For my project I focused on the data about baby names throughout the years. I went with this topic/dataset as I have always found it interesting to hear about the most popular baby names each year and how different they are, how common they are, etc.

In my research, I focused mostly on the years and the occurrences for the years. The first challenge I was stopped by was the amount of data that I had in the file and then the categorical versus numerical variables. I did not immediately notice any outliers for my main variables of state, gender, year, name, and number. Another challenge I was faced with as I went through each step of the project was answering my questions. While I originally thought that the variables I was working with were sufficient enough to answer some of the questions I had, I quickly realized that I would have like to have worked more with numerical data as it seemed easier to handle and analyze. Some of my original questions were:

1. What are the most common male and female names throughout the years?
2. Which year had more male's vs females?
3. Over all the years, are there more males or females?
4. Which states have more males? Females?

I realized that not all of the questions could be answered immediately and tried to approach it in a different way. I came up with the question of what year had the most name occurrences?

While working on my exponential and normal distribution plots, I noticed that neither was accurate and gave any more insight than what I could see before. I create two scatter plots next and one was with Birth Year and Occurrences while the other one was Sex and Birth Year to get a better look at where the information fell. In the scatter plots, I could see that a lot of the information overlapped within the points.

I went with a smaller data frame after that. I only looked at the data given from years 2017 and 2018. There was still a lot of data; however, it did seem more manageable. I calculated the covariance and correlation with the limited dataset. I determined the covariance to be -.13 and the correlation to be -.003. This showed that it would be a negative relationship between the number of occurrences and the year of birth and they are not very closely related.

I did the hypothesis testing with the variables of gender and number of occurrences. It was determined that there was no relationship between the gender of the baby and how many times either male or female had occurred within the year.

Overall, with this particular dataset, I couldn't come up with a direct answer and the information my questions remained unanswered. Throughout the analysis, I could determine direct relationships as well as indirect. It can additionally be determined that there is not either year that shows more females were born or more males. They are statistically about the same.