# Module 5: Pymaceuticals Inc. - Data Analysis

## By: Anqa Javed

### \*\*Background:\*\*

Pymaceuticals Inc. is a pharmaceutical company specializing in anti-cancer medications. The dataset provided contains data from an animal study to test the effectiveness of various drugs in treating \*\*squamous cell carcinoma (SCC)\*\* in mice. The study tracks tumor development over a period of 45 days.

The objective of this analysis was to compare the performance of Pymaceuticals' drug of interest, \*\*Capomulin\*\*, against other treatment regimens.

### \*\*Objective:\*\*

The main objectives of this assignment are to:

- Analyze tumor development data to evaluate the effectiveness of different drug regimens.

- Perform data analysis tasks like generating summary statistics, creating plots, and running statistical tests.

- Visualize the results using various types of charts.

### \*\*Steps in the Analysis:\*\*

1. \*\*Data Preparation\*\*:

- Merged the \*\*mouse metadata\*\* and \*\*study results\*\* into a single DataFrame.

- Cleaned the data by removing duplicate entries based on `Mouse ID` and `Timepoint`.

- Displayed the number of unique mice in both the merged and cleaned datasets.

2. \*\*Summary Statistics\*\*:

- Calculated the \*\*mean\*\*, \*\*median\*\*, \*\*variance\*\*, \*\*standard deviation\*\*, and \*\*SEM (Standard Error of the Mean)\*\* for tumor volumes across different drug regimens.

3. \*\*Data Visualization\*\*:

- Generated \*\*bar charts\*\* to show the total number of timepoints for each drug regimen using both \*\*Pandas\*\* and \*\*Matplotlib\*\*.

- Created \*\*pie charts\*\* to visualize the gender distribution of the mice used in the study (female vs. male), using both \*\*Pandas\*\* and \*\*Matplotlib\*\*.

- Created a \*\*box plot\*\* to visualize the distribution of tumor volumes across different treatment groups and identify potential outliers.

4. \*\*Advanced Statistical Analysis\*\*:

- Performed \*\*linear regression\*\* and calculated the \*\*correlation coefficient\*\* between \*\*mouse weight\*\* and \*\*average tumor volume\*\* for the \*\*Capomulin\*\* regimen.

- Visualized the linear regression model overlaid on a scatter plot showing the relationship between \*\*mouse weight\*\* and \*\*tumor volume\*\*.

### \*\*Key Findings:\*\*

- \*\*Capomulin\*\* appears to be the most consistent and effective treatment, with the \*\*lowest average tumor volume\*\* and \*\*least variation\*\* across mice.

- Some \*\*outliers\*\* were observed, especially in the \*\*Infubinol\*\* and \*\*Ramicane\*\* treatment groups, suggesting some anomalies in the data that may require further investigation.

- The \*\*correlation\*\* between \*\*mouse weight\*\* and \*\*tumor volume\*\* for the \*\*Capomulin\*\* regimen was \*\*moderately positive\*\*, though not strong enough to suggest a direct causal relationship.

### \*\*Repository Structure:\*\*

- `pymaceuticals\_starter.ipynb`: Jupyter notebook containing all the analysis and visualizations.

- `data/Mouse\_metadata.csv`: Mouse metadata CSV file.

- `data/Study\_results.csv`: Study results CSV file.

- `anaconda\_projects/db/project\_filebrowser.db`: Database used in the analysis.