 Covid-19 By County

Understanding variance in the rate of cases and deaths in US county populations

CIND 820 Big Data Project

By William Murphy

Table of Contents

[Abstract 3](#_Toc78456537)

[Literature Review 3](#_Toc78456538)

[Dataset 5](#_Toc78456539)

[Approach 6](#_Toc78456540)

[Findings 9](#_Toc78456541)

[Model Selection 9](#_Toc78456542)

[Forward Stepwise Feature Selection 12](#_Toc78456543)

[Model Performance After Feature Selection 14](#_Toc78456544)

[Model Analysis for County Case Rate 14](#_Toc78456545)

[Model Analysis for County death rate 17](#_Toc78456546)

[Summary of Findings 19](#_Toc78456547)

[Conclusions 21](#_Toc78456548)

[Appendix 22](#_Toc78456549)

[References 34](#_Toc78456550)

# Abstract

On January 19th, 2020, a 35 year old man went to a clinic in Snohomish County, Washington State. He had a 4-day history of cough and fever. Two days later, on January 21, 2020, Snohomish County would report the man as testing positive for Covid 19, being the first documented case in the United States. 352 days from the first reported case, on January 8th, 2021, US counties would report 300,777 new cases for that day alone. 302,772 people had died from Covid 19 in the United States and millions were dead world-wide. A pandemic on a global scale had a grip on the world’s population. There are 3142 counties in the United States of America. Each county has a different health, racial, social, political and economic make up and each county experienced Covid 19 differently. What this project attempts to do is use regression to discover if the existing health, social, economic and political data of each US county can explain the variance between counties of the percentage of the county population who became infected with Covid 19 and the percentage of the county population who died of Covid 19 from January 21st, 2020 to January 8th, 2021. This time period was chosen as it represented the period from the first confirmed cases till the maximum number of cases recorded in a day. Finally, it will attempt to discover if different factors contributed to the percentage of cases in a county than to the percentage of deaths in a county. An often used saying during this pandemic is “Covid doesn’t discriminate”, this project will attempt to discover if that is truth or myth.

# Literature Review

#### 1. Poverty, inequality and COVID-19: the forgotten vulnerable[[1]](#endnote-1)

The above study states that a number of factors increase people of low socio-economic status (SES) in getting COVID 19. First, they are more likely to live in overcrowded accommodation. Second, they are often employed in jobs were work from home is not an option. Third, people in low SES groups are more likely to have unstable work which can contribute to stress, which is known to lower a person’s immune system. Fourth, people of low SES tend to seek medical attention at a more advanced stage of illness. Also, access to health is determined by a person’s ability to access the healthcare system with ease, something people of low SES may feel that there are barriers to. They conclude that an increased exposure the virus, increase stress and lower access to healthcare makes people of low SES particularly vulnerable to COVID 19.

#### 2. Income and Poverty in the Covid19 Pandemic[[2]](#endnote-2)

The purpose of the above study was to create real time information as to the level of poverty in the United States and assess the effectiveness of government stimulus. The study is using data from Basic Monthly Current Population Survey (Monthly CPS), which contains high frequency data for a large, representative sample of U.S. families and individuals. The study shows that poverty fell in the early months of the pandemic due to government assistance. However, since some of the government assistance was a one-time payment, the study concludes that the initial drop in poverty was unlikely to continue as the pandemic and recovery occurs.

#### 3.COVID-19, school closures, and child poverty: a social crisis in the making[[3]](#endnote-3)

This paper is of the opinion that a long period of school closures would have negative social and health consequences for children living in poverty and could increase the current inequities in society. One key way that children living in poverty would be affected by school closures is an increase in food insecurity as schools are often a key source of meals for children living in poverty. This project also cites research to suggest that children living in poverty are more affected by breaks and disruptions in school than their peers who are not living in poverty. The project recommends that officials adapt their school closures to address the above points as well as prepare for considerable challenges that will arise once the pandemic is over.

#### 4. Monthly Poverty Rates in the United States during the COVID-19 Pandemic[[4]](#endnote-4)

The above study attempts to produce monthly estimates of poverty in the US. This is done by creating real times estimates of family income. This study found that the monthly poverty rate in the United States increased from 15% to 16.7% from February to September 2020. This increase in poverty has been particularly bad for children, and Black and Hispanic individuals. This study also notes the increase in poverty rates is also due to the expiration of certain benefits that had begun in the early days of the pandemic but had expired in the summer of 2020.

#### 5. Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 States[[5]](#endnote-5)

This study identified 12 US states that reported the race/ethnicity of the people being hospitalized for COVID 19 between April 30 and June 24, 2020. The study found that the share of white people being hospitalized was much smaller than their overall share of the state population in general. On the flip side, the share of black people being hospitalized was larger than their overall share in the state’s representation. The same disparity found in Black patients was also found in Hispanic patients and Native American patients (where that data existed). This pattern was not seen in the Asian population where hospitalization was lower than state representation.

#### 6.Racial, Economic, and Health Inequality and COVID-19 Infection in the United States[[6]](#endnote-6)

The purpose of this study was to report the connection of COVID 19 with respect to race, economic inequality and health in the United States. The study looked at demographic, socioeconomic, and mobility data from 369 US counties in 7 states. What it found is that the risk factors for infection and mortality were different. What it found was that more affluent counties were more prone to infection, but less affluent counties had a higher death rate. African Americans were the more vulnerable than other ethnicities to COIVD 19

#### 7. Impacts of the COVID-19 pandemic on rural America[[7]](#endnote-7)

The impact of COVID 19 in rural areas is under-studied compared to urban areas. The study found that rural unemployment rates increased more than their urban counterparts during the pandemic. They found major decline in the perceptions of local economic health in rural areas, increased reliance on unemployment benefits and negative impacts to mental health. The study predicts a long road to recovery of rural America from the pandemic.

#### 8. Obesity Is a Risk Factor for Greater COVID-19 Severity[[8]](#endnote-8)

This article examples the connection between obesity and COVID 19 severity among patients. The study found that the presence of obesity in a patient made that patient three times more likely to have a severe infection compared to a non-obese patient. The strong relationship between obesity and severity of COIVD 19 infection was not known at the time of the study, however the study notes several technical theories and states that obesity commonly aggravates the severity of respiratory diseases.

#### 9. The effect of smoking on COVID-19 severity: A systematic review and meta-analysis[[9]](#endnote-9)

This study aims to definitively quantify the effects of smoking on COVID 19 severity by analysing all studies that were published regarding the connection between smoking and COVID 19. What they found is that the smokers and former smokers had a significant increase in the chance of having a severe reaction to COVID 19 compared to never smokers.

# Dataset

The dataset will be an amalgamation of several different datasets, however all the datapoints are on a United States county level. Alaska, Hawaii and US territories are excluded. The data is mainly from the US Government but also contains some data from NGOs and data from the New York Times GitHub repository. The Covid 19 data on cases and deaths is daily county data that has been collected and aggregated by The New York Times. This data has been used in other scholarly articles

# Approach

#### Step 1: Creating the dataset

I am using data from multiple sources to create my dataset. As a result, I will need to combine the different sources into one dataset. The data I am using has been exported from various websites as CSV and XLSX files. All the datasets share the same characteristic, namely that they show data on a county level, and the US contains 3142 counties (excluding Puerto Rico and US territories). Once completed the data frame will have a row for each county in the United States, excluding Hawaii and Alaska, for a total of 3108 rows. Each county of the United States has a unique identifier named a Federal Information Process System Code (FIPS code) and the datasets that I am combining all contain the FIPS code. The FIPS code is used to join the different data into one dataset. In total there are 46 different statistics used as independent variables in this project.

The two dependent variables in this project are total county case rate divided by the population of that county and totals deaths per county divided by the population of that county. Going forward, these dependent variables will be referred to as county case rate and county death rate going forward.

#### Step 2: Cleaning the data

The Datasets that I choose for the most part contain complete data for each 3108 counties used for this project. However, some attributes are missing data points for certain counties. Null values were replaced with the mean values of the column.

There is no complete government dataset for COVID deaths by county in the United States, however the New York Times has created their own dataset that combines data from Federal, State and Local data. They have combined the data into a dataset that shows COVID 19 deaths and confirmed cases on a county level. The data does have some challenges that need to be dealt with. Firstly, New York City is in 5 separate US counties, however due to how COVID deaths were reported by local health authorities the case and death data is combined into one geographic location called New York City. To fix this, the total cases and deaths in New York City were assigned to the five counties based on the percentage of that county’s population relative to the total population of the city. As a result, the five counties received the following share of the cases and deaths: The Bronx 17%, Kings 31%, Queens 28%, New York(county) 19% and Richmond 6%. Two urban centers, Kansas City and Joplin are reported as their own geographic location and the cases and deaths that occurred in those two locations are excluded from the respective counties that the cities inhabit. The cases and deaths assigned to Kansas city were divided evenly between the four counties that Kansas City inhabits: Cass, Clay, Jackson and Platte counties in Missouri. Joplin was treated in the same manner with the cases and deaths assigned to evenly between Newton and Jasper counties in Missouri.

Both dependent variables failed a Shaprio-Wilk normal distribution test. Normalization and standardization of the dependent variables were explored but ultimately did not significantly improve the performance of the model.

County death rate had a significant skew of 1.67 and outliers in the upper quartile:

Chart, box and whisker chart

Description automatically generated

To improve the skew and increase the performance of the regression model all results above 0.2967 were removed. A total of 156 observations. This reduced the number observations from 3108 to 2952 and reduced the skew to 0.55.

Similar cleaning was done to the county case rate, while the skew of 0.85 was not as significant as that of the county death rate, there were outliers present in the upper Quartile.

Chart, box and whisker chart

Description automatically generated

To improve the skew and increase the performance of the regression model all results above 13.74 were removed. This reduced the number of observations to 3045. A reduction of 63 observations. The skew was reduced to 0.04.

All analysis and models were done with outliers included and outliers removed. The performance of the models increased with the removal of the outliers.

*Linear regression model with and without outliers:*

|  |  |  |  |
| --- | --- | --- | --- |
| **County Death rate** | | | |
| **Outliers removed** | | **With outliers** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.269 | R-Squared | 0.257 |
| Adj R squared | 0.257 | Adj R squared | 0.245 |
| MSE | 0.003 | MSE | 0.006 |
| MAE | 0.045 | MAE | 0.056 |
| RMSE | 0.058 | RMSE | 0.079 |

*Linear regression model with and without outliers:*

|  |  |  |  |
| --- | --- | --- | --- |
| **County Case Rate** | | | |
| **Outliers removed** | | **With outliers** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.509 | R-Squared | 0.479 |
| Adj R squared | 0.504 | Adj R squared | 0.4711 |
| MSE | 3.163 | MSE | 4.065 |
| MAE | 1.374 | MAE | 1.493 |
| RMSE | 1.775 | RMSE | 2.012 |

#### Step 3: Analysing the Data

This project has used Python to create a regression model with the purpose of explaining the variance between different counties in respect to their county case rate and county death rate. Using python, the correlation between the dependant variables and independent variables were analysed to identify correlation. Multicollinearity was discovered between some of the independent variables and Principal Component Analysis(PCA) and Forward Stepwise Feature Selection were used to address this. Ultimately, PCA was not effective but Forward Stepwise Feature Selection was effective and used for the final model. Five different regression algorithms were tested for performance and Random Forrest was selected for the final model as it performed the best for county case rate and county death rate.

#### Step 4: Presenting the Data

The python code used for this project was uploaded to Github (<https://github.com/wmurphy50/Covid-by-US-County>) along with the dataset as a CSV file. This word document will be used to present my model evaluation, feature selection, correlation analysis and conclusions.

# Findings

## Model Selection

The regression model for county case rate was first created using all 46 independent variables. The model was trained using 10-fold cross validation repeated 3 times. Five different regression algorithms were used and the results were evaluated by five metrics: R-Squared, Adjusted R-Squared, Mean Squared Error(MSE), Mean Absolute Error(MAE) and Root Mean Squared Error(RMSE).

*Linear regression results:*

|  |  |  |  |
| --- | --- | --- | --- |
| **Linear Regression** | | | |
| **Cases vs Population** | | **Deaths vs Population** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.509 | R-Squared | 0.269 |
| Adj R squared | 0.504 | Adj R squared | 0.257 |
| MSE | 3.163 | MSE | 0.003 |
| MAE | 1.374 | MAE | 0.045 |
| RMSE | 1.775 | RMSE | 0.058 |

Principal Component Analysis(PCA) was also used to address the large number of independent variables and multicollinearity of variables. Ultimately, PCA resulted in a much lower R squared when using linear regression.

*PCA Linear Regression results:*

|  |  |  |  |
| --- | --- | --- | --- |
| **PCA Linear Regression** | | | |
| **Cases vs Population** | | **Deaths vs Population** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.219 | R-Squared | 0.063 |
| Adj R squared | 0.2183 | Adj R squared | 0.062 |
| MSE | 6.093 | MSE | 0.008 |
| MAE | 1.86 | MAE | 0.065 |
| RMSE | 2.463 | RMSE | 0.089 |

Ridge and Lasso regression were then used to address any multicollinearity in the data as some of the independent variables are strongly correlated. For example, the ‘Total hospitals in county’ and the ‘Population’ variables, which have a correlation coefficient of 0.92. Ultimately, the Ridge and Lasso algorithms returned practically the same result as Linear Regression

*Ridge Regression results:*

|  |  |  |  |
| --- | --- | --- | --- |
| **Ridge Regression** | | | |
| **Cases vs Population** | | **Deaths vs Population** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.471 | R-Squared | 0.269 |
| Adj R squared | 0.466 | Adj R squared | 0.257 |
| MSE | 4.124 | MSE | 0.003 |
| MAE | 1.508 | MAE | 0.045 |
| RMSE | 2.026 | RMSE | 0.058 |

*Lasso Regression results:*

|  |  |  |  |
| --- | --- | --- | --- |
| **Lasso Regression** | | | |
| **Cases vs Population** | | **Deaths vs Population** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.481 | R-Squared | 0.269 |
| Adj R squared | 0.473 | Adj R squared | 0.257 |
| MSE | 3.101 | MSE | 0.003 |
| MAE | 1.379 | MAE | 0.045 |
| RMSE | 1.76 | RMSE | 0.058 |

Finally, Random Forrest and Gradient Boost Regression were used. Both resulted in better performing models with Random Forrest achieving the best scores of the five different algorithms:

*Gradient Boosting Regression results:*

|  |  |  |  |
| --- | --- | --- | --- |
| **Gradient Boosting Regression** | | | |
| **Cases vs Population** | | **Deaths vs Population** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.512 | R-Squared | 0.292 |
| Adj R squared | 0.505 | Adj R squared | 0.28 |
| MSE | 2.916 | MSE | 0.003 |
| MAE | 1.336 | MAE | 0.044 |
| RMSE | 1.707 | RMSE | 0.057 |

*Random Forrest Regression results:*

|  |  |  |  |
| --- | --- | --- | --- |
| **Random Forrest Regression** | | | |
| **Cases vs Population** | | **Deaths vs Population** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.523 | R-Squared | 0.317 |
| Adj R squared | 0.515 | Adj R squared | 0.306 |
| MSE | 2.864 | MSE | 0.003 |
| MAE | 1.323 | MAE | 0.044 |
| RMSE | 1.69 | RMSE | 0.056 |

As a result, this project will use Random Forrest Regression as its algorithm and will not use Principal Component Analysis to reduce the number of independent variables.

## Forward Stepwise Feature Selection

Due to the large number of independent variables or features and the multicollinearity of some of the independent variables, forward stepwise feature selection was used to reduce the number of features being used to train the model.

For cases by county, 28 features were optimum to create the model:

Chart

Description automatically generated

Despite dropping 18 features, our Adjusted R-Squared value when using Random Forrest regression is relatively unchanged, dropping from 0.5056 to 0.497. Dropping these features allow us to see through some of the noise to determine which features are affecting our model.

The below tables show the remaining features with an absolute correlation coefficient above 0.1 for County case rate:

|  |  |
| --- | --- |
| **Independent Variables** | **Correlation** |
| % Always wear a mask in public | -0.44 |
| % Less Than 18 Years of Age | 0.37 |
| % Sometimes wear a mask in public | 0.34 |
| % voted GOP 2020 | 0.22 |
| % Social Association Rate | 0.21 |
| Median Household Income | -0.17 |
| % Drive Alone to Work | 0.17 |
| % Smokers | 0.16 |
| % Frequently wear a mask in public | 0.16 |
| % Fair or Poor Health | 0.13 |
| % American Indian & Alaska Native | 0.13 |
| % Uninsured | 0.12 |
| % Female | -0.11 |

Death by county population required even fewer variables, as the optimum number was 20 features:

A screenshot of a computer

Description automatically generated with medium confidence

Reducing the number of variables in this case also did not significantly reduce the Adjusted R-Squared value when using Random Forrest Regression. It dropped from .306 to.289, despite dropping 26 features.

The below tables show the remaining features with an absolute correlation coefficient above 0.1 for county death rate:

|  |  |
| --- | --- |
| **Independent Variables** | **Correlation** |
| % Completed High School | -0.27 |
| % Children in Single-Parent Households | 0.27 |
| Median Household Income | -0.25 |
| % Smokers | 0.23 |
| % Black | 0.22 |
| % Adults with Diabetes | 0.22 |
| % Non-Hispanic White | -0.20 |
| Income Inequality | 0.18 |
| % Frequent Mental Distress | 0.14 |
| % Social Association Rate | 0.14 |
| % Less Than 18 Years of Age | 0.14 |
| % Asian | -0.11 |

### Model Performance After Feature Selection

|  |  |  |  |
| --- | --- | --- | --- |
| **Random Forrest Regression after Forward Stepwise Feature selection** | | | |
| **Cases vs Population** | | **Deaths vs Population** | |
| **Metric** | **Score** | **Metric** | **Score** |
| R-Squared | 0.501 | R-Squared | 0.294 |
| Adj R squared | 0.497 | Adj R squared | 0.289 |
| MSE | 2.993 | MSE | 0.003 |
| MAE | 1.357 | MAE | 0.044 |
| RMSE | 1.728 | RMSE | 0.057 |

### Model Analysis for County Case Rate

The model can explain almost half of the variance between county case rate by using a combination of independent variables pertaining to mask use, politics, race, economics, health and social factors.

#### Mask use

The data appears to indicate that community wide mask wearing contributes to the control of Covid 19. It should be noted that the responses of the NY Times survey were asked: How often do you wear a mask in public when you expect to be within six feet of another person? Any answer, other than ‘Always’ has a positive correlation with county case rate, even ‘Frequently’. This data is consistent with other studies that indicate community wide mask wearing may contribution to the control of Covid 19 by reducing the transmission by asymptomatic infected people.[[10]](#endnote-10)

#### The Politics of Covid – 19

Voting for Donald Trump in the 2020 presidential election has a positive correlation with county case rate. Studies have shown that counties that favoured Donald Trump in the 2016 election were less likely to exhibit physical distancing.[[11]](#endnote-11) What this dataset appears to show is that mask use at the time that the New York Times survey was taken is also linked to voting for Donald Trump in 2020.

|  |  |
| --- | --- |
| **Mask Use** | **Correlation to '% voted GOP 2020'** |
| % Always wear a mask in public | -0.55 |
| % Frequently wear a mask in public | 0.24 |
| % Sometimes wear a mask in public | 0.37 |
| % Never wear a mask in public | 0.41 |
| % Rarely wear a mask in public | 0.42 |

**Race**

This dataset appears to confirm one of the articles cited in my literature review: Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 Statesv. Namely that White and Asian populations are under-represented in hospitalization rates for Covid 19 compared to their make-up of state populations. Conversely, American Indian and Alaskan Native populations were substantially overrepresented.

|  |  |
| --- | --- |
| **Independent Race Variables** | **Correlation** |
| **% American Indian & Alaska Native** | .13 |

#### Social Factors

Apart from mask use, age has the strongest correlation with county case rate. There is a positive correlation with the percentage of a county’s population under the age of 18. This may be due to children attending school in higher numbers in counties with a higher percentage of people under 18. It is also observed that the percentage of a population that is female is negatively correlated with county case rate.

There is a positive correlation with the % Social Association Rate in a county. Social association rate include membership organizations such as civic organizations, bowling centers, golf clubs, fitness centers, sports organizations, religious organizations, political organizations, labor organizations, business organizations, and professional organizations.

|  |  |
| --- | --- |
| **Independent Social Variables** | **Correlation** |
| % Less Than 18 Years of Age | 0.37 |
| % Social Association Rate | 0.21 |
| % Drive Alone to Work | 0.17 |
| % Female | -0.11 |

#### Economic Factors

Median household income is negatively correlated with County case rate.

|  |  |
| --- | --- |
| **Independent Economic Variable** | **Correlation** |
| **Median Household Income** | -0.17 |

#### Health Factors

Independent variables related to poor health, smoking and lack of health insurance are positively correlated with County case rate.

|  |  |
| --- | --- |
| **Independent Health Variables** | **Correlation** |
| **% Smokers** | 0.16 |
| **% Fair or Poor Health** | 0.13 |
| **% Uninsured** | 0.12 |

### Model Analysis for County death rate

The independent variables in this dataset are better at explaining the variance of county case rate compared to county death rate. The adjusted R Squared score for our county death rate was 0.289 meaning that almost 30% of the variance between counties was explained by our model. The notable difference is the lack of political or mask use independent variables when compared to cases per county. The independent variables that contribute to this model are race, economic, health and social factors.

#### Economics

Higher median income is negatively correlated to deaths per county. Poverty indicator, Income Inequality had a positive correlation with county death rate.

|  |  |
| --- | --- |
| **Independent Economic Variables** | **Correlation** |
| **Median Household Income** | -0.25 |
| **Income Inequality** | 0.18 |

#### Race

|  |  |
| --- | --- |
| **Independent Race Variables** | **Correlation** |
| **% Black** | 0.22 |
| **% Non-Hispanic White** | -0.20 |
| **% Asian** | -0.11 |
| **% Hispanic** | 0.08 |
| **% American Indian & Alaska Native** | 0.07 |

Numerous articles have shown COVID-19 mortality disproportionally affects black populations in the United States[[12]](#endnote-12) [[13]](#endnote-13) vi v. This project shows a positive correlation between county death rate and the percentage of the black population by county. This data also shows a negative correlation between death per county population and the percentage of White and Asian people by county, along with a positive correlation of deaths by the percentage of Hispanic and American Indian & Alaska Native per county population. These findings are consistent with the findings of the article cited in my literature review, Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 States.v

#### Health

Smoking had a positive correlation with county death rate and has been shown to increase Covid-19 mortality. ix Diabetes has shown to have a close relationship with Covid-19 mortality.[[14]](#endnote-14)The World Health Organization has identified people with mental illness as higher risk for Covid-19 mortality.[[15]](#endnote-15)

|  |  |
| --- | --- |
| **Independent Health Variables** | **Correlation** |
| **% Smokers** | 0.23 |
| **% Adults with Diabetes** | 0.22 |
| **% Frequent Mental Distress** | 0.14 |

While these variables are an indicator of heath, a closer look indicates that they are highly correlated with Median Household Income and can also be used as an indicator of poverty.

|  |  |
| --- | --- |
| **Independent Health Variables** | **Correlation to Median Household Income** |
| **% Smokers** | -0.71 |
| **% Frequent Mental Distress** | -0.69 |
| **% Adults with Diabetes** | -0.48 |

#### Social Factors

Like county case rate, county death rate has a positive correlation with the percentage of a county population under 18 years of age and the Social Association Rate.

|  |  |
| --- | --- |
| **Independent Social Variables** | **Correlation** |
| % Completed High School | -0.27 |
| % Children in Single-Parent Households | 0.27 |
| % Social Association Rate | 0.14 |
| % Less Than 18 Years of Age | 0.14 |

While education and family structure is often considered a social statistic, it is correlated with the independent economic variable, Median Household Income in this model and reinforces the findings that poverty is an important indicator of deaths by Covid 19 in the United States.

|  |  |
| --- | --- |
| **Independent Social Variables** | **Correlation to Median Household Income** |
| % Children in Single-Parent Households | -0.44 |
| % Completed High School | 0.52 |

### Summary of Findings

* A Random Forrest Regression model was selected after it performed better to four other regression algorithms for county case rate and deaths per population
* Principal Component Analysis proved ineffective to increase the performance of our model
* Forward Stepwise Feature Selection was able to effectively reduce the number of independent variables in our model to address multicollinearity and reduce the overall ‘noise’ of the many independent variables without significantly affecting the performance of the models
* Our models explained half of all variance for county case rate and almost 30% of all variance in death per county population. The independent variables selected for this project were better at predicting county case rate than county death rate
* Analysis of the independent variables in the model for county case rate found that variables pertaining to mask use, politics, race, economics, health and social factors contributed to variance between US counties
* Analysis of the independent variables in the model for county death rate found that variables pertaining to race, economics (especially poverty), health and social factors contributed to variance between US counties

# Conclusions

"I don't wear masks like him. Every time you see him, he's got a mask.”[[16]](#endnote-16). These words were spoken by the then President of the United States, Donald Trump, in an apparent attempt to mock his Democratic challenger, Joe Biden during a presidential debate in October 2020. This project shows correlation of mask use and mask hesitancy to voting for Donald Trump in 2020. These words were said 254 days after the first case of Covid 19 was reported in the United States. 98 days later the United States would record 300,777 cases in a single day. A single day record that still stands at the time of writing this.

Of the 46 independent variables used in this data set, the best predictor of variance in the percentage of the population of US counties that became infected with Covid-19 was the ratio of a county’s population that indicated that they would “Always” wear a mask in public when they expect to be within six feet of another person. This data was collected in a New York Times poll in July 2020. Of the five possible answers in the survey, all other answers (“never”,” rarely”, “sometimes” and “frequently”) had a positive correlation with cases of Covid 19 by county population. What this shows and what other studies have also shownx, is that community wide mask use is an effective way to control Covid 19 spread.

The adage that “Covid-19 does not discriminate” is a myth. This project finds what other peer review studies have found: race and poverty were a determinant in the United States during the time period covered in this dataset for cases and deaths per county population. The percentage of White and Asian people in a county population had negative correlations with deaths. The percentage of Black and American Indian and Alaska Native people in a county’s population was found to have a positive correlation with county death rate. Median household income also has negative correlations with both cases and deaths. Social and health factors also associated with poverty had positive correlations with deaths.

This project found that Random Forest was the most effective regression algorithm to model this data set. The independent variables were better able to explain the variance in county case rate (adj R squared: 0.493) compared to county death rate (adj R squared: 0.288). Due to the scope of this project, a large number of independent variables were selected for analysis, forward stepwise feature selection proved to be an effective way of reducing the number of independent variables used to create the models and allowed for more insightful analysis of the independent variables.

# Appendix

Below are the sources, descriptions and descriptive statistics of the dataset used for this project:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Hospitals per county | | | | |
| **Source(s)** | **Healthcare Cost Report Information System (HCRIS) data for hospitals** | | | | |
| **Description** | the number of hospitals that filed a cost report since the beginning of 2018, for each county. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 1.44 | | 1.00 | 2.56 | 74.00 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Poor or fair health | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Percentage of adults reporting fair or poor health (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 20.25 | | 20.00 | 5.13 | 42.00 | 9.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Poor physical health days | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Average number of physically unhealthy days reported in past 30 days (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 4.41 | | 4.41 | 0.78 | 8.30 | 2.37 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Poor mental health days | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Average number of mentally unhealthy days reported in past 30 days (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 4.69 | | 4.73 | 0.67 | 7.29 | 2.69 |

## 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Adult smoking | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Percentage of adults who are current smokers (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 21.36 | | 21.19 | 4.18 | 44.57 | 7.08 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Adult obesity | | | | |
| **Source(s)** | United States Diabetes Surveillance System : 2017 | | | | |
| **Description** | Percentage of the adult population (age 20 and older) that reports a body mass index (BMI) greater than or equal to 30 kg/m2. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 33.48 | | 33.80 | 6 | 58.9 | 11.80 |

## 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Physical inactivity | | | | |
| **Source(s)** | United States Diabetes Surveillance System : 2017 | | | | |
| **Description** | Percentage of adults age 20 and over reporting no leisure-time physical activity. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 26.8 | | 26.5 | 5.84 | 50.4 | 8.9 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Excessive drinking | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System - CDC : 2018 | | | | |
| **Description** | Percentage of adults reporting binge or heavy drinking (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 18.9 | | 18.83 | 3.25 | 28.34 | 6.45 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Uninsured | | | | |
| **Source(s)** | Small Area Health Insurance Estimates US Census: 2018 | | | | |
| **Description** | Percentage of population under age 65 without health insurance. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 18.08 | | 18.44 | 4.15 | 28.34 | 3.55 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Flu Vaccinations | | | | |
| **Source(s)** | Mapping Medicate Disparities Tool | | | | |
| **Description** | Percentage of fee-for service Medicare enrollees that had an annual flu vaccination | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 18.08 | | 18.44 | 4.15 | 28.34 | 3.55 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | High school completion | | | | |
| **Source(s)** | American Community Survey US Census, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of adults ages 25 and over with a high school diploma or equivalent. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 86.79 | | 88.07 | 6.28 | 98.88 | 26.44 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Some college | | | | |
| **Source(s)** | American Community Survey US Census, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of adults ages 25-44 with some post-secondary education. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 57.94 | | 57.97 | 11.99 | 100 | .83 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Unemployment | | | | |
| **Source(s)** | Bureau of Labor Statistics : 2019 | | | | |
| **Description** | Percentage of population ages 16 and older unemployed but seeking work. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 4.01 | | 3.71 | 1.5 | 19.31 | .74 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Children in poverty | | | | |
| **Source(s)** | Small Area Income and Poverty Estimates US Census : 2019 | | | | |
| **Description** | Percentage of people under age 18 in poverty. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 20.15 | | 18.90 | 8.51 | 63.4 | 2.4 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Income inequality | | | | |
| **Source(s)** | American Community Survey US Census, 5-year estimates : 2015-2019 | | | | |
| **Description** | Ratio of household income at the 80th percentile to income at the 20th percentile. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 4.52 | | 4.39 | .77 | 10.50 | 2.41 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Children in single-parent households | | | | |
| **Source(s)** | American Community Survey, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of children that live in a household headed by single parent. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 24.58 | | 23.08 | 10.02 | 75.1 | 0 |

## 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Social associations Rate | | | | |
| **Source(s)** | County Business Patterns US Census: 2018 | | | | |
| **Description** | Number of membership associations per 10,000 population. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 11.57 | | 10.96 | 5.99 | 55.56 | 0 |

## 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Population Density | | | | |
| **Source(s)** | US Census Bureau Covid 19 site  https://covid19.census.gov/datasets/USCensus::average-household-size-and-population-density-county | | | | |
| **Description** | Count population density by square kilometer | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 104.51 | | 17.2666 | 696.34 | 27819.80 | 0.014 |

## 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Severe housing problems | | | | |
| **Source(s)** | Comprehensive Housing Affordability Strategy (CHAS) data : 2013-2017 | | | | |
| **Description** | Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, lack of kitchen facilities, or lack of plumbing facilities. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 13.63 | | 13.06 | 4.56 | 69.14 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Driving alone to work | | | | |
| **Source(s)** | American Community Survey US Census, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of the workforce that drives alone to work. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 79.69 | | 81.22 | 7.7 | 97.10 | 5.7 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Frequent physical distress | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System : 2018 | | | | |
| **Description** | Percentage of adults reporting 14 or more days of poor physical health per month (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 13.68 | | 13.54 | 2.64 | 29.19 | 7.01 |

## 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Frequent mental distress | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System : 2018 | | | | |
| **Description** | Percentage of adults reporting 14 or more days of poor mental health per month (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 15.16 | | 15.16 | 2.37 | 24.67 | 8.84 |

## 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Diabetes prevalence | | | | |
| **Source(s)** | United States Diabetes Surveillance System : 2017 | | | | |
| **Description** | Percentage of adults aged 20 and above with diagnosed diabetes. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 12.38 | | 11.90 | 3.69 | 29.50 | 3.1 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Food insecurity | | | | |
| **Source(s)** | Map the Meal Gap : 2018 | | | | |
| **Description** | Percentage of population who lack adequate access to food. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 13.3 | | 13.1 | 3.75 | 30.40 | 3.6 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Limited access to healthy foods | | | | |
| **Source(s)** | USDA Food Environment Atlas : 2015 | | | | |
| **Description** | Percentage of population who are low-income and do not live close to a grocery store. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 8.74 | | 6.58 | 8.46 | 71.84 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Insufficient sleep | | | | |
| **Source(s)** | Behavioral Risk Factor Surveillance System : 2018 | | | | |
| **Description** | Percentage of adults who report fewer than 7 hours of sleep on average (age-adjusted). | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 36.92 | | 36.94 | 3.96 | 49.06 | 25.62 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Median household income | | | | |
| **Source(s)** | Small Area Income and Poverty Estimates : 2019 | | | | |
| **Description** | The income where half of households in a county earn more and half of households earn less. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 55525.36 | | 53106.00 | 14562.33 | 151806 | 24732 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Children eligible for free or reduced price lunch | | | | |
| **Source(s)** | National Center for Education Statistics : 2018-2019 | | | | |
| **Description** | Percentage of children enrolled in public schools that are eligible for free or reduced price lunch. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 54.77 | | 52.96 | 18.63 | 100 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Homeownership | | | | |
| **Source(s)** | American Community Survey, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of occupied housing units that are owned. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 71.52 | | 72.78 | 8.39 | 93.06 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Population | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Resident population. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 105659.91 | | 25619 | 337990.39 | 10003107 | 86 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % below 18 years of age | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population below 18 years of age. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 21.95 | | 21.98 | 3.5 | 41.68 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % 65 and older | | | | |
| **Source(s)** | Percentage of population ages 65 and older. | | | | |
| **Description** | Census Population Estimates : 2019 | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 19.73 | | 19.39 | 4.82 | 58.17 | 4.86 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Non-Hispanic Black | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is non-Hispanic Black or African American. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 9.25 | | 2.43 | 14.43 | 85.87 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % American Indian & Alaska Native | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is American Indian or Alaska Native. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 2.37 | | 0.65 | 7.72 | 92.41 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Asian | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is Asian. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 1.59 | | 0.75 | 2.99 | 43.36 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Hispanic | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is Hispanic. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 9.92 | | 4.53 | 14.04 | 96.35 | 0.65 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Non-Hispanic White | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is non-Hispanic White. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 75.40 | | 82.63 | 20.33 | 97.83 | 2.69 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % not proficient in English | | | | |
| **Source(s)** | American Community Survey, 5-year estimates : 2015-2019 | | | | |
| **Description** | Percentage of population that is not proficient in English. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 1.72 | | 0.73 | 2.84 | 34.44 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Females | | | | |
| **Source(s)** | Census Population Estimates : 2019 | | | | |
| **Description** | Percentage of population that is female. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 49.90 | | 50.31 | 2.29 | 57.01 | 26.51 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Rural | | | | |
| **Source(s)** | Census Population Estimates : 2010 | | | | |
| **Description** | Percentage of population living in a rural area. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 58.61 | | 59.54 | 31.56 | 100.00 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Provisional COVID Death Counts in the United States by County as of January 8th, 2021 | | | | |
| **Source(s)** | Ny Times - https://github.com/nytimes/COVID-19-data | | | | |
| **Description** | Since late January 2020, The Times has tracked coronavirus cases and deaths as they are announced using data released by countries, states and local health officials. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 126 | | 33 | 496 | 14641 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | Provisional COVID Case Counts in the United States by County as of January 8th, 2021 | | | | |
| **Source(s)** | Ny Times - https://github.com/nytimes/COVID-19-data | | | | |
| **Description** | Since late January 2020, The Times has tracked coronavirus cases and deaths as they are announced using data released by countries, states and local health officials. | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 7795 | | 2000 | 29024 | 1046424 | 0.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | per\_gop | | | | |
| **Source(s)** | MIT Election Data and Science Lab  https://electionlab.mit.edu/data | | | | |
| **Description** | Percent of the population that voted Republican in the 2020 US federal election | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 64.97 | | 68.29 | 16.127 | 96.18 | 5.39 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Never wear a mask in public | | | | |
| **Source(s)** | NY Times Github repository | | | | |
| **Description** | This data comes from a large number of interviews conducted online by the global data and survey firm Dynata at the request of The New York Times. The firm asked a question about mask use to obtain 250,000 survey responses between July 2 and July 14 2020.  Specifically, each participant was asked: How often do you wear a mask in public when you expect to be within six feet of another person? | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 8 | | 6.8 | 5.85 | 43.2 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Rarely wear a mask in public | | | | |
| **Source(s)** | NY Times Github repository | | | | |
| **Description** | This data comes from a large number of interviews conducted online by the global data and survey firm Dynata at the request of The New York Times. The firm asked a question about mask use to obtain 250,000 survey responses between July 2 and July 14 2020.  Specifically, each participant was asked: How often do you wear a mask in public when you expect to be within six feet of another person? | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 8.3 | | 7.3 | 5.55 | 38.4 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Sometimes wear a mask in public | | | | |
| **Source(s)** | NY Times Github repository | | | | |
| **Description** | This data comes from a large number of interviews conducted online by the global data and survey firm Dynata at the request of The New York Times. The firm asked a question about mask use to obtain 250,000 survey responses between July 2 and July 14 2020.  Specifically, each participant was asked: How often do you wear a mask in public when you expect to be within six feet of another person? | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 12.1 | | 11.5 | 5.8 | 42 | .01 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Frequently wear a mask in public | | | | |
| **Source(s)** | NY Times Github repository | | | | |
| **Description** | This data comes from a large number of interviews conducted online by the global data and survey firm Dynata at the request of The New York Times. The firm asked a question about mask use to obtain 250,000 survey responses between July 2 and July 14 2020.  Specifically, each participant was asked: How often do you wear a mask in public when you expect to be within six feet of another person? | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 20.8 | | 20.4 | 6.4 | 54.9 | 2.9 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute Name** | % Always wear a mask in public | | | | |
| **Source(s)** | NY Times Github repository | | | | |
| **Description** | This data comes from a large number of interviews conducted online by the global data and survey firm Dynata at the request of The New York Times. The firm asked a question about mask use to obtain 250,000 survey responses between July 2 and July 14 2020.  Specifically, each participant was asked: How often do you wear a mask in public when you expect to be within six feet of another person? | | | | |
| **Mean** | | **Median** | **Standard Deviation** | **Max** | **Min** |
| 50.8 | | 49.7 | 15.2 | 88.9 | 11.5 |

# References

1. Patel, J. A., Nielsen, F., Badiani, A. A., Assi, S., Unadkat, V. A., Patel, B., Ravindrane, R., & Wardle, H. (2020). Poverty, inequality and COVID-19: the forgotten vulnerable. *Public health*, *183*, 110–111. https://doi.org/10.1016/j.puhe.2020.05.006 [↑](#endnote-ref-1)
2. Income and Poverty in the COVID-19 Pandemic Jeehoon Han, Bruce D. Meyer, and James X. Sullivan NBER Working Paper No. 27729 August 2020 JEL No. H53,I32,J65 [↑](#endnote-ref-2)
3. Published Online April 7, 2020 https://doi.org/10.1016/ S2468-2667(20)30084-0 [↑](#endnote-ref-3)
4. Parolin, Zachary, Megan A. Curran, Jordan Matsudaira, Jane Waldfogel and Christopher Wimer. 2020. “Monthly Poverty Rates in the United States during the COVID-19 Pandemic.” Poverty and Social Policy Discussion Paper. New York, NY: Center on Poverty and Social Policy. [↑](#endnote-ref-4)
5. Karaca-Mandic P, Georgiou A, Sen S. Assessment of COVID-19 Hospitalizations by Race/Ethnicity in 12 States. JAMA Intern Med. 2021;181(1):131–134. doi:10.1001/jamainternmed.2020.3857 [↑](#endnote-ref-5)
6. Abedi, V., Olulana, O., Avula, V., Chaudhary, D., Khan, A., Shahjouei, S., Li, J., & Zand, R. (2021). Racial, Economic, and Health Inequality and COVID-19 Infection in the United States. *Journal of racial and ethnic health disparities*, *8*(3), 732–742. <https://doi.org/10.1007/s40615-020-00833-4> [↑](#endnote-ref-6)
7. Impacts of the COVID-19 pandemic on rural America. J. Tom Mueller, Kathryn McConnell, Paul Berne Burow, Katie Pofahl, Alexis A. Merdjanoff, Justin Farrell. Proceedings of the National Academy of Sciences Jan 2021, 118 (1) 2019378118; DOI: 10.1073/pnas.2019378118 [↑](#endnote-ref-7)
8. Obesity Is a Risk Factor for Greater COVID-19 Severity. Feng Gao, Kenneth I. Zheng, Xiao-Bo Wang, Qing-Feng Sun, Ke-Hua Pan, Ting-Yao Wang, Yong-Ping Chen, Giovanni Targher, Christopher D. Byrne, Jacob George, Ming-Hua Zheng. Diabetes Care Jul 2020, 43 (7) e72-e74; **DOI:** 10.2337/dc20-0682 [↑](#endnote-ref-8)
9. Reddy, RK, Charles, WN, Sklavounos, A, Dutt, A, Seed, PT, Khajuria, A. The effect of smoking on COVID-19 severity: A systematic review and meta-analysis. *J Med Virol*. 2021; 93: 1045– 1056. <https://doi.org/10.1002/jmv.26389> [↑](#endnote-ref-9)
10. Vincent Chi-Chung Cheng, Shuk-Ching Wong, Vivien Wai-Man Chuang, Simon Yung-Chun So, Jonathan Hon-Kwan Chen, Siddharth Sridhar, Kelvin Kai-Wang To, Jasper Fuk-Woo Chan, Ivan Fan-Ngai Hung, Pak-Leung Ho, Kwok-Yung Yuen,

    The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2,

    Journal of Infection,

    Volume 81, Issue 1,

    2020,

    Pages 107-114,

    ISSN 0163-4453,

    https://doi.org/10.1016/j.jinf.2020.04.024.

    (https://www.sciencedirect.com/science/article/pii/S0163445320302358) [↑](#endnote-ref-10)
11. Gollwitzer, A., Martel, C., Brady, W.J. *et al.* Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nat Hum Behav* **4,**1186–1197 (2020). https://doi.org/10.1038/s41562-020-00977-7 [↑](#endnote-ref-11)
12. Ladan Golestaneh, Joel Neugarten, Molly Fisher, Henny H. Billett, Morayma Reyes Gil, Tanya Johns, Milagros Yunes, Michele H. Mokrzycki, Maria Coco, Keith C. Norris, Hector R. Perez, Shani Scott, Ryung S. Kim, Eran Bellin,

    The association of race and COVID-19 mortality,

    EClinicalMedicine,

    Volume 25,

    2020,

    100455,

    ISSN 2589-5370,

    https://doi.org/10.1016/j.eclinm.2020.100455.

    (https://www.sciencedirect.com/science/article/pii/S2589537020301991) [↑](#endnote-ref-12)
13. Di Xiong, Lu Zhang, Gregory L. Watson, Phillip Sundin, Teresa Bufford, Joseph A. Zoller, John Shamshoian, Marc A. Suchard, Christina M. Ramirez,

    Pseudo-likelihood based logistic regression for estimating COVID-19 infection and case fatality rates by gender, race, and age in California,

    Epidemics,

    Volume 33,

    2020,

    100418,

    ISSN 1755-4365,

    https://doi.org/10.1016/j.epidem.2020.100418.

    (https://www.sciencedirect.com/science/article/pii/S1755436520300396) [↑](#endnote-ref-13)
14. Wu, Zh., Tang, Y. & Cheng, Q. Diabetes increases the mortality of patients with COVID-19: a meta-analysis. Acta Diabetol 58, 139–144 (2021). https://doi.org/10.1007/s00592-020-01546-0 [↑](#endnote-ref-14)
15. https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey [↑](#endnote-ref-15)
16. https://www.youtube.com/watch?v=acSFjfyEx-o [↑](#endnote-ref-16)