**Assignment 5 [Matplotlib]**

**Saranya P**

<<Data Analytics >>

**Pymaceuticals Inc and the Power of Plots**

**Table of Contents**

[**Background** 2](#_Toc19547112)

[**Pre-Work** 2](#_Toc19547113)

[**Assignment Instruction** 4](#_Toc19547114)

[**Script Explanation** 5](#_Toc19547115)

[**Importing and Combining the Source Data** 5](#_Toc19547116)

[**Tumor Response to Treatment** 6](#_Toc19547117)

[**Metastatic Response to Treatment** 9](#_Toc19547118)

[**Survival Rates** 12](#_Toc19547119)

[**Summary Bar Graph** 14](#_Toc19547120)

[**Observations:** 16](#_Toc19547121)

[**Tumor Response to Treatment** 16](#_Toc19547122)

[**Metastatic Response to Treatment** 17](#_Toc19547123)

[**Survival Rates** 18](#_Toc19547124)

[**Summary Bar Graph** 19](#_Toc19547125)

[**Conclusion** 19](#_Toc19547126)

[**Hints and Considerations** 20](#_Toc19547127)

[**GitHub** 22](#_Toc19547128)

**# Unit 5 | Assignment - Pymaceuticals Inc and 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.

# **Pre-Work**

1. Create a new repository for this project called `matplotlib-challenge`. \*\*Do not add this homework to an existing repository\*\*.

**<Sara Comment>:**

***A repository matplotlib-Challenge have been created.***

<https://github.com/SaranyaPandiaraj/Matplotlib-Challenge>

2. Clone the new repository to your computer.

**<Sara Comment>:**

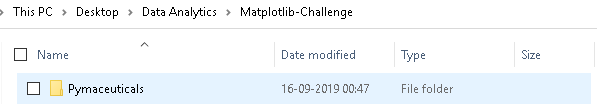
***The Matplotlib-Challenge have been cloned to my computer using the below command***

***<<git clone*** https://github.com/SaranyaPandiaraj/Matplotlib-Challenge.git***>>***

3. Inside your local git repository, create a directory for the assignment \*\*Pymaceuticals\*\*.

**<Sara Comment>:**

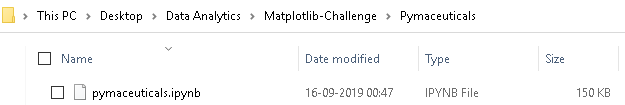
***Folder Pymaceuticals have been created inside my local git repository.***



4. Add your Jupyter notebook to this folder. This will be the main script to run for analysis.

**<Sara Comment>:**

***Jupyter Notebook have been created inside the Pymaceuticals Folder.***



5. Push the above changes to GitHub or GitLab.

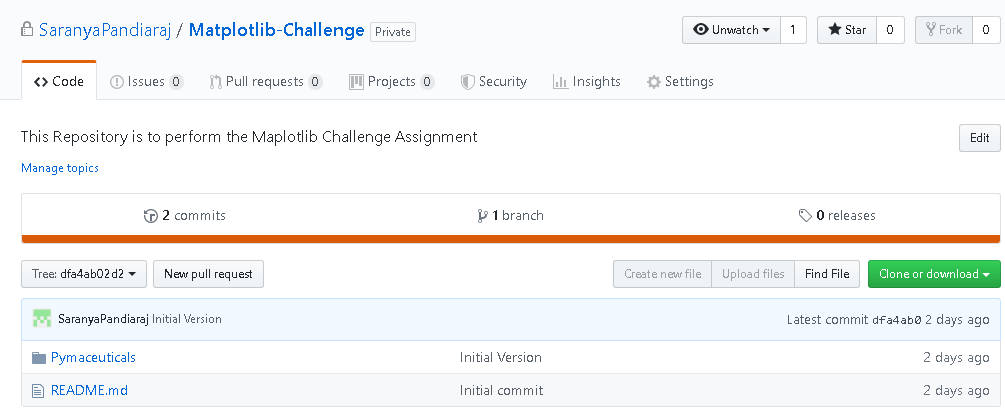
**<Sara Comment>:**

***The above changes have been pushed to GitLab using the below commands.***

***git add.***

***git commit -m “Initial Version”***

***git push***



# **Assignment Instruction**

**Pymaceuticals**



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 regimens 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.***

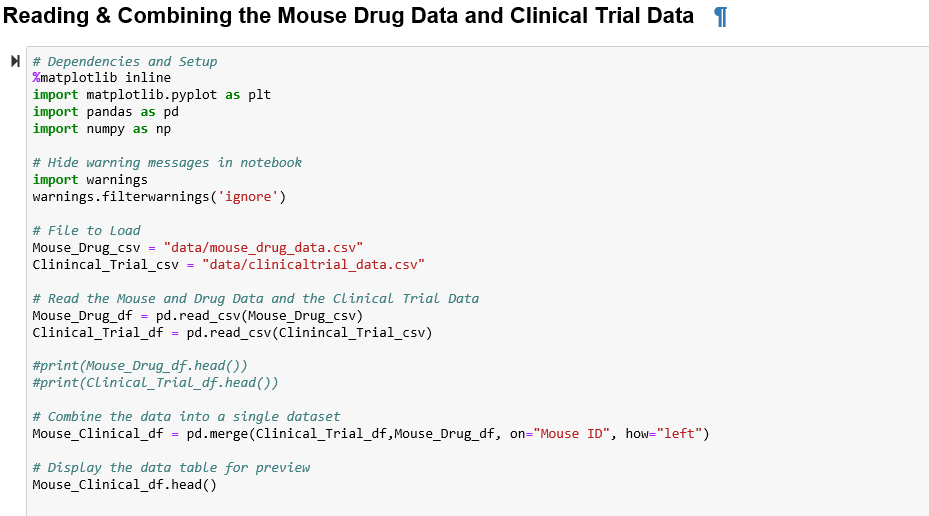
***\* Include 3 observations about the results of the study. Use the visualizations you generated from the study data as the basis for your observations.***

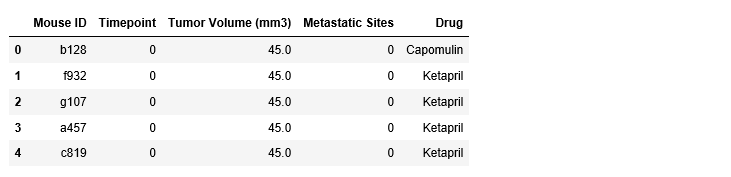
# **Script Explanation**

## **Importing and Combining the Source Data**

**<Sara Comment>:**

* ***The Mouse Drug & Clinical Trial csv file has been loaded into Mouse\_Drug\_df & Clinical\_Trial\_df data frames.***
* ***Both the data frames have been merged using pd.merge() function based on Mouse ID key column and stored into Mouse\_Clinical\_df data frame.***
* ***Preview of the Mouse\_Clinical\_df data frame is displayed using <dataframe>.head() function.***

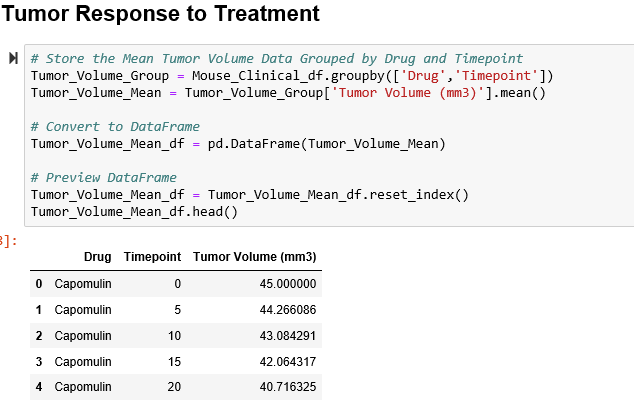




## **Tumor Response to Treatment**

**<Sara Comment>:**

* ***The Mouse\_Clinical\_df data frame has been grouped based on Drug & Timepoint and stored in Tumor\_Volume\_Group.***
* ***The Mean is calculated using mean() function on Tumor Volume (mms) Column based on Tumor\_Volume\_Group result.***
* ***Tumor\_Volume\_Mean is converted to a Dataframe using pd.Dataframe(<data>) and stored in Tumor\_Volume\_Mean\_df.***
* ***The index of the Tumor\_Volume\_Mean\_df has been reset using reset\_index() function.***
* ***Preview of the Tumor\_Volume\_Mean\_df data frame is displayed using <dataframe>.head() function.***



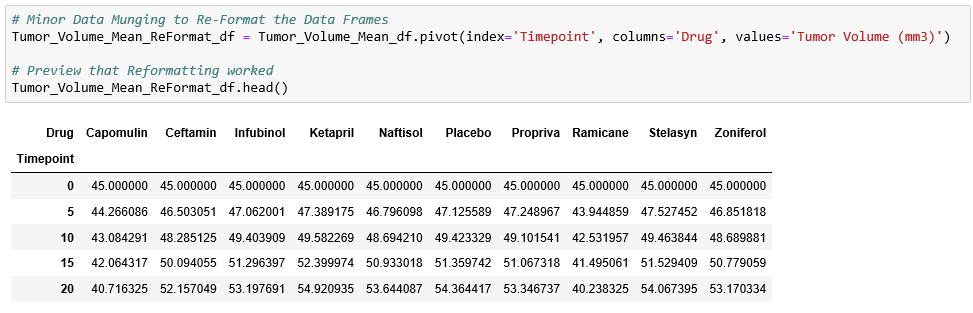
**<Sara Comment>:**

* ***The Mouse\_Clinical\_df data frame has been grouped based on Drug & Timepoint and stored in Tumor\_Volume\_Group.***
* ***The Standard Error is calculated using sem() function on Tumor Volume (mms) Column based on Tumor\_Volume\_Group result.***
* ***Tumor\_Volume\_Sem is converted to a Dataframe using pd.Dataframe(<data>) and stored in Tumor\_Volume\_Sem\_df.***
* ***The index of the Tumor\_Volume\_Sem\_df has been reset using reset\_index() function.***
* ***Preview of the Tumor\_Volume\_Sem\_df data frame is displayed using <dataframe>.head() function.***



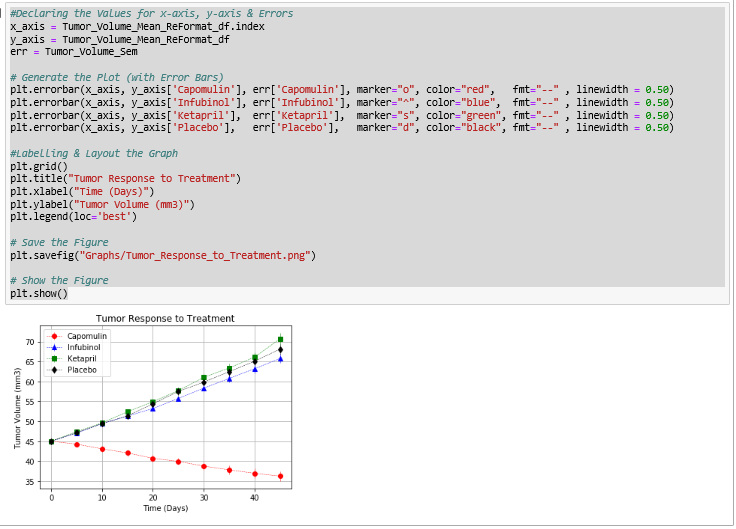
**<Sara Comment>:**

* ***Tumor\_Volume\_Mean\_df dataframe has been reformatted using <dataframe>.pivot(index=<col name>, columns=<col name>, values=<col name>)***
* ***Preview of the Tumor\_Volume\_Mean\_ReFormat\_df data frame is displayed using <dataframe>.head() function.***



**<Sara Comment>:**

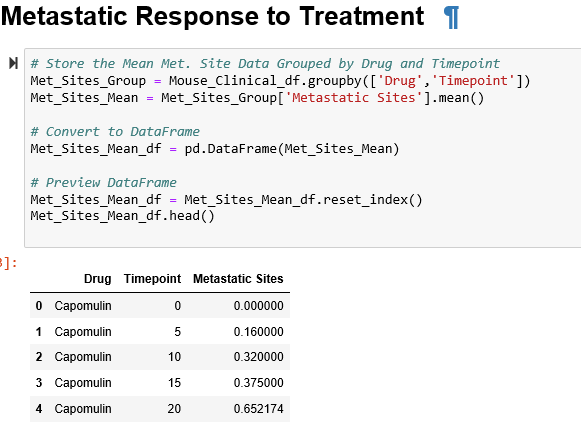
* ***The X-axis, Y-axis & Err values are stored respectively based on Tumor Volume Mean & Standard Error Values.***
* ***Errorbar plot has been generated using plt.errorbar(x\_axis,y\_axis,err,marker=”<value>”, color=”<value>”, fmt=”<value>”, linewidth =<value>) for the four drugs (Capomulin, Infubinol, Ketapril, Placebo).***
* ***Labelling and the Layout of the Graph has been applied using plt.grid(), plt.title(“<titlename>”), plt.xlabel(“<label name>”), plt.ylabel(“<label name>”), plt.legend(loc=’<loc name>’) functions respectively.***
* ***The Graph figure has been saved using plt.savefig(“<location along with the name>”) function.***
* ***The Figure has been showed using plt.show() function.***



## **Metastatic Response to Treatment**

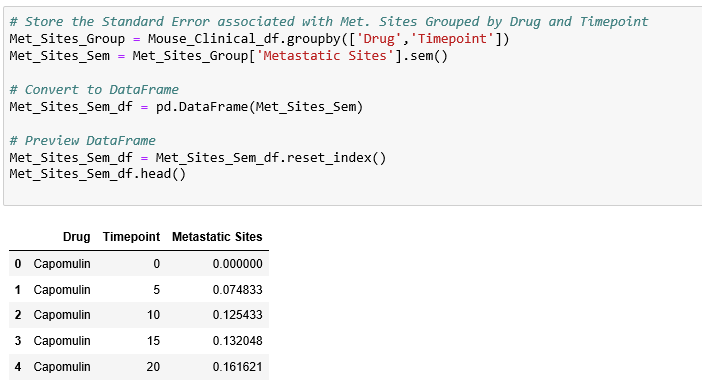
**<Sara Comment>:**

* ***The Mouse\_Clinical\_df data frame has been grouped based on Drug & Timepoint and stored in Met\_Sites\_Group.***
* ***The Mean is calculated using mean() function on Metastatic Sites Column based on Met\_Sites\_Group result.***
* ***Met\_Sites\_Mean is converted to a Dataframe using pd.Dataframe(<data>) and stored in Met\_Sites\_Mean\_df.***
* ***The index of the Met\_Sites\_Mean\_df has been reset using <dataframe>.reset\_index() function.***
* ***Preview of the Met\_Sites\_Mean\_df data frame is displayed using <dataframe>.head() function.***



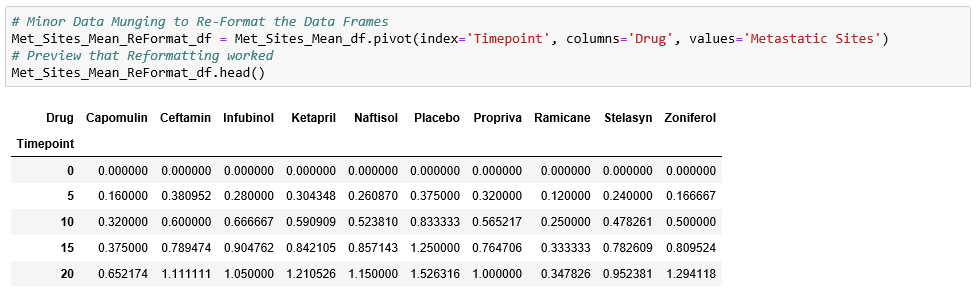
**<Sara Comment>:**

* ***The Mouse\_Clinical\_df data frame has been grouped based on Drug & Timepoint and stored in Met\_Sites\_Group.***
* ***The Mean is calculated using sem() function on Metastatic Sites Column based on Met\_Sites\_Group result.***
* ***Met\_Sites\_Sem is converted to a Dataframe using pd.Dataframe(<data>) and stored in Met\_Sites\_Sem\_df.***
* ***The index of the Met\_Sites\_Sem has been reset using reset\_index() function.***
* ***Preview of the Met\_Sites\_Sem\_df data frame is displayed using <dataframe>.head() function.***



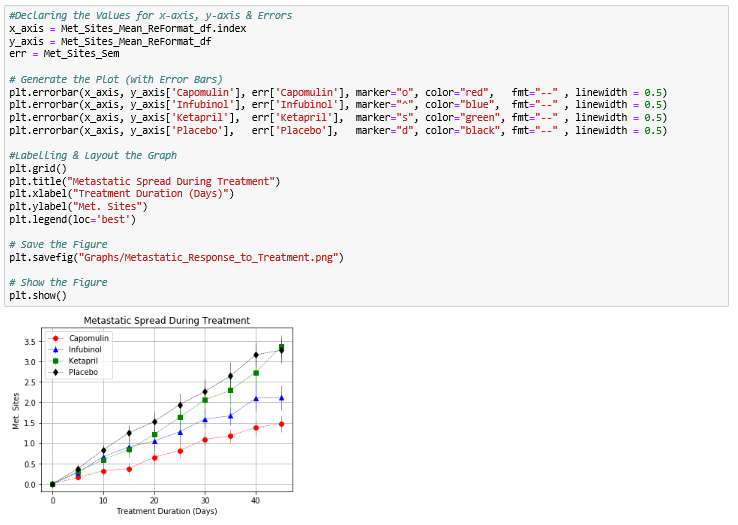
**<Sara Comment>:**

* ***Met\_Sites\_Mean\_df dataframe has been reformatted using <dataframe>.pivot(index=<col name>, columns=<col name>, values=<col name>)***
* ***Preview of the Met\_Sites\_Mean\_ReFormat\_df data frame is displayed using <dataframe>.head() function.***



**<Sara Comment>:**

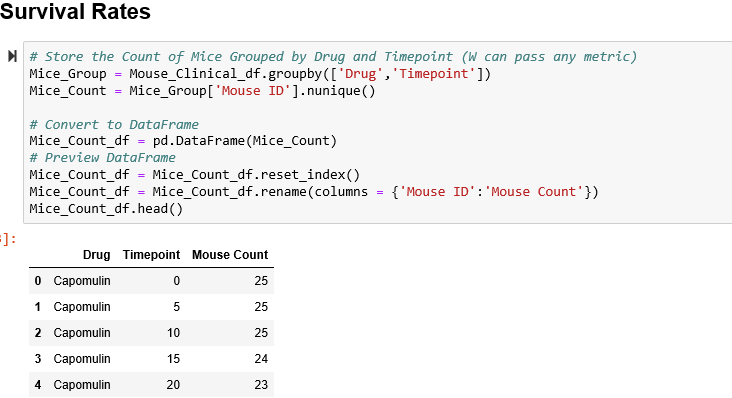
* ***The X-axis, Y-axis & Err values are stored respectively based on Metastatic Sites Mean & Standard Error Values.***
* ***Errorbar plot has been generated using plt.errorbar(x\_axis,y\_axis,err,marker=”<value>”, color=”<value>”, fmt=”<value>”, linewidth =<value>) for the four drugs (Capomulin, Infubinol, Ketapril, Placebo).***
* ***Labelling and the Layout of the Graph has been applied using plt.grid(), plt.title(“<titlename>”), plt.xlabel(“<label name>”), plt.ylabel(“<label name>”), plt.legend(loc=’<loc name>’) functions respectively.***
* ***The Graph figure has been saved using plt.savefig(“<location along with the name>”) function.***
* ***The Figure has been showed using plt.show() function.***



## **Survival Rates**

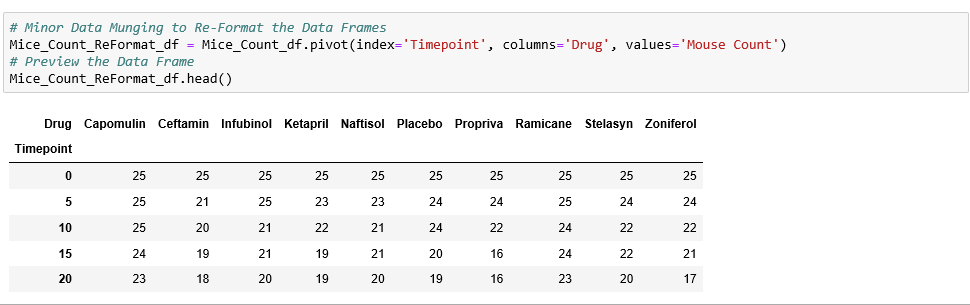
**<Sara Comment>:**

* ***The Mouse\_Clinical\_df data frame has been grouped based on Drug & Timepoint and stored in Mice\_Group.***
* ***The Count is calculated using nunique() function on Mouse ID Column based on Mice\_Group result.***
* ***Mice\_Count is converted to a Dataframe using pd.Dataframe(<data>) and stored in Mice\_Count\_df.***
* ***The index of the Mice\_Count\_df has been reset using <dataframe>.reset\_index() function.***
* ***Preview of the Met\_Sites\_Mean\_df data frame is displayed using <dataframe>.head() function.***



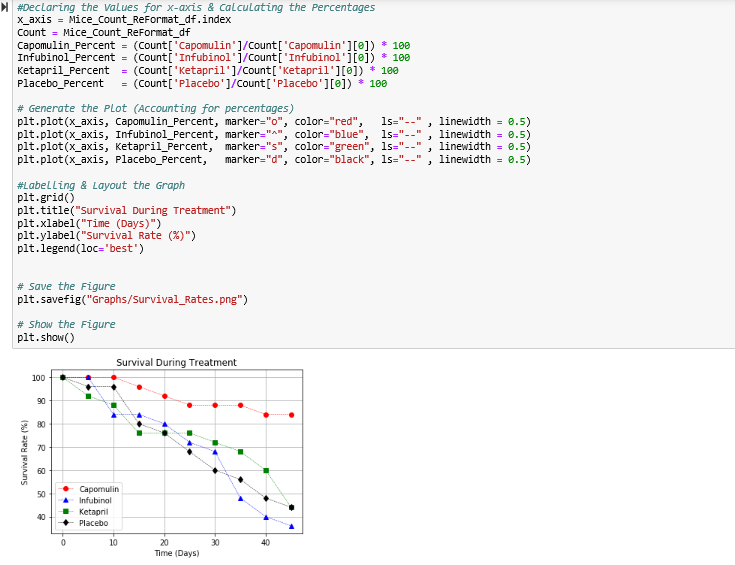
**<Sara Comment>:**

* ***Mice\_Count\_df dataframe has been reformatted using <dataframe>.pivot(index=<col name>, columns=<col name>, values=<col name>)***
* ***Preview of the Mice\_Count\_ReFormat\_df data frame is displayed using <dataframe>.head() function.***



**<Sara Comment>:**

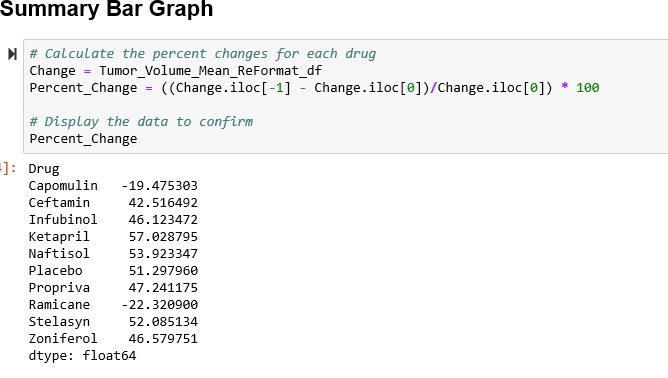
* ***The X-axis values are stored respectively based on Mice Count Values.***
* ***Percentages of the Drugs has been calculated by Count[Drug]/Count[Drug][0] \* 100.***
* ***Errorbar plot has been generated using plt.errorbar(x\_axis,DrugPercent,marker=”<value>”, color=”<value>”, ls=”<value>”, linewidth =<value>) for the four drugs (Capomulin, Infubinol, Ketapril, Placebo).***
* ***Labelling and the Layout of the Graph has been applied using plt.grid(), plt.title(“<titlename>”), plt.xlabel(“<label name>”), plt.ylabel(“<label name>”), plt.legend(loc=’<loc name>’) functions respectively.***
* ***The Graph figure has been saved using plt.savefig(“<location along with the name>”) function.***
* ***The Figure has been showed using plt.show() function.***



## **Summary Bar Graph**

**<Sara Comment>:**

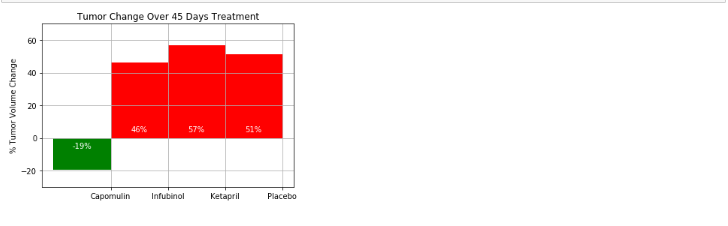
* ***Percent Change is calculated by the change in the Tumor Mean Values and multiplying it by 100.***



**<Sara Comment>:**

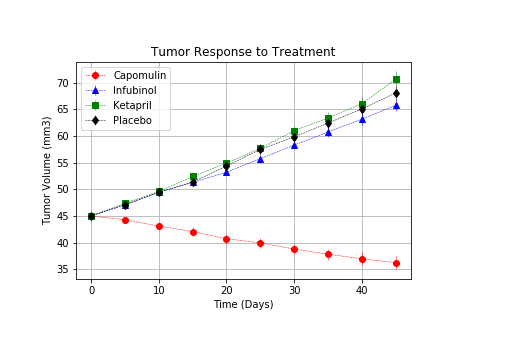
* ***The Four Drugs - (Capomulin, Infubinol, Ketapril, Placebo) are stored in a list and its relevant Percent changes are stored in a tuple.***
* ***If the Percent Change Value is greater than 0, the color in the bar should be red else green. This condition is stored in a Splice variable.***
* ***A Bar graph has been plotted using plt.bar(<DrugName>, <Percent Change Value>, color = Splice, width = -1, align =’edge’) and xticks(Drug\_name) has been applied.***
* ***Labelling and the Layout of the Graph has been applied using plt.grid(), plt.title(“<titlename>”), plt.xlabel(“<label name>”), plt.ylabel(“<label name>”), plt.legend(loc=’<loc name>’), plt.ylim(start\_val,end\_val) functions respectively.***
* ***A function percent\_label has been called using the Bar plot values to label the percentages of changes in the bar graph.***
* ***The Graph figure has been saved using plt.savefig(“<location along with the name>”) function. The Figure has been showed using plt.show() function.***





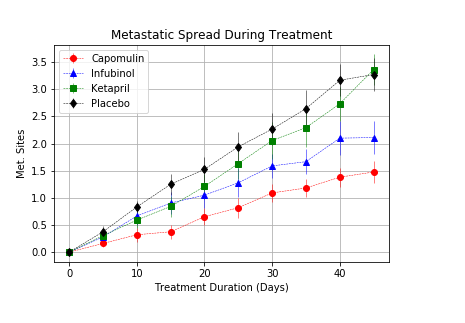
# **Observations:**

## **Tumor Response to Treatment**



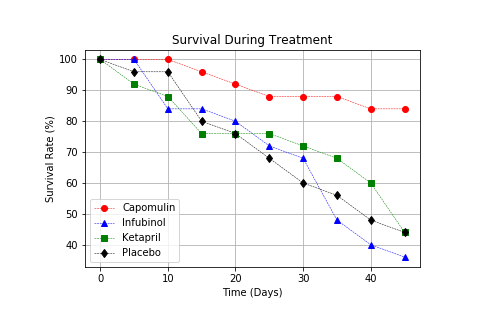
From the Tumor Response to Treatment Graph, we can infer that   
1) The Drug **"Capomulin"** shows the best result compared to other Drugs.   
2) The Tumor Volume of the mice taking the drug Capomulin decreases over the time point whereas for the other drugs, the tumor volume is comparatively increasing with the same rate of time point.

## **Metastatic Response to Treatment**



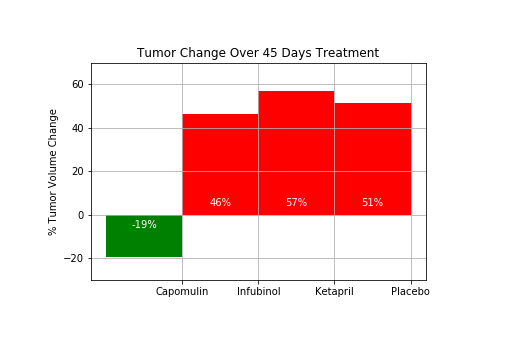
From the Metastatic Spread During Treatment Graph, we can infer that   
1) Metastatic spread is increasing for all the drugs over the time period.   
2) The Mice taking the Capomulin Drug had the least spread in Metastatic rate.  
3) The Mice taking the Placebo Drug had the highest spread in Metastatic rate.

## **Survival Rates**



From the Survival During Treatment Graph, we can infer that   
1) The mice taking the Capomulin drug had the highest survival rate of about 85%.   
2) The mice taking the Infubinol drug had the least survival rate of about 35%.

## **Summary Bar Graph**



From the Tumor Change Over 45 Days Treatment Bar Graph, we can infer that   
1) The mice taking the Capomulin drug reduced the tumor size by 19% over the 45 Days treatment.   
2) The mice taking the other drugs increased the tumor size by almost 45-60% over the time period.

# **Conclusion**

Analysing the 4 Drugs (Capomulin, Infubinol, Ketapril, Placebo), we can conclude that the Drug **Campomulin** is the best anti-cancer drug compared to the other drugs. This drug had reduced the tumor size of the mice by 19% over the 45 Days treatment period. The Survival rate of taking this drug is about 85% with the least Metastatic rate. The Tumor Volume of the mice taking the drug Capomulin decreases over the time point compared to the other 3 drugs.

# **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!

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 labelling 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.)

**<Sara Comment> :**

* ***Pandas Libraries, Matplotlib libraries and Jupyter Notebook has been used for this assignment.***
* ***Proper labelling has been done to the respective plots.***
* ***Scatter Plots are included with the error bars.***
* ***Legends are not overlaid on top of any data.***
* ***Bar Graphs has been indicated the tumor growth as red and Tumor Reduction as green.***

***The Link to Jupyter Notebook with the viewable Data Frame is below.***

***<***<https://github.com/SaranyaPandiaraj/Matplotlib-Challenge/blob/master/Pymaceuticals/pymaceuticals.html>>

***Note to TA: Please Open this file in a Browser.***

***Observations are in Observation Section.***

# **GitHub**

**<Sara Comment>:**

The link to the Pandas Assignment is in the below link.

< <https://github.com/SaranyaPandiaraj/Matplotlib-Challenge> >