**A220A0010 Free Analytics Environment R – Second Assignment – Autumn 2023**

**Part-1**

**Regression Analysis**

| Serial No | Topics covered |
|---|---|
| 1 | Executive summary |
| 2 | Introduction and Goal |
| 3 | Load and study the data set, comment on data set's structure |
| 4 | Conduct explanatory data analysis of Five explanatory variables of choice |
| 5 | Perform correlation analysis between all variables; comment on the variables that have highest linear association with the dependent variable |
| 6 | Visualize on which pair of variables have the highest absolute correlation |
| 7 | Remove explanatory variables with highest correlation value and comment on the reasons |
| 8 | Implement a first linear regression after removing explanatory variables in the previous task |
| 9 | Comment on model improvement and adjustments |
| 10 | Construct regression equation and comment on meaning of all components |
| 11 | Determine whether five property of linear regression using OLS are fulfilled |
| 12 | Conclusion with findings and results |

**Executive summary:**

The analysis work is done for the assignment of the course “A220A0010 Free Analytics Environment R”. On a given data set named “housing_new.csv” the analysis is done. In this task the data set is loaded in the R environment and it was studied based on the topics covered in the lecture. After loading and understanding the data set different methods are used to analyze the data based on the requirement of the assignment. After analyzing the data set proper comment and feedbacks are given. All the work of this task is done individually based on the class lectures and exercises.

**Introduction:**

In this assignment a data set is given. The main task is to load the data and analyze it using the R environment using different methodologies, visualization and comments. The given Data set, “housing_new.csv” is a .CSV file and can be loaded and organized using the R environment. The application used here is Rstudio. In this task the job is to work as an analyst at a housing company. The “housing_new” data set contains information about different houses in a specific city. The job is to make sense of the data set from different perspective and construct a regression model that explains the value of the houses. After receiving the data set the from the company the work of analyst starts.

**Goal:**

The goal of this task is to help the company with predicting the house price in thousands of dollars using the given features or variables. Commenting on different perspective and valuable recommendations.

**Load and study the data set, comment on data set's structure:**

To load and study the data set firstly we will have to remove the global environment from Rstudio using the below command:

rm(list = ls())

Libraries used: To analyze the given data set some library packages we need to install and load in our code. Those are:

library(datasets)

library(ggplot2)

library(dplyr)

library(corrplot)

library(tsoutliers)

Loading data set:

Now for loading the dataset we can use csv.read() function.

#read the csv file

housing=read.csv("housing_new.csv",header =TRUE,sep = ";")

Here, header="TRUE" and sep=";"(separation) is used to get the variables in their rows and columns.

Study the data set:

To study the data set we can get the structure of the data set and from there we can have understanding of the data.

#using str to observe the data

str(housing)

**Missing value:** There is no missing value

Now if we run this command we can see the below results:

```
str(housing)
'data.frame':   506 obs. of  12 variables:
 $ PCCR : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ PRLZ : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ INDUS: num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ NOX  : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ..
.
 $ AVR  : num  6.58 6.42 7.18 7 7.15 ...
 $ AGE  : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ DIS  : num  4.09 4.97 4.97 6.06 6.06 ...
 $ RAD  : int  1 2 2 3 3 3 5 5 5 5 ...
 $ TAX  : int  296 242 242 222 222 222 311 311 311 311 ...
 $ TAX2 : int  77 54 53 60 63 52 67 77 76 73 ...
 $ SUB  : num  0.06 0.41 0.08 0.45 0.27 0.37 0.48 0.02 0.29 0.35 ...
 $ MEDV : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

This provides the data types and name of the columns in the data set which will be useful for future analysis.

Here from the above results we observe:

- There are total 506 objects with 12 variables in this data set
- Here we can see in the variables there are numerical and integer variables
- ‘RAD’, ‘TAX’,’TAX2’ are integer variables
- ‘PCCR’, ‘PRLZ’, ‘INDUS’,’NOX’,’AVR’,’AGE’,’DIS’,’SUB’,’MEDV’ are numerical variables
- Here ‘MEDV’ is the dependent variables and other 11 variables are explanatory variables

**Conduct explanatory data analysis of Five explanatory variables of choice:**

For conducting the explanatory data analysis of five variables, I chose the following five variables:

- PCCR
- PRLZ
- INDUS
- NOX
- AVR

Also dependent variable MEDV analysis is also done.

Summary statistics of the 5 variables we slected:

```
summary(housing$MEDV)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   17.02   21.20   22.53   25.00   50.00
> summary(housing$PCCR)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
> summary(housing$INDUS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.46    5.19    9.69   11.14   18.10   27.74
> summary(housing$NOX)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3850  0.4490  0.5380  0.5547  0.6240  0.8710
> summary(housing$PRLZ)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00   11.36   12.50  100.00
> summary(housing$AVR)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.561   5.886   6.208   6.285   6.623   8.780
```

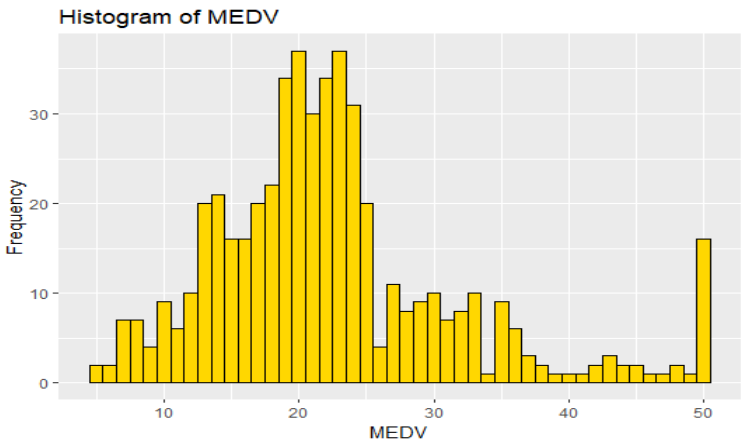
Histogram: Histogram is used using the library ggplot2 and function ggplot() for analysis of five explanatory variables and dependent variable.

Histogram of dependent variable MEDV:

#Histogram of dpendent variable data

```
ggplot(housing, aes(x = MEDV)) +
  geom_histogram(binwidth = 1, fill = "gold", color = "red") +
  labs(x = "MEDV", y = "Frequency") +
  ggtitle("Histogram of dependent variable MEDV")
```

Result:



Here in the X-axis: MEDV and y axis: Frequency of data. It represents the value of houses. From the median value we can see some right skewed distribution.

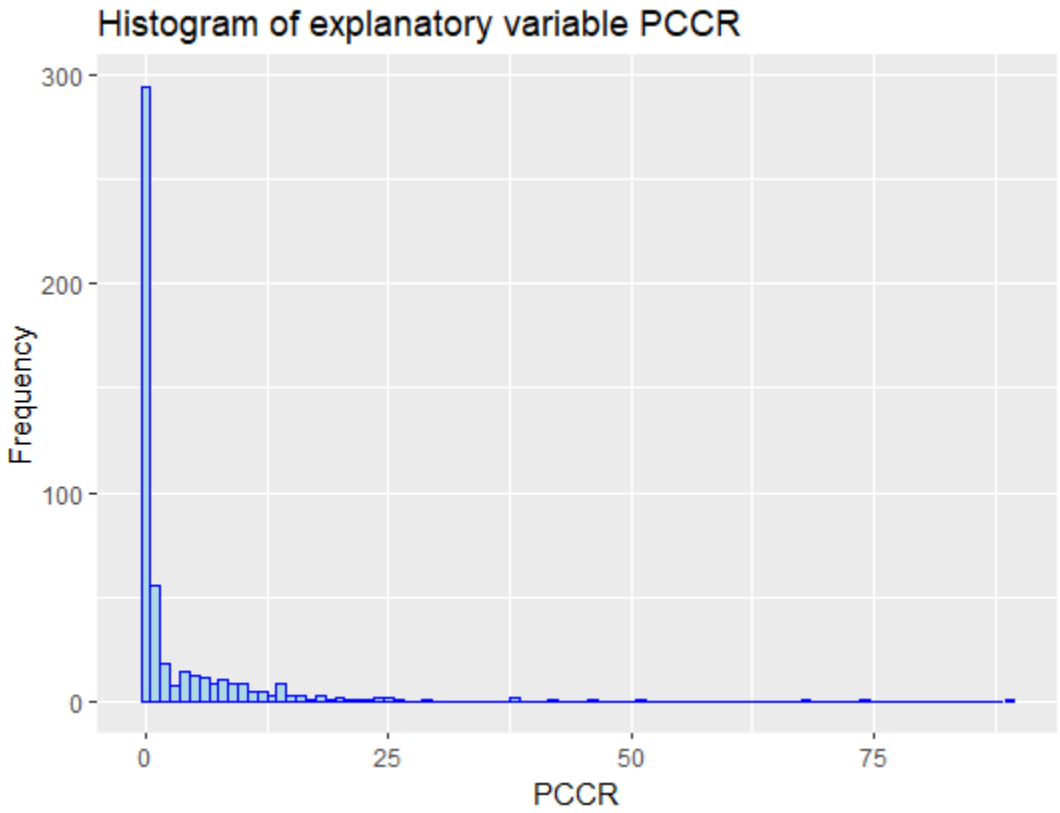
Comment: We can see the histogram where it varies from different value of x. for the value 20 and 23 we can see maximum readings in histogram.

Histogram of explanatory variables:

PCCR:

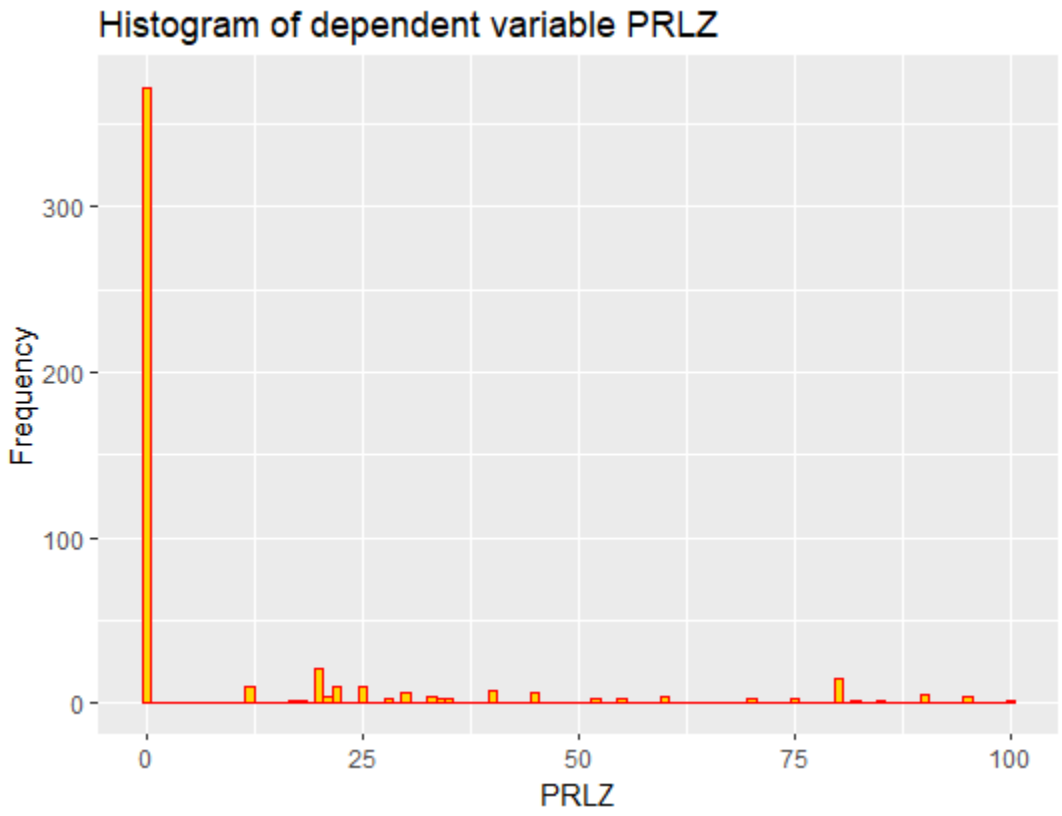
```
ggplot(housing, aes(x = PCCR)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "blue") +
  labs(x = "PCCR", y = "Frequency") +
  ggtitle("Histogram of explanatory variable PCCR")
```

Result:



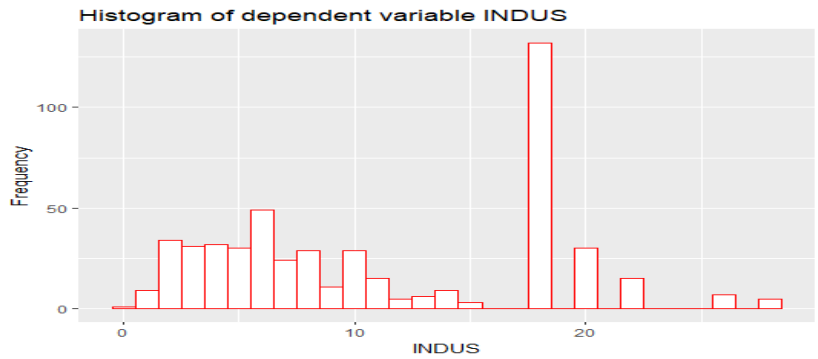
Comment: Here we can see very minimal values in the Histogram where the maximum reading of y is at x=0. We can see majority are has low or minimum crime rates. However some area has high crime rates.

PRLZ:



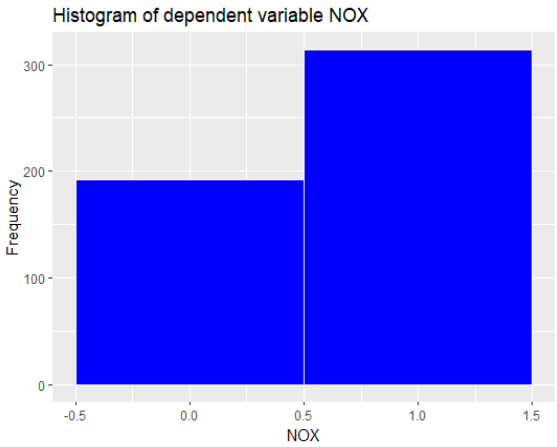

Comment: y value is maximum for x=0 . The proportion of land zoned is low in most of the distributions.

INDUS:



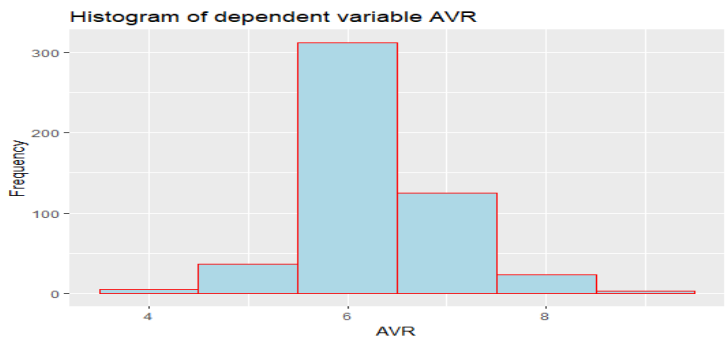

Comment: Histogram of INDUS shows non retails business areas where the distribution is symmetrical apart from a high range value and some minimum values.

NOX:




Histogram of explanatory variable NOX shows the air pollution level. It shows high values with right skewed distribution.

AVR:




Histogram of explanatory variable AVR. Average room dwelling. We can see different levels of AVR.

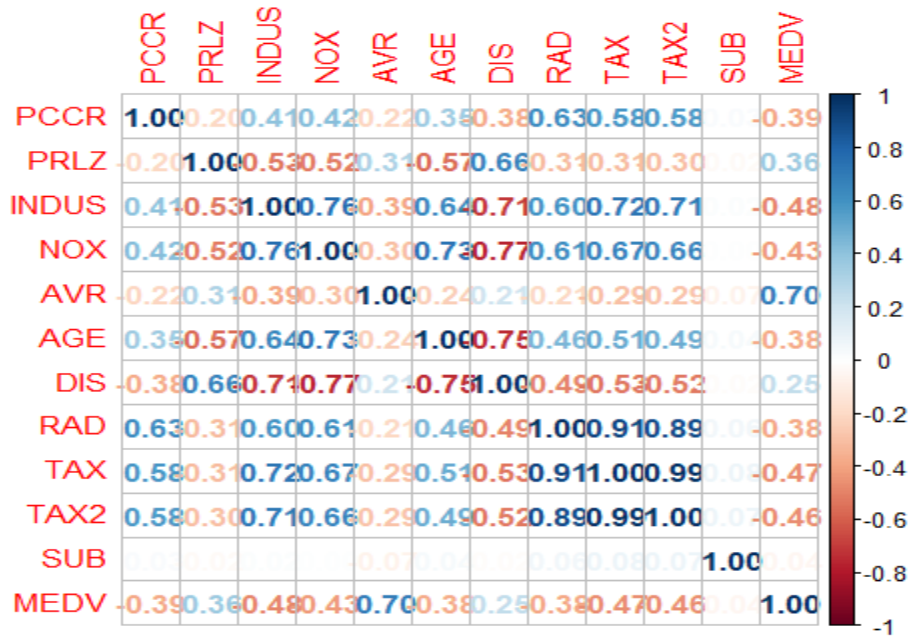
**Perform correlation analysis between all variables:**

If we perform correlation analysis between all the variables and visualize :

#correlation between all variables

#plot correlation

corrplot(cor(housing),"number")



Comment:

The variables that have the highest (absolute) linear associations with the house values (MEDV)

- MEDV has highest absolute correlation with AVR is about 0.70. So we can say that Average Number of Rooms per Dwelling has significant correlation with the house price MEDV.
- Other variables correlation varies from -0.39 to 1
- PCCR has negative correlation with MEDV with -0.39. So par capita crime rate is negatively correlated with MEDV.
- PRLZ has positive correlation with MEDV with 0.36. So proportion of residence land zone is positively correlated with house price.
- NOX , INDUS, AGE, RAD,TAX, TAX2 both have negative correlation with MEDV. So air pollution and Non retail business areas are negatively correlated.
- We know MEDV has highest absolute correlation with itself 1.

**Findings:**

- All the diagonal values are 1
- TAX2 has the highest correlation with TAX 0.99
- PCCR has its highest correlation with both TAX and TAX2 at 0.58
- PRLZ has its highest correlation with DIS 0.66
- INDUS has its highest correlation with NOX at 0.76
- NOX has its highest correlation with DIS
- AVR has negative -0.22 correlation with PCCR
- Most of the explanatory variables have positive correlation with each other except some. So we can come to a conclusion that the explanatory variables are correlated positively with each other except few cases.

**Visualize on which pair of variables have the highest absolute correlation:**

For demonstrating the explanatory variable and their correlations are given without the dependent variable MEDV:

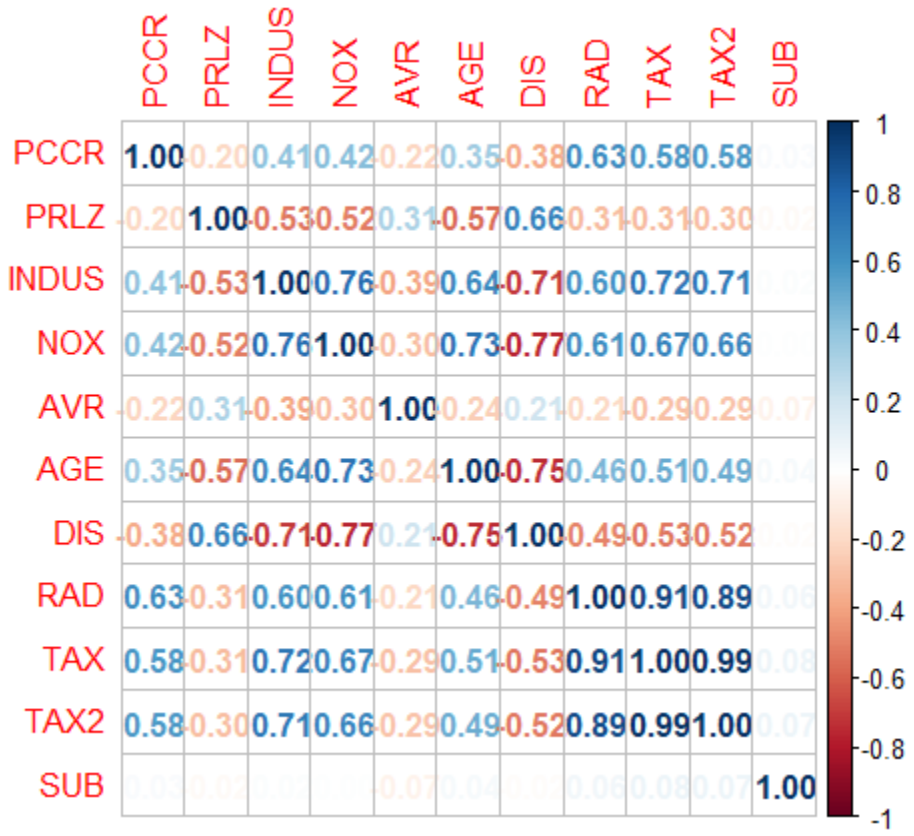
#Now corelation with linear associaciation without MEDV

depvar=housing$MEDV

exvar=housing[,1:11]

corrplot(cor(exvar),"number")

Result:




Correlation of all the explanatory variables without the dependent variable.

The highest absolute pair of variables are correlated TAX with TAX2 = 0.99

So TAX: Full-value property tax rate per $10,000 – TAX2: Additional property tax rate per $10,000 are highly correlated.

RAD is also highly correlated with TAX. 0.91

RAD is also highly correlated with TAX2 0.89

NOX and DIS correlated with 0.77

DIS is negatively correlated with NOX -0.77

**Remove explanatory variables with highest correlation value and comment on the reasons:**

Remove explanatory variables that have high absolute correlation (0.8) from the data set of explanatory variables.

We use a while loop for this:

#Remove explanatory variable with highest absolute value

#remove absolute or values higher than 0.8

```
cormat= abs(cor(exvar))
diag(cormat)=0

#while loop
while (max(cormat)>=0.8) {
  #find explanatory variables with highest corelations
  maxvar= which(cormat==max(cormat),arr.ind = TRUE)

  #select variable with highest average corelation
  maxavg=which.max(rowMeans(cormat[maxvar[,1],]))
  #removal
  exvar=exvar[,-maxvar[maxavg,1]]
```

```
  cormat=cormat[-maxvar[maxavg,1],-maxvar[maxavg,1]]
}
```

**Findings:**

Removing the correlation which is highest absolute value more than 0.80:

Number of explanatory variable removed: 2

Varibales removed: TAX, TAX2

Variables remained in the data set: MEDEV~PCCR+ PRLZ + INDUS + NOX +AVR + AGE +DIS +RAD

| | MEDEV | PCCR | PRLZ | INDUS | NOX | AVR | AGE | DIS | RAD | SUB |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 24.0 | 0.00632 | 18.0 | 2.31 | 0.5380 | 6.575 | 65.2 | 4.0900 | 1 | 0.06 |
| 2 | 21.6 | 0.02731 | 0.0 | 7.07 | 0.4690 | 6.421 | 78.9 | 4.9671 | 2 | 0.41 |
| 3 | 34.7 | 0.02729 | 0.0 | 7.07 | 0.4690 | 7.185 | 61.1 | 4.9671 | 2 | 0.08 |
| 4 | 33.4 | 0.03237 | 0.0 | 2.18 | 0.4580 | 6.998 | 45.8 | 6.0622 | 3 | 0.45 |
| 5 | 36.2 | 0.06905 | 0.0 | 2.18 | 0.4580 | 7.147 | 54.2 | 6.0622 | 3 | 0.27 |
| 6 | 28.7 | 0.02985 | 0.0 | 2.18 | 0.4580 | 6.430 | 58.7 | 6.0622 | 3 | 0.37 |
| 7 | 22.9 | 0.08829 | 12.5 | 7.87 | 0.5240 | 6.012 | 66.6 | 5.5605 | 5 | 0.48 |
| 8 | 27.1 | 0.14455 | 12.5 | 7.87 | 0.5240 | 6.172 | 96.1 | 5.9505 | 5 | 0.02 |
| 9 | 16.5 | 0.21124 | 12.5 | 7.87 | 0.5240 | 5.631 | 100.0 | 6.0821 | 5 | 0.29 |
| 10 | 18.9 | 0.17004 | 12.5 | 7.87 | 0.5240 | 6.004 | 85.9 | 6.5921 | 5 | 0.35 |

TAX and TAX2 have been removed for high absolute correlation 0.99. Thus it is removed from the data set. The new data set is attached above.

**IT is important to remove highly correlated explanatory variables before fitting a linear regression model** because it makes the model challenging to determine individual effects.

- It may generate standard error due to high absolute correlation
- Model interpretation can be unreliable
- Overfitting may happen due to correlation absolute value more than 0.8
- It may occur less of generalization and also may reduce model transparency.
- Extreme values are excluded in linear regression.

**Implement a first linear regression after removing explanatory variables in the previous task:**

After removing the explanatory variables in the previous task let's make a regression model using the variables:

#Make our model

my_data=cbind('MEDEV'=depvar,exvar)

linearregmodel=lm(MEDEV~PCCR+ PRLZ + INDUS + NOX +AVR + AGE +DIS +RAD, data =my_data)

Hence the model is constructed, now summarize the model:

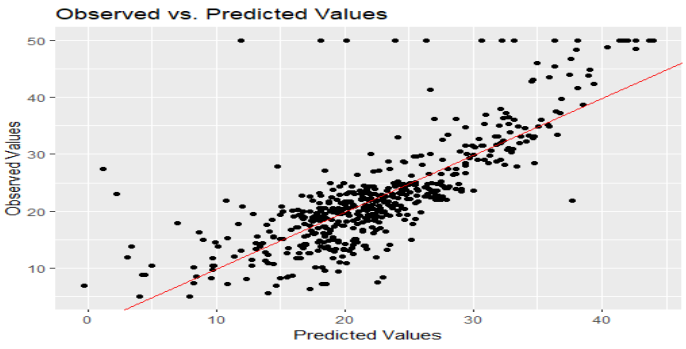
```
summary(linearregmodel)
```

```
Call:
lm(formula = MEDEV ~ PCCR + PRLZ + INDUS + NOX + AVR + AGE +
    DIS + RAD, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-15.808  -3.065  -0.666   2.062  38.058

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.31945    4.09090  -0.323 0.747184
PCCR         -0.18960    0.03877  -4.891 1.36e-06 ***
PRLZ          0.06209    0.01541   4.030 6.45e-05 ***
INDUS        -0.20164    0.06535  -3.085 0.002145 **
NOX         -12.32881    4.31920  -2.854 0.004492 **
AVR           6.99462    0.41206  16.975  < 2e-16 ***
AGE          -0.05495    0.01493  -3.681 0.000258 ***
DIS          -1.78329    0.23963  -7.442 4.40e-13 ***
RAD          -0.05302    0.04445  -1.193 0.233511
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.739 on 497 degrees of freedom
Multiple R-squared:  0.6168,   Adjusted R-squared:  0.6107
F-statistic:   100 on 8 and 497 DF,  p-value: < 2.2e-16
```

Now if u see the observations we find the below observations:

- We can see the estimated values which are almost zero for PRLZ, for AVR the estimated value is around 6.99 which indicates for one unit change in AVR causes 6.99 increase in MEDV
- Other estimate values are negative this shows decrease in MEDV for changes in the variables
- P value determines whether the coefficients are other than zero. Smaller the p-value it is said that better the model.
- A p value smaller than 0.5 or less or almost zero indicates strong evidence that the model does not accept the null hypothesis.
- Here the significant variables p-values are indicated with(*) symbol. More significant variables have more (*) symbols.
- PRLZ P-value =1.36e-06 which is almost zero which indicates that p value is significant and it rejects the null hypothesis.
- For AVR the P-value is extreme < 2e-16 which shows that the model as a whole is statistically significant
- F-statistic: 100 on 8 and 497 DF defines the significance of the model
- R^2 value to the adjusted R^2 value is used for adjusting the overfitting

Let's see the predicted values and the original values in a regression model plotting:



We can see the predicted values are good fit except some points which are not good fit.

Now if we calculate the residual means we get:

```
mean(residuals(linearregmodel))
[1] 2.496905e-16
```

So here we can see the residual mean is very small and almost close to 0 which is 2.496905e-16

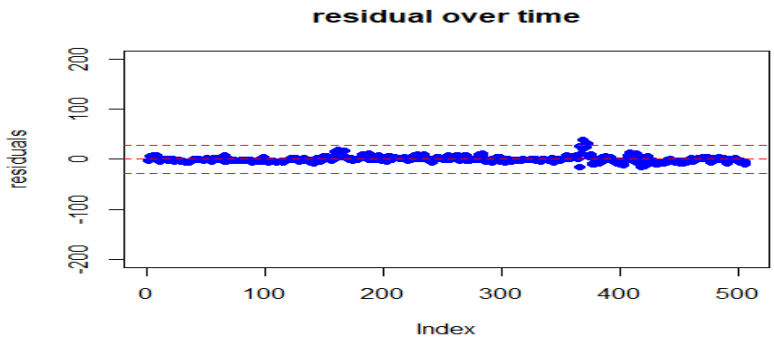
- So it says the model predicted values are well observed and almost close to the original or observed values.
- So it describes a well fitted model

**Comment on model improvement and adjustments:**

**Now let's plot the residuals for Homoskedasticity:**

```
#Homoskedasticity & residual linear independent

plot(residuals(linearregmodel), type="p", col="blue",ylim =c(-200,200),pch=16,

    ylab ="residuals",main = "residual over time")

abline(a=5*sd(residuals(linearregmodel)), b=0,col="red", lty=2)

abline(a=-5*sd(residuals(linearregmodel)), b=0,col="red", lty=2)

abline(a=0,b=0,col="red",lty=2)
```

Plot:




So from the Homoskedasticity we can see the model is a good fit and very well fitted.

If we calculate the residual correlation with five variables we get:

```
              INDUS          AVR           PRLZ          NOX          PCCR
[1,] 8.675542e-17 2.790478e-17 -9.343405e-18 4.351961e-17 -3.51299e-17
```

So we can say that this model can yet be improved:

- Feature selection and assess the significance of the variables
- High p-value variables can be removed and remodel
- Additional data or collecting and inserting new data can be useful for this model
- Residual analysis can be done. The residual analysis uses residuals to check if there is any pattern or heteroscedasticity. If there is any then it may be omitted.

**Construct regression equation and comment on meaning of all components:**

Regression equation: MEDEV = -1.31945 - 0.18960 * PCCR + 0.06209 * PRLZ - 0.20164 * INDUS - 12.32881 * NOX + 6.99462 * AVR - 0.05495 * AGE - 1.78329 * DIS - 0.05302 * RAD

Comments:

Here,

Intercept: -1.31945 ;this is the intercept which represents the estimated value of the dependent variable MEDV

PCCR: the coefficient of PCCR is - 0.18960. which indicated the MEDV decreases by 0.18960 for 1 unit of PCCR

PRLZ: The coefficient is 0.06209 which indicates that the MEDV increases with PRLZ units

INDUS: Indus also has a negative coefficient

NOX: it also has negative coefficient of -12.32

AVR: This coefficient is positive and for 1 unit of AVR ,the value of MEDV increases by 6.9942 times

AGE: the coefficient is negative and it causes decrease in MEDV value.

DIS: Negative coefficient

RAD: Negative coefficient.

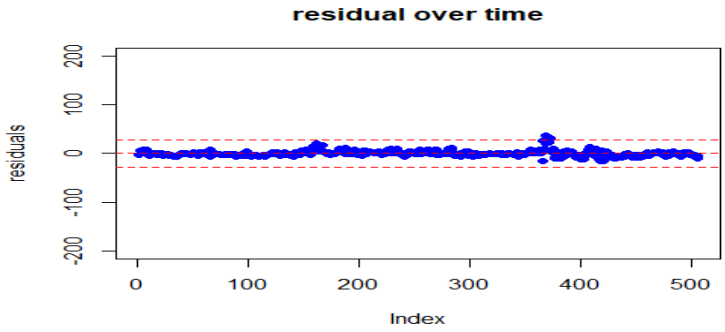
Now, if we exclude the variable RAD as it's higher p value and not significant then we get

MODEL 2: Second regression

```
summary(linearregmodel1)

Call:
lm(formula = MEDEV ~ PCCR + PRLZ + INDUS + NOX + AVR + AGE +
    DIS, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-16.168  -3.061  -0.560   2.102  37.432

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.54702    4.04104  -0.135 0.892378
PCCR         -0.21288    0.03351  -6.352 4.80e-10 ***
PRLZ          0.06215    0.01541   4.032 6.39e-05 ***
INDUS        -0.21995    0.06355  -3.461 0.000585 ***
NOX         -13.61827    4.18350  -3.255 0.001210 **
AVR           6.95125    0.41063  16.928  < 2e-16 ***
AGE          -0.05421    0.01492  -3.633 0.000310 ***
DIS          -1.79775    0.23943  -7.509 2.78e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression equation for the second model: MEDEV = -0.54702 - 0.21288 * PCCR + 0.06215 * PRLZ - 0.21995 * INDUS - 13.61827 * NOX + 6.95125 * AVR - 0.05421 * AGE - 1.79775 * DIS

This is how we can exclude the insignificant variables and make the model better.




**Determine whether five property of linear regression using OLS are fulfilled:**

- **Lineraity:** As the regression model is linear and it does not violate any regulation of linearity we can say that the linearity of the linear regression model is true
- **Errors are independent:** We have already shown that there is no dependence and significant correlation between the errors. Errors are independent of variables.
- **Homoskedasticity:** We have shown homoskedasticity in this regression model and we have shown that the variances of the residuals are constants. We can see residual standard error is provided as 5.739 and it's constant.
- **Normality of residuals:** Residuals are normally distributed and the mean is zero
- **Independent variables are not highly correlated:** Highly absolute correlated variables are excluded from the model so we can state that the independent explanatory variables are not highly correlated.

So we can say that the five property of linear regression using OLS are fulsilled.

**Conclusion with findings and results:**

The linear regression model was constructed to predict the median home value (MEDV) based on several independent explanatory variables, including PCCR, PRLZ, INDUS, NOX, AVR, AGE, DIS, and RAD. There were two more variables which were excluded due to their high(>0.8) correlation.

- PCCR, PRLZ, INDUS, NOX, AVR, AGE,, DIS have statistically significant coefficients which indicates that changes in these variables are linked with the changes in MEDV.
- PCCR, PRLZ, AVR have positive relationships and coefficients with MEDV, while INDUS, NOX, AGE, and DIS have negative relationships and coefficients.
- RAD does not have significant coefficient and p value is 0.23511 , so RAD may not be meaningful predictor for this regression model.
- The model is a good fit in terms of the observed values
- Residual mean is 2.496905e-16 which states significance of the model
- P value `p-value: < 2.2e-16 which states that this model is significant and does not accept the null hypothesis.`
- This model has shown homoskedasticity and thus this can be a good fit model

So we can conclude that, The linear regression model that is constructed is a successful analysis of the data “housing_new”. This data set was loaded successfully and all the analysis is done with success. It can be called a good fit model.

## PART-2

## CLUSTERING

| Serial no | TOPICS |
|---|---|
| 1 | Load and study the dataset. |
| 2 | Conduct exploratory data analysis for all variables |
| 3 | Conduct a correlation analysis of the variables. |
| 4 | Create a new data frame that contains the original data normalized |
| 5 | Using the k-means algorithm together with four different methods: Elbow method, Silhouette method, Calinski-Harabasz Index and Gap statistic method |
| 6 | Use now the optimal number of clusters determined in the previous step to run the k-means algorithm (with nstart at least 25). |
| 7 | Give a detailed conclusion on findings and results. |

**Introduction:**

A data set of wholesale_2023 is given for analysis. As an analyst the job is to analyze the data set and its variables, also group the data sets in different clusters and then plot them. Also to determine the number of optimal clusters. And four clustering methods: Elbow method, Silhouette method, Calinski-Harabasz Index and Gap statistic method. The Goal is to determine the number of clusters and do k-means and normalization. Also optimal number of clusters to be found out.

**Load and study the dataset:**

Let's load the data set "wholesale_2023.csv"

# Read the CSV file with the correct delimiter (e.g., assuming a comma as the delimiter)

data=read.csv("wholesale_2023.csv", header = TRUE, sep = ";")

# Check the structure and summary of the dataset

str(data)

Result:

```
'data.frame':  440 obs. of  9 variables:
 $ Channel         : int  2 2 2 3 2 2 2 2 1 2 ...
 $ Region          : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Fresh           : int  12669 7057 6353 13265 22615 9413 12126 7579 5963 60
06 ...
 $ Milk            : int  214 1762 2405 6404 3915 666 480 1669 425 1159 ...
 $ Grocery         : int  7561 9568 7684 4221 7198 5126 6975 9426 6192 18881
...
 $ Frozen          : int  9656 9810 8808 1196 5410 8259 3199 4956 3648 11093
...
 $ Detergents_Paper: int  1337 1647 1758 254 889 898 1570 1661 858 3713 ...
 $ Delicassen      : int  1338 1776 7844 1788 5185 1451 545 2566 750 2098 ...
 $ Beverages       : int  98 855 1181 2915 1812 303 235 816 198 567 ...
```

- Here all the variables are integer type

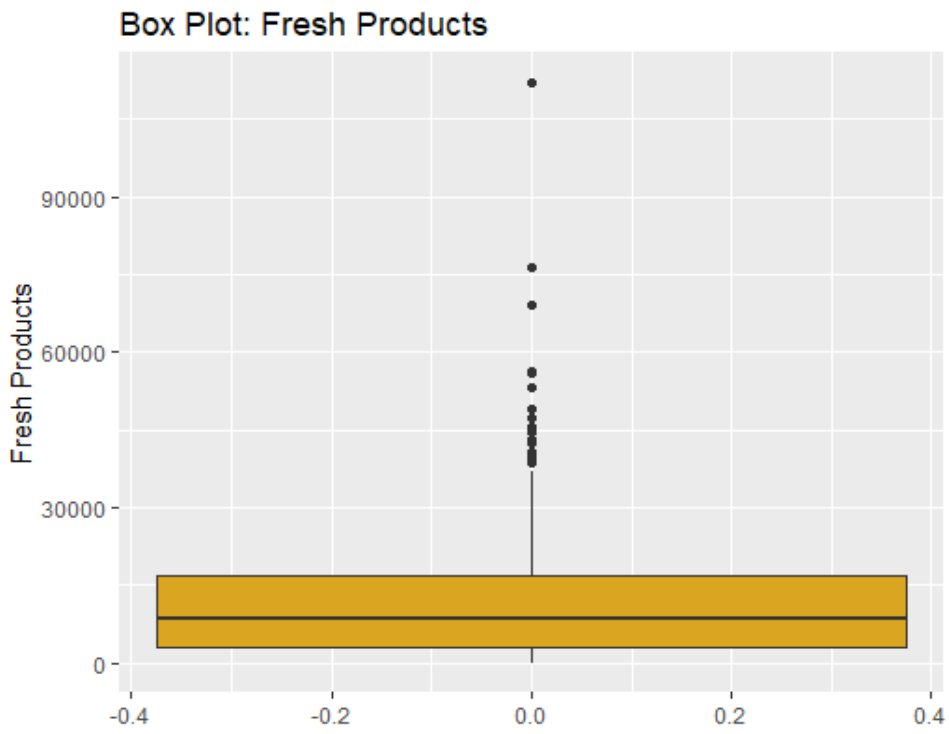
- There are 9 variables

  **Check whether there is any missing value:**

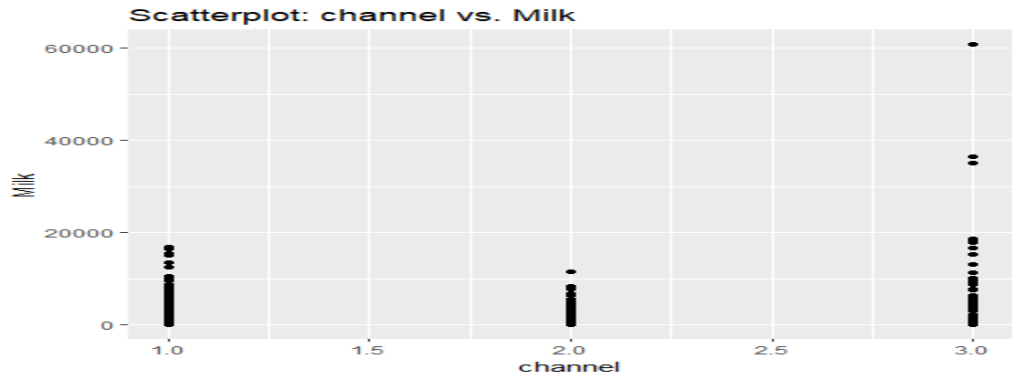
  There is no missing value. Data is not needed to be scaled.

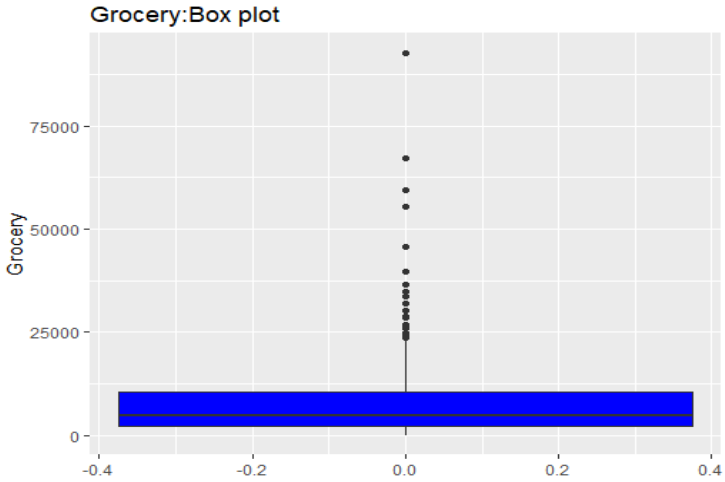
## Conduct exploratory data analysis for all variables

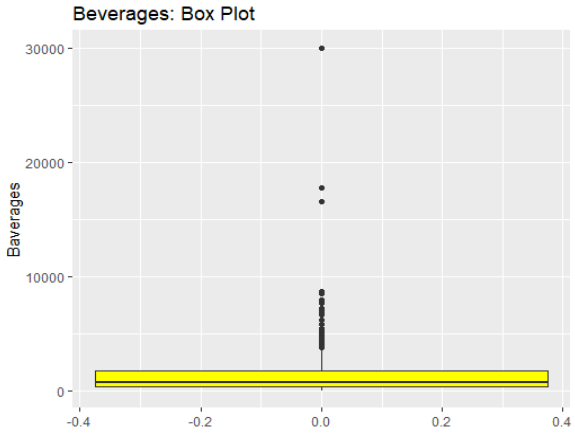
Box plot of fresh products:

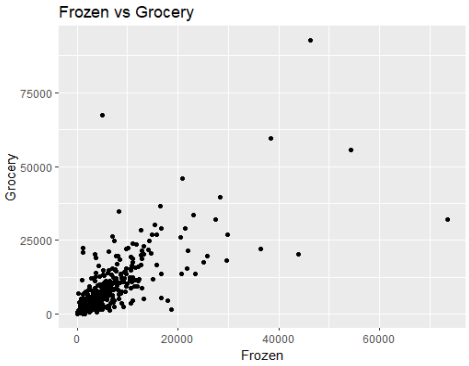



Plot Channel vs milk:


Scatterplot: channel vs. Milk
60000
40000
20000
0
Milk
1.0
1.5
2.0
2.5
3.0
channel


Grocery:Box plot
75000
50000
25000
0
Grocery
-0.4
-0.2
0.0
0.2
0.4


Beverages: Box Plot
30000
20000
10000
0
Baverages
-0.4
-0.2
0.0
0.2
0.4



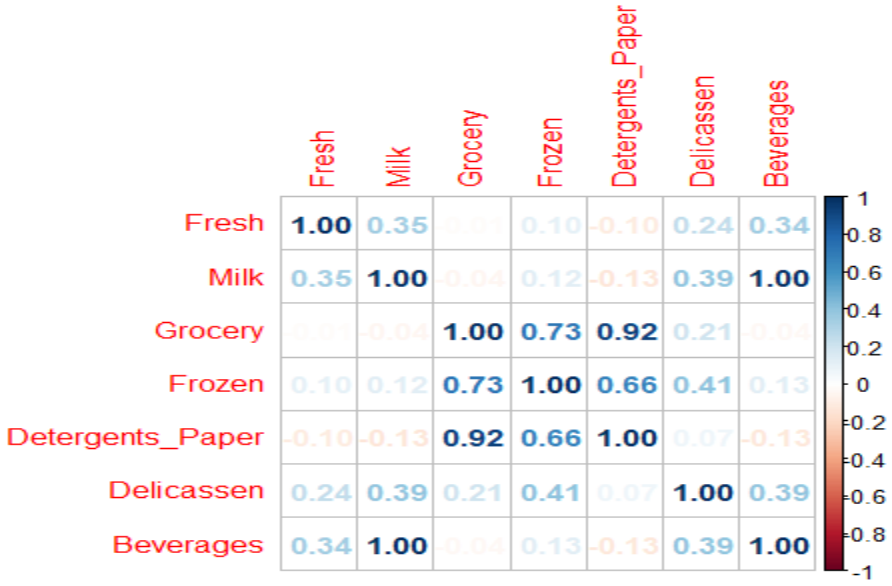

From the above plotting analysis we can have observations on the variables.

- Most of the beverages are below 10000 frequency
- Grocery shows medium except one extreme value
- Milk vs channel shows in three different groups channel 1,2 ,3
- Frozen vs Grocery in plotting shows dense scatter in low axes points

**Here the variable analysis states different plotting techniques on different variables. We can identify the values in boxplot or scatter plot.**

**Conduct a correlation analysis of the variables.**

Correlation of the variables:



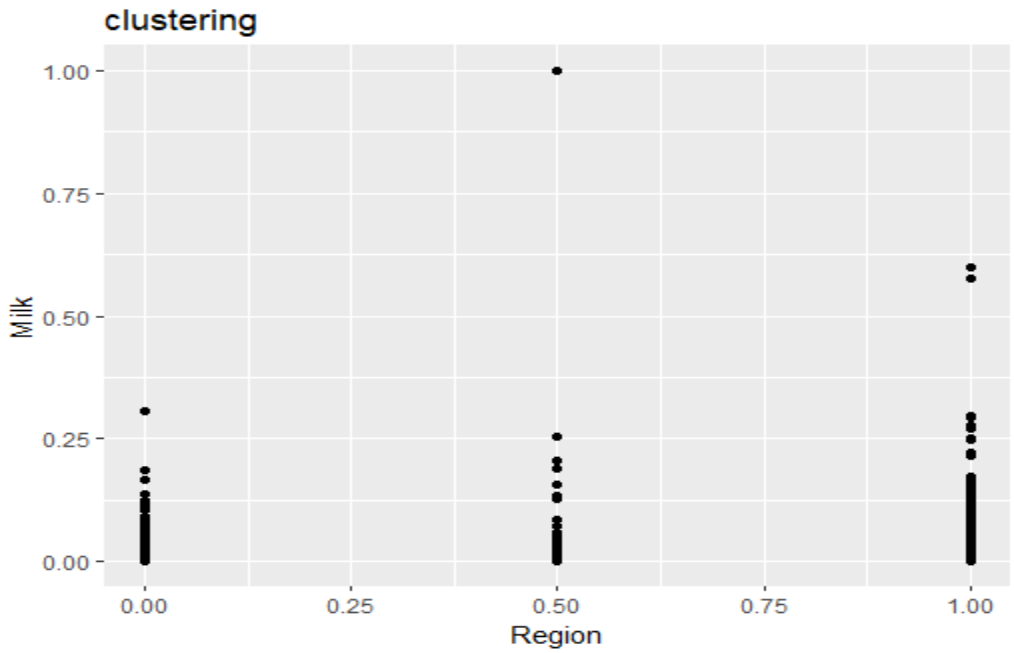

Here we can see the correlations . We can see the highest correlation between two variables are,Milk and Beverages =1 and Grocery and Detergents_Paper=0.92.

Diagonally the correlation values are 1 and there are more positive correlations among the variables. Although there are some negative correlations like Detergent_paper has -0.10 correlation value with Fresh.

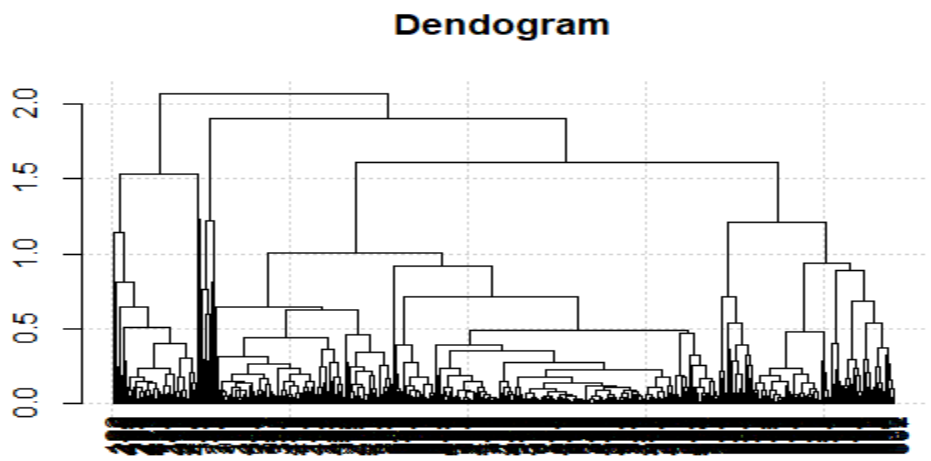
**Create a new data frame that contains the original data normalized**

```
#scaling with min max Normalization

data= apply(whole_sale,2,rescale, to=c(0,1))

#plot data
ggplot(data.frame(data),aes(x=Region,y=Milk))+
  geom_point()+
  ggtitle("clustering")
```

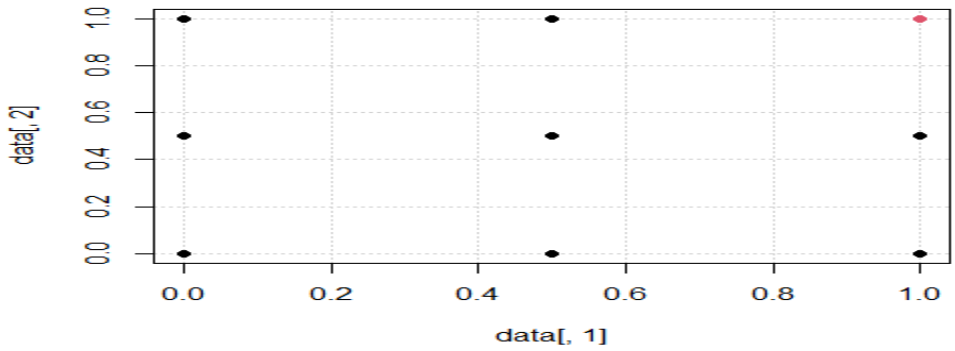
Result:




Dendogram of the data:




Plot of the membership:




K-means clusters:

#kmeans clustering

```
kmeansmod= kmeans(data, centers=2, nstart = 25)


kmeansmod$cluster
kmeansmod$centers
plot(data[,1],data[,2],col=kmeansmod$cluster, pch=16, main = "kmeans cluster")
kmeansmod$size
kmeansmod$withinss
```

kmeansmod$tot.withinss

Results:

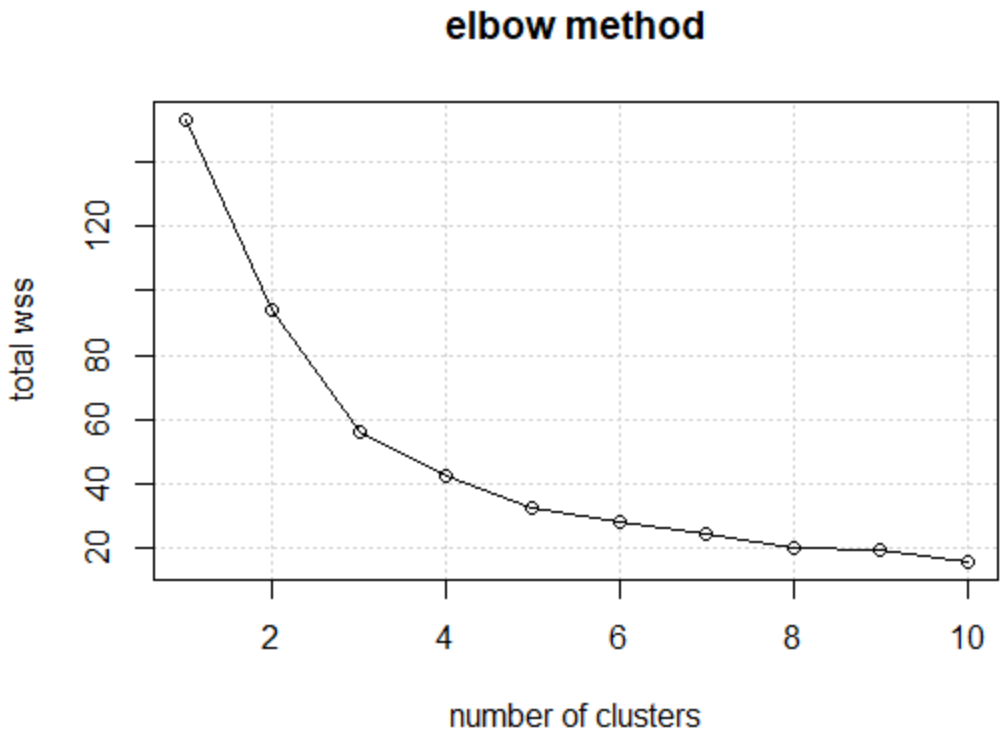
```
Channel    Region      Fresh       Milk    Grocery     Frozen Detergents_Pape
r
1 0.3104839 0.1895161 0.09486183 0.05541114 0.08717913 0.07189886       0.074
47430
2 0.3212025 1.0000000 0.11173156 0.04798493 0.08507889 0.08063512       0.068
93655
  Delicassen  Beverages
1 0.02665647 0.05361922
2 0.03374220 0.04654072
>     plot(data[,1],data[,2],col=kmeansmod$cluster, pch=16, main = "kmeans cl
uster")
>     kmeansmod$size
[1] 124 316
>     kmeansmod$withinss
[1] 32.22630 62.06186
>     kmeansmod$tot.withinss
[1] 94.28816
```

**Using the k-means algorithm together with four different methods: Elbow method, Silhouette method, Calinski-Harabasz Index and Gap statistic method**

**Elbow method:**

```
#Elbow method
tot_within_ss= map_dbl(1:10, function(k){
 model=kmeans(data,centers = k,nstart = 25)
 model$tot.withinss

})

plot(1:10,tot_within_ss,type="o", xlab = "number of clusters",
   ylab = "total wss", main="elbow method",panel.first = grid())
```

Plot:




**Findings:**

- From the plot we can see normal decreasing after point 3
- At point 3 we can see sudden downfall. After point 3 the downfall is observed normal
- We can see the elbow point at 3
- After point 3 we can see monotonous progression

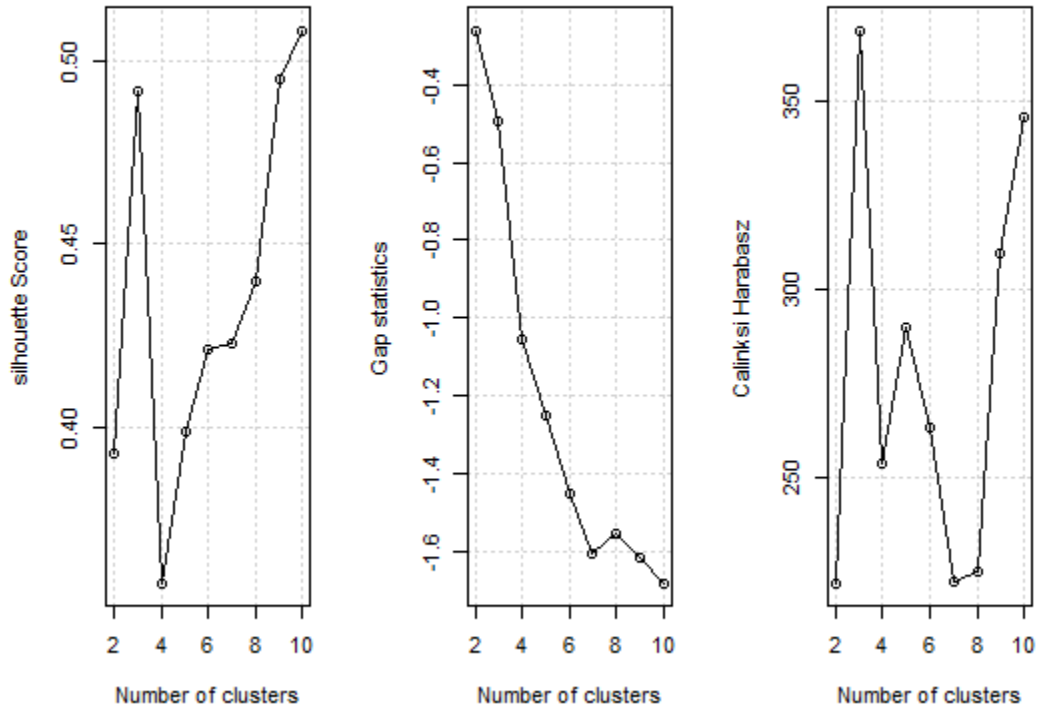
**Thus by elbow method the number of clusters should be 3**

**Silhouette method, Calinski-Harabasz Index and Gap statistic method**

```
#silclustering method

silclust=NbClust(data,distance = "euclidean",min.nc = 2,max.nc = 10,
                 method = "kmeans",index = "silhouette")
```

```
#Gap clustering


Gap_clust= NbClust(data,distance = "euclidean",min.nc = 2,max.nc = 10,
                   method = "kmeans",index = "gap")


#CHclust


CH_clust= NbClust(data,distance = "euclidean",min.nc = 2,max.nc = 10,
                  method = "kmeans",index = "ch")



#plot the methods
par(mfrow=c(1,3))
plot(2:10,silclust$All.index,type = "o", xlab = "Number of clusters",
    ylab = "silhouette Score", panel.first = grid())
plot(2:10,Gap_clust$All.index,type = "o", xlab = "Number of clusters",
    ylab = "Gap statistics", panel.first = grid())


plot(2:10,CH_clust$All.index,type = "o", xlab = "Number of clusters",
    ylab = "Calinksi Harabasz", panel.first = grid())
```

**Plots:**




- In siihouette score we can see the peak at the point 3
- In Gap statistics point 3 is the point of downfall
- In CH method the maximum value point hits at point 3
- After Analyzing the above methods we can say that thre should be 3 clusters

**Use now the optimal number of clusters determined in the previous step to run the k-means algorithm (with nstart at least 25).**

```
summarise_all(list(avg= mean))
# A tibble: 3 x 10
  member Channel_avg Region_avg Fresh_avg Milk_avg Grocery_avg Frozen_avg
  <fct>        <dbl>      <dbl>     <dbl>    <dbl>       <dbl>      <dbl>
1 1             1.90       2.59    36156.    6811.       6367.      6124.
2 2             1.54       2.55     8280.    2564.       5311.      3817.
3 3             2          2.45     8120.    2011.      27745.     18813.
```

If we observe the mean values we can see:

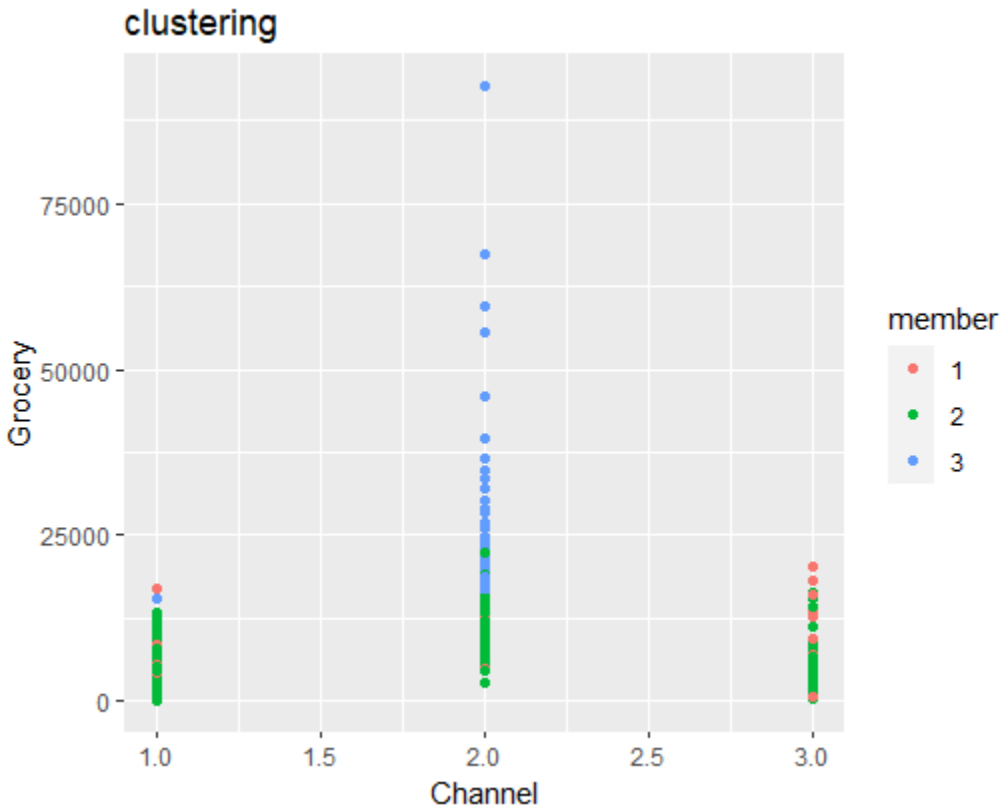
- There are 3 optimal number of clusters
- Fresh, Milk, Grocery, Frozen have large means
- Average/mean values are not very distinct in three member classes

We can see (mean, std) also

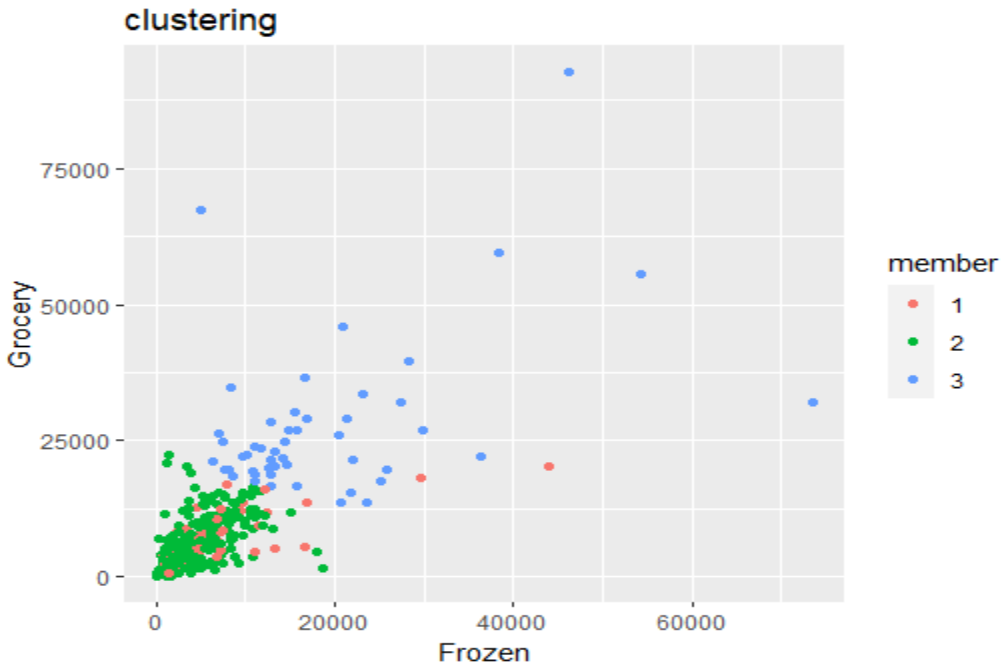
```
+   summarise_all(list(avg= mean, std= sd))
# A tibble: 3 × 19
  member Channel_avg Region_avg Fresh_avg Milk_avg Grocery_avg Frozen_avg
  <fct>        <dbl>      <dbl>     <dbl>    <dbl>       <dbl>      <dbl>
1 1             1.90       2.59    36156.    6811.       6367.      6124.
2 2             1.54       2.55     8280.    2564.       5311.      3817.
3 3             2          2.45     8120.    2011.      27745.     18813.
```

Now if we plot the clusters we can find below plots:

Channel vs Grocery

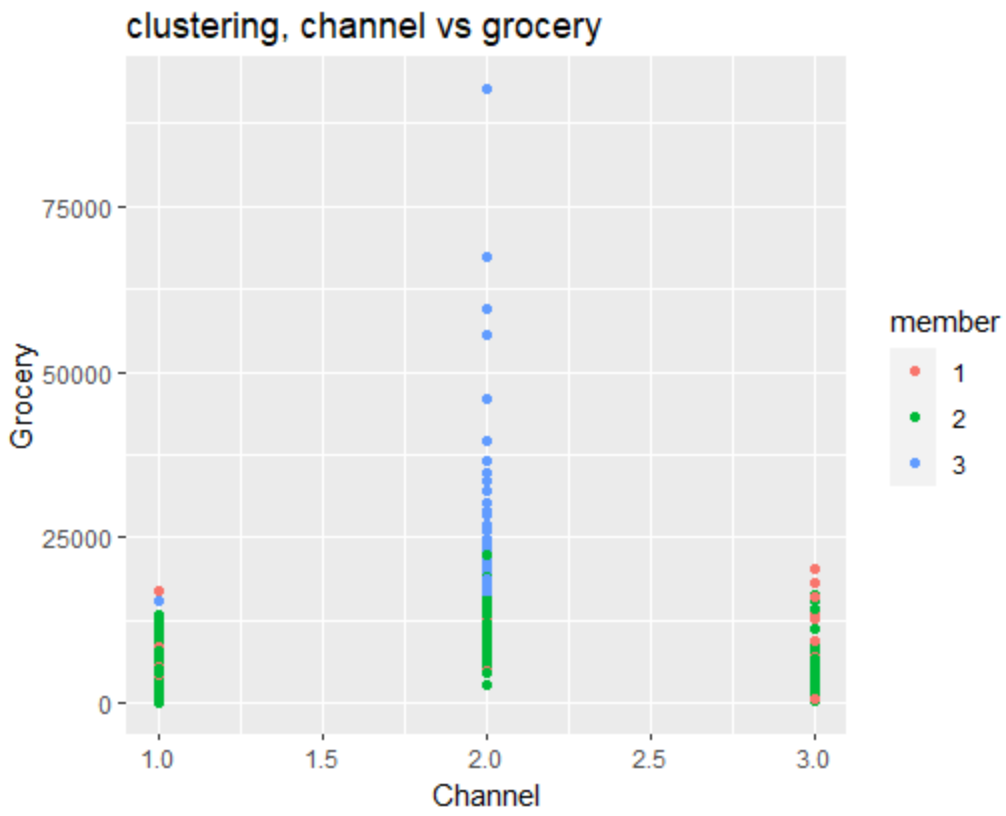


Here we can see the data points are grouped in three different clusters but this is not a good clustering as one cluster member has entered in another cluster member.

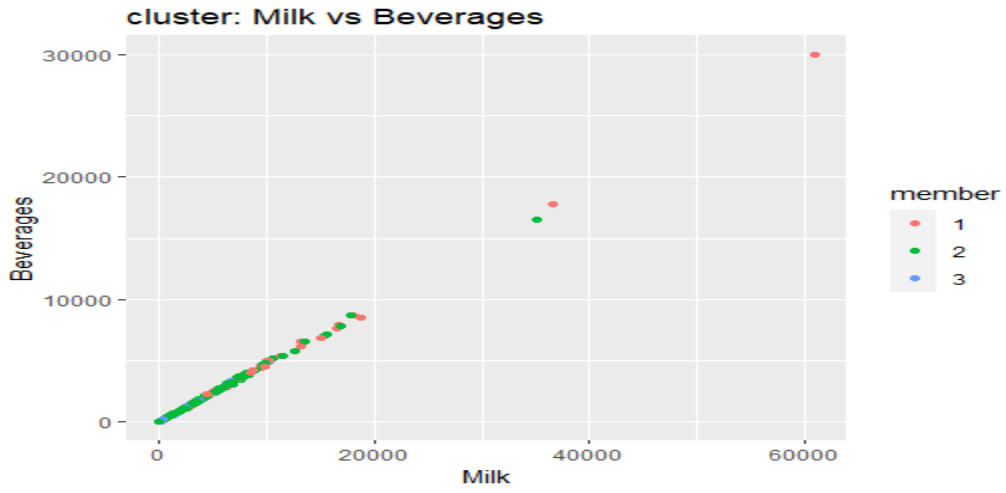



Frozen vs Grocery

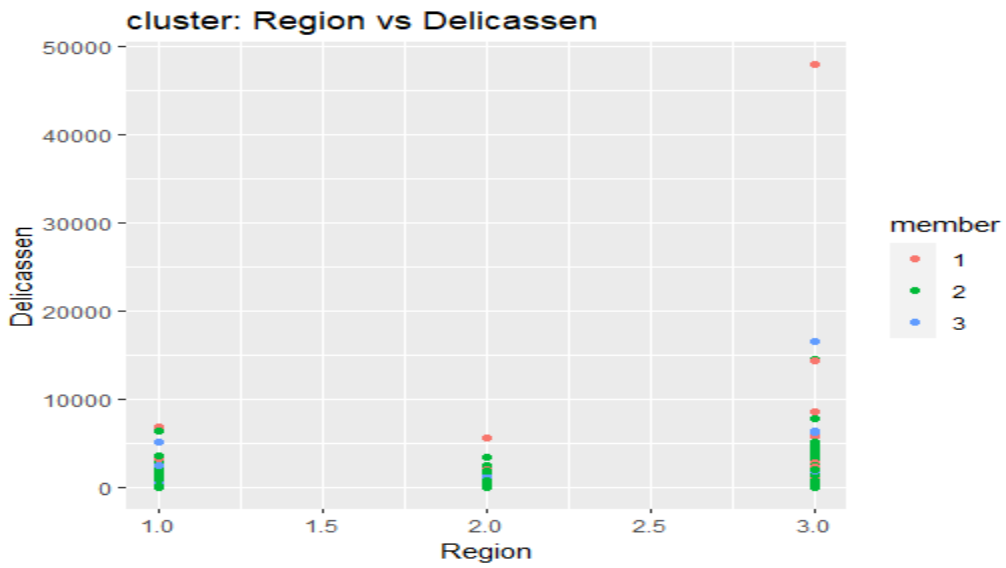
Frozen vs grocery is also not a good clustering example as the data are scattred and combined in different regions. In same region there are dense data points.




Here we can see three different members but the data points are colliding or sometimes combining. Not  a good cluster example




Here the cluster members are colliding.




Findings:

- Here we can see that the number of optimal classes are 3.
- But in the clustering we can see the points of different clusters combining or colliding
- So we can say that clustering is not good for this data set.
- Different data points have positive correlations and they collide often.

**Give a detailed conclusion on findings and results.**

From the given data set we can see the data has integer variables. I've anayze the variables in box plot and scatter plot to analyze the nature of the data. We can see the data set has similar values.

**Missing value and scaling:** The data set doesn't have any missing value. The data is scaled.

**Analysis and visualization:** We can analyze the bloxpot and the scatterplot to analyze. Most of the variables have densely pointed data points.

**Normalization:** Normalization is done using min-max method.

**K-means and four methods**: Using the k-means algorithm together with four different methods: Elbow method, Silhouette method, Calinski-Harabasz Index and Gap statistic method we have analyzed the number of clusters. All the methods support that there are three appropriate clusters.

**K-means algorithm:** K-means algorithm is implemented and different variable and variables and their relationships are plotted and analyzed.

So from the above analysis we can see the distinct relationship between various variable whole sale data and relationship of different variables with Region, channel and one another etc.

The findings can help the wholesale company sales in below ways:

- The boxplot shows sales of different wholesale products and their frequencies.
- The correlation will help the company to know which are the variables positively correlated and which pair of variables are negatively correlated.
- Negatively correlated variable can save the company from some decrease in sales.
- The clustering method can help the company wih number of cluster and members.
- Company can determine that clustering does not play a vital role in this sales data set.
- If the negative coefficient can be excluded then there will be increase in sales.
- The mean/average of the variables of the members are calculated and shown thus the company can show the average sales of different products

So the clustering data analysis using the “wholesale_2023” is successfully done and it can be very useful for the company for future sales.