**Unsupervised Learning: K-Means Clustering**

**Author:** Harshita Singh, 8786434

**Course:** PROG8430

**Background**

**Data Source**

1. **Data Transformation:**

* Quick Normalization Function(Min-max normalization): It does not handle outliers and it will have same scale with the features



* Different Normalization Function (Z-score normalization): It handles outliers and it will not normalize data with the same scale.



**A Different standardization function ()**

*normDf\_HS <- function(x)*

*{*

*return((x-min(x)/sd(x)))*

*}*

#Standardization Food

ExpenseData\_HS$Food\_HS\_MinMax <- normDf\_HS(ExpenseData\_HS$Food\_HS)

#Standardization Entrance

ExpenseData\_HS$Entr\_HS\_MinMax <- normDf\_HS(ExpenseData\_HS$Entr\_HS)

#Standardization Education

ExpenseData\_HS$Educ\_HS\_MinMax <- normDf\_HS(ExpenseData\_HS$Educ\_HS)

#Standardization Transportation

ExpenseData\_HS$Tran\_HS\_MinMax <- normDf\_HS(ExpenseData\_HS$Tran\_HS)

#Standardization Work Related Expenses

ExpenseData\_HS$WorkRE\_HS\_MinMax <- normDf\_HS(ExpenseData\_HS$Work\_HS)

#Standardization House

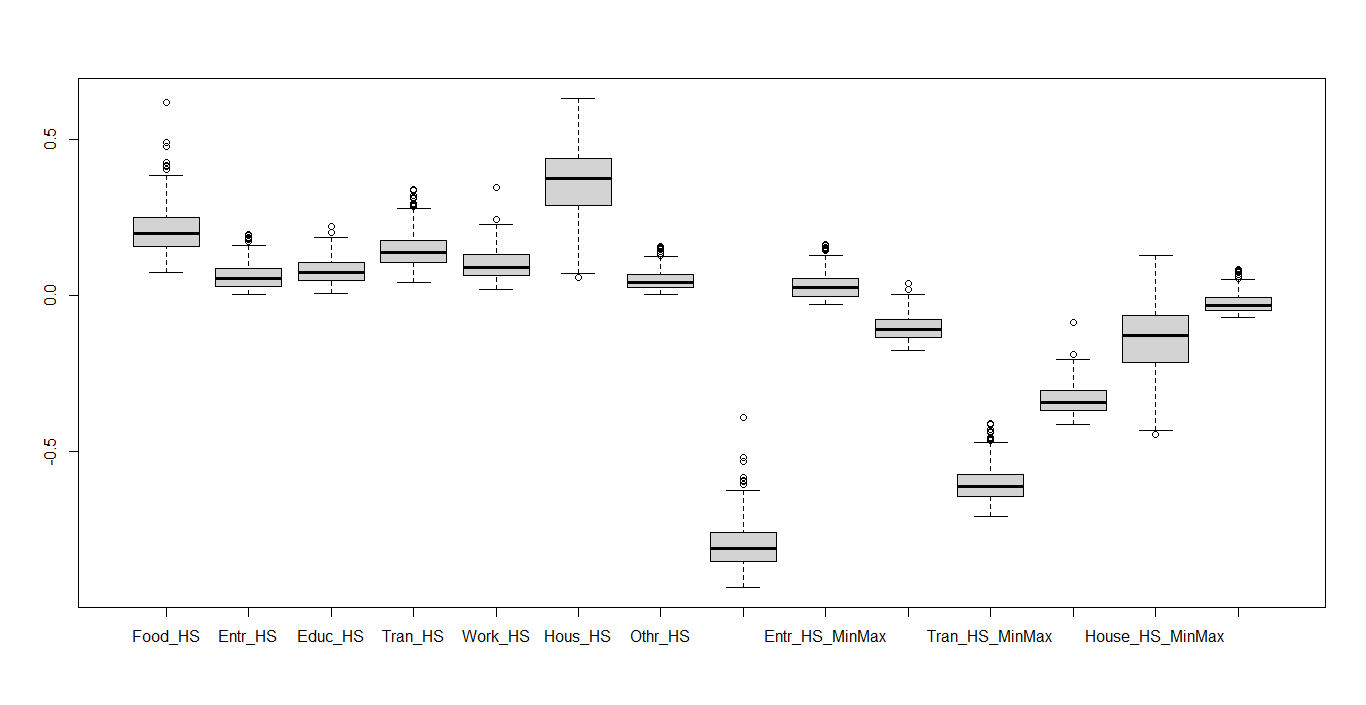
ExpenseData\_HS$House\_HS\_MinMax <- normDf\_HS(ExpenseData\_HS$Hous\_HS)

#Standardization Other Expenses

ExpenseData\_HS$Other\_HS\_MinMax <- normDf\_HS(ExpenseData\_HS$Othr\_HS)

* I choose **A Different standardization function** because it usually have few outliers and it does not need clipping.
* Mean of the standardized variable are close to 0, hence they are normally distributed.
* From the summary statistics I conclude that the percentage of income was distributed properly. Also, all of the data look reasonable.

1. Descriptive Data Analysis:

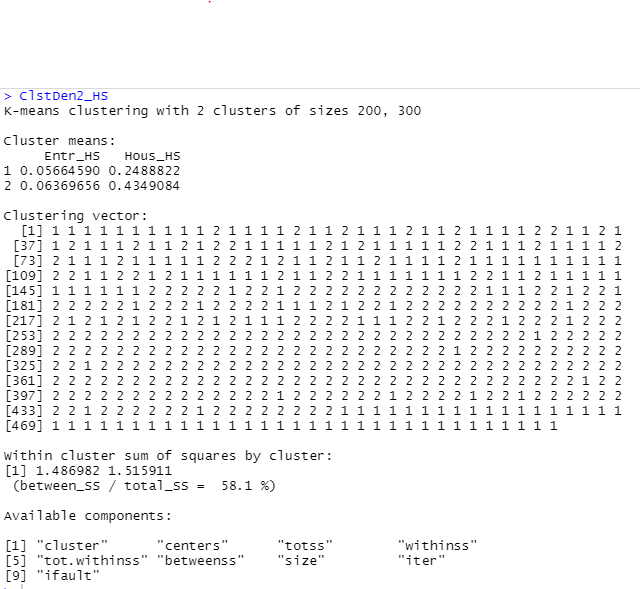
****

Boxplot represents the visual distribution of the quantitate data. Outlier observed: Food\_HS, Food\_HS\_MinMax, Tran\_HS\_MinMax they have 1.5 interquartile ranges away from Q1 or Q3 percentiles.

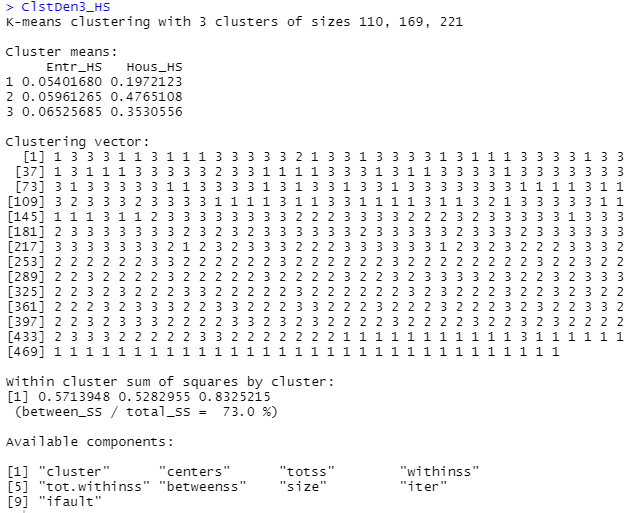
1. Clustering:

Part 1:

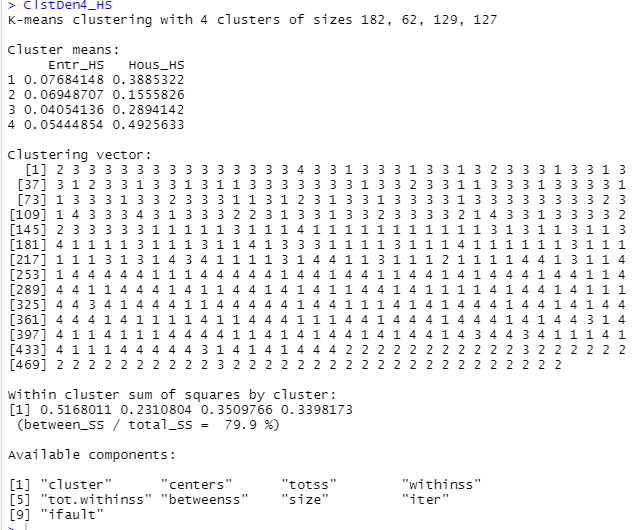
Cluster =2; 58.1% of variance is described by the cluster



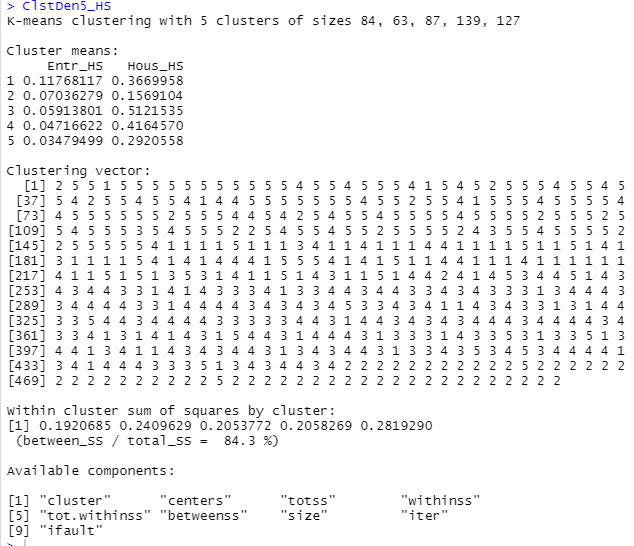
Cluster =3; 73% of variance is described by the cluster



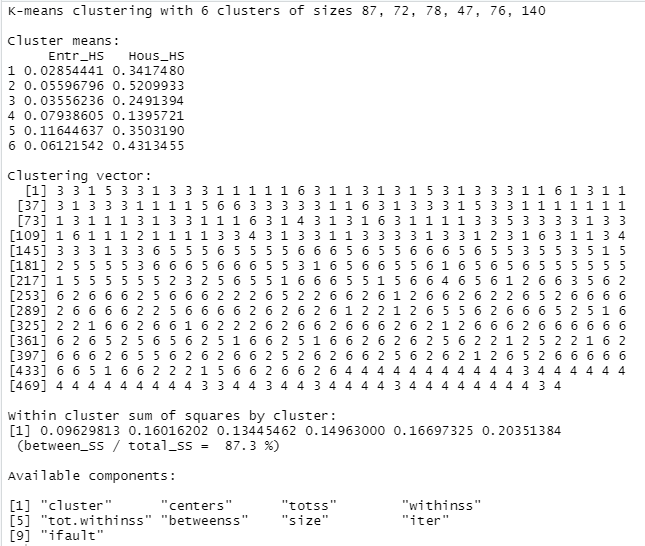
Cluster = 4; 79.9% of variance is described by the cluster

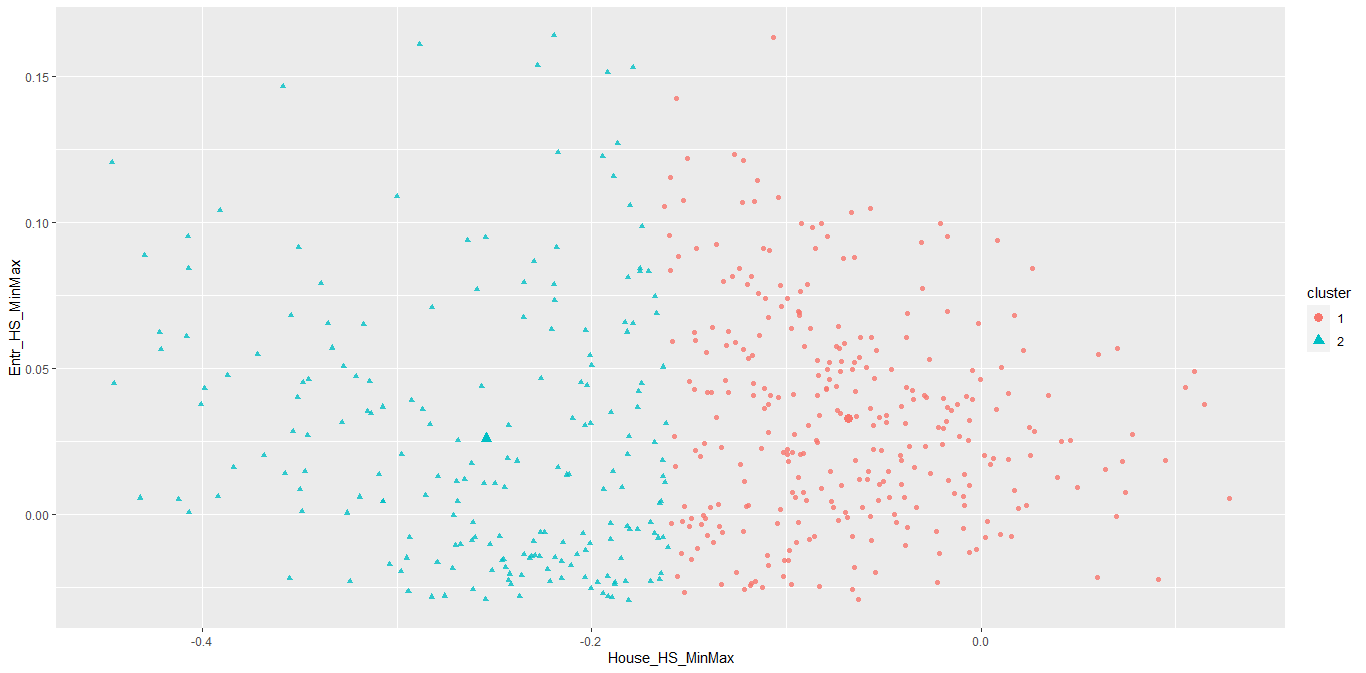


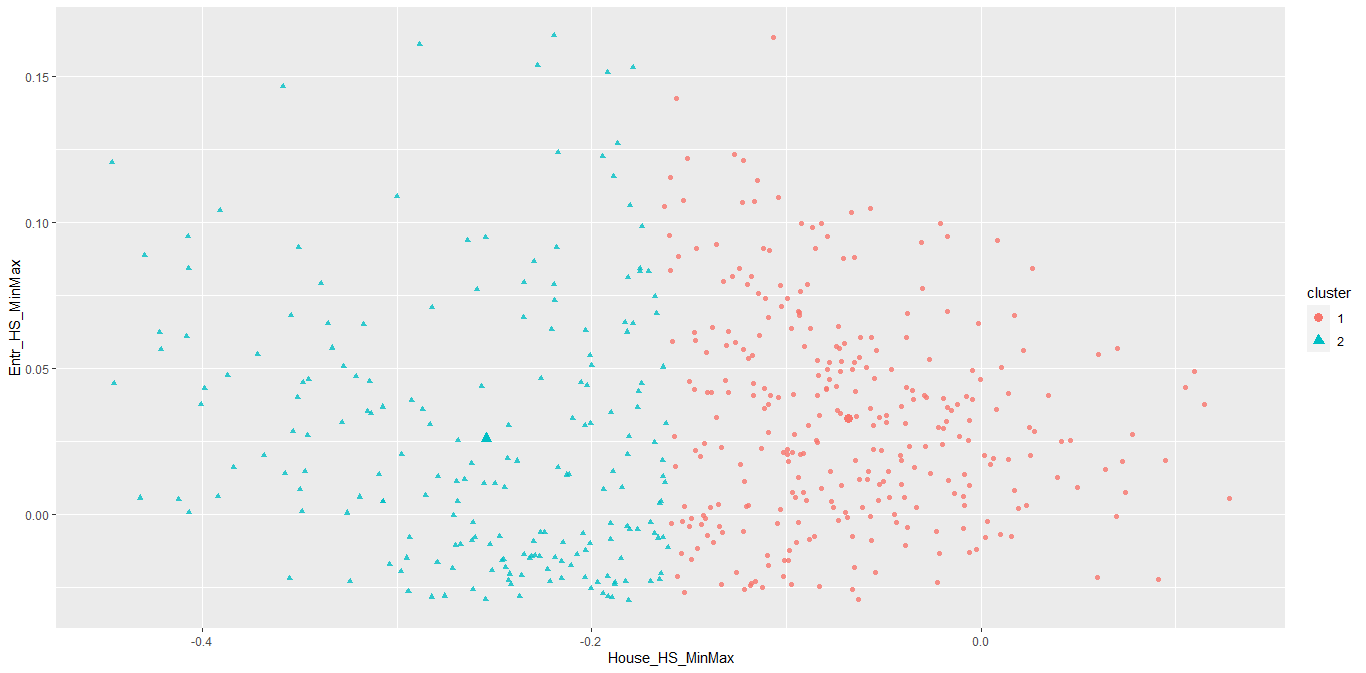
Cluster =5; 84.3% of variance is described by the cluster



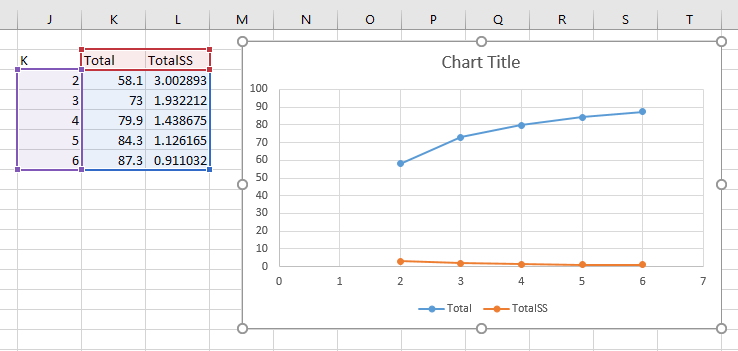
Cluster =6; 87.3% of variance is described by the cluster







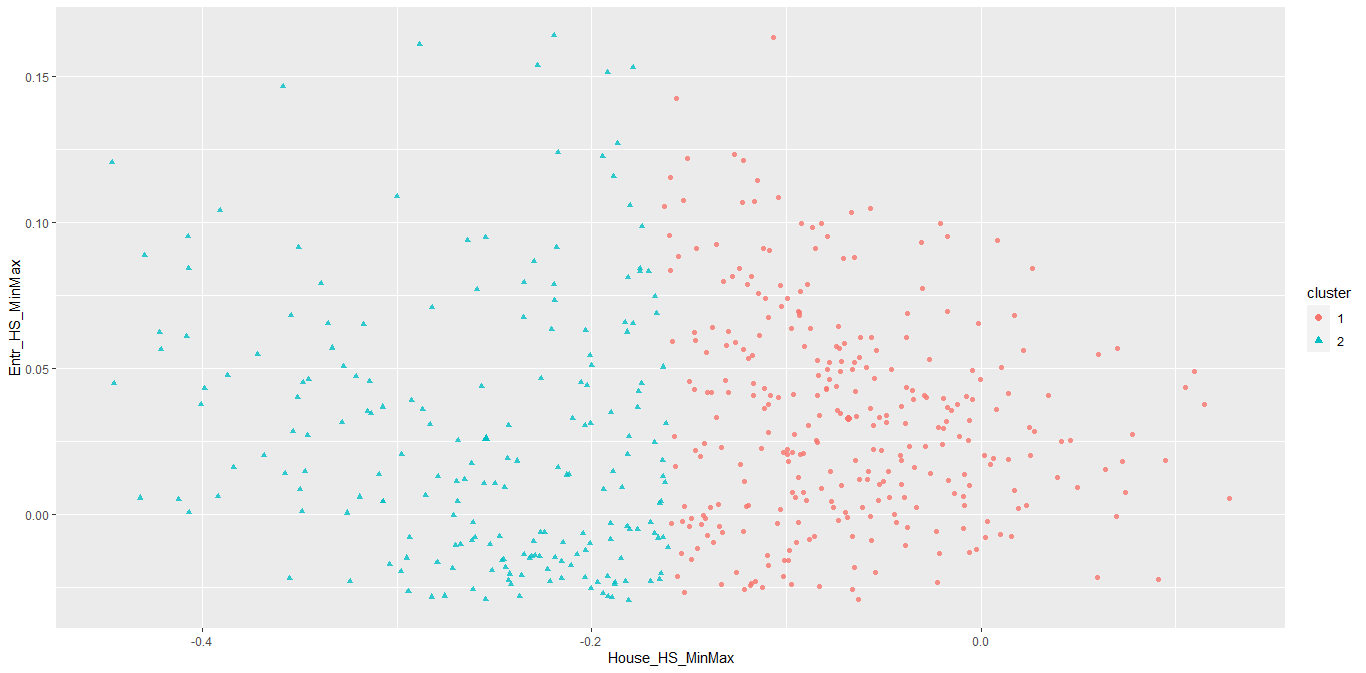
**Part 2: WSS Plot**



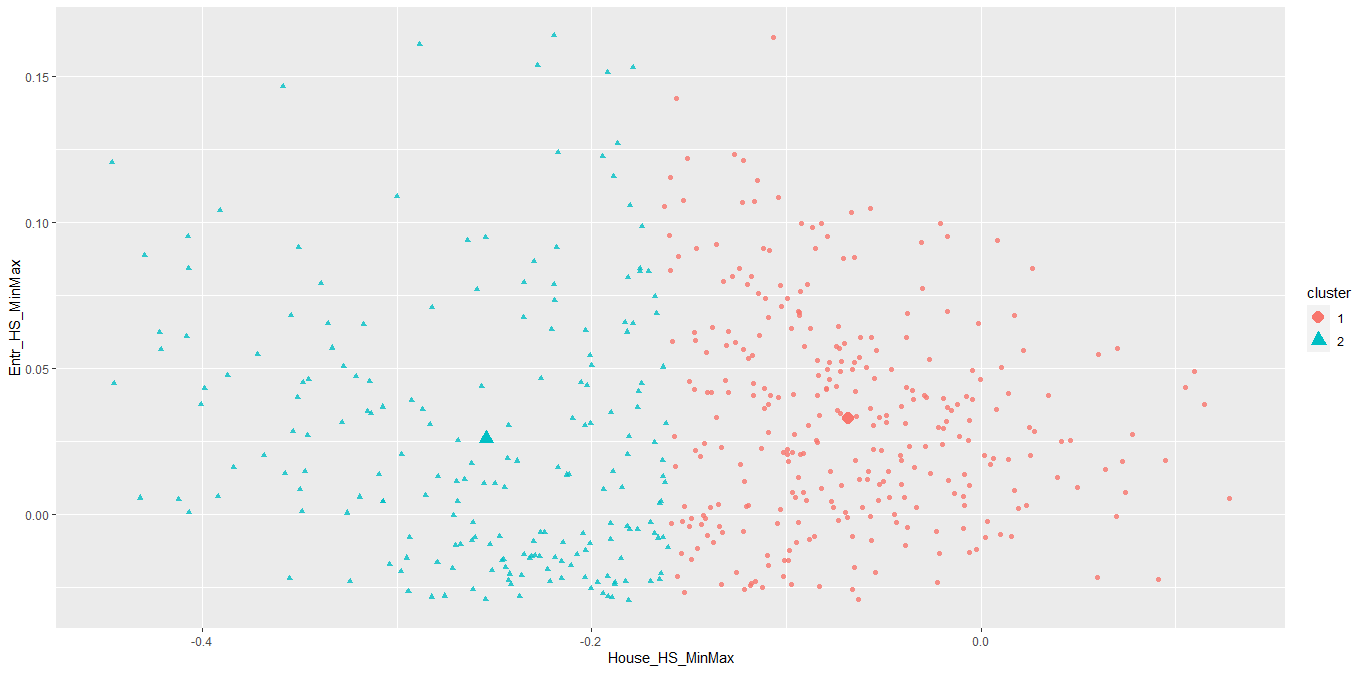
1. Evaluation of Clusters:

I think, elbow is at k=2. Then K-1 =1; K+1 =3

**K-1:**

****

**K+1:**

****

**Part 2:**

Where K=2; it best describes the data.