Quinn Shim, David Ding
CS 396
Milestone 1

1. Summarize your project and data of interests

In this project, we will be exploring whether mask usage, Covid case, and Covid deaths are correlated with the political outcomes of the 2020 election. To do this, we will be looking covid and mask data for counties in America. For this project, we will be using the county name, population, id, and vote outcome from the 2020 election dataset. We will also be using covid cases and deaths per county as well as the mask usage percentage from the New York Times dataset.

This data will help us solve the problem in our proposal because by having Covid data in each county as well as the outcome of the election in each county, we will be able to make an assessment of whether Covid is correlated with the 2020 election.

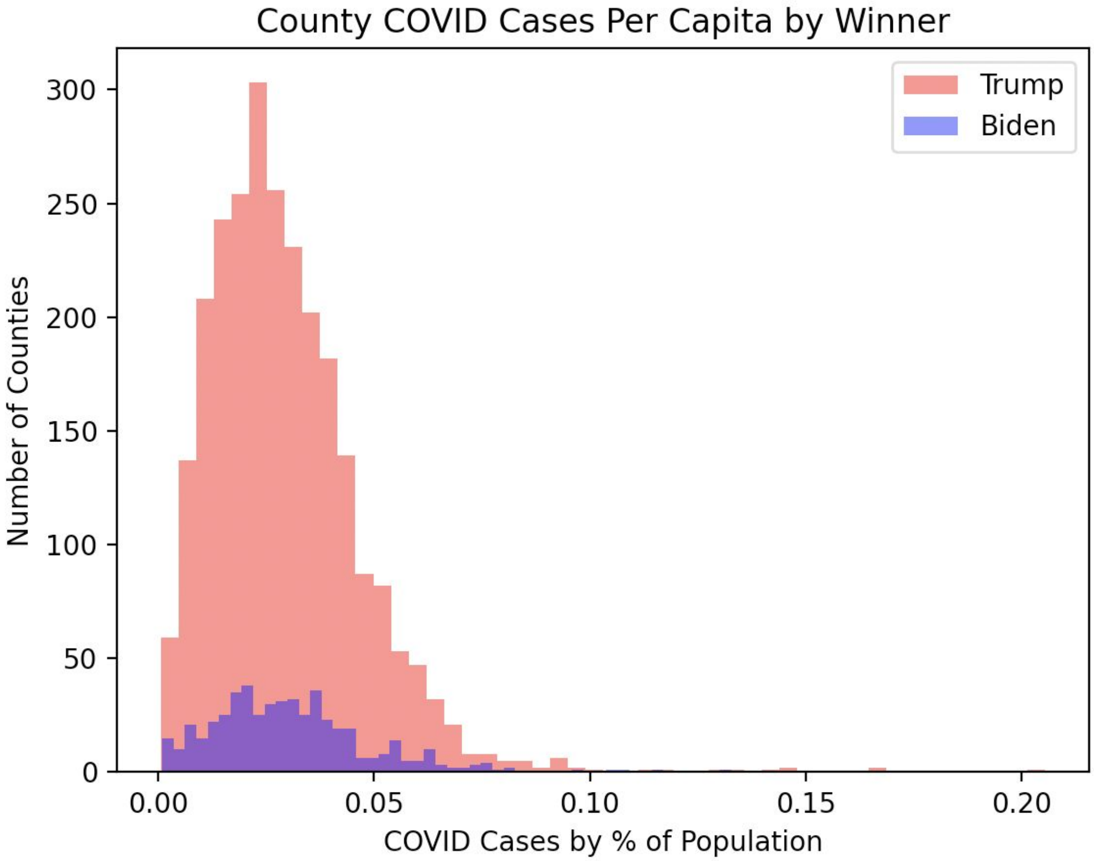
2. Data cleaning and management

For the data cleaning, we needed to remove lots of unnecessary data from the County Election Result dataset. It included many rows of data that would not be useful to us, such as longitude and latitude of the county. We also needed to reformat the state column for the election data to be consistent with the other datasets (i.e. turning "Colorado" into "CO"). For the covid cases and deaths dataset, we needed to grab only a day's worth of data for each county. For this, we picked 11/2/20 since we wanted the most up to date statistics before election day. Finally, we needed to clean the FIPS column so that we could easily combine the mask and cases datasets by their FIPS column. To merge the election dataset which did not have a FIPS column, we instead merged off of a county-state pair since both the election data and the covid cases data had a column with those attributes. Some data was dropped from the election dataset when merging, however this was due to certain counties being spelled differently and thus having no data. For example, certain counties had abbreviations in their name but had no data associated with them, so their rows were empty. These rows were at the bottom of the dataset, so they were dropped when merging.

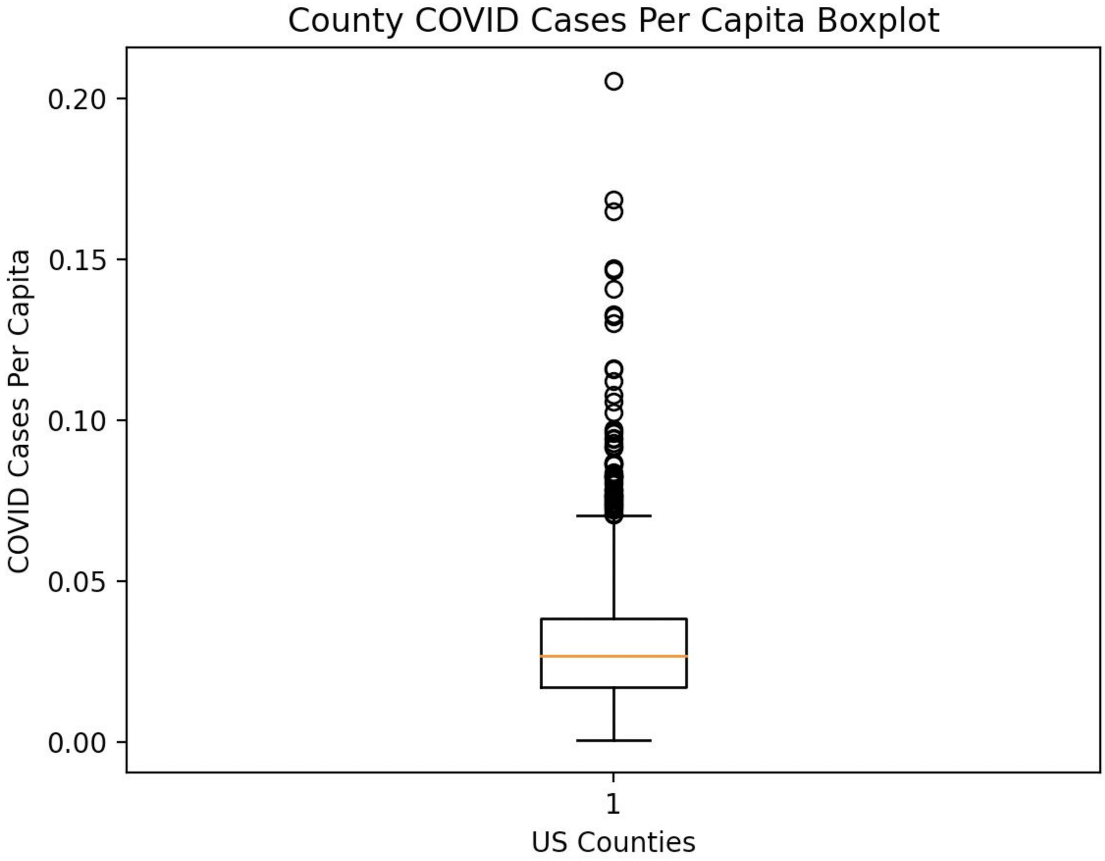
3. Explorative Data Analysis

We conducted four different analyses to get an initial understanding of our data as part of the EDA phase.

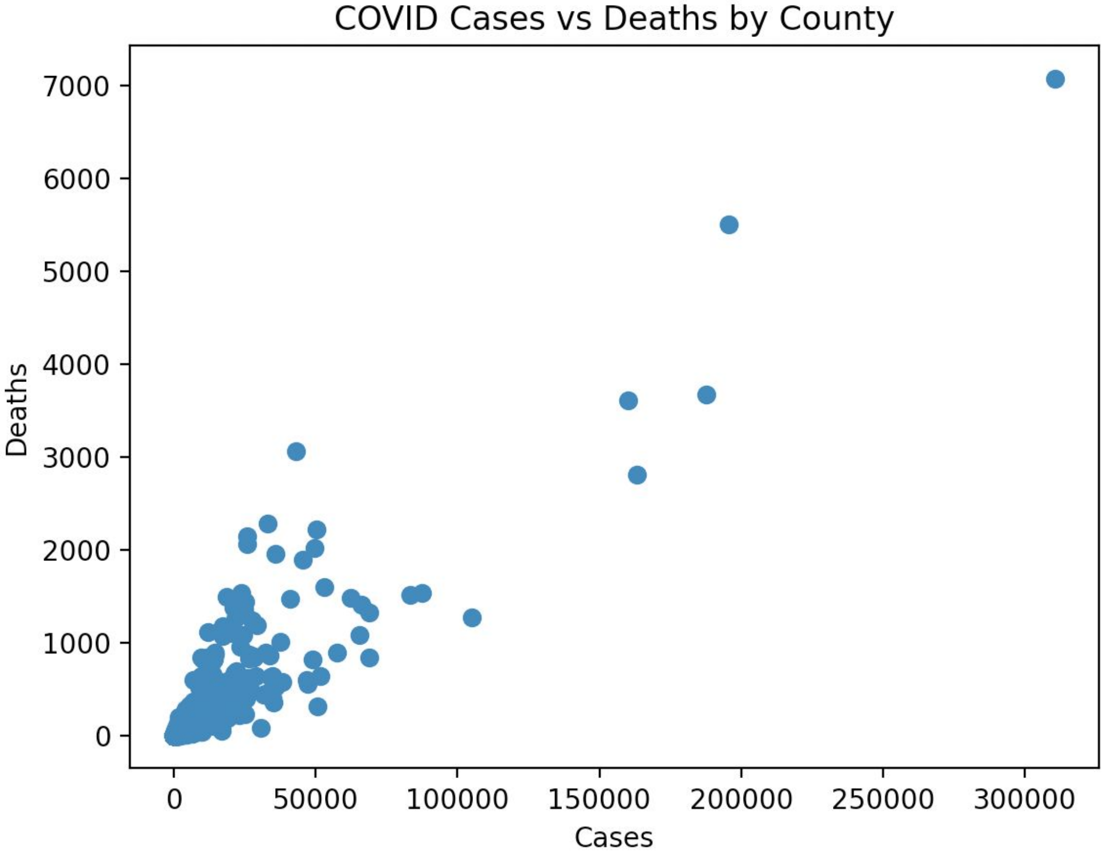
First, we graphed a histogram of COVID cases per capita in counties where Trump got the majority vote in the 2020 election and counties where Biden got the majority vote.



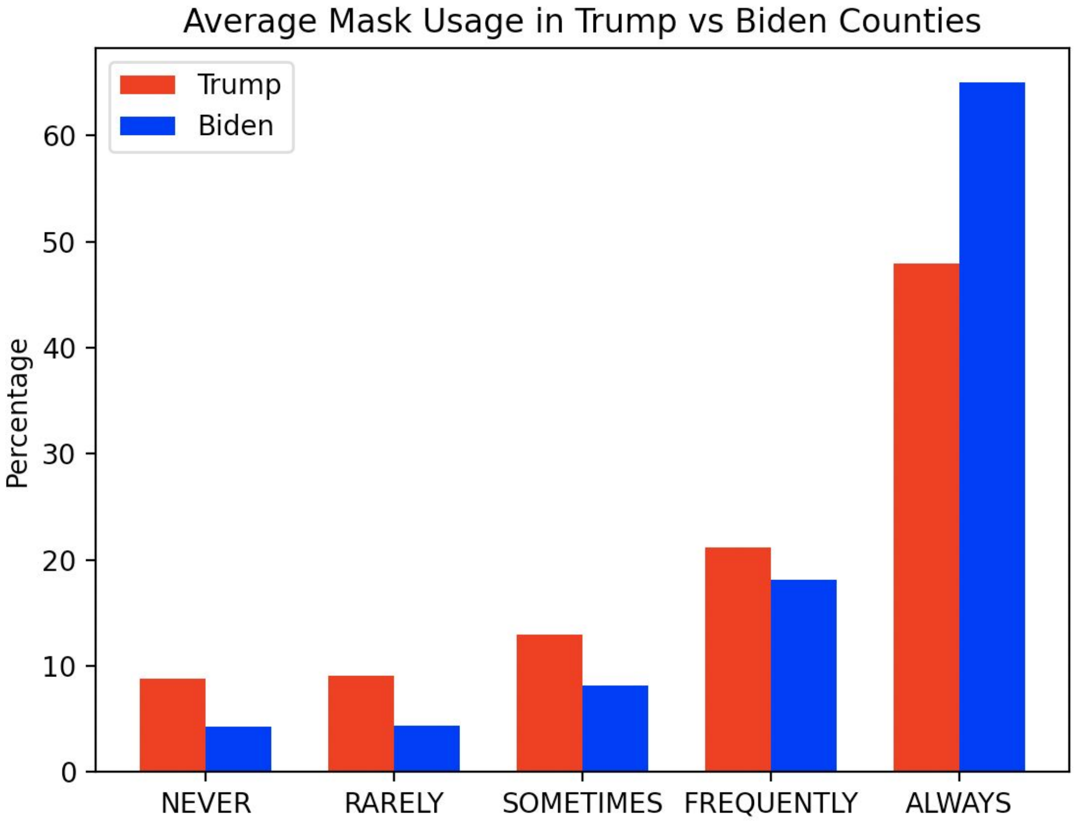
From this plot, we can see that the distributions for the two winning parties appear to be similar, with an average per capita case rate of around 3%. We can also see that there were more counties where Trump got the majority vote, and that the two distributions skew right with a few outlier counties that have a very high COVID case percentage.



Our second plot is a box plot that shows the IQR, median, and outlier statistics of COVID cases for all counties in the US. From this plot, we can confirm that the median percentage of COVID cases in US counties is around 2.7%, with a large number of outlier counties that have a high infection percentage.



Our third plot is a scatterplot that shows COVID cases against COVID deaths in US counties. As we can see, most counties have less than 100,000 cases and 3000 deaths. There also appears to be a positive linear correlation between the number of COVID cases and COVID deaths, which we expected.



Lastly, we plotted self-reported mask usage data. We averaged the data for all counties where Trump got the majority vote, and all counties where Biden got the majority vote. As we can see, “Always” is the most popular category when people reported how often they wear masks. We can also see that the distribution is similar for both categories, except Biden counties are more likely to always wear a mask compared to Trump counties.

From this exploratory data analysis, we were able to visualize some initial insights about our data, which will guide our decisions as we perform further analysis.