Final Project Summary

For my Project I choose the data related to the ranking of military strength of the country. My null hypothesis is that there is no relationship between defense budget and total military personnel for a country. In my EDA process the correlation of the variable that I choose was not very strong, it was good to look after the scatter plot, histogram, pmf, cdf and hypothesis testing. While these variables are correlated, we cannot tell if one of them is causing another one to increase. We need more analysis to find the correlation and causation. I was able to find out there was outliers in my data set with the help of histogram. Looking at the PMF we see that the there is more chance of seeing that most of the country has the total military which less than 0.5 million. There is a greater number of countries that has a smaller number of military than the country that have high number of military personnel.

The CDF for the Total\_Military\_Personnel variable shows that the most common number is around the left side of the graph. Most country has the value less than 0.25 million so I would say the common number is 0.25 million. The highest value is around 4-5 million. The curve for the total military personnel deviates from the expected model because most of our data are concentrated in the left and there are very few countries in the right. Since we have very few records in the right the model and the actual curve are not lining together. The curve for the total military personnel deviates from the expected model because most of our data are concentrated in the left and there are very few countries in the right. Since we have very few records in the right the model and the actual curve are not lining together.

A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. The correlation calculated is 0.29 It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Here we have p value equal to 0.03, that means we have strong evidence against the null hypothesis hence there is relationship between the total military personnel and the defense budget.

The higher the R-squared, the better the model fits our data. R square indicate the percentage of the variance in the dependent variable that the independent variables explain collectively. Here we see that R squared value to be 0.09 that means 9% variance is done by total military personnel count towards the defense budget. This means that military personnel count only counts for 9% relation towards the defense budget, and we can tell that there are other factors that affects the defense budget like the military combat equipment, maintenance, and mission.

The dataset has the top ranked countries to have the value variables in very high range in all the variables. There are some outliers in our dataset. This includes around ten countries out of 133 so this might affect the model that we try to predict. I am not removing them to see how it affects the model and they also have most significant data that determines the military strength ranking of the country. I feel like the model prediction was not very strong with only 9% variability of the independent variable. This could be because we included the outliers in our analysis which threw us from predicting the model. If I had chosen multiple regression rather than just one variable, then the model would be predicting to give us the intended result. While I was trying to draw histogram using thinkstat2 package my graph was not showing anything, so I needed to google for other options to display the histogram. Also, I was not sure why the R squared was very low for my analysis. Overall, I was able to learn many new concepts in statistics with some challenges that is needed for a data scientist.