**Problem Set 1**

PSCI 107

17 February 2021

21249856

**Question 1:**

**a.** I created three vectors *state*, *pop*, and *cases* using the concatenate function *c().* Please refer to the R script.

**b.** I combined the three vectors in 1a into the object *ne.covid* using the function *cbind().* Please refer to the R script.

**c.** In order to find some descriptive statistics on the second column in *ne.covid* (i.e. *pop* or population), I used the R functions *mean(), median(), min(),* and *max().* Please refer to the R script.

**d.** Similarly to 1c, I used the R functions *mean(), median(), min(),* and *max()* on the third column of *ne.covid*, which is *cases*. Please refer to the R script.

**e.** In order to create a new vector, *cases.per.1000*, I first created a vector *pop.1000*, which converts *pop* from state population in millions to state population in thousands. I just used simple arithmetic for this (refer to R script). Then I divided *cases* by my new vector *pop.1000* to create *cases.per.1000*. This final vector will give me the number of COVID cases per 1000 state residents.

**f.** To create a scatter plot, I used the *plot()* function, with *pop.1000* (Population in thousands)on the x-axis and *cases.per.1000* (Cases per thousand)on the y-axis. Here is the plot:

Chart, scatter chart

Description automatically generated

(Bonus): Here is my bonus plot, in which I used the *text()* function to label the points with the state abbreviations; I created a new vector *state.names* in order to do this. I also set the x limit with *xlim* to better fit the plotted points:

Chart

Description automatically generated

**g.** According to the above scatter plot, it seems like New York (NY) is the most affected by COVID-19; it is the state with the highest case rate per 1000 residents, and it has the highest population.

it seems like Pennsylvania (PA) is the least affected by COVID-19; despite having the second-highest population, it is the state with the lowest case rate per 1000 residents.

**h.** If I could get one more piece of information to determine how affected these states are by COVID-19, I would like to know the population density of each state−−that is, the population of the state divided by its area. Since COVID-19 is a virus, it spreads more easily in densely populated areas. So, perhaps a state with a high population density could be more affected by COVID-19.

**Question 2:**

**a.** I created five vectors *age, gender, hours.daily, pref.network,* and *much.time* using *c()*. Please refer to the R script.

**b.** I created an object *soc.media* using *cbind.data.frame()*, since the class is not the same for each variable. The class of the variables are numeric, character, numeric, character, and logical, respectively.

**c.** I used the *dim()* command to find how many men spend 3 or more hours per day on social media, and 5 men do in this dataset. I also found that their preferred networks are TikTok, Twitter, Twitter, Instagram, and Facebook, respectively.

**d.** Similarly to 2c, I used the *dim()* command to find how many women over 40 whose preferred networks are Facebook or Instagram, and 3 women meet these criteria.

**e.** I subset the data to include only those who think they spend too much time on social media (*much.time == TRUE).* There are 7 people in this subset *soc.media.much*.

**f.** I used the *plot()* command to create a scatter plot of ages on the y-axis by daily social media use (in hours) on the x-axis for *soc.media.much.* I set the color to “darkblue”, which will be relevant later on:

Chart, scatter chart

Description automatically generated

**g.** Again, I used the *plot()* function to create a scatter plot for the whole dataset, with age on the y-axis and daily social media use (in hours) on the x-axis. I used “red” for these points:

Chart, scatter chart

Description automatically generated

**h.** Comparing the plots I created for the subset (2f) and the whole dataset (2g), it seems like the subset plot has a narrower x and y default scale. For instance, the x scale for plot 2f is from 1 to 6 hours, while the x scale for plot 2g is from 0 to 6 hours. Likewise, the y scale for plot 2f is from 20 to 55 years old, in increments of 5, while the y scale for plot 2g is from 20 to 60 years old, in increments of 10.

For plot 2f, there is no clear relationship between age and daily social media use, but for plot 2g, there seems to be a moderate negative relationship between age and daily social media use. So, those with the highest social media use tend to be lower in age.

From these observations, I can tell that the subset is narrower in scope compared to the whole dataset; people who spend 0 hours on social media, for instance, are very unlikely to say that they spend too much time on social media. Thus, the subset plot does not need to include 0 hours in the x scale. Furthermore, the subset was not as conclusive as to a relationship between age and daily social media use; with so few data points, there was no clear trend among them. This contrasts to the perceptible relationship with the whole dataset.

(Bonus): Below is one plot that visualizes and highlights the differences between the two relationships. I combined the plots using the *points()* command, and I also created a legend using the *legend()* command. The subset points are in dark blue, and the whole dataset points are in red:

Chart, scatter chart

Description automatically generated

**Question 3:**

Random sampling is a vital component of survey research because it allows the sample to adequately represent and apply to the whole population in question. If the sample were *not* randomly selected from the population, then it could not be generally applied; such samples could have certain characteristics which might not represent the population at large. Especially for very large populations, random sampling ensures that the researcher has a representative “snapshot” without impractically surveying every single person in that population.

If random sampling is truly random (that is, everyone in the population is equally likely to be chosen), then a relatively small sample can be sufficient to measure a very large population. Furthermore, the margin of error is determined by sample size, not population size; a sample of around 1000 to 2000 respondents will sufficiently limit the margin of error for almost any population. That is why a 1500-person sample can be enough to gauge the whole U.S. population.