fmlfinalproject2

Aksa Taniya

2022-12-13

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

First CSV file and Required Packages are loaded

In this project, we use the k-means clustering technique to do a non-hierarchical cluster analysis. The goal here is to divide the data into homogeneous clusters from which we may extract meaningful information. Let's start by loading the required packages and the original data set.

```
#packages are loaded
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.2
```

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

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
## Loading required package: lattice
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.2
```

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

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.3     v forcats 0.5.2
```

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.2.2
```

```
#Reading the dataset
library(readr)
fuel_receipts_costs_eia923 <-read.csv("fuel_receipts_costs_eia923.csv")
head(fuel_receipts_costs_eia923)
colMeans(is.na(fuel_receipts_costs_eia923))
  View(fuel_receipts_costs_eia923)

  library(dplyr)
fuel_receipts_costs_numerical<-fuel_receipts_costs_eia923[,c(1,2,12,13,15,16,17,18,20)]

fuel_receipts_costs_numerical<-sample_n(fuel_receipts_costs_numerical,12000)

Na<- fuel_receipts_costs_numerical %>% replace(.=="",NA)
new<-na.omit(Na)
```

1)Using only the numerical variables (1,2,12,13,15,16,17,18,20) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

Focusing on a subset of the original data set that it only contains numerical variables for the first part of the assignment.

Normalizing and Clustering the data

I compute the distance between each observation in this part. Because the Euclidean distance metric is utilized by default and is scale sensitive, data must first be modified.

```
#normalizing data
norm.fuel_receipts_costs_eia923 <- scale(new)
```

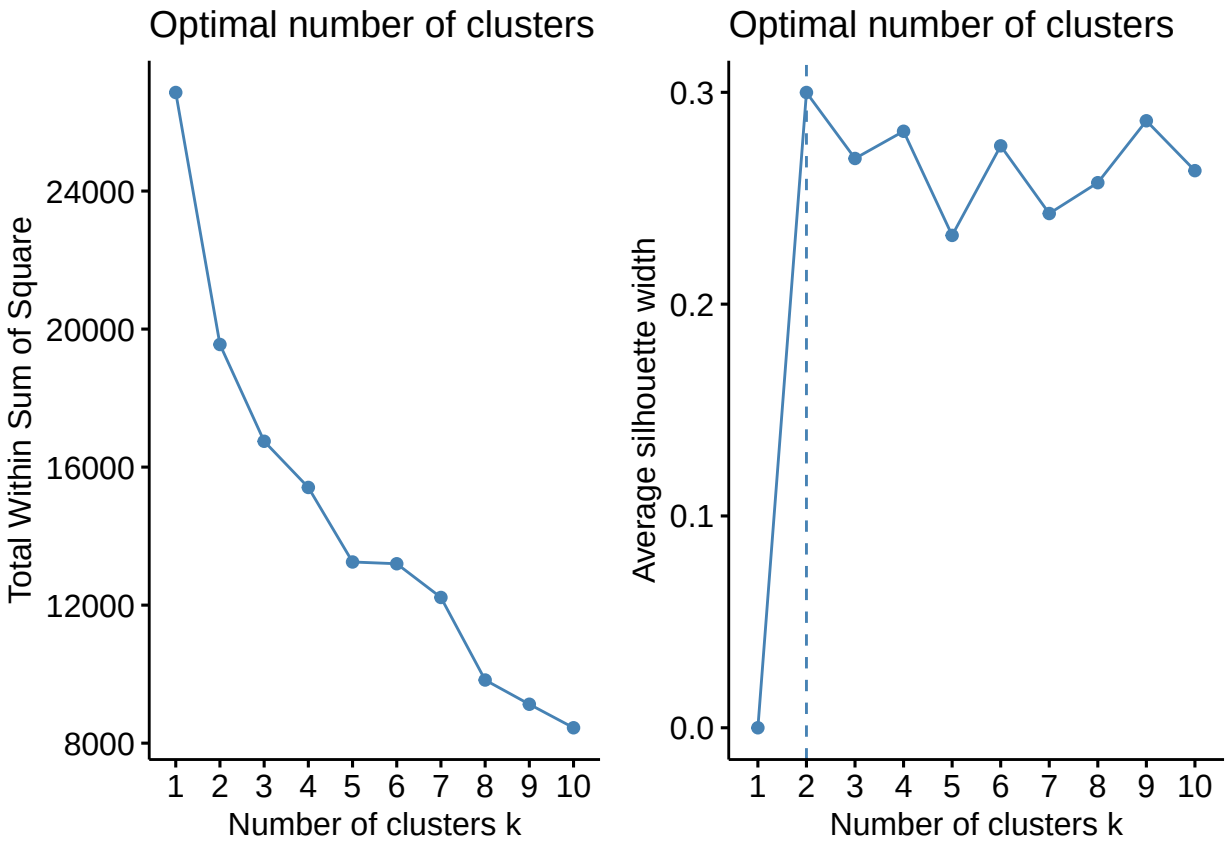
The graph depicts how intensity of color varies with distance. The diagonal, as we would predict, has a value of zero since it represents the distance between two observations.

##Finding the optimal K value

When there are no external factors, the Elbow chart and the Silhouette Method are two of the most effective approaches for calculating the number of clusters for the k-means model. The former shows how cluster heterogeneity decreases when more clusters are introduced. The latter compares an object's similarity to its cluster to the other clusters.

```
#Using elbow chart and silhouette method
WSS <- fviz_nbclust(norm.fuel_receipts_costs_eia923, kmeans, method = "wss")

Silho <- fviz_nbclust(norm.fuel_receipts_costs_eia923, kmeans, method = "silhouette")
plot_grid(WSS, Silho)
```



The plotted charts show that in the elbow method line occurs when k=2. I am using the k-means method with k=2.

```
#using k-means with k=2 for making clusters
set.seed(123)
KMeans_model <- kmeans(norm.fuel_receipts_costs_eia923, centers = 2)
KMeans_model$centers
```

```
##            X plant_id_eia mine_id_pudl mine_id_pudl_label fuel_received_units
## 1  0.1455741    0.06228530   -0.6193534         -0.6193534            0.3761962
## 2 -0.1038740  -0.04444353    0.4419382          0.4419382           -0.2684339
##   fuel_mmbtu_per_unit sulfur_content_pct ash_content_pct fuel_cost_per_mmbtu
## 1          -0.9858866         -0.7843499       -0.715365          -0.6291658
## 2           0.7034771          0.5596710        0.510447           0.4489398
```

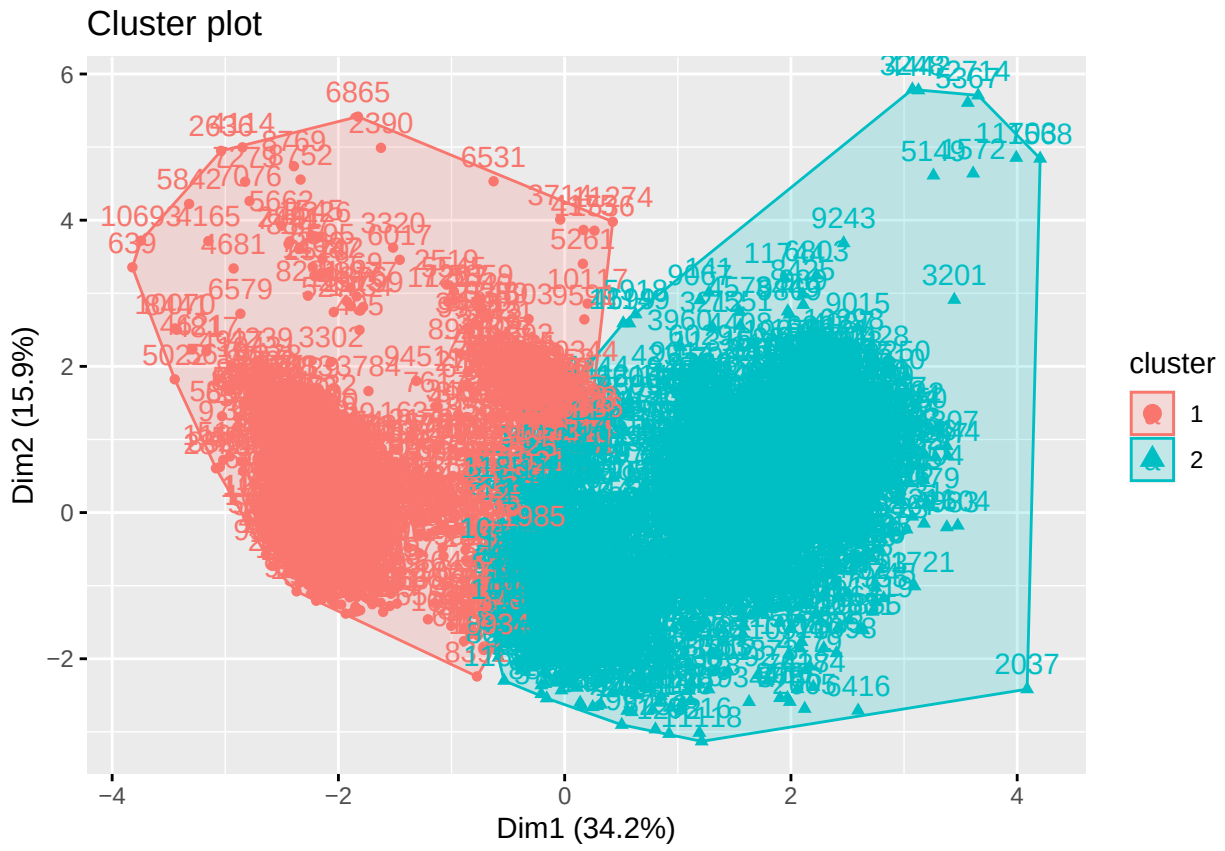
```
KMeans_model$size
```

```
## [1] 1243 1742
```

```
KMeans_model$withinss
```

```
## [1]  7533.019 12020.342
```

```
fviz_cluster(KMeans_model, data = norm.fuel_receipts_costs_eia923)
```




```
mean_k3<- new %>% mutate(Clusters=KMeans_model$cluster) %>% group_by(Clusters) %>% summarise_all("mean")

View(mean_k3)
```