SLHD.Assessement-DSTI

Gustavo Chinchayan

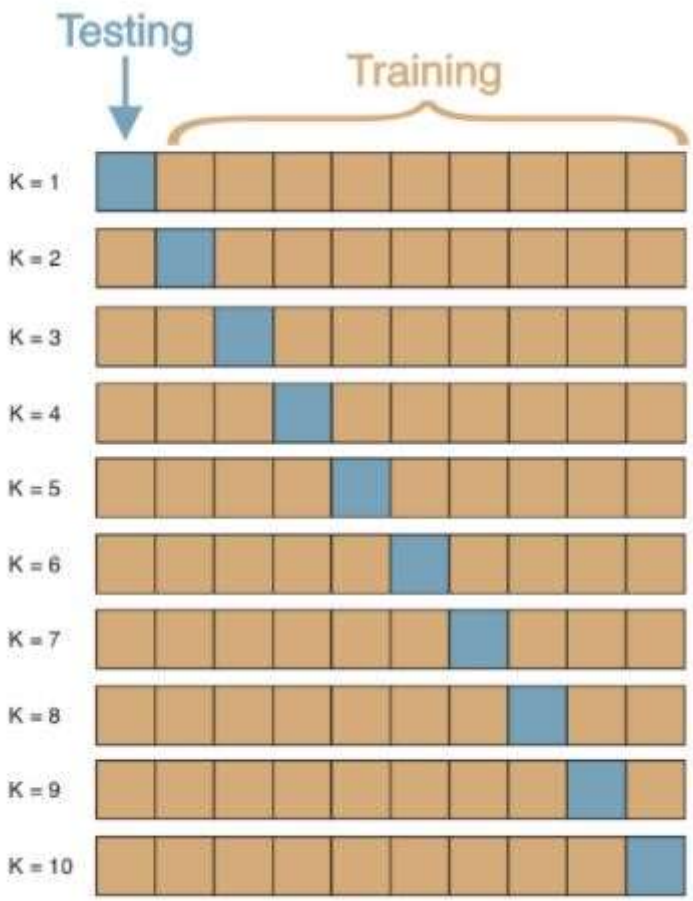
8/30/2021

Exercise 1: general questions

1.

A resampling method can be used for parameter tuning. This is a tool consisting of repeatedly drawing samples from a training data set and calculating statistics and metrics on each of those samples to obtain further information about the performance of a model. One of the methods of effective resampling technique is known as V-fold cross validation (CV). The V-fold cross validation is a robust method for estimating the accuracy of a model. The V-fold CV method starts by:

- Randomly splitting the data set into V number of subsets (5 or 10).
- One subset is reserved, and all the other subsets are used to train the model.
- Test the model on the reserved subset and record the prediction error
- Repeat this process until each of the V subsets has served as the test set.
- Compute the average of the V recorded error (also called the cross-validation error) used as the performance metric.



sample of cross-validation

Other Resampling methods are:

Leave one out Method: Which each observation is considered as the validation set and the rest (N-1) observations are considered as the training set.This is a special case of V-fold CV, its known to reduce bias and randomness.

BootStrapping: Independently sampling with replacement from an existing sample data with same sample size and performing inference among these resampled data. Examples of statistical inferences can be estimating standard error and obtaining confidence interval.

2.

A Gaussian mixture model is a model-based clustering where each data point has a probability of belonging to each cluster. In combination with the EM algorithm for inference, it is used to find the local maximum likelihood. The Gaussian mixture model comes with a family of different sub-models defined by their constraints of the covariance (spherical, diagonal, tied, and full). One such technique to find the number of clusters is known as the Bayesian Information Criterion (BIC). The BIC takes the log. likelihood of the model and adds a penalty which is dependent on the complexity of the model. It is the count of the scalar parameters for each sub models with constraints. In practice, several sub-models are compared at the same time with different number of K (number of groups) resulting in different BIC, therefore the model with the highest value of BIC is selected.

Exercise 2: hierchical clustering

1.

A Criterion to measure the clustering quality (Determines how well each object lies within its cluster) is the following:

J(K) = $\frac{B(maximize)}{W(minimize)}$ in practice we do $\frac{B}{S}\frac{B}{S}$, which both are equivalent (same optimization) In Clustering context:

A. SS is the general variance (this is fixed)
B. BB is the between group variance
C. WW is the within group variance

Data Scientists should choose the amount of K based on their understanding of the problem.

2.

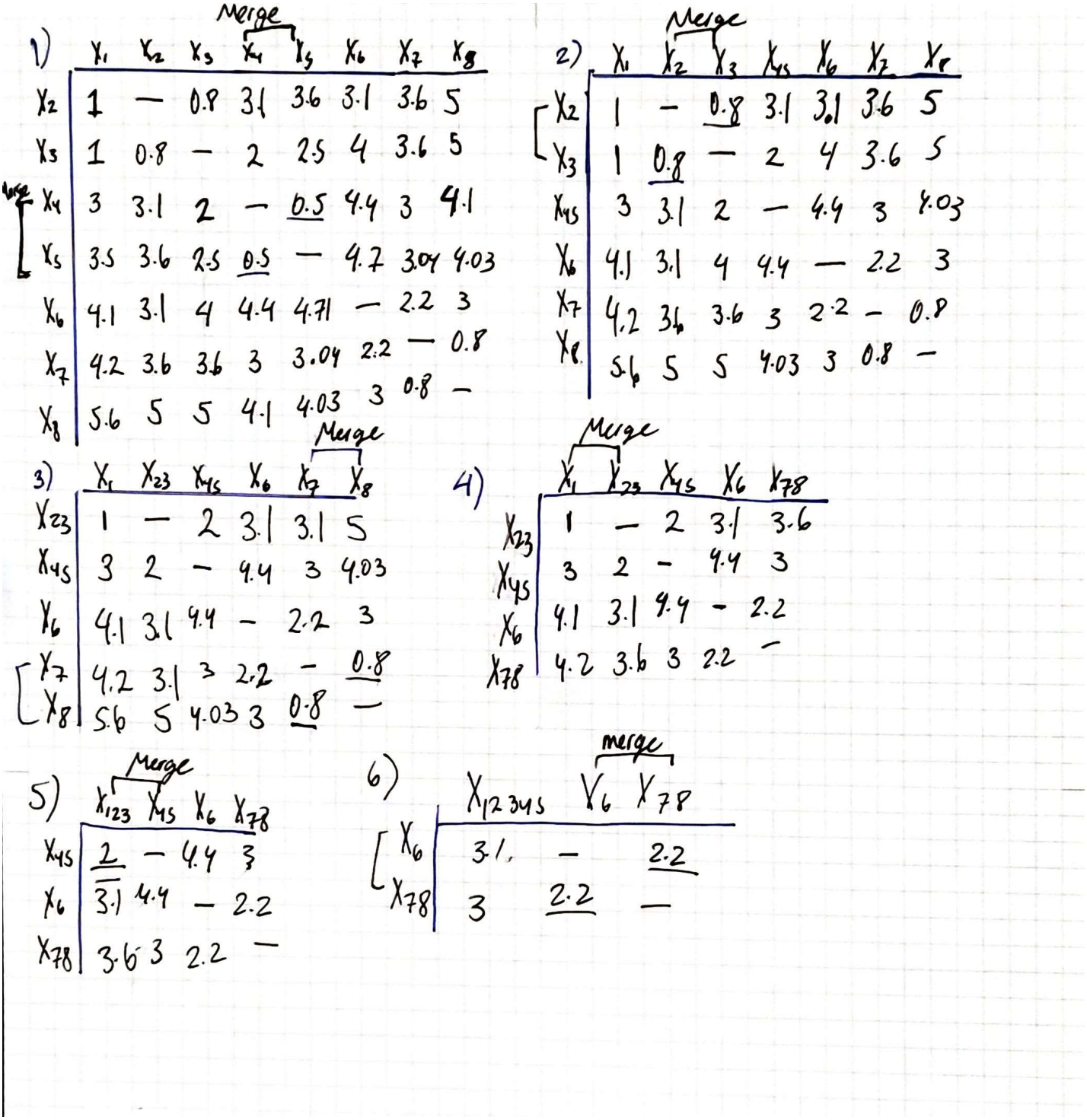
```
h.cl <- matrix(c(0, 1, 0, 2, 1, 1, 3, 1, 3.5, 1, 1, 5, 3, 4, 4, 5), nrow = 8, byrow = TRUE)
```

in $R^2 R^2$ with single linkage to calculate the distance between two points.

Calculating distance following this formula:

$$\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots . (q_n - p_n)^2}$$

$$\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots . (q_n - p_n)^2}$$



Math Data points attached for reference

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.0.5
```

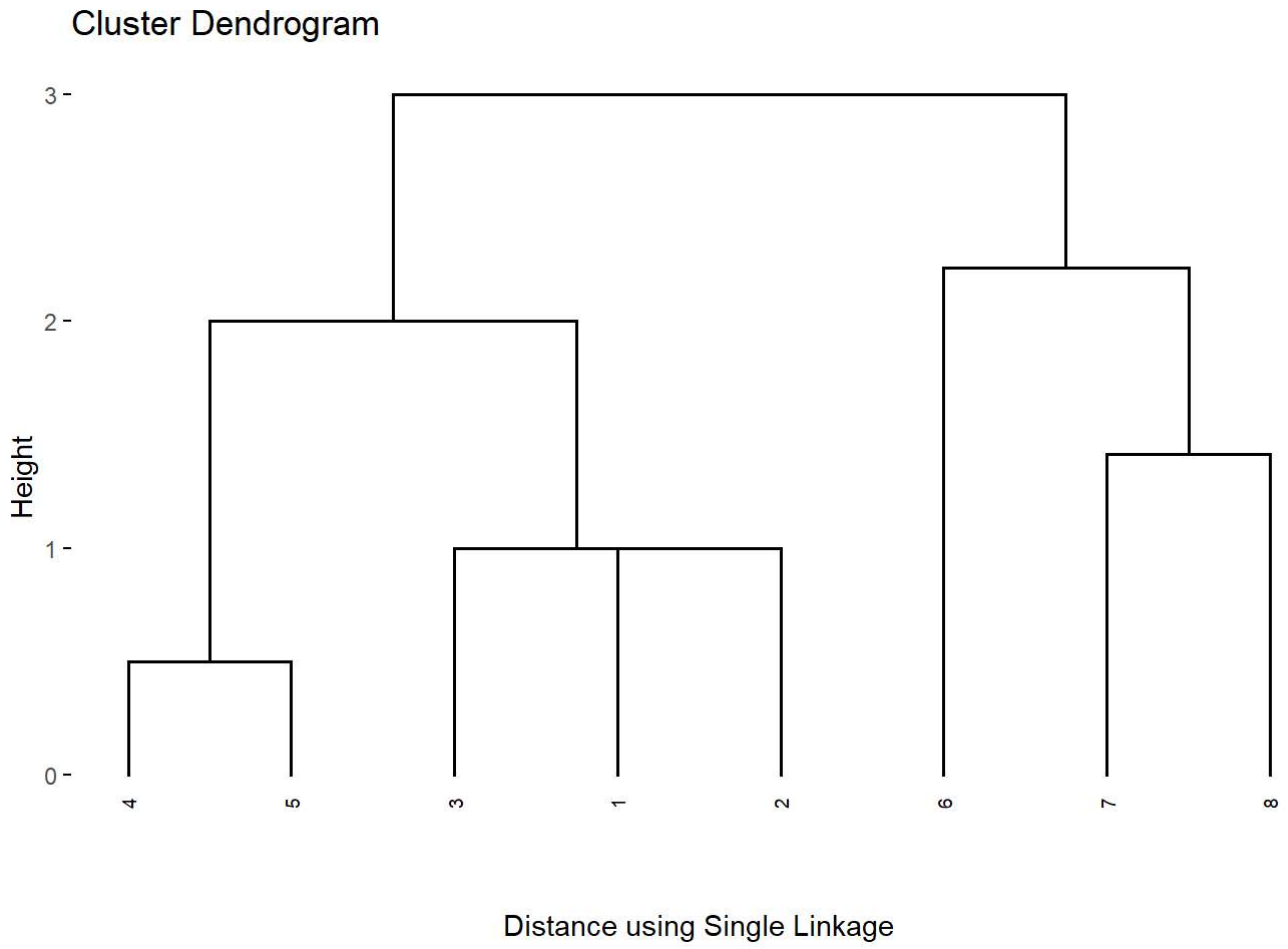
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
h.clust <- dist(h.cl, method = "euclidean")
singl.hc <- hclust(d = h.clust, method = "single")
fviz_dend(singl.hc, cex = 0.5, xlab = "Distance using Single Linkage")
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



Dendrogram generated by fviz_dend() function in factorextra package

Exercise 3: the Velib Data

3.1 Loading the Data

```
load("C:/Users/USER/Documents/R/R Exercize/velib.Rdata")
```

3.2 Pretreatment et descriptive analysis

Setting up the dataset

```
X <- velib$data
colnames(X) <- velib$dates
rownames(X) <- paste(c(1:NROW(velib$names)),velib$names)
X$means <- rowMeans(X)
```

Pretreatment

Checking the completeness of Dataframe and verifying the class,structure, and dimensions

```
class(X)
```

```
## [1] "data.frame"
```

```
dim(X)
```

```
## [1] 1189  182
```

Checking whether null values exist in this DF

```
is.null(X)
```

```
## [1] FALSE
```

Descriptive Analysis

Descriptive Statistics

```
summary(X)
```

```
##      Dim-11           Dim-12           Dim-13           Dim-14
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0625   1st Qu.:0.05263   1st Qu.:0.05882   1st Qu.:0.05556
##  Median :0.2500   Median :0.23077   Median :0.20000   Median :0.20513
##  Mean   :0.3698   Mean   :0.35481   Mean   :0.34735   Mean   :0.35250
##  3rd Qu.:0.6596   3rd Qu.:0.64000   3rd Qu.:0.62000   3rd Qu.:0.64103
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Dim-15           Dim-16           Dim-17           Dim-18
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.04762   1st Qu.:0.04286   1st Qu.:0.04348   1st Qu.:0.05085
##  Median :0.18462   Median :0.15789   Median :0.15789   Median :0.16667
##  Mean   :0.34459   Mean   :0.32828   Mean   :0.32347   Mean   :0.33054
##  3rd Qu.:0.64706   3rd Qu.:0.60000   3rd Qu.:0.56000   3rd Qu.:0.58065
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Dim-19           Dim-20           Dim-21           Dim-22
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0625   1st Qu.:0.07692   1st Qu.:0.1000   1st Qu.:0.1034
##  Median :0.2069   Median :0.25000   Median :0.2917   Median :0.3030
##  Mean   :0.3410   Mean   :0.36003   Mean   :0.3848   Mean   :0.3947
##  3rd Qu.:0.5926   3rd Qu.:0.61765   3rd Qu.:0.6522   3rd Qu.:0.6765
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##      Dim-23           Lun-00           Lun-01           Lun-02
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.09091   1st Qu.:0.09375   1st Qu.:0.09302   1st Qu.:0.09677
##  Median :0.30612   Median :0.31250   Median :0.33333   Median :0.34286
##  Mean   :0.39831   Mean   :0.40150   Mean   :0.40877   Mean   :0.41172
##  3rd Qu.:0.68421   3rd Qu.:0.69444   3rd Qu.:0.70588   3rd Qu.:0.70833
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Lun-03           Lun-04           Lun-05           Lun-06
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1000   1st Qu.:0.1087   1st Qu.:0.1143   1st Qu.:0.1143
##  Median :0.3500   Median :0.3571   Median :0.3636   Median :0.3600
##  Mean   :0.4152   Mean   :0.4171   Mean   :0.4216   Mean   :0.4199
##  3rd Qu.:0.7105   3rd Qu.:0.7000   3rd Qu.:0.7000   3rd Qu.:0.7000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      Lun-07           Lun-08           Lun-09           Lun-10
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.1176   1st Qu.:0.1143   1st Qu.:0.04762   1st Qu.:0.03571
##  Median :0.3600   Median :0.3200   Median :0.21538   Median :0.22222
##  Mean   :0.4157   Mean   :0.3847   Mean   :0.32633   Mean   :0.37170
##  3rd Qu.:0.6875   3rd Qu.:0.6364   3rd Qu.:0.57500   3rd Qu.:0.71429
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Lun-11           Lun-12           Lun-13           Lun-14
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.04167   1st Qu.:0.04545   1st Qu.:0.04762   1st Qu.:0.03846
##  Median :0.23333   Median :0.20455   Median :0.23077   Median :0.20833
##  Mean   :0.38307   Mean   :0.37759   Mean   :0.37246   Mean   :0.36972
##  3rd Qu.:0.75000   3rd Qu.:0.73333   3rd Qu.:0.70000   3rd Qu.:0.72000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Lun-15           Lun-16           Lun-17           Lun-18
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.03846   1st Qu.:0.0400   1st Qu.:0.04167   1st Qu.:0.0500
##  Median :0.19565   Median :0.1875   Median :0.20000   Median :0.2000
```

```
##  Mean   :0.37281   Mean   :0.3678   Mean   :0.36002   Mean   :0.3311
##  3rd Qu.:0.76000   3rd Qu.:0.7241   3rd Qu.:0.69643   3rd Qu.:0.5849
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##      Lun-19            Lun-20            Lun-21            Lun-22
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.05882   1st Qu.:0.06897   1st Qu.:0.07692   1st Qu.:0.07692
##  Median :0.20000   Median :0.22727   Median :0.27273   Median :0.28125
##  Mean   :0.30935   Mean   :0.33747   Mean   :0.37851   Mean   :0.39414
##  3rd Qu.:0.50000   3rd Qu.:0.55000   3rd Qu.:0.66667   3rd Qu.:0.73333
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Lun-23            Mar-00            Mar-01            Mar-02
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.06897   1st Qu.:0.06667   1st Qu.:0.0678   1st Qu.:0.06897
##  Median :0.27586   Median :0.27692   Median :0.3043   Median :0.33333
##  Mean   :0.39751   Mean   :0.40009   Mean   :0.4111   Mean   :0.41590
##  3rd Qu.:0.72727   3rd Qu.:0.72727   3rd Qu.:0.7500   3rd Qu.:0.74545
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##      Mar-03            Mar-04            Mar-05            Mar-06
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.07143   1st Qu.:0.07407   1st Qu.:0.08333   1st Qu.:0.08824
##  Median :0.34783   Median :0.35714   Median :0.36842   Median :0.37500
##  Mean   :0.42059   Mean   :0.42176   Mean   :0.42652   Mean   :0.42656
##  3rd Qu.:0.74286   3rd Qu.:0.75000   3rd Qu.:0.73913   3rd Qu.:0.73333
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Mar-07           Mar-08           Mar-09           Mar-10
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.1000   1st Qu.:0.1176   1st Qu.:0.07407   1st Qu.:0.05263
##  Median :0.3636   Median :0.3333   Median :0.25000   Median :0.26087
##  Mean   :0.4217   Mean   :0.3864   Mean   :0.31495   Mean   :0.36878
##  3rd Qu.:0.7073   3rd Qu.:0.6250   3rd Qu.:0.50000   3rd Qu.:0.65455
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Mar-11            Mar-12            Mar-13            Mar-14
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.04762   1st Qu.:0.05556   1st Qu.:0.05882   1st Qu.:0.05263
##  Median :0.26316   Median :0.25000   Median :0.25000   Median :0.24444
##  Mean   :0.37800   Mean   :0.37616   Mean   :0.37178   Mean   :0.36880
##  3rd Qu.:0.69697   3rd Qu.:0.68889   3rd Qu.:0.67857   3rd Qu.:0.67857
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Mar-15            Mar-16            Mar-17            Mar-18
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.05128   1st Qu.:0.04545   1st Qu.:0.05263   1st Qu.:0.05556
##  Median :0.24000   Median :0.22222   Median :0.20513   Median :0.21053
##  Mean   :0.37263   Mean   :0.36768   Mean   :0.36031   Mean   :0.33624
##  3rd Qu.:0.70000   3rd Qu.:0.68293   3rd Qu.:0.68182   3rd Qu.:0.60000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Mar-19            Mar-20            Mar-21            Mar-22
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.05556   1st Qu.:0.05556   1st Qu.:0.0600   1st Qu.:0.06667
##  Median :0.17949   Median :0.20833   Median :0.2400   Median :0.27273
##  Mean   :0.30928   Mean   :0.33607   Mean   :0.3728   Mean   :0.39259
##  3rd Qu.:0.52174   3rd Qu.:0.58824   3rd Qu.:0.6571   3rd Qu.:0.71698
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##      Mar-23           Mer-00           Mer-01           Mer-02
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
```

```
##  1st Qu.:0.0625   1st Qu.:0.05882   1st Qu.:0.06667   1st Qu.:0.07407
##  Median :0.2581   Median :0.25000   Median :0.27273   Median :0.30435
##  Mean   :0.3923   Mean   :0.39325   Mean   :0.40637   Mean   :0.41683
##  3rd Qu.:0.7273   3rd Qu.:0.73333   3rd Qu.:0.76000   3rd Qu.:0.76000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Mer-03           Mer-04           Mer-05            Mer-06
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0800   1st Qu.:0.0800   1st Qu.:0.08333   1st Qu.:0.09091
##  Median :0.3200   Median :0.3548   Median :0.37500   Median :0.36364
##  Mean   :0.4212   Mean   :0.4243   Mean   :0.42868   Mean   :0.42871
##  3rd Qu.:0.7576   3rd Qu.:0.7500   3rd Qu.:0.75000   3rd Qu.:0.74194
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Mer-07           Mer-08           Mer-09            Mer-10
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.09677   1st Qu.:0.1071   1st Qu.:0.06667   1st Qu.:0.05128
##  Median :0.37500   Median :0.3500   Median :0.21739   Median :0.23529
##  Mean   :0.42408   Mean   :0.3871   Mean   :0.31278   Mean   :0.36163
##  3rd Qu.:0.73469   3rd Qu.:0.6364   3rd Qu.:0.50000   3rd Qu.:0.64000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Mer-11           Mer-12           Mer-13            Mer-14
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.04545   1st Qu.:0.04762   1st Qu.:0.05263   1st Qu.:0.0625
##  Median :0.25000   Median :0.25000   Median :0.25000   Median :0.2500
##  Mean   :0.37457   Mean   :0.37536   Mean   :0.36517   Mean   :0.3696
##  3rd Qu.:0.68750   3rd Qu.:0.69048   3rd Qu.:0.67391   3rd Qu.:0.6667
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      Mer-15           Mer-16           Mer-17            Mer-18
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.05128   1st Qu.:0.05085   1st Qu.:0.04762   1st Qu.:0.05263
##  Median :0.23333   Median :0.22222   Median :0.22727   Median :0.21429
##  Mean   :0.37228   Mean   :0.36857   Mean   :0.36303   Mean   :0.33426
##  3rd Qu.:0.71875   3rd Qu.:0.68293   3rd Qu.:0.67347   3rd Qu.:0.57895
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Mer-19           Mer-20           Mer-21            Mer-22
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.05556   1st Qu.:0.05556   1st Qu.:0.0625   1st Qu.:0.07692
##  Median :0.18182   Median :0.20000   Median :0.2500   Median :0.25000
##  Mean   :0.30647   Mean   :0.34074   Mean   :0.3800   Mean   :0.39570
##  3rd Qu.:0.50000   3rd Qu.:0.58696   3rd Qu.:0.6957   3rd Qu.:0.73684
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##      Mer-23           Jeu-00           Jeu-01            Jeu-02
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0625   1st Qu.:0.0625   1st Qu.:0.06667   1st Qu.:0.06897
##  Median :0.2500   Median :0.2653   Median :0.28571   Median :0.32075
##  Mean   :0.3943   Mean   :0.3974   Mean   :0.40858   Mean   :0.41925
##  3rd Qu.:0.7391   3rd Qu.:0.7297   3rd Qu.:0.76087   3rd Qu.:0.76190
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Jeu-03           Jeu-04           Jeu-05            Jeu-06
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.07143   1st Qu.:0.07692   1st Qu.:0.07895   1st Qu.:0.08333
##  Median :0.33333   Median :0.35714   Median :0.36667   Median :0.36364
##  Mean   :0.42457   Mean   :0.42825   Mean   :0.43285   Mean   :0.43149
##  3rd Qu.:0.76923   3rd Qu.:0.76923   3rd Qu.:0.76000   3rd Qu.:0.76000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
```

```
##      Jeu-07            Jeu-08           Jeu-09            Jeu-10
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.09375   1st Qu.:0.1111   1st Qu.:0.05882   1st Qu.:0.05882
##  Median :0.36842   Median :0.3182   Median :0.25000   Median :0.27273
##  Mean   :0.42520   Mean   :0.3844   Mean   :0.33215   Mean   :0.38313
##  3rd Qu.:0.75000   3rd Qu.:0.6316   3rd Qu.:0.55556   3rd Qu.:0.69565
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Jeu-11            Jeu-12            Jeu-13            Jeu-14
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.05556   1st Qu.:0.06061   1st Qu.:0.07407   1st Qu.:0.06452
##  Median :0.28302   Median :0.28000   Median :0.28000   Median :0.28000
##  Mean   :0.39166   Mean   :0.38225   Mean   :0.38193   Mean   :0.38159
##  3rd Qu.:0.73684   3rd Qu.:0.70000   3rd Qu.:0.68421   3rd Qu.:0.69231
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Jeu-15            Jeu-16            Jeu-17            Jeu-18
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0625   1st Qu.:0.06061   1st Qu.:0.06897   1st Qu.:0.06061
##  Median :0.2727   Median :0.26415   Median :0.25000   Median :0.22727
##  Mean   :0.3843   Mean   :0.38051   Mean   :0.37192   Mean   :0.34304
##  3rd Qu.:0.7097   3rd Qu.:0.70000   3rd Qu.:0.67857   3rd Qu.:0.60000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Jeu-19            Jeu-20            Jeu-21            Jeu-22
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.06061   1st Qu.:0.06061   1st Qu.:0.06452   1st Qu.:0.06667
##  Median :0.21053   Median :0.22222   Median :0.25926   Median :0.27273
##  Mean   :0.32786   Mean   :0.34946   Mean   :0.38724   Mean   :0.39751
##  3rd Qu.:0.54688   3rd Qu.:0.60714   3rd Qu.:0.72000   3rd Qu.:0.73684
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Jeu-23            Ven-00            Ven-01            Ven-02
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.05714   1st Qu.:0.05882   1st Qu.:0.06061   1st Qu.:0.06383
##  Median :0.25926   Median :0.27692   Median :0.30769   Median :0.32836
##  Mean   :0.39060   Mean   :0.39660   Mean   :0.40866   Mean   :0.41742
##  3rd Qu.:0.75000   3rd Qu.:0.75000   3rd Qu.:0.77273   3rd Qu.:0.76596
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Ven-03            Ven-04            Ven-05            Ven-06
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.06977   1st Qu.:0.07407   1st Qu.:0.08333   1st Qu.:0.08824
##  Median :0.34615   Median :0.36364   Median :0.38889   Median :0.38462
##  Mean   :0.42513   Mean   :0.43077   Mean   :0.43711   Mean   :0.43540
##  3rd Qu.:0.79070   3rd Qu.:0.77778   3rd Qu.:0.76744   3rd Qu.:0.76000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Ven-07            Ven-08           Ven-09           Ven-10
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.09302   1st Qu.:0.1111   1st Qu.:0.0625   1st Qu.:0.05882
##  Median :0.38462   Median :0.3478   Median :0.2581   Median :0.29630
##  Mean   :0.42786   Mean   :0.3884   Mean   :0.3393   Mean   :0.38685
##  3rd Qu.:0.74194   3rd Qu.:0.6364   3rd Qu.:0.5556   3rd Qu.:0.69767
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##      Ven-11            Ven-12            Ven-13            Ven-14
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0625   1st Qu.:0.05882   1st Qu.:0.07143   1st Qu.:0.0625
##  Median :0.3095   Median :0.30357   Median :0.30435   Median :0.3043
##  Mean   :0.3966   Mean   :0.38720   Mean   :0.38506   Mean   :0.3845
```

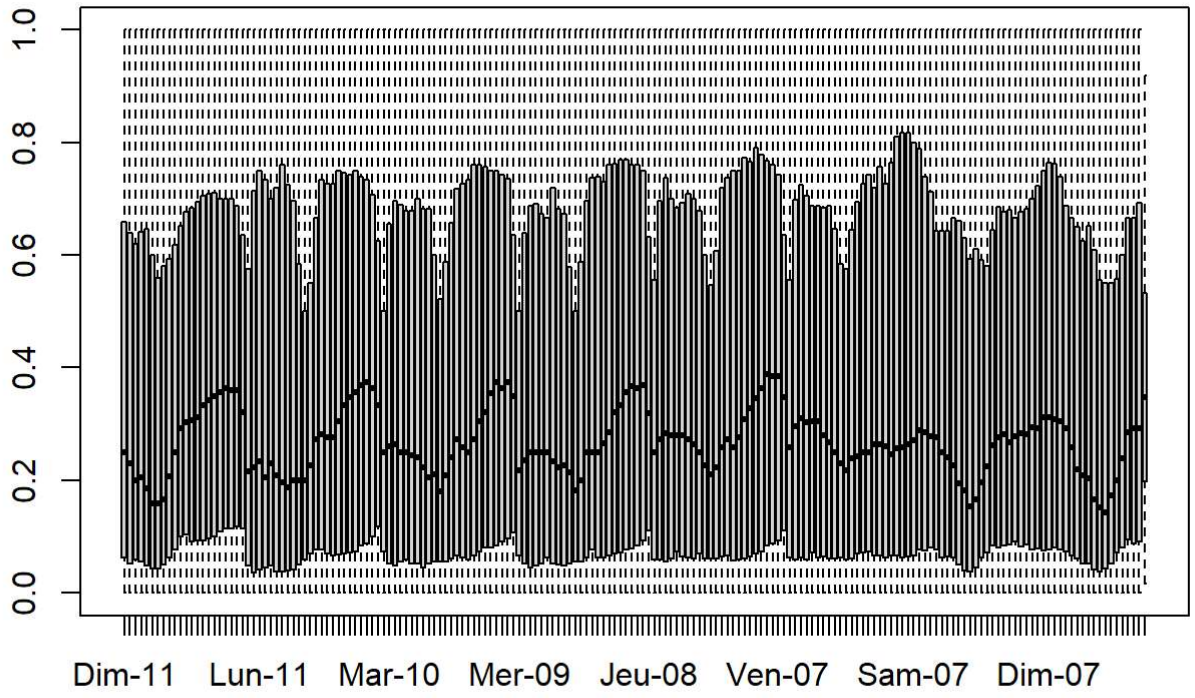
```
##  3rd Qu.:0.7241   3rd Qu.:0.70588   3rd Qu.:0.68750   3rd Qu.:0.6875
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      Ven-15            Ven-16            Ven-17            Ven-18
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.06452   1st Qu.:0.0600   1st Qu.:0.06061   1st Qu.:0.0625
##  Median :0.28000   Median :0.2667   Median :0.25000   Median :0.2308
##  Mean   :0.38198   Mean   :0.3761   Mean   :0.36205   Mean   :0.3391
##  3rd Qu.:0.68421   3rd Qu.:0.6875   3rd Qu.:0.64706   3rd Qu.:0.5833
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##      Ven-19            Ven-20            Ven-21            Ven-22
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.05882   1st Qu.:0.06061   1st Qu.:0.06897   1st Qu.:0.07143
##  Median :0.21739   Median :0.23810   Median :0.24242   Median :0.25000
##  Mean   :0.33910   Mean   :0.36088   Mean   :0.38303   Mean   :0.39209
##  3rd Qu.:0.57576   3rd Qu.:0.64516   3rd Qu.:0.69444   3rd Qu.:0.72727
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Ven-23            Sam-00            Sam-01            Sam-02
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.07407   1st Qu.:0.06667   1st Qu.:0.06667   1st Qu.:0.0625
##  Median :0.25000   Median :0.26316   Median :0.26415   Median :0.2609
##  Mean   :0.39527   Mean   :0.39051   Mean   :0.39614   Mean   :0.3935
##  3rd Qu.:0.74194   3rd Qu.:0.72000   3rd Qu.:0.75676   3rd Qu.:0.7273
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      Sam-03            Sam-04            Sam-05            Sam-06
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.06667   1st Qu.:0.06667   1st Qu.:0.0625   1st Qu.:0.06452
##  Median :0.24590   Median :0.25714   Median :0.2581   Median :0.26316
##  Mean   :0.39706   Mean   :0.40424   Mean   :0.4076   Mean   :0.41293
##  3rd Qu.:0.76471   3rd Qu.:0.80952   3rd Qu.:0.8182   3rd Qu.:0.81818
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.00000
##      Sam-07            Sam-08            Sam-09            Sam-10
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.06667   1st Qu.:0.07692   1st Qu.:0.0750   1st Qu.:0.0800
##  Median :0.27027   Median :0.28947   Median :0.2857   Median :0.2778
##  Mean   :0.41326   Mean   :0.40920   Mean   :0.4010   Mean   :0.3869
##  3rd Qu.:0.80000   3rd Qu.:0.78947   3rd Qu.:0.7391   3rd Qu.:0.7121
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##      Sam-11            Sam-12            Sam-13            Sam-14
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.07692   1st Qu.:0.0625   1st Qu.:0.06452   1st Qu.:0.0625
##  Median :0.27692   Median :0.2500   Median :0.24000   Median :0.2273
##  Mean   :0.37260   Mean   :0.3615   Mean   :0.35875   Mean   :0.3606
##  3rd Qu.:0.64286   3rd Qu.:0.6429   3rd Qu.:0.64286   3rd Qu.:0.6667
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##      Sam-15            Sam-16            Sam-17            Sam-18
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0500   1st Qu.:0.0400   1st Qu.:0.03704   1st Qu.:0.04545
##  Median :0.1944   Median :0.1818   Median :0.15385   Median :0.16667
##  Mean   :0.3474   Mean   :0.3328   Mean   :0.32363   Mean   :0.32592
##  3rd Qu.:0.6600   3rd Qu.:0.6296   3rd Qu.:0.59259   3rd Qu.:0.61111
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Sam-19            Sam-20            Sam-21            Sam-22
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0625   1st Qu.:0.07143   1st Qu.:0.08333   1st Qu.:0.08108
```

```
##  Median :0.1970   Median :0.22500   Median :0.26190   Median :0.27692
##  Mean   :0.3331   Mean   :0.34363   Mean   :0.37239   Mean   :0.38962
##  3rd Qu.:0.5909   3rd Qu.:0.58000   3rd Qu.:0.64444   3rd Qu.:0.68571
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Sam-23            Dim-00            Dim-01            Dim-02
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.08333   1st Qu.:0.08571   1st Qu.:0.09091   1st Qu.:0.08163
##  Median :0.28125   Median :0.26667   Median :0.27778   Median :0.28261
##  Mean   :0.39040   Mean   :0.38611   Mean   :0.38963   Mean   :0.38747
##  3rd Qu.:0.67742   3rd Qu.:0.68000   3rd Qu.:0.66667   3rd Qu.:0.67742
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Dim-03            Dim-04            Dim-05            Dim-06
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.08571   1st Qu.:0.07692   1st Qu.:0.07812   1st Qu.:0.0750
##  Median :0.28125   Median :0.29412   Median :0.29167   Median :0.3111
##  Mean   :0.39162   Mean   :0.39874   Mean   :0.40740   Mean   :0.4130
##  3rd Qu.:0.68182   3rd Qu.:0.70000   3rd Qu.:0.72222   3rd Qu.:0.7500
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##      Dim-07            Dim-08           Dim-09            Dim-10
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.07692   1st Qu.:0.0800   1st Qu.:0.07692   1st Qu.:0.07407
##  Median :0.31250   Median :0.3077   Median :0.30508   Median :0.29167
##  Mean   :0.41508   Mean   :0.4127   Mean   :0.40486   Mean   :0.38872
##  3rd Qu.:0.76471   3rd Qu.:0.7619   3rd Qu.:0.73913   3rd Qu.:0.68750
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Dim-11            Dim-12            Dim-13            Dim-14
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.06667   1st Qu.:0.06061   1st Qu.:0.05263   1st Qu.:0.05128
##  Median :0.25806   Median :0.21875   Median :0.20833   Median :0.20408
##  Mean   :0.37093   Mean   :0.35203   Mean   :0.34820   Mean   :0.34615
##  3rd Qu.:0.66667   3rd Qu.:0.65000   3rd Qu.:0.62500   3rd Qu.:0.65116
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Dim-15            Dim-16            Dim-17            Dim-18
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.04167   1st Qu.:0.03846   1st Qu.:0.04348   1st Qu.:0.05263
##  Median :0.16667   Median :0.15152   Median :0.14286   Median :0.17241
##  Mean   :0.32553   Mean   :0.31133   Mean   :0.30640   Mean   :0.31473
##  3rd Qu.:0.60870   3rd Qu.:0.55556   3rd Qu.:0.55000   3rd Qu.:0.55000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##      Dim-19            Dim-20           Dim-21            Dim-22
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.07143   1st Qu.:0.0800   1st Qu.:0.09524   1st Qu.:0.08696
##  Median :0.20000   Median :0.2381   Median :0.28571   Median :0.29268
##  Mean   :0.33408   Mean   :0.3535   Mean   :0.38598   Mean   :0.39255
##  3rd Qu.:0.55814   3rd Qu.:0.6000   3rd Qu.:0.66667   3rd Qu.:0.66667
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      Dim-23            means
##  Min.   :0.00000   Min.   :0.01675
##  1st Qu.:0.09091   1st Qu.:0.19758
##  Median :0.29268   Median :0.34775
##  Mean   :0.39753   Mean   :0.37979
##  3rd Qu.:0.69231   3rd Qu.:0.53244
##  Max.   :1.00000   Max.   :0.91909
```

Summary shows the daily velib bike station occupancy with range, quartiles, median and mean to get a better idea of the distribution of variables in the dataset

Boxplot

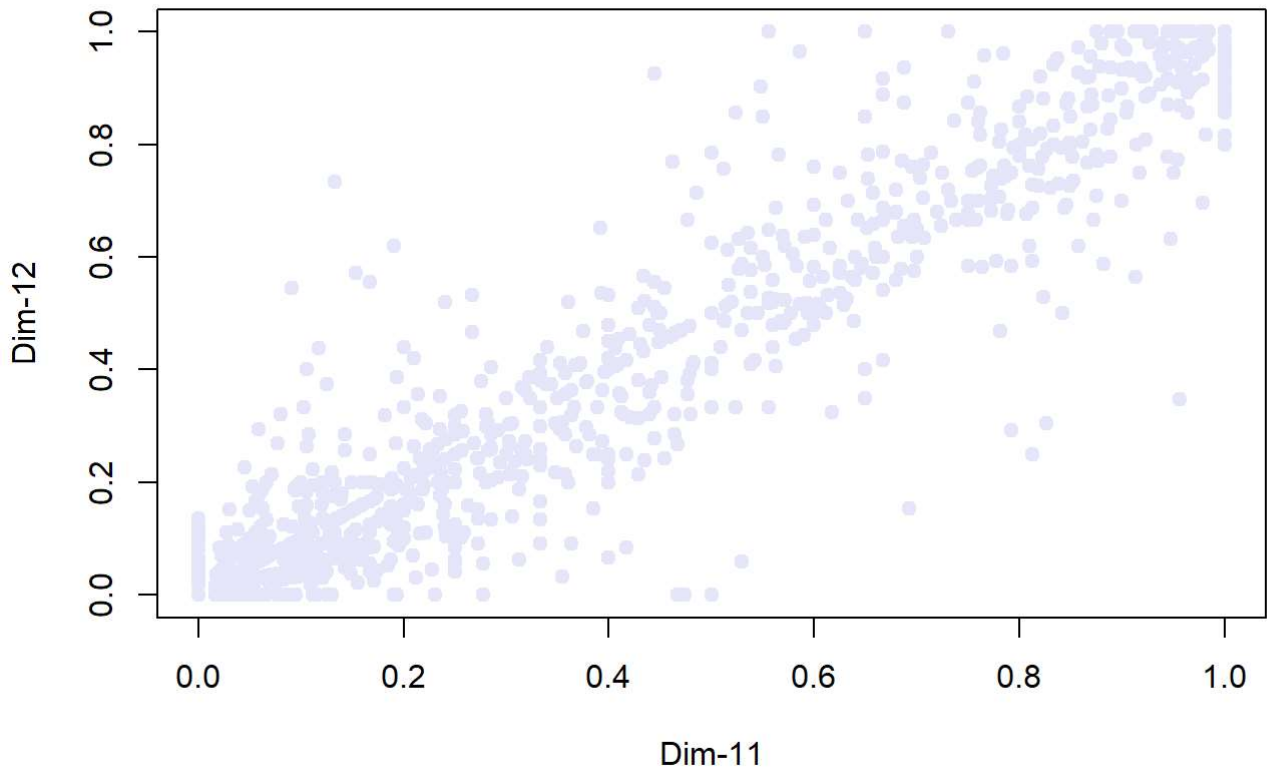
```
boxplot(X)
```



As we can see, there is a clear frequency of bike station occupancy on a through out the day and night. Seems like customers of velib take out bikes from the bikestations at specific time periods

Scatter plot between Dim-11 and Dim-12

```
plot(X[,1:2],pch = 19, col='lavender')
```



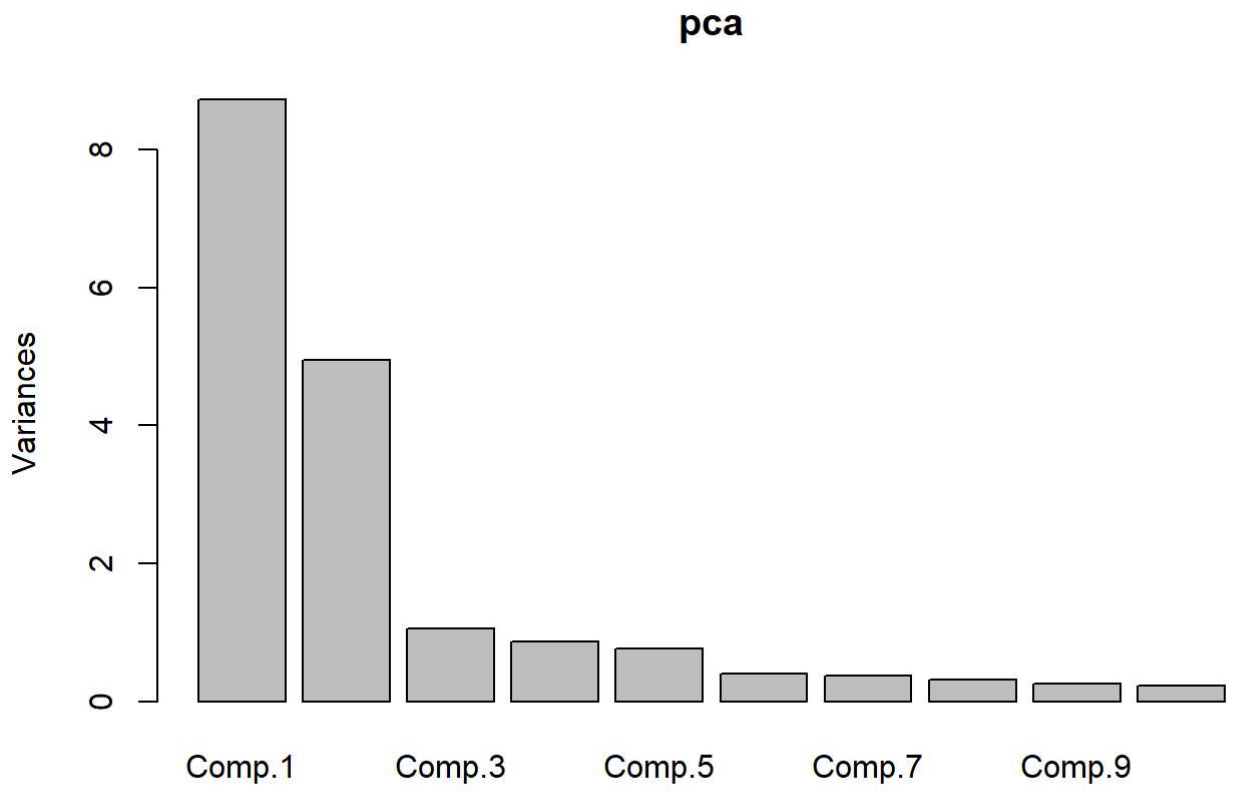
Comparing the 2 sundays there is similar correlation when it comes to bike station occupancy rate.

3.3 Data visualization

Principal component Analysis

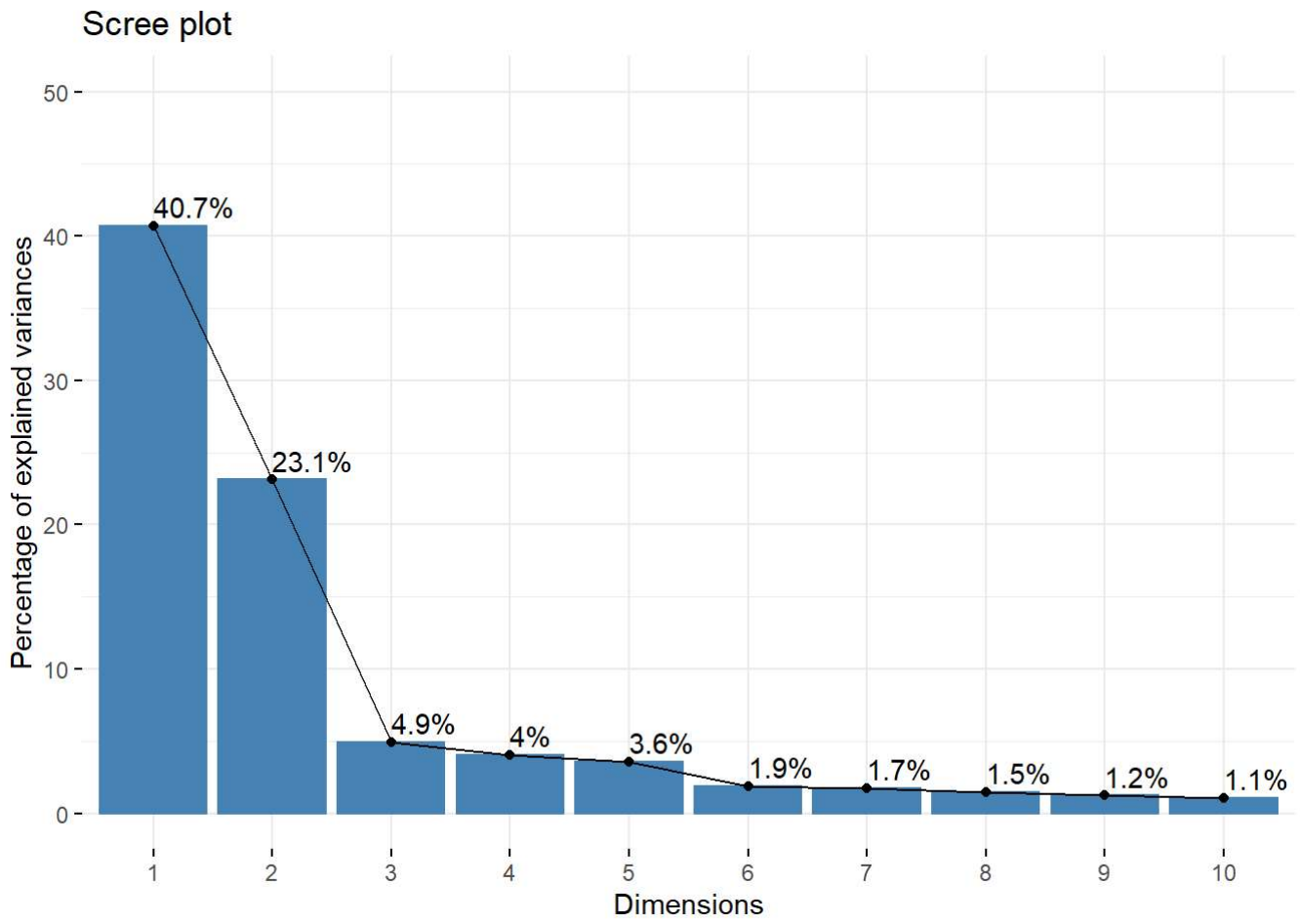
by using princomp to perform PCA

```
pca <- princomp(X)
Yhat = predict(pca)
screeplot(pca)
```



Loading Factoextra Library for comparative purposes to visualize the Screeplot in order to be certain in selecting the right number of dimensions for this test

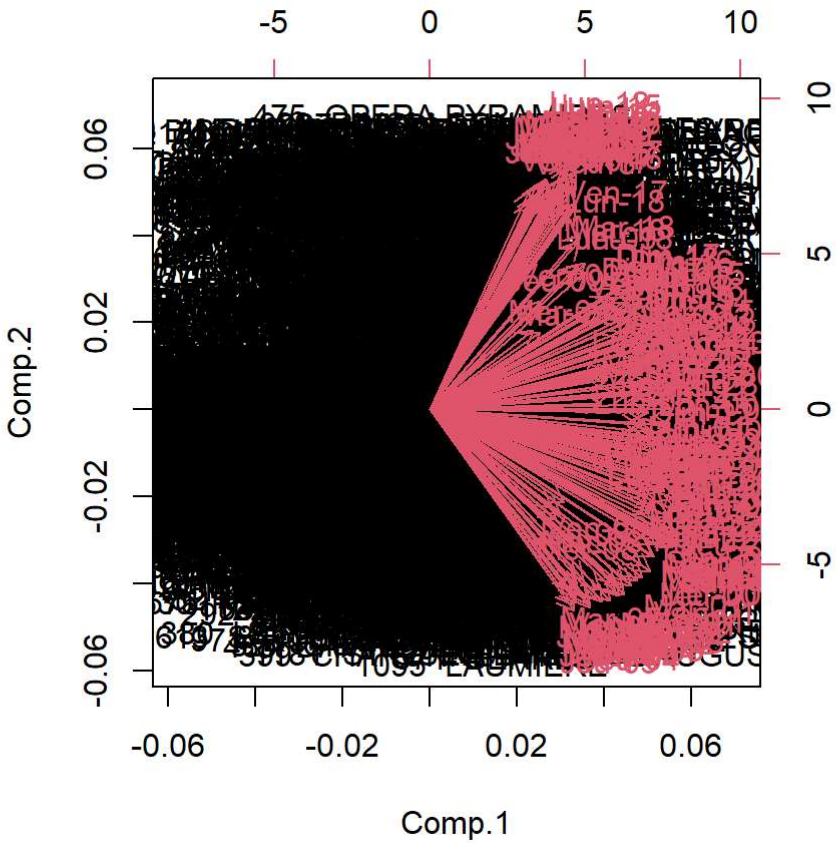
```
fviz_screeplot(pca, addlabels = TRUE, ylim = c(0, 50))
```



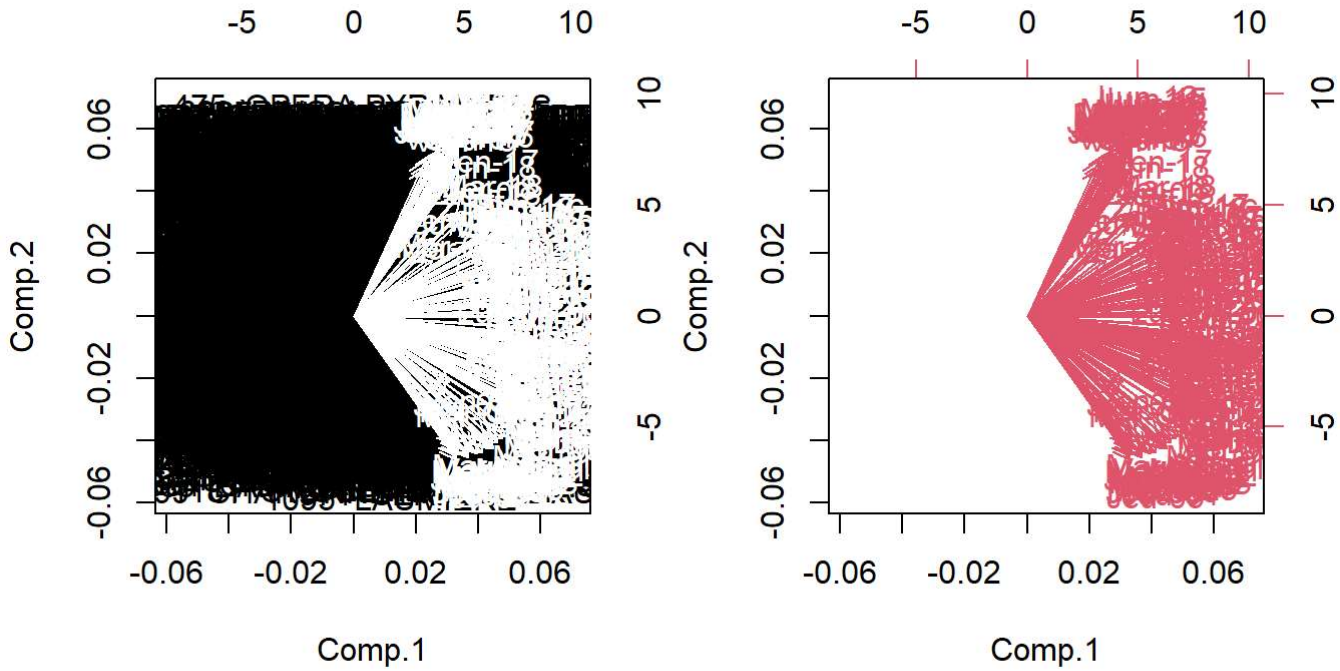
By looking at both Scree tests for comparison purposes, its safe to say that using 2 components seem to capture most amount data

After selecting the number components its best to look at the correlation circle to understand the relationships between the PCA axes and its original variables, we first place them in a biplot. Then we split the two plots in order to achieve clarity.

```
biplot(pca)
```



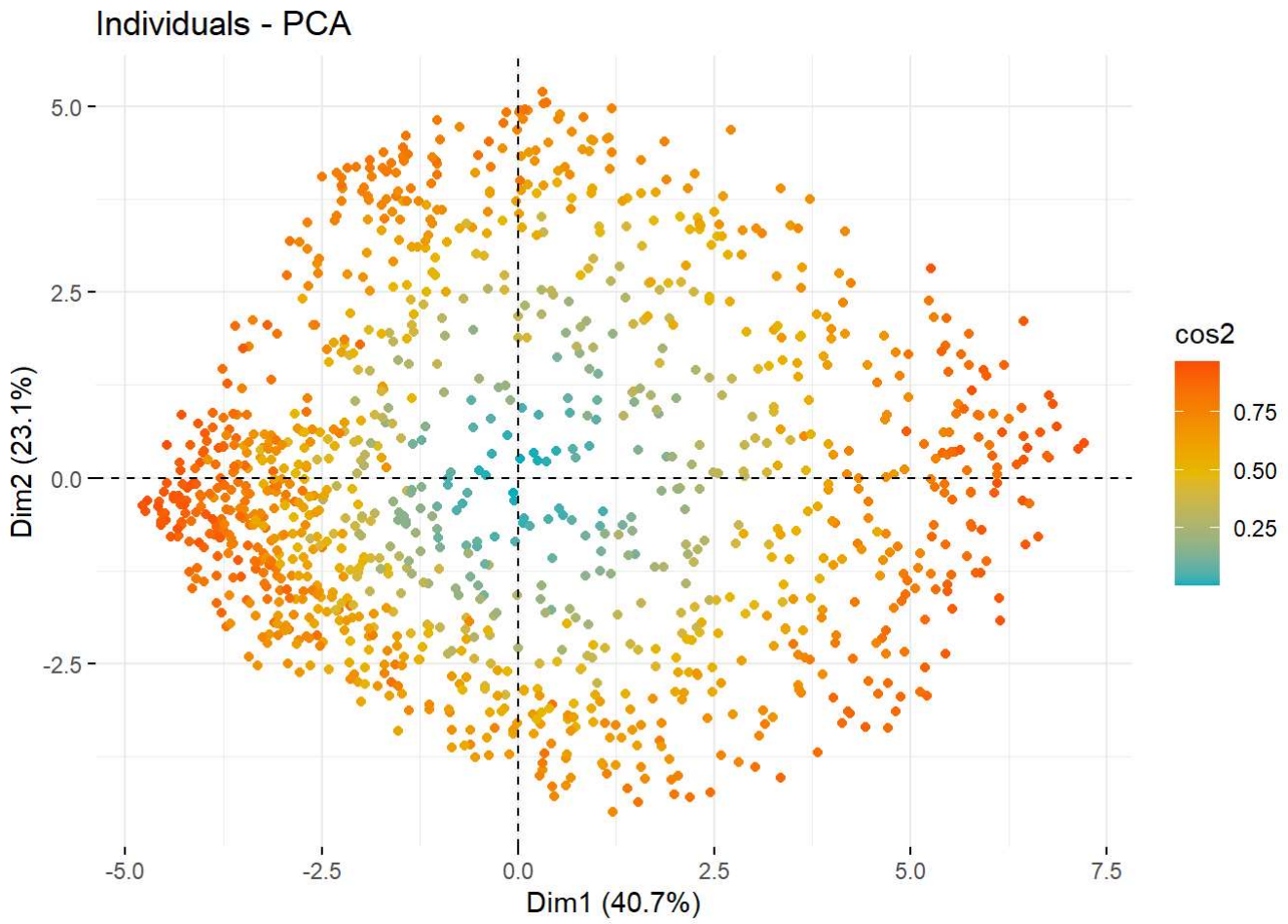

```
par(mfrow = c(1,2))
biplot(pca,col = c(1,0))
biplot(pca,col = c(0,2)); box()
```



By splitting the biplot datasets, and only focusing on the right hand-side graph that contains the dates (labeled in pink). There seems to be a positive correlation, but with the help of a different library, it can help me explain things better.

Factoextra Library is again used for comparative purpose agaisnt the earlier biplots

```
fviz_pca_ind(pca,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,    # Avoid text overlapping.
             label = FALSE
             )
```



As we can notice in both the Biplot graphs as well as the fviz_pca_ind graph it seems that most of the individual clusters are aligned in the right direction, meaning that there is positive correlation between components 1 and 2.

3.4 Clustering

3.4.1 Hierarchical clustering

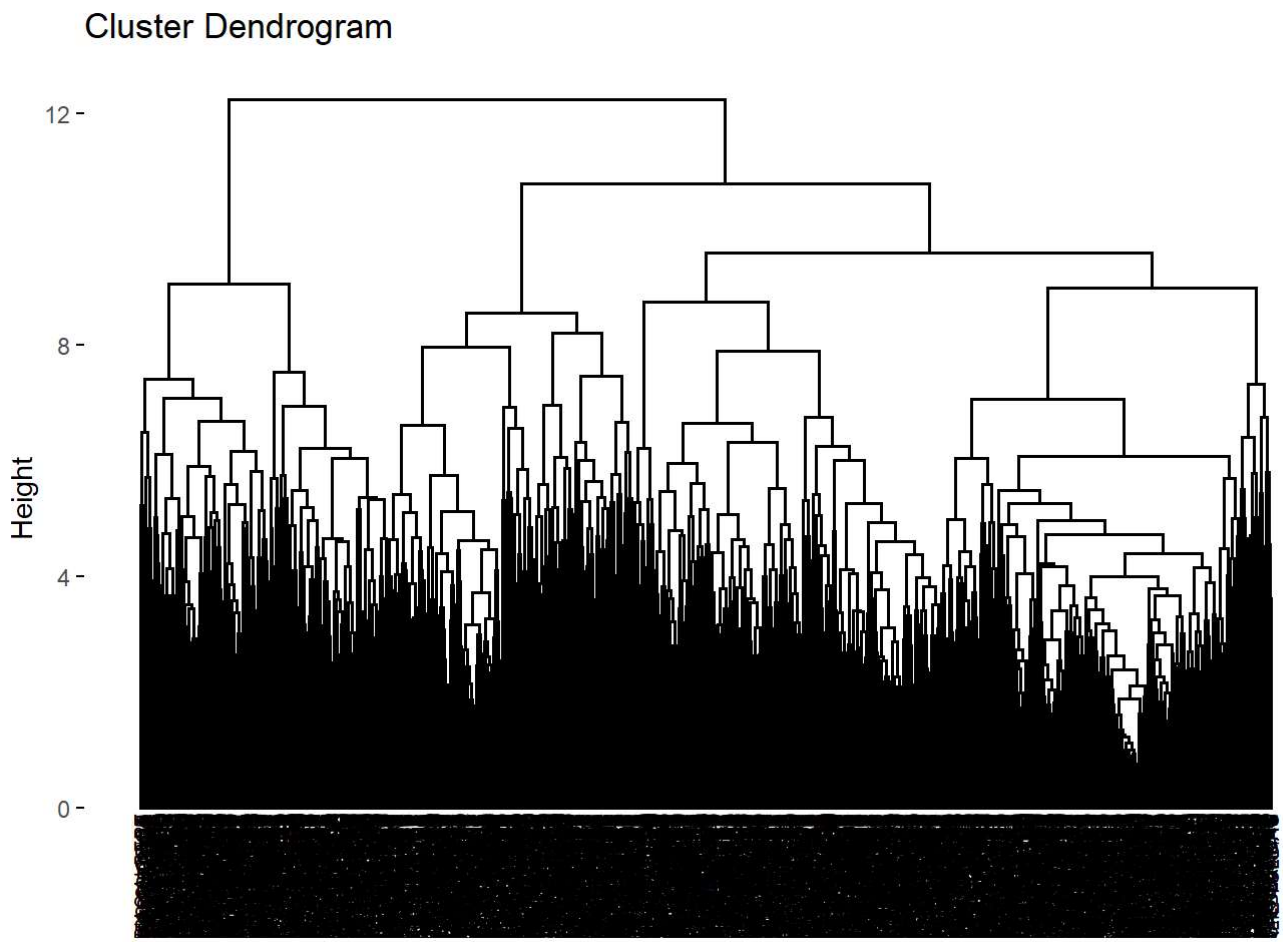
```
library(leaflet)
```

```
## Warning: package 'leaflet' was built under R version 4.0.5
```

```
velibmap <- velib$position
velibmap$bonus <- velib$bonus
```

```
res.dist <- dist(X, method = "euclidean")
res.hc <- hclust(d = res.dist, method = "complete")
fviz_dend(res.hc, cex = 0.5)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

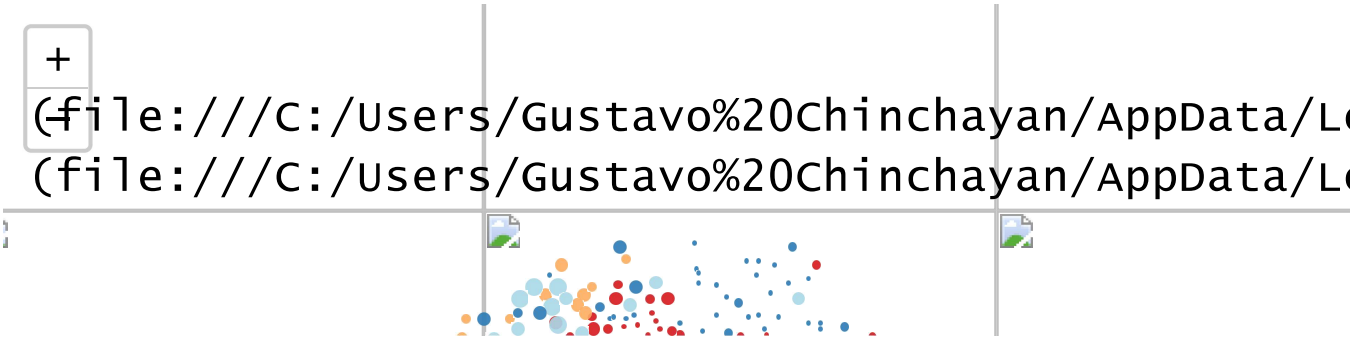


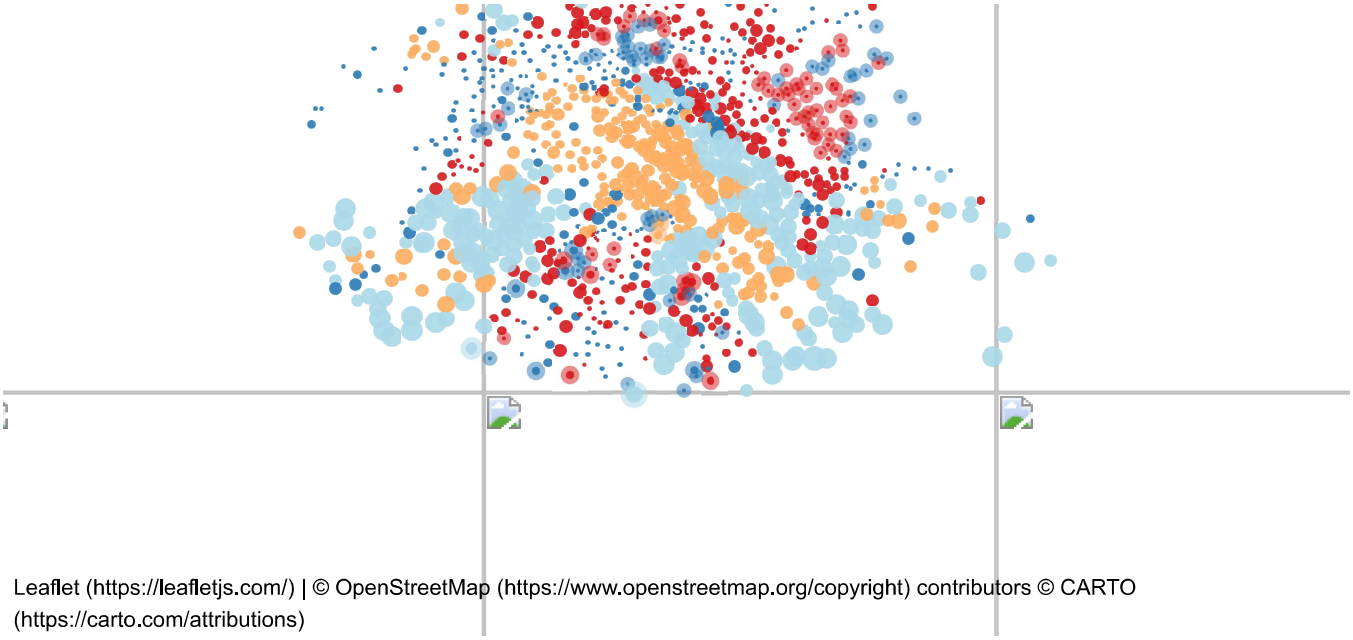
Looking at the Hirearchical clustering results with the help of the fviz_dend() function in factorextra package to produce this Dendrogram, I am almost certain that the best number of clusters should be 4 but it can also be 3. The K-means test can be conducted subsequently to finalize this decision.

```
h.clust <- cutree(res.hc, k = 4)
palette <- colorFactor("RdYlBu", domain = NULL)
```

```
leaflet(velibmap) %>% addProviderTiles(providers$CartoDB.Positron) %>%
  addCircleMarkers(radius = (X$means)*6,
                   color = palette(h.clust),
                   stroke = ~ifelse(velibmap$bonus == "1", TRUE, FALSE),
                   label = ~paste(row.names(X), sep = " - Clus.:", ... = h.clust),
                   fillOpacity = 0.9)
```

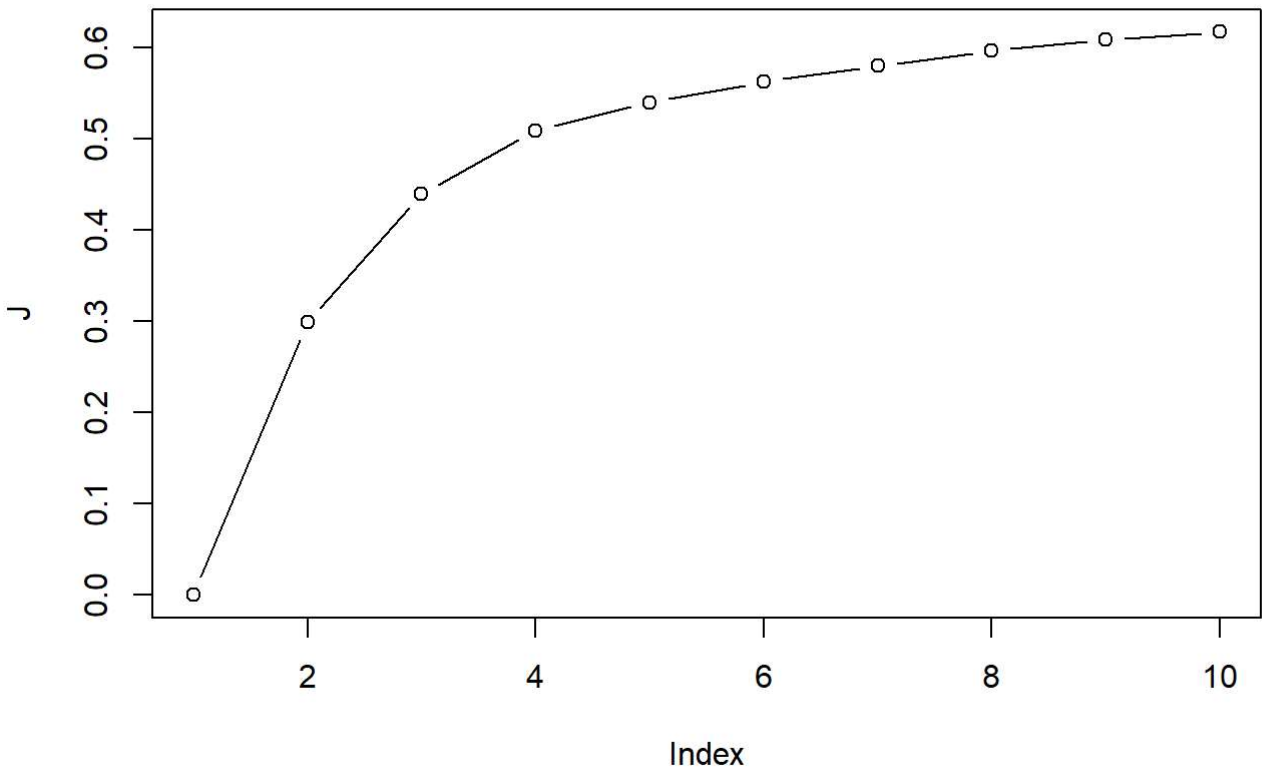
```
## Assuming "longitude" and "latitude" are longitude and latitude, respectively
```





3.4.2 k-means

```
Kmax = 10
J = rep(NA,Kmax)
for (k in 1:Kmax){
  out = kmeans(X, centers = k, nstart = 10)
  J[k] = out$betweenss / out$totss
}
plot(J,type='b')
```



Based on the result of this K-means test, it seems certain that 4 clusters it the optimal amount (where the bend occurs also known as the elbow). This can further confirm that the right number of clusters is 4 instead of 3 as previously observed in the hierarchical clustering test.

```
K.means <- 4
K.combined <- kmeans(X,centers = K.means, nstart = 10)
```

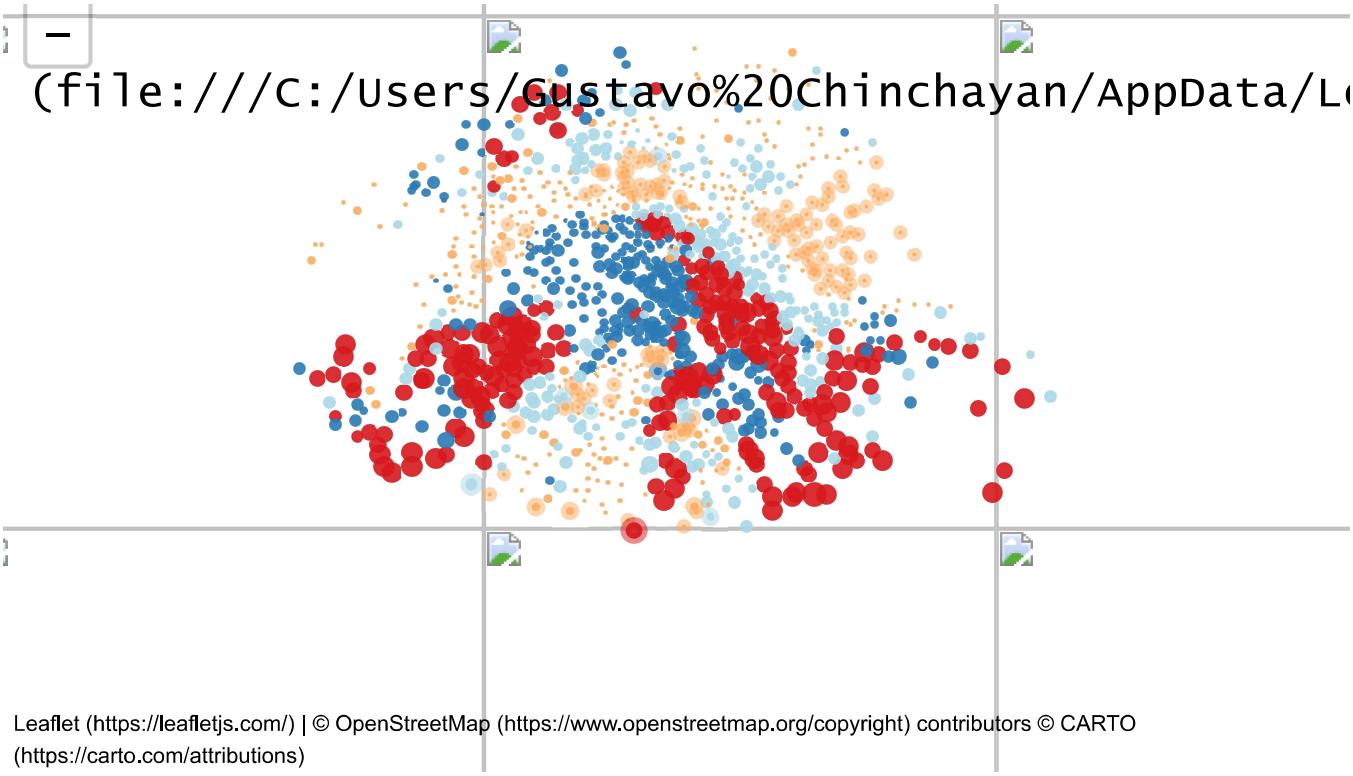
We can therefore build the same map before by adding the K means cluster number to the previous analysis using leaflet package

```
palette.2 <- colorFactor("RdYlBu", domain = NULL)
leaflet(velibmap) %>% addProviderTiles(providers$CartoDB.Positron) %>%
  addCircleMarkers(radius = (X$means)*6,
                   color = palette.2(K.combined$cluster),
                   stroke = ~ifelse(velib$bonus == "1", TRUE, FALSE),
                   label = ~paste(row.names(X), sep = " - Clus.:", K.combined),
                   fillOpacity = 0.9)
```

```
## Assuming "longitude" and "latitude" are longitude and latitude, respectively
```

+

(file:///C:/Users/Gustavo%20Chinchayan/AppData/L

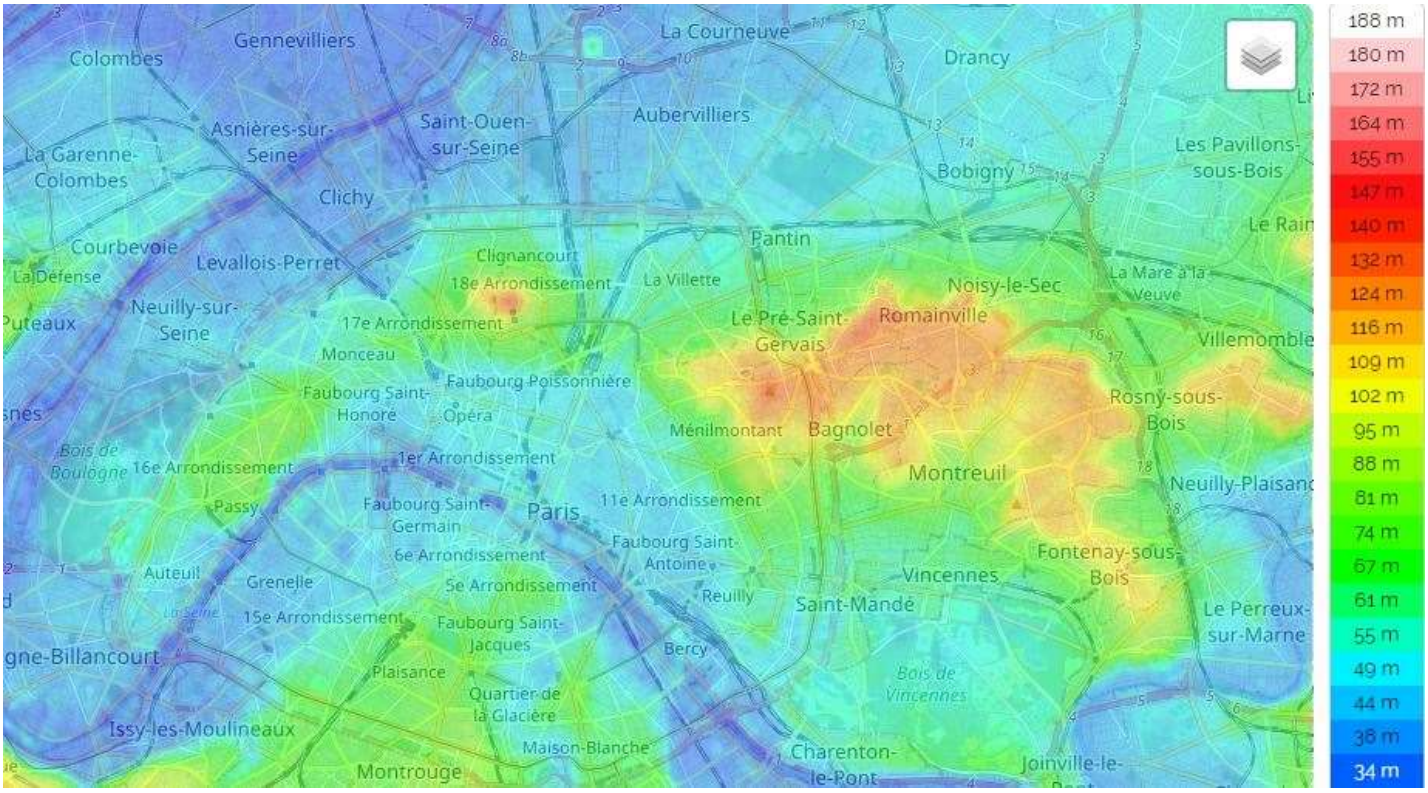


3.5 Summary

In Summary. we see that both Hierarchical clustering as well as K-means clustering exhibit similar results of 4 clusters when visualizing the data set using the leaflet package.

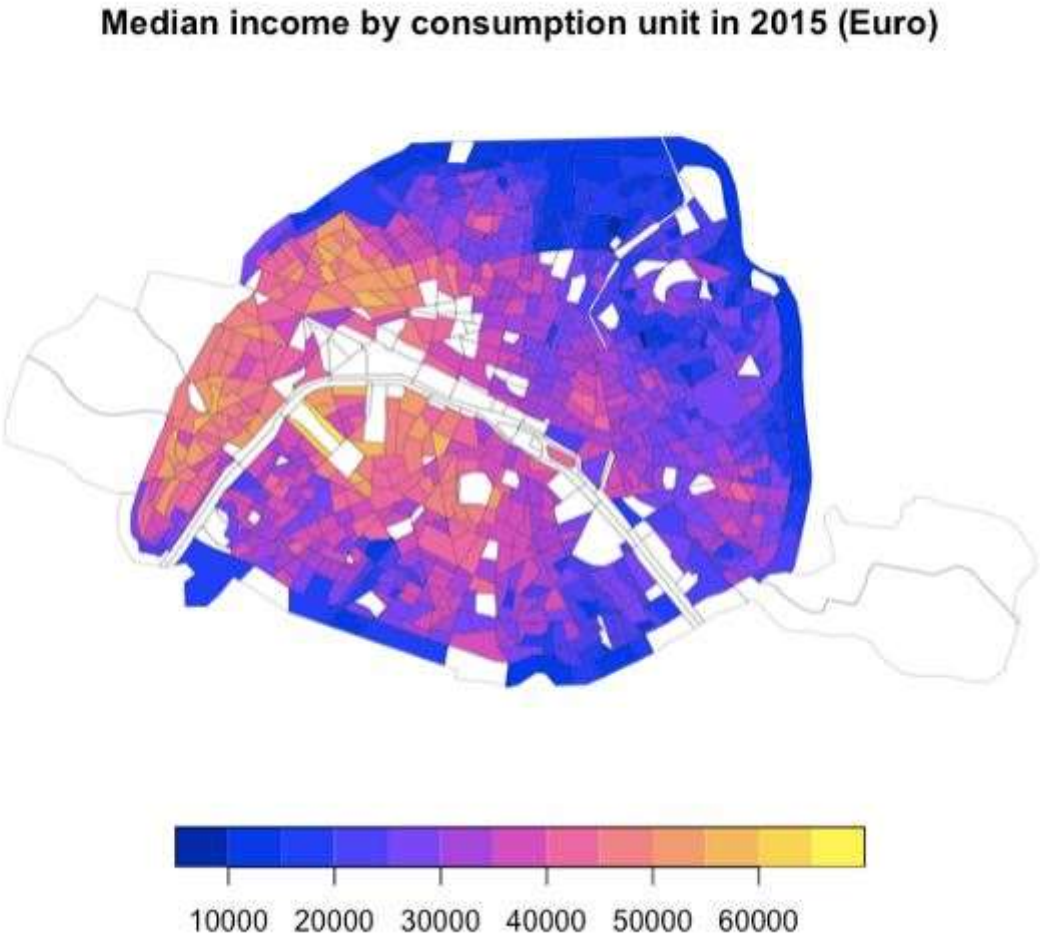
The 4 clusters allow us to paint the picture that the 2 clusters nearest to the city centre of Paris are closely related, while the other 2 clusters are spread throughout the city and its borders. There is a few reasons why bike station availability is more in certain areas than the others:

1)On a geographical standpoint, it can be mentioned that bike station availability seems to be more prevalent in zones of higher elevation areas compared to the ones in the city center closer to the Seine River of Paris. Bikes are being used a lot more often in areas of the city center as opposed to the outside areas. If you look closely at the map provided below, areas where elevation is higher bike station availability is highly concentrated in those clusters. However it cannot explain why bike station availability is present in the southern western region of Paris where elevation is not a factor.



Paris Elevation Map

2. On an income standpoint, another reason as to why bike station availability is prevalent in those areas is because they are in zones of low income compared those in the city centre. If you look at the map below (Could not find a recent map of after 2015), Low median income areas are in the north and north eastern areas of Paris where bikes are highly available. Whereas in city center of Paris bikes are being used a lot more often. Bike station availability is still unexplained in the south western region of Paris



Paris Median Income 2015

I think going forward, It would be interesting to know other factors that influence bike station availability in the south western region on Paris. This could possibly be due to other reasons perhaps different preferences in mode of transportation