Brendan Golden

Module 5 Analysis

This module introduces operations within Python concerning plotting a multitude of graphs and implementing statistical methods. These concepts are used for calculating central tendencies, conducting normal tests, determining standard deviation and variance, manipulating data sets into samples, and ultimately combining the two units into one cohesive project. The introductory portion of the assignment required applying basic concepts of merging data sets, with some minor data wrangling to deduce a usable data frame for further calculations. We were required to apply concepts learned in class to calculate the necessary statistical values of mean, median, variance, standard deviation, and standard error and combine them into a uniform, summarized data frame. The next part of the assignment required the application of procedures within Pandas and Matplotlib to construct two sets of bar graphs and pie charts. We were required to apply basic and advanced concepts to intuitively produce the graphs, based on matters that were initially presented during lectures. Moving into the more analytical portion of the module we practiced filtering data structures based on specific columns using the group by function, and once again merging two distinct data sets into a more relevant, applicable table. We then needed to go back to concepts covered primarily during the first Python module, when we were introduced to lists and for loops, to calculate the quartiles, the interquartile range, lower and upper bounds, and the possible outliers in the pharmaceutical data. We were able to find out, with the use of a box plot, that one of the drugs, Infubinol, did have a statistically significant outlier as it was located outside of the lower bound. The penultimate part of the assignment required us to sort and filter the data based on a specific column, as we needed to visualize the relationship between the tumor volume over time for a specific mouse in the data. As we filtered and merged data we calculated the average weight of the tumor, for all of the mice, and plotted a scatter plot, that seems to be positively correlated as the weight of the mice tends to increase the size or volume of the tumor also tends to increase. The final component of the module introduced the ability to calculate the actual correlation coefficient, which was nearly one indicating there is a strong correlation, and the ability to plot a linear regression line, ultimately giving us, as data analysts, the ability to predict data for the future. I enjoyed this module as I felt that it was extremely applicable to real-world examples, as I can see myself using these skills and methods to be able to produce aesthetically pleasing graphs, to be able to help determine if there is a correlation between a dependent and independent variable and to be able to use regression methods to accurately help predict linear relationships in the future for all types of data.