**Statistical Question**: At the start of a national pandemic, will people search more federal government websites to try to find more information out more about COVID-19 and the stimulus bill than after a certain point when people start adapting to the economic unknowns?

**Hypothesis**: I believe people will want to find out more about health issues related to COVID-19, the stimulus bill (i.e. stimulus checks, aid for small businesses), and every day uses (i.e. USPS) at the beginning of the national pandemic than after a certain point of time.

**Summary**

Before I started doing more of my statistical analysis, I had to do a little cleanup on the datasets before I joined them together. There were over 2,000 records and I only wanted to analyze the top 50 records from the 4/5/20 data pull with the second dataset matching the domains of the first. After I joined the two datasets on the Domain (website), I removed the last part of the extra rows that did not include the first 50 Domains (empty information for the 4/5/20 data pull). Then, I had to update the dates, remove unwanted fields, add zeros where there was NaN, and calculated the difference of variables between the 4/5/20 and 5/17/20 data pulls. What surprised me about reviewing the initial data after joining the data, was some of the websites that people were going to such as ncbi.nlm.nih.gov (looks like it may point to different links where data is pulled from) and sa.www4.irs.gov (Access denied message).

While reviewing the histograms, there were some outliers, but I believe they should be kept. If I was evaluating more data in the datasets, the outliers would not appear as outliers. When I reviewed the means, variance, standard deviations, and mode, I noticed most of the numbers decreased between the different periods of time. I used the Page Views Per Session variables from the two data pulls while analyzing the PMF. There was not much of a difference between the means so the graphs look pretty much identical. The CDF graph curves from the bottom right towards the top of the right-hand side. It looks there is a fairly high likelihood the average session duration will be 400 or more.

The type of analytical distribution that I used was log normal. I based the data on page views per a session for data pull 4/5/20. It looks like it has a pretty close to normal distribution as the line curves from the upper left-hand side and slopes down to the bottom of the graph.

After reviewing the VisitsDiff and ExitsDiff scatter plot, most of the dots are centered around the zero area of the graph. For the most part, there will be a correlation between these variables as people have to go to the websites and exit them. On the next scatterplot, I analyzed the avg\_session\_duration1 and avg\_session\_duration2. There is a small correlation based on events that were going on during certain periods of time but the duration of staying on websites, will vary.

I used the Pearson’s Correlation test to test the hypothesis. I used the variables pageviews\_per\_session1 and pageviews\_per\_session2. The correlation test showed that the variable will most likely be dependent.

During the analysis, I believe more data should have been used. I would have been able to see more of the trends and I could have potentially created more variables that looked at the data a different way. Some variables that could have helped include breaking up the data based on certain parameters such as above a number of page views, below page views, number of visits and exits above a certain amount and below a certain amount, and the percentage of pages views and exits based on the totals. I do not think there were really any assumptions that I made that were incorrect. After comparing the datasets, it was noticeable how there were not as many people looking at certain websites than when the first dataset was pulled.

Overall, this was a great project to apply my newly acquired skills and translating the data from what was discussed in the class throughout the session. There were some challenges such as trying to find a dataset that would work with what we were learning. Another challenge was finding ways to apply the information and put it into an understandable format that others could see.