Analysis of Fruit Preference Data

siju.swamy@saintgits.org

2023-10-13

Abstract: This article is a short analysis of the data collected from the course participants of the `Fundamentals of Data Analytics using R Programming`. A baseline descriptive analysis is conducted and the results are tested using hypothesis testing for generalizations.

Key words: Baseline analysis, categorical data, Likert scaled items, correlation testing, regression models, box plot, bar plot, percentage analysis, χ^2-test, ANOVA.

```
# loading the .csv file containing data

df=read.csv("https://raw.githubusercontent.com/sijuswamy/Data-Analytics-using-R/main/Fruit_data.csv",hea
df=df[,-1] # select all rows but removes first column
size=nrow(df)
#df
```

Introduction

This article is prepared during the add-on course - Data analytics using R programming `Statistical Foundations for Food Engineering`. A questionnaire is prepared and administered through *Google form*. Total number of samples collected in the study is 255.

Data cleaning and Wrangling

As the first stage of the data pre-processing, the row data is cleaned and wrangled to make it ready for statistical analysis. Main processes involved are:

1. Removing unnecessary columns
2. Rename the attributes for make then clear and precise
3. Mapping of data types for statistical analysis
4. Rhttps://raw.githubusercontent.com/sijuswamy/Data-Analytics-using-R/main/Fruit_data.csv Day-1: Introduction to R programminge-frame the structure of the data if required

Rename the attribute names

Since the column titles obtained through the *Google forms* are the questions given in the questionnaire, it will be not suitable to represent an attribute. There are two ways to correct it- manually correct in the downloaded excel file or rename the column names programatically.

In this article, the later approach is demonstrated.

```
colnames(df) <- c("Gender","Age","Weight","Orange","Grapes","Banana","Apple","Mango","Cherry","Height")
#df
```

Descriptive Analysis

A cleaned data is examined with basic statistical tools to understand the distribution of various attributes and its relationship between control variables. At this initial stage, fundamental tools like frequency tables, cross tabulations and percentage analysis followed by proper visualizations will be used. All the socio-demographic variables and control variables will be analysed statistically.

Geder

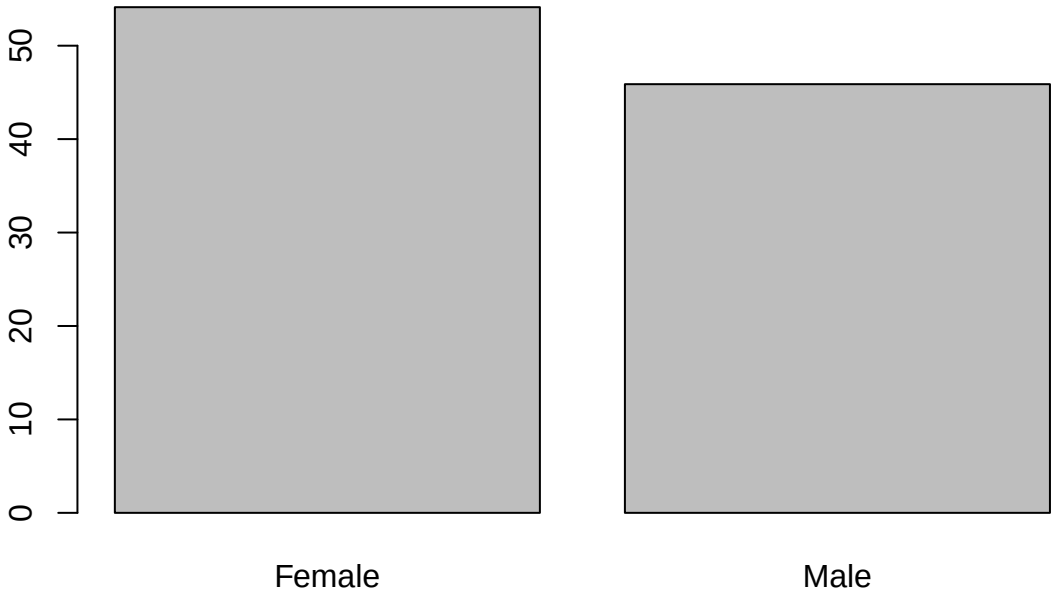
The gender-wise distribution of the responses is shown in the following Table.

```
Gtab=round(prop.table(table(df$Gender))*100,2)
#Gtab
```

The percentage analysis shows that majority of the respondents are from Female category.

A barplot showing this distribution is shown in the following Figure.

```
barplot(Gtab)
```

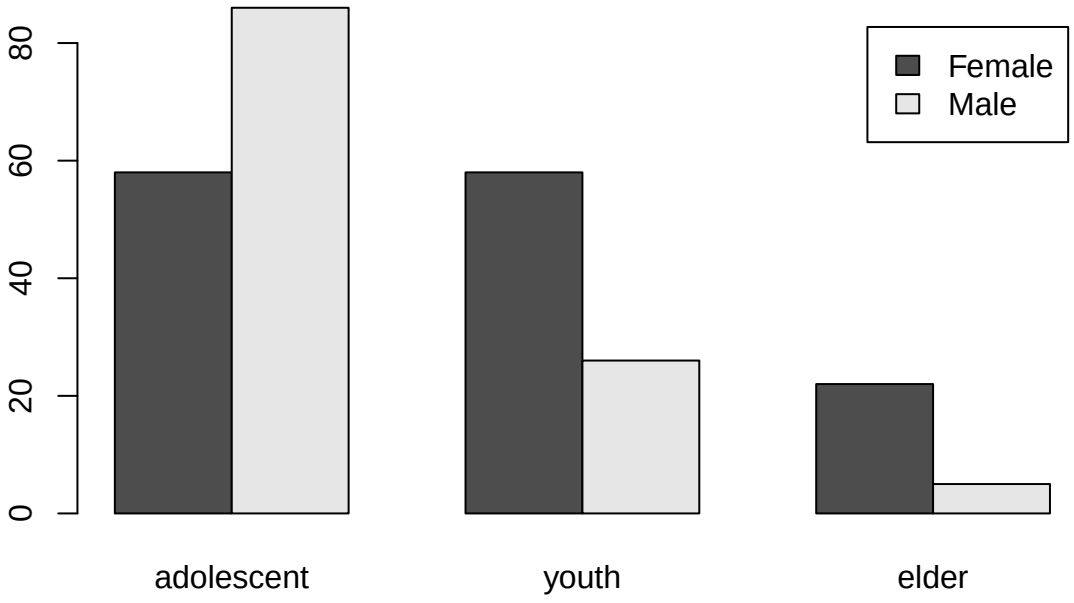


Age

Completed age of the respondents are collected through the form. Interest of respondents may be varied over age range rather than individual ages. So the continuous variable `age` shall be converted to a categorical variable for reasonable use of this attribute. Respondents with age upto 20 is consider as 'adolescent', between 20 and 30 as 'youth' and above 30 is considered as the 'elder'. A new variable `Age_group` is created as follows.

```
df$Age_group <- cut(df$Age,
                    breaks=c(-Inf, 20, 30, Inf),
                    labels=c("adolescent","youth","elder"))
#df
```

```
df$Gender=as.factor(df$Gender)
barplot(table(df$Gender,df$Age_group),beside = TRUE,legend.text = levels(df$Gender))
```



A contegency table of the gender-wise distribution over age-group is shown in the Table below:

```
knitr::kable(table(df$Age_group,df$Gender))
```

	Female	Male
adolescent	58	86
youth	58	26
elder	22	5

Conversion of data into long format

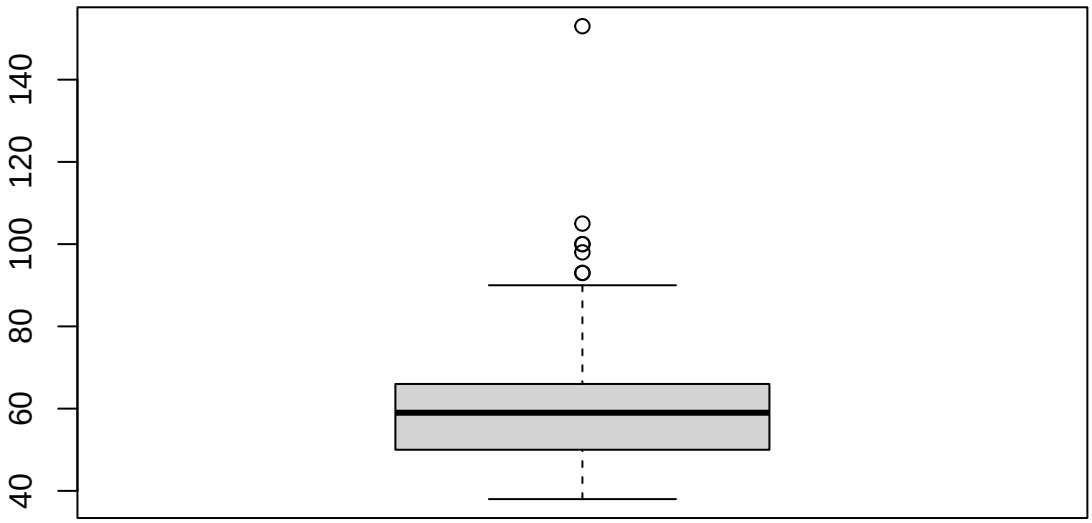
The basic objective of this work is to compare the respondent preference of fruit based on their interest shown in the survey. So responses regarding preference of various fruits under the study should br bring together before statistical investigations. To achieve this goal, the *wide formatted** data is transformed into the long format with a columns for fruits and the corresponding preference level.

```
library(reshape2)
library(plyr)
data_long <- melt(df,
        # ID variables - all the variables to keep but not split apart on
    id.vars=c("Gender","Age","Weight","Height","Age_group"),
        # The source columns
    measure.vars=c("Orange", "Grapes", "Banana","Apple","Mango","Cherry" ),
        # Name of the destination column that will identify the original
        # column that the measurement came from
    variable.name="Fruit",
    value.name="Rating"
)
data_long <- arrange(data_long,data_long$Fruit, data_long$Rating)# to get a sorted view
#data_long
```

Weight

The distribution of respondent's weights is shown in the `Boxplot` shown below:

```
boxplot(df$Weight,notch = F)
```



The five point summary of the attribite 'weight' is shown in the Table below:

```
summary(df$Weight)
```

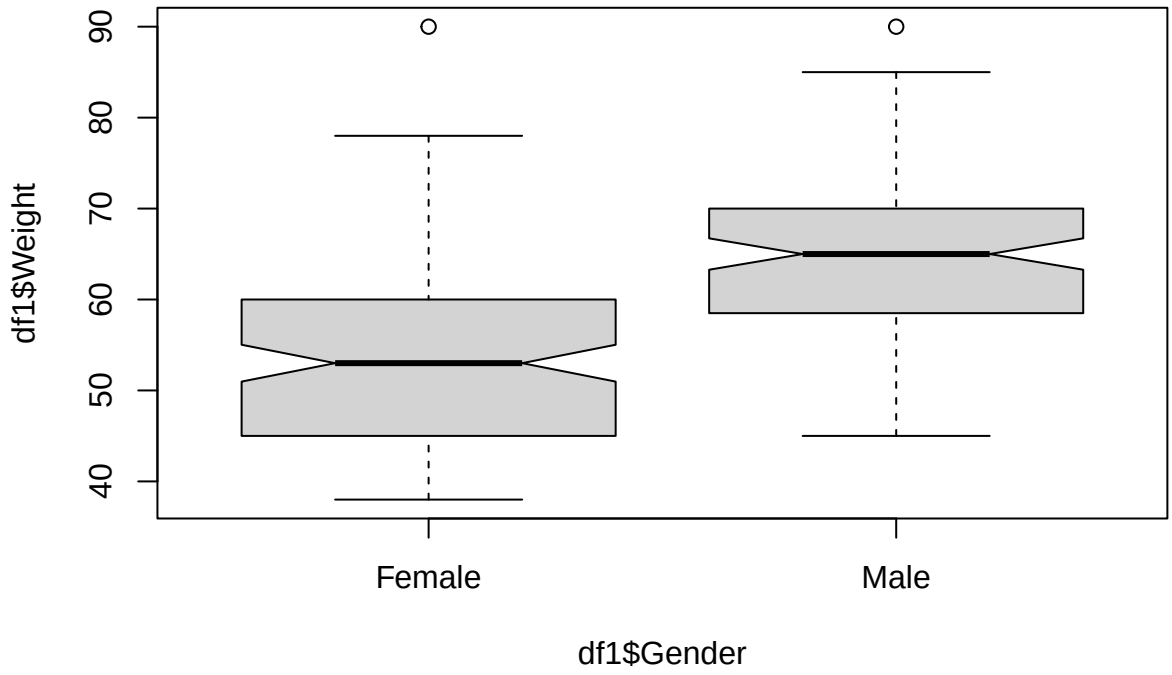
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   38.00   50.00   59.00   59.86   66.00  153.00
```

It is noted that the box plot shows the presence of outliers in the weights. These samples can be removed as follows:

```
outliers <- boxplot(df$Weight, plot=FALSE)$out
df1<-df;
df1<- df1[-which(df1$Weight %in% outliers),]
#dim(df1)
```

A gender-wise comparison of weights is shown in the following figure. It is clear from the plot that, the majority of the female category is above the average body weight category!

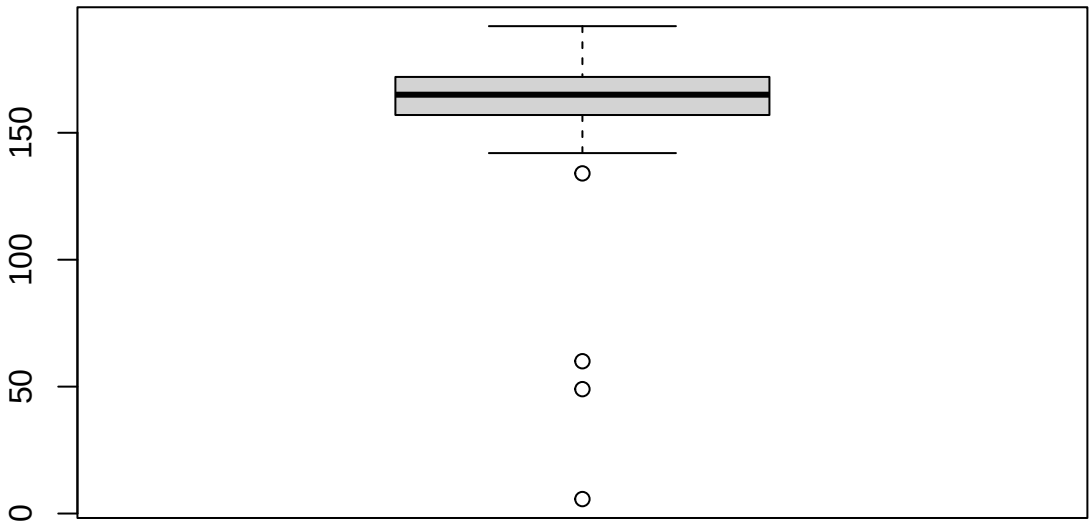
```
boxplot(df1$Weight~df1$Gender,notch = T)
```



Height

The distribution of respondent's height is shown in the Boxplot shown below:

```
boxplot(df$Height,notch = F)
```



The five point summary of the attribute 'Height' is shown in the Table below:

```
summary(df$Height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.7   157.0   165.0   163.5   172.0   192.0
```

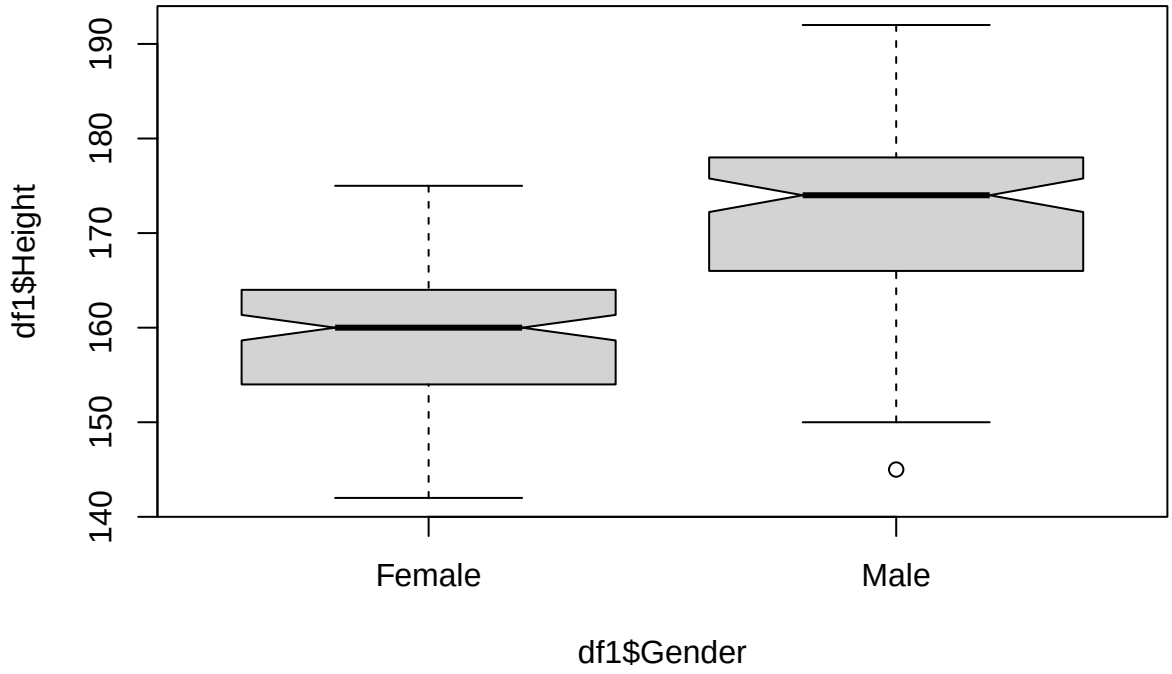
It is noted that the box plot shows the presence of outliers in the weights. These samples can be removed as follows:

```
outliers <- boxplot(df$Height, plot=FALSE)$out
df1<-df;
```

```r
df1<- df1[-which(df1$Height %in% outliers),]
#dim(df1)
```

A gender-wise comparison of heights is shown in the following figure. It is clear from the plot that, the majority of the female category is below the average height category!

```r
boxplot(df1$Height~df1$Gender,notch = T)
```



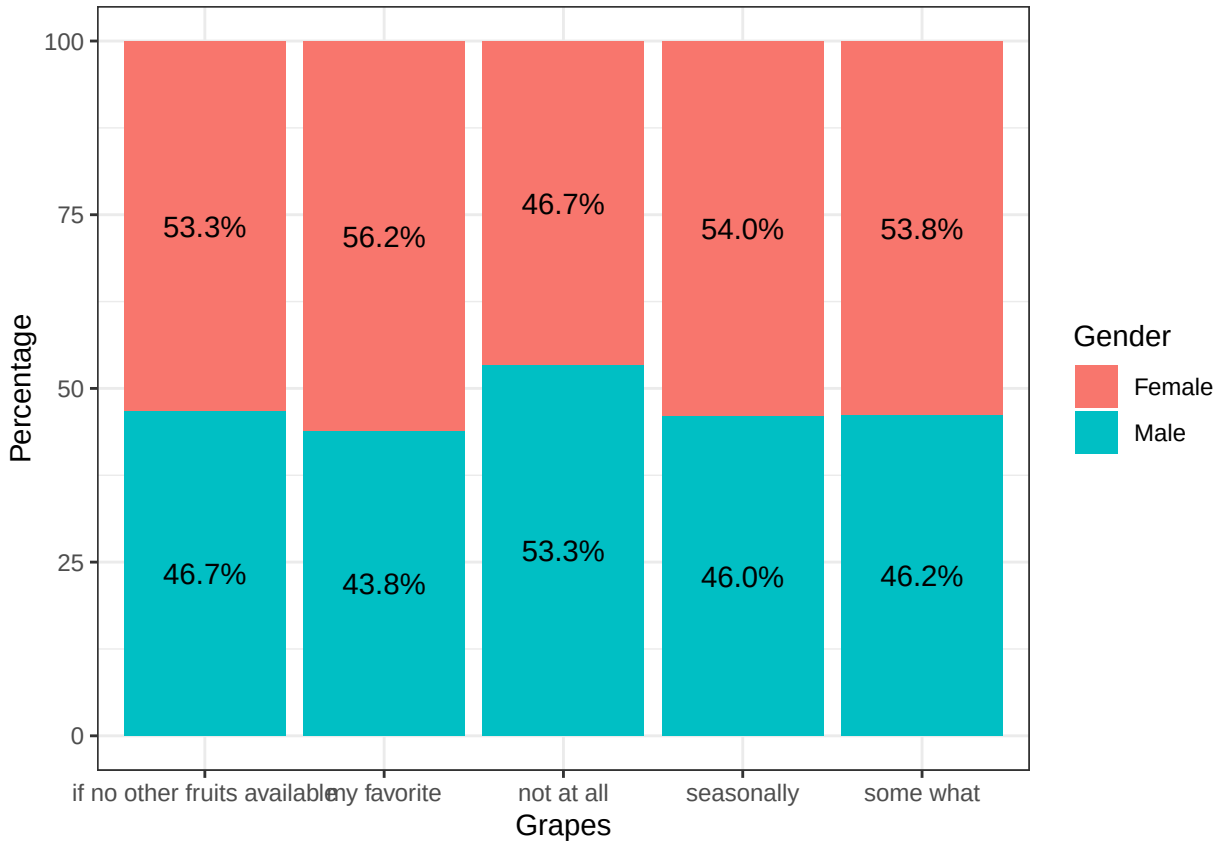
Gender-wise preference of fruitshttps://raw.githubusercontent.com/sijuswamy/Data-Analytics-using-R/main/Fruit_data.csv

Day-1: Introduction to R programming

A percentage analysis of gender-wise preference of various fruits is given in theis section.

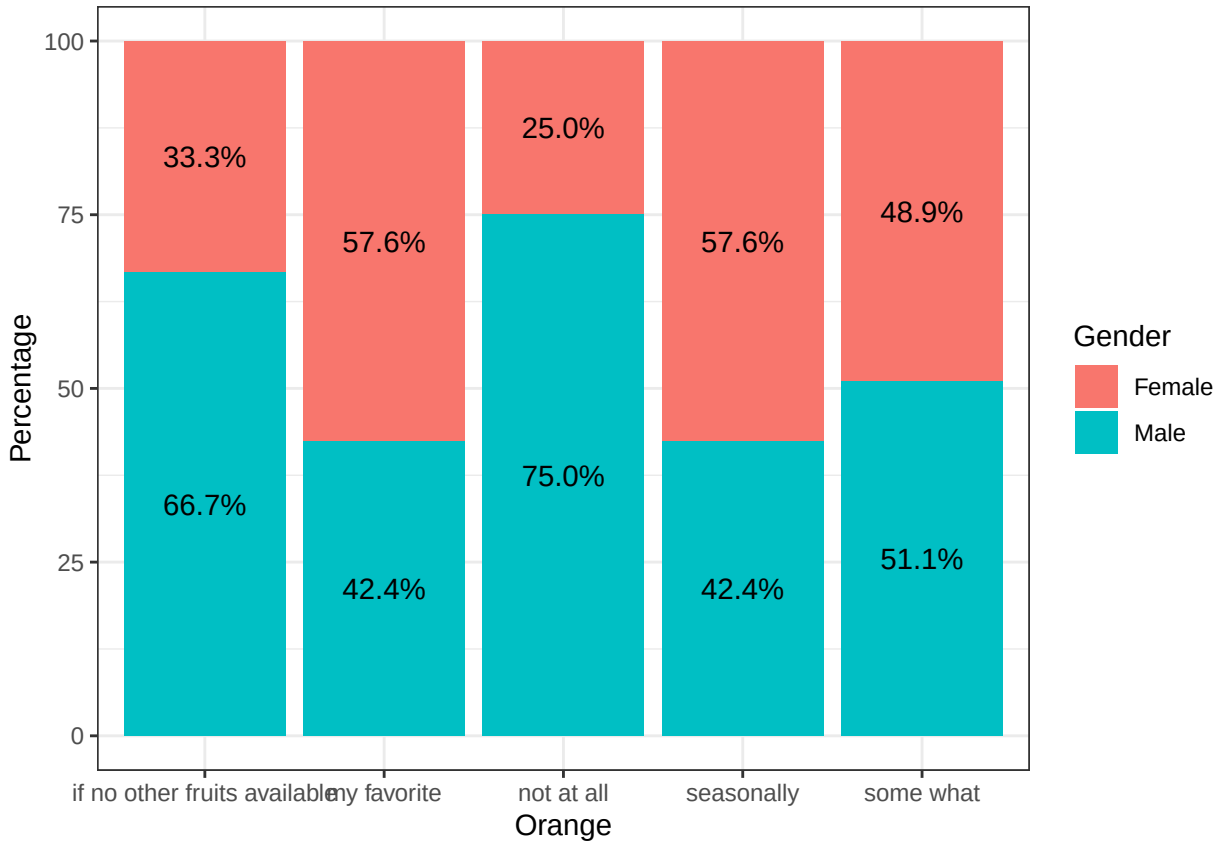
```r
#gender-wise
library(dplyr)
library(ggplot2)

df %>%
  count(Grapes,Gender) %>%
  group_by(Grapes) %>%
  mutate(pct= prop.table(n) * 100) %>%
  ggplot() + aes(Grapes, pct, fill=Gender) +
  geom_bar(stat="identity") +
  ylab("Number of respondents") +
  geom_text(aes(label=paste0(sprintf("%1.1f", pct),"%")),
            position=position_stack(vjust=0.5)) +labs(x ="Grapes", y = "Percentage",fill="Gender")+
  theme_bw()
```



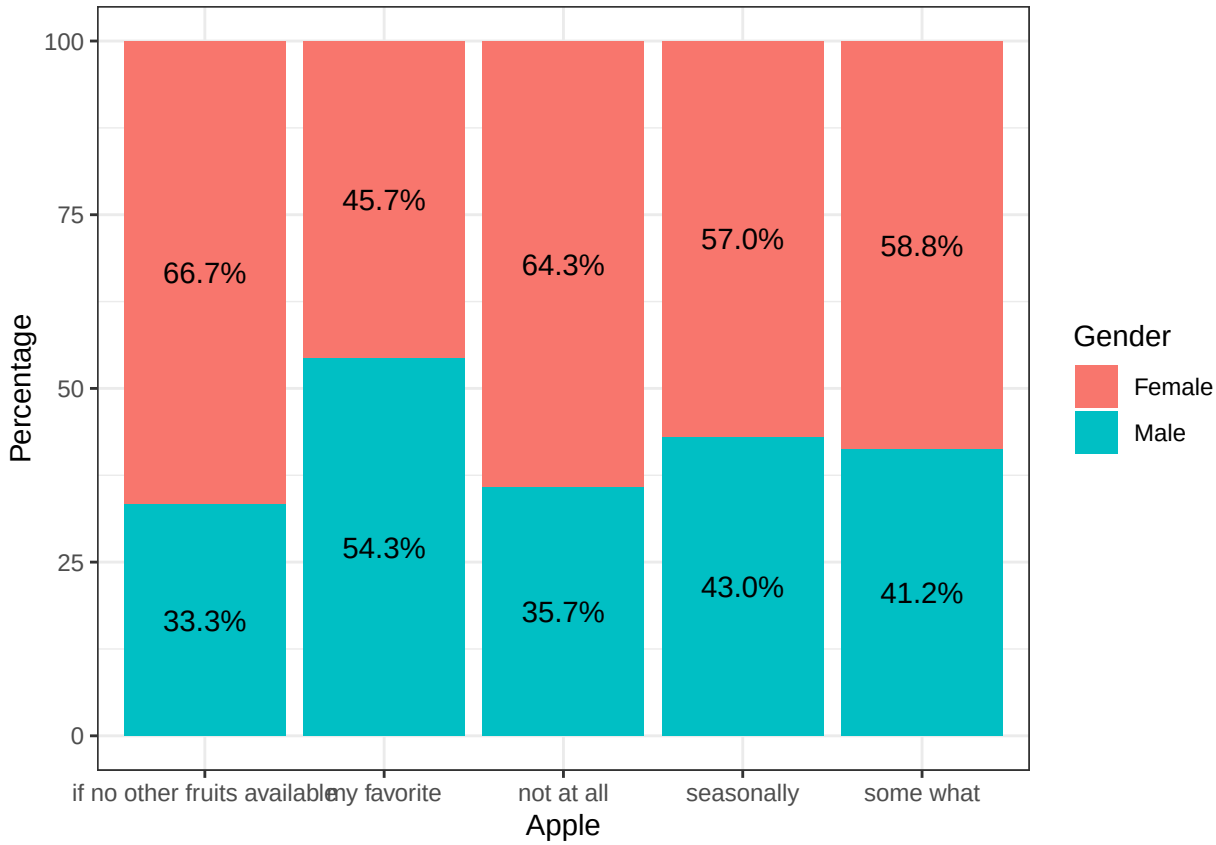

```
#gender-wise
library(dplyr)
library(ggplot2)

df %>%
  count(Orange,Gender) %>%
  group_by(Orange) %>%
  mutate(pct= prop.table(n) * 100) %>%
  ggplot() + aes(Orange, pct, fill=Gender) +
  geom_bar(stat="identity") +
  ylab("Number of respondents") +
  geom_text(aes(label=paste0(sprintf("%1.1f", pct),"%")),
            position=position_stack(vjust=0.5)) +labs(x ="Orange", y = "Percentage",fill="Gender")+
  theme_bw()
```



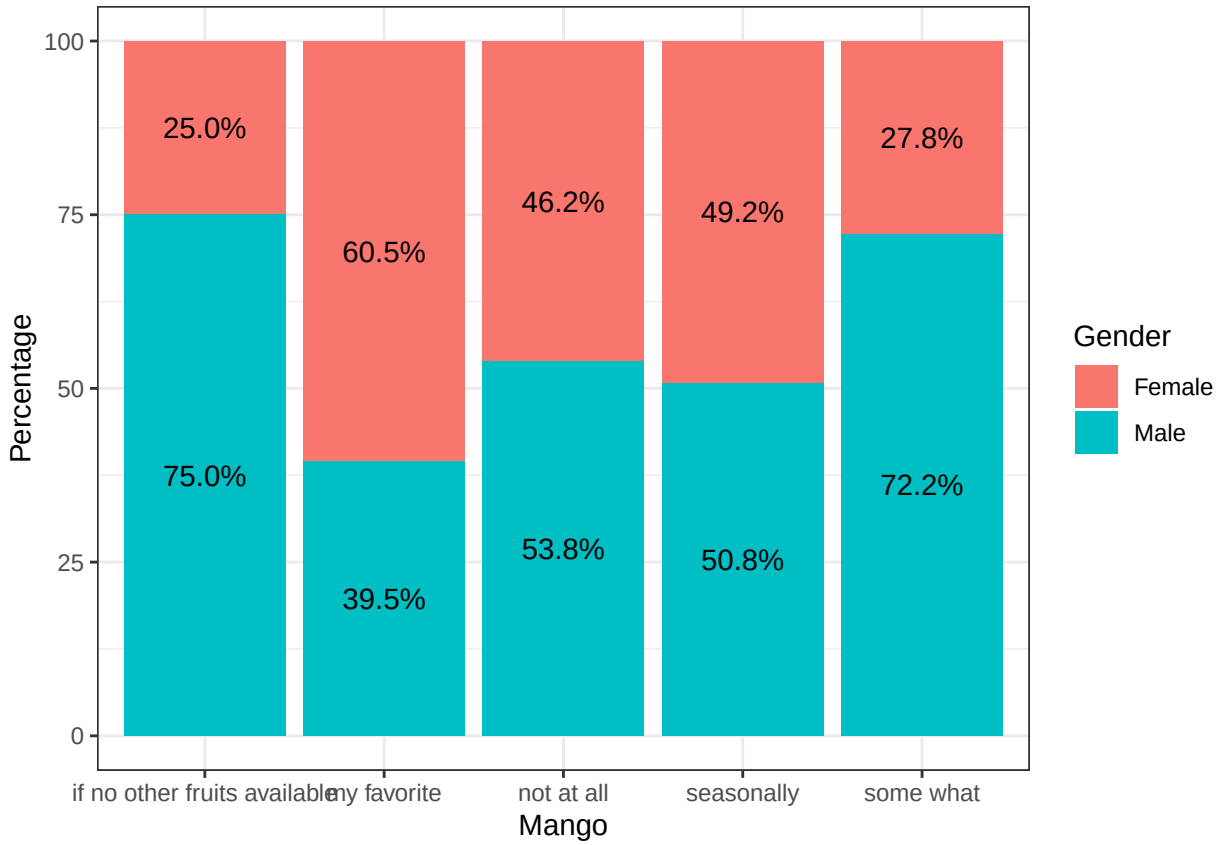

```
#gender-wise
library(dplyr)
library(ggplot2)

df %>%
  count(Apple,Gender) %>%
  group_by(Apple) %>%
  mutate(pct= prop.table(n) * 100) %>%
  ggplot() + aes(Apple, pct, fill=Gender) +
  geom_bar(stat="identity") +
  ylab("Number of respondents") +
  geom_text(aes(label=paste0(sprintf("%1.1f", pct),"%")),
            position=position_stack(vjust=0.5)) +labs(x ="Apple", y = "Percentage",fill="Gender")+
  theme_bw()
```



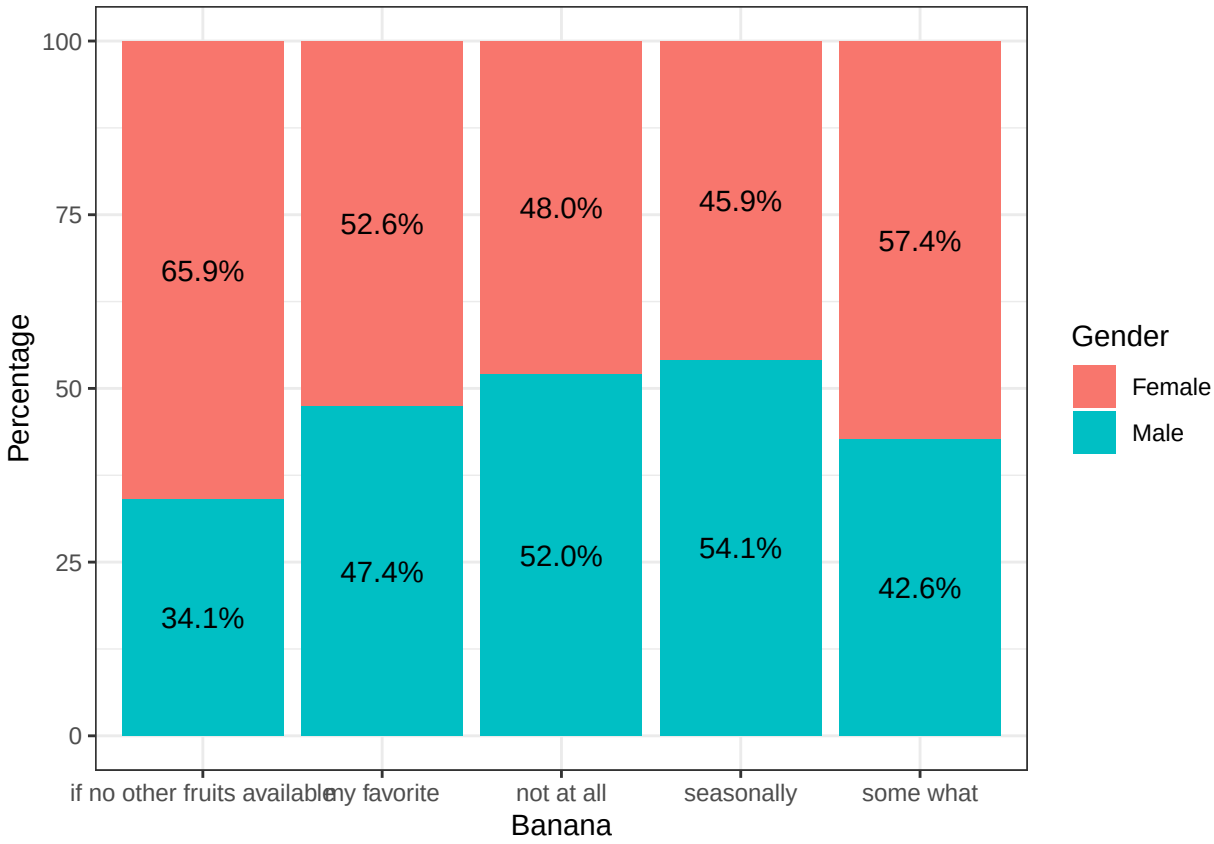

```
#gender-wise
library(dplyr)
library(ggplot2)

df %>%
  count(Mango,Gender) %>%
  group_by(Mango) %>%
  mutate(pct= prop.table(n) * 100) %>%
  ggplot() + aes(Mango, pct, fill=Gender) +
  geom_bar(stat="identity") +
  ylab("Number of respondents") +
  geom_text(aes(label=paste0(sprintf("%1.1f", pct),"%")),
            position=position_stack(vjust=0.5)) +labs(x ="Mango", y = "Percentage",fill="Gender")+
  theme_bw()
```




```
#gender-wise
library(dplyr)
library(ggplot2)

df %>%
  count(Banana,Gender) %>%
  group_by(Banana) %>%
  mutate(pct= prop.table(n) * 100) %>%
  ggplot() + aes(Banana, pct, fill=Gender) +
  geom_bar(stat="identity") +
  ylab("Number of respondents") +
  geom_text(aes(label=paste0(sprintf("%1.1f", pct),"%")),
            position=position_stack(vjust=0.5)) +labs(x ="Banana", y = "Percentage",fill="Gender")+
  theme_bw()
```

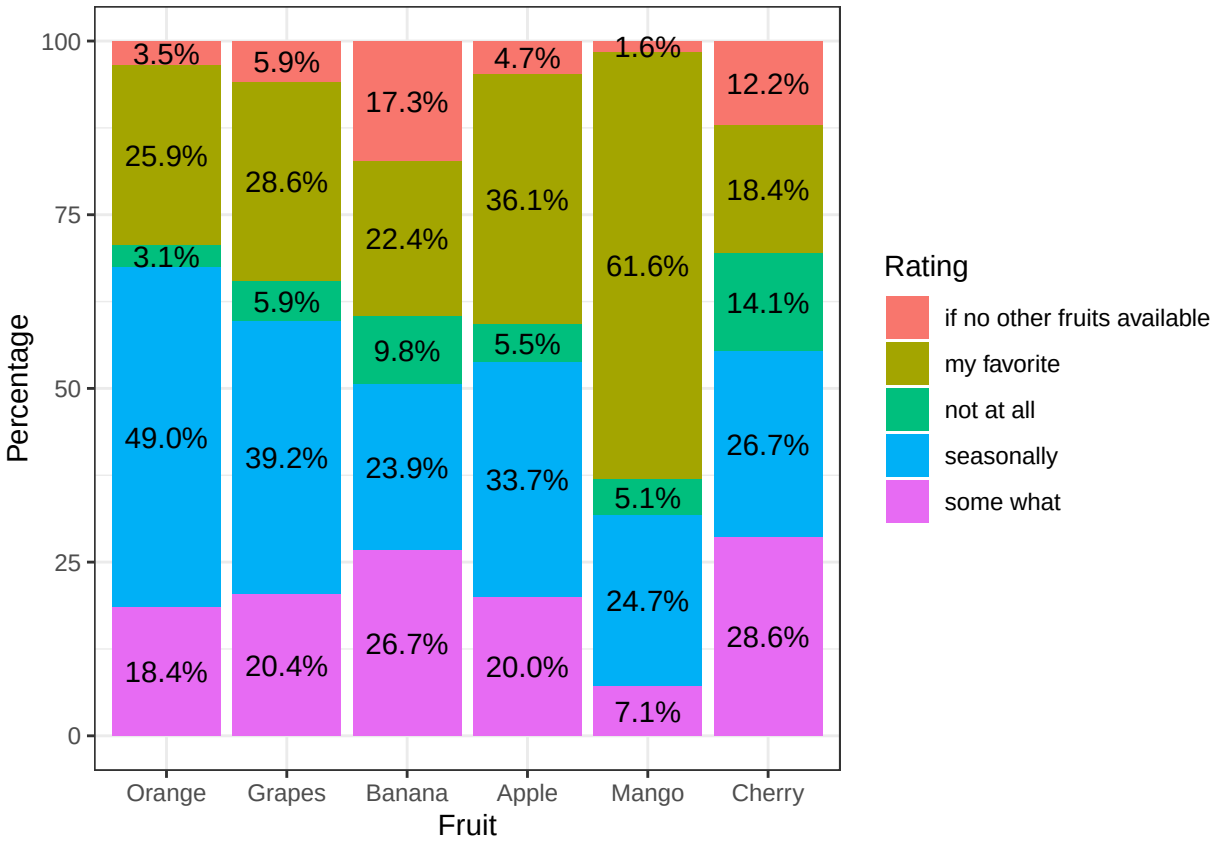


Overall preference over fruits

A percentage analysis of overall preference over various fruits is shown in the following figure.

```
#gender-wise
library(dplyr)
library(ggplot2)

data_long%>%
  count(Fruit,Rating) %>%
  group_by(Fruit) %>%
  mutate(pct= prop.table(n) * 100) %>%
  ggplot() + aes(Fruit, pct, fill=Rating) +
  geom_bar(stat="identity") +
  ylab("Number of respondents") +
  geom_text(aes(label=paste0(sprintf("%1.1f", pct),"%")),
            position=position_stack(vjust=0.5)) +labs(x ="Fruit", y = "Percentage",fill="Rating")+
  theme_bw()
```



Finding descriptive statistics of continuous variables

Weight over Gender

```
library(dplyr)
Fa=group_by(df, Gender) %>%
  summarise(
    count = n(),
    mean =round( mean(Weight, na.rm = TRUE),2),
    sd = round(sd(Weight, na.rm = TRUE),2)
  )
knitr::kable(Fa)
```

Gender	count	mean	sd
Female	138	54.80	12.96
Male	117	65.83	11.88

Height over Gender

```
library(dplyr)
Fa=group_by(df, Gender) %>%
  summarise(
    count = n(),
    mean =round( mean(Height, na.rm = TRUE),2),
    sd = round(sd(Height, na.rm = TRUE),2)
  )
```

```
knitr::kable(Fa)
```

Gender	count	mean	sd
Female	138	158.42	11.34
Male	117	169.40	20.58

Height over Age Group

```
library(dplyr)
Fa=group_by(df, Age_group) %>%
  summarise(
    count = n(),
    mean =round( mean(Height, na.rm = TRUE),2),
    sd = round(sd(Height, na.rm = TRUE),2)
  )
knitr::kable(Fa)
```

Age_group	count	mean	sd
adolescent	144	164.56	19.09
youth	84	162.55	15.42
elder	27	160.41	8.85

```
library(dplyr)
Fa=group_by(df, Gender) %>%
  summarise(
    count = n(),
    mean =round( mean(Age, na.rm = TRUE),2),
    sd = round(sd(Age, na.rm = TRUE),2)
  )
knitr::kable(Fa)
```

Gender	count	mean	sd
Female	138	23.41	6.40
Male	117	20.65	3.39

Rename the levels (For Numerical calculation)

```
df$Orange <- factor(df$Orange,
                    levels = c("not at all","if no other fruits available","some what","seasonally","my fa
                    labels = c("1","2","3","4","5"))
head(df)
```

```
##   Gender Age Weight Orange     Grapes                     Banana       Apple
## 1   Male  42     68      5  some what                  seasonally   some what
## 2 Female  39     63      4 seasonally                  seasonally  seasonally
## 3   Male  36     65      4  some what                 my favorite  seasonally
## 4   Male  21     60      4 seasonally                  seasonally  seasonally
## 5   Male  21    105      4 seasonally                  not at all  seasonally
## 6   Male  21     65      3 seasonally if no other fruits available my favorite
##         Mango                      Cherry Height Age_group
```

```
## 1  seasonally                   not at all    145     elder
## 2 my favorite                   seasonally    155     elder
## 3 my favorite if no other fruits available    162     elder
## 4  seasonally                   seasonally    174     youth
## 5 my favorite                   seasonally    172     youth
## 6  seasonally                   not at all    175     youth
```

```
mean(as.integer(df$Orange))
```

```
## [1] 3.909804
```

Inferential Analysis

The inferential analysis is the generalization part of statistics. This phase focuses on possibilities of generalization of observations in the descriptive analysis. This is achieved through the `hypothesis testing` aspects of inferential statistics.

Testing of significance of difference in mean weight over gender

Significance of difference in mean of continuous variable over two categories can be tested using the `t-test`. The null hypothesis of the t-test is;

H_0 : there is no significance difference in the mean

The alternative hypothesis is:

H_1 : there is significance difference in the mean

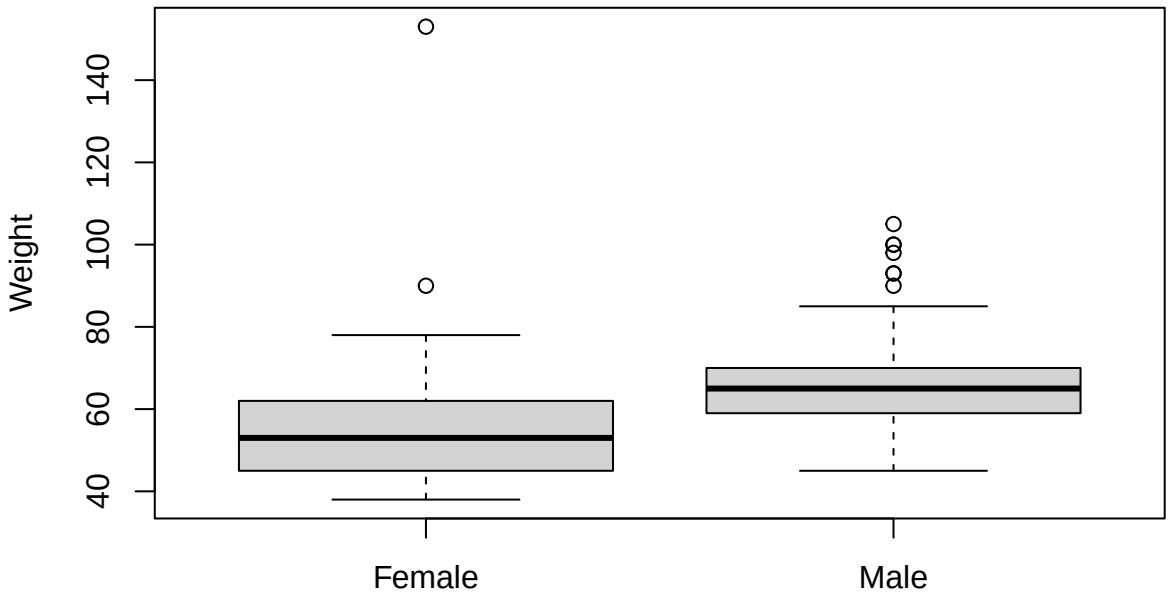
If the `p-value` of the test result is less than 0.05, the null hypothesis is rejected at 5% level of significance. Otherwise the null hypothesis can't be rejected.

As an example, let us investigate whether there is significant difference in the mean weight over gender.

```
t.test(Weight~Gender,alternative="less",data=df)
```

```
##
##  Welch Two Sample t-test
##
## data:  Weight by Gender
## t = -7.0845, df = 251.45, p-value = 6.99e-12
## alternative hypothesis: true difference in means between group Female and group Male is less than 0
## 95 percent confidence interval:
##       -Inf -8.459822
## sample estimates:
## mean in group Female   mean in group Male
##             54.79710             65.82735
```

```
plot(Weight~Gender,data=df)
```



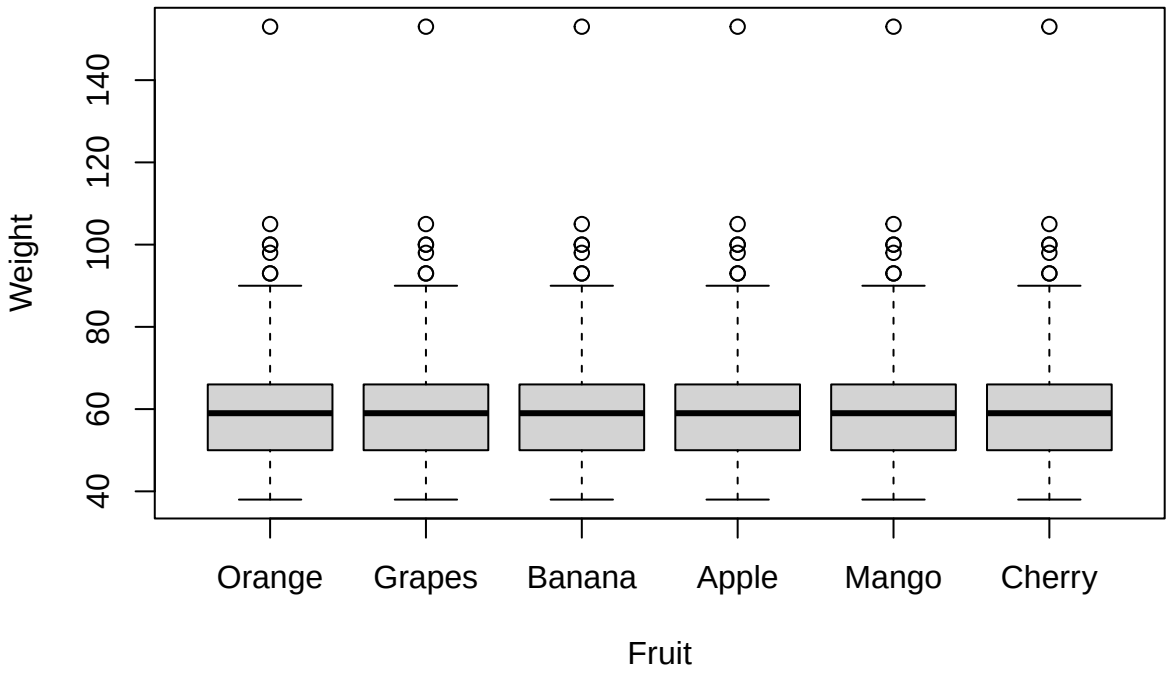
Since the p-value is less than 0.05, the null hypothesis is rejected. So it is statistically reasonable to conclude that the mean weight of female respondents is less than male respondents.

Significance of difference in mean weight over Fruit interest

```
# Compute the analysis of variance
res.aov <- aov(Weight ~ Fruit, data = data_long)
# Summary of the analysis
summary(res.aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Fruit          5      0     0.0       0      1
## Residuals   1524 282596   185.4
```

```
plot(Weight~Fruit,data=data_long)
```



Testing of significance difference in rating over Fruit category

As a first step a percentage analysis is conducted on the fruit preference data as shown below:

Percentage analysis of the rating of fruit category

```
prop.table(table(data_long$Fruit, data_long$Rating))*100
```

```
##
##          if no other fruits available my favorite not at all seasonally
##   Orange                    0.5882353   4.3137255  0.5228758  8.1699346
##   Grapes                    0.9803922   4.7712418  0.9803922  6.5359477
##   Banana                    2.8758170   3.7254902  1.6339869  3.9869281
##   Apple                     0.7843137   6.0130719  0.9150327  5.6209150
##   Mango                     0.2614379  10.2614379  0.8496732  4.1176471
##   Cherry                    2.0261438   3.0718954  2.3529412  4.4444444
##
##           some what
##   Orange  3.0718954
##   Grapes  3.3986928
##   Banana  4.4444444
##   Apple   3.3333333
##   Mango   1.1764706
##   Cherry  4.7712418
```

χ^2 test for confirmation of difference in rating over fruit category

```
# chi square
chisq.test(table(data_long$Fruit,data_long$Rating))
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  table(data_long$Fruit, data_long$Rating)
## X-squared = 259.37, df = 20, p-value < 2.2e-16
```

Since the p-value is less than 0.05, the null hypothesis that there is no significant difference in rating over fruit category is rejected. So it is statistically reasonable to conclude that the respondent's preference over fruits is statistically significant.

Fruit preference over Age group

Significance of difference in fruit rating over age group is tested using the chi-squared test.

```
# chi square
chisq.test(table(data_long$Fruit,data_long$Age_group))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(data_long$Fruit, data_long$Age_group)
## X-squared = 0, df = 10, p-value = 1
```

Since the p-value is greater than 0.05, the null hypothesis that there is no significant difference in the fruit rating over age group.

Conclusion

Based on the statistical analysis the following findings are elicited.

1. There is significant difference in the mean weight of the respondents
2. There is significant difference in the mean weight over fruit preference.
3. There is significant difference in the fruit ratings over fruit type.
4. There is no significant difference in the fruit preference over age group.