Final Project

Yan Gao

19 March, 2019

Data loading and formatting

I load the data from JSE data set and formatting the data from fixed width table txt file.

```
rm(list=ls())
lines <- readLines("http://jse.amstat.org/datasets/airport.dat.txt")

w <- list(c(1,21), c(22,43), c(44,49), c(51,56), c(58,65), c(67,75), c(77,85))
ns <- c('Airport','City','Scheduled_departures','Performed_departures','Enplaned_passengers','Enplaned_
for(i in 1:length(w)) {
     assign(ns[i], str_trim(substring(lines, w[[i]][1], w[[i]][2])))
}
obj.list <- lapply(ns, get)
names(obj.list) <- ns
dat <- data.frame(obj.list)
rm(Airport,City,Scheduled_departures,Performed_departures,Enplaned_passengers,Enplaned_revenue_tons_of_
```

This is a very standard formatted data frame. No missing value or "dirty" value exist. The only thing we need to modify is the data type of each column.

```
dat[,1:2] <- as.character(unlist(dat[,1:2]))
dat[,3:5] <- as.integer(as.character(unlist(dat[,3:5])))
dat[,6:7] <- as.numeric(as.character(unlist(dat[,6:7])))
```

Data source and definitions

SOURCE: U.S. Federal Aviation Administration and Research and Special Programs Administration, 'Airport Activity Statistics' (1990). I find it on http://jse.amstat.org/datasets/airport.dat.txt .

This data set consists of all 135 large and medium sized air hubs in the United States as defined by the Federal Aviation Administration. And it consists with 7 columns:

(1) Airport: The name of the corresponding air hubs, which is unique for each row.
(2) City: The name of city which the air hub belongs to.
(3) Scheduled_departures: The number of scheduled departure flights at the air hub.
(4) Performed_departures: The number of performed departure flights at the air hub.
(5) Enplaned_passengers: The number of enplaned_passengers, which is the most important air traffic metric because the majority of airport revenues are generated directly or indirectly from enplaned passengers.
(6) Enplaned_revenue_tons_of_freight: The number of revenue tons of freight loaded on all aircrafts in the air hub including originating and transfer tons.
(7) Enplaned_revenue_tons_of_mail: The number of revenue tons of mail loaded on all aircrafts in the air hub including originating and transfer tons.

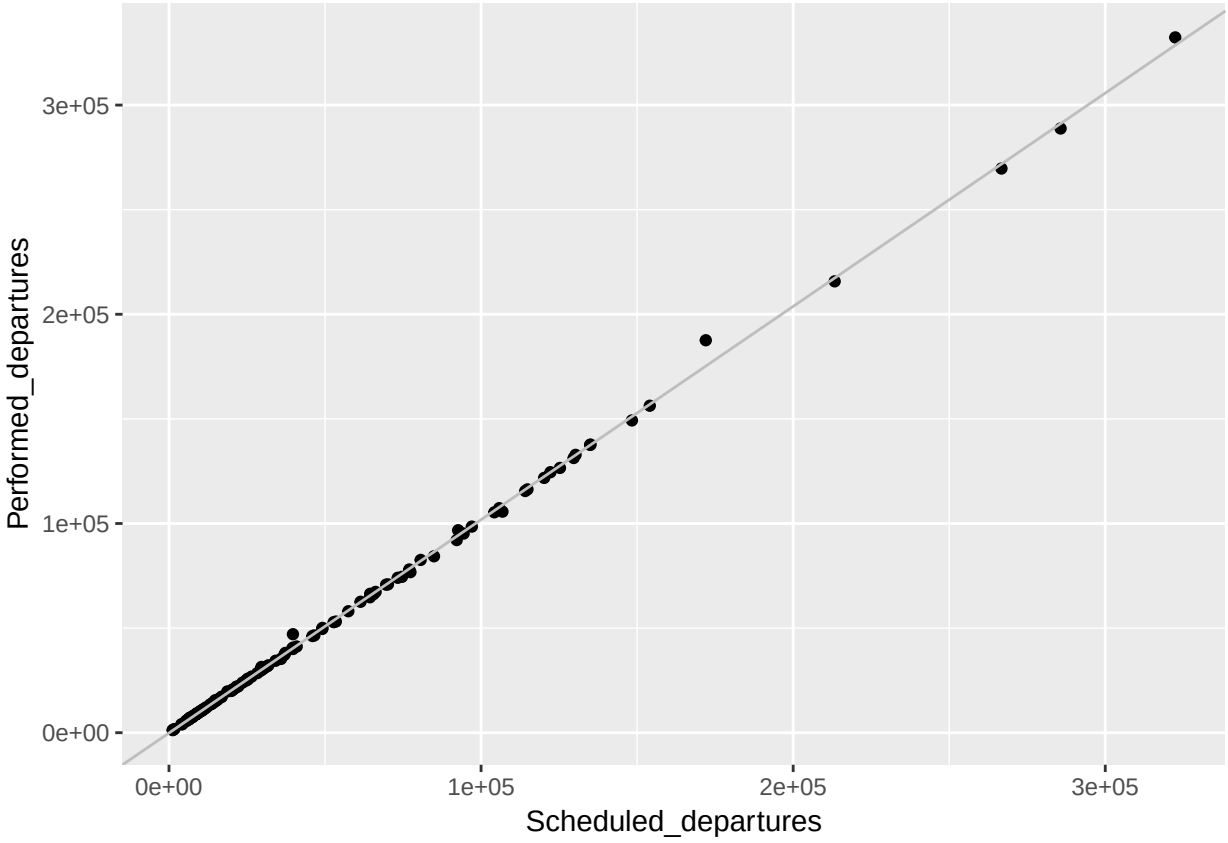
A brief description from the data contributor is in link: http://jse.amstat.org/datasets/airport.txt .

Main features of dataset

Some features of the dataset are visulized as an exploration of the data.

Visual check of Scheduled_departures and Performed_departures

```
lire <- lm(dat$Performed_departures ~ dat$Scheduled_departures)
g <- ggplot(data=dat,aes(x=Scheduled_departures,y=Performed_departures)) + geom_point() +
  geom_abline(intercept = lire$coefficients[1], slope = lire$coefficients[2], colour = 'gray')
g
```




```
summary(lire)
```

```
##
## Call:
## lm(formula = dat$Performed_departures ~ dat$Scheduled_departures)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3145.7  -308.2    38.2   133.7 12355.8
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.413e+02  1.596e+02  -0.886    0.377
## dat$Scheduled_departures  1.020e+00  2.203e-03 462.750   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1439 on 133 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9994
## F-statistic: 2.141e+05 on 1 and 133 DF,  p-value: < 2.2e-16
```

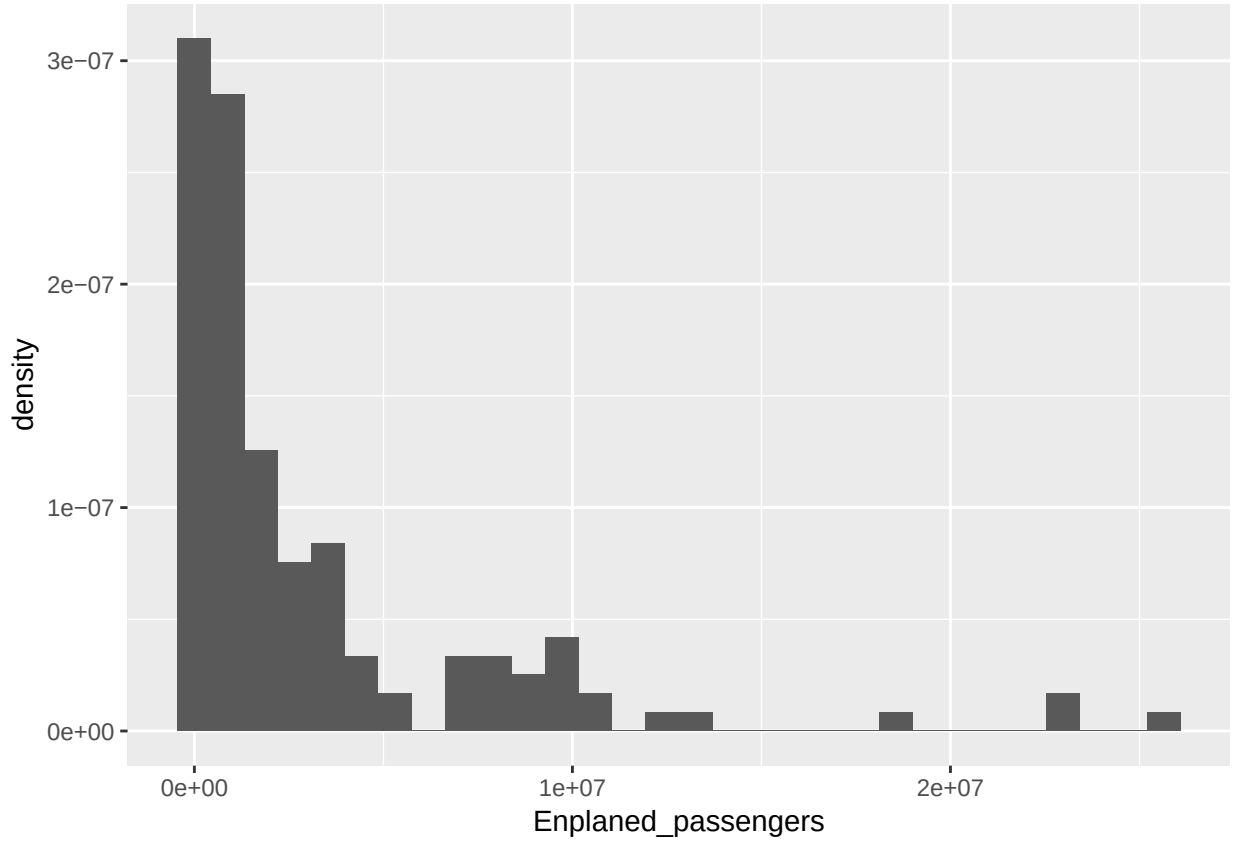
The relation between the Performed_departures and the Scheduled_departures is pretty linear. There seems no intervention like climates, geography or other factors could affect that dramatically. A slightly systematic decrease could exist. They are similar distributions with a bit shift.

The dustribution of Enplaned_passengers

As described in column definitions, the enplaned passengers is the most important air traffic metric. I want to look at its distribution to find if there is any pattern indicates that how many clusters should we make.

```
ggplot(data = dat, aes(x=Enplaned_passengers)) + geom_histogram(aes(y=..density..))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the plot, I can tell that if the Enplaned_passengers column is used to cluster the air hub sets of large sized and medium size, it's likely to find some boundary with enough distance to achieve that. But it seems more reasonable to cluster this into 3 parts, which could be called medium, medium-large and large.

Research Question

1. Explore if the Scheduled_departures and Performed_departures are the same distribution.

2. Explore if the data set could be clustered as two parts, large size and medium sized air hub, as the dataset claimed, or more reasonable clustering method could work.

Research Method

1. For question 1, The strong linearity with R-squared $0.9994 > 0.8$ shows that the two column should have the same distribution.I intend to use the Mann Whitney U test to check whether the distributions of two groups, Scheduled_departures and Performed_departures, have the same mean, or there exists a considerable amount of shift.
2. For question 2, I intend to use hierarchical clustering. Although k-means clustering runs quite faster than hierarchical clustering, it is quite random and could produces different result, which is vague to test appropriate parameters like the number of clusters. The hierarchical clustering is more clear to show the priority of different number clusters and which one is more reasonable.

Data satisfaction of requirements of method

1. It's obvious that the Scheduled_departures and Performed_departures groups data are quite similar. We can tell this from the plot in the Main Feature section. This indicates that the two groups should have same distribution. So we could apply the Mann Whitney U test here to check if they have the same mean.
2. The hierarchical clustering is not so sensitive with data set size. The most important thing is to determine the contribution of each columns to the distance of each data point. We should not easily just applied the Euclidean distance method as each columns may have different percentage of contributions. So a better way is to define the distance by ourselves.

Method application and interpration

The analysis of intervention between Scheduled_departures and Performed_departures

```
wilcox.test(dat$Scheduled_departures,dat$Performed_departures)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dat$Scheduled_departures and dat$Performed_departures
## W = 8993.5, p-value = 0.8535
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is much greater than 0.05. So the difference between the two distributions' means is statistically insignificant.

The analysis of Hierarchical Clustering

Before I implement the hierarchical clustering, we have to merge the influence of some columns which have the same property.

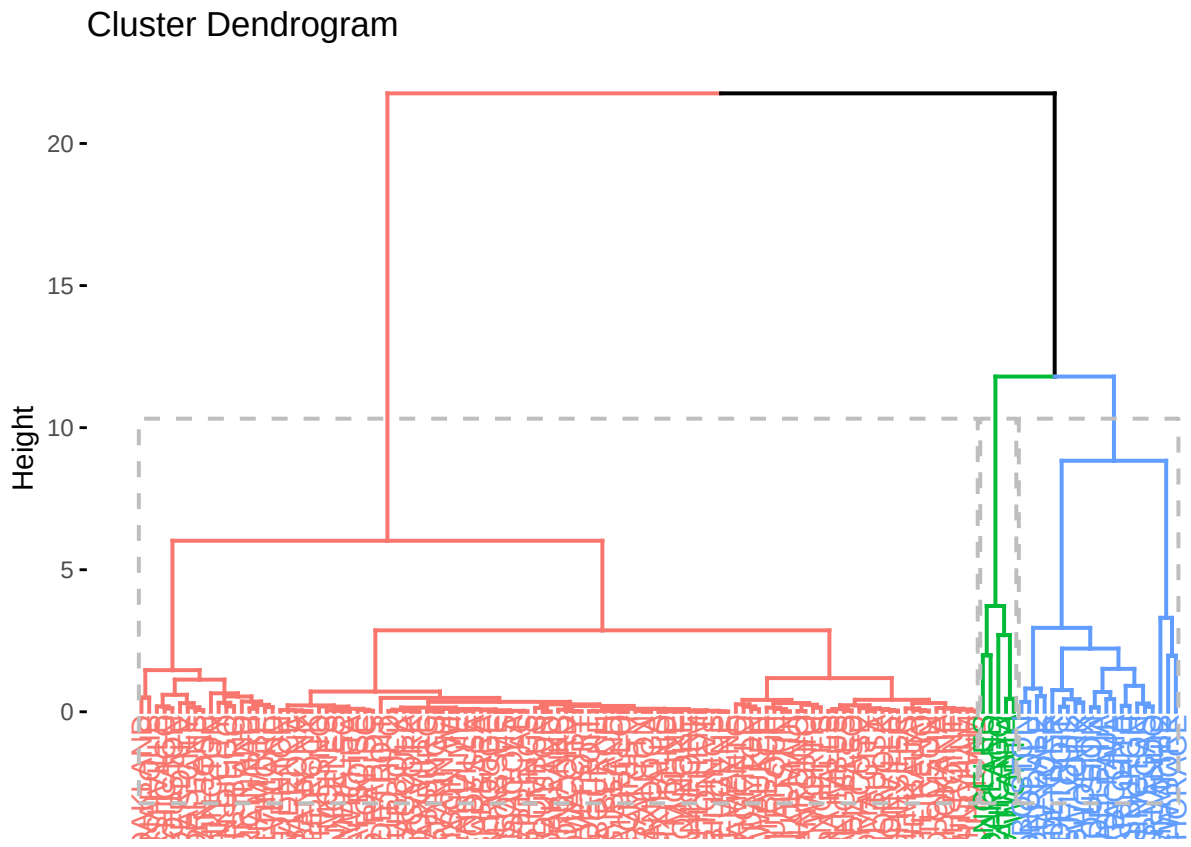
1. The Scheduled_departures and Performed_departures have exactly same distribution, so I only choose the later column to avoid double counting.
2. The Enplaned_passengers, as told before, is the most important metric for scale the size of the air hub and it has no relation with other columns. So I keep it alone.

3. Both the Enplaned_revenue_tons_of_freight and the Enplaned_revenue_tons_of_mail are the revenue been charged by tons. So I add them together to take the total revenue by tons into count.
4. There are same airport names in the Airport column, INTERNATIONAL for example. To treat each airport as unique, I add each city it belongs to to the airport name and set it as the row name.

After building a new dataframe, it turns to determine each column's weight. I haven't found any clues of how to determine the weight of each column. So I just treat them as same scale and use the default Euclidean distance. There should be a way to count the weight. I used to get it by using fuzzy clustering in MATLAB.

```
dat.cluster <- data.frame( dat$Performed_departures )
dat.cluster$Enplaned_passengers <- dat$Enplaned_passengers
dat.cluster$Enplaned_revenue_tons <- dat$Enplaned_revenue_tons_of_freight + dat$Enplaned_revenue_tons_o
row.names(dat.cluster) <- paste(dat$Airport,dat$City)

df <- scale(dat.cluster) # normalize data
res.hc <- eclust(df, "hclust",k = 3) # compute hclust
fviz_dend(res.hc, rect = TRUE) # dendrogam
```




```
l1 <- res.hc$cluster[res.hc$cluster == 1]
names(l1)
```

```
## [1] "HARTSFIELD INTL ATLANTA"
## [2] "O'HARE INTL CHICAGO"
## [3] "DALLAS/FT WORTH INTL DALLAS/FT WORTH"
## [4] "LOS ANGELES INTL LOS ANGELES"
## [5] "SAN FRANCISCO INTL SAN FRANCISCO/OAKLAND"
```

The results seem really good. The Cluster Dendrogram shows that, by dividing the air hubs into three clusters, how would each air hub belongs to different clusters. Two clusters method is also acceptable (check this in Appendix section). But the distribution feature of the Enplaned_passengers in Main Features section shows that the clustering could be more informational with 3 clusters.

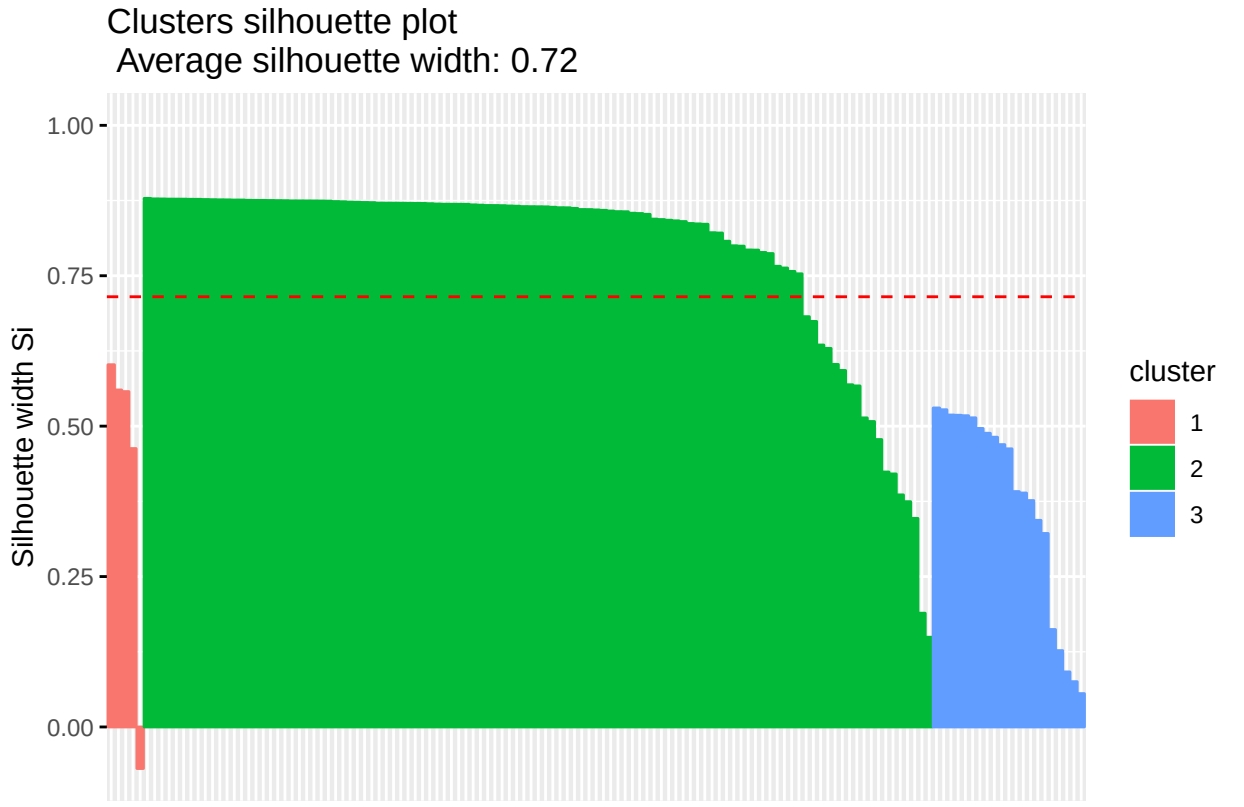
To confirm that, I print the airport names in the top cluster result. It's easily to check that they are still the biggest airports in USA. Almost all the international travelers will pass one of the airports among them. They are more outstanding than other airports of same cluster in the two clusters implementation. It even matches the result I find on the website (httpd://.market watch.com/story/this-was-the-busiest-us-airport-in-2018-2019-02-04). This website has the latest result in 2018. There is no surprise that SAN FRANCISCO INTL SAN FRANCISCO/OAKLAND in my 1990 dataset is replaced by Denver International Airport. Others are just the same.

Appendix

The Silhouette plot and scatter plot to scale how well the clustering works

```
fviz_silhouette(res.hc)
```

```
##   cluster size ave.sil.width
## 1       1    5          0.42
## 2       2  109          0.79
## 3       3   21          0.37
```




```
fviz_cluster(res.hc,repel = T,geom = 'point') # scatter plot
```

The rest clustring results

```
(l2 <- res.hc$cluster[res.hc$cluster == 2])
```

```
##                   BALTO/WASH INTL BALTIMORE
##                                           2
##                               MIDWAY CHICAGO
##                                           2
##                  LOVE FIELD DALLAS/FT WORTH
##                                           2
##                        DETROIT CITY DETROIT
##                                           2
##                          WILLOW RUN DETROIT
##                                           2
##                                HOBBY HOUSTON
##                                           2
##                     ELLINGTON FIELD HOUSTON
##                                           2
##                HOLLYWOOD-BURBANK LOS ANGELES
##                                           2
##                     LONG BEACH LOS ANGELES
##                                           2
##                   ORANGE COUNTY LOS ANGELES
##                                           2
##      FT LAUDERDALE INTL MIAMI/FT LAUDERDALE
```

```
##                                          2
##          SALT LAKE CITY INTL SALT LAKE CITY
##                                          2
##                    SAN DIEGO INTL SAN DIEGO
##                                          2
##        BUCHANAN FIELD SAN FRANCISCO/OAKLAND
##                                          2
##    OAKLAND METRO INTL SAN FRANCISCO/OAKLAND
##                                          2
##              TAMPA INTL TAMPA/ST PETERSBURG
##                                          2
##                      DULLES INTL WASHINGTON
##                                          2
##                ALBUQUERQUE INTL ALBUQUERQUE
##                                          2
##                         MUELLER MUNI AUSTIN
##                                          2
##  GREATER BUFFALO INTL BUFFALO/NIAGARA FALLS
##                                          2
##               GREATER CINCINNATI CINCINNATI
##                                          2
##                      HOPKINS INTL CLEVELAND
##                                          2
##                PORT COLUMBUS INTL COLUMBUS
##                                          2
##                       LOCKBURN AFB COLUMBUS
##                                          2
##                      COX/DAYTON INTL DAYTON
##                                          2
##                        EL PASO INTL EL PASO
##                                          2
##                        SOUTHWEST FORT MYERS
##                                          2
##           BRADLEY INTL HARTFORD/SPRINGFIELD
##                                          2
##              INDIANAPOLIS INTL INDIANAPOLIS
##                                          2
##              JACKSONVILLE INTL JACKSONVILLE
##                                          2
##                                KAHULUI MAUI
##                                          2
##                  INTERNATIONAL KANSAS CITY
##                                          2
##                                LIHUE KAUAI
##                                          2
##                 GENERAL MITCHELL MILWAUKEE
##                                          2
##                     METROPOLITAN NASHVILLE
##                                          2
##              INTL/MOISANT FIELD NEW ORLEANS
##                                          2
##                    NORFOLK REGIONAL NORFOLK
##                                          2
##             WILL ROGERS WORLD OKLAHOMA CITY
```

```
##                                         2
##         ONTARIO INTL ONTARIO/SAN BERNARDINO
##                                         2
##                      PORTLAND INTL PORTLAND
##                                         2
##           RALEIGH-DURHAM RALEIGH/DURHAM/RTP
##                                         2
##                               RENO INTL RENO
##                                         2
##              ROCHESTER-MONROE CTY ROCHESTER
##                                         2
##                 SACRAMENTO METRO SACRAMENTO
##                                         2
##                SAN ANTONIO INTL SAN ANTONIO
##                                         2
##                      SAN JOSE MUNI SAN JOSE
##                                         2
##                             HANCOCK SYRACUSE
##                                         2
##                           TUCSON INTL TUCSON
##                                         2
##                             TULSA INTL TULSA
##                                         2
##             PALM BEACH INTL WEST PALM BEACH
##                                         2
##                   AKRON-CANTON AKRON/CANTON
##                                         2
##                    ALBANY COUNTY ALBANY, NY
##                                         2
##      ALLENTOWN-BETHEHEM ALLENTOWN/BETHLEHEM
##                                         2
##                            AMARILLO AMARILLO
##                                         2
##                             RYAN BATON ROUGE
##                                         2
##                         LOGAN FIELD BILLINGS
##                                         2
##                  BIRMINGHAM MUNI BIRMINGHAM
##                                         2
##                    BOISE AIR TERMINAL BOISE
##                                         2
##            HARLINGEN INDUSTRIAL BROWNSVILLE
##                                         2
##              BURLINGTON INTL BURLINGTON, VT
##                                         2
##    CEDAR RAPIDS MUNI CEDAR RAPIDS/IOWA CITY
##                                         2
##          CHARLESTON AFB/MUNI CHARLESTON, SC
##                                         2
##                    LOVELL FIELD CHATTANOOGA
##                                         2
##             PETERSON FIELD COLORADO SPRINGS
##                                         2
##                 COLUMBIA METRO COLUMBIA, SC
```

```
##                                       2
##          CORPUS CHRISTI INTL CORPUS CHRISTI
##                                       2
##             DAYTONA BEACH REG DAYTONA BEACH
##                                       2
##                 DES MOINES MUNI DES MOINES
##                                       2
##                  MAHLON SWEET FIELD EUGENE
##                                       2
##                   FAIRBANKS INTL FAIRBANKS
##                                       2
##                 MUNI/BAER FIELD FORT WAYNE
##                                       2
##                 FRESNO AIR TERMINAL FRESNO
##                                       2
##                   KENT COUNTY GRAND RAPIDS
##                                       2
##  GREENSBORO-HP-WS REG GREENSBORO/HIGH POINT
##                                       2
## GREENVILLE/SPARTANBG GREENVILLE/SPARTANBURG
##                                       2
##             HARRISBURG INTL HARRISBURG/YORK
##                                       2
##                   GENERAL LYMAN FIELD HILO
##                                       2
##               MADISON COUNTY HUNTSVILLE, AL
##                                       2
##        PALM SPRINGS MUNI INDIO/PALM SPRINGS
##                                       2
##     LONG ISLAND-MCARTHUR ISLIP, LONG ISLAND
##                                       2
##                 THOMPSON FIELD JACKSON, MS
##                                       2
##                       KE-AHOLE KAILUA-KONA
##                                       2
##                    MCGHEE TYSON KNOXVILLE
##                                       2
##              BLUE GRASS LEXINGTON/FRANKFORT
##                                       2
##                   ADAMS FIELD LITTLE ROCK
##                                       2
##                STANDIFORD FIELD LOUISVILLE
##                                       2
##                   LUBBOCK REGIONAL LUBBOCK
##                                       2
##                    TRUAX FIELD MADISON, WI
##                                       2
##                MUNICIPAL MANCHESTER/CONCORD
##                                       2
##                  CAPE KENNEDY REG MELBOURNE
##                                       2
##             MIDLAND REGIONAL MIDLAND/ODESSA
##                                       2
##               BATES FIELD MOBILE/PASCAGOULA
```

```
##                                       2
##                           QUAD CITY MOLINE
##                                       2
##                        EPPLEY AIRFIELD OMAHA
##                                       2
##                  PENSACOLA REGIONAL PENSACOLA
##                                       2
##            PORTLAND INTL JETPRT PORTLAND, ME
##                                       2
##               FRANCIS GREEN STATE PROVIDENCE
##                                       2
##                    BYRD FLYING FIELD RICHMOND
##                                       2
##                         ROANOKE MUNI ROANOKE
##                                       2
##                    TRI CITY SAGINAW/BAY CITY
##                                       2
##                  SANTA BARBARA SANTA BARBARA
##                                       2
##       SARASOTA-BRADENTON SARASOTA/BRADENTON
##                                       2
##                       SAVANNAH INTL SAVANNAH
##                                       2
##                SHREVEPORT REGIONAL SHREVEPORT
##                                       2
##                       FOSS FIELD SIOUX FALLS
##                                       2
##                  MICHIANA REGIONAL SOUTH BEND
##                                       2
##                         SPOKANE INTL SPOKANE
##                                       2
##              TALLAHASSEE REGIONAL TALLAHASSEE
##                                       2
##                        MID-CONTINENT WICHITA
##                                       2
```

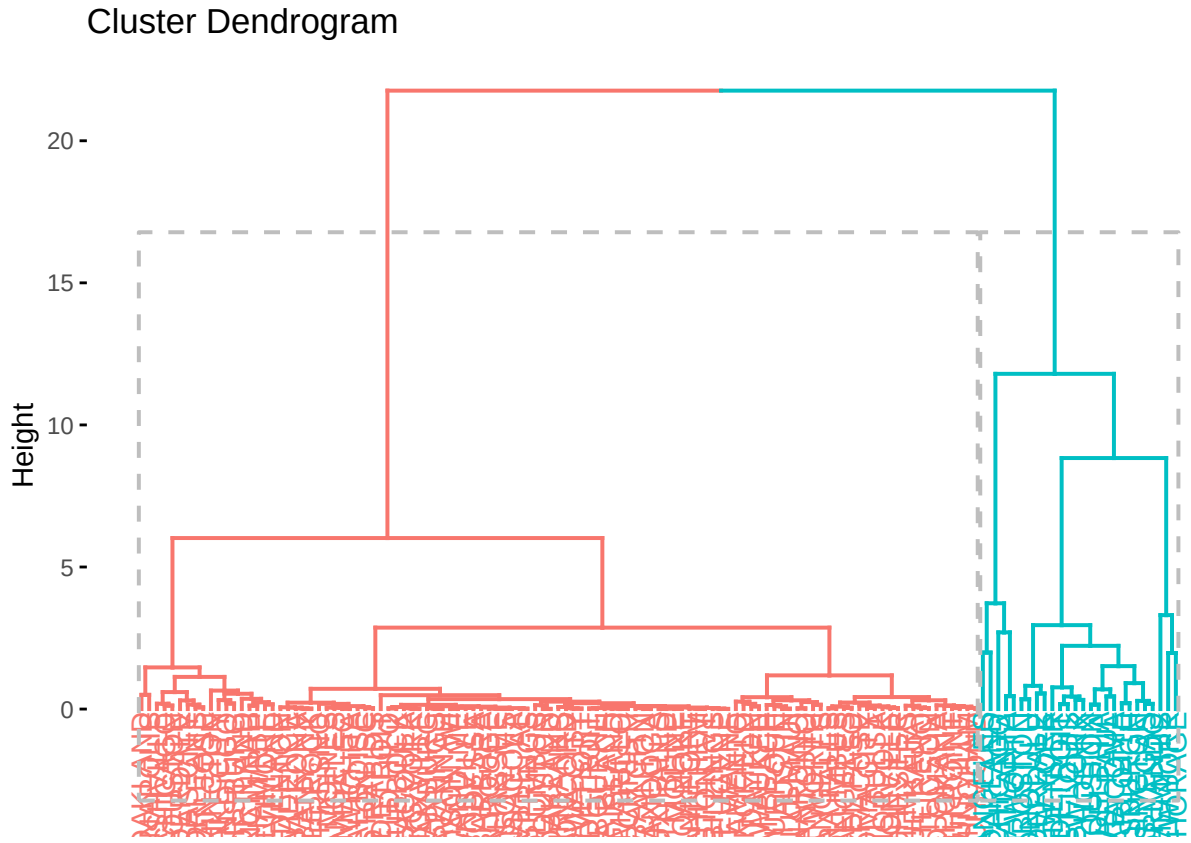
```
(l3 <- res.hc$cluster[res.hc$cluster == 3])
```

```
##                         LOGAN INTL BOSTON
##                                       3
##                    DOUGLAS MUNI CHARLOTTE
##                                       3
##                     STAPLETON INTL DENVER
##                                       3
##                      WAYNE COUNTY DETROIT
##                                       3
##                    HONOLULU INTL HONOLULU
##                                       3
##                  INTERCONTINENTAL HOUSTON
##                                       3
##                  MC CARRAN INTL LAS VEGAS
##                                       3
##            MIAMI INTL MIAMI/FT LAUDERDALE
##                                       3
## MINNEAPOLIS/ST PAUL MINNEAPOLIS/ST PAUL
```

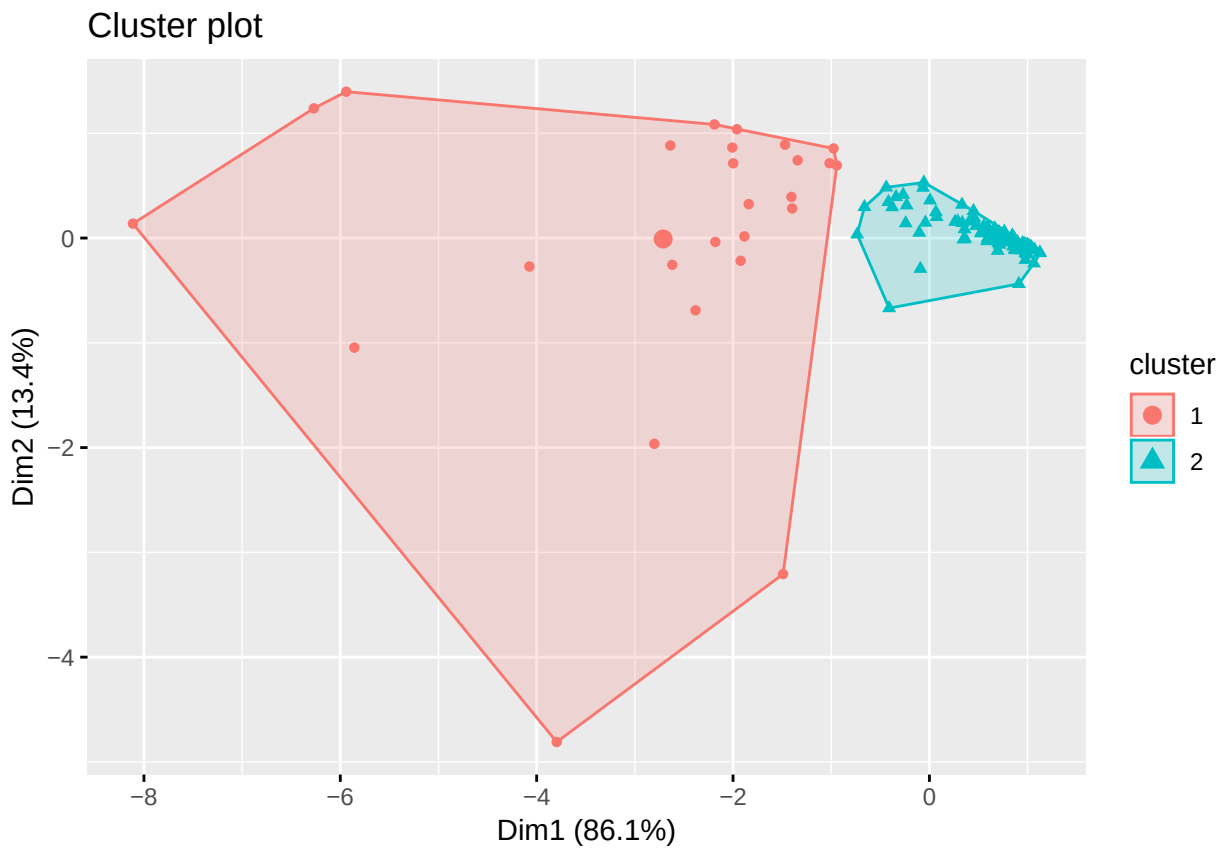
```
##                                     3
##                   NEWARK NEWARK/NEW YORK
##                                     3
##             JOHN F KENNEDY INTL NEW YORK
##                                     3
##                      LA GUARDIA NEW YORK
##                                     3
##                     ORLANDO INTL ORLANDO
##                                     3
##               INTERNATIONAL PHILADELPHIA
##                                     3
##                  SKY HARBOR INTL PHOENIX
##                                     3
##            GREATER PITTSBURGH PITTSBURGH
##                                     3
##                LAMBERT-ST LOUIS ST LOUIS
##                                     3
##       SEATTLE-TACOMA INTL SEATTLE/TACOMA
##                                     3
##           WASHINGTON NATIONAL WASHINGTON
##                                     3
##                 ANCHORAGE INTL ANCHORAGE
##                                     3
##                     MEMPHIS INTL MEMPHIS
##                                     3
```

Two clusters' Hierarchical Clustering

```
res.hc <- eclust(df, "hclust",k = 2) # compute hclust
fviz_dend(res.hc, rect = TRUE) # dendrogam
```

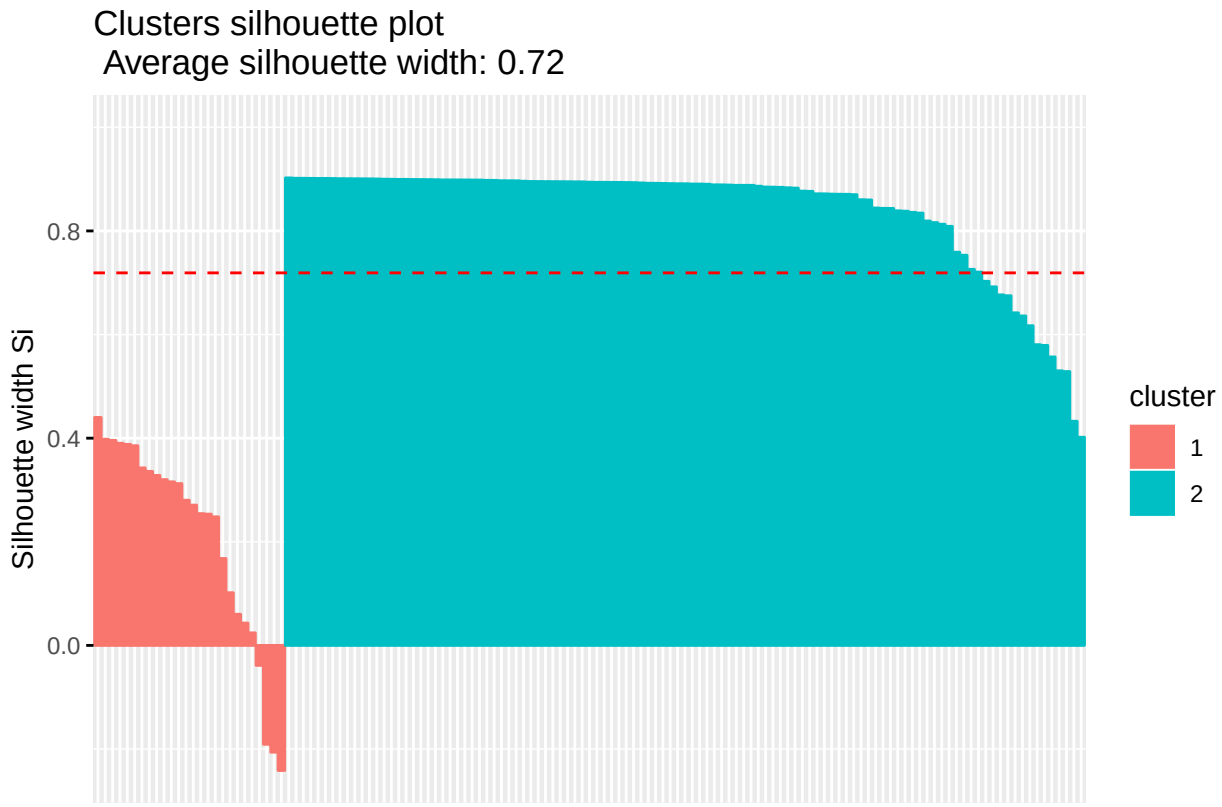



```
fviz_cluster(res.hc,repel = T,geom = 'point') # scatter plot
```




```
fviz_silhouette(res.hc) # silhouette plot
```

```
##   cluster size ave.sil.width
## 1       1   26          0.21
## 2       2  109          0.84
```


Clusters silhouette plot
Average silhouette width: 0.72
Silhouette width Si
0.8
0.4
0.0
cluster
1
2