# Homework 4 Report

## Problem 1: Load the dataset as a new data frame called caffeine.

```
#Load data set and rename as "caffine"

caffine <- read.table("https://raw.githubusercontent.com/ndphillips/ThePiratesGu
                      header = TRUE)
```

Steps:

- Used function read.table() and inserted link to data set
- Renamed data set as “caffine”

## Problem 2: Write R code to calculate and print:

### a. The mean age for each gender

Code:

```
#grouped caffine data set by gender and found mean age for each gender
#renamed this "Avg.age.gender"

Avg.age.gender <- caffine%>%
  group_by(gender)%>%
  summarize(mean_age = mean(age))

print(Avg.age.gender)
```

Result:

```
# A tibble: 2 × 2
  gender mean_age
  <chr>      <dbl>
1 female      24.6
2 male        24.9
```

Steps:

- Grouped the caffine dataset by gender
- Used summarize function to find the mean age for each gender
- Renamed this “Avg.age.gender”

Conclusions:

- Females in the study had an average age of 24.6
- Males in the study had an average age of 24.9
- Males had a slighter higher average age

### b. The mean age for each drink

Code:

```
#grouped caffine data set by drink and found mean age for each drink
#renamed this "Avg.age.drink"

Avg.age.drink <- caffine%>%
  group_by(drink)%>%
  summarize(mean_age = mean(age))

print(Avg.age.drink)
```

Result:

```
# A tibble: 2 × 2
  drink     mean_age
  <chr>        <dbl>
1 coffee        25.2
2 greentea      24.3
```

Steps:

- Grouped the caffine data set by drink
- Used summarize function to find the mean age for each drink
- Renamed this “Avg.age.drink”

Conclusions:

- The average age of people who tested coffee was 25.2
- The average age of people who tested green tea was 24.3

### c. The mean age for each combined level of both gender and drink

Code:

```
#grouped caffine by gender and drink. Found the mean age for each combined level
#renamed this "Avg.age.combined"

Avg.age.combined <- caffine%>%
  group_by(gender,drink)%>%
  summarize(mean_avg = mean(age))

print(Avg.age.combined)
```

Result:

```
  gender drink     mean_avg
  <chr>  <chr>        <dbl>
1 female coffee        25.2
2 female greentea      23.7
3 male   coffee        25.2
4 male   greentea      24.8
```

Steps:

- Grouped the caffine data set by gender and drink
- Used the summarize function to mean age for each combined level
- Renamed this “Avg.age.combined”

Conclusions:

- The average age for females who drank coffee was 25.2
- The average age for females who drank green tea was 23.7
- The average age for males who drank coffee was 25.2
- The average age for males who drank green tea was 24.8
- The average ages for males and females who drank coffee was the highest

### d. The median score for each age

Code:

```
#grouped caffine dataset by age and found median score for each age
#renamed this "Med.score"

Med.score <- caffine%>%
  group_by(age)%>%
  summarize(mid_score = median(score))

print(Med.score)
```

Result:

| | age <int> | mid_score <dbl> |
|---|---|---|
| 1 | 18 | 21.2 |
| 2 | 19 | 20.3 |
| 3 | 20 | 30.2 |
| 4 | 21 | 20.0 |
| 5 | 22 | 21.8 |
| 6 | 23 | 32.9 |
| 7 | 24 | 42.7 |
| 8 | 25 | 20.5 |
| 9 | 26 | 46.2 |
| 10 | 27 | 23.1 |
| 11 | 28 | 32.3 |
| 12 | 29 | 24.9 |
| 13 | 30 | 13.0 |
| 14 | 31 | 9.88 |
| 15 | 32 | 7.85 |
| 16 | 33 | 14.0 |

Steps:

- Grouped caffine data set by age
- Used summarize function to find the median score for each age
- Renamed this “Med.score”

Conclusions:

- The lowest median score was 7.85 for age group 32
- The highest median score was 46.2 for age group 26

## Problem 3: For mean only, write R code to calculate and print the maximum score for each age

Code:

```
#filtered for only males. grouped by age. found max score for each age
#renamed this "caffine_men"

caffine_men <- caffine%>%
  filter(gender == "male")%>%
  group_by(age)%>%
  summarize(max_score = max(score))

print(caffine_men)
```

Result:

| | age <int> | max_score <dbl> |
|---|---|---|
| 1 | 18 | 21.2 |
| 2 | 21 | 55.8 |
| 3 | 22 | 11.0 |
| 4 | 23 | 56.9 |
| 5 | 24 | 55.2 |
| 6 | 25 | 72.7 |
| 7 | 26 | 61.2 |
| 8 | 27 | 38.5 |
| 9 | 28 | 62.6 |
| 10 | 29 | 38.3 |
| 11 | 30 | 9.51 |
| 12 | 33 | 14.0 |

Steps:

- Filtered caffine data set for only men
- Grouped by age
- Used summarize function to find the max score for each age
- Renamed this “caffine_men”

Conclusions:

- Men’s highest score is 72.7 for age group 25
- Men’s lowest score is 9.51 for age group 30

## Problem 4: Create a data frame showing, for each level of drink, the mean, median, maximum, and standard deviation of scores.

Code:

```
#grouped caffine data set by drink
#found mean, median, max and standard deviation of the score for each drink
#named this data set "caffine_df"

caffine_df <- caffine%>%
  group_by(drink)%>%
  summarize(Mean = mean(score),
            Median = median(score),
            Max = max(score),
            SD = sd(score))

#made caffine_df into a dataframe

df<- data.frame(caffine_df)
print(df)
```

Result:

```
     drink    Mean Median   Max       SD
1   coffee 35.2146  36.40 73.64 25.68715
2 greentea 29.9718  28.68 46.73 10.24395
```

Steps:

- Grouped caffine data set by drink
- Used the summarize function to find the mean, median, max and standard deviation of scores for each drink
- Renamed this “caffine_df”
- Made this into a data frame

Conclusions:

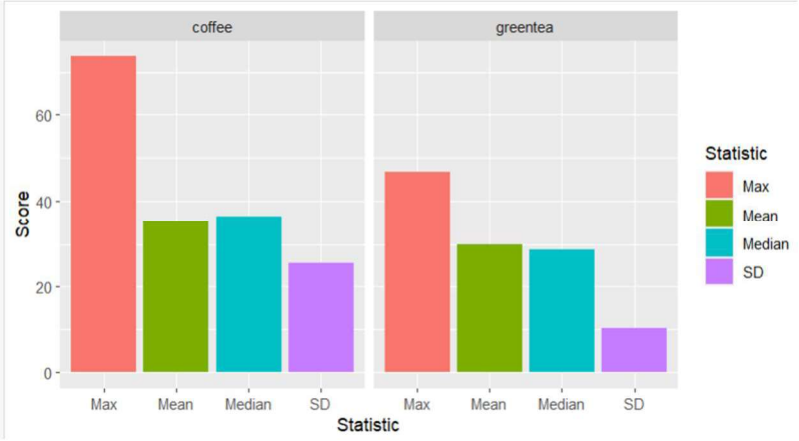
- Coffee has a higher mean, median, max, and standard deviation score than green tea
- Green tea has a lower mean, median, max, and standard deviation score than coffee

## Problem 5: Write R code to plot the contents of the data frame created in the previous step in a visually pleasant and informative way.

Code:

```
#reshaped the df dataframe so there are three columns:
##  x (drink), y (values from df) and group

df_reshaped <- data.frame(drink = df$drink,
                          y = c(df$Mean, df$Median, df$Max, df$SD),
                          Statistic = c(rep("Mean", nrow(df)),
                                        rep("Median", nrow(df)),
                                        rep("Max", nrow(df)),
                                        rep("SD", nrow(df))))

#made bar graph with x as Statistic and y as Score
#for each drink I graphed four bars: the max, mean, median and SD
#separated the drink with facet wrap

ggplot(df_reshaped, mapping = aes(x = Statistic, y, fill = Statistic))+
  geom_col(position = "dodge")+
  facet_wrap(~ drink)+
  labs(x = "Statistic",
       y = "Score")
```

Result:



Steps:

- Reshaped the dataframe so there are three columns:
  - X (for drink)
  - Y (for values from df)
  - Statistic for the type of statistical measure performed
- Renamed this df_reshaped
- Made bar graph out of df_reshaped data with x as Statistic and y as Score
- For each drink I graphed four bars: the max, mean, median, and standard deviation. I did this with the function "position = dodge"
- With the facet wrap function, I made a ribbon of panels separated by drink.

Conclusions:

- Coffee has a higher mean, median, max, and standard deviation score than green tea
- Green tea has a lower mean, median, max, and standard deviation score than coffee

## Problem 6: Only for females above the age of 20, create a table showing for each level of drink and cups, the mean, median, maximum, and standard deviation of scores. Also, include a column showing how many people were in each group.

Code:

```
#filtered caffine data set for females above the age of 20
#grouped the data set by drink and cups
#found the mean, median, max, SD, and count of each level with summarize function
#renamed this "caffine_fem"

caffine_fem <-caffine%>%
  filter(gender == "female", age > 20)%>%
  group_by(drink, cups)%>%
  summarize(Mean = mean(score),
            Median = median(score),
            Max = max(score),
            SD = sd(score),
            Count = n())
caffine_fem
```

Result:

| | drink | cups | Mean | Median | Max | SD | Count |
|---|---|---|---|---|---|---|---|
| | <chr> | <int> | <dbl> | <dbl> | <dbl> | <dbl> | <int> |
| 1 | coffee | 1 | 9.40 | 8.67 | 14.2 | 2.46 | 13 |
| 2 | coffee | 5 | 59.8 | 58.7 | 73.6 | 6.15 | 11 |
| 3 | greentea | 1 | 22.1 | 20.6 | 29.5 | 4.48 | 11 |
| 4 | greentea | 5 | 38.6 | 38.4 | 45.1 | 4.71 | 9 |

Steps:

- Filtered caffine data set for females above the age of 20
- Grouped the data set by drink and cups
- Found the mean, median, max, standard deviation, and count of each level with the summarize function

- Renamed this “caffine_fem”

Conclusions:

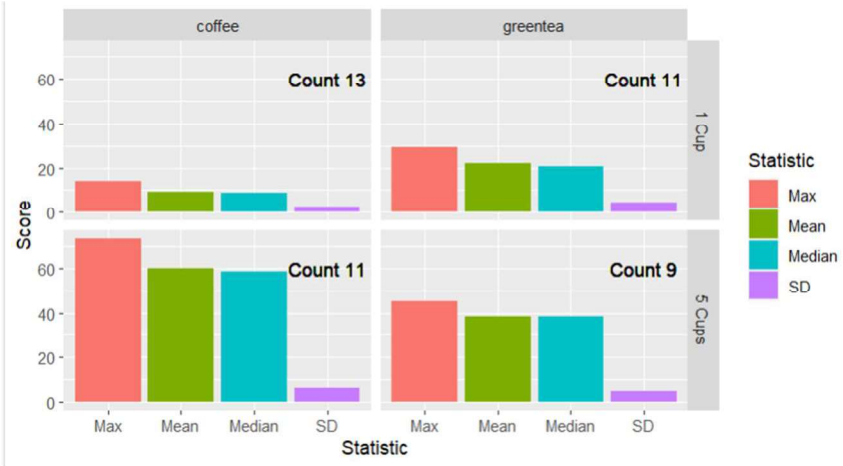
- 5 cups of coffee have the highest mean, median, max, and standard deviation
- 1 cup of coffee has the lowest mean, median, max, and standard deviation

## Problem 7: Write R code to plot the contents of the table created in the previous step in a visually pleasant and informative way

Code:

```
#reshaped caffine_fem so there are 5 columns:
# drink, cups, SampleSize, y (statistical values), Statistic (type of statistic)

caffine_fem_reshaped <- data.frame(drink = c(caffine_fem$drink),
                                   Cups = c(caffine_fem$cups),
                                   SampleSize = c(caffine_fem$Count),
                                   y = c(caffine_fem$Mean,
                                         caffine_fem$Median,
                                         caffine_fem$Max,
                                         caffine_fem$SD),
                                   Statistic = c(rep("Mean", nrow(caffine_fem)),
                                                 rep("Median", nrow(caffine_fem)),
                                                 rep("Max", nrow(caffine_fem)),
                                                 rep("SD", nrow(caffine_fem))))

#renamed "1" and "5" to "1 cup" and "5 cups" in Cups column

caffine_fem_reshaped$Cups <- factor(caffine_fem_reshaped$Cups, levels = c("1", "5"),
                                    labels = c("1 Cup", "5 Cups"))

#renamed "13", "11", and "9" to "Count 13", "Count 11", and "Count 9" in SampleSize column

caffine_fem_reshaped$SampleSize <- factor(caffine_fem_reshaped$SampleSize,
                                          levels = c("13", "11", "9"),
                                          labels = c("Count 13", "Count 11", "Count 9"))

#made bar graph using caffine_fem_reshaped data set
#x as Statistic, y as score, and fill by statistic
#separated the drink and # of cups using facet grid
#added text to graph with each groups sample size

ggplot(caffine_fem_reshaped, mapping = aes(x = Statistic, y,
                                           fill = Statistic))+
  geom_col(position = "dodge")+
  facet_grid(Cups ~ drink)+
  geom_text(x = 4, y = 60, label = caffine_fem_reshaped$SampleSize)+
  labs(x = "Statistic",
       y = "Score")
```

Result:




Steps:

- Reshaped the caffine_fem data set so there are 5 coulmns:
  - Drinks

    - Cups
    - SampleSize
    - Y (statistical values)
    - Statistic (type of statistic)
- Renamed this caffine_fem_reshaped
- In the cups column, renamed values “1” and “5” to “1 cup” and “5 cups”
- In the SampleSize column, renamed values “13”, “11”, and “9” to “Count 13”, “Count 11”, “Count 9”
- Made a bar graph using caffine_fem_reshaped data set with X as statistic and Y as Score
- For each drink I graphed four bars: the max, mean, median, and standard deviation. I did this with the function “position = dodge”
- With the facet wrap function, I made a ribbon of panels separated by drink and cups.
- Added text to the graph with each groups sample size, or count

Conclusions:

- 5 cups of coffee have the highest mean, median, max, and standard deviation
- 1 cup of coffee has the lowest mean, median, max, and standard deviation

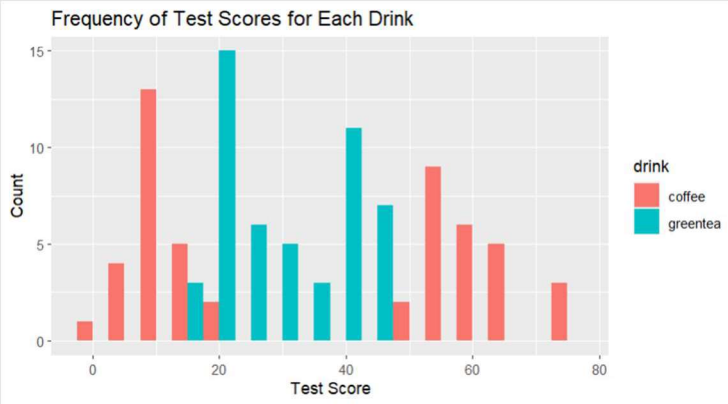
## Problem 8: Write R code to answer the following questions:

1. What is the correlation between test scores and type of drink?

Code:

```
### Part 1 (correlation between test scores and type of drink)


#Made a histogram to show correlation between test scores and type of drink
#used caffine data set, and x as score. y is the frequency of each score.
#categorized drink by color

ggplot(caffine, mapping = aes(x = score, fill = drink))+
  geom_histogram(position = "dodge", binwidth = 5)+
  labs(title = "Frequency of Test Scores for Each Drink",
       x = "Test Score",
       y = "Count")

#Coffee has the more extreme scores while greentea has moderate scores.
```

Result:




Steps:

- Made a histogram to show the correlation between test scores and type of drink
- Used the caffine data set with x as score and y as frequency of each score
- Categorized drink by color

Conclusions:

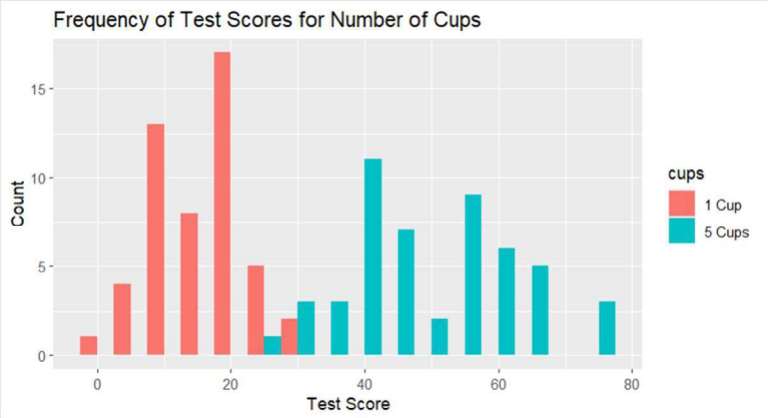
- Coffee results in the more extreme scores (low and high)
- Green tea results in moderate scores that fall in the middle

## 2. What is the correlation between test scores and the amount of drink?

Code:

```
### Part 2 (correlation between test scores and amount of drink)

#Renamed values 1 and 5 to 1 Cup and 5 Cups
caffine$cups <- factor(caffine$cups, levels = c("1", "5"),
                       labels = c("1 Cup", "5 Cups"))


#Made a histogram to show correlation between test scores and amount of drink
#used caffine data set, and x as score. y is the frequency of each score.
#categorized cups by color

ggplot(caffine, mapping = aes(x = score, fill = cups))+
  geom_histogram(position = "dodge", binwidth = 5)+
  labs(title = "Frequency of Test Scores for Number of Cups",
       x = "Test Score",
       y = "Count")

# 5 Cups results in higher scores and 1 cup results in lower scores
```

Result:




Steps:

- Renamed values 1 and 5 in the cups column to “1 Cup” and “5 Cups”
- Made a histogram to show the correlation between test scores and amount of drink
- Used the caffine data set with x as score and y as frequency of each score
- Categorized cups by color

Conclusions:

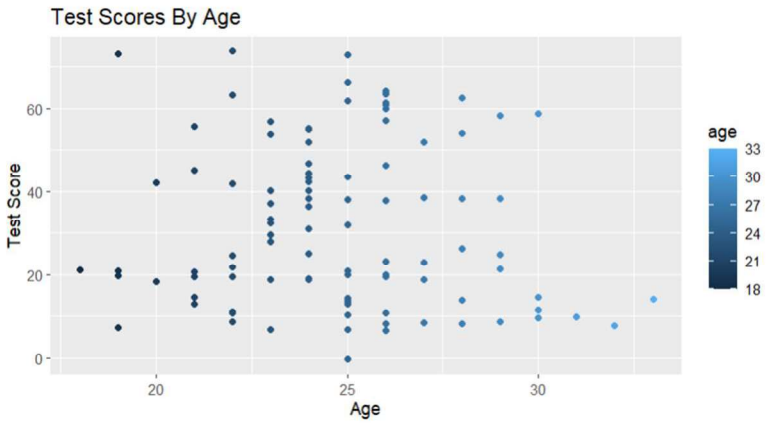
- 5 cups of beverage result in higher test scores
- 1 cup of beverage result in lower test scores
- Larger amount of drink consumed correlates with higher test scores

## 3. What is the correlation between test scores and age?

Code:

```
### Part 3 (correlation between test scores and age)

#Made scatter plot graph to show the correlation between test scores and age
#used caffine data set. x as age and y as score
#categorized age by color

ggplot(caffine, mapping = aes(x = age, y = score, color = age))+
  geom_point()+
  labs(title = "Test Scores By Age",
       x = "Age",
       y = "Test Score")

#No correlation between test scores and age
```

Result:




Steps:

- Made a scatter plot graph to show the correlation between test scores and age
- Used the caffine data set with x as Age and y as Test Scores
- Categorized age by color

Conclusions:

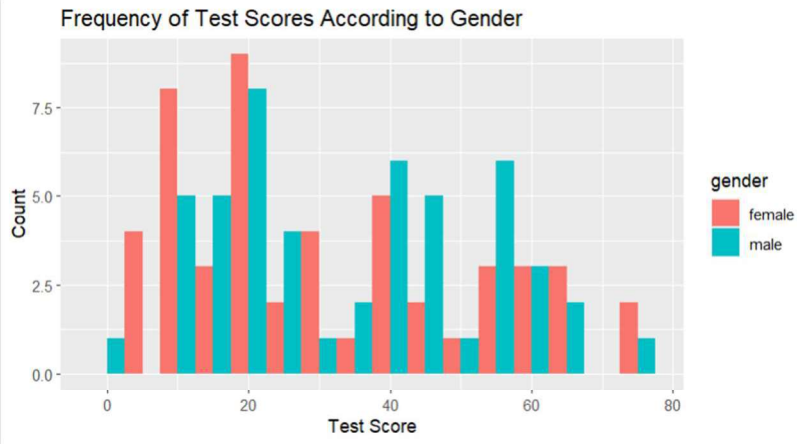
- No correlation between test scores and age

## 4. What is the correlation between test scores and gender?

Code:

```
### Part 4 (correlation between test scores and gender)

#Made a histogram to show the correlation between test scores and gender
#used caffine data set, and x as score. y is the frequency of each score.
#categorized gender by color

ggplot(caffine, mapping = aes(x = score, fill = gender))+
  geom_histogram(position = "dodge", binwidth = 5)+
  labs(title = "Frequency of Test Scores According to Gender",
       x = "Test Score",
       y = "Count")

#No correlation between test scores and gender
```

Result:




Steps:

- Made a histogram to show the correlation between test scores and gender
- Used caffine data set with x as score and y as frequency of test scores
- Categorized gender by color

Conclusions:

- No correlation between test scores and gender

5. What can we conclude from this study? (and how does the data support each conclusion?)

I can conclude from this study that the number of drink cups has the greatest effect on test scores. This effect is most obvious with coffee. So, the more coffee a person consumes, the more likely they will score higher. This is not true for green tea. The score difference doesn't range as greatly as it does with coffee.

I can also conclude that there is no correlation between test scores and age, and test scores and gender.

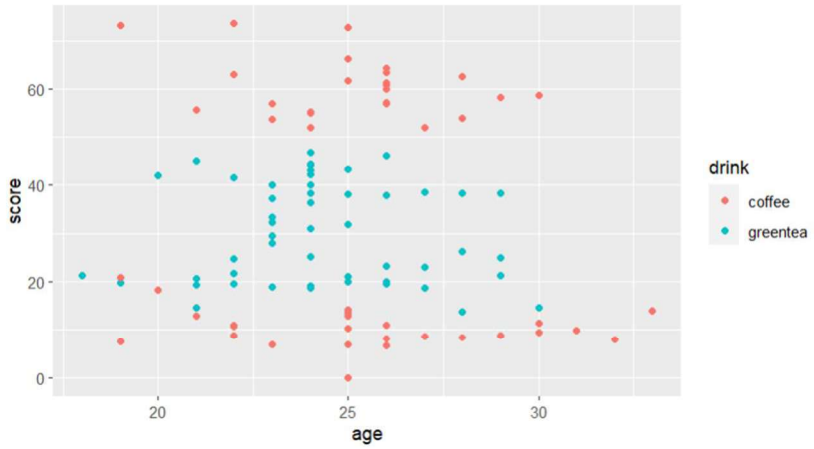
## Problem 9:

1. Write R code to generate a scatterplot of the performance versus age, for different types of drinks and comment on the result.

Code:

```
###Part 1

#Made scatter plot with x as age, y as score
#categorized drink by color

ggplot(caffine, mapping = aes(x = age, y = score, color = drink))+
  geom_point()
```

Result:




Steps:

- Made scatter plot using caffine data set with x as age and y as score
- Categorized drink by color

Conclusions:

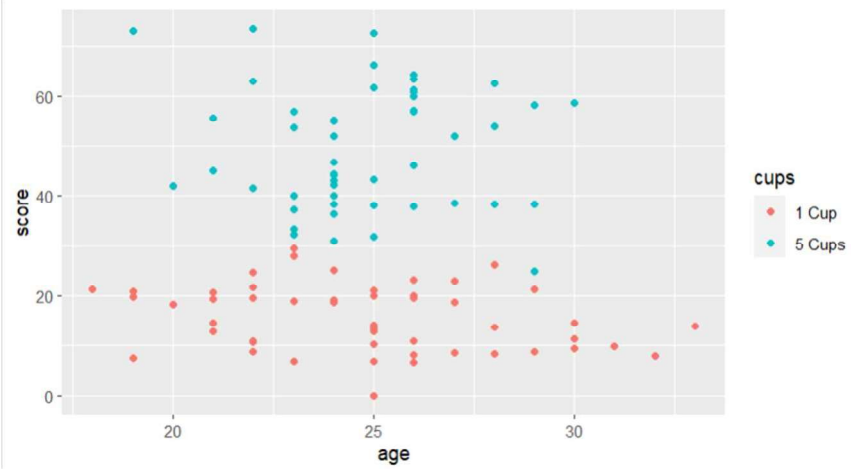
- No correlation between test scores and age
- Coffee has a larger range of scores

2. Write R code to generate a scatterplot of the performance versus age, for different amounts of drink and comment on the result.

Code:

```
###Part 2

#Made scatter plot with x as age, y as score
#categorized cups by color

ggplot(caffine, mapping = aes(x =  age, y = score, color = cups))+
  geom_point()
```

Result:




Steps:

- Made scatter plot using caffine data set with x as age and y as score
- Categorized number of cups by color

Conclusions:

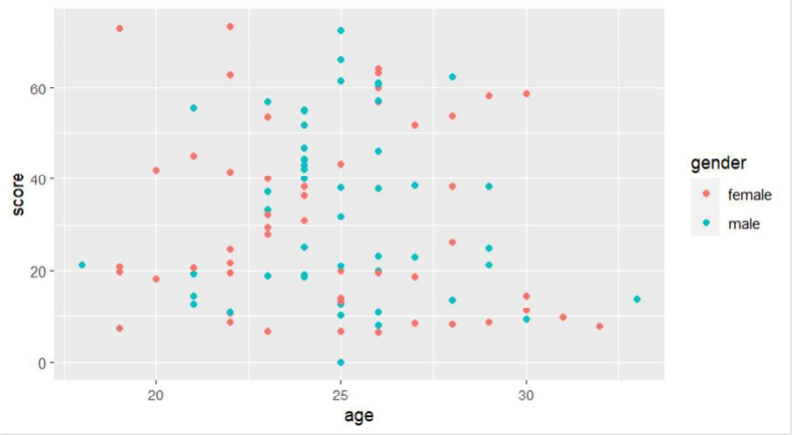
- No correlation between age and test scores
- 5 cups result in a higher score
- 1 cup results in a lower score

3. Write R code to generate a scatterplot of the performance versus age, for different genders and comment on the result.

Code:

```
###Part 3

#Made scatter plot with x as age, y as score
#categorized gender by color

ggplot(caffine, mapping = aes(x = age, y = score, color = gender))+
  geom_point()
```

Result:




Steps:

- Made scatter plot using caffine data set with x as age and y as score
- Categorized gender by color

Conclusions:

- No correlation between age and test scores
- No correlation between gender and test scores

4. What can you conclude from this study based on the three scatterplots above?

I can conclude that there is no correlation between age and test scores for any of the different drinks, cups, or gender. The number of cups has the most effect on test scores. As the number of cups increases, test score increases. Coffee has the largest range in test scores. So, the number of cups of coffee influences test score. Green tea has a short range in test scores. So, the number of cups of green tea does not have a great effect on green tea.