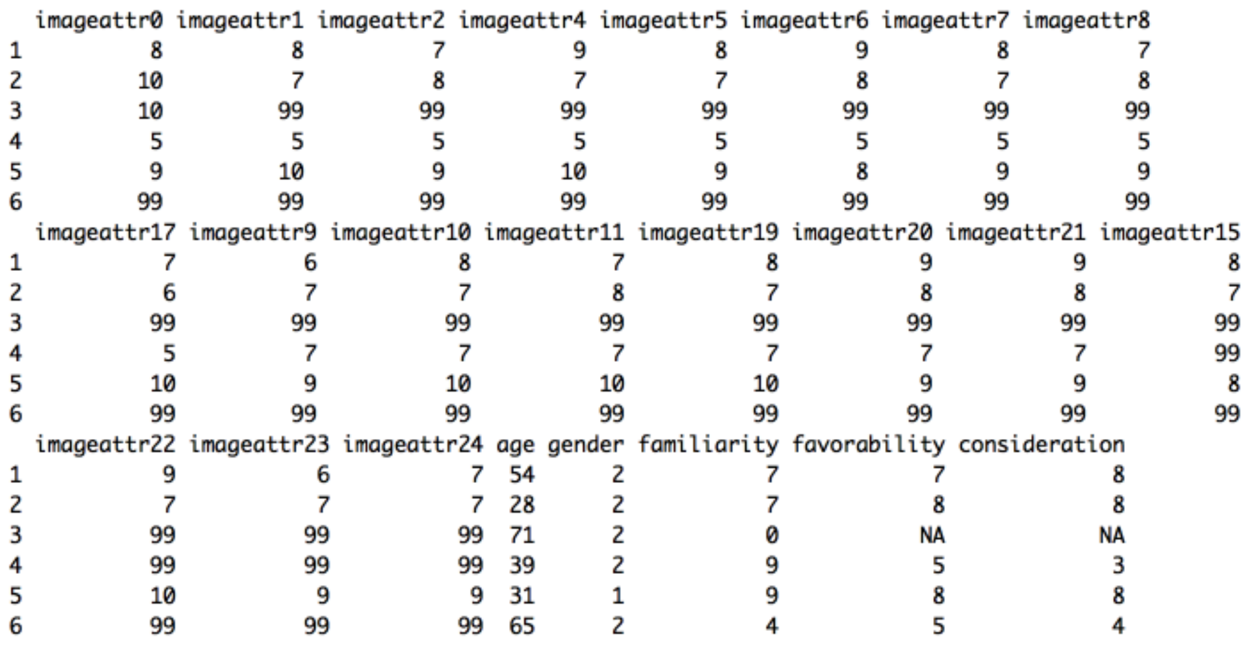
Question 1

Syntax:

```
# Check the file exist in working directory

if(!file.exists("ME_DA_Test.txt")){
    # Read the the file if not available in the working directory
    row_data <- read.table(file.choose(), sep = "\t", header = TRUE)
    head(row_data)
}else{

   row_data <- read.table("ME_DA_Test.txt", sep = "\t", header = TRUE)
   head(row_data)
}
```

Output:



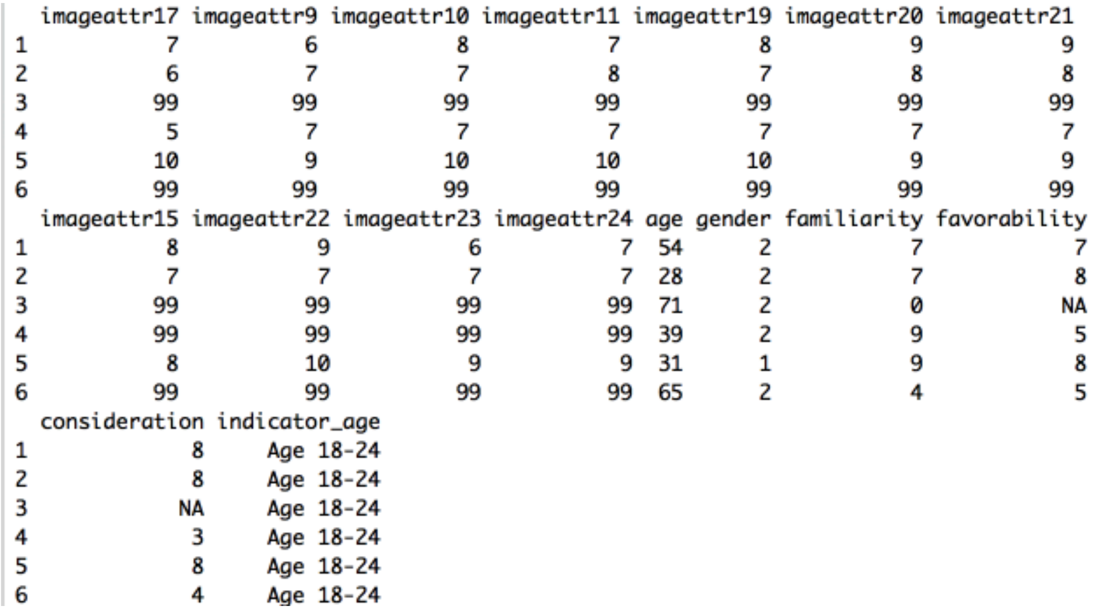
```
  imageattr0 imageattr1 imageattr2 imageattr4 imageattr5 imageattr6 imageattr7 imageattr8
1          8          8          7          9          8          9          8          7
2         10          7          8          7          7          8          7          8
3         10         99         99         99         99         99         99         99
4          5          5          5          5          5          5          5          5
5          9         10          9         10          9          8          9          9
6         99         99         99         99         99         99         99         99
  imageattr17 imageattr9 imageattr10 imageattr11 imageattr19 imageattr20 imageattr21 imageattr15
1           7          6           8           7           8           9           9           8
2           6          7           7           8           7           8           8           7
3          99         99          99          99          99          99          99          99
4           5          7           7           7           7           7           7          99
5          10          9          10          10          10           9           9           8
6          99         99          99          99          99          99          99          99
  imageattr22 imageattr23 imageattr24 age gender familiarity favorability consideration
1           9           6           7  54      2           7            7             8
2           7           7           7  28      2           7            8             8
3          99          99          99  71      2           0           NA            NA
4          99          99          99  39      2           9            5             3
5          10           9           9  31      1           9            8             8
6          99          99          99  65      2           4            5             4
```

Question 2

Syntax:

```
# function to identify age category
age_indicator_function <- function(data,col){
    no_row = nrow(data);
    new_col = c()
    for(i in 1:no_row)
    {
       if((data[i,col] >= 18) && (data[i,col] <= 24))
       {
          new_col[i] <- "Age 18-24"
       }
       else if((data[i,col] >= 25) && (data[i,col] <= 44))
       {
          new_col[i] <- "Age 18-24"
       }
       else if((data[i,col] >= 45) && (data[i,col] <= 64))
       {
          new_col[i] <- "Age 18-24"
       }
       else
       {
          new_col[i] <- "Age 18-24"
       }
    }
    return (new_col)
 }
# function call
indicator_age <- age_indicator_function(row_data,"age")
row_data <- cbind(row_data,indicator_age)  #Attach new column to data using cbind function
head(row_data)
```

Output:



```
  imageattr17 imageattr9 imageattr10 imageattr11 imageattr19 imageattr20 imageattr21
1           7          6           8           7           8           9           9
2           6          7           7           8           7           8           8
3          99         99          99          99          99          99          99
4           5          7           7           7           7           7           7
5          10          9          10          10          10           9           9
6          99         99          99          99          99          99          99
  imageattr15 imageattr22 imageattr23 imageattr24 age gender familiarity favorability
1           8           9           6           7  54      2           7            7
2           7           7           7           7  28      2           7            8
3          99          99          99          99  71      2           0           NA
4          99          99          99          99  39      2           9            5
5           8          10           9           9  31      1           9            8
6          99          99          99          99  65      2           4            5
  consideration indicator_age
1             8     Age 18-24
2             8     Age 18-24
3            NA     Age 18-24
4             3     Age 18-24
5             8     Age 18-24
6             4     Age 18-24
```

Question 3

Syntax (Function 1 to 3)

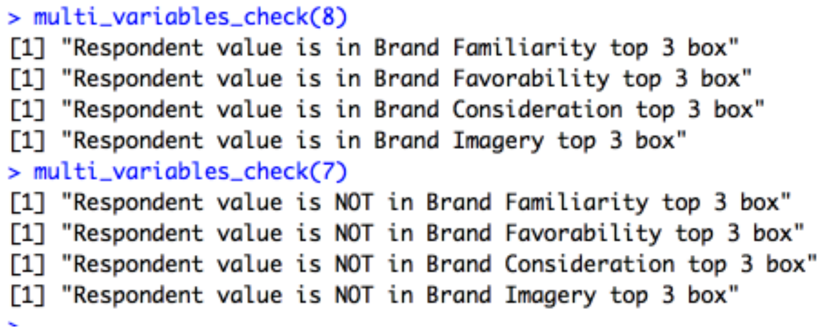
Function 1: max3_values function

```
#Returns top 3 box
max3_values <- function(dataset)
{
   # Return top 4 values in specified database
   max_3 <- head(unique(sort(dataset, decreasing = TRUE, index.return = FALSE)),4)
   # print(max_3)
   # check correct value.
   if (any(max_3==99))
   {
      #print("99 is avalable")
      max_3 <- max_3[max_3 != 99]
      #print(max_3)
   }else{
      #print("99 is not avalable")

      max_3 <- head(max_3,3)
      #print(max_3)
   }
}
```

Function 3: check_top_3 function

```
# Returns the respond value is top 3 box
check_top_3 <- function(respondent_value,data_set)
{
   top_3 <- max3_values(data_set)
   if(any(top_3 == respondent_value))
   {
      return(1)
   }
   else{
      return(0)
   }
}
```

Output:



```
> multi_variables_check(8)
[1] "Respondent value is in Brand Familiarity top 3 box"
[1] "Respondent value is in Brand Favorability top 3 box"
[1] "Respondent value is in Brand Consideration top 3 box"
[1] "Respondent value is in Brand Imagery top 3 box"
> multi_variables_check(7)
[1] "Respondent value is NOT in Brand Familiarity top 3 box"
[1] "Respondent value is NOT in Brand Favorability top 3 box"
[1] "Respondent value is NOT in Brand Consideration top 3 box"
[1] "Respondent value is NOT in Brand Imagery top 3 box"
```

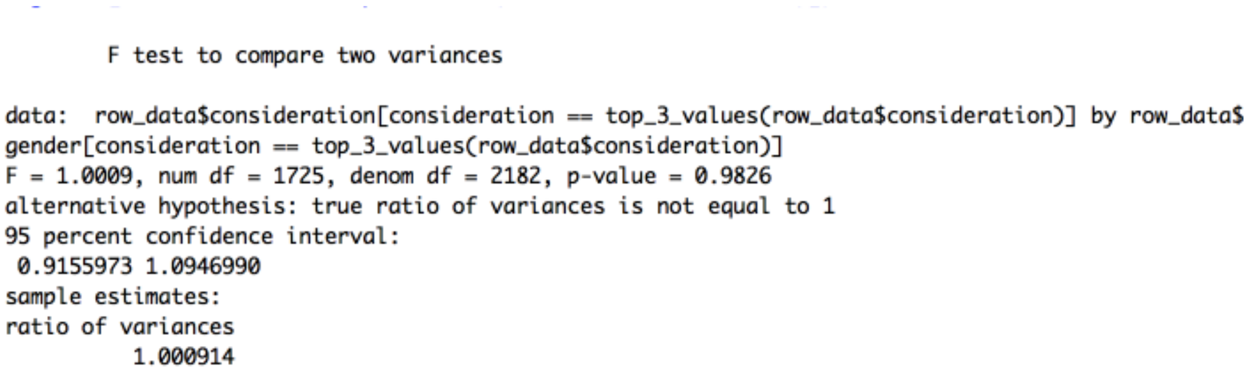
Function 3: **multi_variable_ckeck function**

```
# Multi variable check function
multi_variables_check <- function(respondent_value)
{
   #Brand Familiarity
   check_brand_familiarity <- check_top_3(respondent_value, row_data$familiarity)
   if(check_brand_familiarity)
   {print("Respondent value is in Brand Familiarity top 3 box")}
   else
   {print("Respondent value is NOT in Brand Familiarity top 3 box")  }

   #Brand Favorability
   check_brand_favorability <- check_top_3(respondent_value, row_data$favorability)
   if(check_brand_favorability)
   { print("Respondent value is in Brand Favorability top 3 box")}

   else
   {print("Respondent value is NOT in Brand Favorability top 3 box")}

   #Brand Consideration
   check_brand_consideration <- check_top_3(respondent_value, row_data$consideration)
   if(check_brand_consideration)
   { print("Respondent value is in Brand Consideration top 3 box")}
   else
   {  print("Respondent value is NOT in Brand Consideration top 3 box") }

  #Brand Imagery
  imageatt0 <- check_top_3(respondent_value, row_data$imageattr0)
  imageatt1 <- check_top_3(respondent_value, row_data$imageattr1)
  imageatt2 <- check_top_3(respondent_value, row_data$imageattr2)
  imageatt4 <- check_top_3(respondent_value, row_data$imageattr4)
  imageatt5 <- check_top_3(respondent_value, row_data$imageattr5)
  imageatt6 <- check_top_3(respondent_value, row_data$imageattr6)
  imageatt7 <- check_top_3(respondent_value, row_data$imageattr7)
  imageatt8 <- check_top_3(respondent_value, row_data$imageattr8)
  imageatt9 <- check_top_3(respondent_value, row_data$imageattr9)
  imageatt10 <- check_top_3(respondent_value, row_data$imageattr10)
  imageatt11 <- check_top_3(respondent_value, row_data$imageattr11)
  imageatt15 <- check_top_3(respondent_value, row_data$imageattr15)
  imageatt17 <- check_top_3(respondent_value, row_data$imageattr17)
  imageatt19 <- check_top_3(respondent_value, row_data$imageattr19)
  imageatt20 <- check_top_3(respondent_value, row_data$imageattr20)
  imageatt21 <- check_top_3(respondent_value, row_data$imageattr21)
 imageatt22 <- check_top_3(respondent_value, row_data$imageattr22)
 imageatt23 <- check_top_3(respondent_value, row_data$imageattr23)
 imageatt24 <- check_top_3(respondent_value, row_data$imageattr24)

 if (imageatt0 || imageatt1 || imageatt2 || imageatt4 || imageatt5 || imageatt6 || imageatt7 || imageatt8 || imageatt9 ||
imageatt10 || imageatt11 || imageatt15 || imageatt17 || imageatt19 || imageatt20 || imageatt21 || imageatt22 || imageatt23 ||
imageatt24 )
 { print("Respondent value is in Brand Imagery top 3 box") }
else
{ print("Respondent value is NOT in Brand Services  top 3 box")}
}
```

Question 4

For this problem, we need to compare difference between male and female So we can either use ANOVA test or T test. I have used T test to compare significance difference.
Case 1: Find significant difference between means in both male and female have selected values top 3 box consideration box.

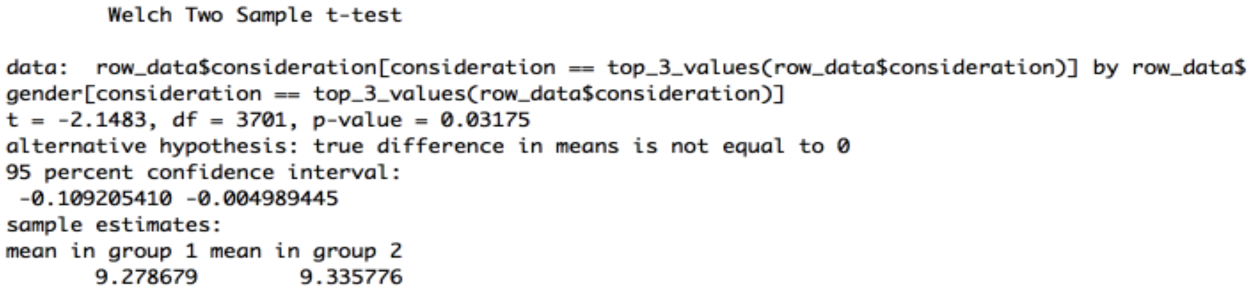
```
# get top 3 values
top_3_values <- function(data_set){
   return(head(unique(sort(data_set, decreasing = TRUE, index.return = FALSE)),3))
}
# Assumption 1: samples are independent
# Assumption 2: Given sample data equal variance
# H0 : variance for both male and female have selected top 3 box consideration box value is same.
# H1 : variance for both male and female have selected top 3 box consideration box value is NOT same.

var.test(row_data$consideration[consideration == top_3_values(row_data$
consideration)]~row_data$gender[consideration == top_3_values(row_data$ consideration)])
```



```
        F test to compare two variances

data:  row_data$consideration[consideration == top_3_values(row_data$consideration)] by row_data$
gender[consideration == top_3_values(row_data$consideration)]
F = 1.0009, num df = 1725, denom df = 2182, p-value = 0.9826
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9155973 1.0946990
sample estimates:
ratio of variances
         1.000914
```

Explanation, As, P value is greater than 0.05 so we fail to reject null hypothesis, in other word variance for both male and female have selected top 3 box consideration box value are equally accepted.

```
# Independent sample test
#H0 :  There is no significant difference between means in both male and female have selected values top 3 box
consideration box.
#H1 : There is significant difference between means in both male and female have selected values top 3 box
consideration box.
t.test(row_data$consideration[consideration == top_3_values(row_data$
consideration)]~row_data$gender[consideration == top_3_values(row_data$ consideration)])
```

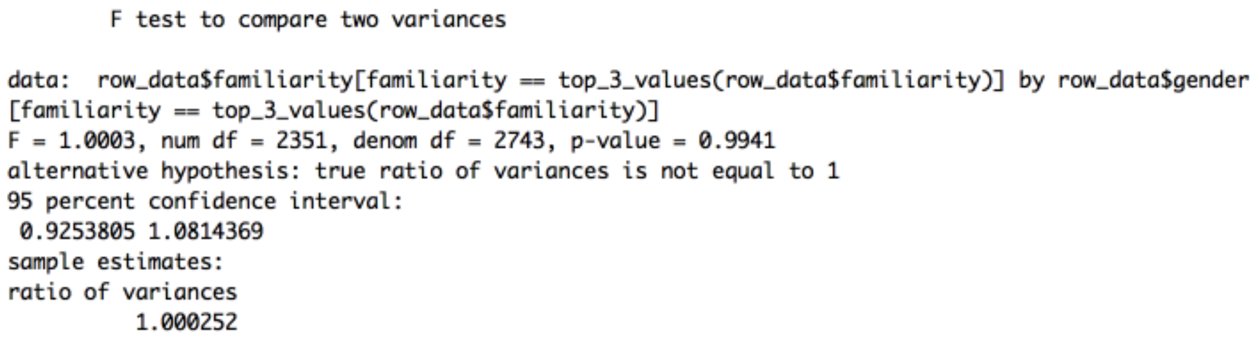


```
        Welch Two Sample t-test

data:  row_data$consideration[consideration == top_3_values(row_data$consideration)] by row_data$
gender[consideration == top_3_values(row_data$consideration)]
t = -2.1483, df = 3701, p-value = 0.03175
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.109205410 -0.004989445
sample estimates:
mean in group 1 mean in group 2
       9.278679        9.335776
```

p-value is less than 0.05, we reject the null hypothesis that there's no difference between the means in both male and female have selected values top 3 box consideration box and conclude that There is significant difference between means in both male and female have selected values top 3 box consideration box.

Case 2: significant difference between means in both male and female have selected values top 3 box familiarity box.

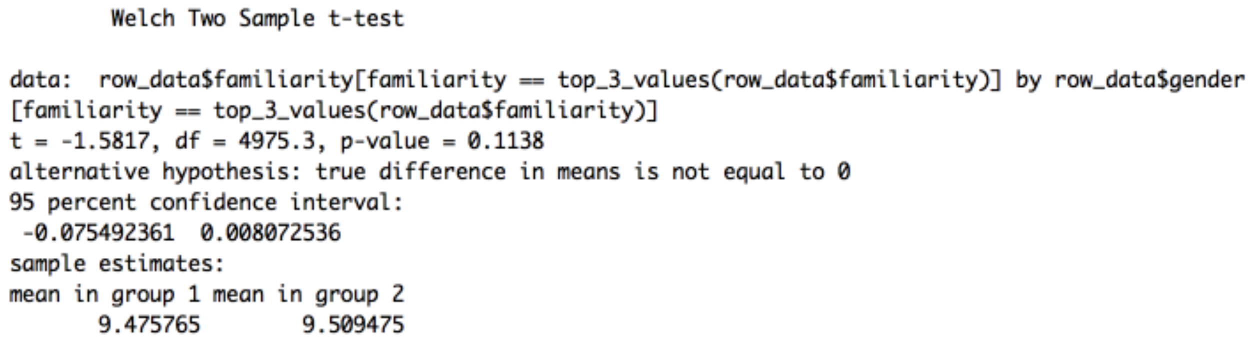
```
# Assumption 1: Given sample data equal variance
# H0 : variance for both male and female have selected top 3 box familiarity box value is same.
# H1 : variance for both male and female have selected top 3 box familiarity box value is NOT same.

var.test(row_data$familiarity[familiarity == top_3_values(row_data$familiarity)]~row_data$gender[familiarity ==
top_3_values(row_data$familiarity)])
```



```
        F test to compare two variances

data:  row_data$familiarity[familiarity == top_3_values(row_data$familiarity)] by row_data$gender
[familiarity == top_3_values(row_data$familiarity)]
F = 1.0003, num df = 2351, denom df = 2743, p-value = 0.9941
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9253805 1.0814369
sample estimates:
ratio of variances
          1.000252
```

As, P value is greater than 0.05 so we fail to reject null hypothesis, in other word variance for both male and female have selected top 3 box familiarity box values are equally accepted.

```
# Independent sample test
# H0 :  There is no significant difference between means in both male and female have selected values top 3 box
familiarity box.
# H1: There is significant difference between means in both male and female have selected values top 3 box familiarity
box.
t.test(row_data$familiarity[familiarity == top_3_values(row_data$familiarity)]~row_data$gender[familiarity ==
top_3_values(row_data$familiarity)])
```

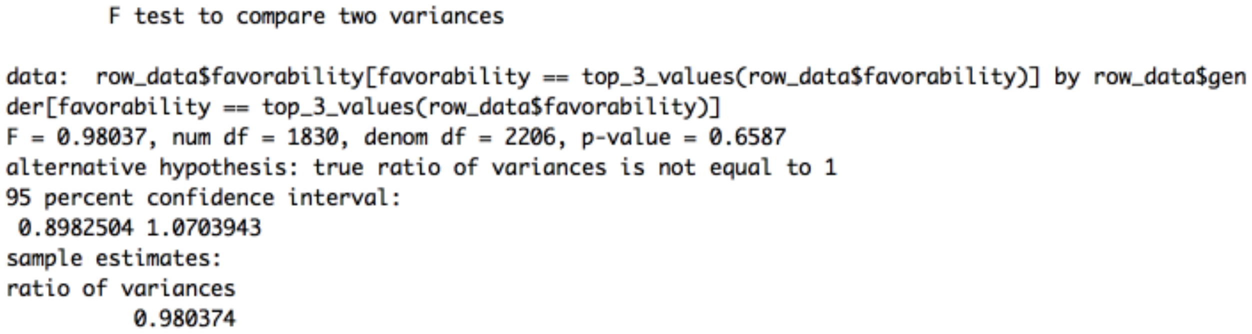


```
        Welch Two Sample t-test

data:  row_data$familiarity[familiarity == top_3_values(row_data$familiarity)] by row_data$gender
[familiarity == top_3_values(row_data$familiarity)]
t = -1.5817, df = 4975.3, p-value = 0.1138
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.075492361  0.008072536
sample estimates:
mean in group 1 mean in group 2
       9.475765        9.509475
```

p-value is greater than 0.05, we keep the null hypothesis that there's no significant difference between the means in both male and female have selected values top 3 box familiarity box.

Case 3: significant difference between means in both male and female have selected values top 3 box favorability box.

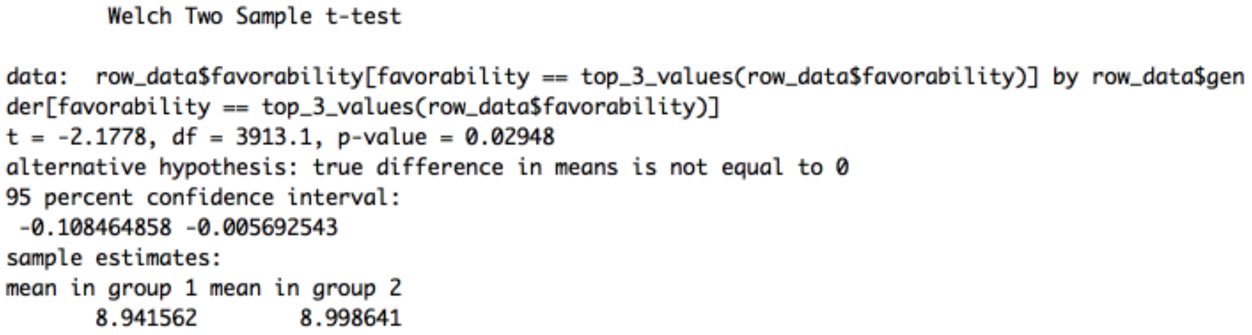
```
# Assumption 1: Given sample data equal variance
#H0 :  There is no significant difference between means in both male and female have selected values top 3 box favorability box.
#H1 : There is significant difference between means in both male and female have selected values top 3 box favorability box.

var.test(row_data$favorability[favorability == top_3_values(row_data$favorability)]~row_data$gender[favorability == top_3_values(row_data$favorability)])
```



```
        F test to compare two variances

data:  row_data$favorability[favorability == top_3_values(row_data$favorability)] by row_data$gen
der[favorability == top_3_values(row_data$favorability)]
F = 0.98037, num df = 1830, denom df = 2206, p-value = 0.6587
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8982504 1.0703943
sample estimates:
ratio of variances
          0.980374
```

As, P value is greater than 0.05 so we fail to reject null hypothesis, in other word variance for both male and female have selected top 3 box familiarity box values are equally accepted.

```
# Independent sample test
#H0 : Both male and female have selected values top 3 box favorability box are same.
#H1 : Both male and female have selected values top 3 box favorability box value are NOT same.

#Welch Two Sample t-test
t.test(row_data$favorability[favorability == top_3_values(row_data$favorability)]~row_data$gender[favorability == top_3_values(row_data$favorability)])
```



```
        Welch Two Sample t-test

data:  row_data$favorability[favorability == top_3_values(row_data$favorability)] by row_data$gen
der[favorability == top_3_values(row_data$favorability)]
t = -2.1778, df = 3913.1, p-value = 0.02948
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.108464858 -0.005692543
sample estimates:
mean in group 1 mean in group 2
       8.941562        8.998641
```

p-value is less than 0.05, we reject the null hypothesis that there's no difference between the means in both male and female have selected values top 3 box favorability box and conclude that There is significant difference between means in both male and female have selected values top 3 box favorability box.

Question 5

The aim of this study is to find significant relationship between Brand Consideration in a young audience with brand attitudes by proving or disproving the marketing team's hypothesis that improving brand image and attitudes among the young audience will have better impact on Consideration.

Assumption 1: The young audience consider age group between 18 to 24 ages.
Assumption 2: The top three boxes are calculated by arranging records in descending order and get top three values.
Assumption 3: Consideration is dependent variable, and both familiarity and favorability are independent variable in above given problem.

Find correlation between the dependent variable independent variable.

```
cor(data_filter$consideration,data_filter$favorability+data_filter$familiarity)
```

Output:

```
 [1] 1
```

Correlation can take values between -1 to +1. The value is 1 suggest a strong relationship consideration with the favorability and familiarity in young audience.

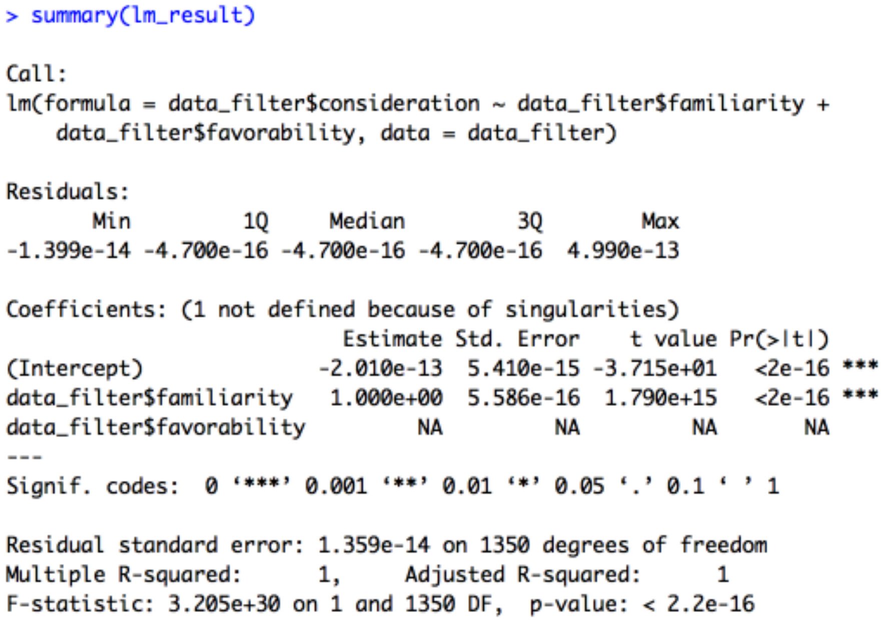
Now, find significant relationship between the consideration with familiarity and favorability in young audience. So, the following hypothesis need to prove using liner regression analysis using linear regression model.

H0: There is no relationship between overall Brand Consideration in a young audience with Improving brand attitudes
H1: There is no relationship between overall Brand Consideration in a young audience with Improving brand attitudes

Step 1: Subset data as per required column familiarity, favorability, consideration, indicator_age
Step 2: Filter data with top three box and age group between Age 18-24
Step 3: Perform liner regression model on filtered data.
Step 4: Analyze result

```
data_filter <- subset(row_data, select = c(familiarity,favorability,consideration,indicator_age))
data_filter <- subset(data_filter,consideration == top_3_values(data_filter$consideration) & familiarity ==
top_3_values(data_filter$familiarity) & indicator_age == "Age 18-24" & favorability ==
top_3_values(data_filter$favorability))
lm_result <- lm(data_filter$consideration~data_filter$familiarity+data_filter$favorability,data=data_filter)
summary(lm_result)
```

Output:



```
> summary(lm_result)

Call:
lm(formula = data_filter$consideration ~ data_filter$familiarity +
    data_filter$favorability, data = data_filter)

Residuals:
       Min         1Q     Median         3Q        Max
-1.399e-14 -4.700e-16 -4.700e-16 -4.700e-16  4.990e-13

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error    t value Pr(>|t|)
(Intercept)             -2.010e-13  5.410e-15 -3.715e+01   <2e-16 ***
data_filter$familiarity  1.000e+00  5.586e-16  1.790e+15   <2e-16 ***
data_filter$favorability        NA         NA         NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.359e-14 on 1350 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.205e+30 on 1 and 1350 DF,  p-value: < 2.2e-16
```

As NA as coefficient indicates that the favorability linearly related to the familiarity. As the p-value is much less than 0.05 for both familiarity and favorability. Secondly t value for the both variable is far away from the zero and closed to Std. Error. So, we reject the null hypothesis that β = 0. Hence there is a significant relationship between the familiarity and favorability with consideration in young audience in the linear regression model of the data set faithful. In other word, **we can say that marketing team's hypothesis that improving brand image and attitudes among the young audience will have better impact on consideration is right and proven**. The R2 we get is 1. Or roughly 100% of the variance found in the response variable (consideration) can be explained by the predictor variable (familiarity and favorability) in young age group.

Question 6

1. Baseline probability of likelihood to purchase is 0%.
Answer:
As no value coefficient is zero, baseline probability of likelihood to purchase is not 0%.

2. Younger age groups are more likely to purchase than older age groups.
Answer:
As t value coefficient is negative, there is decrease in likelihood to purchase with increase in age. Hence conclusion is true.

3. Probability of Older age group purchasing is exp(-0.33) = 72%.
Answer:
Probability of Older age group in logistic regression is 1/1+e^-e
False, As we are using above formula to calculate probability in logistic regression.
1/(1+exp(-0.33)) = 0.5817594
So actually probability of Older age group is 58.18%

4. Lower income groups are less likely to purchase across the board.
Answer:
Negative t value interprets the decrease trend in likelihood to purchase corresponding to lower income groups.

5. It is better to improve Attribute B than Attribute A.
Answer:
True. As T value depends on slope of regression line. More the t value, more is the linear trend of increase. Attribute B has more t value, so it's better to improve attribute B.

6. We should move marketing dollars away from males and low-income groups.
Answer:

7. Sports TV is effective for females only.
Answer: