# Does Population Wealth Indicate COVID-19 Vulnerability?

# Anita Adams

# Randy Geszvain

# Michael O’Keefe

# IST 707

# Introduction

As of June 12th, 2020, the United States has the most confirmed coronavirus cases and the most deaths of any country in the world. Brazil is a distant second in confirmed cases, while the UK remains second in deaths. The infection is still spreading despite unprecedented measures including mandatory statewide quarantines, social distancing, required use of facial protection, and drastic travel restrictions. The disease is still not well understood due to its novelty and sudden appearance, and consequently, an effective, safe treatment has yet to be widely implemented.

While the scientific community races to develop an effective and safe vaccine, government officials are developing a whole-of-government approach to COVID-19 infection prevention and management. Key to such an approach is an understanding of where COVID-19 is most likely to spread, most likely to spread the fastest and the risk factors associated with an increase in mortality. At the federal level, having this understanding would allow for better allocation of nationally controlled resources such as testing kits, emergency reserve and newly manufactured ventilators, and military medical augmentation. At the state level, medical assets such as personnel (doctors, nurses, and respirator technical staff), medical equipment, beds, supplies, and medicine could be allocated proactively to predicted hotspots across counties and even across state lines where counties with similar demographics border one another. Local governments could better prepare, warn citizens, and tighten recommended or required preventative measures at both the state and county levels.

One hypothesis, among many, is that population demographics related to wealth might provide insight into COVID-19 spread. The government agency responsible for the whole-of-government approach has collected data by state and county regarding population wealth and COVID cases and deaths for two points in time during the pandemic. The agency is interested to know if population wealth is an indicator of per capita COVID cases and deaths, death rate from confirmed cases, rate of change of these factors, and spread to adjacent principalities. This analysis is focused on one of the questions: is population wealth an indicator of death rate from confirmed COVID cases.

# Analysis and Models

## About the Data

The data is provided by kaggle.[1] The data is merged According to the kaggle author “James”, the data includes “County level indicators for the general population including race, poverty level, housing size, sources of income, employment status, whether living alone, language barriers, immigration status, and disability status, modes of transportation stats, and industry stats.” The data was created by merging attributes from the following datasets:

* Johns Hopkins University (JHU): 4/16/2020 and 4/22/2020 daily spread estimates from the JHU Coronavirus Resource Center
* Center for Disease Control (CDC): Population density from the CDC Control and Prevention’s Social Vulnerability Index
* American Community Survey (ACS): population information and housing data from the US Census Bureaus’ 2018 ACS

The original data is provided in .csv format and has 1476 observations of 133 variables. Each observation represents a county and state in the United States with non-zero confirmed cases of coronavirus on 4/16/2020 or 4/22/2020

[1]<https://www.kaggle.com/jtourkis/us-county-level-acs-features-for-covid-analysis#ACS_full_data_wo_over_60.csv>

## Models

### EDA Models

#### EDA Model 1: Dimensionality of the Data

Figure 1 graphically depicts the dimensionality of the original data.

#### 

Figure Data Dimensionality

#### EDA Model 2: Missing Data / NAs

Volume of missing data and which variables have the most missing data is presented in Figure 2. The total number of NAs is shown in Figure 3.

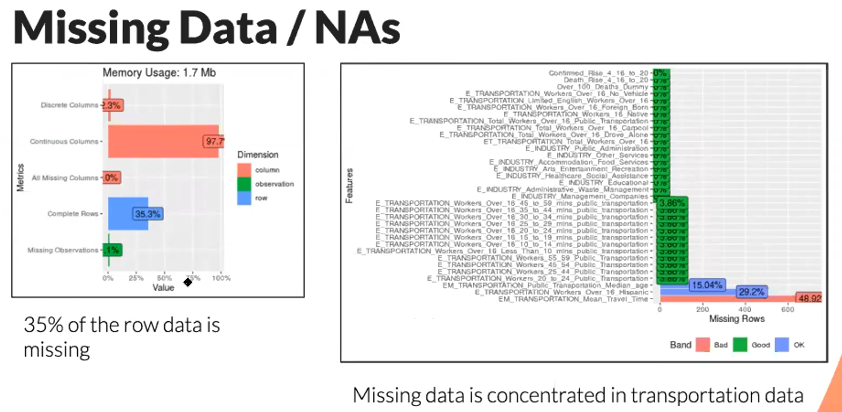


Figure Missing Data / NAs



Figure NAs

#### EDA Model 3: Reduce Dimensionality Based on Variable Applicability to Analytic Question

All variables in the original data set are presented in Figure 3. Variables in green and black are retained for analysis and the variables in red are removed. Figure 4 summarizes the types of data eliminated and why. Figure 4 summarizes the types of data retained for analysis.

| **Column** | **Kaggle Defn** |
| --- | --- |
| ID | to Match Data to County |
|  |  |
| STATE | STATE |
| COUNTY | COUNTY |
| Pop\_Density | Population Density |
| 4\_22\_Confirmed | John Hopkins Total Confirmed Confirmed as of 4/22 |
| 4\_22\_Deaths | John Hopkins Total Confirmed Deaths as of 4/22 |
| 4\_16\_Confirmed | John Hopkins Total Confirmed Confirmed as of 4/16 |
| 4\_16\_Deaths | John Hopkins Total Confirmed Deaths as of 4/16 |
| ET\_Total\_Population | Estimated Total Population |
| EM\_Total\_Pop\_Median\_Age | Estimated Median Age |
|  |  |
| **E\_Total\_Pop\_SEX\_Male** | **Estimated Total Males** |
| **E\_Total\_Pop\_SEX\_Female** | **Estimated Total Females** |
|  |  |
| **E\_Total\_Pop\_RACE\_One\_Race** | **Estimated Total One Race** |
| **E\_Total\_Pop\_RACE\_White** | **Estimated Total White** |
| **E\_Total\_Pop\_RACE\_Black** | **Estimated Total Black** |
| **E\_Total\_Pop\_RACE\_Native\_Pop** | **Estimated Total Native Population** |
| **E\_Total\_Pop\_RACE\_Asian** | **Estimated Total Asian** |
| **E\_Total\_Pop\_RACE\_Pacific\_Islander** | **Estimated Total Pacific Islander** |
| **E\_Total\_Pop\_RACE\_Other\_Race** | **Estimated Total Other Race** |
| **E\_Total\_Pop\_RACE\_Two\_or\_More\_Races** | **Estimated Total Two or More Races** |
| **E\_Total\_Pop\_RACE\_Hispanic** | **Estimated Total Hispanic** |
| **E\_Total\_Pop\_RACE\_White\_Alone** | **Estimated Total White Alone** |
|  |  |
| **E\_Total\_Pop\_in\_Households\_Householder\_Spouse** | **Estimated Total Householder Spouse** |
| **E\_Total\_Pop\_in\_Households\_Householder\_Parent** | **Estimated Total Householder Parent** |
| **E\_Total\_Pop\_in\_Households\_Householder\_Other\_Relative** | **Estimated Total Householder Other Relative** |
| **E\_Total\_Pop\_in\_Households\_Householder\_Nonrelative** | **Estimated Total Householder Nonrelative** |
| **E\_Total\_Pop\_in\_Households\_Householder\_Unmarried\_Partner** | **Estimated Total Householder Unmarried Partner** |
|  |  |
| **E\_Total\_Households\_TYPE\_Family** | **Estimated Total Household Type Family** |
| **E\_Total\_Households\_TYPE\_Family\_Married\_Couple** | **Estimated Total Household Type Family Married Couple** |
| **E\_Total\_Households\_TYPE\_Family\_Female\_Householder** | **Estimated Total Household Type Family Female Householder** |
| **E\_Total\_Households\_TYPE\_Nonfamily** | **Estimated Total Household Type Nonfamily** |
| **E\_Total\_Households\_TYPE\_Nonfamily\_Householder\_Living\_Alone** | **Estimated Total Household Type Nonfamily Lives Alone** |
|  |  |
| **E\_Total\_Pop\_Over\_15\_MARITAL\_STATUS\_Married** | **Estimated Total Married** |
| **E\_Total\_Pop\_Over\_15\_MARITAL\_STATUS\_Widowed** | **Estimated Total Widowed** |
| **E\_Total\_Pop\_Over\_15\_MARITAL\_STATUS\_Divorced** | **Estimated Total Divorced** |
| **E\_Total\_Pop\_Over\_15\_MARITAL\_STATUS\_Separated** | **Estimated Total Separated** |
| **E\_Total\_Pop\_Over\_15\_MARITAL\_STATUS\_Never\_Married** | **Estimated Total Never Married** |
|  |  |
| **E\_Total\_Pop\_Over\_15\_EDUCATION\_Less\_Than\_High\_School** | **Estimated Total Less Than High School** |
| **E\_Total\_Pop\_Over\_15\_EDUCATION\_High\_School\_Grad** | **Estimated Total High School Grad** |
| **E\_Total\_Pop\_Over\_15\_EDUCATION\_Some\_College** | **Estimated Total Some College or Associates** |
| **E\_Total\_Pop\_Over\_15\_EDUCATION\_Bachelors\_Degree\_or\_Higher** | **Estimated Total College Grad or Higher** |
|  |  |
| **E\_Total\_Pop\_Over\_30\_RESPONSIBLE\_FOR\_GRANDCHILDREN\_Live\_Together** | **Estimated Total Live With Grandchildren** |
| **E\_Total\_Pop\_Over\_30\_RESPONSIBLE\_FOR\_GRANDCHILDREN\_Live\_Together\_and\_Responsible** | **Estimated Total Live With Grandchildren and Responsible for Grandchildren** |
| **E\_Total\_Pop\_Over\_18\_VETERAN\_STATUS\_Civilian\_Veteran** | **Estimated Total Veteran** |
|  |  |
| **E\_Total\_Pop\_DISABILITY\_STATUS\_Yes** | **Estimated Total Disabled** |
| **E\_Total\_Pop\_DISABILITY\_STATUS\_No** | **Estimated Total Not Disabled** |
|  |  |
| **E\_Total\_Pop\_Over\_1\_Place\_of\_Residence\_Last\_Year\_Same\_House** | **Estimated Total Lived in Same House Last Year** |
| **E\_Total\_Pop\_Over\_1\_Place\_of\_Residence\_Last\_Year\_Different\_US\_House** | **Estimated Total Lived in Different US Home Last Year** |
| **E\_Total\_Pop\_Over\_1\_Place\_of\_Residence\_Last\_Year\_Different\_Foreign\_House** | **Estimated Total Lived in Different Foreign Home Last Year** |
|  |  |
| **ET\_Total\_Pop\_CITIZENSHIP\_Native** | **Estimated Total Native Born** |
| **ET\_Total\_Pop\_CITIZENSHIP\_Foreign\_Born** | **Estimated Total Foreign Born** |
| **E\_Total\_Pop\_CITIZENSHIP\_Foreign\_Born\_Post\_2010\_Entry** | **Estimated Total Foreign Born Entered After 2010** |
| **E\_Total\_Pop\_CITIZENSHIP\_Foreign\_Born\_Btwn\_2000\_2009\_Entry** | **Estimated Total Foreign Born Entered Btwn 2000 & 2009** |
| **E\_Total\_Pop\_CITIZENSHIP\_Foreign\_Born\_Before\_2000\_Entry** | **Estimated Total Foreign Born Entered Before 2000** |
| **E\_Total\_Pop\_CITIZENSHIP\_Foreign\_Born\_US\_Citizen** | **Estimated Total Foreign Born US Citizens** |
| **E\_Total\_Pop\_CITIZENSHIP\_Foreign\_Born\_Not\_US\_Citizen** | **Estimated Total Foreign Born Not US Citizens** |
|  |  |
| **E\_Total\_Pop\_Over\_5\_LANGUAGE\_English\_Only\_At\_Home** | **Estimated Total English Only At Home** |
| **E\_Total\_Pop\_Over\_5\_LANGUAGE\_Nonenglish\_At\_Home** | **Estimated Total English Nonenglish At Home** |
| **E\_Total\_Pop\_Over\_5\_LANGUAGE\_Limited\_English** | **Estimated Total Limited English** |
|  |  |
| **ET\_Total\_Pop\_Over\_16\_EMPLOYMENT\_STATUS\_In\_Labor\_Force** | **Estimated Total in Labor Force** |
| **E\_Total\_Pop\_Over\_16\_EMPLOYMENT\_STATUS\_In\_Civilian\_Labor\_Force** | **Estimated Total in Civilian Labor Force** |
| **E\_Total\_Pop\_Over\_16\_EMPLOYMENT\_STATUS\_In\_Civilian\_Labor\_Force\_Employed** | **Estimated Total Employed** |
| **E\_Total\_Pop\_Over\_16\_EMPLOYMENT\_STATUS\_In\_Civilian\_Labor\_Force\_Unemployed** | **Estimated Total Unemployed** |
| **E\_Total\_Pop\_Over\_16\_EMPLOYMENT\_STATUS\_In\_Civilian\_Labor\_Force\_Unemployed\_Percent\_of\_Labor\_Force** | **Estimated Total Workforce Members Unemployed** |
| **E\_Total\_Pop\_Over\_16\_EMPLOYMENT\_STATUS\_In\_Armed\_Forces** | **Estimated Total in Armed Forces** |
| **E\_Total\_Pop\_Over\_16\_EMPLOYMENT\_STATUS\_Not\_In\_Labor\_Force** | **Estimated Total Not in Labor Force** |
|  |  |
| **EM\_Total\_Households\_WITH\_INCOME\_Mean\_Earnings** | **Mean Household Income in Last 12 Months** |
| **EM\_Total\_Households\_WITH\_INCOME\_Mean\_Social\_Security** | **Mean Household Social Security in Dollars** |
| **EM\_Total\_Households\_WITH\_INCOME\_Mean\_Suppl\_Security** | **Mean Household Supplemental Security in Dollars** |
| **EM\_Total\_Households\_WITH\_INCOME\_Mean\_Retirement\_Income** | **Mean Household Retirement Income in Dollars** |
| **EM\_Total\_Households\_WITH\_INCOME\_Mean\_Cash\_Public\_Asst** | **Mean Household Public Assistance in Dollars** |
| **E\_Total\_Households\_WITH\_INCOME\_WITH\_EARNINGS** | **Estimated Households with Income in Last 12 Months** |
| **E\_Total\_Households\_WITH\_INCOME\_WITH\_Social\_Security** | **Estimated Households with with Social Security** |
| **E\_Total\_Households\_WITH\_INCOME\_WITH\_Suppl\_Security** | **Estimated Households with Supplemental Security Income** |
| **E\_Total\_Households\_WITH\_INCOME\_WITH\_Cash\_Public\_Asst** | **Estimated Households with Public Assistance Income** |
| **E\_Total\_Households\_WITH\_INCOME\_WITH\_Retirement\_Income** | **Estimated Households with Retirement Income** |
| **E\_Total\_Households\_WITH\_INCOME\_WITH\_SNAP** | **Estimated Households with Food Stamps/SNAP** |
|  |  |
| **E\_Total\_Pop\_POVERTY\_STATUS\_Below\_100\_Percent** | **Estimated Total Below 100 Percent Poverty Level** |
| **E\_Total\_Pop\_POVERTY\_STATUS\_Btwn\_100\_149\_Percent** | **Estimated Total Between 100 to 149 Percent Poverty Level** |
| **E\_Total\_Pop\_POVERTY\_STATUS\_Above\_150\_Percent** | **Estimated Total Above 150 Percent** |
|  |  |
| **E\_Total\_HOUSING\_UNITS\_Owner\_Occupied** | **Estimated Total Owned Units** |
| **E\_Total\_HOUSING\_UNITS\_Rented** | **Estimated Total Rented Units** |
| **E\_Total\_HOUSING\_UNITS\_No\_Telephone** | **Estimated Units with No Phone** |
| **E\_Total\_HOUSING\_UNITS\_More\_Than\_One\_Occupant\_Per\_Room** | **Estimated Units with Average Occupants Per Room Over 1** |
| **EM\_Total\_HOUSING\_UNITS\_Avg\_Household\_Size\_Rented** | **Average Household Size For Rented Units** |
| **EM\_Total\_HOUSING\_UNITS\_Avg\_Household\_Size\_Owned** | **Average Household Size For Owned Units** |
|  |  |
| **ET\_INDUSTRY\_Total\_Civilian\_Employed\_Population\_Over\_16** | **Estimated Total Population Working Over 16** |
| **E\_INDUSTRY\_Agriculture\_Forestry\_Fishing\_and\_Hunting** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Mining\_Quarrying\_and\_Oil\_and\_Gas\_Extraction** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Construction** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Manufacturing** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Wholesale\_Trade** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Retail\_Trade** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Transportation\_and\_Warehousing** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Utilities** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Information** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Finance\_an\_Insurance** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_REAL\_ESTATE** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Professional\_Scientific\_Technical** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Management\_Companies** | **Estimated Total Population in Specified Industry** |
| **E\_INDUSTRY\_Administrative\_Waste\_Management** |  |
| **E\_INDUSTRY\_Educational** |  |
| **E\_INDUSTRY\_Healthcare\_Social\_Assistance** |  |
| **E\_INDUSTRY\_Arts\_Entertainment\_Recreation** |  |
| **E\_INDUSTRY\_Accommodation\_Food\_Services** |  |
| **E\_INDUSTRY\_Other\_Services** |  |
| **E\_INDUSTRY\_Public\_Administration** |  |
|  |  |
| **ET\_TRANSPORTATION\_Total\_Workers\_Over\_16** |  |
| **E\_TRANSPORTATION\_Total\_Workers\_Over\_16\_Drove\_Alone** |  |
| **E\_TRANSPORTATION\_Total\_Workers\_Over\_16\_Carpool** |  |
| **E\_TRANSPORTATION\_Total\_Workers\_Over\_16\_Public\_Transportation** |  |
| **E\_TRANSPORTATION\_Workers\_20\_to\_24\_Public\_Transportation** |  |
| **E\_TRANSPORTATION\_Workers\_25\_44\_Public\_Transportation** |  |
| **E\_TRANSPORTATION\_Workers\_45\_54\_Public\_Transportation** |  |
| **E\_TRANSPORTATION\_Workers\_55\_59\_Public\_Transportation** |  |
| **EM\_TRANSPORTATION\_Public\_Transportation\_Median\_age** |  |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_Hispanic** |  |
| **E\_TRANSPORTATION\_Workers\_16\_Native** |  |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_Foreign\_Born** |  |
| **E\_TRANSPORTATION\_Limited\_English\_Workers\_Over\_16** |  |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_Less\_Than\_10\_mins\_public\_transportation** | |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_10\_to\_14\_mins\_public\_transportation** | |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_15\_to\_19\_mins\_public\_transportation** | |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_20\_to\_24\_mins\_public\_transportation** | |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_25\_to\_29\_mins\_public\_transportation** | |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_30\_to\_34\_mins\_public\_transportation** | |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_35\_to\_44\_mins\_public\_transportation** | |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_45\_to\_59\_mins\_public\_transportation** | |
| **EM\_TRANSPORTATION\_Mean\_Travel\_Time** |  |
| **E\_TRANSPORTATION\_Workers\_Over\_16\_No\_Vehicle** |  |

Figure Original Variables Retained and Eliminated

|  |  |
| --- | --- |
| Variable Category Removed | Rationale |
| Industry | Not fine grained enough to reflect essential services |
| Housing Units | Correlated with Poverty Status which is retained |
| Income Source | Correlated with Poverty Status which is retained |
| Household Composition | Not an indicator of wealth |
| Marital Status | Not an indicator of wealth |
| Grandchildren in Home | Not an indicator of wealth |
| Veteran Status | Not an indicator of wealth |
| Moved Last Year | Not an indicator of wealth |
| Transportation | Too much missing data |

Figure Categories of Variables Removed and Rationale

|  |
| --- |
| Variable Category Retained |
| State and County Identifiers |
| COVID Confirmed Cases and Deaths |
| Population Counts and Ages |
| Race Data |
| Educational Data |
| Disability Status |
| Language Spoken |
| Employment Status |
| Poverty Status |

Figure Categories of Variables Retained for Analysis

#### EDA Model 4: Transform the Data

A new variable was created (Deaths\_Per\_Confirmed\_4\_16 = Deaths\_Per\_Capita\_4\_16 / Confirmed\_Per\_Capita\_4\_16).

The remaining variables were consolidated. For example, instead of including all poverty variables (Total\_Pop\_POVERTY\_STATUS\_Below\_100\_Percent, Total\_Pop\_POVERTY\_STATUS\_Btwn\_100\_149\_Percent and Total\_Pop\_POVERTY\_STATUS\_Above\_150\_Percent), only the variable representing the most impoverished (Total\_Pop\_POVERTY\_STATUS\_Below\_100\_Percent) was retained.

The remaining variables were transformed to be per county capita as depicted in Figure 6. The numeric variables were then normalized.



Figure Variable Transformation

#### EDA Model 5: Resultant Data

The cleaned data has all complete rows as shown in Figure 8. The new data structure is visualized in Figure 9.

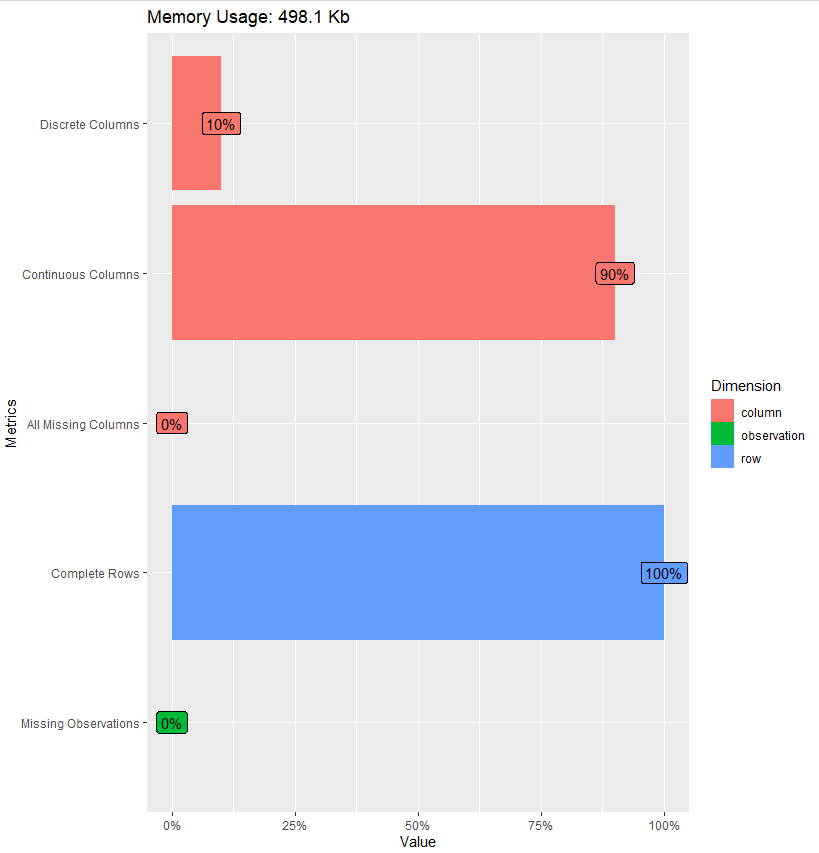


Figure Cleaned Data

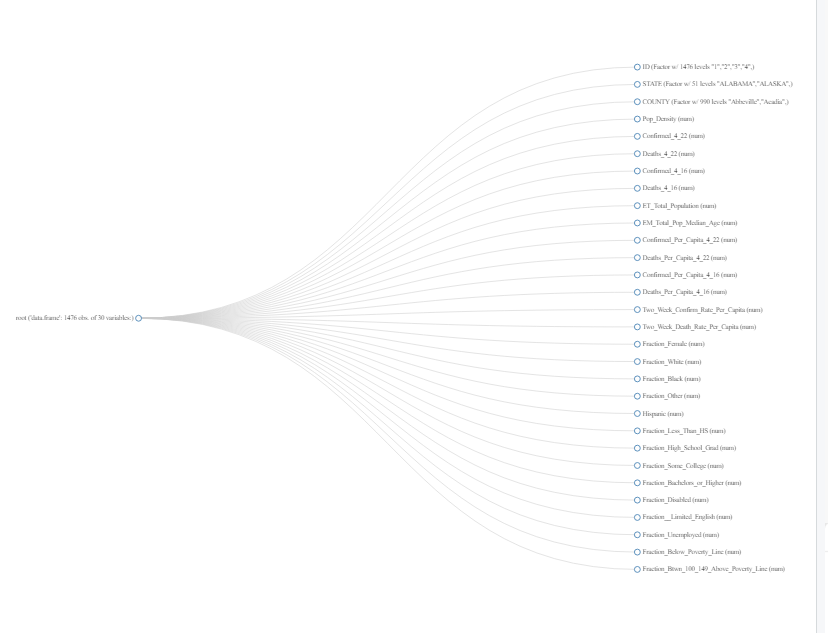


Figure New Data Dimensionality

#### EDA Model 6: Outliers

Figure 10 depicts deaths per confirmed cases vs confirmed cases. Figure 11 depicts deaths per confirmed cases vs confirmed cases with outliers removed.

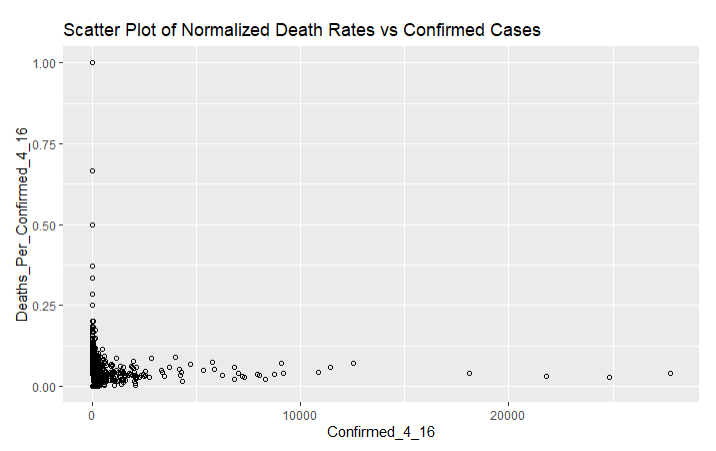


Figure Deaths Per Confirmed Cases vs Confirmed Cases

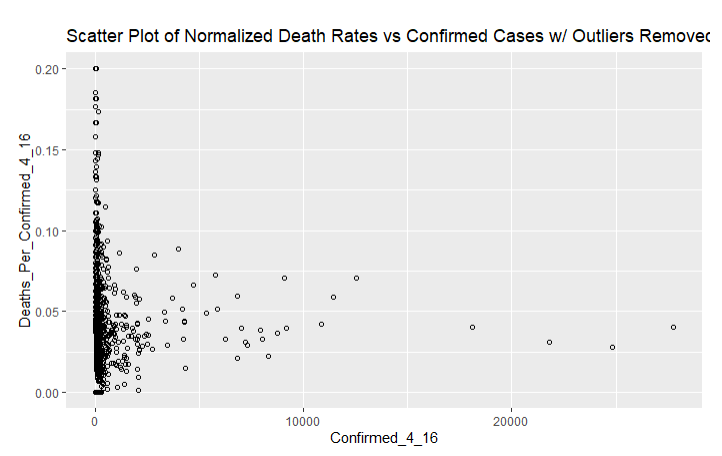


Figure Deaths Per Confirmed Cases vs Confirmed Cases - Outliers Removed

#### EDA Model 7: Discretization

Figure 12 depicts Deaths\_Per\_Confirmed\_4\_16 binned into 10 equal bins. Figure 13 depicts Deaths\_Per\_Confirmed\_4\_16 binned into 200 equal bins. Figure 14 depicts Deaths\_Per\_Confirmed\_4\_16 binned into 11 custom bins.

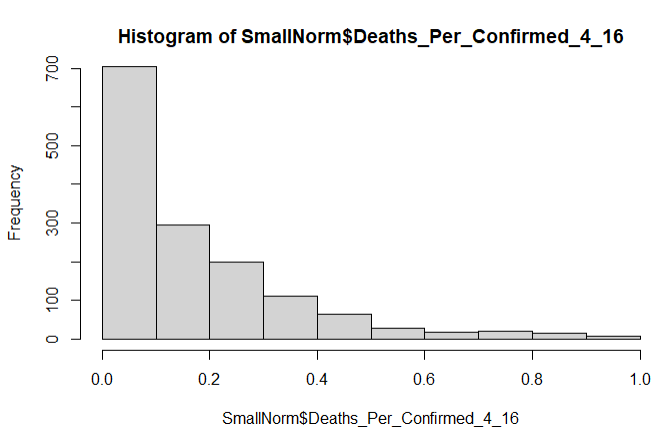


Figure Deaths Per Confirmed Discretized Using 10 Equal Bins

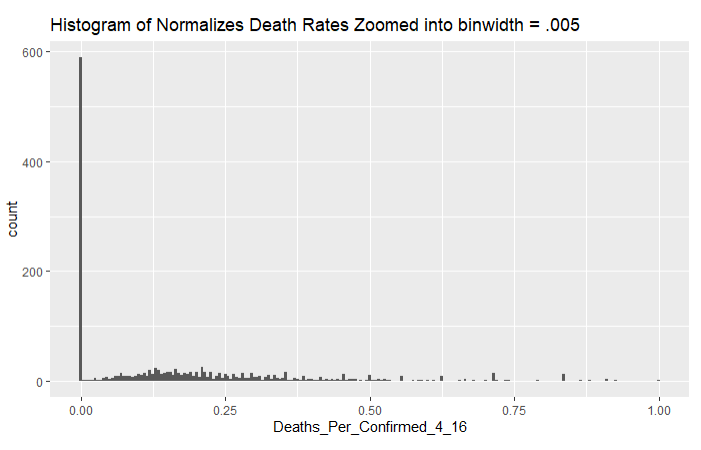


Figure Deaths Per Confirmed Discretized Using 200 Equal Bins

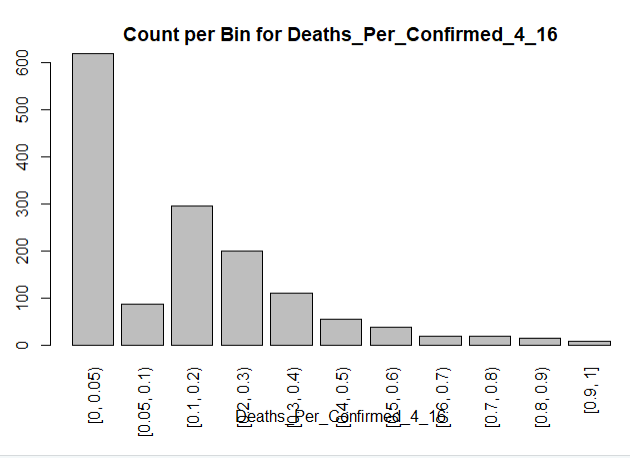


Figure Deaths Per Confirmed Discretized Using 11 Custom Bin

### Classification Models

#### Classification Model 1: Classification Results with Base Data Set

The base data set was used. The following algorithms from function “train” in package “caret” were applied and tuned using k-fold validation:

* Naive Bayesian
* K Nearest Neighbor (KNN)
* Random Forest
* Support Vector Machine (kernel - Linear)
* Support Vector Machine (kernel - Radial Basis Function)

The actual test data and the model results to predict the test data are shown in Figure 15

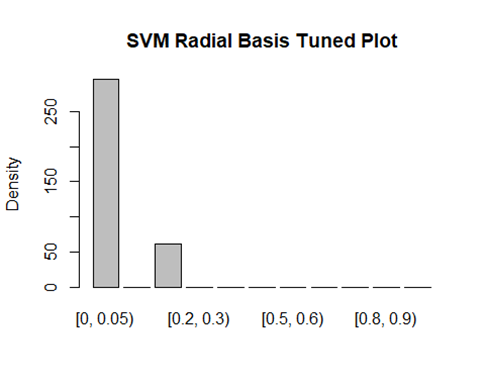
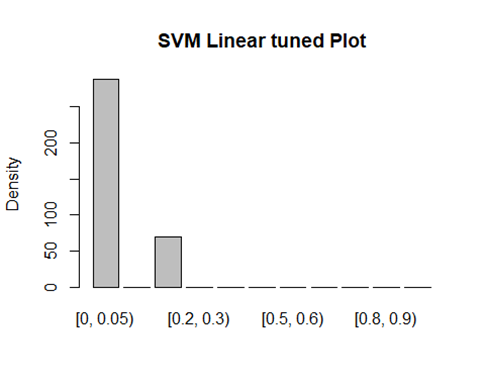
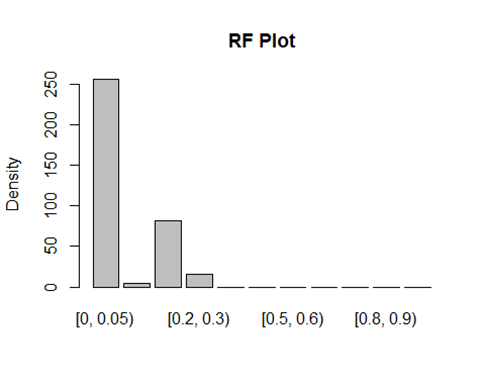
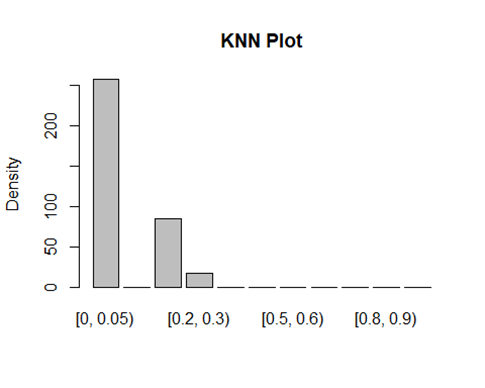
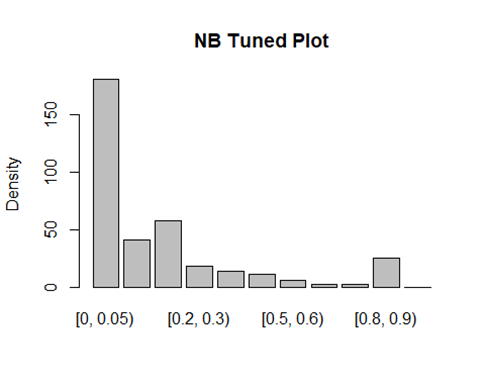
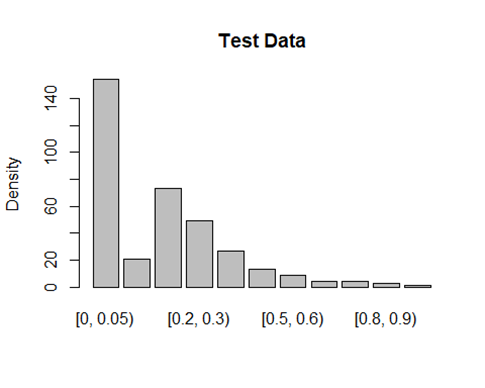


Figure Basic Classification Results

#### Classification Model 2: Classification Details With Base Data Set

Classification accuracies, Kappas and tuning variables are presented in Figure 15.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Accuracy (Test/Train) | Kappa (Test/Train) | Tuning |
| NB | 39% / 44% | 17% / 14% | Laplace |
| KNN | 47% / 49% | 18% / 20% | K |
| RF | 49% / 54% | 23% / 27% | mtry\* |
| SVM- Linear | 49% / 53% | 18% / 23% | c |
| SVM - RBF | 48% / 51% | 16% / 20% | sigma |

Figure Classification Accuracies, Kappas and Tuning Variables

#### Classification Model 3: Variable Importance for RF

The variable importance for the RF model is depicted in Figure 17.

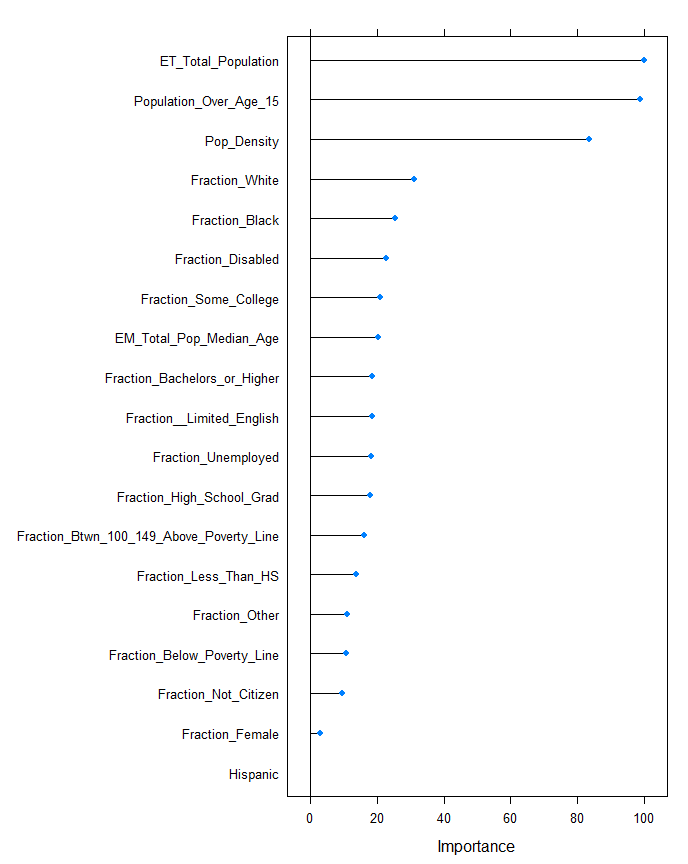


Figure Variable Importance for the RF model

#### Classification Model 4: Classification Results for Data Set Excursions

The following data set excursions were conducted using the most accurate model (RF)

* PCA dimensionality reduction
* Take out low death rate / very high instances bin
* Don’t normalize fractions
* Remove non-wealth variables

The RF algorithm was applied and tuned using k-fold validation.

Classification Results for these excursions are presented in Figure 15.

Figure Classification Results for Dataset Excursions

#### Classification Model 5: Classification Details for Dataset Excursions

Classification accuracies, Kappas and tuning variables are presented in Figure 19.

|  |  |  |
| --- | --- | --- |
| Model | Accuracy (Train/ Test) | Kappa (Train/Test) |
| RF baseline | 49% / 54% | 23% / 27% |
| PCA | 47% / 49% | 21% / 23% |
| Remove Bin1 | 30% / 38% | 2% / 13% |
| No Norm of Fractions | 49% / 52% | 21% / 25% |
| Remove Non-Wealth Variables | 48% / 53% | 21% / 28% |

Figure Result Details for Data Set Excursions

### Clustering Models

The next model used k-means clustering to see if counties that have a high mortality rate share similar characteristics, and also to determine if there were specific types of counties that were particularly high in fatality rates. The goal of the analysis was to establish county profiles that are particularly at risk for experiencing a high percentage of deaths per confirmed cases. Two R packages were used, “factoextra” and “stats”, for clustering whilte “ggplot2” was used for visualization.

First the dataset was divided into two sets, “high mortality” and “low mortality”, based on the national fatality rate for coronavirus, which is about 5.6%. Outliers were then eliminated, as there were quite a few counties where there were only a few cases and one death (which drove up the mortality rate to as high as 100%). Therefore, only counties with a large number of cases were used for clustering.

#### Clustering Model 1: High Mortality Rates

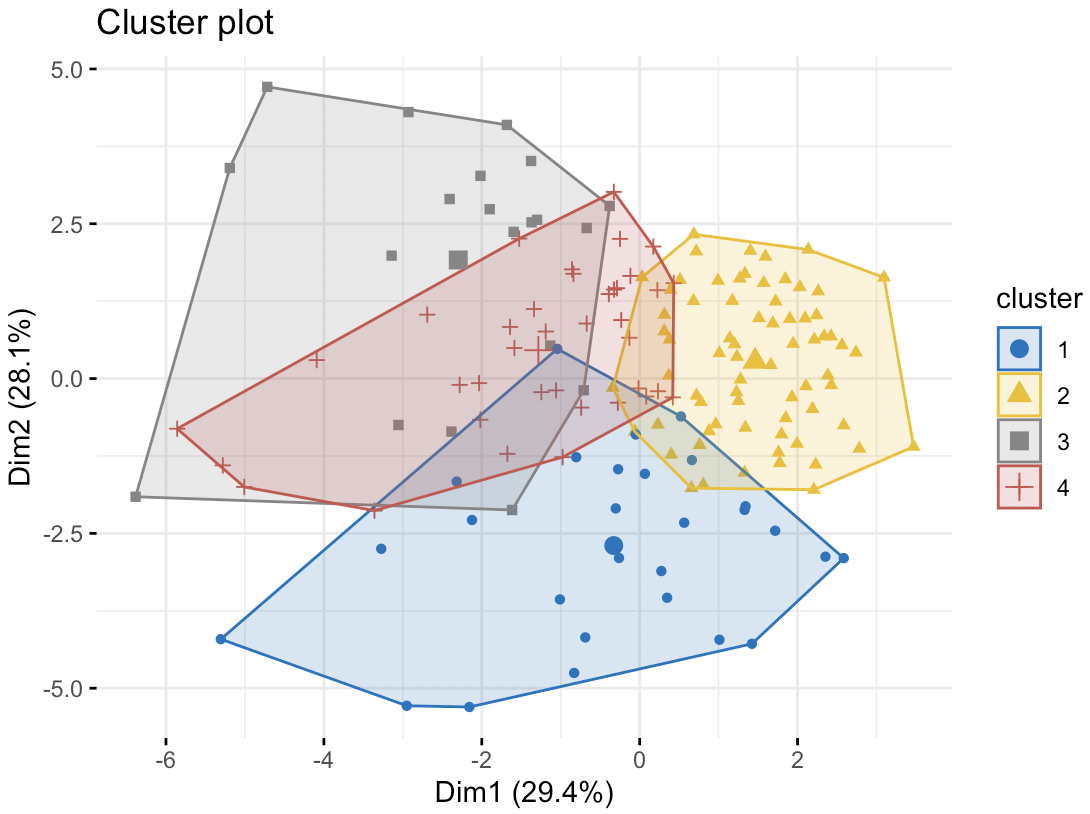


Figure Cluster Analysis, High Mortality

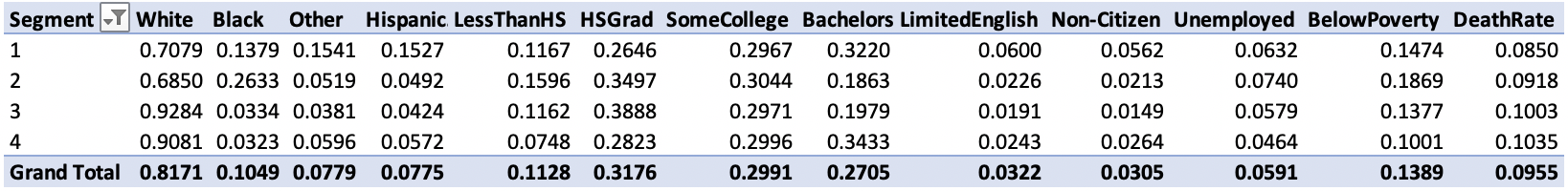


Figure Cluster Analysis Table, High Mortality

### 

#### Clustering Model 2: Low Mortality Rates

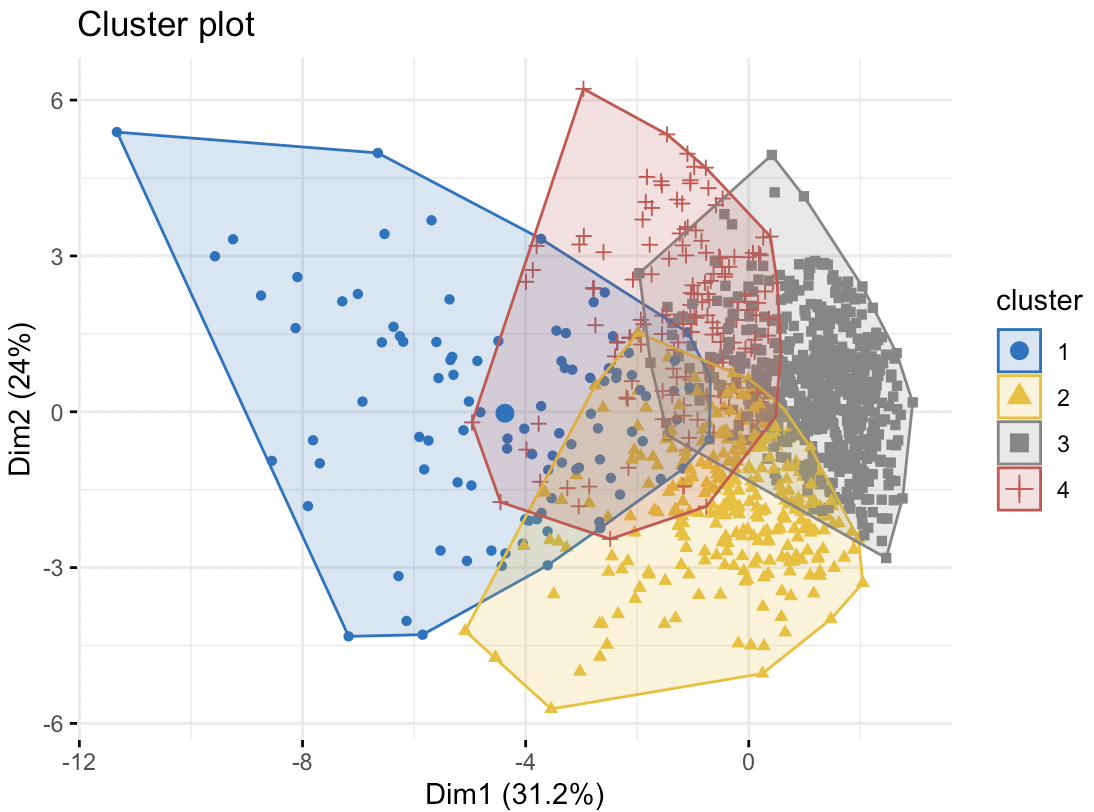


Figure Cluster Analysis, Low Mortality

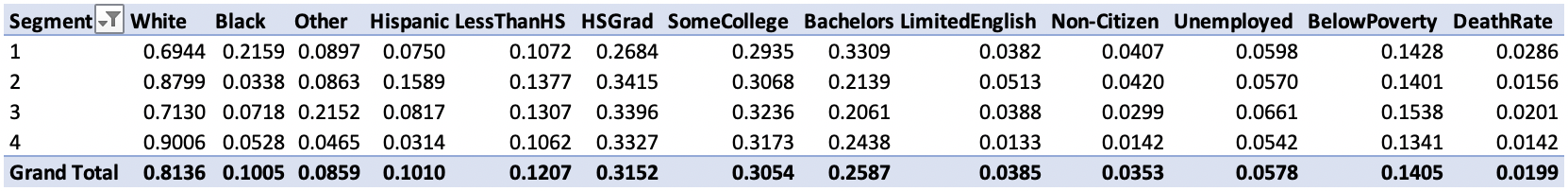


Figure Cluster Analysis, Low Mortality Table

### Association Rule Mining Models

ARM was used as the final model for analysis, to determine if certain demographic aspects of populations were associated with certain levels of mortality rates. The data was first normalized to be between 0 and 1, with lower ranges representing a demographic that is less populated with that specific variable. Three key measures - support, confidence and lift - were used to determine the relative strength of any rules found:

* The Support measure was used to calculate how frequent an itemset 𝑋 appeared in the dataset as the antecedent/left-hand-side (demographic variable)
* The Confidence measure was used to rank the likelihood of how frequently the association rule {𝑋}⇒{𝑌} was true in the dataset
* The Lift factor (which measured how much the co-occurrence of A and B exceeded the expected probability of A and B co-occurring, had they been independent) was used to determine the chance of A and B occurring together

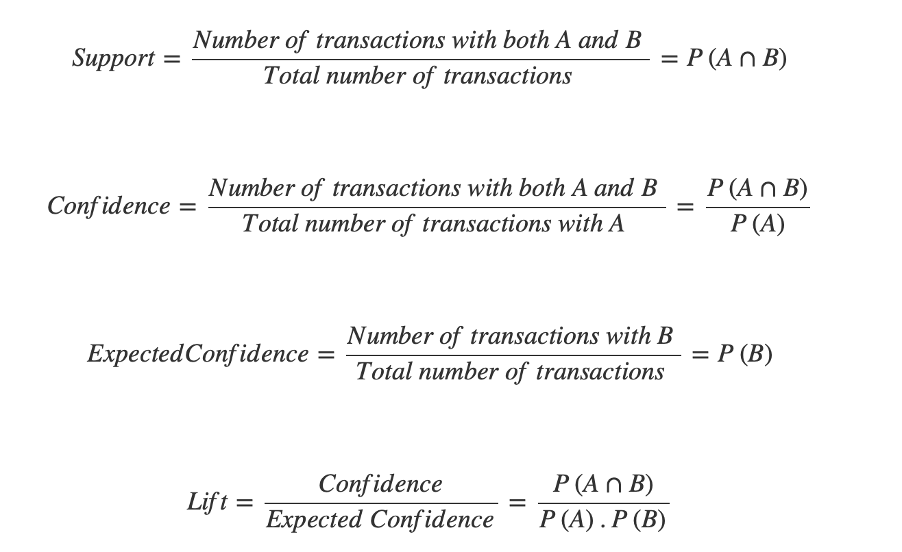


Figure Association Rules Key Measurements Formulas

#### ARM Model 1: Lowest Mortality Rate

Support: 0.001369863

Confidence: 1

Lift: 2.358643

RHS: Deaths\_Per\_Confirmed\_4\_16=[0, 0.05)

Rules:

|  |
| --- |
| EM\_Total\_Pop\_Median\_Age=[0.4, 0.5)  Fraction\_Female=[0.3, 0.4)} |
| Fraction\_Other=[0.9, 1]  Fraction\_Below\_Poverty\_Line=[0.9, 1] |
| Fraction\_Other=[0.9, 1]  Fraction\_Btwn\_100\_149\_Above\_Poverty\_Line=[0.7, 0.8) |
| Fraction\_Unemployed=[0.4, 0.5)  Fraction\_Not\_Citizen=[0.8, 0.9) |
| Fraction\_Disabled=[0.3, 0.4)  Fraction\_Not\_Citizen=[0.8, 0.9) |

Figure Lowest Mortality Rate

#### ARM Model 2: Low Mortality Rate

Support: 0.001369863

Confidence: 1

Lift: 4.949153

RHS: Deaths\_Per\_Confirmed\_4\_16=[0.1, 0.2)

Rules:

|  |
| --- |
| Fraction\_Bachelors\_or\_Higher=[0.9, 1] |
| Pop\_Density=[0.6, 0.7),Pop\_Density=[0.7, 0.8) |
| Fraction\_White=[0.7, 0.8)  Fraction\_Bachelors\_or\_Higher=[0.9, 1] |
| Fraction\_Female=[0.8, 0.9)  Fraction\_Bachelors\_or\_Higher=[0.9, 1] |
| EM\_Total\_Pop\_Median\_Age=[0.3, 0.4)  Fraction\_Bachelors\_or\_Higher=[0.9, 1] |

Figure Low Mortality Rate

#### ARM Model 3: Median Mortality Rate

Support: 0.001369863

Confidence: 1

Lift: 7.336683

RHS: Deaths\_Per\_Confirmed\_4\_16=[0.2, 0.3)

Rules:

|  |
| --- |
| Fraction\_White=[0.7, 0.8)  Fraction\_High\_School\_Grad=[0.5, 0.6) |
| Fraction\_Other=[0.5, 0.6)  Fraction\_Below\_Poverty\_Line=[0.5, 0.6) |
| Fraction\_Other=[0.4, 0.5)  Hispanic=[0.4, 0.5) |
| Fraction\_White=[0.4, 0.5)  Fraction\_Other=[0.4, 0.5) |
| Hispanic=[0.6, 0.7)  Fraction\_Not\_Citizen=[0.6, 0.7) |

Figure Median Mortality Rate

#### ARM Model 4: Above Median Mortality Rate

Support: 0.001369863

Confidence: 1

Lift: 13.27273

RHS: Deaths\_Per\_Confirmed\_4\_16=[0.3, 0.4)

Rules:

|  |
| --- |
| Fraction\_White=[0.6, 0.7)  Hispanic=[0.5, 0.6)  Fraction\_Unemployed=[0.4, 0.5) |
| Fraction\_White=[0.6, 0.7)  Fraction\_Black=[0, 0.1)  Hispanic=[0.5, 0.6) |
| Fraction\_Disabled=[0.5, 0.6)  Fraction\_Below\_Poverty\_Line=[0.5, 0.6)  Fraction\_Btwn\_100\_149\_Above\_Poverty\_Line=[0.8, 0.9) |
| Fraction\_White=[0.6, 0.7)  Fraction\_Other=[0.3, 0.4)  Fraction\_Not\_Citizen=[0.5, 0.6) |

Figure Above Median Mortality Rate

#### ARM Model 5: High Mortality Rate

Support: 0.001369863

Confidence: 1

Lift: 26.54545

RHS: Deaths\_Per\_Confirmed\_4\_16=[0.4, 0.5)

Rules:

|  |
| --- |
| Fraction\_Female=[0.7, 0.8)  Fraction\_Btwn\_100\_149\_Above\_Poverty\_Line=[0.7, 0.8)  Fraction\_Not\_Citizen=[0.5, 0.6) |
| Fraction\_White=[0.5, 0.6)  Fraction\_Unemployed=[0.4, 0.5)  Fraction\_Btwn\_100\_149\_Above\_Poverty\_Line=[0.7, 0.8) |
| Fraction\_Female=[0.7, 0.8)  Fraction\_\_Limited\_English=[0.3, 0.4)  Fraction\_Btwn\_100\_149\_Above\_Poverty\_Line=[0.7, 0.8)  Fraction\_Not\_Citizen=[0.5, 0.6) |

Figure High Mortality Rate

#### ARM Model 6: Highest Mortality Rate

Support: 0.001369863

Confidence: 1

Lift: 39.45946

RHS: Deaths\_Per\_Confirmed\_4\_16=[0.5, 0.6)

Rules:

|  |
| --- |
| EM\_Total\_Pop\_Median\_Age=[0.3, 0.4)  Fraction\_Limited\_English=[0.2, 0.3)  Fraction\_Below\_Poverty\_Line=[0.4, 0.5) |
| Fraction\_Female=[0.7, 0.8)  Fraction\_Black=[0.2, 0.3)  Fraction\_Unemployed=[0.5, 0.6) |
| Fraction\_Female=[0.7, 0.8)  Fraction\_White=[0.7, 0.8)  Fraction\_Black=[0.2, 0.3)  Fraction\_Unemployed=[0.5, 0.6) |

Figure Highest Mortality Rate

# Results

## EDA Results

### EDA Results 1: Dimensionality

* There are too many variables: need to reduce them

### EDA Results 2: Missing Data / NAs

* 35% of the rows are missing data and there are 2063 NAs.
* Most of the missing data is in the transportation-related data
* Delete variables related to transportation data from the analysis

### EDA Results 3: Reduce Dimensionality Based on Variable Applicability to Analytic Question

* Eight categories of variables were eliminated because they do not bear on the analytic question, are redundant, or have too much missing data
* Nine categories of variables were retained.

### EDA Results 4: Transform the Data

* The new variable - Deaths\_Per\_Confirmed\_4\_16 is the “y” variable for our analyses.
  + We will analyze if there is clustering of the other variables around Deaths\_Per\_Confirmed\_4\_16
  + We will analyze if there are other variables associated with Deaths\_Per\_Confirmed\_4\_16 through Association Rule Mining.
  + We will analyze if we can predict Deaths\_Per\_Confirmed\_4\_16 using a variety of classification algorithms.
* After the transportation variables were removed, only two rows had NAs and they were removed from the dataset.
* The dataset variables are almost prepared for analysis after consolidating, transforming, and normalizing the numeric data.

### EDA Results 5: Resultant Data

* There are only complete rows (no NAs) and the dimensionality is greatly reduced.

### EDA Results: Outliers

* Most Deaths\_Per\_Confirmed\_4\_16 values are under 25% and there appear to be just a few outliers when the x-axis , Confirmed, is very small. This would male sense. If one person is confirmed in a county and that one person dies, that is a 100% death rate
* The 3 standard deviation Z-score was computed and 14 data points were found where the Z-score exceeded that value. Those 14 data points (rows) were removed from the analysis as outliers
* Deaths\_Per\_Confirmed\_4\_16 was renormalized between 0 and 1

### EDA Results 7: Discretization

* Deaths\_Per\_Confirmed\_4\_16 had to be discretized for both Association Rule Mining and Classification
* The histogram with values of Deaths\_Per\_Confirmed\_4\_16 binned into ten equal values showed almost half of the values falling in the first bin (Deaths\_Per\_Confirmed\_4\_16 <=.1)
  + Note after Deaths\_Per\_Confirmed\_4\_16 outliers were removed, the death rates were all under .21. But because it was then normalized, the meaning has changed and Deaths\_Per\_Confirmed\_4\_16 = X no longer means the actual rate is X
* The histogram with bins = .005 (200 equal bins) shows almost 600 (~ 40%) in Bin1 (<=.005) and no other big spikes. Separate the spike into its own bin
* The bins chosen for Deaths\_Per\_Confirmed\_4\_16 were
  + Bin1 : d<=.005,
  + Bin2 : .005>d<=.1.,
  + Bin3 : .1<Bin3<=.2
  + And Bin4 through Bin11 each increasing increments of .1
* Association Rule Mining required all the “x” variables be discretized as well. They were all discretized into 10 equal bins

## Classification Results

### Classification Results 1: Classification Results With Base Data Set

* The question posed is “Can Deaths Per COVID Case Be Predicted by Wealth Indicators?”
* The Naïve Bayesian (NB) appears to have the best result because it most closely matches the pattern of the actual test data
* K Nearest Neighbor(KNN) and Random Forest (RF) both predicted most values in Bin1, Bin3 and Bin4
* Both SVM models predicted values only in Bin1 and Bin3

### Classification Model 2: Classification Details With Base Data Set

* While NB may have appeared visually to be the best match, it actually has the least accuracy.
* The variables are not independent which is an assumption for NB
* RF and SVM Linear have highest training accuracy at 49%.
* RF has slightly better test accuracy of 54% (vs 53%)
* Random Forest has the best Kappa.
* Random Forest is the Best model based on Accuracy and Kappa but it, and all this other models except NB, are “lazy” in the sense that they predict the Bins with the highest frequency in the training data at the expense of predicting any in the other bins.
  + This could be acceptable if average accuracy is the goal
  + But if understanding how to predict higher death rates ois of interest, this could be a problem.

### Classification Results 3: Variable Importance for RF

* Total Population, Size of Population over 15 (which is strongly correlated to total Population), and Population Density drive the predictions.
  + None of these are indicators of wealth
* Race, Disability and Education are the indicators of wealth that have most importance
  + They are followed by Age which is counter-intuitive given all of the information provided by US government and others that the elderly are at most risk for death
* Poverty level income, Alien status, Sex are last in importance
* Being Hispanic has no predictive importance

### Classification Results 4: Classification Results for Data Set Excursions

* “Test Data” and “RF Plot” depict the actual test data and the baseline RF results are depicted for comparison.
* An excursion was made with PCA Dimensionality reduction.
  + The point of this excursion is that reducing dimensionality can sometimes improve results.
  + “ PCA RF Plot” depicts dimensionality reduction using PCA.
  + This reduced the x variables to 5 in number.
  + While it still heavily favors Bin1, the distribution of result is a little broader
* An excursion was made without Bin1 data.
  + The point of this excursion if to see if a model can be developed that does a better job of predicting higher death rates.
  + RF w/o Bin1 Plot” represents taking out the favored bin, Bin1, that represents very low death rate and very high instances bin.
  + The shape of the results parallels the test data so this looks promising.
* An excursion was made where fractions were not normalized.
  + The point of this excursions was to see if normalizing to one fractions which are already a form of normalization was somehow losing the signal in the data.
  + “Un-normalizing” the Y variable meant it needed to be discretized again to fit between 0 and .2, rather than 0 and 1.
  + “Test Data with New Bins” represents the unnormalized test data in the new but equivalent bins
  + “RF No Norm Fraction Plot” is the result.
  + The results looks very similar to the baseline RF results so normalization does not appear to have impacted results
* An excursion was made to remove non-wealth variables
  + The point of this excursion is to see is if dropping the non-wealth that had most importance in the baseline RF Model would allow the wealth variables to predict better.
  + “RF Wealth Data Only” represents taking out those non-wealth variables (Total Population, Size of Population over 15, and Population Density)
  + The shape of the results parallels the test data so this looks promising.
  + The results looks very similar to the baseline RF results so the non-wealth and non-wealth variables predict essentially the same thing.

### Classification Results 5: Classification Details for Dataset Excursions

* All of the data set excursions except “Remove Bin1” performed comparably (w/in 2%) but worse than the baseline RF model
* While the shape of the results for removing Bin1 was interesting, the accuracy was very bad – 30%.
  + Much of that accuracy was just chance – the Kappa was only 2%
* The overall conclusion is that we cannot predict death rate with much accuracy at all, and that the baseline RF model is the best model we have.

## Clustering Results

### Clustering Results: High Mortality Rates

* Cluster 2 (yellow) has highest average fraction of population below poverty line out of allclusters, as well as:
  + Lowest fraction of college-educated individuals
  + Highest unemployment
  + Largest fraction black population
  + Is also the most compact cluster, which could indicate higher correlation between these variables and higher than average mortality
* Other clusters did not appear to reveal meaningful inferences

### Clustering Results: Low Mortality Rates

* Clusters in this category did not appear to show meaningful patterns, except cluster 3 which had highest average fraction of race “other” out of *all* of the clusters, by a fairly large margin (further analysis would be required in order to draw inferences)

## Association Rule Mining Results

### Association Rules: High Mortality Rates

* Results were sorted by LIFT, and several rules stood out between demographic data and normalized numeric values targeting deaths confirmed rate
* LIFT value gets higher while the rate of deaths per confirmed rises
* With RHS rule as “Death Per Confirmed Cases”, association rules for higher mortality rates included:
  + Higher female population
  + Higher fraction of population speaking limited English
  + Higher unemployment

### Association Rules: Low Mortality Rates

* With RHS rule as “Death Per Confirmed Cases”, association rules for lower mortality rates included:
  + Lower female population
  + Higher fraction of population not citizens
  + Higher fraction of population above poverty line
  + Higher fraction of population disabled
* Some rules were found to be fairly ambiguous/unclear, but some correlations were found between mortality rate and “wealth” indicators

# Conclusion

The research question, "Does wealth and wealth-related factors influence mortality rate of coronavirus," may require additional analysis to be answered. Some of the results show that indicators of poverty may be associated with higher death rates. The reason could be due to a one-off in the analysis, or it could be a sign of a real pattern in the way COVID-19 affects different people of different demographics.

However, when it comes to being able to use demographic data alone to predict coronavirus's mortality rate, the results are not accurate enough to make meaningful conclusions. Instead, we can only surmise why cases with higher mortality rate are in specific populations (such as in the cluster analysis where unemployment, poverty, and percentage of the black population all seemed to correlate with higher mortality rate). Lack of healthcare, or a lack of willingness to get treatment because of cost, may prevent specific populations from receiving treatment for potentially life-threatening illnesses. Lower-income individuals might live in crowded public housing, and when taking into account the fact that EDA found population density to be strongly correlated with higher cases and higher deaths, it seems that poverty is related, at least indirectly, with increased COVD mortality.

There may be other factors that influenced the somewhat ambiguous results of the data analysis. For one, as of today, there are no remarkably effective treatments for COVID. Whether they live in the Bronx or the Upper East Side of New York City, coronavirus can still take lives regardless of someone's income level or whether they are college-educated or not. The problem cannot be solved by throwing money at it. The development of treatments takes time, effective vaccines need to be put through rigorous trials, and there are no miracle cures that only the elite have access to. Even if you can, at the very least, afford a hospital visit, the data shows that many people who end up in the hospital don't leave alive.

Recommendations for further analysis include eliminating variables dependent on one another, for example, in the dataset race as a percentage of the total population, which is added to one in every case. Using a single race variable per test and performing the same test multiple times on different races might result in more clear outcomes. Additionally, using a different dataset that provides more granularity might help draw stronger correlations between economic status and COVID mortality rate. The current dataset, which groups coronavirus infections by county, might fail to pick up on the massive variations that can exist within just one county. Aggregating data by zip code, for example, might lead to better results as zip codes tend to be more homogeneous in their demographics. Even more ideal would be a dataset that includes detailed information on individual patients; however, that is likely not to be published due to HIPAA laws. Finally, using another predictor variable might suit the analysis better. For example, using confirmed cases or the rate of increase in cases from week to week could potentially reveal patterns not seen when studying the mortality rate.