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

pdf_document: default

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
```

Introduction: Pymaceuticals Inc., a fictional burgeoning pharmaceutical company based out of San Diego, CA, specializes in drug-based, anti-cancer pharmaceuticals.They have provided the data to test the efficacy of potential drug treatments for squamous cell carcinoma. 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 analyze the data to show how four treatments (Capomulin, Infubinol, Ketapril, and Placebo) compare.

Research question:

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

Question: Is Capomulin more effective than the three other drugs in the dataset? Is there a correlation between the age, weight and the effectiveness of capomulin?

Null Hyothesis: There is no difference between the effectiveness of the four drug regimens.

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

Perform linear regression to study the correlation between various variables.

Approach for answering the research question will be:

1- Identify the distribution of tumor growth for each mouse_id within each drug category to ensure that there is a trend available to make conclusions. 2- Then perform statistical analysis whether the change is due to age, sex or weight. 3- And finally compare the four population against each other

Cases:

What are the cases? How many different drug treatments are there? How many total sample size as well as the sample size by drug treatments are there?

Answer: The metadata_df contain 249 unique mouse id and so are the number of cases that treated with variety of drug regimem .The results_df dataset holds the tumor growth measurments observed for each Mouse ID and carries 1,893 rows results. There are 10 different drug treatments. The total sample size of mouse_id for four treatments (Capomulin, Infubinol, Ketapril, and Placebo) is 100 and the sample size of mouse_id by drug treatments is 25 each.

Data collection:

Describe the method of data collection.

Answer: Data is collected by the fictitious pharmaceutical company who was testing the efficacy of potential drug treatments for squamous cell carcinoma. I import the data into my .Rmd file from github.

Type of study:

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

Answer: This is a experimental study.A group of 249 mice were monitored after administration of a variety of drug regimens over a 45-day treatment period. The impact of Capomulin on tumor growth, metastasis and survival rates were monitored, along with Infubinol, Ketapril, and Placebo.

Data Source:

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

Answer: The citation and data collection links are as follows.

citation/link: https://c-l-nguyen.github.io/web-design-challenge/

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)?

Answer: 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)?

Answer: The explanatory variable is the "Drug.Regimen" and it holds a categorical data and "Timepoint" which holds numerical data. The 'Timepoint' unit is 'days'.

Relevant summary statistics: (Tables and Charts)

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} summary(metadata_df)

```
Summary Statistic

```{r}
summary(results_df)
```

Sample Sizes for metadata_df

```{r} nrow(metadata_df)

```
Sample Sizes for results_df

```{r}
nrow(results_df)
```

How many drug treatments are there?

```{r}

drug_count <- unique(metadata_df$Drug.Regimen)

drug_count
```

```{r}

length(drug_count)
```

Sample sizes of mouse_id by drug treatment

```{r} capomulin_df <- filter(metadata_df, Drug.Regimen=="Capomulin")

capomulin_df

```{r}
nrow(capomulin_df)
```

```{r} infubinol_df <- filter(metadata_df, Drug.Regimen=="Infubinol")

nrow(infubinol_df)

```{r}
ketapril_df <- filter(metadata_df, Drug.Regimen=="Ketapril")

nrow(ketapril_df)
```

```{r} placebo_df <- filter(metadata_df, Drug.Regimen=="Placebo")

nrow(placebo_df)

### Performing full outer join, so that no data is lost

```{r}
merge_df <- merge(x = metadata_df, y = results_df, all = TRUE)

merge_df
```

```{r} glimpse(merge_df)

###Dropping the NA rows

```{r}
merge_df <- merge_df %>% drop_na()

merge_df
```

Change colnames of some columns

assigning new names to the columns of the merged data frame

olnames(df)[2] <- "new_col2"

```{r} colnames(merge_df)[1] <- c("Mouse_Id") colnames(merge_df)[2] <- c("Drug_Regimen") colnames(merge_df)[5] <- c("Weight_g") colnames(merge_df)[7] <- c("Tumor_Volume_mm3") colnames(merge_df)[8] <- c("Metastatic_Sites")

merge_df

```{r}
merge_df %>% group_by(Mouse_Id, Timepoint)

view(merge_df)
```

```{r} df1 <- select(merge_df, Drug_Regimen, Tumor_Volume_mm3) df1

```{r}
df1 <- group_by(df1, Drug_Regimen)
df1
```

Finding the summary statistics of Tumor_Volume

```{r} stats_df <- df1 %>% summarise( Tumor_Volume_mean = mean(Tumor_Volume_mm3), Tumor_Volume_median = median(Tumor_Volume_mm3), Tumor_Volume_sd = sd(Tumor_Volume_mm3), Tumor_Volume_se = sd(Tumor_Volume_mm3)/sqrt(length((Tumor_Volume_mm3))))

stats_df

```
Filter by four drug of interest(Capomulin, Infubinol, Ketapril, and Placebo)

Drug_Regimen == "Capomulin"

```{r}
data_t <- stats_df %>% filter(Drug_Regimen == "Capomulin")


 data_t2 <- data_t %>%

  pivot_longer(cols = c(2:5),

               names_to = "Variable",

               values_to = "Value")



p <- ggplot(data = data_t2, aes(x = Variable, y = Value)) +

  geom_bar(stat = "identity", fill="lightblue") + coord_flip() + theme_dark()

p
```

Drug_Regimen == "Infubinol"

```{r} data_t <- stats_df %>% filter(Drug_Regimen == "Infubinol")

data_t2 <- data_t %>%

pivot_longer(cols = c(2:5),

```
            names_to = "Variable",

           values_to = "Value")
```

p <- ggplot(data = data_t2, aes(x = Variable, y = Value)) +

geom_bar(stat = "identity", fill="lightgreen") + coord_flip() + theme_dark()

p
```

Drug_Regimen == "Ketapril"

```{r}
data_t <- stats_df %>% filter(Drug_Regimen == "Ketapril")


 data_t2 <- data_t %>%

  pivot_longer(cols = c(2:5),

               names_to = "Variable",

               values_to = "Value")



p <- ggplot(data = data_t2, aes(x = Variable, y = Value)) +

  geom_bar(stat = "identity", fill="pink") + coord_flip() + theme_dark()

p
```

Drug_Regimen == "Placebo"

```{r} data_t <- stats_df %>% filter(Drug_Regimen == "Placebo")

data_t2 <- data_t %>%

pivot_longer(cols = c(2:5),

```
            names_to = "Variable",

           values_to = "Value")
```

p <- ggplot(data = data_t2, aes(x = Variable, y = Value)) +

geom_bar(stat = "identity", fill="purple") + coord_flip() + theme_dark()

p

```
Finding the count of each Drug

```{r}
count_df <- df1 %>% count(Drug_Regimen)

count_df
```

```{r}

# library

library(treemap)

# treemap

treemap(count_df, index="Drug_Regimen", vSize="n", type="index" )
```

To view the complete dataset

```{r}

view(merge_df)
```

remove duplicate rows across entire data frame

```{r}

merge_df <- merge_df[!duplicated(merge_df), ]

merge_df
```

filter by Capomulin, Infubinol, Ketapril, and Placebo

```{r}

capomulin_df <- filter(merge_df, Drug_Regimen == "Capomulin")
infubinol_df <- filter(merge_df, Drug_Regimen == "Infubinol")
ketapril_df <- filter(merge_df, Drug_Regimen == "Ketapril")
placebo_df <- filter(merge_df, Drug_Regimen == "Placebo")


capomulin_df
```

```{r} arrange(capomulin_df, Timepoint)

```{r}
capomulin_df <- group_by(capomulin_df, Timepoint)

capomulin_df
```
```