Alex Arnold
SMU Data Science Bootcamp
November 2, 2020

Matplotlib Homework - The Power of Plots

Purpose:

While your data companions rushed off to jobs in finance and government, you remained adamant that science was the way for you. Staying true to your mission, you've joined Pymaceuticals Inc., a burgeoning pharmaceutical company based out of San Diego. Pymaceuticals specializes in anti-cancer pharmaceuticals. In its most recent efforts, it began screening for potential treatments for squamous cell carcinoma (SCC), a commonly occurring form of skin cancer.

As a senior data analyst at the company, you've been given access to the complete data from their most recent animal study. In this study, 249 mice identified with SCC tumor growth were treated through a variety of drug regimens. Over the course of 45 days, tumor development was observed and measured. The purpose of this study was to compare the performance of Pymaceuticals' drug of interest, Capomulin, versus the other treatment regimens. You have been tasked by the executive team to generate all of the tables and figures needed for the technical report of the study. The executive team also has asked for a top-level summary of the study results.

Data:

Merged Mouse Data with Study Results

Analysis:

- Clean Mouse Data by removing duplicated Mouse ID.
- Generate a summary statistics table consisting of the mean, median, variance, standard deviation, and SEM of the tumor volume for each drug regimen.
- Generate a bar plot using both Pandas's DataFrame.plot() and Matplotlib's pyplot that shows the number of total mice for each treatment regimen throughout the course of the study.
- Generate a pie plot using both Pandas's DataFrame.plot() and Matplotlib's pyplot that shows the distribution of female or male mice in the study.
- Calculate the final tumor volume of each mouse across four of the most promising treatment regimens: Capomulin, Ramicane, Infubinol, and Ceftamin. Calculate the quartiles and IQR and quantitatively determine if there are any potential outliers across all four treatment regimens.
- Using Matplotlib, generate a box and whisker plot of the final tumor volume for all four treatment regimens and highlight any potential outliers in the plot by changing their color and style.

- Select a mouse that was treated with Capomulin and generate a line plot of tumor volume vs. time point for that mouse.
- Generate a scatter plot of mouse weight versus average tumor volume for the Capomulin treatment regimen.
- Calculate the correlation coefficient and linear regression model between mouse weight and average tumor volume for the Capomulin treatment. Plot the linear regression model on top of the previous scatter plot.

Observations

Capomulin and Ramicane are the most successful drug regimens in their effectiveness of decreasing tumor volume as identified with final tumor volume (Figure 1).

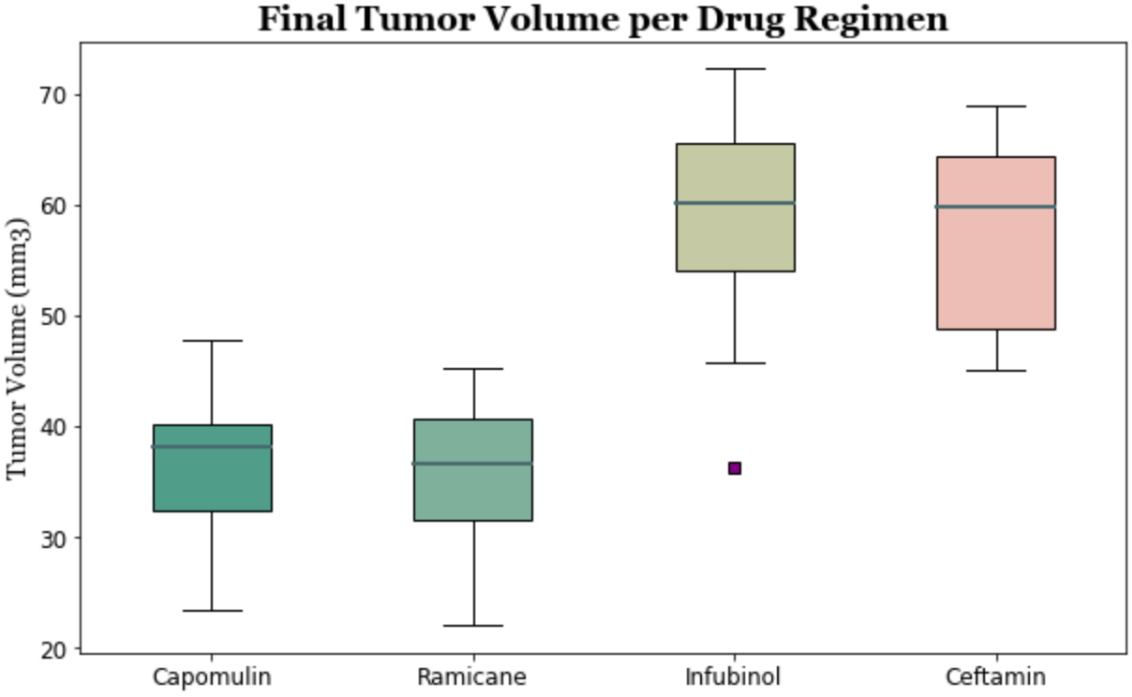

Figure 1. Final Tumor Volume per Drug Regimen

Capomulin and Ramicane have the greatest number of mice per timepoints throughout the study, highlighting the effectiveness of the treatments over time (Figure 2).

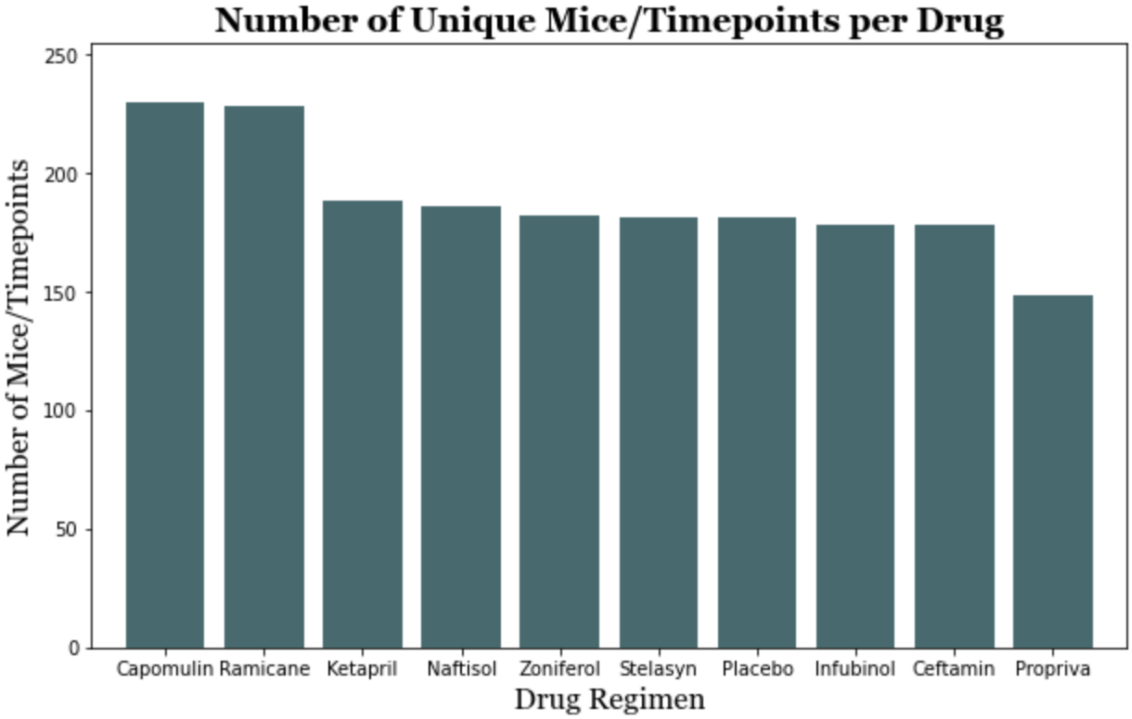

Figure 2. Number of Unique Mice/Timepoints per Drug

There is a positive correlation between Capomulin Mice Weight and Tumor Size (Figure 3).

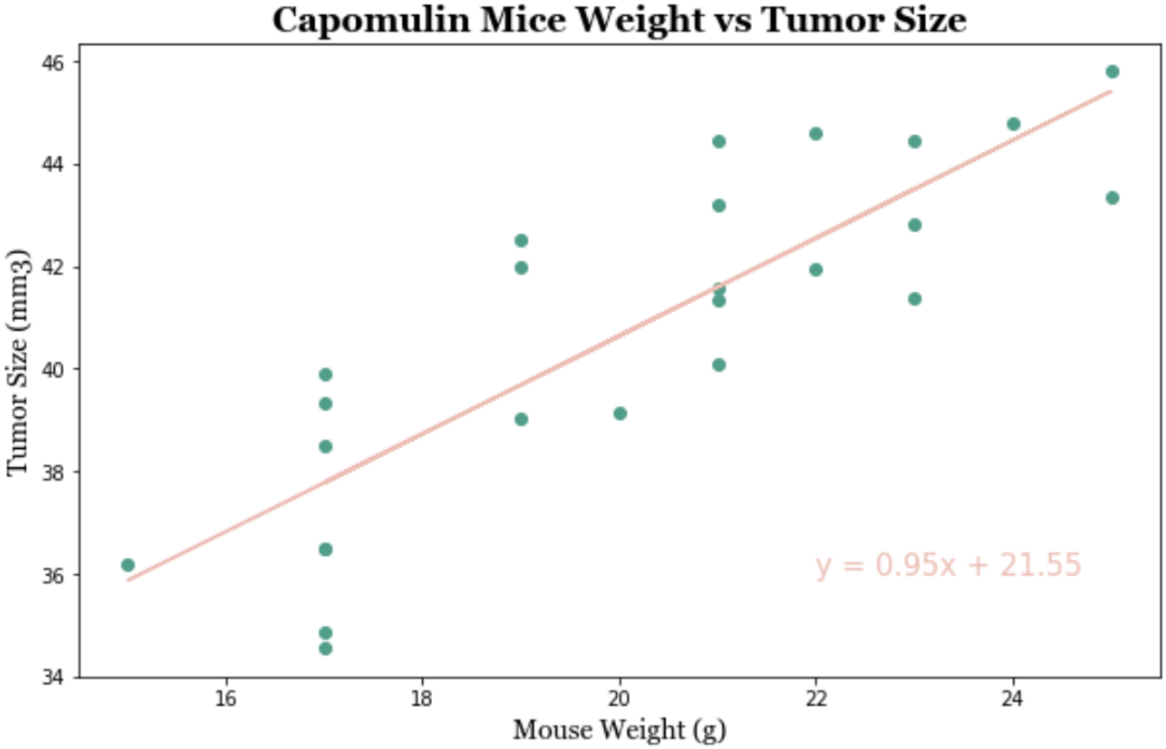

Figure 3. Capomulin Mice Weight vs. Tumor Size