What good is data without a good plot to tell the story ?

I have just joined Pymaceuticals, Inc., a new pharmaceutical company that specializes in anti-cancer medications. The company recently 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, I've been given access to the complete data from their most recent animal study. In this study, 249 mice who were identified with SCC tumors, received treatment with a range 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, against the other treatment regimens.

The executive team has tasked me with generating all the tables and figures needed for the technical report of the clinical study. They have also asked for a top-level summary of the study results.

To achieve this objective, I will be running statistical tests to look for correlations between data sets and creating plots to help visualize my findings.

**Tools**

I will be using Jupyter Notebook/Python, Pandas in Python, Matplotlib and scipy to display data for my pharmaceutical company.

**Project Steps**

Step 1: Prepare the data

Two data sets were provided in Excel format, one showed the metadata for each mouse and the other showed the study results.

*Looking for Duplicates:* I identified and extracted data for Mouse g989 as it had duplicate entries.

Step 2: Generate summary statistics

I generated a summary statistics table of mean, median, variance, standard deviation, and SEM of the tumor volume for each regimen.

A screenshot of a computer screen

Description automatically generated

Step 3: Create bar charts and pie charts

*Bar Chart* - Generated a bar plot showing the total number of rows (Mouse ID/Timepoints) for each drug regimen using Pandas. This showed the total number of timepoints for all mice tested for each drug regimen.

A graph of a number of drugs

Description automatically generated

The bar graphs showed that the Drug Regimen Capomulin had the maximum number of mice and Propriva had the smallest number of mice.

Next, I generated a pie plot showing the distribution of female versus male mice using Pandas

A blue and orange circle with text

Description automatically generated

The 2 pie charts showed that 51% of mice tested were male and 49% were female.

Step 4: Calculate quartiles, find outliers, and create a box plot

Calculated the final tumor volume of each mouse across four of the treatment regimens: Capomulin, Ramicane, Infubinol, and Ceftamin

A screenshot of a computer screen

Description automatically generated

*Boxplots* - Generated a box plot to show the distribution of the tumor volume for all the mice in each treatment group.

A diagram of a number of boxes

Description automatically generated

The box plot indicated an outlier: the Infubinol drug. An outlier is a data point that is significantly different from the rest of the data i.e. they could represent errors or unusual observations. However, not all extreme values are necessarily errors; they sometimes represent valid data points that deviate from the central trend.

Step 5: Create a line plot and a scatter plot

*Line Graph* - Generated a line plot of tumor volume vs. time point for a **single** mouse treated with Capomulin.

This can be easily used to show results for any given mouse.

A graph of a line

Description automatically generated

The line plot for Capomulin treatment of mouse I509 showed the tumor volume still increasing steadily up to Day 20, after which it dropped up to Day 25. Then the tumor started to increase again for 5 days followed by a significant drop up to Day 35. From Day 35, the tumor started increasing.

*Scatter Plot* - Generated a scatter plot of mouse weight vs. the average observed tumor volume for the entire Capomulin regimen

A graph of blue dots

Description automatically generated

Step 6: Calculate correlation and regression

Calculated the correlation coefficient and a linear regression model for mouse weight and average observed tumor volume for the entire Capomulin regimen.

A red line with blue dots

Description automatically generated

The correlation between mouse weight and the average tumour volume is 0.84.

An R-value of 1 indicates a perfect positive linear relationship, while an R-value of -1 indicates a perfect negative linear relationship. An R-value of 0 suggests no linear relationship.

The regression analysis showed us how much the average tumor volume changed when the weight of the mouse changed. The R-value was 0.84, which means 84% of the model fit the data. This is fairly good to predict the data from the model. The heavier the mouse, the greater the tumor volume.

In this instance, the correlation between mouse weight and average tumor volume indicates a moderately strong positive linear relationship between the mouse weight and average tumor volume.

This means that the lines are reasonably good fits for the data, and can be used to make predictions about weight versus tumor volume.

Our analyses showed that the Capomulin and Ramicane drugs reduce the size of tumors best whilst the Propiva drug causes more harm than good.

Capomulin would be the preferred drug.