title: "Data 606 - Final Project" output: pdf_document: default

html_document: default

```{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)

```

---
title: "DATA 606 Data Project Proposal"
output: html_document

### Libraries Imported

```{r}
library(tidyverse)
library(dplyr)
library(plotly)
library(tidyr)
library(stringr)
library(psych)
library(ggplot2)
```

Data Preparation

load dataset1

```{r } metadata_df <- read.delim("https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Mouse_metadata.csv", header=T, sep=",") metadata_df

```
 ### Grouping by Drug.Regimen

```{r}
metadata_df %>%
  group_by(Drug.Regimen)
```

Load dataset2

```{r} results_df <- read.delim("https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Study_results.csv", header=T, sep=",") results_df

```

### Research question

### You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

In this study, 249 mice identified with Squamous cell carcinoma (SCC) tumor growth, kind of skin cancer,
were treated through a variety of drug regimens. Over the course of 45 days, tumor development was observed and measured.
The objective is to compare the performance of Pymaceuticals' drug of interest, Capomulin, versus the other treatment regimens
and find out which treatment is most effective.

Null Hyothesis: There is no difference on the results of drug treatment regimens.

Alternate Hyothesis: Capomulin does impact the results and is effective in the treatment of SCC tumor growth.

Also perform some kind of regression if possible.

### Cases

###What are the cases, and how many are there?

There are 249 unique mouse id and so are the number of cases that treated with variety if drug regimem.
```
```

~~~
Therefore there are 249 row count in the metadata_df
The results_df dataset holds the tumor growth measurments observed for each Mouse ID and carries 1,893 rows results.



### Data collection

### Describe the method of data collection.

Data must be collected by the Pymaceutical company. But I do not have the source of the data.
I got it from github.



### Type of study
### What type of study is this (observational/experiment)?

This is a experimental study.



### Data Source
### If you collected the data, state self-collected. If not, provide a citation/link.

Data is collected from GitHub Repsitory from the following links:
https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Mouse_metadata.csv
https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Study_results.csv

### Response
### What is the response variable, and what type is it (numerical/categorical)?

The response variable is the size of tumor, "Tumor.Volume..mm3." and it holds a numerical data.

### Explanatory

What is the explanatory variable, and what type is it (numerical/categorical)?

The explanatory variable is the "Drug.Regimen" and it holds a categorical data and "Timepoint" which holds numerical data

### Relevant summary statistics

Provide summary statistics relevant to your research question. For example,
if you’re comparing means across groups provide means, SDs, sample sizes of each group.
This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```{r}
describe(metadata_df)
~~~

Summary Statistic

```{r} describe(results_df)

````

### Sample Sizes for metadata_df

```{r}
nrow(metadata_df)
````

Sample Sizes for results_df

```{r} nrow(results_df)

````

### Histogram of Tumor Size

```{r}
ggplot(results_df, aes(x=Tumor.Volume..mm3.)) + geom_histogram()
````