UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

FACULTY OF ENGINEERING DESIGN AND TECHNOLOGY

COURSE: BACHELOR OF SCIENCE IN COMPUTER SCIENCE (BSCS)

COURSE UNIT: DATA SCIENCE DSC 2103

LECTURER: DR. DAPHINE NYACHAKI BITALO (PhD GENETICS AND BIOINFORMATICS)

NAME: MUGANGA CHARLES

Access no: A96447

REG no: J22B23/032

Assignment 3 Write Up.

Assignment 1 markdown

libraries used

- tidyverse
- ggplot2
- readxl

Part a

Approach

- Load the dataset
- Summarize dataset
- calculate the mode of the overall performance
- Calculate the mode using the user function.
- print the mode

Mode

```
## [1] 7.5
```

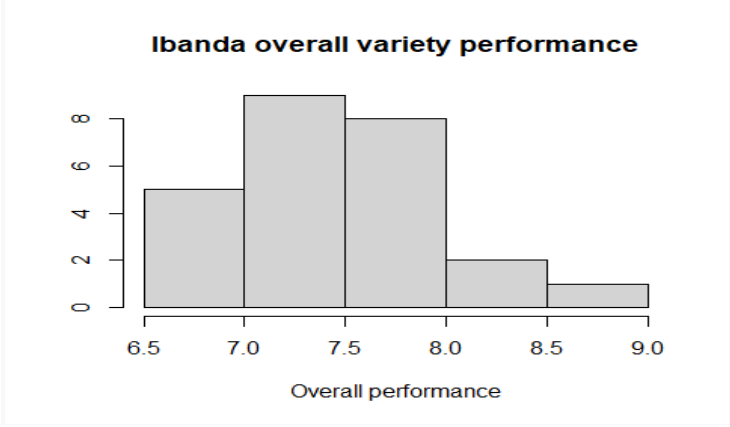
breaking mean and median

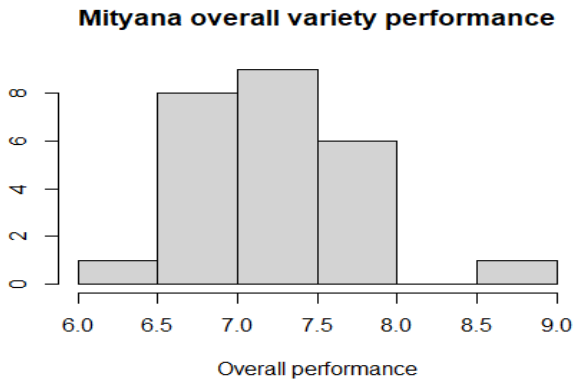
results

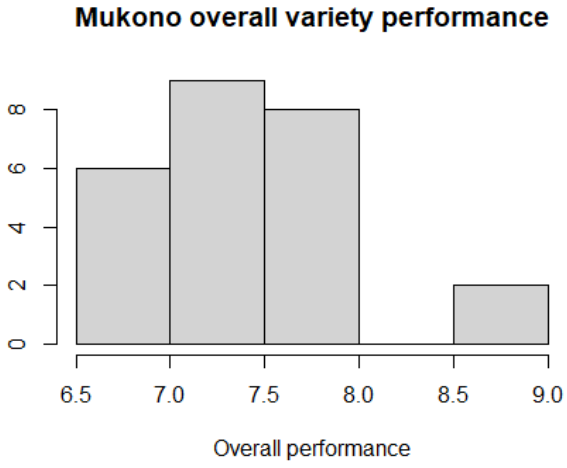
- Ibanda, Mityana and MUkono respectively.

```
summary(Ibanda$OVERALL)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.75    7.25    7.50    7.59    8.00    8.75

summary(Mityana$OVERALL)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.000   7.000   7.500   7.402   7.750   8.750

summary(Mukono$OVERALL)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.500   7.250   7.500   7.524   7.750   9.000
```

The graphs







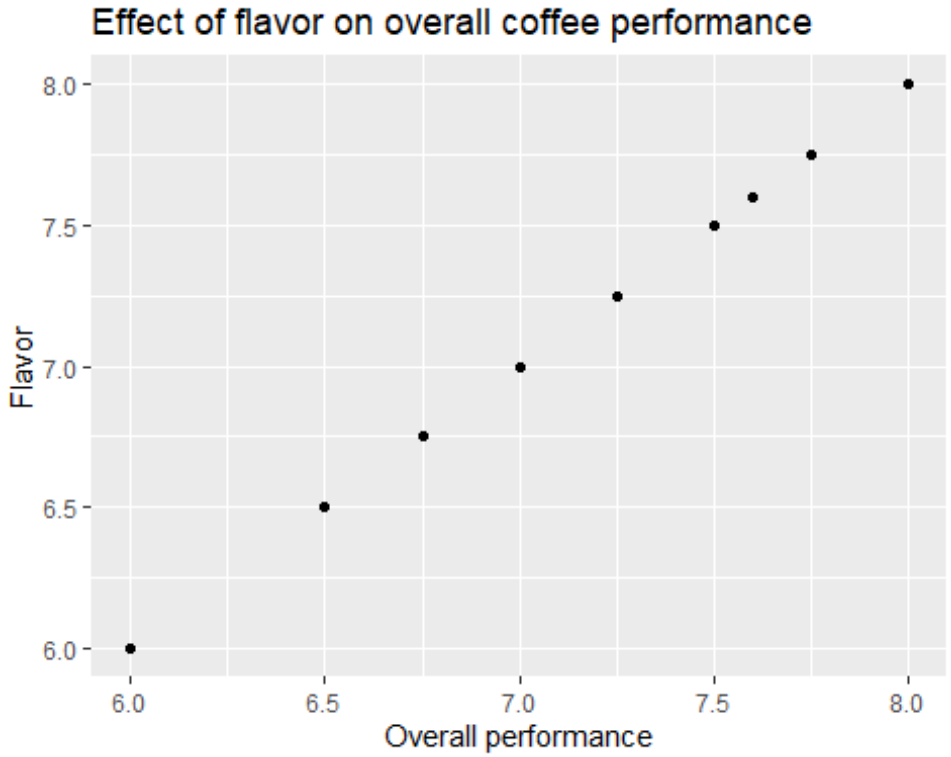
Part b

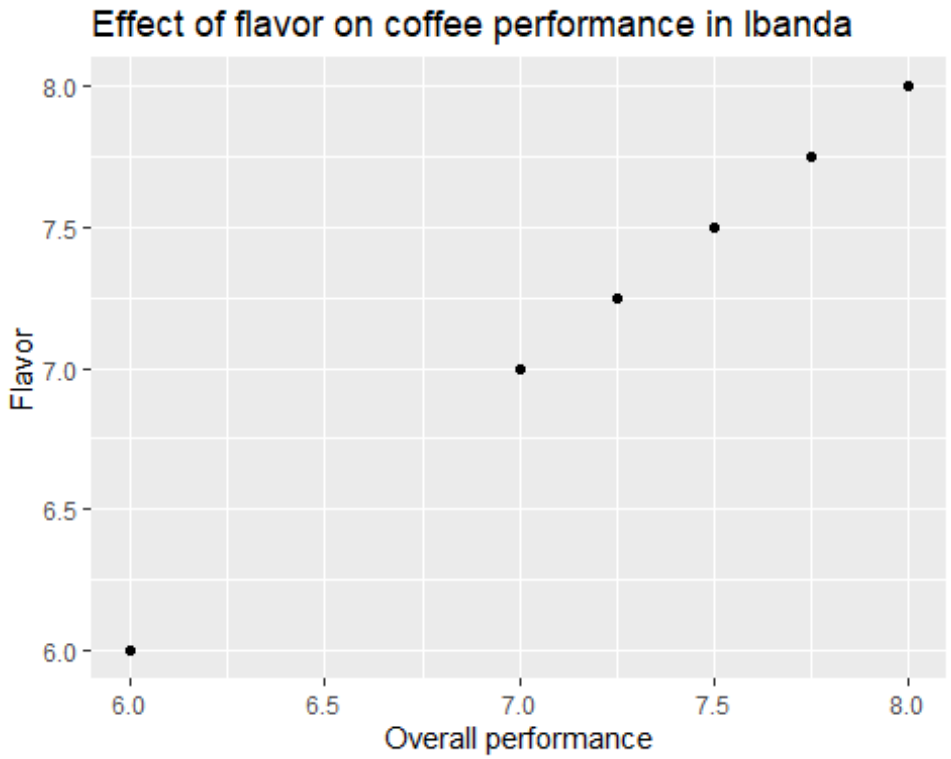
Approach

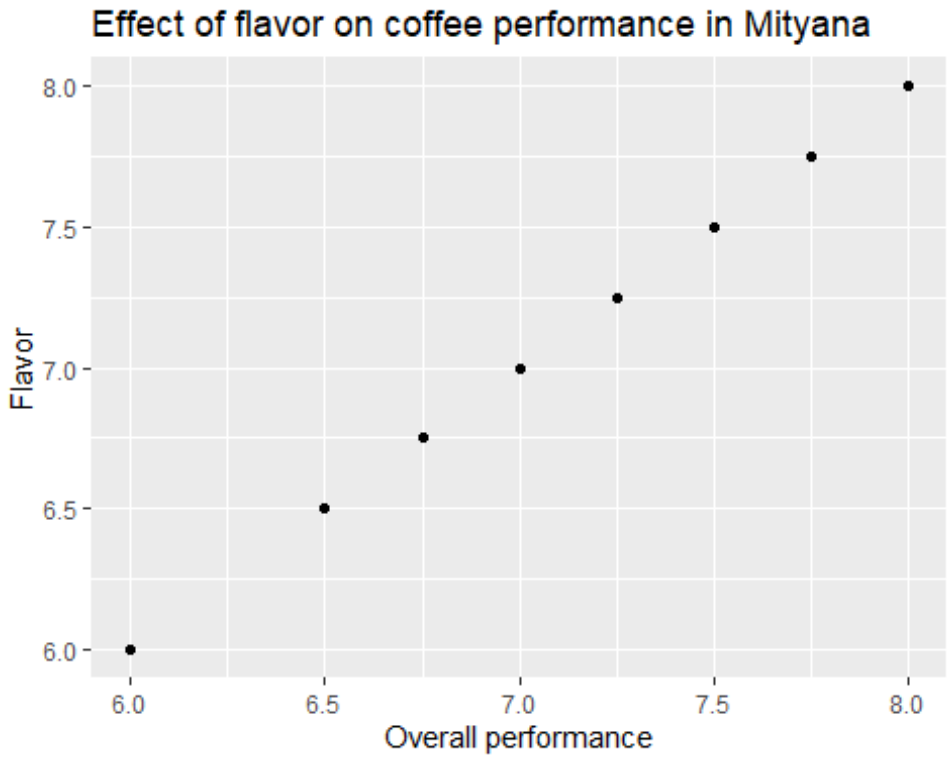
- relationship between the variables
- Looking at effect of flavor on performance per district
- Looking at effect of aroma on performance per district

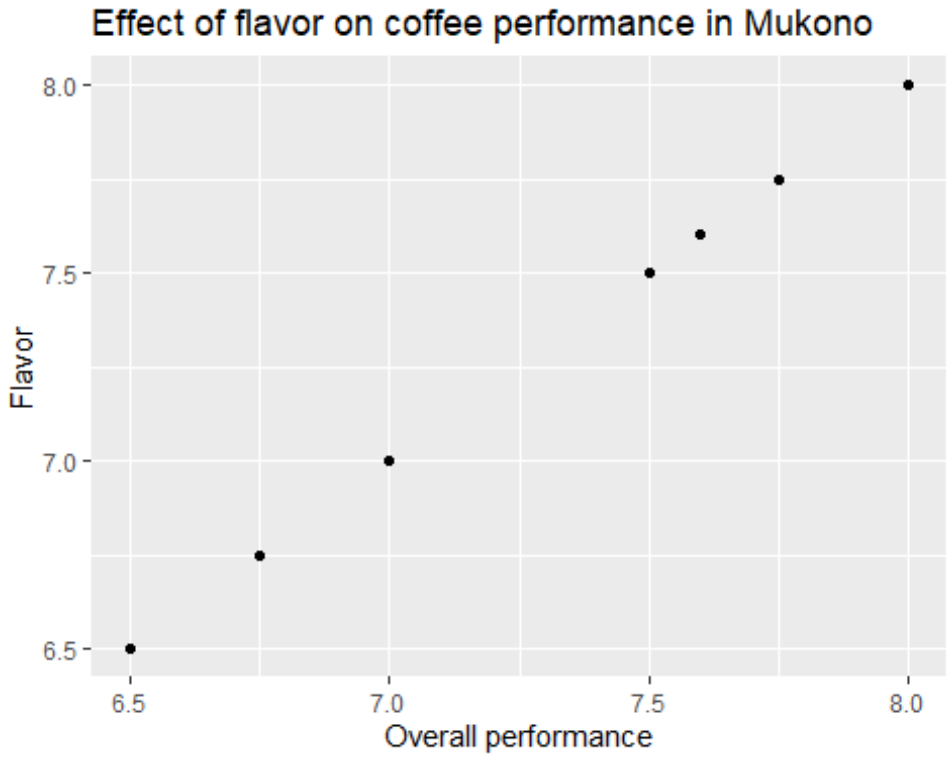
- Looking at effect of aftertaste on performance per district
- Looking at effect of salt/acid on performance per district

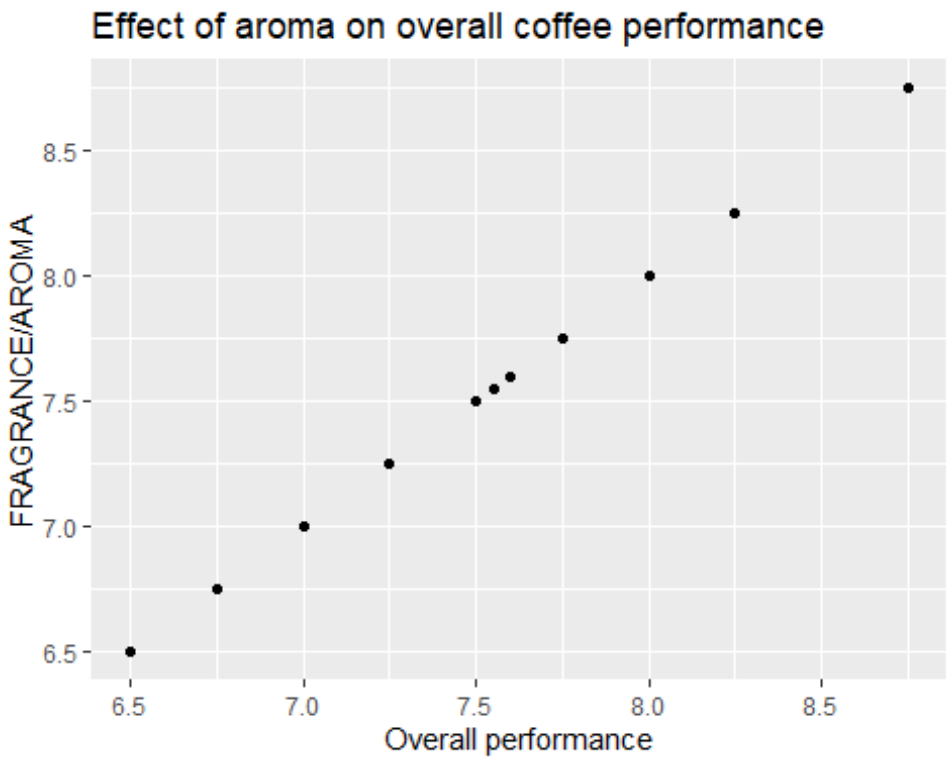
The graphs.

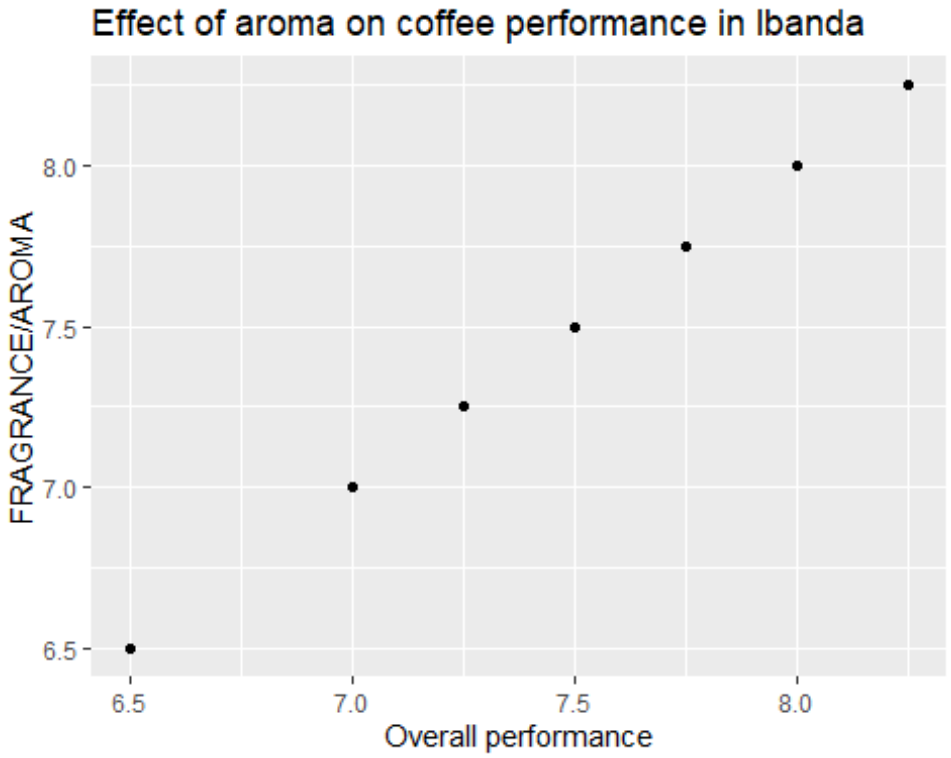


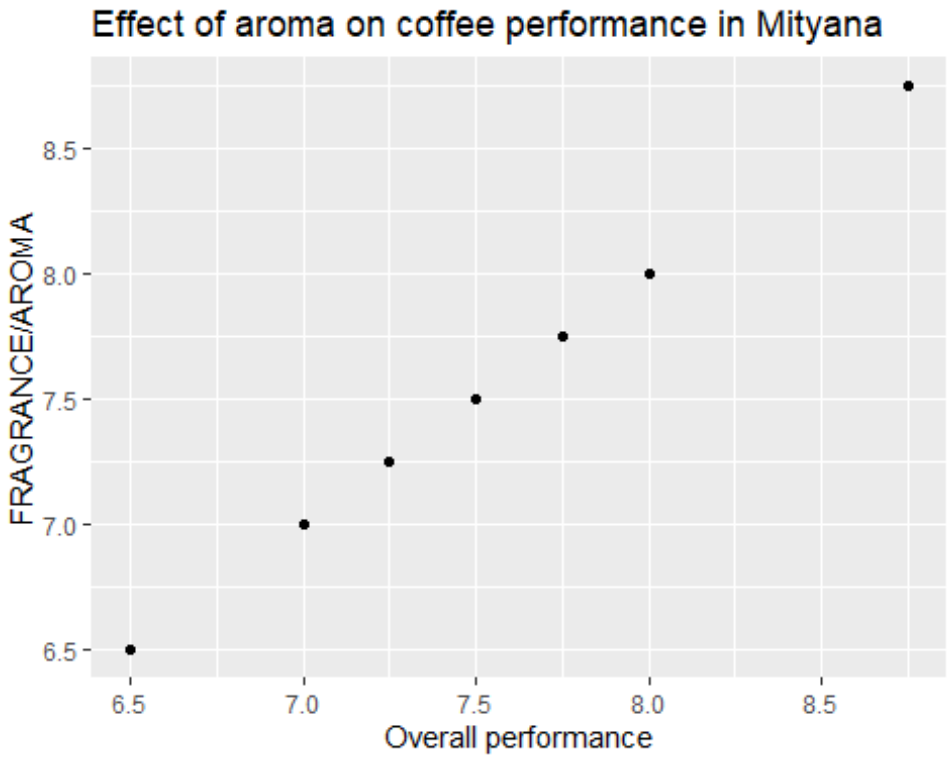


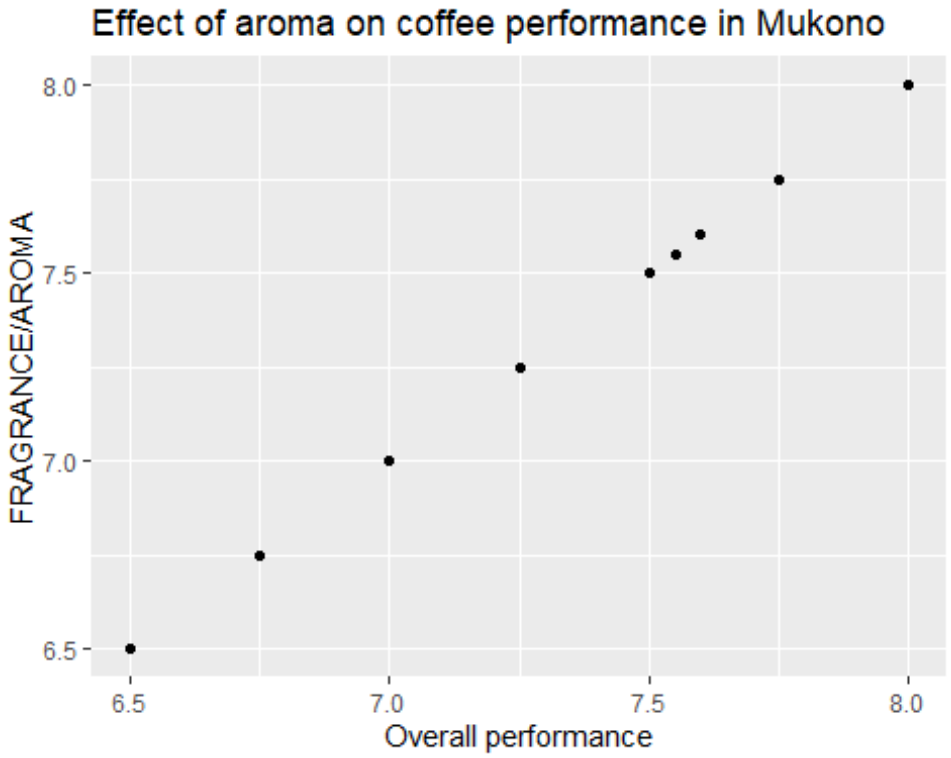

Effect of flavor on coffee performance in Mityana
Flavor
8.0
7.5
7.0
6.5
6.0
6.0
6.5
7.0
7.5
8.0
Overall performance


Effect of flavor on coffee performance in Mukono
Flavor
8.0
7.5
7.0
6.5
6.5
7.0
7.5
8.0
Overall performance

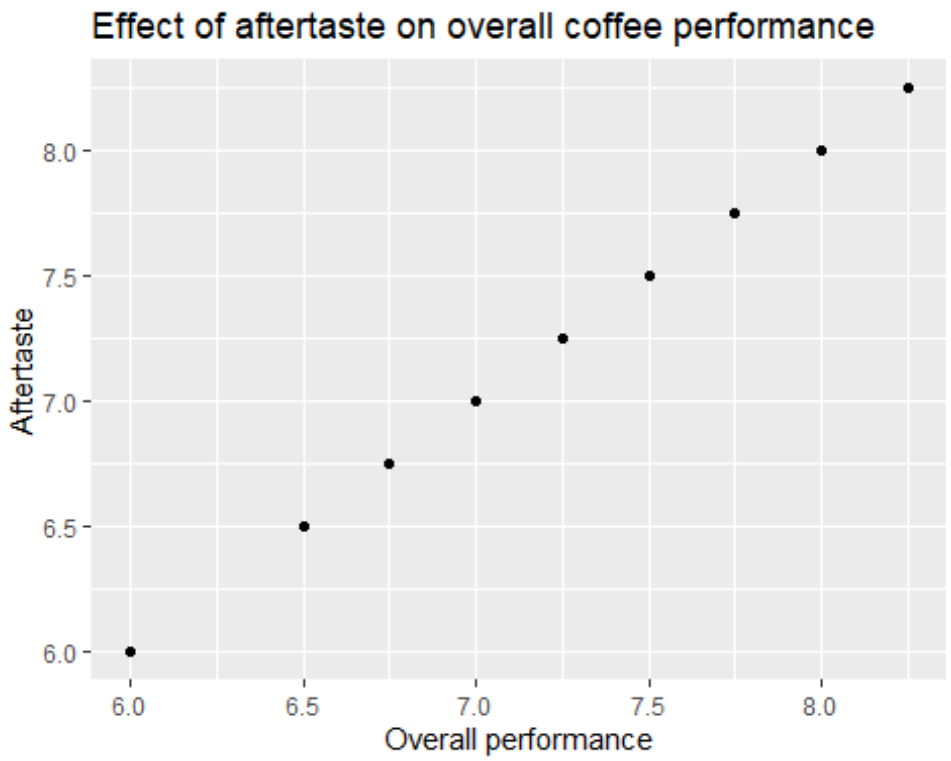

Effect of aroma on overall coffee performance
FRAGRANCE/AROMA
8.5
8.0
7.5
7.0
6.5
6.5
7.0
7.5
8.0
8.5
Overall performance

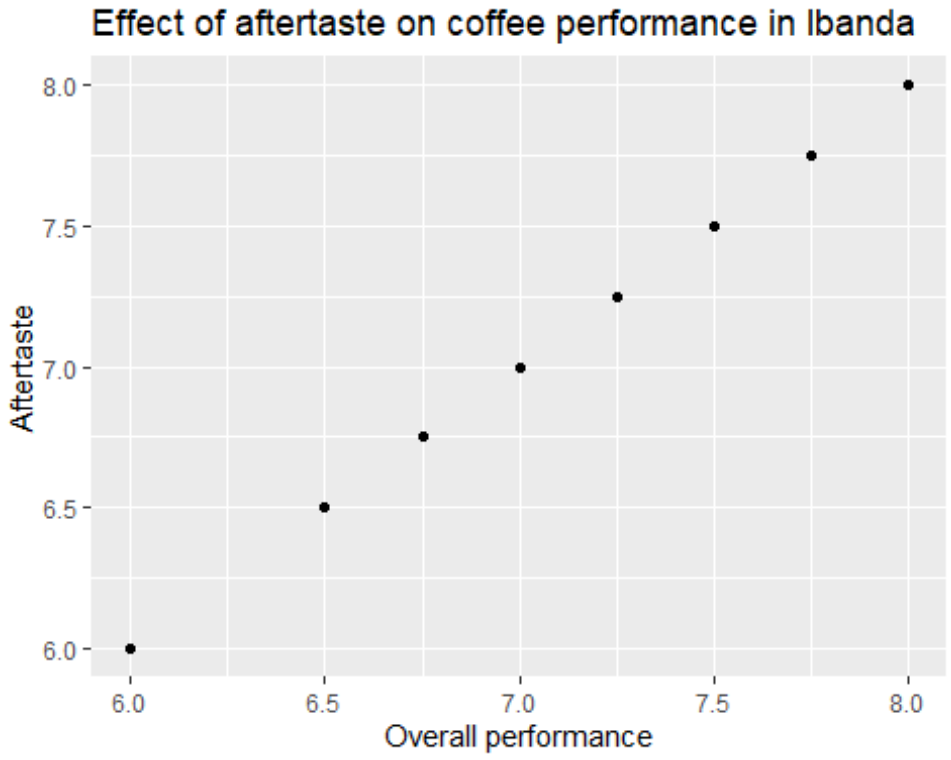

Effect of aroma on coffee performance in Ibanda
FRAGRANCE/AROMA
8.0
7.5
7.0
6.5
6.5
7.0
7.5
8.0
Overall performance

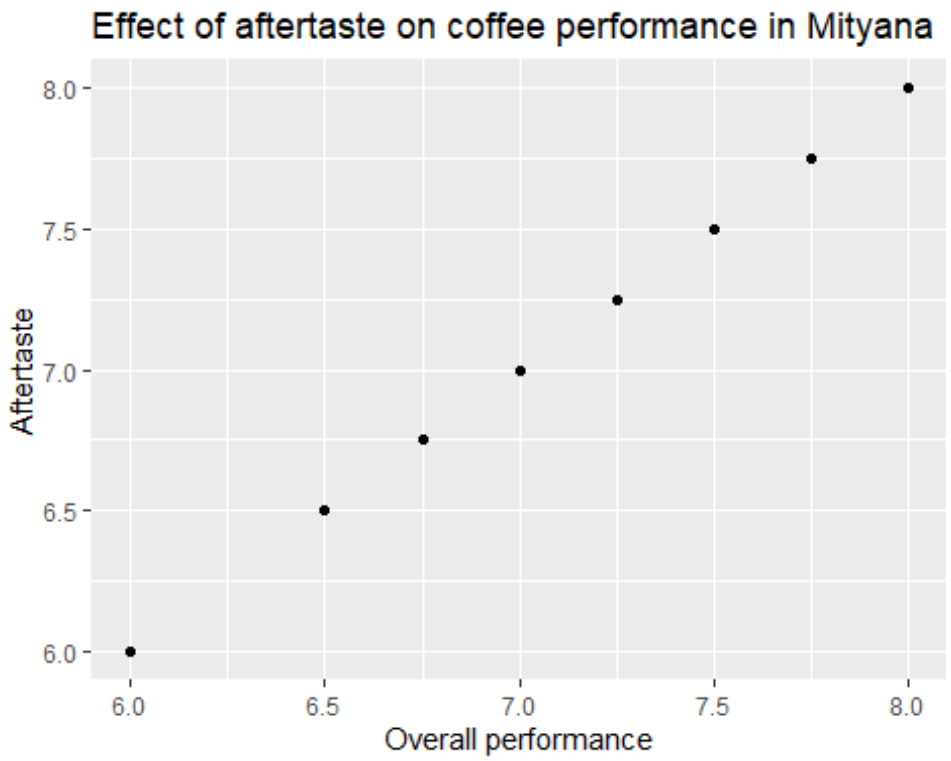


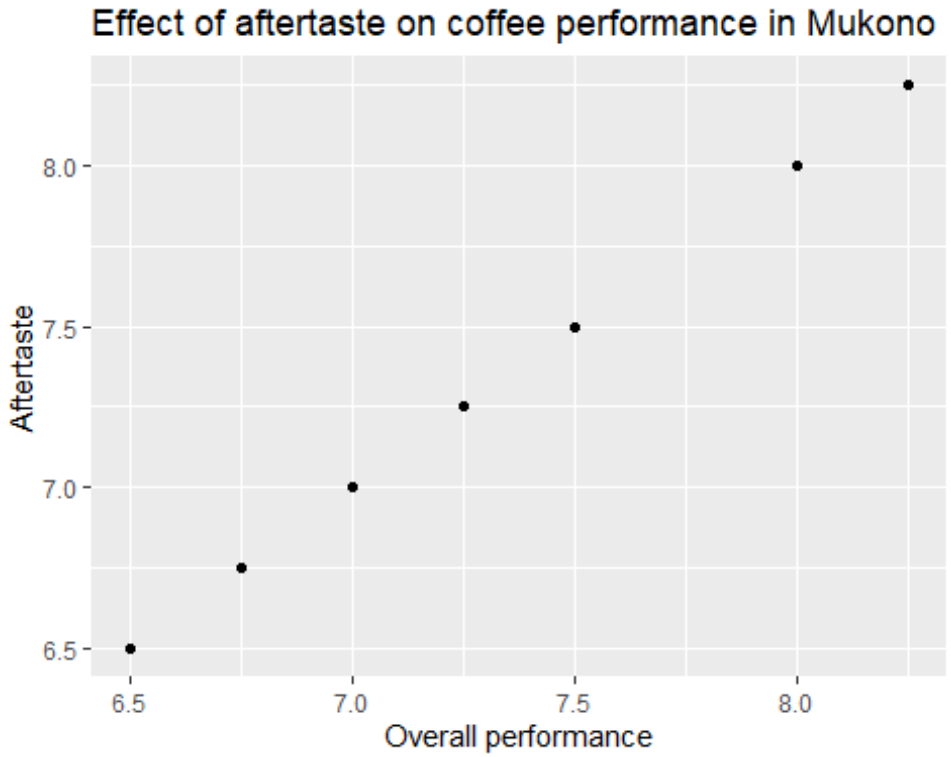


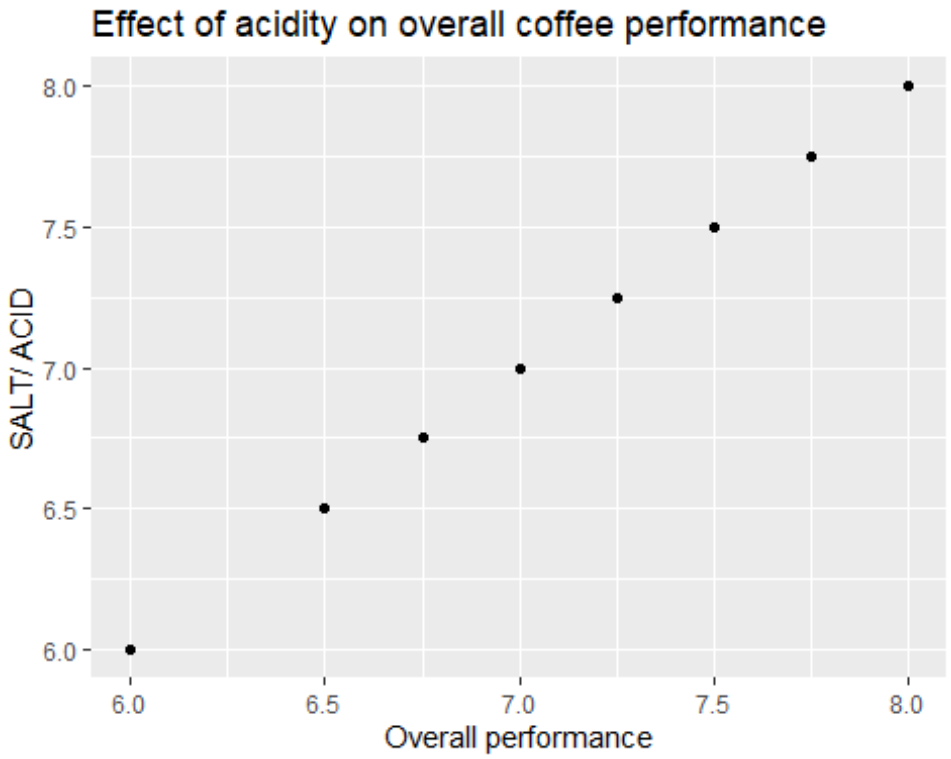

```
## Warning: Use of `coffe_dataset$AFTERTASTE` is discouraged.
## ℹ Use `AFTERTASTE` instead.
```

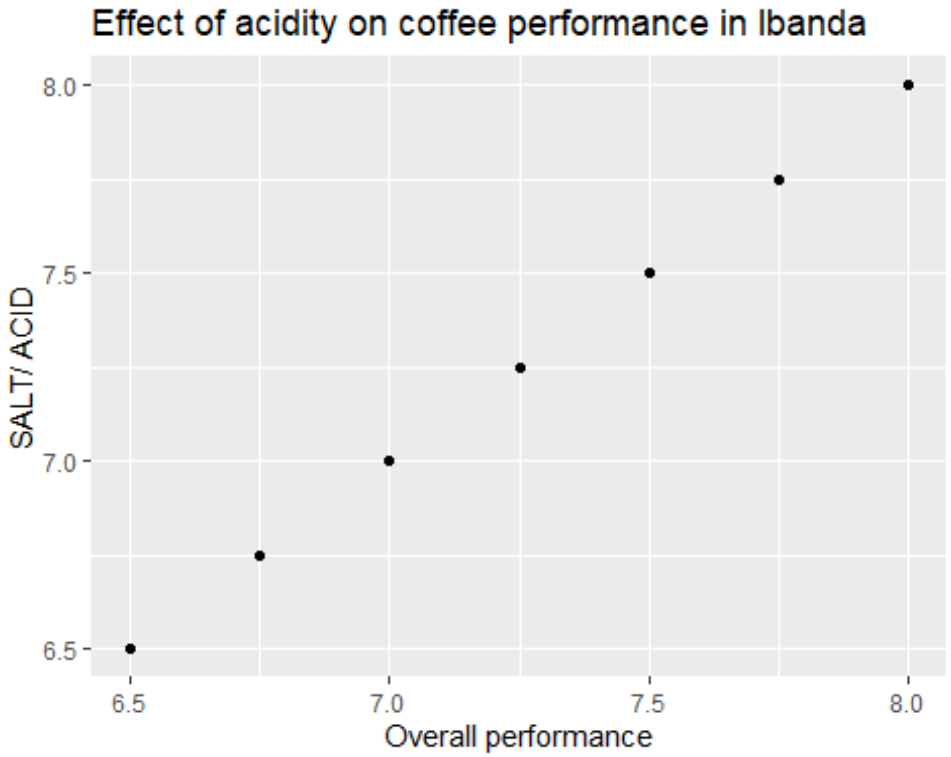

Effect of aftertaste on overall coffee performance
Aftertaste
8.0
7.5
7.0
6.5
6.0
6.0
6.5
7.0
7.5
8.0
Overall performance


Effect of aftertaste on coffee performance in Ibanda
Aftertaste
8.0
7.5
7.0
6.5
6.0
6.0
6.5
7.0
7.5
8.0
Overall performance


Effect of aftertaste on coffee performance in Mityana
Aftertaste
8.0
7.5
7.0
6.5
6.0
6.0
6.5
7.0
7.5
8.0
Overall performance

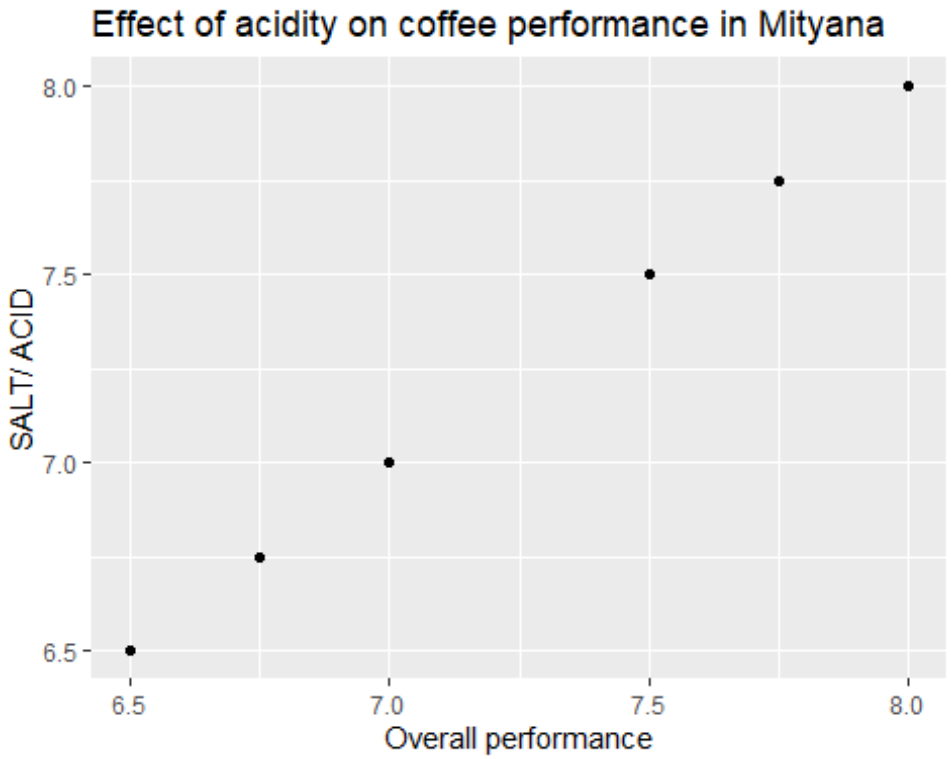

Effect of aftertaste on coffee performance in Mukono
Aftertaste
8.0
7.5
7.0
6.5
6.5
7.0
7.5
8.0
Overall performance

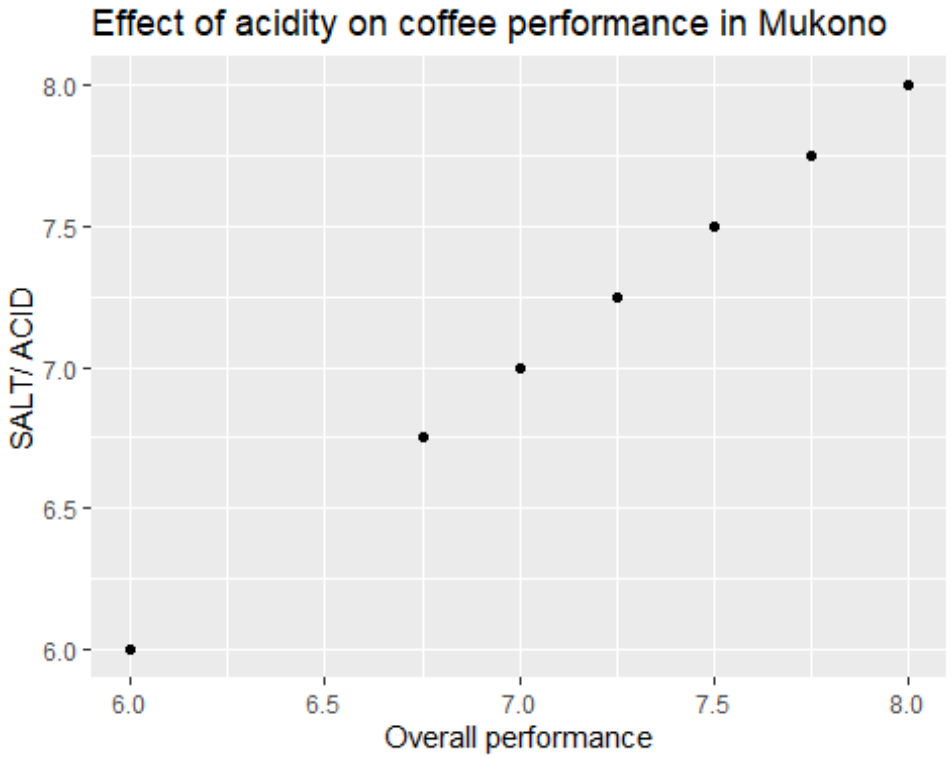





```
## Warning: Use of `Mityana$"SALT/ ACID"` is discouraged.
## ℹ Use `SALT/ ACID` instead.
```

```
## Warning: Use of `Mityana$"SALT/ ACID"` is discouraged.
## ℹ Use `SALT/ ACID` instead.
```

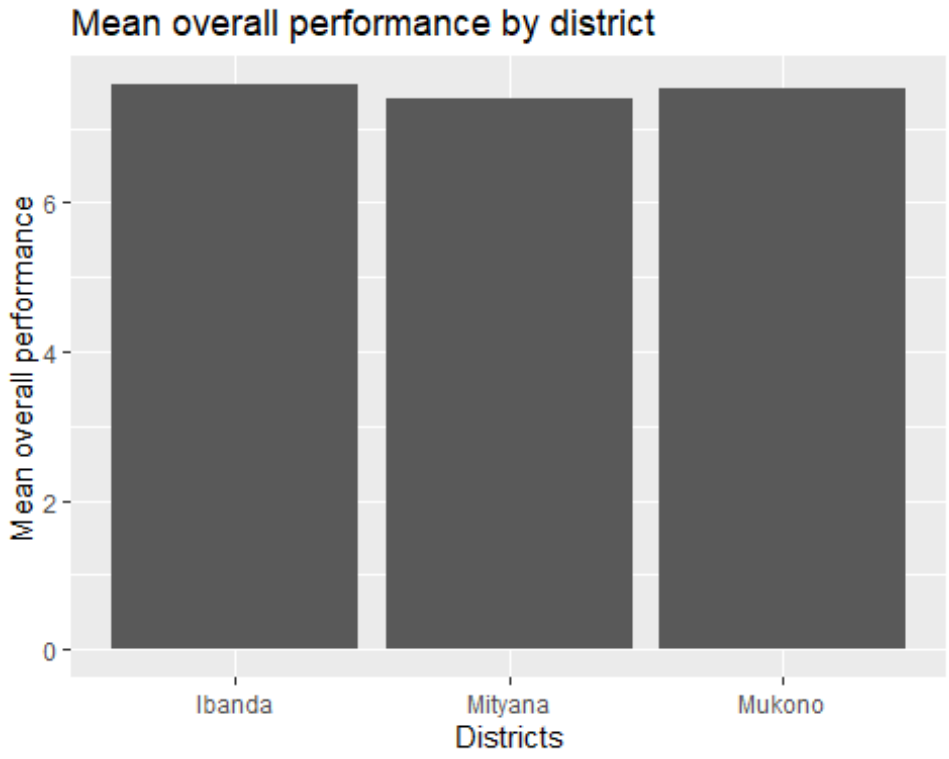




Part c(i)

PERFORMANCE BY VARIETY AND DISTRICT

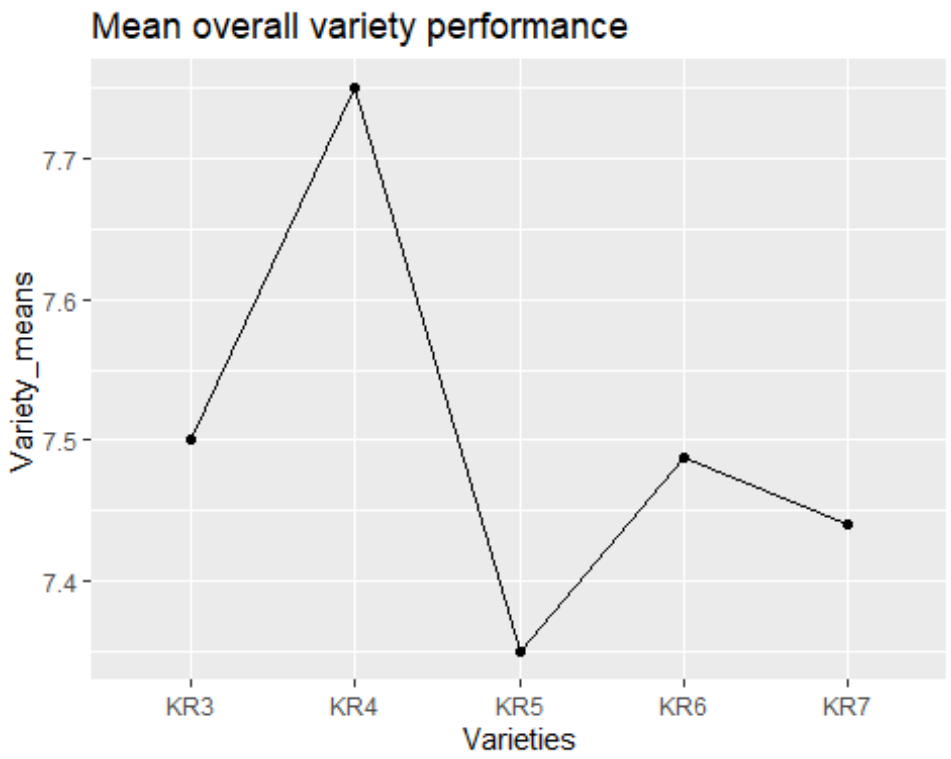
concentrating on the mean of the overall performance



Part c(ii)

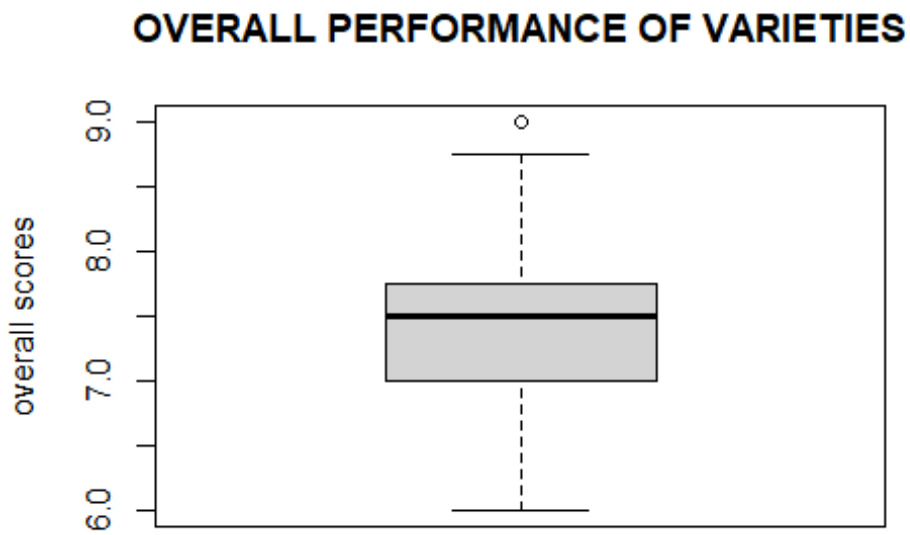
Performance of each variety

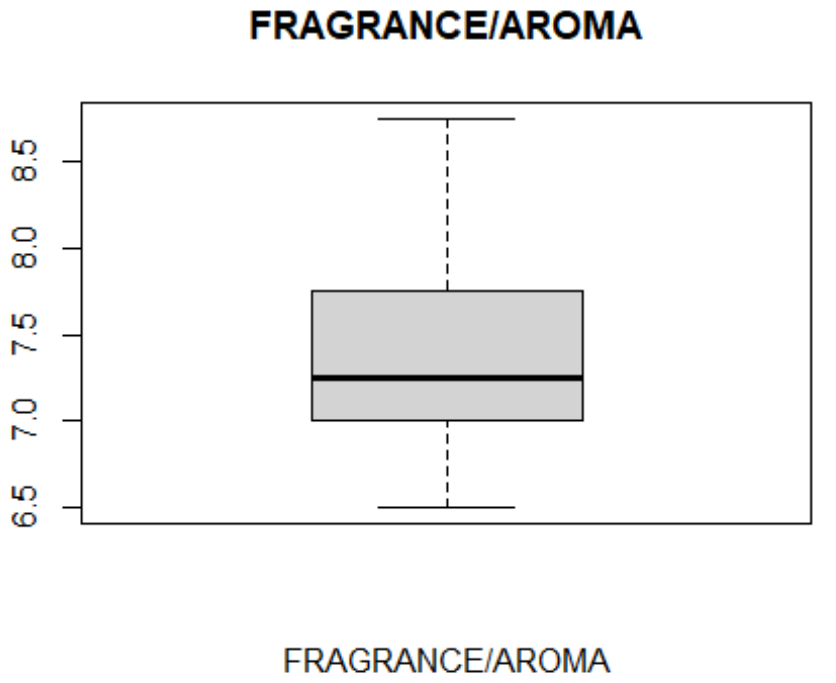
- Transform the coffe_dataset to subselected varieties individually
- calculating the mean of each variety
- Plot the means

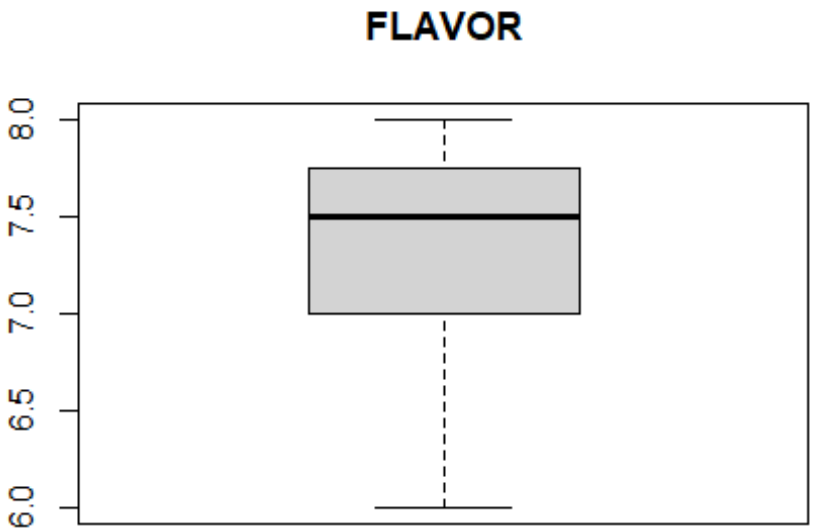

Mean overall variety performance
Variety_means
7.7
7.6
7.5
7.4
KR3
KR4
KR5
KR6
KR7
Varieties

Part d The distibution central tendency of the overall performance

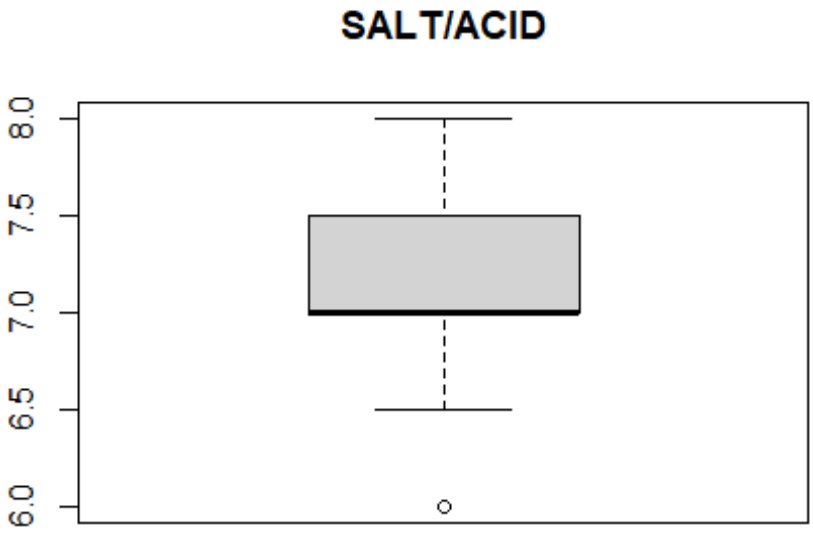
the distibution central tendency of each variable



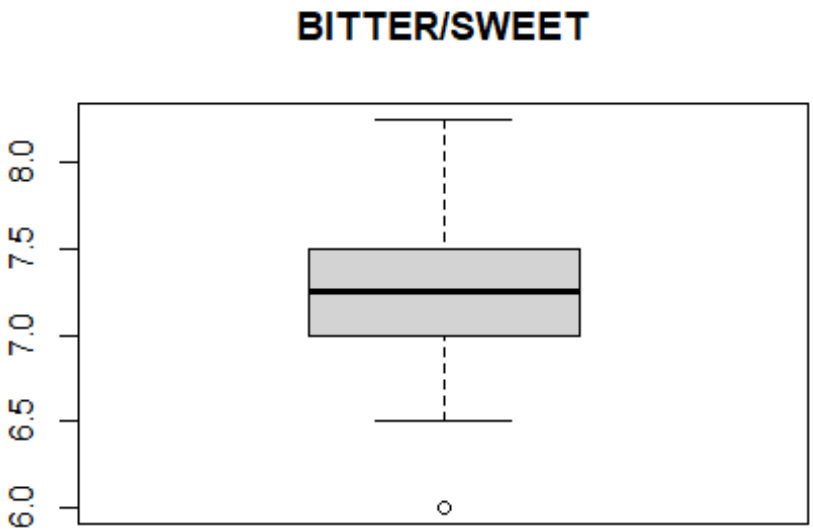


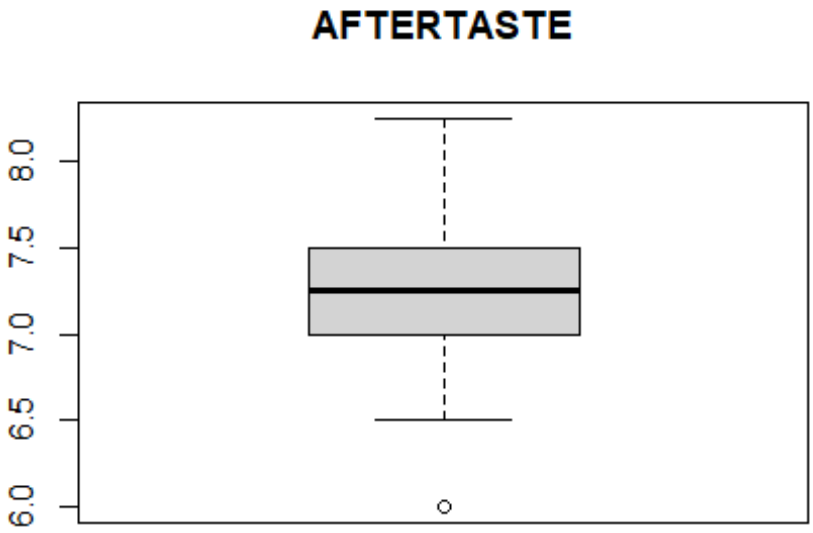


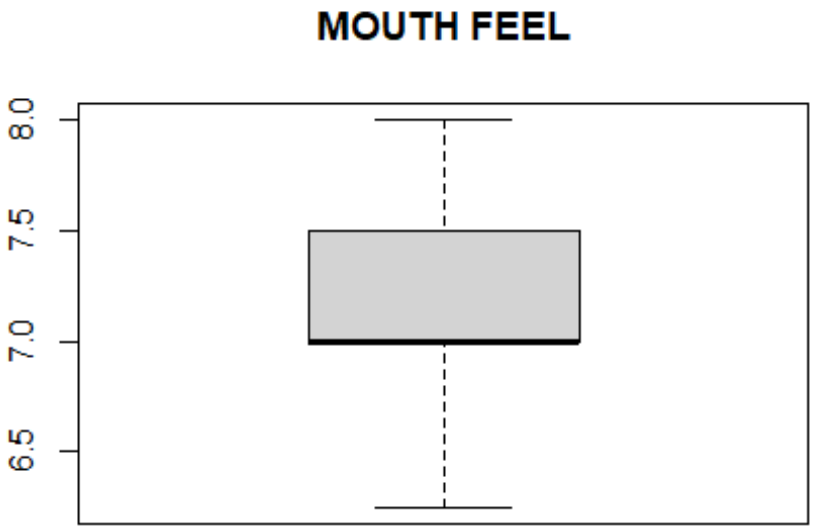
FLAVOR

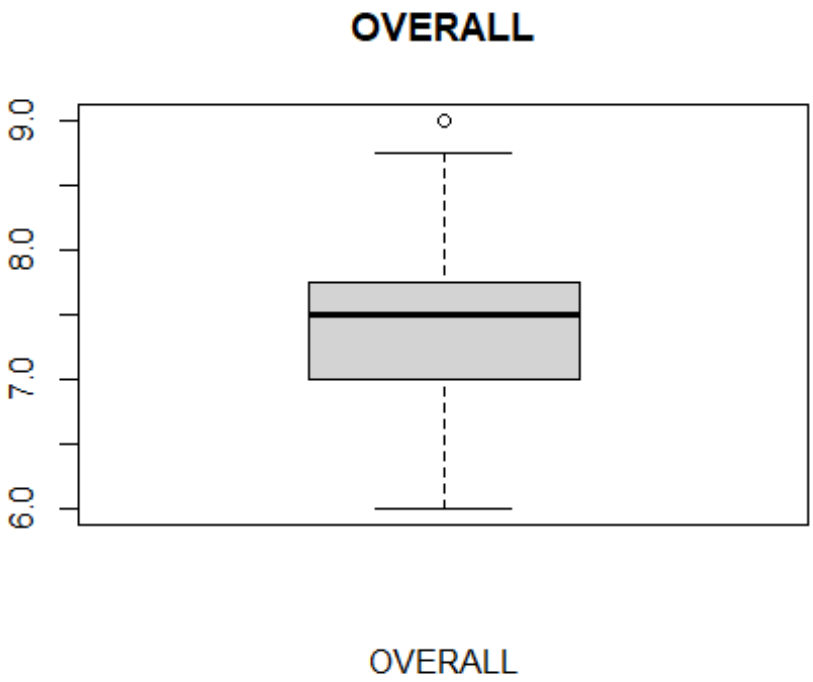


SALT/ACID


BITTER/SWEET
8.0
7.5
7.0
6.5
6.0
BITTER/SWEET


AFTERTASTE
8.0
7.5
7.0
6.5
6.0
AFTERTASTE





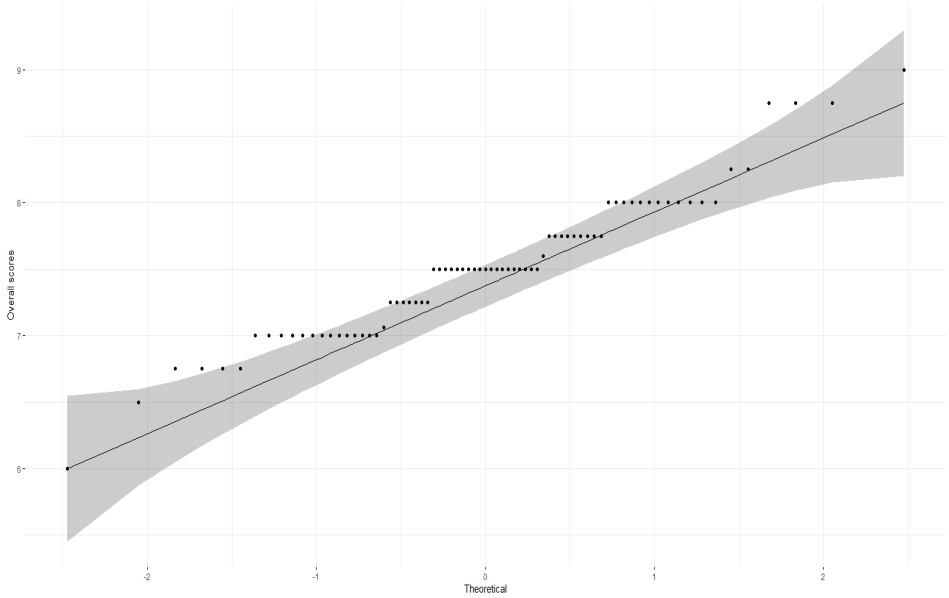
Part e normal distribution tests

- Test the overall variable for normal distribution using the Shapiro-wilk test

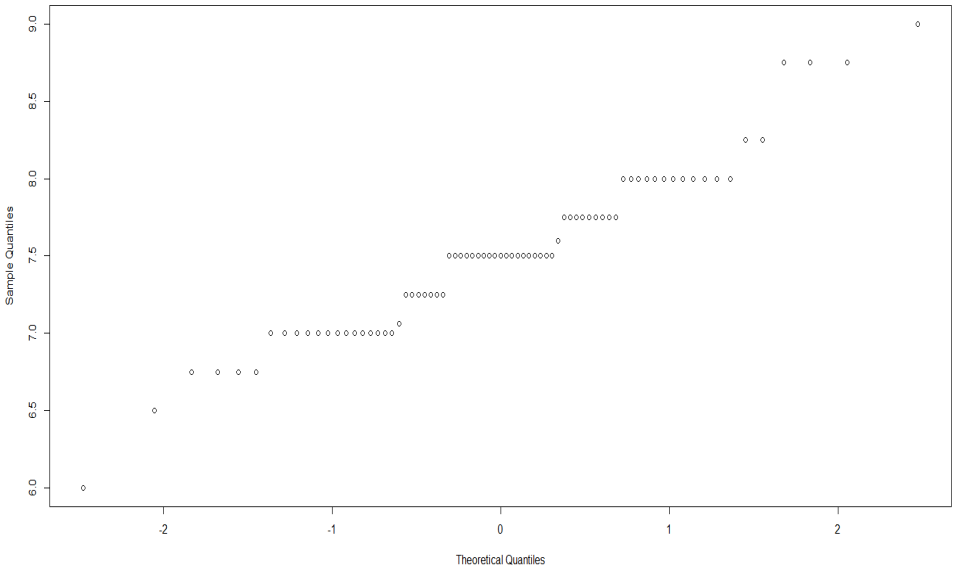
- Null hypothesis based on research question: The overall performance of varieties across - districts is normally distributed. p-value >= 0.05
- Alternative hypothesis: Overall performance is not normally distributed across districts. - p-value =<0.05

```
Shapiro-Wilk normality test

data:  coffe_dataset$OVERALL
 W = 0.95674, p-value = 0.01189
```

- p-value = 0.01189
- Since the p-value is less than 0.05, we therefore fail to reject the null hypothesis.



Visualize the distribution of the overall variable



- the data is not normally distributed therefore we will use the non-parametric test
- One-sample Wilcoxon signed-rank test

```
  Wilcoxon signed rank test with continuity correction

 data:  coffe_dataset$OVERALL
 V = 2850, p-value = 4.371e-14
alternative hypothesis: true location is not equal to 0.8
```

Assignment two markdown

Packages used include;

- readxl
- tidyverse
- ggplot2
- ggpubr
- dplyr

Importing data

Question 1

removing the missing values

```
work <- na.omit(work)
```

Question 2

Show the relationship between the prices and perception change.

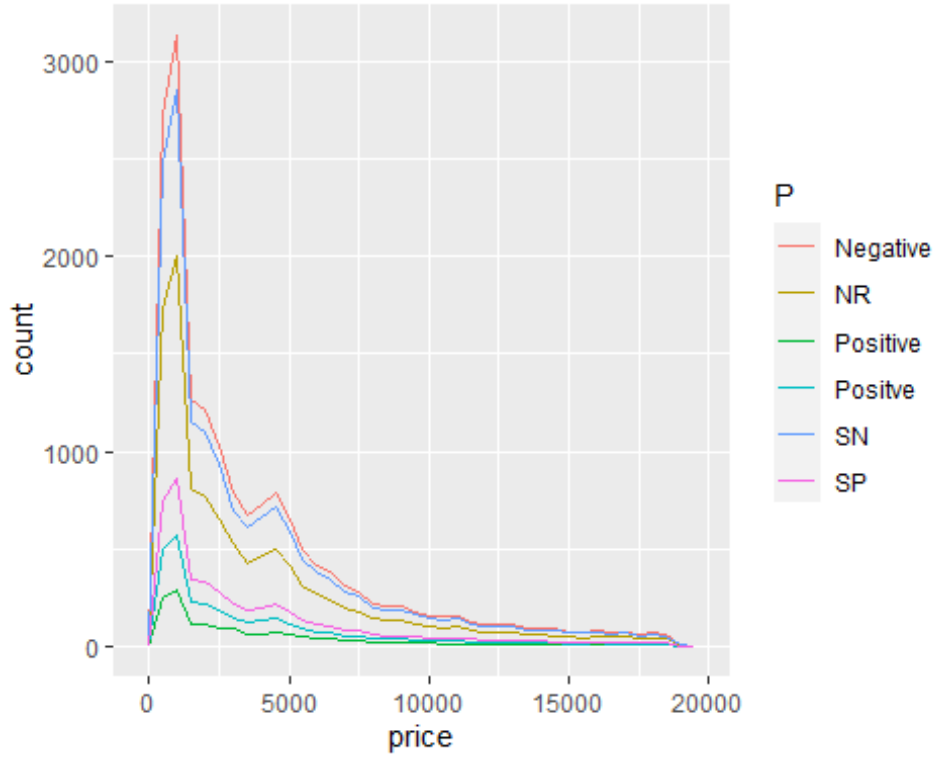
Appoach

- define the perception values.
- compute the percentage of each perception in response to price.

Results as percentages of all perceptions.

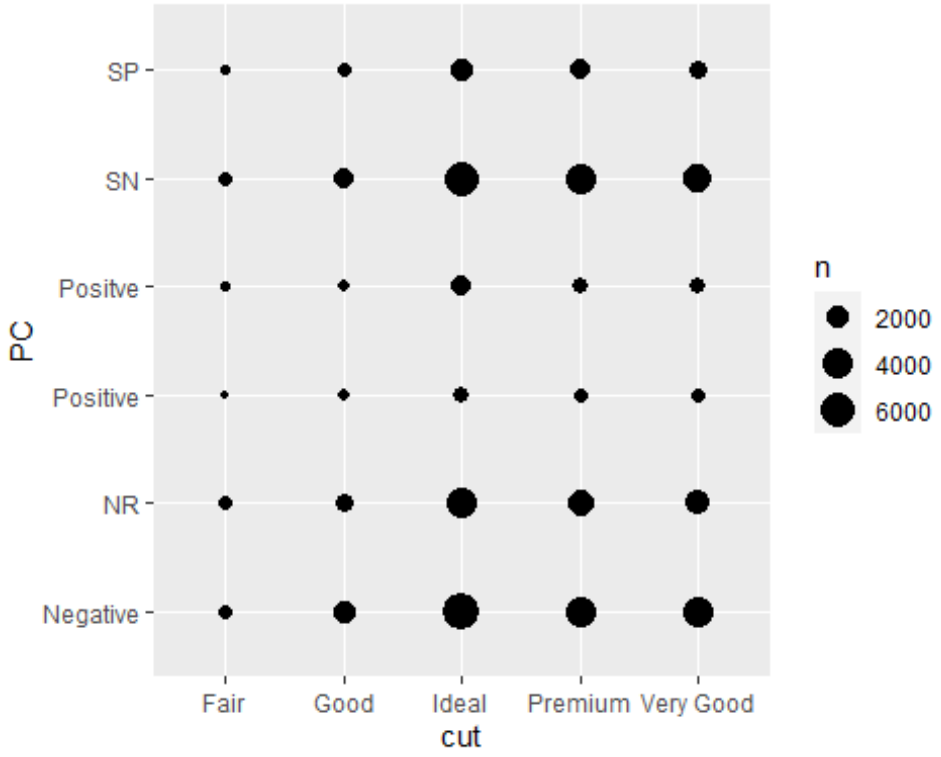
```
Positive

[1] 2.977934

Negative

[1] 32.33265

partial_postive

[1] 8.831819

partial_negative

[1] 29.34545

Nuetral

[1] 20.63416
```

Displaying the covariation between the perception and price



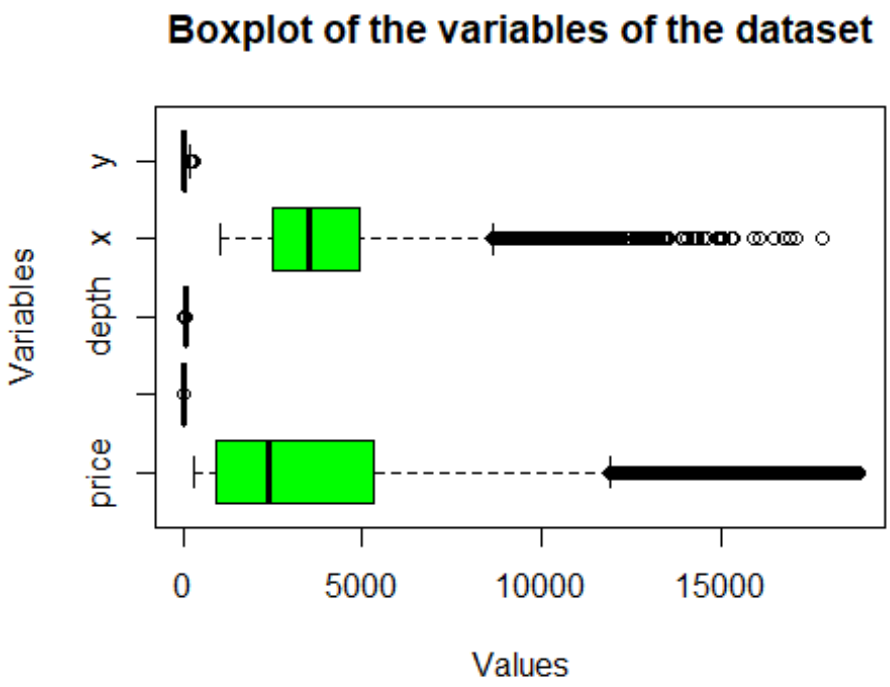
Question 3

Compare the perception change and diamond quality



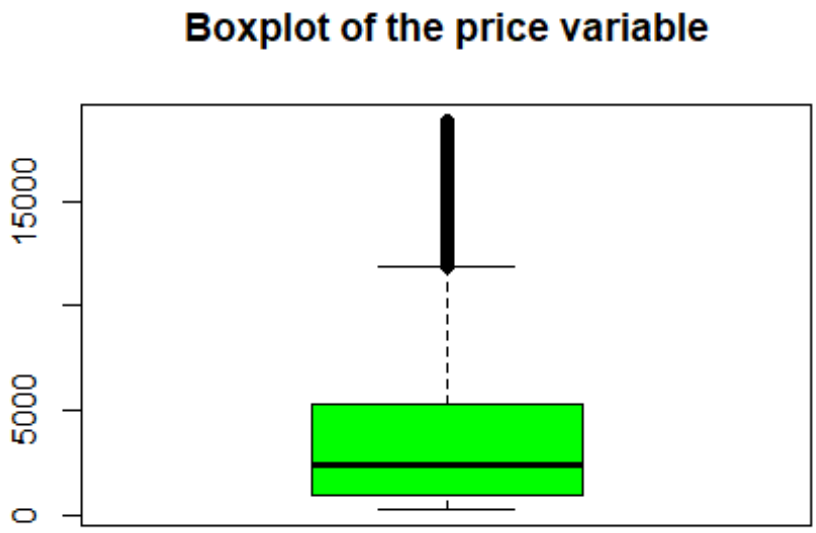
Question 4

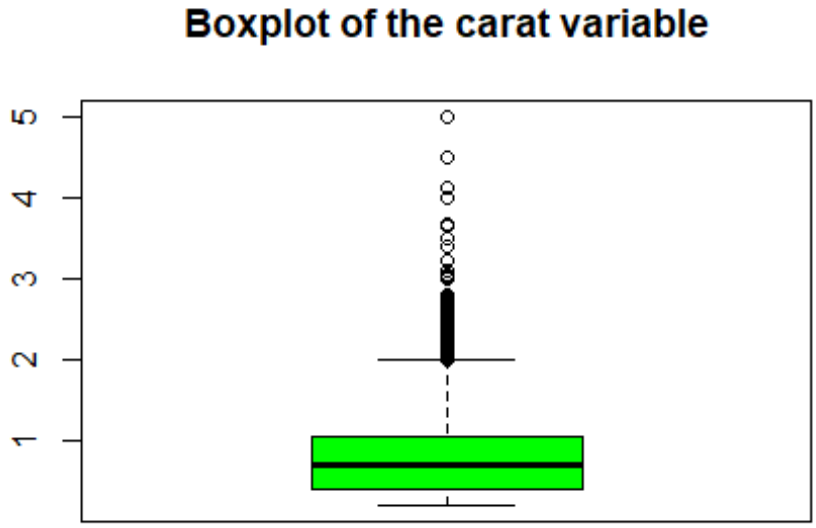
Generating a boxplot of all the variables

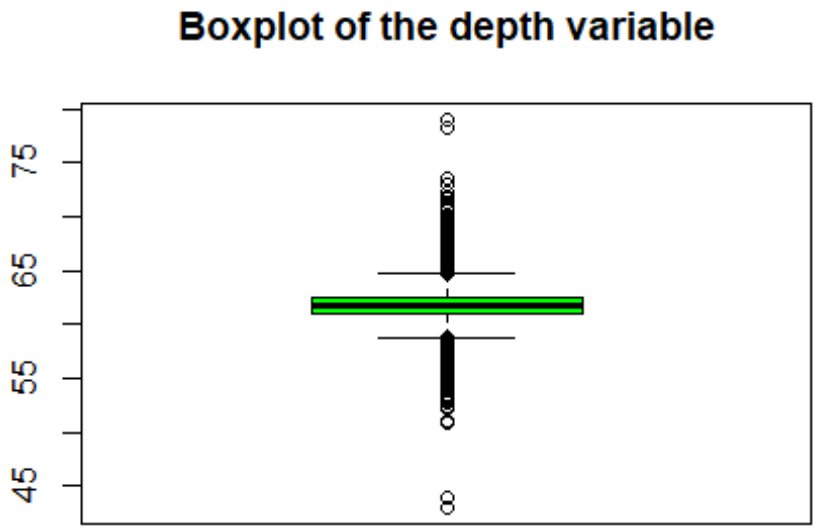


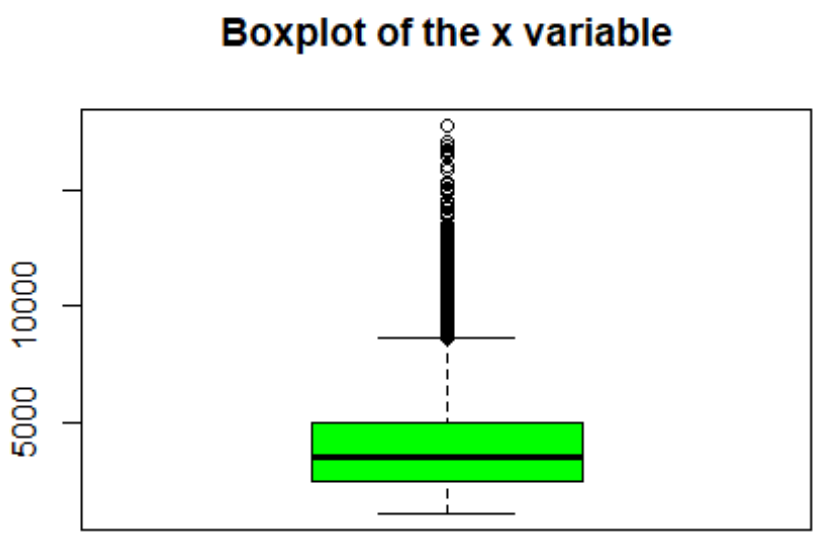
Question 5

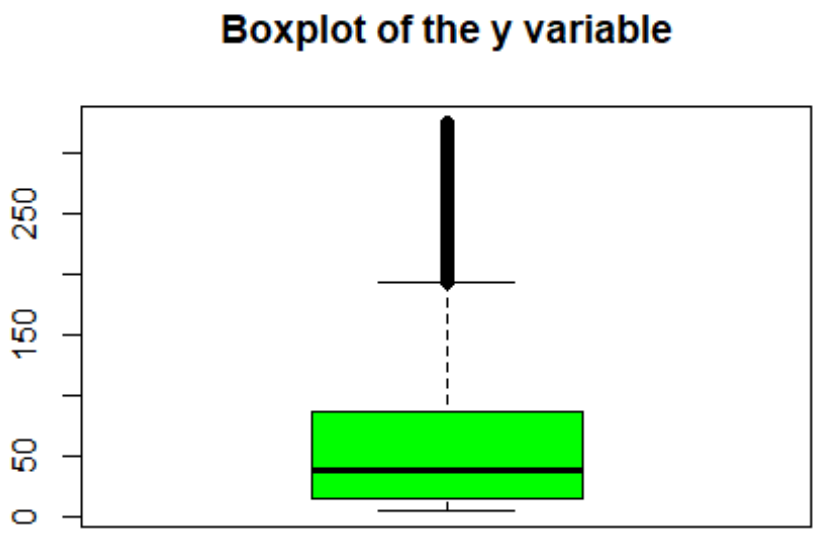
Generating induvidual boxplots of the variables






Boxplot of the depth variable
75
65
55
45


Boxplot of the x variable
10000
5000



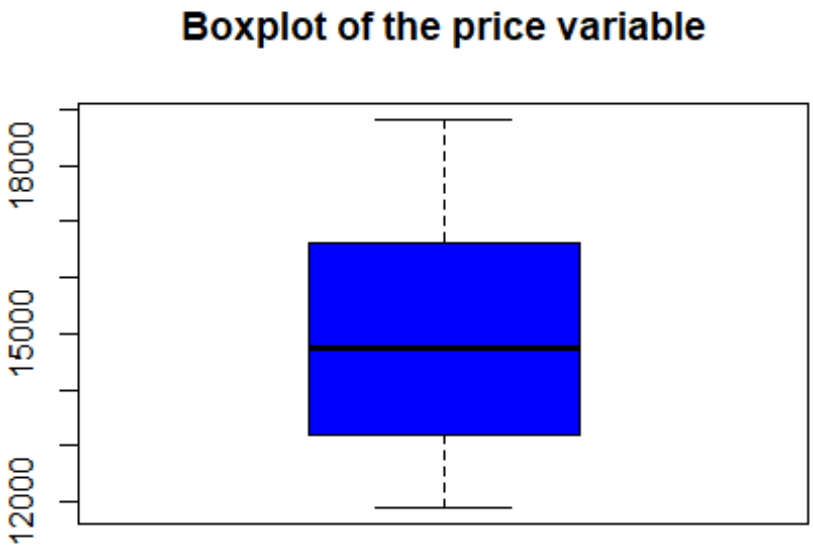
A box plot that labels out the outliers

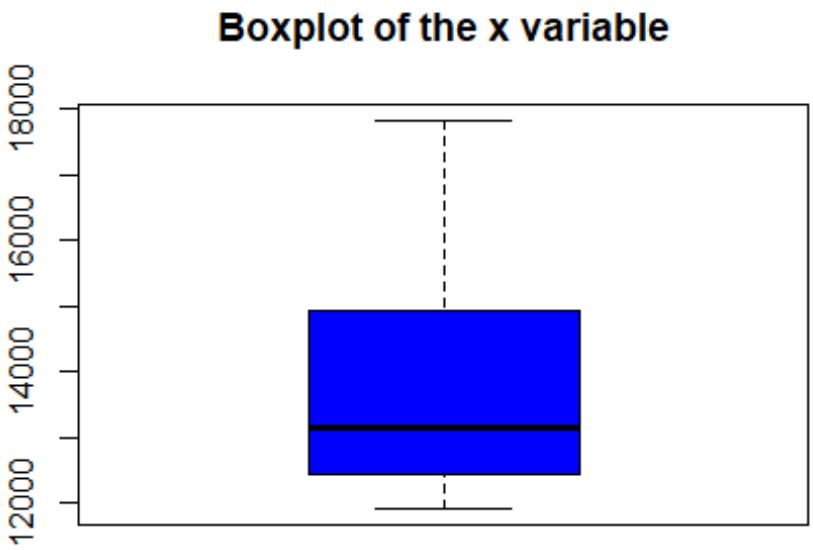
Using the interquartile range to identify the outliers

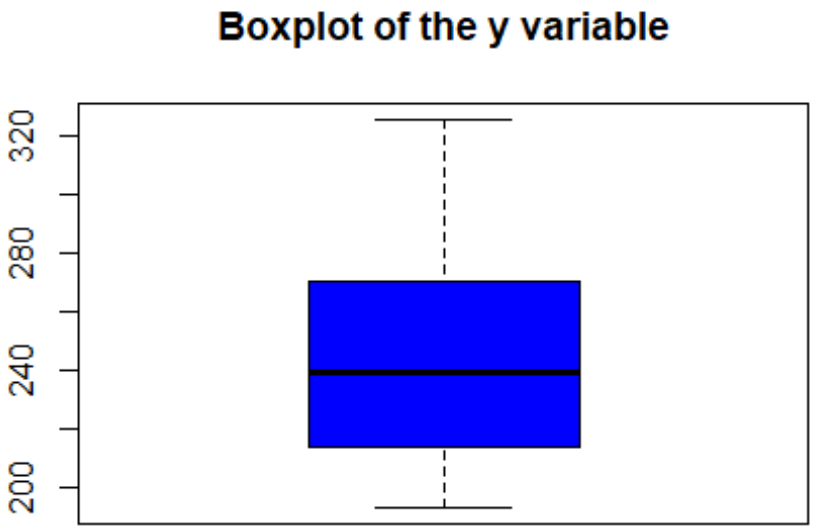
Approach

- Compute the interquartile range.
- Compute the upper and lower limits.
- Identify the outliers.

Plots without outliers


Boxplot of the price variable
18000
15000
12000


Boxplot of the x variable
18000
16000
14000
12000



Question 6

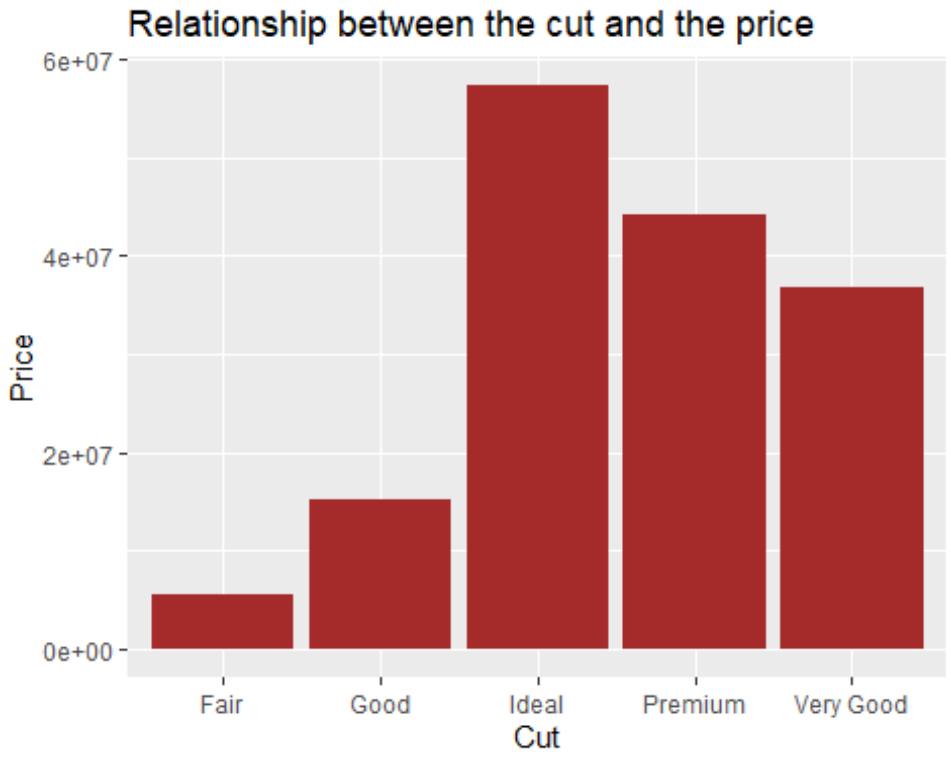
creating a new csv file with the outliers removed

```
write.csv(new_work, file="D:/R/Assignment2_MugangaCharles.csv")
```

Question 7

Display the relationship between one qualitative variable and one finite variable in the dataset

- the selected variables are cut and price.



Question 8

Compute the variance between three groups; diamond carat, perception change and price

Appoach

- Comparing mean, median and mode
- Calculating the mean and median of the variable “carat” in the diamonds dataset

```
## [1] 0.7239025
```

```
## [1] 0.7
```

results

- Mean =0.72, Median = 0.7.
- The performance of carat is positively skewed.

Question 9

Appoach

- Compute the variance between three groups; diamond carat, perception change and price

- the groups are carat, price and perception change
- the variables are carat, PC and price
- The null hypothesis is that the groups have the same variance
- The alternative hypothesis is that the groups have different variances

Steps

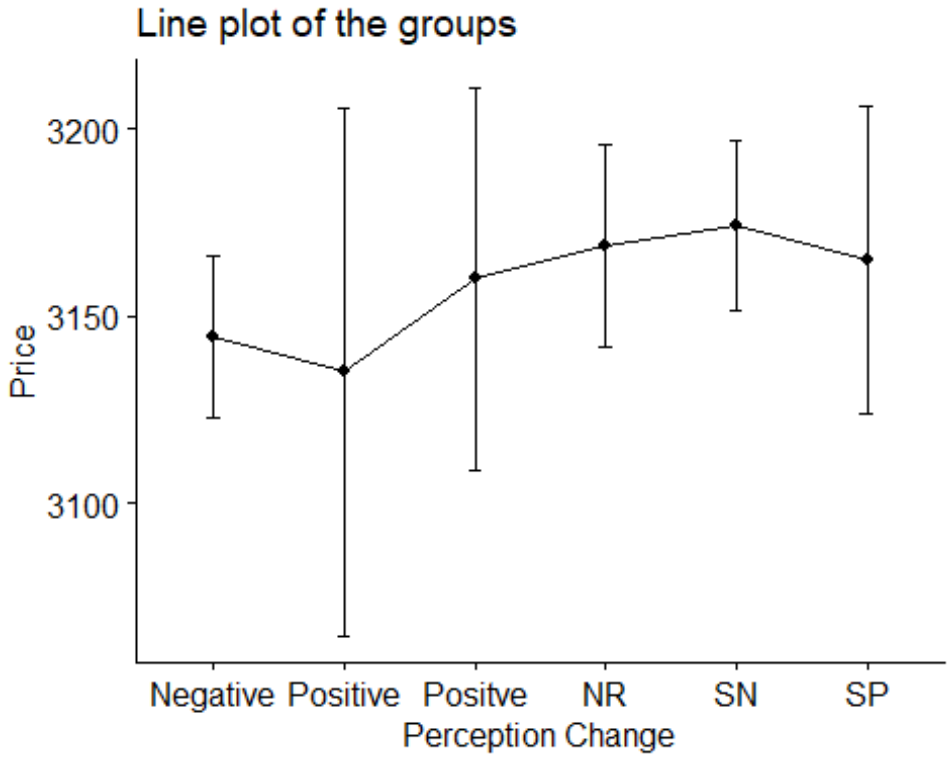
- view the groups

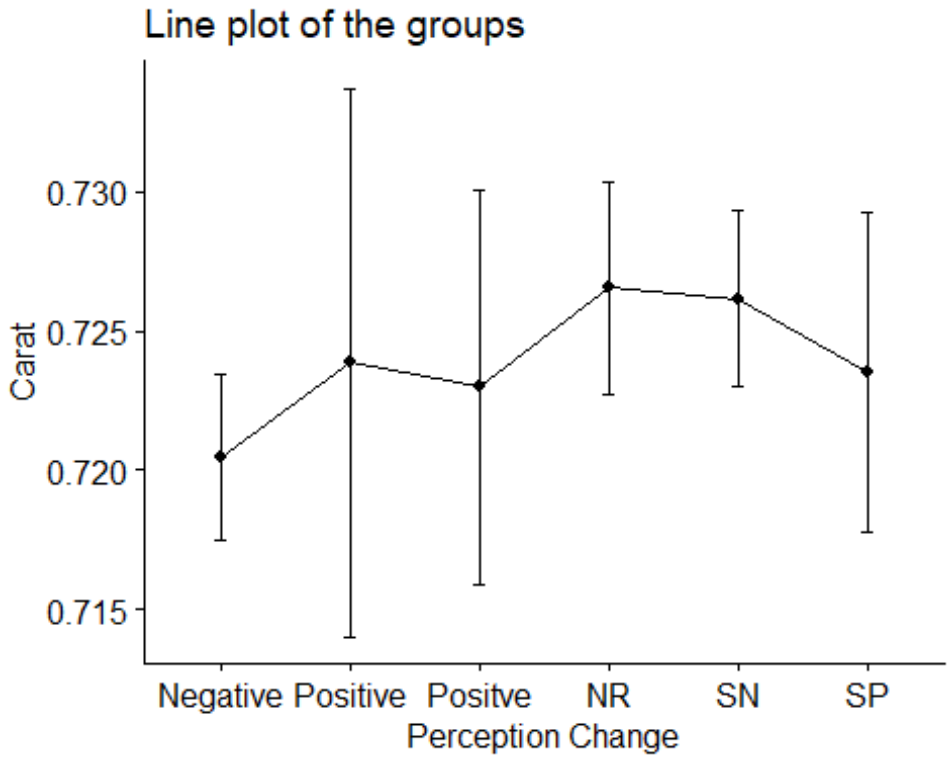
```
## # A tibble: 50,393 × 3
##    carat PC       price
##    <dbl> <chr>    <dbl>
##  1  0.23 Negative   326
##  2  0.21 Negative   326
##  3  0.23 Negative   327
##  4  0.29 Negative   334
##  5  0.31 Negative   335
##  6  0.24 Negative   336
##  7  0.24 Negative   336
##  8  0.26 Negative   337
##  9  0.22 Negative   337
## 10  0.23 Negative   338
## # … with 50,383 more rows
```

- Generate the random sample of the data.

```
## # A tibble: 6 × 4
##   PC        mean    sd     n
##   <chr>    <dbl> <dbl> <int>
## 1 Negative 3144. 2768. 16291
## 2 NR       3169. 2767. 10404
## 3 Positive 3135. 2741.  1501
## 4 Positve  3160. 2779.  2964
## 5 SN       3174. 2771. 14789
## 6 SP       3165. 2755.  4444
```

Plots


Line plot of the groups
Price
3200
3150
3100
Negative
Positive
Positve
NR
SN
SP
Perception Change


Line plot of the groups
Carat
0.730
0.725
0.720
0.715
Negative
Positive
Positve
NR
SN
SP
Perception Change

Computng the variance betwen the groups

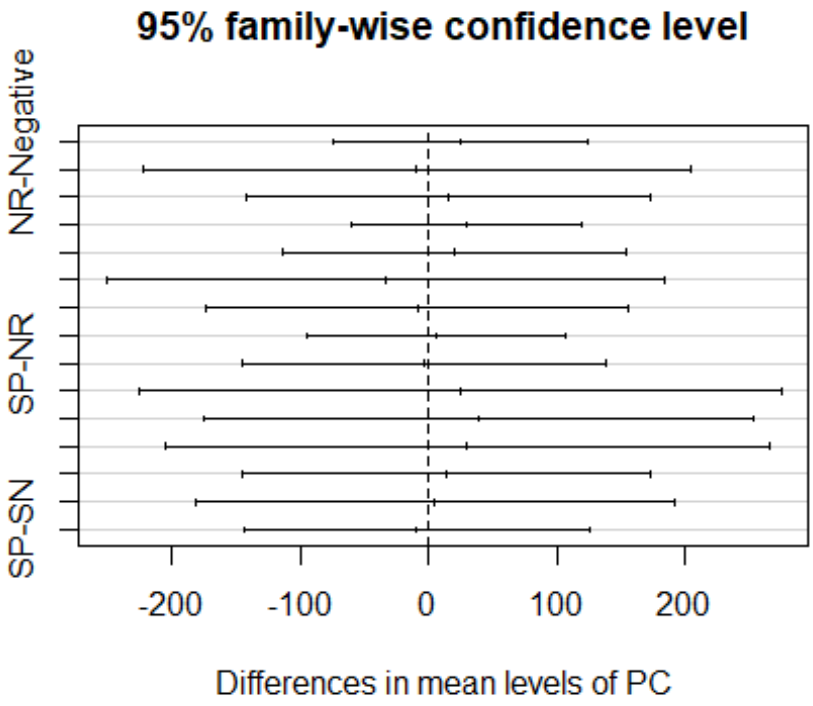
```
##                Df    Sum Sq Mean Sq F value Pr(>F)
## PC              5 8.637e+06 1727312   0.226  0.952
## Residuals   50387 3.859e+11 7658723
```

Commenting on the results

- The p-value is 0.954 which is greater than 0.05 and therefore we fail to reject the null hypothesis therefore statistically not significant
- The posthoc test used is the Tukey HSD test
- The null hypothesis is that the groups have the same variance
- The alternative hypothesis is that the groups have different variances by aleast one group having a variance not equal to the others groups

Question 9(b)

```
TukeyHSD(anova, conf.level = .95)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = price ~ PC, data = groups)
##
## $PC
##                          diff        lwr      upr     p adj
## NR-Negative         24.169212  -74.80432 123.1427 0.9824596
## Positive-Negative   -9.318170 -222.04699 203.4106 0.9999958
## Positve-Negative    15.538464 -141.94572 173.0227 0.9997645
## SN-Negative         29.562176  -60.01047 119.1348 0.9360071
## SP-Negative         20.484454 -112.98113 153.9500 0.9979834
## Positive-NR        -33.487382 -251.23446 184.2597 0.9979638
## Positve-NR          -8.630748 -172.83035 155.5689 0.9999896
## SN-NR                5.392965  -95.52036 106.3063 0.9999887
## SP-NR               -3.684758 -145.01169 137.6422 0.9999997
## Positve-Positive    24.856634 -224.98191 274.6952 0.9997547
## SN-Positive         38.880347 -174.75787 252.5186 0.9954693
## SP-Positive         29.802624 -205.63546 265.2407 0.9992041
## SN-Positve          14.023713 -144.68674 172.7342 0.9998632
## SP-Positve           4.945990 -182.08024 191.9722 0.9999997
## SP-SN               -9.077723 -143.98807 125.8326 0.9999644

plot(TukeyHSD(anova, conf.level = .95))
```



- The Tukey HSD test shows that the groups have different variances by aleast one group having a variance not equal to the others groups.