Module 5 Challenge

Start Assignment

- Due Thursday by 11:59pm
- Points 100
- Submitting a text entry box or a website url

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

In this assignment, you'll apply what you've learned about Matplotlib to a real-world situation and dataset.

Background

You've just joined Pymaceuticals, Inc., a new pharmaceutical company that specializes in anti-cancer medications. Recently, 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 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 you with generating all of the tables and figures needed for the technical report of the clinical study. They have also asked you for a top-level summary of the study results.

Files

Download the following files to help you get started:

Module 5 Challenge files ⤓ (https://static.bc-edx.com/data/dl-1-2/m5/lms/starter/Starter_Code.zip)

Instructions

This assignment is broken down into the following tasks:

- Prepare the data.

- Generate summary statistics.

- Create bar charts and pie charts.

- Calculate quartiles, find outliers, and create a box plot.

- Create a line plot and a scatter plot.

- Calculate correlation and regression.

- Submit your final analysis.

Prepare the Data

1. Run the provided package dependency and data imports, and then merge the `mouse_metadata` and `study_results` DataFrames into a single DataFrame.

2. Display the number of unique mice IDs in the data, and then check for any mouse ID with duplicate time points. Display the data associated with that mouse ID, and then create a new DataFrame where this data is removed. Use this cleaned DataFrame for the remaining steps.

3. Display the updated number of unique mice IDs.

Generate Summary Statistics

Create a DataFrame of summary statistics. Remember, there is more than one method to produce the results you're after, so the method you use is less important than the result.

Your summary statistics should include:

- A row for each drug regimen. These regimen names should be contained in the index column.

- A column for each of the following statistics: mean, median, variance, standard deviation, and SEM of the tumor volume.

Create Bar Charts and Pie Charts

1. Generate two bar charts. Both charts should be identical and show the total total number of rows (Mouse ID/Timepoints) for each drug regimen throughout the study.

   - Create the first bar chart with the Pandas `DataFrame.plot()` method.

   - Create the second bar chart with Matplotlib's `pyplot` methods.

2. Generate two pie charts. Both charts should be identical and show the distribution of female versus male mice in the study.

   - Create the first pie chart with the Pandas `DataFrame.plot()` method.

   - Create the second pie chart with Matplotlib's `pyplot` methods.

Calculate Quartiles, Find Outliers, and Create a Box Plot

1. Calculate the final tumor volume of each mouse across four of the most promising treatment regimens: Capomulin, Ramicane, Infubinol, and Ceftamin. Then, calculate the quartiles and IQR, and determine if there are any potential outliers across all four treatment regimens. Use the following substeps:

   - Create a grouped DataFrame that shows the last (greatest) time point for each mouse. Merge this grouped DataFrame with the original cleaned DataFrame.

   - Create a list that holds the treatment names as well as a second, empty list to hold the tumor volume data.

   - Loop through each drug in the treatment list, locating the rows in the merged DataFrame that correspond to each treatment. Append the resulting final tumor volumes for each drug to the empty list.

   - Determine outliers by using the upper and lower bounds, and then print the results.

2. Using Matplotlib, generate a box plot that shows the distribution of the final tumor volume for all the mice in each treatment group. 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.

Create a Line Plot and a Scatter Plot

1. Select a single mouse that was treated with Capomulin, and generate a line plot of tumor volume versus time point for that mouse.

2. Generate a scatter plot of mouse weight versus average observed tumor volume for the entire Capomulin treatment regimen.

Calculate Correlation and Regression

1. Calculate the correlation coefficient and linear regression model between mouse weight and average observed tumor volume for the entire Capomulin treatment regimen.

2. Plot the linear regression model on top of the previous scatter plot.

Requirements

Prepare the Data (20 points)

- The datasets are merged into a single DataFrame. (6 points)
- The number of mice are shown from the merged DataFrame. (2 points)
- Each duplicate mice is found based on the Mouse ID and Timepoint. (6 points)
- A clean DataFrame is created with the dropped duplicate mice. (4 points)
- The number of mice are shown from the clean DataFrame. (2 points)

Generate Summary Statistics (15 points)

- The mean of the tumor volume for each regimen is calculated using `groupby`. (2 points)
- The media of the tumor volume for each regimen is calculated using `groupby`. (2 points)
- The variance of the tumor volume for each regimen is calculated using `groupby`. (2 points)
- The standard deviation of the tumor volume for each regimen is calculated using `groupby`. (2 points)
- The SEM of the tumor volume for each regimen is calculated using `groupby`. (2 points)
- A new DataFrame is created with using the summary statistics. (5 points)

Create Bar Charts and Pie Charts (15 points)

- A bar plot showing the total number of timepoints for all mice tested for each drug regimen using Pandas is generated. (4.5 points)

- A bar plot showing the total number of timepoints for all mice tested for each drug regimen using pyplot is generated. (4.5 points)

- A pie plot showing the distribution of female versus male mice using Pandas is generated. (3 points)

- A pie plot showing the distribution of female versus male mice using pyplot is generated. (3 points)

Calculate Quartiles, Find Outliers, and Create a Box Plot (30 points)

- A DatFrame that has the last timepoint for each mouse ID is created using `groupby`. (5 points)

- The index of the DataFrame is reset. (2 points)

- Retrieve the maximum timepoint for each mouse. (2 points)

- The four treatment groups, Capomulin, Ramicane, Infubinol, and Ceftamin, are put in a list. (3 points)

- An empty list is created to fill with tumor volume data. (3 points)

- A `for` loop is used to display the interquartile range (IQR) and the outliers for each treatment group (10 points)

- A box plot is generated that shows the distribution of the final tumor volume for all the mice in each treatment group. (5 points)

Create a Line Plot and a Scatter Plot (10 points)

- A line plot is generated that shows the tumor volume vs. time point for one mouse treated with Capomulin. (5 points)

- A scatter plot is generated that shows average tumor volume vs. mouse weight for the Capomulin regimen. (5 points)

Calculate Correlation and Regression (10 points)

- The correlation coefficient and linear regression model are calculated for mouse weight and average tumor volume for the Capomulin regimen. (10 points)

Grading

This assignment will be evaluated against the requirements and assigned a grade according to the following table:

Grade	Points
A (+/-)	90+
B (+/-)	80−89
C (+/-)	70−79
D (+/-)	60−69
F (+/-)	< 60

Submission

Review all the figures and tables that you generated in this assignment. Write at least three observations or inferences that can be made from the data. Include these observations? at the top of your notebook.

To submit your Challenge assignment, click Submit, and then provide the URL of your GitHub repository for grading.

NOTE

You are allowed to miss up to two Challenge assignments and still earn your certificate. If you complete all Challenge assignments, your lowest two grades will be dropped. If you wish to skip this assignment, click Next, and move on to the next module.

Comments are disabled for graded submissions in Bootcamp Spot. If you have questions about your feedback, please notify your instructional staff or your Student Success Advisor. If you would like to resubmit your work for an additional review, you can use the Resubmit Assignment button to upload new links. You may resubmit up to three times for a total of four submissions.

IMPORTANT

It is your responsibility to include a note in the README section of your repo specifying code source and its location within your repo. This applies if you have worked with a peer on an assignment, used code in which you did not author or create sourced from a forum such as Stack Overflow, or you received code outside curriculum content from support staff such as an Instructor, TA, Tutor, or Learning Assistant. This will provide visibility to grading staff of your circumstance in order to avoid flagging your work as plagiarized.

If you are struggling with a challenge assignment or any aspect of the academic curriculum, please remember that there are student support services available for you:

1. Ask the class Slack channel/peer support.

2. AskBCS Learning Assistants exists in your class Slack application.

3. Office hours facilitated by your instructional staff before and after each class session.

4. **Tutoring Guidelines** ⏎
   (https://docs.google.com/document/d/1hTldEfWhX21B_Vz9ZentkPeziu4pPfnwiZbwQB27E90/edit?usp=sharing) -
   schedule a tutor session in the Tutor Sessions section of Bootcampspot - Canvas

5. If the above resources are not applicable and you have a need, please reach out to a member of your instructional team, your Student Success Advisor, or submit a support ticket in the Student Support section of your BCS application.

Hints and Considerations

- Use the code comments in the provided starter file to guide you through this assignment.

- Use proper labeling for your plots. Include plot titles, axis labels, legend labels, *x*-axis and *y*-axis limits, etc.

- As you work on this assignment, refer to Stack Overflow and the Matplotlib documentation for guidance. These are essential tools in every data analyst's tool belt.

- Remember that there are many ways to approach a data problem: One way is to break up your task into micro tasks. For example, ask yourself questions like the following:

  - How does my DataFrame need to be structured so it has the correct *x*-axis and *y*-axis?

  - How do I build a basic scatter plot?

  - How do I add a label to a scatter plot?

  - Where in the DataFrame can I find the names that will go into the labels?

  - Get help when you need it! Your instructional team is there for you.

References

Data generated by **Mockaroo** ⤷ **(https://mockaroo.com/)**, LLC (2022). Realistic Data Generator.