

# DATA 606 Data Project Proposal

## Libraries Imported

```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.3

## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.5      v purrr  0.3.4
## v tibble  3.1.2      v dplyr  1.0.7
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.2

## Warning: package 'stringr' was built under R version 4.1.2

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(dplyr)
library(plotly)

## Warning: package 'plotly' was built under R version 4.1.3

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(tidyr)
library(stringr)
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.2
```

```
##
```

```
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
```

```
##
```

```
##      %>%, alpha
```

```
library(ggplot2)
```

## Data Preparation

```
metadata_df <- read.delim("https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Mouse_m")
head(metadata_df)
```

```
load dataset1
```

```
##   Mouse.ID Drug.Regimen   Sex Age_months Weight..g.
## 1    k403    Ramicane   Male         21         16
## 2    s185    Capomulin Female          3         17
## 3    x401    Capomulin Female         16         15
## 4    m601    Capomulin   Male         22         17
## 5    g791    Ramicane   Male          11         16
## 6    s508    Ramicane   Male           1         17
```

## Grouping by Drug.Regimen

```
df <- metadata_df %>%
  group_by(Drug.Regimen)
```

```
head(df)
```

```
## # A tibble: 6 x 5
```

```
## # Groups:   Drug.Regimen [2]
```

```
##   Mouse.ID Drug.Regimen Sex   Age_months Weight..g.
##   <chr>    <chr>      <chr>    <int>    <int>
## 1 k403    Ramicane    Male        21        16
## 2 s185    Capomulin   Female         3        17
## 3 x401    Capomulin   Female        16        15
## 4 m601    Capomulin   Male        22        17
## 5 g791    Ramicane    Male        11        16
## 6 s508    Ramicane    Male         1        17
```

## Load dataset2

```
results_df <- read.delim("https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Study_results.csv")
head(results_df)
```

| ##   | Mouse.ID | Timepoint | Tumor.Volume..mm3. | Metastatic.Sites |
|------|----------|-----------|--------------------|------------------|
| ## 1 | b128     | 0         | 45                 | 0                |
| ## 2 | f932     | 0         | 45                 | 0                |
| ## 3 | g107     | 0         | 45                 | 0                |
| ## 4 | a457     | 0         | 45                 | 0                |
| ## 5 | c819     | 0         | 45                 | 0                |
| ## 6 | h246     | 0         | 45                 | 0                |

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 1: Is Capomulin more effective than the three other drugs in the dataset?

Question 2: Is there a correlation between the age, weight and the effectiveness of capomulin?

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

Alternate Hypothesis: Capomulin is more effective than the other three drug regimens on treating SCC tumor growth.

Approach for answering the research question will be:

- 1- Perform linear regression to study the correlation between various variables and calculating the correlation coefficient.
- 2- And finally compare the four population against each other.
- 3- Perform Hypothesis testing

### 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 regimen. The results\_df dataset holds the tumor growth measurements 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 an 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.

In my search for the experimental datasets, I found the Mouse\_metadata and the Study\_results on the GitHub link provided below:

[https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Mouse\\_metadata.csv](https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Mouse_metadata.csv)

[https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Study\\_results.csv](https://raw.githubusercontent.com/rfpoulos/pymaceuticals/master/data/Study_results.csv)

Upon further research in finding the original source of the the dataset, I found that these datasets are provided by Pymaceuticals Inc., a fictional burgeoning pharmaceutical company based out of San Diego, CA, specializes in drug-based, anti-cancer pharmaceuticals. Below is the link for the original source of the datasets.

<https://c-l-nguyen.github.io/web-design-challenge/index.html>

### **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.

```
summary(metadata_df)
```

```
##      Mouse.ID      Drug.Regimen      Sex      Age_months
## Length:249      Length:249      Length:249      Min.   : 1.00
## Class :character Class :character Class :character 1st Qu.: 6.00
## Mode  :character Mode  :character Mode  :character Median :13.00
##                                           Mean  :12.73
##                                           3rd Qu.:19.00
##                                           Max.   :24.00
##      Weight..g.
## Min.   :15.00
## 1st Qu.:25.00
## Median :27.00
## Mean   :26.12
## 3rd Qu.:29.00
## Max.   :30.00
```

### Summary Statistic

```
summary(results_df)
```

```
##      Mouse.ID      Timepoint      Tumor.Volume..mm3. Metastatic.Sites
## Length:1893      Min.   : 0.00      Min.   :22.05      Min.   :0.000
## Class :character 1st Qu.: 5.00      1st Qu.:45.00      1st Qu.:0.000
## Mode  :character Median :20.00      Median :48.95      Median :1.000
##                                           Mean   :19.57      Mean   :50.45      Mean   :1.022
##                                           3rd Qu.:30.00      3rd Qu.:56.29      3rd Qu.:2.000
##                                           Max.   :45.00      Max.   :78.57      Max.   :4.000
```

### Sample Sizes for metadata\_df

```
nrow(metadata_df)
```

```
## [1] 249
```

### Sample Sizes for results\_df

```
nrow(results_df)
```

```
## [1] 1893
```

How many drug treatments are there?

```
drug_count <- unique(metadata_df$Drug.Regimen)
```

```
drug_count
```

```
## [1] "Ramicane" "Capomulin" "Infubinol" "Placebo" "Ceftamin" "Stelasyn"  
## [7] "Zoniferol" "Ketapril" "Propriva" "Naftisol"
```

```
length(drug_count)
```

```
## [1] 10
```

Sample sizes of mouse\_id by drug treatment

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

```
head(capomulin_df)
```

```
##   Mouse.ID Drug.Regimen   Sex Age_months Weight..g.  
## 1    s185   Capomulin Female         3         17  
## 2    x401   Capomulin Female        16         15  
## 3    m601   Capomulin  Male        22         17  
## 4    f966   Capomulin  Male        16         17  
## 5    u364   Capomulin  Male        18         17  
## 6    y793   Capomulin  Male        17         17
```

```
nrow(capomulin_df)
```

```
## [1] 25
```

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

```
nrow(infubinol_df)
```

```
## [1] 25
```

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

```
nrow(ketapril_df)
```

```
## [1] 25
```

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

```
nrow(placebo_df)
```

```
## [1] 25
```

Performing full outer join, so that no data is lost

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

head(merge_df)
```

```
##   Mouse.ID Drug.Regimen   Sex Age_months Weight..g. Timepoint
## 1    a203   Infubinol Female      20      23      20
## 2    a203   Infubinol Female      20      23      25
## 3    a203   Infubinol Female      20      23      15
## 4    a203   Infubinol Female      20      23      10
## 5    a203   Infubinol Female      20      23      35
## 6    a203   Infubinol Female      20      23       0
##   Tumor.Volume..mm3. Metastatic.Sites
## 1          55.17334              1
## 2          56.79321              1
## 3          52.77787              1
## 4          51.85244              1
## 5          61.93165              2
## 6          45.00000              0
```

```
glimpse(merge_df)
```

```
## Rows: 1,893
## Columns: 8
## $ Mouse.ID      <chr> "a203", "a203", "a203", "a203", "a203", "a203", "a2~
## $ Drug.Regimen  <chr> "Infubinol", "Infubinol", "Infubinol", "Infubinol", ~
## $ Sex           <chr> "Female", "Female", "Female", "Female", "Female", "~
## $ Age_months    <int> 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, ~
## $ Weight..g.    <int> 23, 23, 23, 23, 23, 23, 23, 23, 23, 23, 25, 25, 25, ~
## $ Timepoint     <int> 20, 25, 15, 10, 35, 0, 30, 5, 45, 40, 5, 40, 35, 45~
## $ Tumor.Volume..mm3. <dbl> 55.17334, 56.79321, 52.77787, 51.85244, 61.93165, 4~
## $ Metastatic.Sites <int> 1, 1, 1, 1, 2, 0, 1, 0, 2, 2, 0, 1, 1, 1, 1, 1, ~
```

Dropping the NA rows

```
merge_df <- merge_df %>% drop_na()

head(merge_df)
```

```
##   Mouse.ID Drug.Regimen   Sex Age_months Weight..g. Timepoint
## 1    a203   Infubinol Female      20      23      20
## 2    a203   Infubinol Female      20      23      25
## 3    a203   Infubinol Female      20      23      15
## 4    a203   Infubinol Female      20      23      10
## 5    a203   Infubinol Female      20      23      35
## 6    a203   Infubinol Female      20      23       0
##   Tumor.Volume..mm3. Metastatic.Sites
## 1          55.17334              1
```

```
## 2      56.79321      1
## 3      52.77787      1
## 4      51.85244      1
## 5      61.93165      2
## 6      45.00000      0
```

Change colnames of some columns

assigning new names to the columns of the merged data frame

```
Colnames(df)[2] <- "new_col2"
```

```
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")
```

```
head(merge_df)
```

```
##   Mouse_Id Drug_Regimen Sex Age_months Weight_g Timepoint Tumor_Volume_mm3
## 1    a203   Infubinol Female      20      23      20      55.17334
## 2    a203   Infubinol Female      20      23      25      56.79321
## 3    a203   Infubinol Female      20      23      15      52.77787
## 4    a203   Infubinol Female      20      23      10      51.85244
## 5    a203   Infubinol Female      20      23      35      61.93165
## 6    a203   Infubinol Female      20      23       0      45.00000
##   Metastatic_Sites
## 1                1
## 2                1
## 3                1
## 4                1
## 5                2
## 6                0
```

```
merge_df %>% group_by(Mouse_Id, Timepoint)
```

```
## # A tibble: 1,893 x 8
## # Groups:   Mouse_Id, Timepoint [1,888]
##   Mouse_Id Drug_Regimen Sex Age_months Weight_g Timepoint Tumor_Volume_mm3
##   <chr>    <chr>      <chr>    <int>    <int>    <int>      <dbl>
## 1 a203     Infubinol Female      20      23      20      55.2
## 2 a203     Infubinol Female      20      23      25      56.8
## 3 a203     Infubinol Female      20      23      15      52.8
## 4 a203     Infubinol Female      20      23      10      51.9
## 5 a203     Infubinol Female      20      23      35      61.9
## 6 a203     Infubinol Female      20      23       0      45
## 7 a203     Infubinol Female      20      23      30      59.5
## 8 a203     Infubinol Female      20      23       5      48.5
## 9 a203     Infubinol Female      20      23      45      68.0
## 10 a203    Infubinol Female      20      23      40      63.6
## # ... with 1,883 more rows, and 1 more variable: Metastatic_Sites <int>
```



```
head(merge_df)
```

```
##   Mouse_Id Drug_Regimen   Sex Age_months Weight_g Timepoint Tumor_Volume_mm3
## 1     a203   Infubinol Female      20      23        20      55.17334
## 2     a203   Infubinol Female      20      23        25      56.79321
## 3     a203   Infubinol Female      20      23        15      52.77787
## 4     a203   Infubinol Female      20      23        10      51.85244
## 5     a203   Infubinol Female      20      23        35      61.93165
## 6     a203   Infubinol Female      20      23         0      45.00000
##   Metastatic_Sites
## 1                1
## 2                1
## 3                1
## 4                1
## 5                2
## 6                0
```

```
df1 <- select(merge_df, Drug_Regimen, Tumor_Volume_mm3, Age_months, Weight_g)
head(df1)
```

```
##   Drug_Regimen Tumor_Volume_mm3 Age_months Weight_g
## 1   Infubinol      55.17334      20      23
## 2   Infubinol      56.79321      20      23
## 3   Infubinol      52.77787      20      23
## 4   Infubinol      51.85244      20      23
## 5   Infubinol      61.93165      20      23
## 6   Infubinol      45.00000      20      23
```

```
df1 <- group_by(df1, Drug_Regimen)
head(df1)
```

```
## # A tibble: 6 x 4
## # Groups:   Drug_Regimen [1]
##   Drug_Regimen Tumor_Volume_mm3 Age_months Weight_g
##   <chr>          <dbl>      <int>    <int>
## 1 Infubinol      55.2        20      23
## 2 Infubinol      56.8        20      23
## 3 Infubinol      52.8        20      23
## 4 Infubinol      51.9        20      23
## 5 Infubinol      61.9        20      23
## 6 Infubinol      45          20      23
```

## Finding the summary statistics of Tumor\_Volume

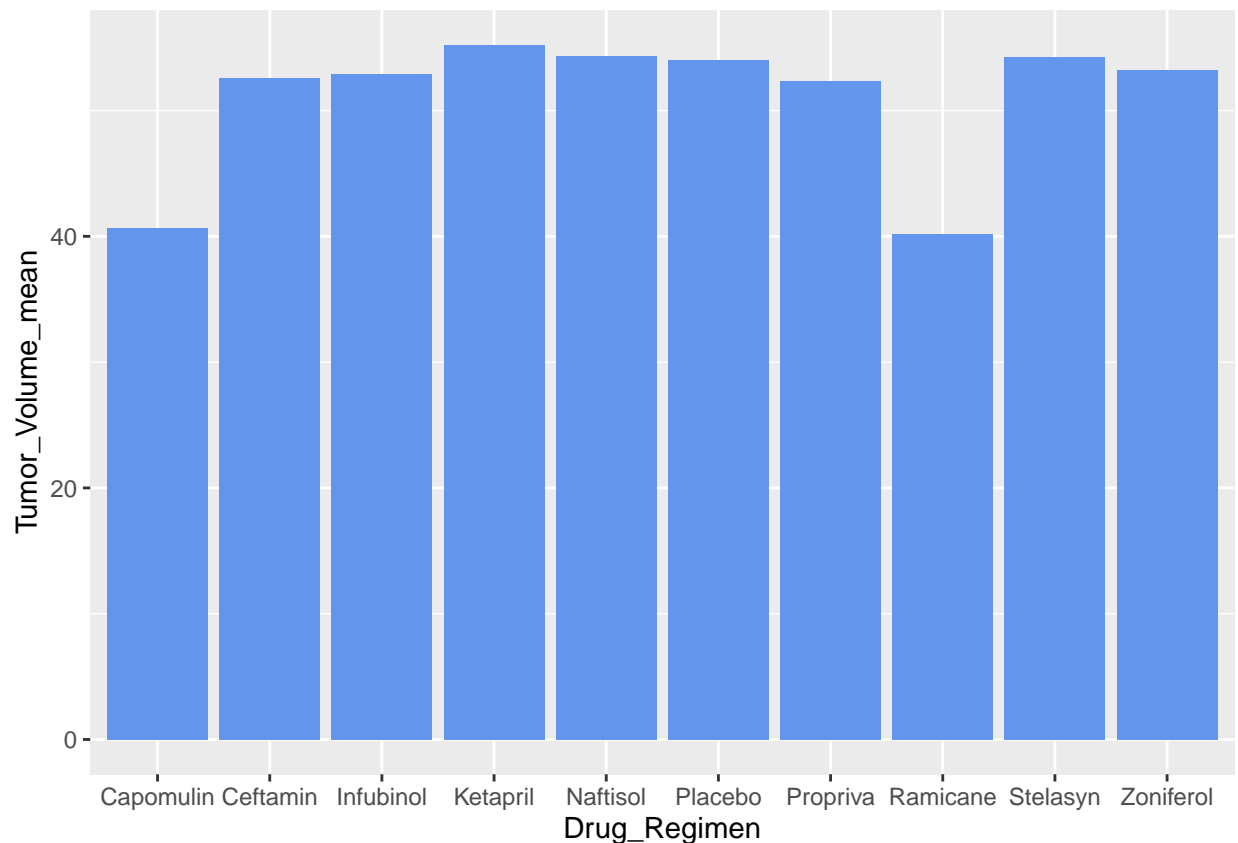
```
stats_df <- df1 %>% summarise(
  Tumor_Volume_mean = mean(Tumor_Volume_mm3), Tumor_Volume_median = median(Tumor_Volume_mm3), Tumor_Vol
head(stats_df)
```

```
## # A tibble: 6 x 5
##   Drug_Regimen Tumor_Volume_me~ Tumor_Volume_me~ Tumor_Volume_sd Tumor_Volume_se
##   <chr>         <dbl>         <dbl>         <dbl>         <dbl>
## 1 Capomulin      40.7           41.6           4.99           0.329
## 2 Ceftamin       52.6           51.8           6.27           0.470
## 3 Infubinol      52.9           51.8           6.57           0.492
## 4 Ketapril       55.2           53.7           8.28           0.604
## 5 Naftisol       54.3           52.5           8.13           0.596
## 6 Placebo       54.0           52.3           7.82           0.581
```

Comparing means of tumor size by drug treatment.

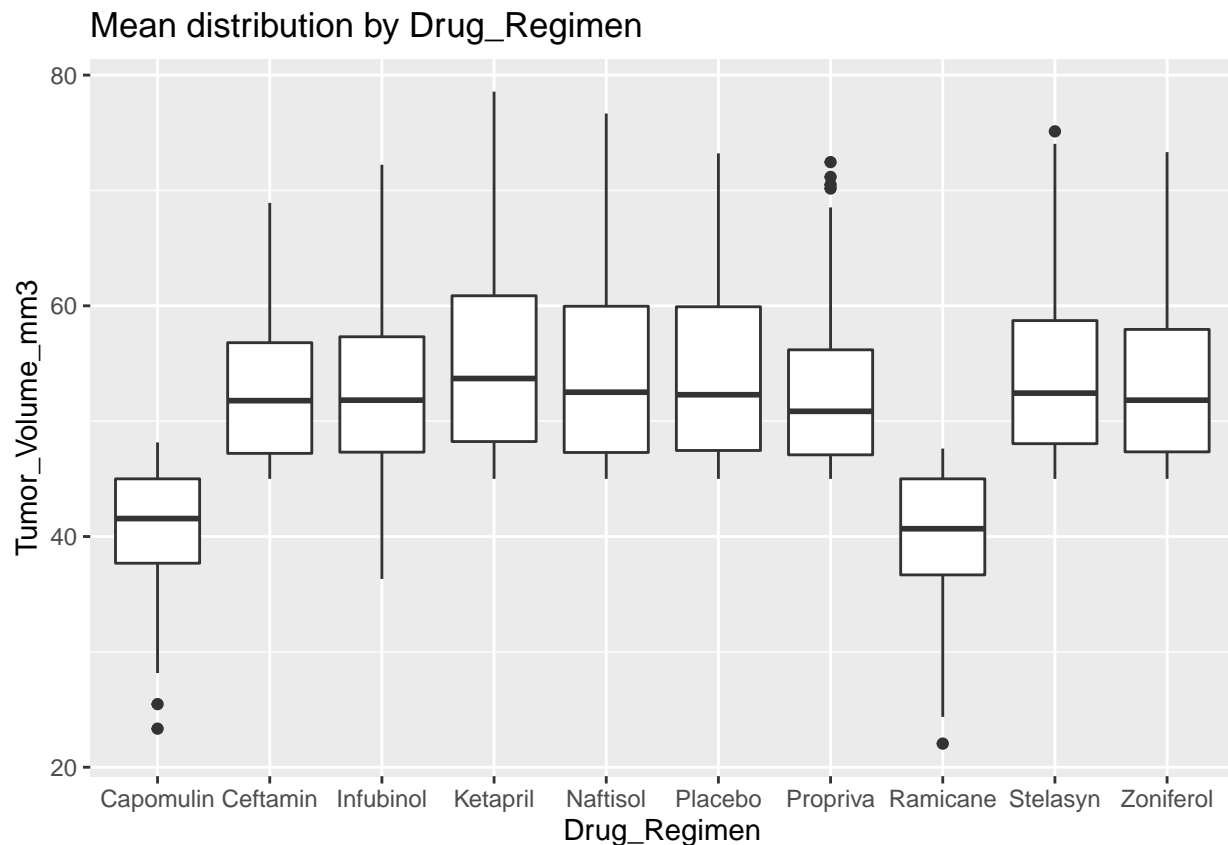
```
library(ggplot2)

# plot mean salaries
ggplot(stats_df,
  aes(x = Drug_Regimen,
      y = Tumor_Volume_mean)) +
  geom_bar(stat = "identity", fill = "cornflowerblue")
```



Side-by-side box plots are very useful for comparing groups (i.e., the levels of a categorical variable) on a numerical variable. Outliers are prominent for Drug\_Regimen Capomulin, Propriva, Ramicane and Stelasyn.

```
ggplot(merge_df,
      aes(x = Drug_Regimen,
          y = Tumor_Volume_mm3)) +
  geom_boxplot() +
  labs(title = "Mean distribution by Drug_Regimen")
```



Finding the mice count of each Drug Regimen

```
count_df <- df1 %>% count(Drug_Regimen)
```

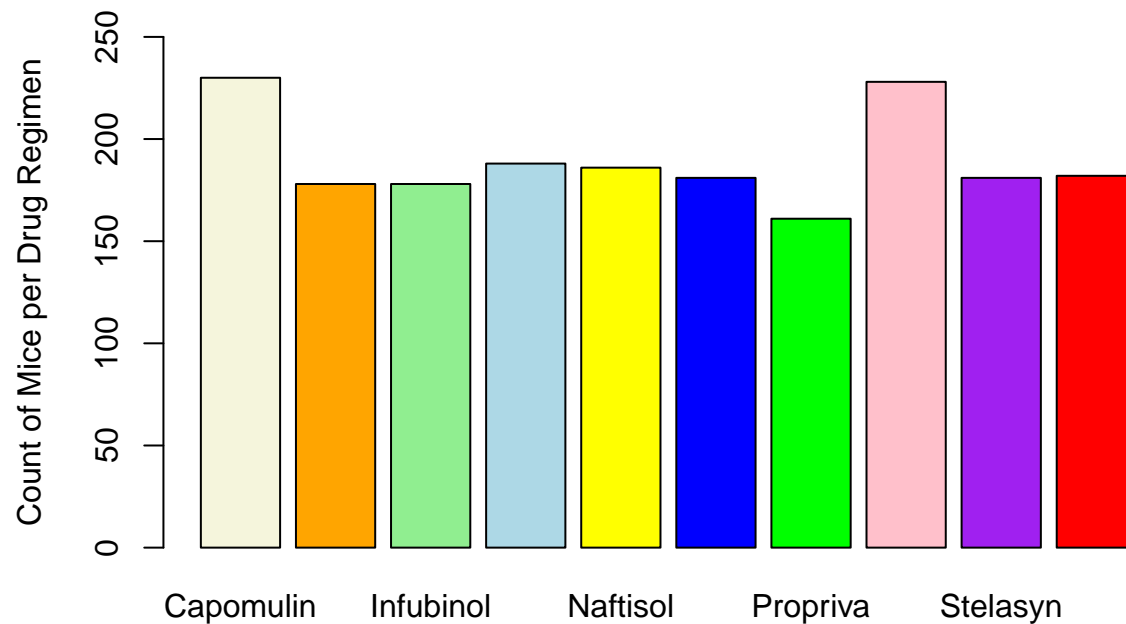
```
count_df
```

```
## # A tibble: 10 x 2
## # Groups:   Drug_Regimen [10]
##   Drug_Regimen     n
##   <chr>         <int>
## 1 Capomulin      230
## 2 Ceftamin       178
```

```
## 3 Infubinol      178
## 4 Ketapril       188
## 5 Naftisol       186
## 6 Placebo        181
## 7 Propriva       161
## 8 Ramicane       228
## 9 Stelasyn      181
## 10 Zoniferol     182
```

Plotting the number of mice in each drug regimen

```
barplot(c(230, 178, 178, 188, 186, 181, 161, 228, 181, 182),
        names.arg=c("Capomulin","Ceftamin","Infubinol","Ketapril","Naftisol", "Placebo", "Propriva", "Ramicane", "Stelasyn", "Zoniferol"),
        ylim=c(0,250),
        col=c("beige","orange","lightgreen","lightblue","yellow", "blue", "green", "pink", "purple", "red"),
        ylab="Count of Mice per Drug Regimen")
```



Remove duplicate rows across entire data frame

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

|      | Mouse_Id         | Drug_Regimen | Sex    | Age_months | Weight_g | Timepoint | Tumor_Volume_mm3 |
|------|------------------|--------------|--------|------------|----------|-----------|------------------|
| ## 1 | a203             | Infubinol    | Female | 20         | 23       | 20        | 55.17334         |
| ## 2 | a203             | Infubinol    | Female | 20         | 23       | 25        | 56.79321         |
| ## 3 | a203             | Infubinol    | Female | 20         | 23       | 15        | 52.77787         |
| ## 4 | a203             | Infubinol    | Female | 20         | 23       | 10        | 51.85244         |
| ## 5 | a203             | Infubinol    | Female | 20         | 23       | 35        | 61.93165         |
| ## 6 | a203             | Infubinol    | Female | 20         | 23       | 0         | 45.00000         |
| ##   | Metastatic_Sites |              |        |            |          |           |                  |
| ## 1 |                  | 1            |        |            |          |           |                  |
| ## 2 |                  | 1            |        |            |          |           |                  |
| ## 3 |                  | 1            |        |            |          |           |                  |
| ## 4 |                  | 1            |        |            |          |           |                  |
| ## 5 |                  | 2            |        |            |          |           |                  |
| ## 6 |                  | 0            |        |            |          |           |                  |

filter by Capomulin, Infubinol, Ketapril, and Placebo

```
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")
```

```
head(capomulin_df)
```

|      | Mouse_Id         | Drug_Regimen | Sex    | Age_months | Weight_g | Timepoint | Tumor_Volume_mm3 |
|------|------------------|--------------|--------|------------|----------|-----------|------------------|
| ## 1 | b128             | Capomulin    | Female | 9          | 22       | 5         | 45.65133         |
| ## 2 | b128             | Capomulin    | Female | 9          | 22       | 25        | 43.26214         |
| ## 3 | b128             | Capomulin    | Female | 9          | 22       | 35        | 37.96764         |
| ## 4 | b128             | Capomulin    | Female | 9          | 22       | 10        | 43.27085         |
| ## 5 | b128             | Capomulin    | Female | 9          | 22       | 0         | 45.00000         |
| ## 6 | b128             | Capomulin    | Female | 9          | 22       | 40        | 38.37973         |
| ##   | Metastatic_Sites |              |        |            |          |           |                  |
| ## 1 |                  | 0            |        |            |          |           |                  |
| ## 2 |                  | 1            |        |            |          |           |                  |
| ## 3 |                  | 1            |        |            |          |           |                  |
| ## 4 |                  | 0            |        |            |          |           |                  |
| ## 5 |                  | 0            |        |            |          |           |                  |
| ## 6 |                  | 2            |        |            |          |           |                  |

To generate a scatter plot of average tumor volume vs. mouse weight for all mice in the Capomulin regimen.

First we calculate the final tumor volume of each mouse\_id across four of the treatment regimens:

(Capomulin, Infubinol, Ketapril, and Placebo)

Since not all mice lived until timepoint 45, we start by getting the last (greatest) timepoint for each mouse

capomulin\_df:

```
capo_df1 <- select(capomulin_df, Mouse_Id, Timepoint, Tumor_Volume_mm3) %>%
  group_by(Mouse_Id) %>%
  filter(Timepoint == max(Timepoint, na.rm=TRUE))

head(capo_df1)
```

```
## # A tibble: 6 x 3
## # Groups:   Mouse_Id [6]
##   Mouse_Id Timepoint Tumor_Volume_mm3
##   <chr>      <int>      <dbl>
## 1 b128         45         39.0
## 2 b742         45         38.9
## 3 f966         20         30.5
## 4 g288         45         37.1
## 5 g316         45         40.2
## 6 i557         45         47.7
```

Find the average weight by mice\_id in Capomulin\_df

```
capo_df2 <- select(capomulin_df, Mouse_Id, Weight_g) %>%
  group_by(Mouse_Id) %>%
  summarise(Average_weight = mean(Weight_g, na.rm=TRUE))

head(capo_df2)
```

```
## # A tibble: 6 x 2
##   Mouse_Id Average_weight
##   <chr>      <dbl>
## 1 b128         22
## 2 b742         21
## 3 f966         17
## 4 g288         19
## 5 g316         22
## 6 i557         24
```

## Joining the two df's for adding average weight

```
capo_df <- capo_df1 %>% inner_join(capo_df2, by = "Mouse_Id")  
head(capo_df)
```

```
## # A tibble: 6 x 4  
## # Groups:   Mouse_Id [6]  
##   Mouse_Id Timepoint Tumor_Volume_mm3 Average_weight  
##   <chr>         <int>         <dbl>         <dbl>  
## 1 b128          45          39.0          22  
## 2 b742          45          38.9          21  
## 3 f966          20          30.5          17  
## 4 g288          45          37.1          19  
## 5 g316          45          40.2          22  
## 6 i557          45          47.7          24
```

## Find the average age by mice\_id in Capomulin\_df

```
capo_df3 <- select(capomulin_df, Mouse_Id, Age_months) %>%  
  group_by(Mouse_Id) %>%  
  summarise(Average_age = mean(Age_months, na.rm=TRUE))  
head(capo_df3)
```

```
## # A tibble: 6 x 2  
##   Mouse_Id Average_age  
##   <chr>         <dbl>  
## 1 b128          9  
## 2 b742          7  
## 3 f966         16  
## 4 g288          3  
## 5 g316         22  
## 6 i557          1
```

## Joining the two df's for adding average age

```
capo_df <- capo_df %>% inner_join(capo_df3, by = "Mouse_Id")  
head(capo_df)
```

```
## # A tibble: 6 x 5  
## # Groups:   Mouse_Id [6]  
##   Mouse_Id Timepoint Tumor_Volume_mm3 Average_weight Average_age  
##   <chr>         <int>         <dbl>         <dbl>         <dbl>  
## 1 b128          45          39.0          22           9  
## 2 b742          45          38.9          21           7  
## 3 f966          20          30.5          17          16
```

```
## 4 g288          45          37.1          19          3
## 5 g316          45          40.2          22          22
## 6 i557          45          47.7          24          1
```

summerize the Tumor\_Volume\_mm3

```
capo_df$Tumor_Volume_mm3 %>%
  summary()
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  23.34   32.38   38.13   36.67   40.16   47.69
```

Standard Deviation

```
capo_df$Tumor_Volume_mm3 %>% sd()
```

```
## [1] 5.715188
```

For project proposal, plotting correlation matrices with all the relevant variables for Capomulin drug to analyze.

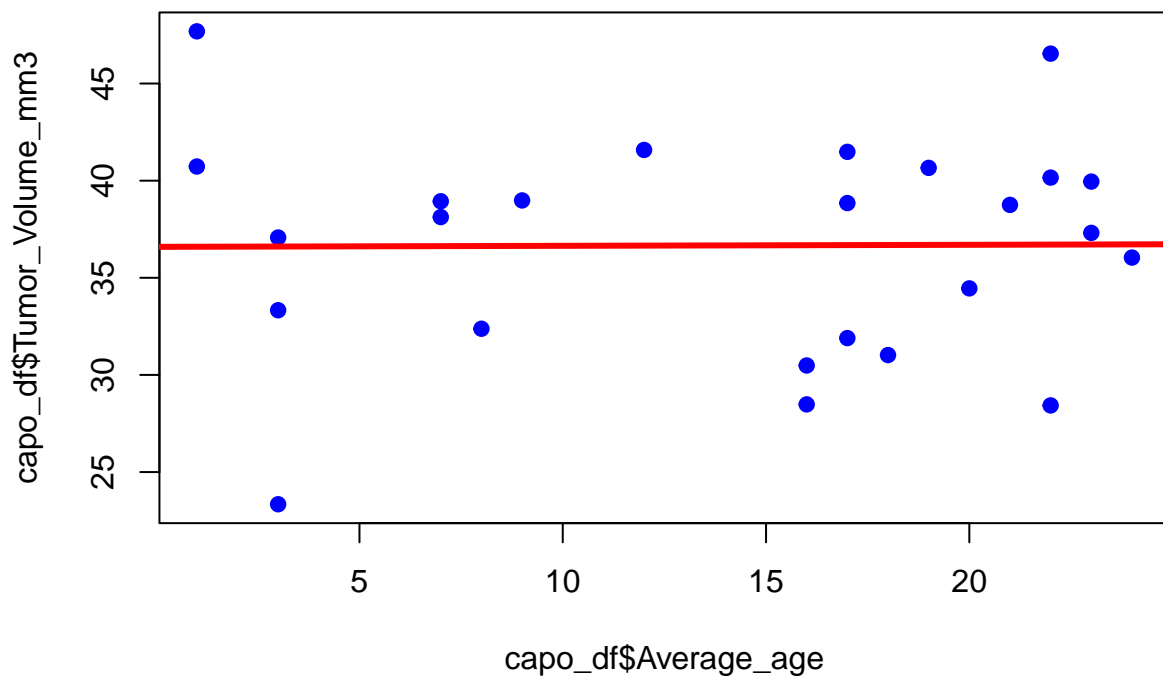
capomulin\_df Vs Age\_months

```
# Creating the plot
plot(capo_df$Average_age, capo_df$Tumor_Volume_mm3, pch = 19, col = "blue")

# Regression line
abline(lm(capo_df$Tumor_Volume_mm3 ~ capo_df$Average_age), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(capo_df$Average_age, capo_df$Tumor_Volume_mm3), 2)), x = 25, y = 9)
```



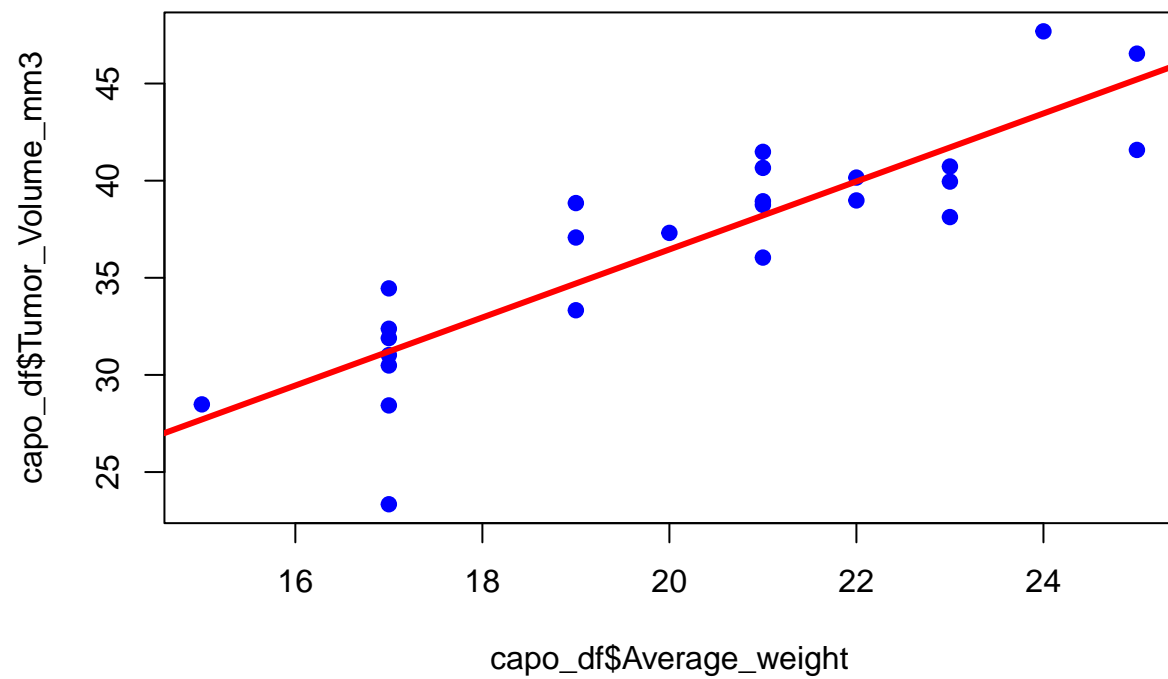


capomulin\_df Vs Weight\_g

```
# Creating the plot
plot(capo_df$Average_weight, capo_df$Tumor_Volume_mm3, pch = 19, col = "blue")

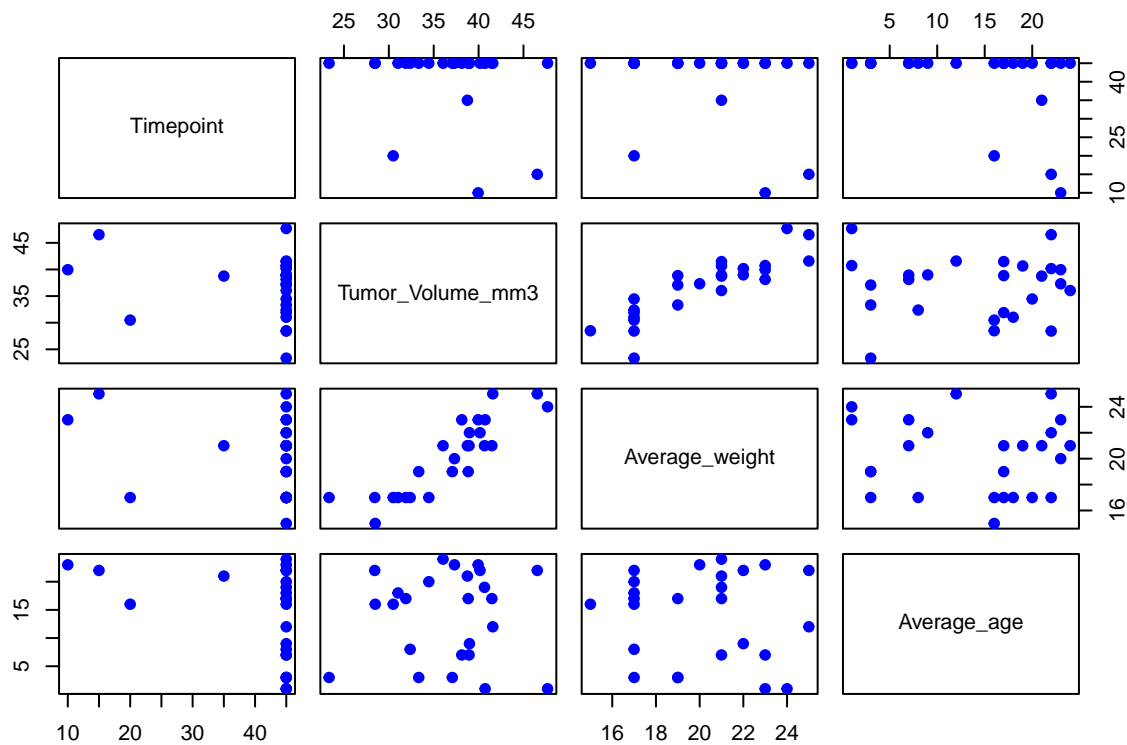
# Regression line
abline(lm(capo_df$Tumor_Volume_mm3 ~ capo_df$Average_weight), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(capo_df$Average_weight, capo_df$Tumor_Volume_mm3), 2)), x = 25, y = 45)
```



Correlation Matrix

```
pairs(capo_df[,2:5], pch = 19, col = "blue")
```



Infubinol\_df:

```
infu_df1 <- select(infubinol_df, Mouse_Id, Timepoint, Tumor_Volume_mm3) %>%
  group_by(Mouse_Id) %>%
  filter(Timepoint == max(Timepoint, na.rm=TRUE))
```

### Find the average weight by mice\_id in Infubinol\_df

```
infu_df2 <- select(infubinol_df, Mouse_Id, Weight_g) %>%
  group_by(Mouse_Id) %>%
  summarise(Average_weight = mean(Weight_g, na.rm=TRUE))
```

### Joining the two df's for adding average weight

```
infu_df <- infu_df1 %>% inner_join(infu_df2, by = "Mouse_Id")
```

### Find the average age by mice\_id in Capomulin\_df

```
infu_df3 <- select(infubinol_df, Mouse_Id, Age_months) %>%
  group_by(Mouse_Id) %>%
  summarise(Average_age = mean(Age_months, na.rm=TRUE))
```

### Joining the two df's for adding average age

```
infu_df <- infu_df %>% inner_join(infu_df3, by = "Mouse_Id")

head(infu_df)
```

```
## # A tibble: 6 x 5
## # Groups:   Mouse_Id [6]
##   Mouse_Id Timepoint Tumor_Volume_mm3 Average_weight Average_age
##   <chr>      <int>      <dbl>          <dbl>      <dbl>
## 1 a203         45         68.0           23         20
## 2 a251         45         65.5           25         21
## 3 a577         30         57.0           25          6
## 4 a685         45         66.1           30          8
## 5 c139         45         72.2           28         11
## 6 c326          5         36.3           25         18
```

summerize the Tumor\_Volume\_mm3

```
infu_df$Tumor_Volume_mm3 %>%
  summary()
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   36.32  54.05   60.17   58.18  65.53   72.23
```

Standard Deviation

```
infu_df$Tumor_Volume_mm3 %>% sd()
```

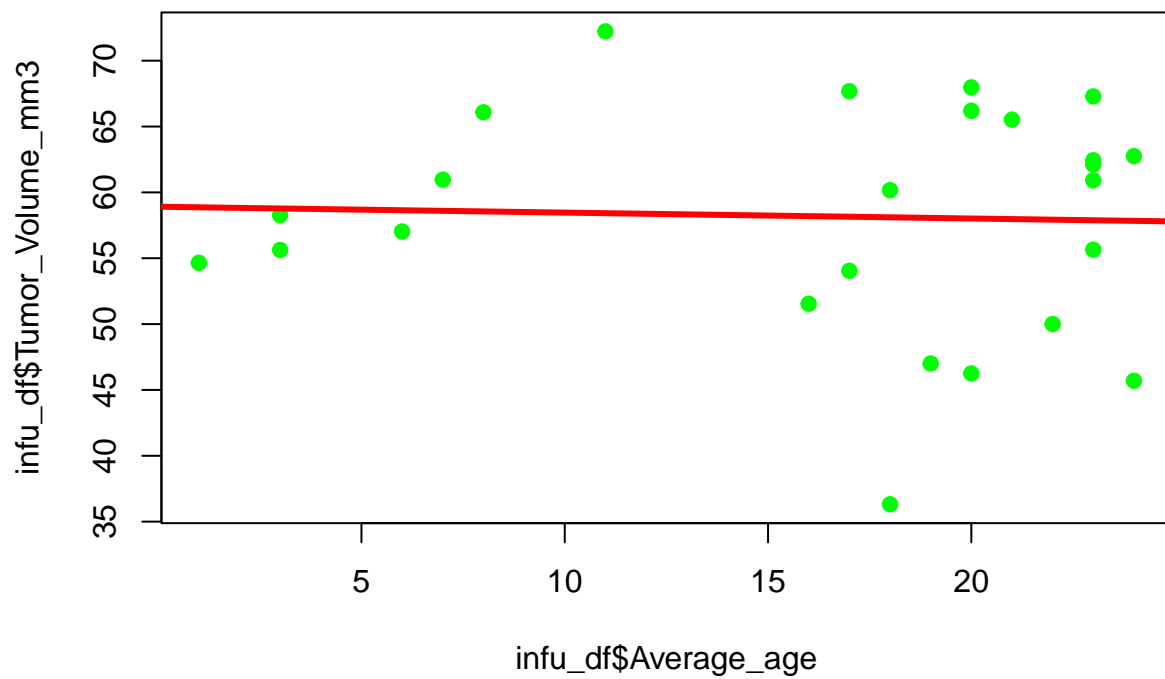
```
## [1] 8.602957
```

infubinol\_df Vs Age\_months

```
# Creating the plot
plot(infu_df$Average_age, infu_df$Tumor_Volume_mm3, pch = 19, col = "green")

# Regression line
abline(lm(infu_df$Tumor_Volume_mm3 ~ infu_df$Average_age), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(infu_df$Average_age, infu_df$Tumor_Volume_mm3), 2)), x = 25, y = 9)
```

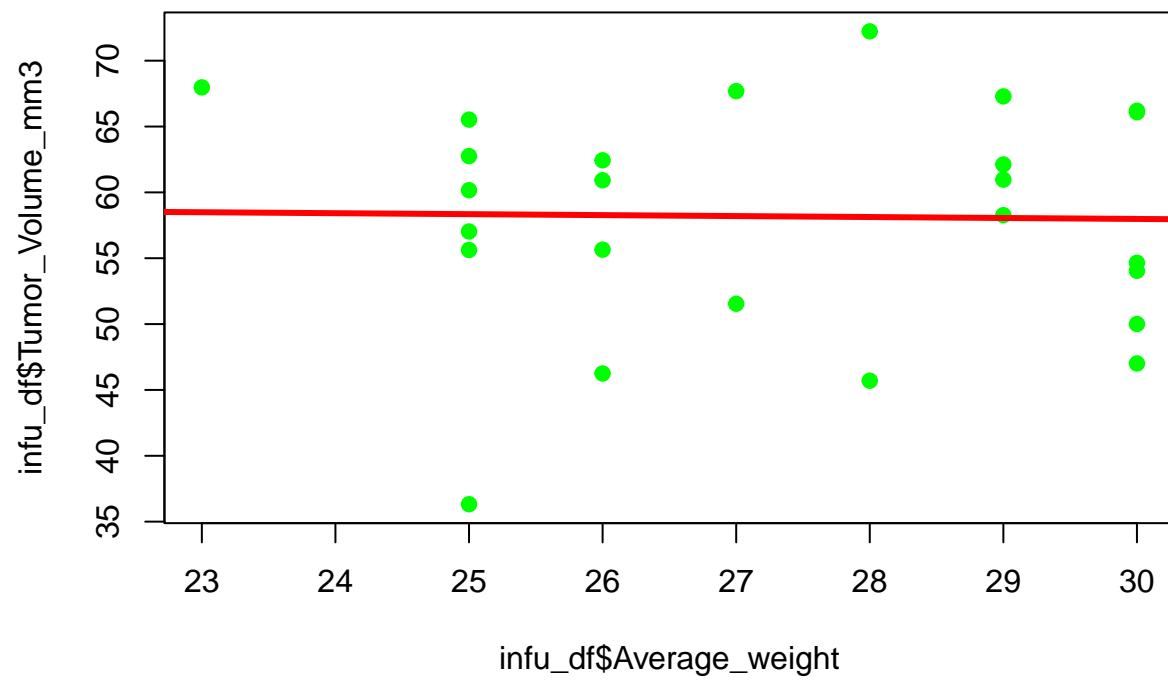


infubinol\_df Vs Weight\_g

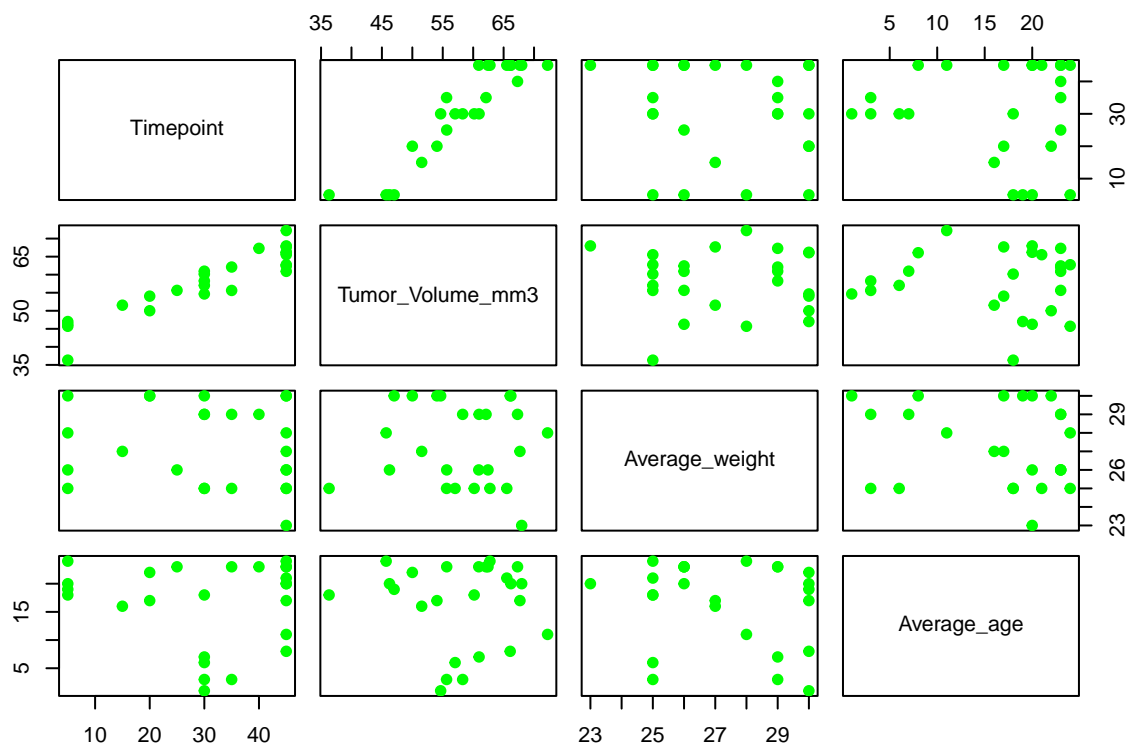
```
# Creating the plot
plot(infu_df$Average_weight, infu_df$Tumor_Volume_mm3, pch = 19, col = "green")

# Regression line
abline(lm(infu_df$Tumor_Volume_mm3 ~ infu_df$Average_weight), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(infu_df$Average_weight, infu_df$Tumor_Volume_mm3), 2)), x = 25, y = 70)
```



```
pairs(infu_df[,2:5], pch = 19, col = "green")
```



ketapril\_df:

```
keta_df1 <- select(ketapril_df, Mouse_Id, Timepoint, Tumor_Volume_mm3) %>%
  group_by(Mouse_Id) %>%
  filter(Timepoint == max(Timepoint, na.rm=TRUE))
```

### Find the average weight by mice\_id in Infubinol\_df

```
keta_df2 <- select(ketapril_df, Mouse_Id, Weight_g) %>%
  group_by(Mouse_Id) %>%
  summarise(Average_weight = mean(Weight_g, na.rm=TRUE))
```

### Joining the two df's for adding average weight

```
keta_df <- keta_df1 %>% inner_join(keta_df2, by = "Mouse_Id")
```

### Find the average age by mice\_id in Capomulin\_df

```
keta_df3 <- select(ketapril_df, Mouse_Id, Age_months) %>%
  group_by(Mouse_Id) %>%
  summarise(Average_age = mean(Age_months, na.rm=TRUE))
```

### Joining the two df's for adding average age

```
keta_df <- keta_df %>% inner_join(keta_df3, by = "Mouse_Id")

head(keta_df)
```

```
## # A tibble: 6 x 5
## # Groups:   Mouse_Id [6]
##   Mouse_Id Timepoint Tumor_Volume_mm3 Average_weight Average_age
##   <chr>      <int>      <dbl>          <dbl>      <dbl>
## 1 a457         10         49.8           30         11
## 2 c580         30         58.0           25         22
## 3 c819         40         62.2           25         21
## 4 c832         45         65.4           29         18
## 5 d474         40         60.2           27         18
## 6 f278         5          48.2           30         12
```

summerize the Tumor\_Volume\_mm3

```
keta_df$Tumor_Volume_mm3 %>%
  summary()
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  45.00  56.72   64.49   62.81  69.87   78.57
```

Standard Deviation

```
keta_df$Tumor_Volume_mm3 %>% sd()
```

```
## [1] 9.94592
```

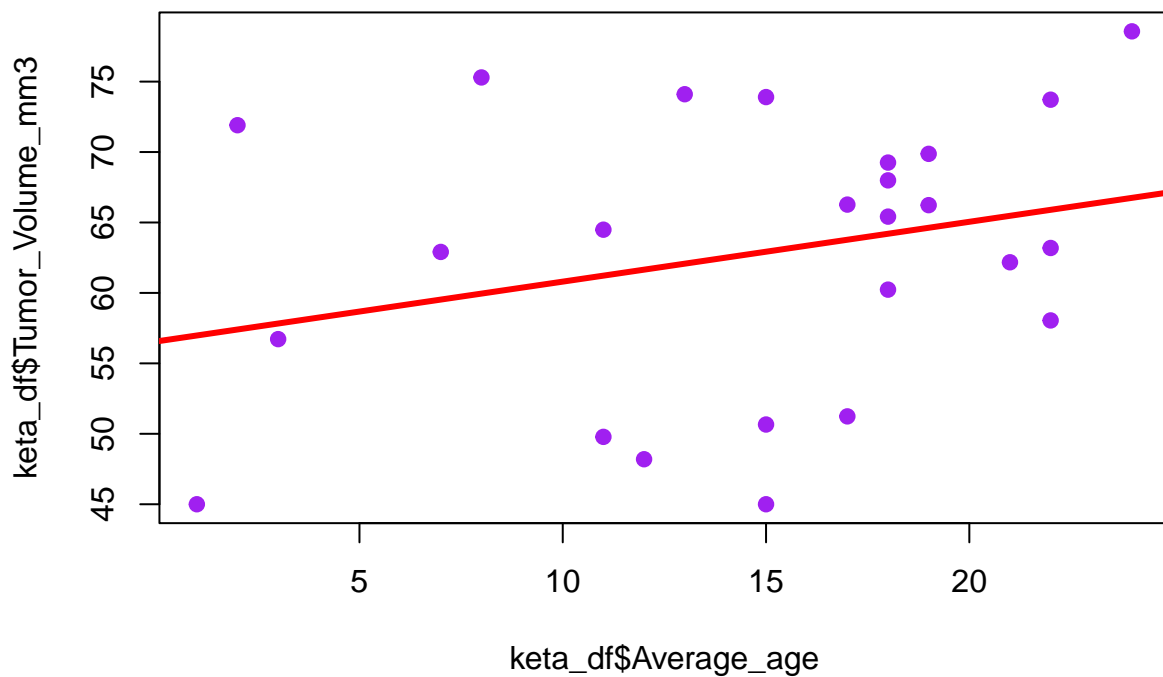
ketapril\_df Vs Age\_months

```
# Creating the plot
plot(keta_df$Average_age, keta_df$Tumor_Volume_mm3, pch = 19, col = "purple")

# Regression line
abline(lm(keta_df$Tumor_Volume_mm3 ~ keta_df$Average_age), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(keta_df$Average_age, keta_df$Tumor_Volume_mm3), 2)), x = 25, y = 90)
```



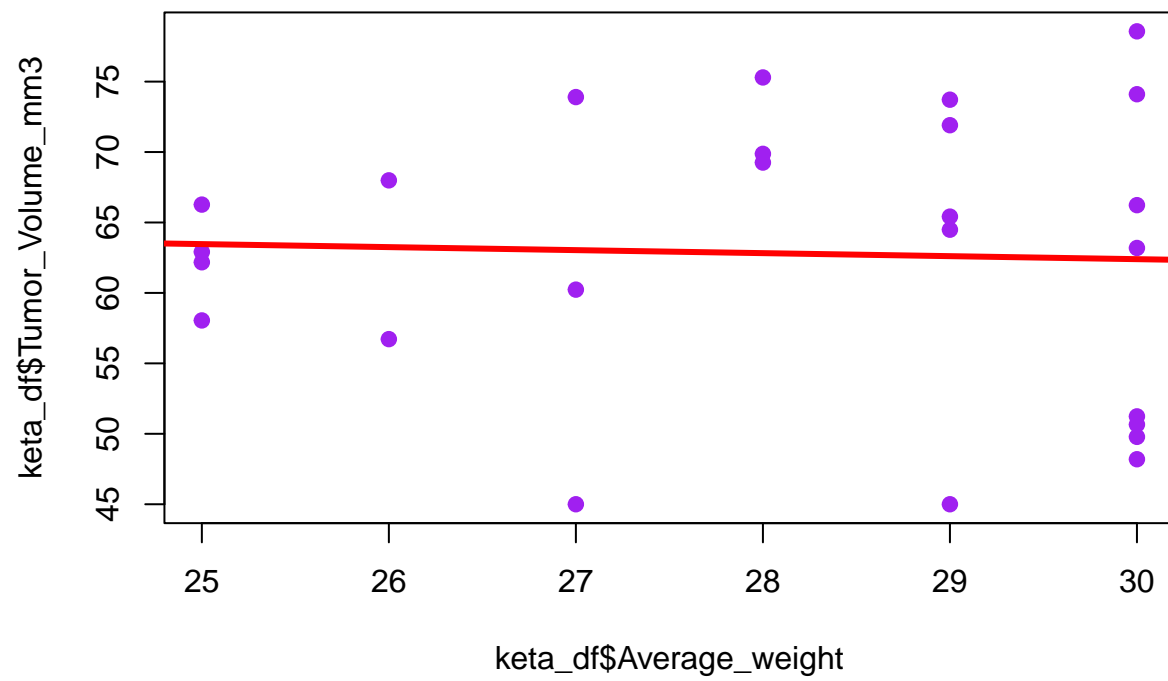


ketapril\_df Vs Weight\_g

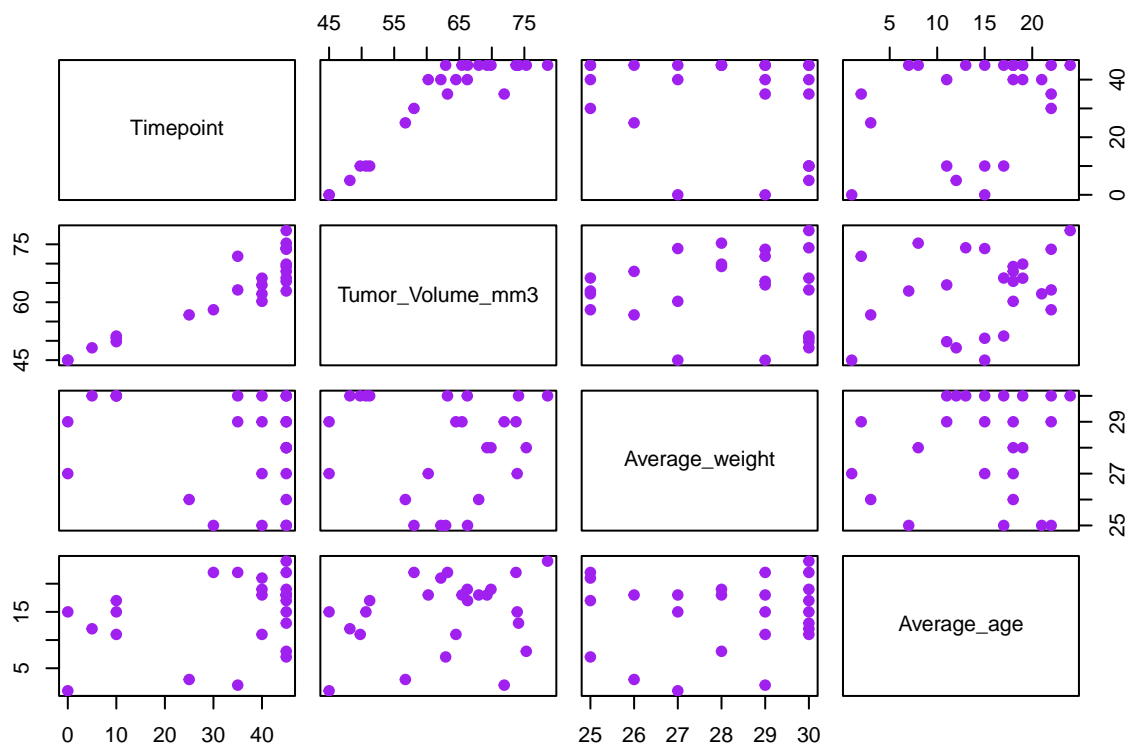
```
# Creating the plot
plot(keta_df$Average_weight, keta_df$Tumor_Volume_mm3, pch = 19, col = "purple")

# Regression line
abline(lm(keta_df$Tumor_Volume_mm3 ~ keta_df$Average_weight), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(keta_df$Average_weight, keta_df$Tumor_Volume_mm3), 2)), x = 25, y = 75)
```



```
pairs(keta_df[,2:5], pch = 19, col = "purple")
```



placebo\_df:

```
plac_df1 <- select(placebo_df, Mouse_Id, Timepoint, Tumor_Volume_mm3) %>%
  group_by(Mouse_Id) %>%
  filter(Timepoint == max(Timepoint, na.rm=TRUE))
```

### Find the average weight by mice\_id in Infubinol\_df

```
plac_df2 <- select(placebo_df, Mouse_Id, Weight_g) %>%
  group_by(Mouse_Id) %>%
  summarise(Average_weight = mean(Weight_g, na.rm=TRUE))
```

### Joining the two df's for adding average weight

```
plac_df <- plac_df1 %>% inner_join(plac_df2, by = "Mouse_Id")
```

### Find the average age by mice\_id in Capomulin\_df

```
plac_df3 <- select(placebo_df, Mouse_Id, Age_months) %>%
  group_by(Mouse_Id) %>%
  summarise(Average_age = mean(Age_months, na.rm=TRUE))
```

### Joining the two df's for adding average age

```
plac_df <- plac_df %>% inner_join(plac_df3, by = "Mouse_Id")

head(plac_df)
```

```
## # A tibble: 6 x 5
## # Groups:   Mouse_Id [6]
##   Mouse_Id Timepoint Tumor_Volume_mm3 Average_weight Average_age
##   <chr>      <int>      <dbl>          <dbl>      <dbl>
## 1 a262         45        70.7            29         17
## 2 a897         45        72.3            28          7
## 3 c282         45        65.8            27         12
## 4 c757         45        69.0            27          9
## 5 c766         45        69.8            26         13
## 6 e227         45        73.2            30          1
```

summerize the Tumor\_Volume\_mm3

```
plac_df$Tumor_Volume_mm3 %>%
  summary()
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  45.00  52.94   62.03   60.51  68.13   73.21
```

Standard Deviation

```
plac_df$Tumor_Volume_mm3 %>% sd()
```

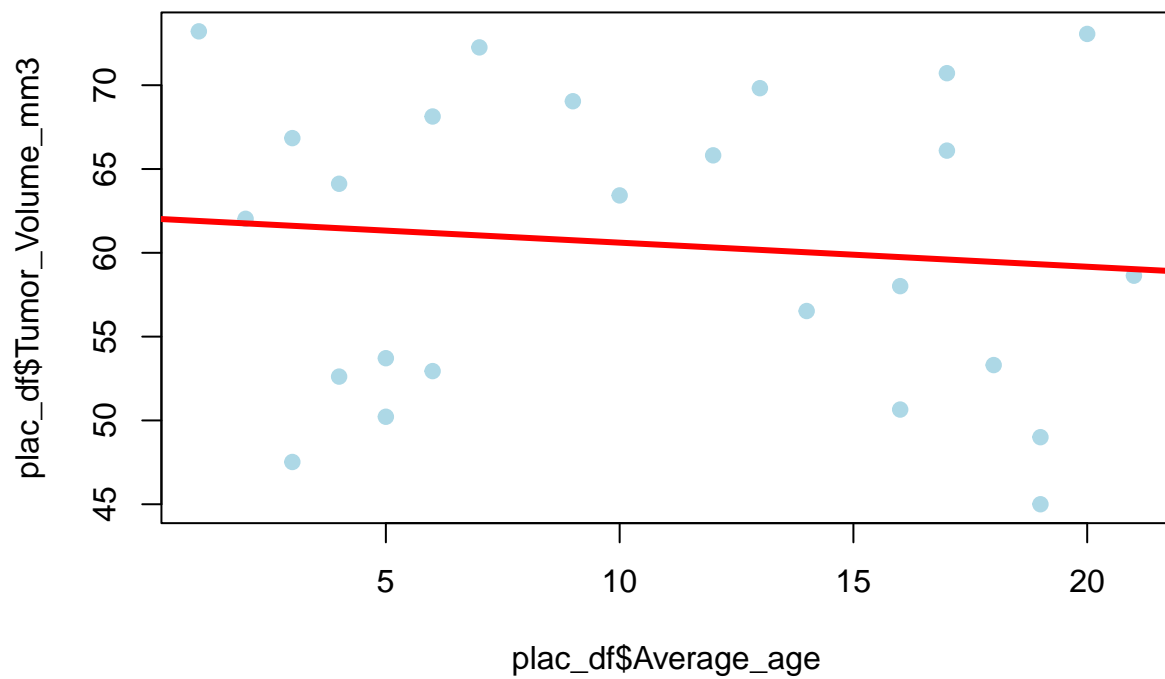
```
## [1] 8.874672
```

placebo\_df Vs Age\_months

```
# Creating the plot
plot(plac_df$Average_age, plac_df$Tumor_Volume_mm3, pch = 19, col = "lightblue")

# Regression line
abline(lm(plac_df$Tumor_Volume_mm3 ~ plac_df$Average_age), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(plac_df$Average_age, plac_df$Tumor_Volume_mm3), 2)), x = 25, y = 9)
```

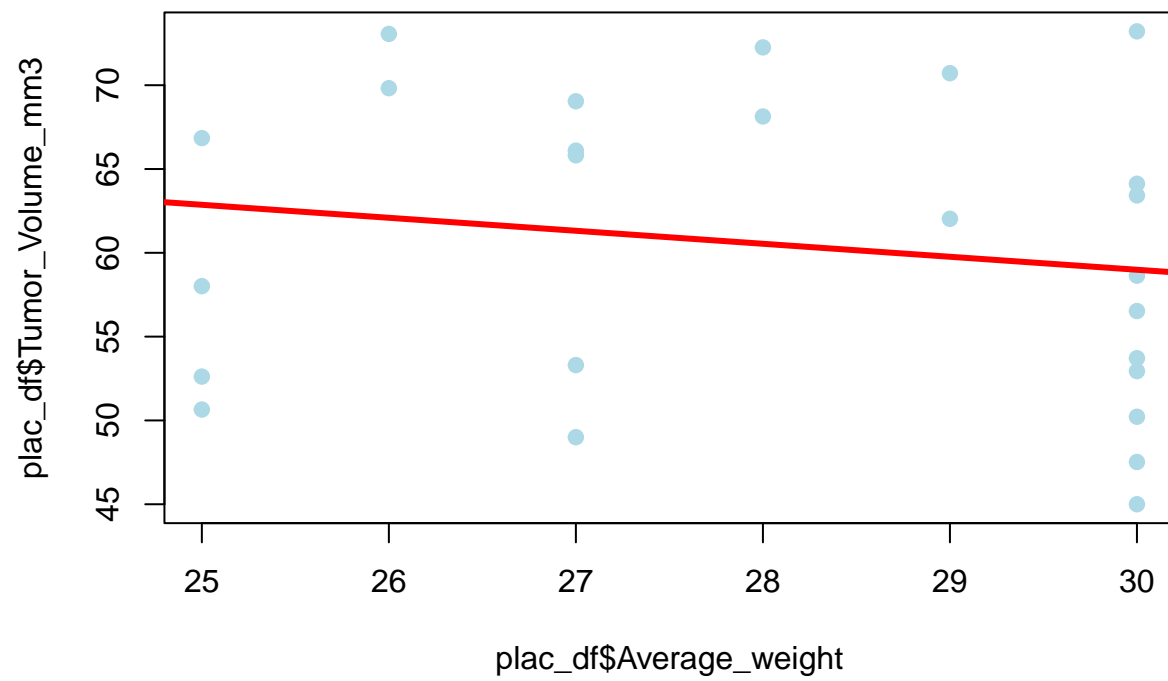


placebo\_df Vs Weight\_g

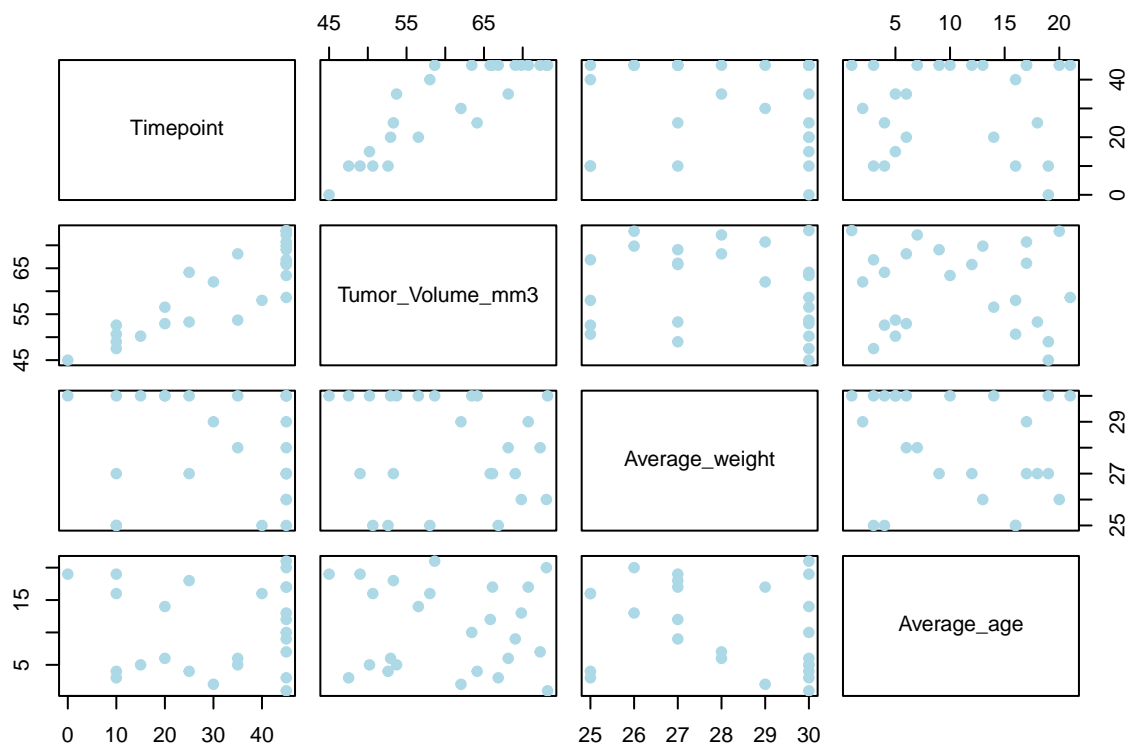
```
# Creating the plot
plot(plac_df$Average_weight, plac_df$Tumor_Volume_mm3, pch = 19, col = "lightblue")

# Regression line
abline(lm(plac_df$Tumor_Volume_mm3 ~ plac_df$Average_weight), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(plac_df$Average_weight, plac_df$Tumor_Volume_mm3), 2)), x = 25, y = 75)
```



```
pairs(plac_df[,2:5], pch = 19, col = "lightblue")
```



### Conclusion:

From the plots above, there seems a correlation between weight and Tumor size for capomulin drug regimen but will be checked by calculating the correlation coefficient.