF model found that Amerindian including Hispanic identification is important for predicting a county's risk level. Overall, the model had an Out-of-Bag estimate of an error rate of 43.33%. This means that, on average, the model can correctly classify roughly 56.67% of the counties in the dataset. We then gave the RF model the test dataset. We generated a confusion matrix and the resulting confusion matrix is shown in the table below. After running through the test data, the model had an accuracy of 44.08%. Below is the map of the Random Forest model's classification.

	Actual		
Predicted	**High**	**Medium**	**Low**
High	135	55	88
Medium	89	104	96
Low	45	123	68

Table 13. Confusion Matrix for Random Forest Model on Test Dataset.

The 3x3 confusion matrix can tell us the number of correctly classified observations for each category. For example, our RF model correctly classified 135 counties as high risk, 104 as medium risk, and 68 as low risk. The confusion matrix can also tell us more about incorrect predictions. For example, our RF model misclassified 55 counties that are actually classified as high-risk. Similarly, our model misclassified 89 counties that should have been classified as low-risk.

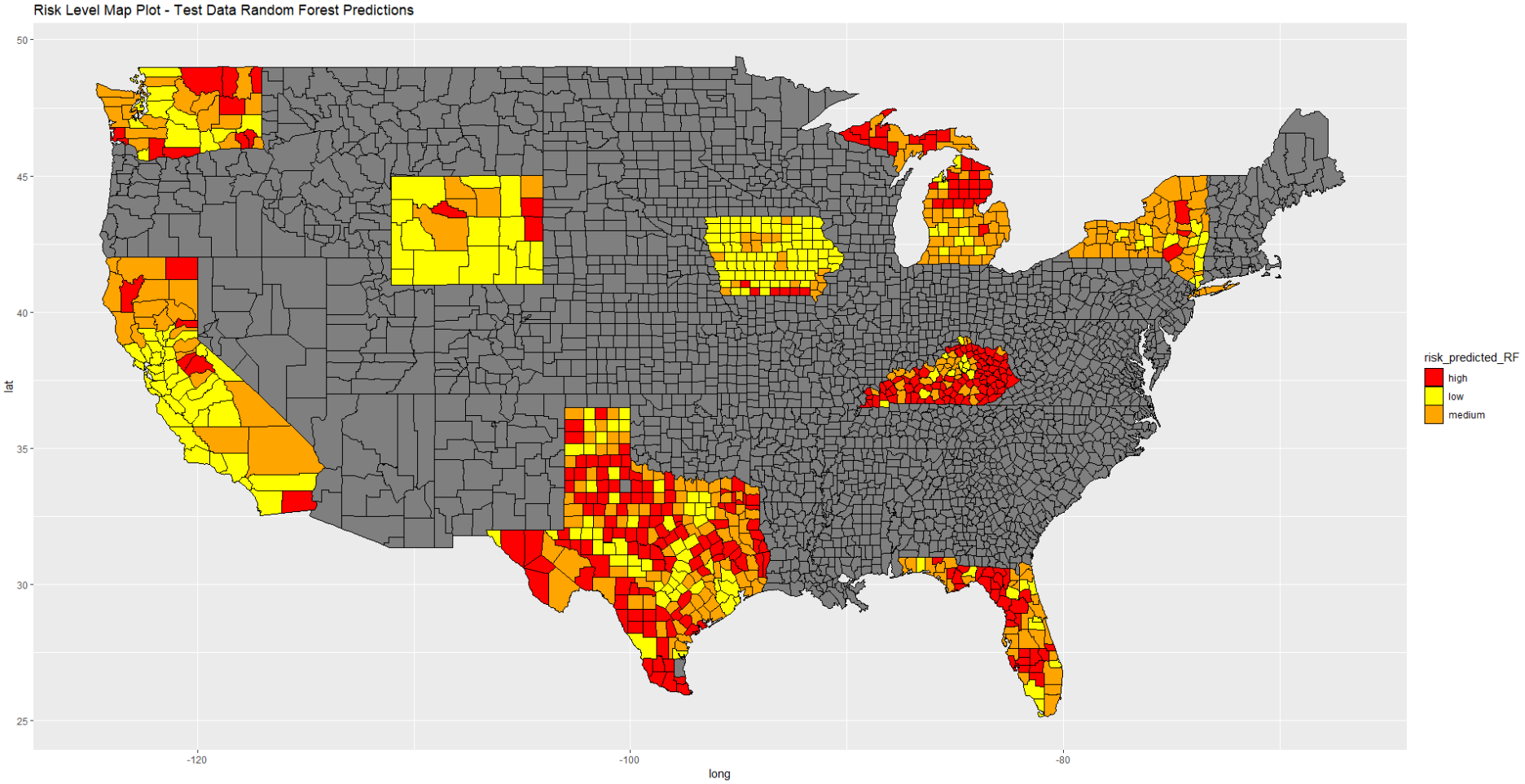

Figure 12. Random Forest Test Classifications.

Artificial Neural Network Model

Next, we built an Artificial Neural Network model and trained it on our training dataset. Again, we subset the training dataset to select all columns except for the columns "county" and "state". We trained our ANN model on the training dataset with 10-fold cross-validation and 5 values tested for each tuning parameter. Some of the hyperparameters that can be tuned include the number of hidden layers, the number of nodes in each hidden layer, the activation function used, and the weight decay parameter. The goal of tuning is to find the optimal parameter values that result in the best performance on the validation set, which improves the generalization performance of the model on the test data. We used accuracy to select the optimal model. The final values used for the model were size = 3 and decay = 0.1. In the table below, we show 5 sets of values to illustrate that our selection results in the greatest accuracy. Our final ANN model was a 34-3-3 network with 117 weights.

Size	Decay	Accuracy	Kappa
1	0e+00	0.3673375	0.002113003
1	1e-01	0.4500656	0.154374598
3	0e+00	0.3694779	0.005762061
3	1e-01	0.5015262	0.244361016
5	0e+00	0.3814896	0.028665161

Table 14. Resampling Results Across Tuning Parameters for Artificial Neural Network (ANN) Model.

We then gave the model the test dataset. We generated a confusion matrix and the resulting confusion matrix is shown in the table below. Overall, the model had an accuracy of 44.08%. The following figure is a map of the ANN model's classifications.

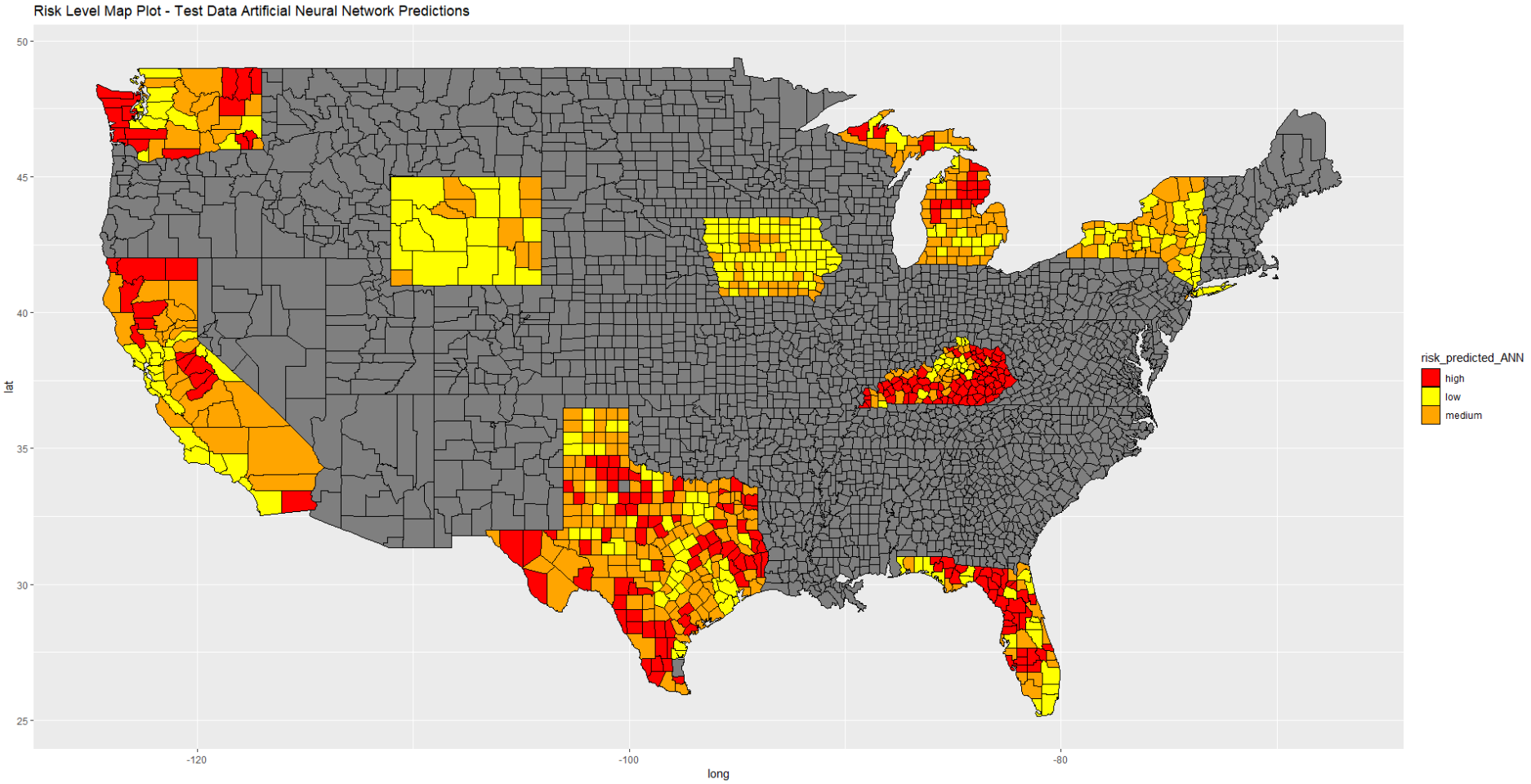

Figure 13. ANN Test Classifications.

	Actual		
Predicted	**High**	**Medium**	**Low**
High	100	56	73
Medium	134	81	109
Low	35	145	70

Table 15. Confusion Matrix for Artificial Neural Network Model on Test Dataset.

The 3x3 confusion matrix can tell us the number of correctly classified observations for each category. For example, our model correctly classified 100 counties as high risk, 81 as medium risk, and 70 as low risk. The confusion matrix can also tell us more about incorrect predictions. For example, our model misclassified 56 counties that are actually classified as high-risk. Similarly, our model misclassified 35 counties that should have been classified as low-risk.

Comparing All Three Models

In order to compare the 3 models, we first created a fixed sampling scheme with 10-folds to compare the models using the same folds. Then we compared the accuracy and kappa values over all folds. Below is the ground truth map that our models were graded on.

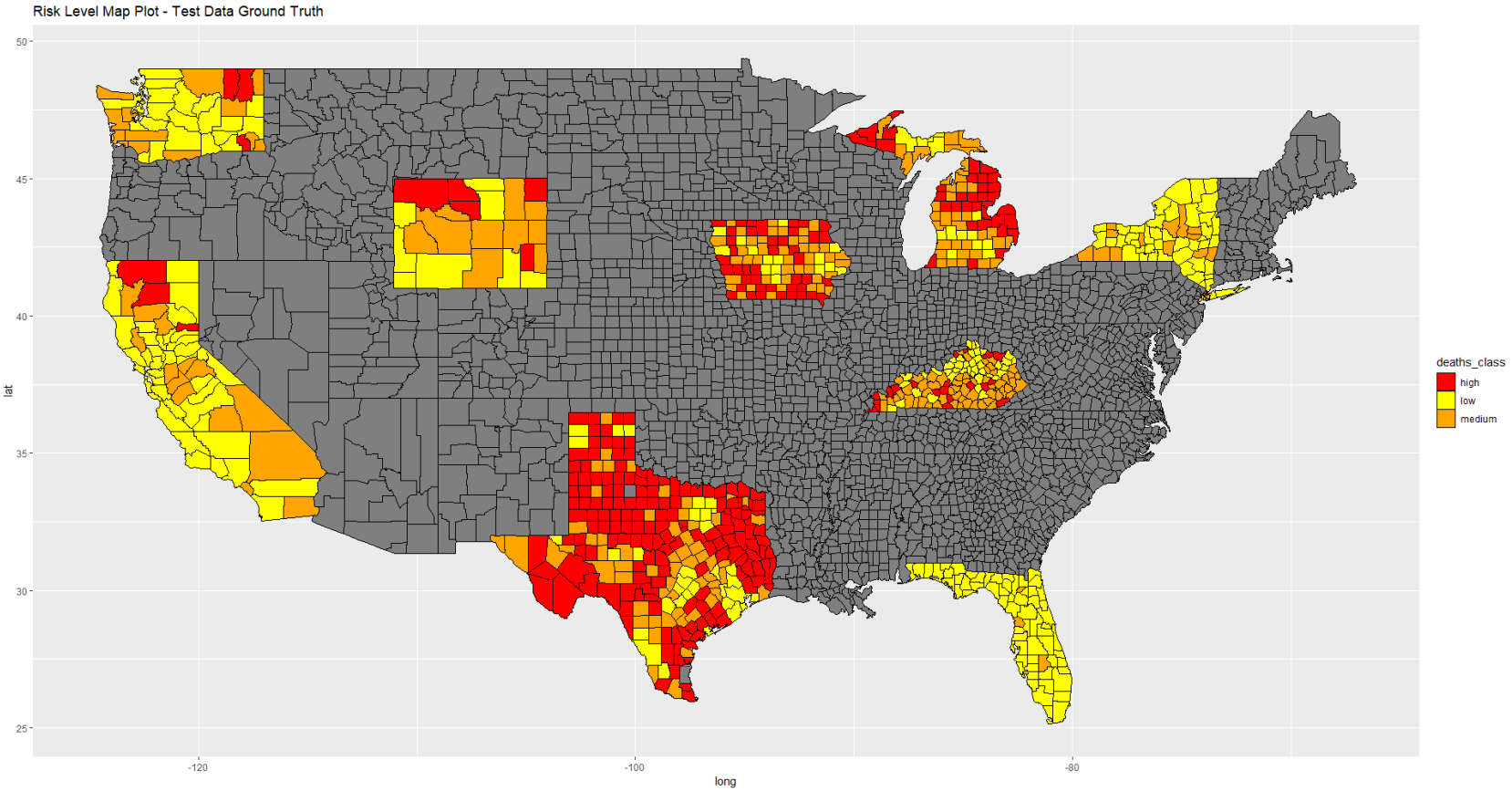

Figure 14. Ground Truth of the Map.

The tables below show the accuracy and kappa values for all 3 models.

Model Comparison - Accuracy

Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
K-Nearest Neighbors	0.9188	0.9399	0.9507	0.9502	0.9592	0.9828	0
Random Forest	0.5278	0.5476	0.5717	0.5701	0.5905	0.6094	0
Artificial Neural Network	0.3776	0.4667	0.5320	0.5015	0.5530	0.5708	0

Table 16. Statistical Summary of Each Model's Accuracy.

The accuracy of a model measures the proportion of correctly classified instances out of all instances in the dataset. You can calculate the accuracy by dividing the number of correct predictions by the number of total predictions. The accuracy simply tells us how often a model can correctly predict a county's risk level.

Our measurements show the KNN model consistently surpasses the other two models' accuracies across the board. There is a significant drop in accuracy between KNN's mean accuracy of 0.95 and the next best model, Random Forest's accuracy of 0.57. The Artificial Neural Network performed worse than both other models, with a maximum accuracy of 0.57 and an average of 0.50.

Model Comparison - Kappa

Model	Min	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
K-Nearest Neighbors	0.8779	0.9096	0.9259	0.9251	0.9387	0.9741	0
Random Forest	0.2902	0.3194	0.3563	0.3545	0.3859	0.4150	0
Artificial Neural Network	0.0214	0.2107	0.2983	0.2443	0.3311	0.3544	0

Table 17. Statistical Summary of Each Model's Kappa Values.

The Kappa value is a metric that takes into account the possibility of agreement occurring by chance. Kappa measures the agreement between the predicted labels and the actual labels, after adjusting for chance agreement. It is the proportion of agreement beyond what would be expected by chance. Kappa ranges from -1 to 1 when 1 indicates perfect agreement, 0 indicates chance agreement and a value less than 0 indicates poor agreement. Thus, while accuracy is

useful for evaluating the model's performance, it is important to also consider Kappa to ensure that the model's predictions are significantly better than random chance.

Our measurements in Table X show that the KNN model has significantly higher Kappa values across the board, with a mean of 0.9251. Meanwhile, the Random Forest and ANN models lag far behind with Kappa values entirely under 0.5. From these numbers, we can conclude that the KNN model outperformed the other two models. To further illustrate the comparisons of accuracy and kappa values for the three models, we generated a box and whisker plot.

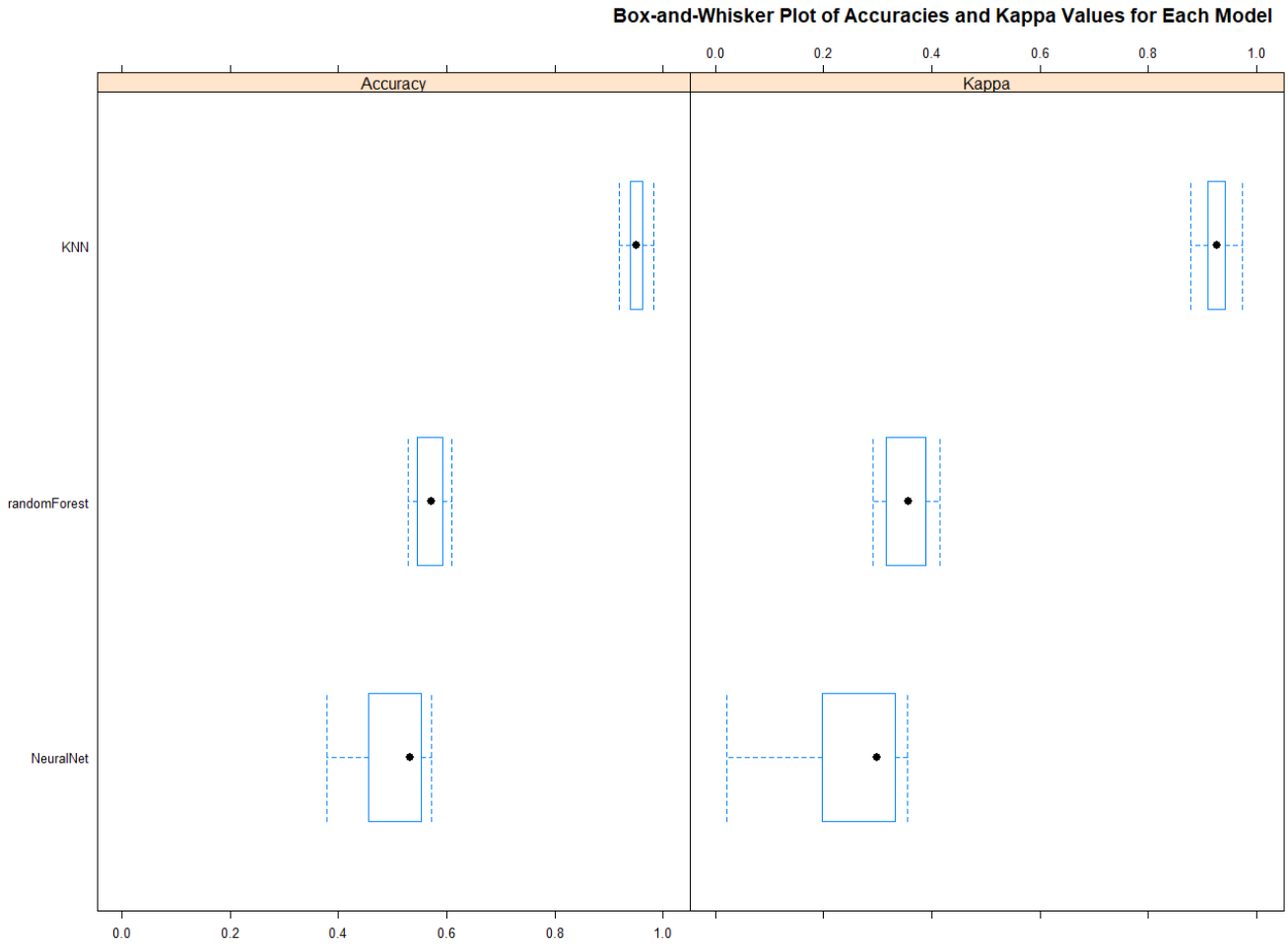

Figure 11. Accuracies and Kappa values.

As shown in Figure 11, the K-Nearest Neighbors consistently outperforms both the Random Forest model and the Artificial Neural Network model with 10-folds. We can securely conclude that the K-Nearest Neighbors model is the best model of the 3 models we generated. The Random Forest is marginally better than the Artificial Neural Network Model.

Evaluation

Training

To run any sort of classification algorithm, we must first train our models. The following graph visualizes the data that we trained on.

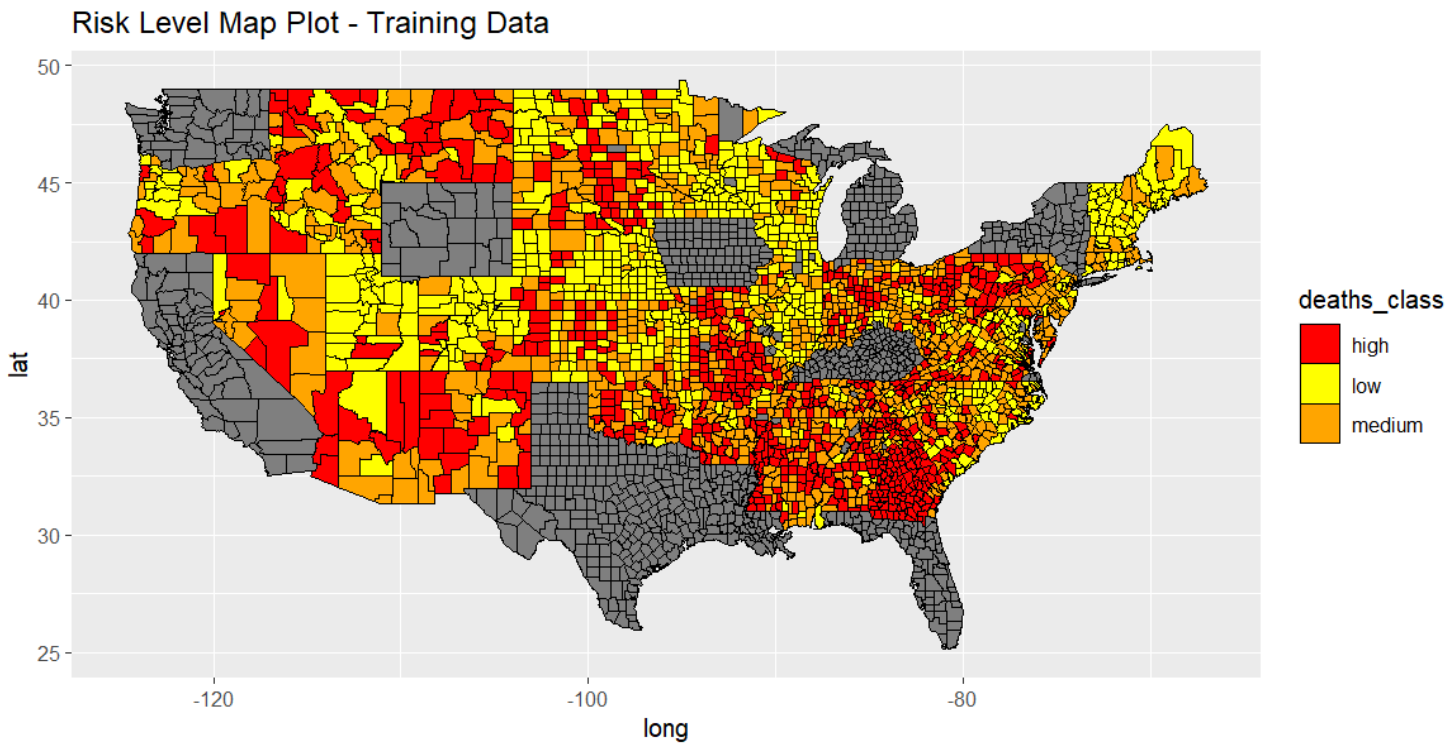

Figure 12. Training Data of Risk Level.

Our training data contains counties from across the United States, excluding a few states. The states we excluded were California, Florida, Iowa, Kentucky, Louisiana, Michigan, New York, Texas, Washington, and Wyoming. These states represent a small sliver of the United States, but a huge part of the population and gross domestic product. California, Florida, New York, and Texas may be 4 states, but they contain over a third of the population and contributed to approximately 34.5% of the United States Gross Domestic Product. On the other hand, Wyoming is the smallest state in the Union. The population is a little under 600,000, and the state contributes to approximately 0.4% of the United States GDP. This large discrepancy provides an interesting challenge for our analysis. The model must be able to accurately handle the largest and smallest states when using only the middle few states.

Usefulness And Value

The use of machine learning models such as K-Nearest Neighbors (KNN), Random Forest, and Artificial Neural Networks (ANN) can be invaluable for stakeholders like the Centers for Disease Control and Prevention (CDC) in predicting and managing future waves of COVID-19. These models can help the CDC identify high-risk, medium-risk, and low-risk areas for potential outbreaks, enabling them to allocate resources more effectively and implement targeted interventions to mitigate the spread of the virus.

The K-Nearest Neighbors model is a simple yet powerful classification algorithm that works by identifying the K most similar instances in the training dataset to a new, unclassified instance. By examining the risk levels of these neighboring instances, the model can assign a risk class to the new instance. KNN is particularly useful for its simplicity and ease of interpretation, which can help the CDC understand the relationships between various factors and the risk levels they contribute to. Moreover, KNN can be easily updated with new data, allowing the model to adapt to the evolving pandemic situation.

The Random Forest model is an ensemble learning method that constructs multiple decision trees and combines their predictions to determine the final risk class. Random Forest models are known for their accuracy and robustness, as well as their ability to handle large datasets with many variables. By capturing complex relationships and interactions among variables, Random Forest models can provide the CDC with a more comprehensive understanding of the factors driving the spread of COVID-19. Additionally, the model's ability to estimate feature importance can help the CDC identify key variables that contribute the most to the risk levels, informing their decision-making process and guiding resource allocation.

Artificial Neural Networks, inspired by the structure and function of biological neural networks, are capable of learning complex patterns and relationships in the data. ANNs are particularly well-suited for tasks like pattern recognition and classification, making them a valuable tool for predicting risk levels associated with future COVID-19 waves. With their ability to handle large datasets and adapt to new data, ANNs can provide the CDC with accurate and up-to-date predictions. Furthermore, ANNs can be fine-tuned using various hyperparameters, allowing the model to be tailored to the specific needs and challenges of the COVID-19 pandemic.

To ensure the effectiveness of these models, it is crucial to maintain balanced training and testing datasets, as unbalanced data can negatively impact model performance. By carefully selecting states for the training and testing datasets and verifying their balance, the models can be trained and assessed more accurately, ultimately leading to more reliable predictions.

By incorporating these machine learning models into their decision-making processes, the CDC can benefit from a data-driven approach to pandemic management. The models can help identify areas at higher risk for a potential 5th wave of COVID-19, allowing the CDC to prioritize resource allocation, plan vaccination campaigns, and implement targeted public health measures. Furthermore, the models' ability to adapt to new data ensures that the CDC remains informed about the evolving situation, enabling them to respond effectively to changes in the pandemic landscape.

Further Improvements

As with any model, the accuracy and performance of these algorithms depend heavily on the quality and diversity of the training data. In this context, we will discuss the challenges faced when using a training dataset predominantly composed of midsized states in the USA, lacking representation from larger states such as California, Texas, New York, and Florida. We will also explore potential improvements to achieve better performance in these models.

A critical aspect of training our machine learning models is ensuring that the training dataset is representative of the problem space, i.e., it covers diverse examples from various categories or classes. When our training data is mostly composed of midsized states, it may lead to biased models that perform poorly on larger, more diverse states. This lack of diversity can be attributed to the models' inability to generalize well on unseen data from these larger states, as they are not exposed to the unique characteristics and patterns present in these regions during the training phase. Consequently, the accuracies of KNN, Random Forest, and ANN models may be low when applied to data from California, Texas, New York, and Florida.

To address this issue, we can start by restructuring the training data to include more examples from these larger states. This would involve collecting additional data points from California, Texas, New York, and Florida, thereby ensuring a more balanced representation of the different regions within the USA. By doing so, our machine learning models can learn the unique features and patterns present in these states, allowing them to generalize better and improve their overall performance.

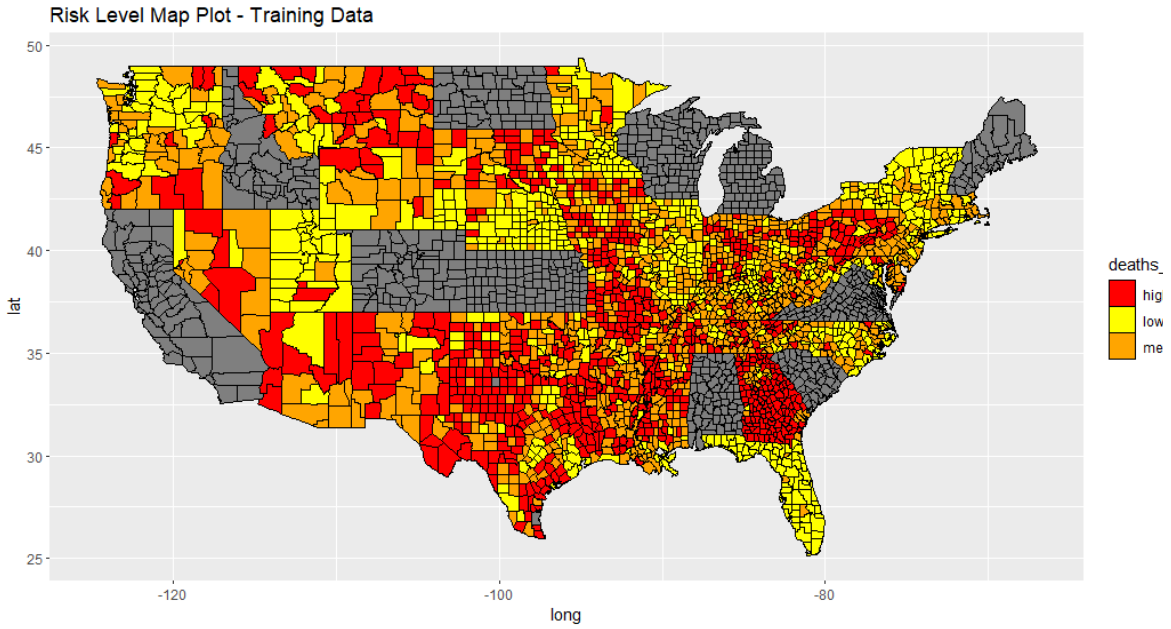

Figure 13. New Training Data.

The following maps show the new classifications when the K-nearest-neighbors, Random Forest, and Artificial Neural Networks are run on the new training data.

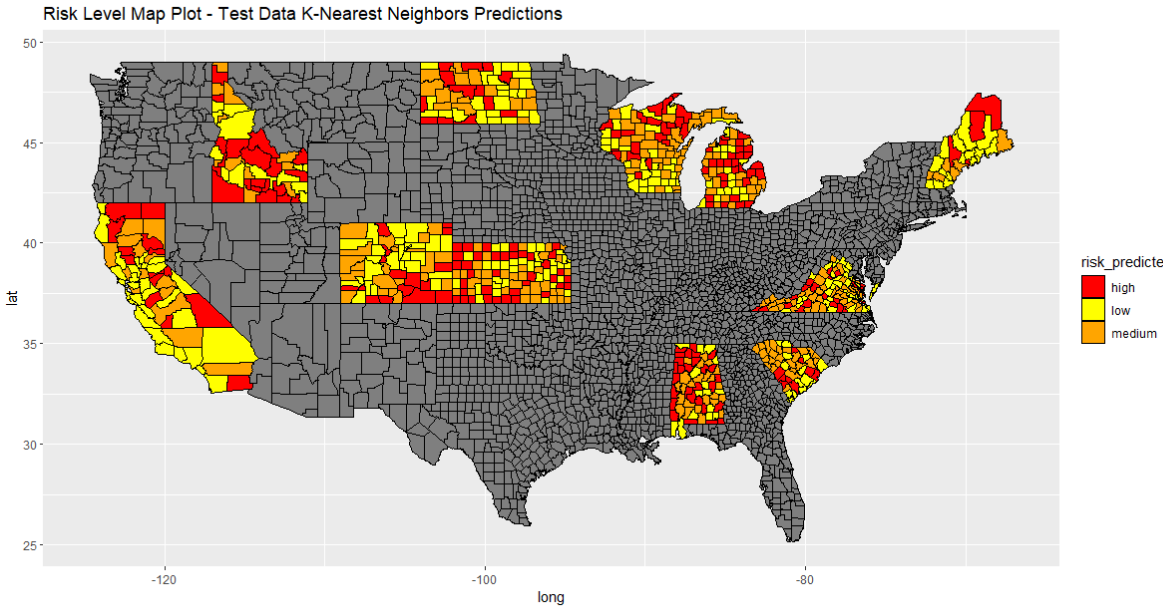

Figure 14. KNN Run on New Training Data.

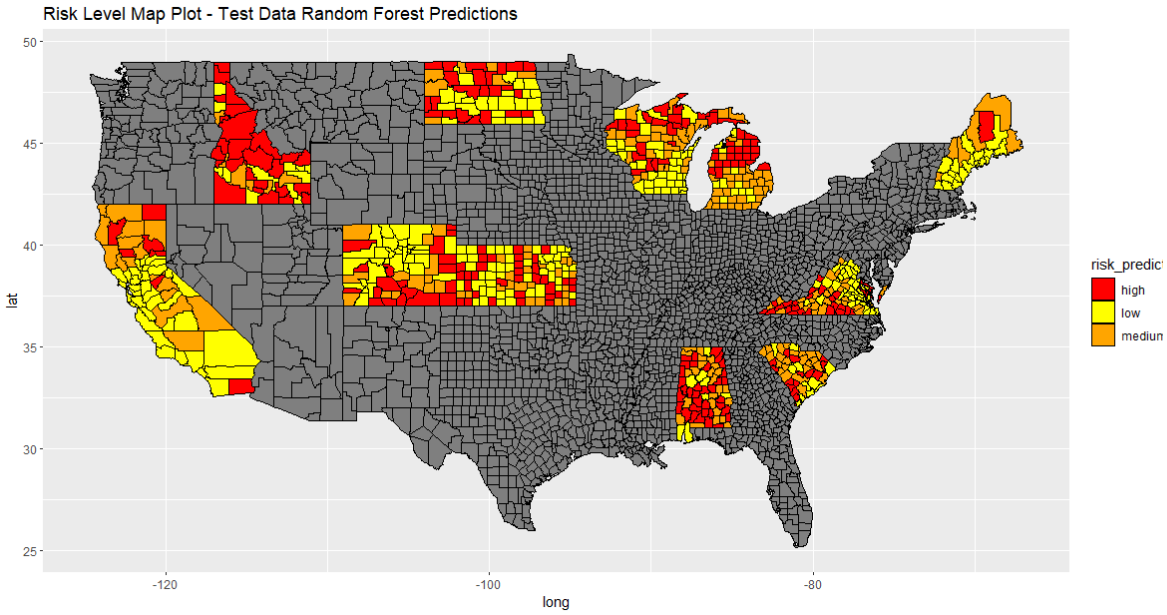

Figure 15. RF Run on New Training Data.

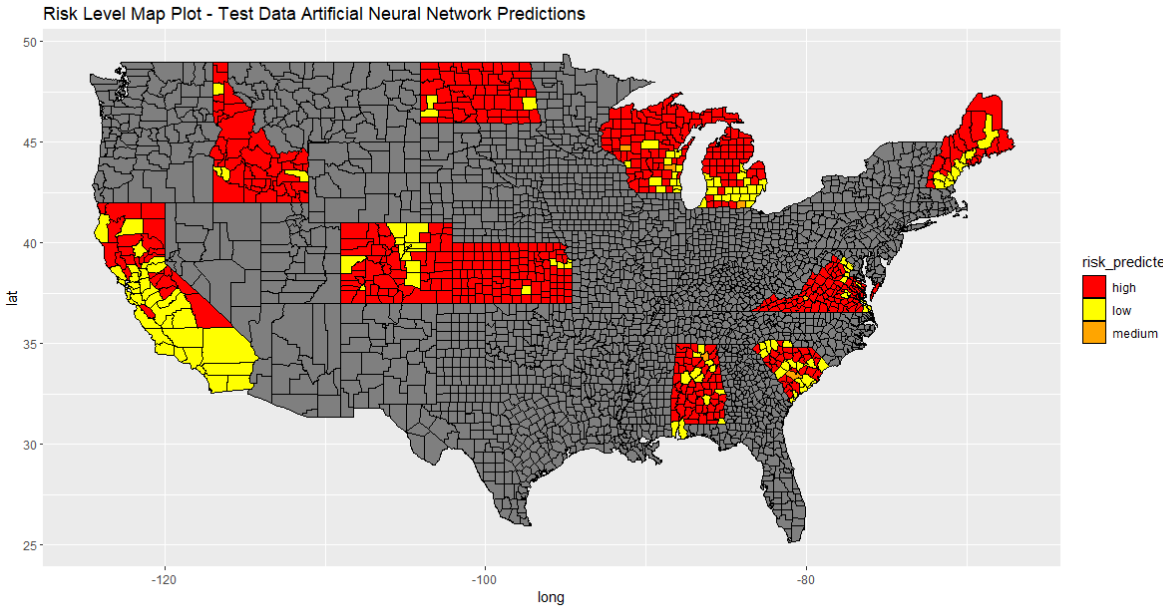

Figure 16. ANN Run on New Training Data.

Our new accuracies when run on this test set are as follows: 45.81% for KNN, 56.46% for Random Forest, and 37.68% for Artificial Neural Network. We can generally see a slight improvement when performing classification with this new data set. In addition to restructuring the training data, there are other potential improvements that can be made to the KNN, Random Forest, and ANN models. For the KNN model, we can explore different distance metrics to better capture the relationships between the data points. This may involve testing different distance measures, such as Euclidean, Manhattan, or Minkowski distances, to determine which one best captures the underlying structure of the data.

For Random Forest, we can fine-tune various hyperparameters, such as the number of trees, the maximum depth of the trees, and the minimum number of samples required to split a node. By optimizing these hyperparameters, we can reduce overfitting and improve the model's overall performance. Furthermore, we can also explore techniques like feature selection and feature engineering to create more informative input features for the model, potentially leading to better classification results.

Lastly, for ANN models, we can experiment with different architectures, such as increasing the number of layers or neurons in the hidden layers, to create a more expressive model capable of capturing complex patterns in the data. We can also investigate various optimization algorithms, activation functions, and regularization techniques to reduce overfitting and improve the model's generalization capabilities.

In conclusion, the performance of K nearest neighbors, Random Forest, and Artificial Neural Networks models can be significantly improved by restructuring the training data to include more examples from larger states like California, Texas, New York, and Florida. This ensures that our models are exposed to diverse data points, enabling them to generalize better on unseen data from these regions. Additionally, by fine-tuning various model-specific parameters and exploring different techniques, we can further enhance the performance of these machine-learning models, leading to more accurate and robust predictions.

Deployment

If the machine learning models were deployed, the CDC could use them to inform their decision-making process and develop targeted strategies to manage the COVID-19 pandemic more effectively. The models can serve multiple purposes, including identifying high-risk, medium-risk, and low-risk areas, predicting potential outbreaks, and guiding resource allocation. Here are some ways the CDC could utilize the models:

1. Resource allocation: By identifying areas at higher risk for future COVID-19 waves, the CDC can prioritize the distribution of resources such as vaccines, personal protective equipment (PPE), testing kits, and medical personnel. This targeted approach can help ensure that resources are directed to the areas that need them the most, ultimately contributing to a more efficient response.
2. Public health measures: The models can help the CDC determine which public health measures are most effective in controlling the spread of the virus in different regions. By analyzing the relationships between various factors and risk levels, the CDC can develop tailored strategies for social distancing, mask-wearing, and other preventive measures based on local risk profiles.
3. Vaccination campaigns: The models can inform the planning and execution of vaccination campaigns by identifying regions with higher vulnerability to COVID-19. Targeting these areas for vaccination efforts can help reduce the overall impact of the pandemic and protect vulnerable populations more effectively.
4. Surveillance and monitoring: The models can serve as an early warning system, allowing the CDC to monitor the risk levels in various regions and detect potential outbreaks before they become widespread. This can enable the CDC to act proactively and implement targeted interventions to curb the spread of the virus.

To maintain and update the models with the latest patterns, several steps can be taken:

1. Regular data updates: Ensure that the models are trained on the most recent data available by regularly updating the dataset with new cases, deaths, vaccinations, and other relevant information. This will help the models adapt to the evolving pandemic situation and maintain their accuracy.
2. Continuous model evaluation: Periodically assess the performance of the models by comparing their predictions with the actual outcomes. This can help identify any discrepancies or weaknesses in the models and inform potential improvements.
3. Model fine-tuning: Adjust the hyperparameters of the models, such as the number of trees in the Random Forest or the learning rate in the Artificial Neural Network, to optimize

their performance. This can be achieved through techniques like grid search or random search, which systematically explore different combinations of hyperparameters.

4. Feature engineering: As new information becomes available or new factors are identified as potentially relevant to the spread of COVID-19, consider incorporating these features into the models. This can help capture additional insights and improve the models' predictive capabilities.
5. Collaboration with experts: Collaborate with epidemiologists, public health experts, and data scientists to ensure that the models are up-to-date and informed by the latest scientific knowledge. This interdisciplinary approach can help ensure the models remain relevant and useful in the rapidly changing landscape of the pandemic.

By deploying these machine learning models and maintaining their accuracy through regular updates and evaluations, the CDC can harness the power of data-driven decision-making to navigate the complex challenges posed by the COVID-19 pandemic more effectively.

Conclusions

In conclusion, the COVID-19 pandemic has highlighted the crucial role of data-driven decision-making in managing public health crises. Throughout our discussions, we have emphasized the importance of thorough data preprocessing, analysis, and the application of machine learning models to inform the Centers for Disease Control and Prevention (CDC) and other stakeholders. The primary goal is to predict and manage future waves of COVID-19 effectively, allocate resources efficiently, and implement targeted interventions to protect public health and mitigate the virus's spread.

We began by examining the need for data cleaning, standardization, and normalization, specifically focusing on county names and handling missing or unreliable data. By employing R code for these tasks, we ensured the integrity and reliability of the dataset used for analysis. Furthermore, we discussed the significance of correlation matrices in identifying relationships between various demographic, socioeconomic, and COVID-19-related variables. Normalization was shown to be a crucial step in ensuring accurate comparisons and avoiding misleading correlations due to different scales or magnitudes.

To create a predictive model for future COVID-19 waves, we classified counties into "high risk," "medium risk," and "low risk" categories based on their normalized deaths per case values. We then verified the balance of our classes, ensuring the dataset was suitable for training machine learning models. After splitting the dataset into training and testing subsets, we selected a diverse range of states to be included in each subset to increase model accuracy.

We then explored the application of three machine learning models: K-Nearest Neighbors (KNN), Random Forest, and Artificial Neural Networks (ANN). The KNN model was noted for its simplicity and ease of interpretation, making it a useful tool for understanding relationships between factors and risk levels. The Random Forest model, known for its accuracy and robustness, can capture complex relationships and interactions among variables, providing a more comprehensive understanding of the factors driving the pandemic. Finally, the ANN model, with its ability to learn complex patterns and adapt to new data, can provide accurate and up-to-date predictions tailored to the specific needs and challenges of the COVID-19 pandemic.

By incorporating these machine learning models into the decision-making process, the CDC and other stakeholders can benefit from a data-driven approach to pandemic management. Identifying areas at higher risk for future COVID-19 waves allows for prioritizing resource allocation, planning vaccination campaigns, and implementing targeted public health measures. Moreover, these models' adaptability to new data ensures that the CDC remains informed about

the evolving situation, enabling them to respond effectively to changes in the pandemic landscape.

In summary, the effective management of the COVID-19 pandemic requires a deep understanding of the relationships between various factors and their impact on the risk levels of future waves. By leveraging data preprocessing, analysis, and machine learning models, stakeholders like the CDC can make informed decisions and implement targeted interventions. This data-driven approach is essential in navigating the complex challenges posed by the pandemic, ultimately contributing to the protection of public health and the well-being of individuals around the world.

References

[1] “Who coronavirus (COVID-19) dashboard,” *World Health Organization*. [Online]. Available: https://covid19.who.int/. [Accessed: 28-Feb-2023].

[2] “Guidance and tips for tribal community living during COVID-19,” *Centers for Disease Control and Prevention*, 23-Aug-2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/community/tribal/social-distancing.html. [Accessed: 28-Feb-2023].

[3] By, “Coronavirus in the U.S.: Latest Map and case count,” *The New York Times*, 03-Mar-2020. [Online]. Available: https://www.nytimes.com/interactive/2021/us/covid-cases.html. [Accessed: 28-Feb-2023].

[4] J. H. Cullum Clark, “The Texas Triangle: A rising megaregion unlike all others,” *George W. Bush Presidential Center*, 19-May-2021. [Online]. Available: https://www.bushcenter.org/publications/the-texas-triangle-a-rising-megaregion-unlike-all-others. [Accessed: 28-Feb-2023].

[5] “Texas triangle,” *Austin Capital Advisors*. [Online]. Available: https://www.austincapitaladvisors.com/texas-triangle. [Accessed: 28-Feb-2023].