Unit 5 | Assignment - 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 some real-world situations. For this assignment, you'll need to complete **1 of 2** Data Challenges. As always, it's your choice which you complete. *Perhaps*, choose the one most relevant to your future career.

Option 1: Pyber

The ride sharing bonanza continues! Seeing the success of notable players like Uber and Lyft, you've decided to join a fledgling ride sharing company of your own. In your latest capacity, you'll be acting as Chief Data Strategist for the company. In this role, you'll be expected to offer data-backed guidance on new opportunities for market differentiation.

You've since been given access to the company's complete recordset of rides. This contains information about every active driver and historic ride, including details like city, driver count, individual fares, and city type.

Your objective is to build a Bubble Plot (https://en.wikipedia.org/wiki/Bubble_chart) that showcases the relationship between four key variables:

- Average Fare ($) Per City
- Total Number of Rides Per City
- Total Number of Drivers Per City
- City Type (Urban, Suburban, Rural)

In addition, you will be expected to produce the following three pie charts:

- % of Total Fares by City Type
- % of Total Rides by City Type
- % of Total Drivers by City Type

As final considerations:

- You must use the Pandas Library and the Jupyter Notebook.
- You must use the Matplotlib library.
- You must include a written description of three observable trends based on the data.
- You must use proper labeling of your plots, including aspects like: Plot Titles, Axes Labels, Legend Labels, Wedge Percentages, and Wedge Labels.
- Remember when making your plots to consider aesthetics!

    - You must stick to the Pyber color scheme (Gold, Light Sky Blue, and Light Coral) in producing your plot and pie charts.
    - When making your Bubble Plot, experiment with effects like `alpha`, `edgecolor`, and `linewidths`.
    - When making your Pie Chart, experiment with effects like `shadow`, `startangle`, and `explosion`.

- See Starter Workbook (Pyber/pyber_starter.ipynb) for a reference on expected format.

Option 2: Pymaceuticals Inc

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 since joined Pymaceuticals Inc., a burgeoning pharmaceutical company based out of San Diego, CA. Pymaceuticals specializes in drug-based, anti-cancer pharmaceuticals. In their most recent efforts, they've since begun screening for potential treatments to squamous cell carcinoma (SCC), a commonly occurring form of skin cancer.

As their Chief Data Analyst, you've been given access to the complete data from their most recent animal study. In this study, 250 mice were treated through a variety of drug regimes over the course of 45 days. Their physiological responses were then monitored over the course of that time. Your objective is to analyze the data to show how four treatments (Capomulin, Infubinol, Ketapril, and Placebo) compare.

To do this you are tasked with:

- Creating a scatter plot that shows how the tumor volume changes over time for each treatment.
- Creating a scatter plot that shows how the number of metastatic (https://en.wikipedia.org/wiki/Metastasis) (cancer spreading) sites changes over time for each treatment.
- Creating a scatter plot that shows the number of mice still alive through the course of treatment (Survival Rate)
- Creating a bar graph that compares the total % tumor volume change for each drug across the full 45 days.

As final considerations:

- You must use the Pandas Library and the Jupyter Notebook.
- You must use the Matplotlib library.
- You must include a written description of three observable trends based on the data.
- You must use proper labeling of your plots, including aspects like: Plot Titles, Axes Labels, Legend Labels, X and Y Axis Limits, etc.
- Your scatter plots must include error bars (https://en.wikipedia.org/wiki/Error_bar). This will allow the company to account for variability between mice. You may want to look into `pandas.DataFrame.sem` (http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.sem.html) for ideas on how to calculate this.
- Remember when making your plots to consider aesthetics!

    - Your legends should not be overlaid on top of any data.
    - Your bar graph should indicate tumor growth as red and tumor reduction as green.
      It should also include a label with the percentage change for each bar. You may want to consult this tutorial (http://composition.al/blog/2015/11/29/a-better-way-to-add-labels-to-bar-charts-with-matplotlib/) for relevant code snippets.

- See Starter Workbook (Pymaceuticals/pymaceuticals_starter.ipynb) for a reference on expected format. (Note: For this example, you are not required to match the tables or data frames included. Your only goal is to build the scatter plots and bar graphs. Consider the tables to be potential clues, but feel free to approach this problem, however, you like.)

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 **guideline** of steps you may wish to follow as you complete the assignment. You do not have to follow them step-for-step. Do not get bogged down in trying to interpret and accomplish each step.

- Between these two exercises, the Pymaceuticals one is significantly more challenging. So choose that one only if you feel somewhat comfortable with the material covered so far. The Pymaceuticals example *will* require you to research a good bit on your own for hacked solutions to problems you'll experience along the way. If you end up choosing this exercise, feel encouraged to constantly refer to Stack Overflow and the Pandas Documentation. These are needed tools in every data analyst's arsenal.

- 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 re-visit old problems.

- Remember: There are many ways to skin a cat, and similarly there are many ways to approach a data problem. The key throughout, however, is to break up your task into micro tasks. Try answering questions like: "How does my Data Frame need to be structured for me to have the right X 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 smaller, 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." Good luck!

Copyright