**Midterm**

PSCI 107

12 March 2021

21249856

**Question 1:**

**a.** I loaded the data and named the dataset “monmouth”. The unit of analysis is an individual respondent, and there are 867 observations in the data.

**b.** See R Script.

**c.** I started by recoding the relevant variables, *party* and *ideology,* to more descriptive names, excluding any missing values. For ideology, I consolidated “Very liberal” and “Somewhat liberal” into one label, “Liberal”, which was fitting for the purposes of my analysis. I then subset the Democrats and used *prop.table()* to find the proportions of Democrats who say they were “Liberal” (i.e. “Very” or “Somewhat” liberal). I converted this into percentages and found that 43.797% of Democrats say they are very or somewhat liberal.

Vice versa, I found that 87.817% of liberals (i.e. those who responded “Very” or “Somewhat” liberal) identified themselves as Democrats.

**d.** Similarly to **1c**, I recoded gender to “Man” and “Women” and subset the data into women and men. There were a higher percentage of women who reported increased daily stress levels during the pandemic: 61.416%, compared to 50.118% of men.

**e.** 27 states had 10 or more respondents. South Carolina had the highest percentage of respondents who were somewhat or very confident that the country would limit the pandemic’s impact, at 75%. Connecticut had the least percentage of confidence, at about 14% of respondents saying they were somewhat or very confident the country could limit the impact of COVID-19.

**Question 2:**

**a.** Unit of analysis is individual counties, and each observation/unit shows up only once in the data. The final 29 columns are the total numbers of COVID-19 cases in American counties for each day in April 2020.

**b.** I reshaped the data from wide to long format, so the new unit of analysis is the date, April 1 through April 30.

**c.** I cleaned the date variable, then made sure it was read as a date using *lubridate*. The number of cases in New York County in New York was 16,617 on April 15, 17,091 on April 16, 17,490 on April 17, and 17,932 on April 18.

**d.** The mean of 206.43 cases is substantially higher than the mean of 8 cases, which indicates that there are a number of high outliers which skew the data to the right. The plot confirms my suspicions, because there is a steep drop-off of density very close to 0, even though cases go up to 50,000. That means that there are a select number of counties which have really high COVID-19 cases, thus dragging up the mean.

Chart

Description automatically generated

**e.** I found the mean to be .81 cases per 1000 residents, and median to be .302 cases per 1000 residents.

Chart

Description automatically generated

The above graph does give a somewhat better understanding, because the mean is not so drastically different than the median, and the peak density is more visible. The data Is still very right skewed. This leads me to believe that the spread of COVID-19 was not that much in April, because most of the counties in the data did not experience more than 5 cases per 1000 residents throughout the month. There were a couple counties, though, that sustained very high numbers of COVID cases.

**f.** I used a *for* loop to return the median number of COVID cases per 1000 residents over the month of April. I then created a scatter plot with the date on the x-axis and the median number of cases per 1000 residents on the y-axis.

Chart, scatter chart

Description automatically generated

This graph has a surprising, very strong correlation between later dates and the median number of cases per 1000 residents. This makes sense, because the virus spread rapidly in March/April and the numbers were growing daily.

**g.** After subsetting the data to include only the data for April 30, I used separate *for* loops to find the median population density and median cases per 1000 residents for each state. The scatterplot is below:

**Chart

Description automatically generated**

From this graph, it appears that the median cases increase as the median population density increases. So when population density increases, the COVID case incidence seems to increase.

**h.** The obvious outlier in **2g** was DC, so I removed DC and recreated the scatterplot below:

Chart, scatter chart

Description automatically generated

Below is a close-up of this graph, when I narrowed the *xlim* and *ylim*:

Chart, scatter chart

Description automatically generated

From these graphs, the relationship between population density and COVID case incidence still seems to have a positive correlation, but the relationship is not as strong as **2g.** This relationship can be misleading, though, because many of the states have a median case incidence near 0 per 1000 residents, where the population density has little to no relationship. The positive relationship may not actually be as strong as it seems on the graph when you consider certain ranges of population density.

**i.** After creating a double loop to determine the within-state correlation between population density and median cases per 1000 residents, I created the below box plot for every state (excluding DC):

Chart, scatter chart

Description automatically generated

According to this plot, there is a very low correlation between population and median COVID case incidence for each state. This leads me to a similar conclusion in that **2h** may be misleading; this plot shows there is essentially no positive relationship between these two variables per state.