**5-2 Milestone Two: Data Validation and Discovery**

Michael Surdek

Southern New Hampshire University

DAT-500 Data & Information Management

Dr. Arman Kanooni

July 26, 2020

**Summary:** In RStudio, I used the summary function to get the descriptive statistics for all of the variables from both the Firefighter and Police data. This was easier to perform and outputs much more information than just calculating the min, max, and mean of two variables in Excel, like we did in Milestone 1. One calculation of the summary function that provides more information for this analysis is the median of each variable. For example, I noticed that the median amount of overtime compensation is lower than the mean for both police and firefighters. On the other hand, the median amount of total compensation is higher than the mean for both. This gives me additional insight for the shape of the distributions of these variables. In Milestone One, the first “next step” that I identified was to calculate the medians, so my findings of the summary function did not change my plans, but it allowed me to continue with my analysis of the data.

**Data Validation:** It is ideal that the results from all 3 methods of calculating the min, max, and mean of the variables turned out to be identical. In my opinion, these results confirm what I found in Milestone One, which was that the min, max, and mean of firefighters’ overtime and total compensation amounts are higher than those of police officers. One avenue that I would now like to pursue would be to visualize the distributions of these variables. I would like to create histogram or density plots to see how the overtime and total compensation variables are centered and spread.

**Data Discovery:** When I calculated the min, max, and mean, I found that the firefighter who earns the most overtime earns over $40,000 more in overtime than the police officer who earns the most overtime. I also found that the average firefighter earns nearly triple the amount of overtime dollars than the average police officer. This led me to wonder if the overtime earnings are consistently higher for firefighters than police officers at every percentile of earner. From the output of the summary function, I was able to determine that firefighters in the 25th, 50th, and 75th percentile of overtime compensation earn more than police officers of similar earning percentiles. This also happens to be true for the total compensation variables as well. At this point, I can safely say that the firefighters in the dataset received more total compensation at every percentile of earnings. In order to dig deeper into the data, I would like to dig deeper into some of the individual data points to make sure that this analysis is being done properly. It concerned me that the minimum amounts of total compensation are only $12 for firefighters and $100 for police. Obviously, the individuals behind these earnings would not be considered to have worked full time. I am not sure exactly how to do it, but one of my next steps would be to try to only compare firefighters and police officers who work full time.