Michael Surdek 3-2 Milestone One: Data Analysis Plan

The dataset in the Milestone One excel file appears to contain information regarding compensation for hundreds of firefighters and police officers. The first worksheet contains information on 912 firefighters and the second on 889 police officers. The Total Compensation is broken down into Salary and Benefits. Salary consists of general salary, overtime, and other salaries. Benefits consists of retirement, health/dental, and other benefits. This data will allow us to analyze the distributions of each element of compensation and to compare firefighters and police officers.

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Job** | **Overtime** | **Total Compensation** |
| Mean | Firefighter | $24,583.63 | $173,389.87 |
| Minimum | Firefighter | $0.00 | $12.27 |
| Maximum | Firefighter | $132,728.94 | $316,195.42 |
| Mean | Police Officer | $8,965.23 | $110,966.88 |
| Minimum | Police Officer | $0.00 | $100.62 |
| Maximum | Police Officer | $88,976.01 | $277,685.39 |

I chose to calculate the mean, minimum, and maximum statistics for the columns Overtime and Total Compensation. I chose the Overtime column because I know some police officers make a large portion of their income from overtime and I wanted to see if it is similar for firefighters. My findings show that the firefighter who earns the most overtime earns over $40,000 more in overtime than the police officer who earns the most overtime. My findings also show that the average firefighter earns nearly triple the amount of overtime dollars than the average police officer. This result is interesting to me and leads me to wonder if the overtime earnings are consistently higher for firefighters than police officers at every percentile of earner. I chose to calculate the statistics for the Total Compensation column because I wanted to see more information on the full amount of salary plus benefits for these occupations. My findings show that the mean, minimum, and maximum of total compensation is higher for firefighters than police officers. When I calculated these statistics, I noticed that the minimums for total compensation are only $12 for a firefighter and $100 for a police officer. These results might indicate that there is missing data or obscurities in the dataset or that there are some employees who only worked a small number of hours in the given year.

My first impressions and initial analysis have put me in a good position to further analyze the data. One of the first steps I would take would be to calculate the median of each column. Although the mean can be a good measure of center, I think the median might be a better indicator if there are outliers in the data such as a firefighter who only earns $12 in a year. Another next step I would take is to double check the Employee Identifier column for any duplicates. It is possible that there were mistakes when recording some of this information and finding duplicates could be a simple way to identify that issue. Finally, the last next step that I have in mind would be visualize the data in each column with histogram or density plots to get a better sense of the full distributions. By doing this for the total compensation column, I could find out if firefighters at each level can expect to earn more than police officers at similar levels. This would provide more information on top of descriptive statistics such as the mean, minimum, and maximum, and it would be a good way to move into the next phase of the project which is data validation and discovery.