Matplotlib Homework - The Power of Plots

Background

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

So, let's take what you've learned about Python Matplotlib and apply it to a real-world situation and dataset:

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.

Instructions

Your tasks are to do the following:

- Before beginning the analysis, check the data for any mouse ID with duplicate time points and remove any data associated with that mouse ID.
- Use the cleaned data for the remaining steps.
- 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 total number of measurements taken for each treatment regimen throughout the course of the study.
    - **NOTE:** These plots should look identical.
- 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.
    - **NOTE:** These plots should look identical.
- 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.

  Hint: All four box plots should be within the same figure. Use this Matplotlib documentation page (https://matplotlib.org/gallery/pyplots/boxplot_demo_pyplot.html#sphx-glr-gallery-pyplots-boxplot-demo-pyplot-py) for help with changing the style of the outliers.
- 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.
- Look across all previously generated figures and tables and write at least three observations or inferences that can be made from the data. Include these observations at the top of notebook.

Here are some final considerations:

- You must use proper labeling of your plots, to include properties such as: plot titles, axis labels, legend labels, *x*-axis and *y*-axis limits, etc.
- See the starter workbook (Pymaceuticals/pymaceuticals_starter.ipynb) for help on what modules to import and expected format of the notebook.

Hints and Considerations

- Be warned: These are very challenging tasks. Be patient with yourself as you trudge through these problems. They will take time and there is no shame in fumbling along the way. Data visualization is equal parts exploration, equal parts resolution.
- You have been provided a starter notebook. Use the code comments as a reminder of steps to follow as you complete the assignment.
- Don't get bogged down in small details. Always focus on the big picture. If you can't figure out how to get a label to show up correctly, come back to it. Focus on getting the core skeleton of your notebook complete. You can always revisit old problems.
- While you are trying to complete this assignment, feel encouraged to constantly refer to Stack Overflow and the Pandas documentation. These are needed tools in every data analyst's tool belt.
- Remember, there are many ways to approach a data problem. The key is to break up your task into micro tasks. Try answering questions like:
    - How does my DataFrame need to be structured for me to have the right *x*-axis and *y*-axis?
    - How do I build a basic scatter plot?
    - How do I add a label to that scatter plot?
    - Where would the labels for that scatter plot come from?

  Again, don't let the magnitude of a programming task scare you off. Ultimately, every programming problem boils down to a handful of bite-sized tasks.
- Get help when you need it! There is never any shame in asking. But, as always, ask a *specific* question. You'll never get a great answer to "I'm lost."