FINAL PROJECT MATH 4800

Using K-means and Hierarchical clustering to weigh the severity of Covid-19 in Ontario.

SUPERVISOR – *WESLEY BURR*

INSTRUCTOR – *STEFAN BILANIUK*

BY:

NAME – *HIDAYA PATEL*

STUDENT ID – *0685595*

TABLE OF CONTENTS

INTRODUCTION

All of us have been affected by the current COVID-19 epidemic. Our reactions to the epidemic and its consequences, on the other hand, differ depending on its effects and severity on our community. Throughout human history, we have lacked the ability to recognize patterns resulting from massive volumes of data. We lacked the necessary expertise and computing capacity to fully use data. But now we have that computational capability. So the capability of analyzing big data can be used, as the COVID-19 pandemic spreads across continents, all countries impacted are implementing major public health measures. The rapid spread of COVID-19 has highlighted the need of studying how demographic characteristics interact with pandemics.

During the epidemic, statisticians have used the power of statistics to improve their efforts in battling and minimizing the effects of covid on society. One of the abilities is clustering tactics, which may be very useful during pandemic periods when there is a lot of anxiety about whether current health services will have the resources and facilities to deal with the epidemic. Clustering is a typical statistical data analysis approach. Within a data collection, data clustering is used to identify diverse groups of data, referred to as clusters.

A cluster is a group of data items that have been grouped together based on shared characteristics. Two types of clustering techniques I will be using in my project are K-Means clustering and Hierarchical clustering. The goal of K-means is to identify patterns by grouping related datapoints. K-means seeks out a predetermined number (k) of clusters in a dataset to

accomplish this goal. Hierarchical clustering requires constructing groups in a certain sequence, and itis divided into two categories: Agglomerative and Divisive. In the former, data points are clustered individually, whereas in the latter, all data points are considered as one huge cluster, and the clustering process entails splitting the one big cluster into multiple tiny clusters.

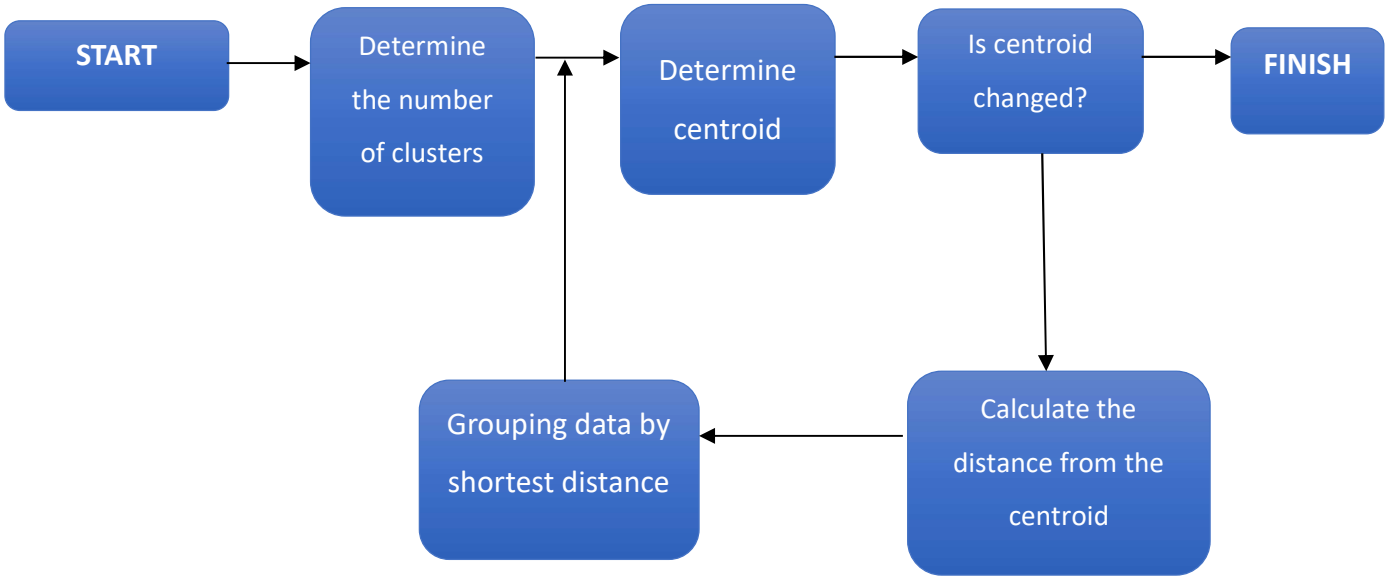
When these techniques are used to such situations, they assist us in categorizing and analyzing data that would otherwise look unprocessed. When these tactics are used to situations like these, they help us undertake a more complete study of the problem and produce solutions. On the COVID-19 Ontario cities dataset, I will use the clustering algorithms outlined above for this study. The information gained from this clustering application will be used to establish measures such as community activity restrictions or other regulations aimed at reducing the spread of COVID-19 in Ontario.

DATASET

- Retrieved data from Ontario Data catalogue.
- The original dataset link: - https://data.ontario.ca/dataset/f4112442-bdc8-45d2-be3c-12efae72fb27/resource/455fd63b-603d-4608-8216-7d8647f43350/download/conposcovidloc.csv
- I have used two sets of data for this project, first one with Toronto and the second one without Toronto, which is basically the modification of the first set.
- The first dataset has 34 rows and 4 columns
- The modified data set has 33 rows and 4 columns

K- MEANS CLUSTERING

- Two randomly chosen plots are the center points or AKA “centroids”.
- Then this process draws a line between the centroids and each of the plots.
- The plots that have the shortest line to the centroids are grouped together.
- After the initial cluster formation, the centroid for each cluster is re-calculated based on the location of the points that were assigned to it. It will be located at the place where the distance between all points and the centroid is at its lowest.
- Then the points get re-assigned to a new centroid based on the closest distance.
- The process starts all over again calculating the distance between the points in the plot and the new “centroids”.
- Then it relocated the centroid to the place in the plot where the sum of the new distances for the plots assigned to the “centroid” are the lowest (minimum).
- This process is repeatedly done until the location for the centroids does not change very much.



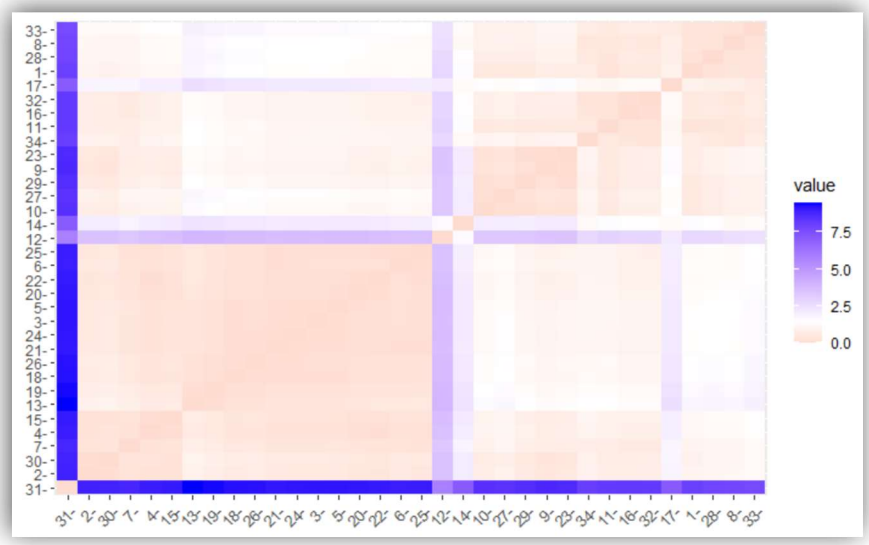
EUCLIDEAN DISTANCE

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

It is just a distance measure between a pair of samples p and q in an n-dimensional feature space. As you can imagine, there are many approaches to estimate the distance between observations. However, the most common is the Euclidean Distance. The Euclidean distance process determines the proximity between observations by drawing a straight line between pairs of observations. Therefore, this process measures the distance between observations by looking at the length of this line between observations.

DISTANCE MATRIX

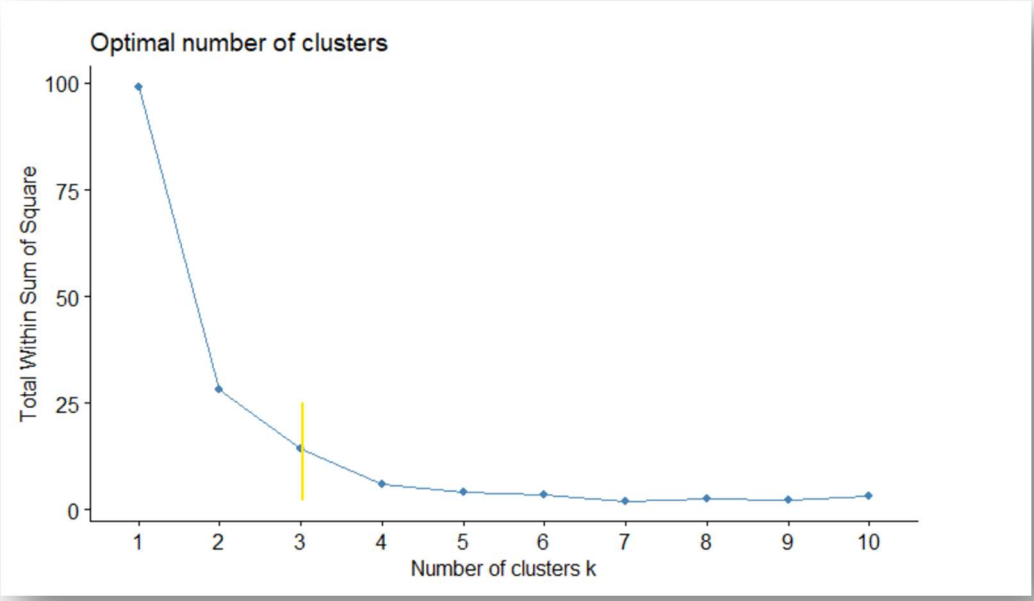
A distance matrix is a matrix that shows the distance between pairs of objects. For the dataset used, the distance matrix looks as follows. This matrix shows us the shortest distance between each observation from the dataset.



There are different methods to predict the number of clusters which are optimal. The two ways I have used are Elbow method and Silhouette method.

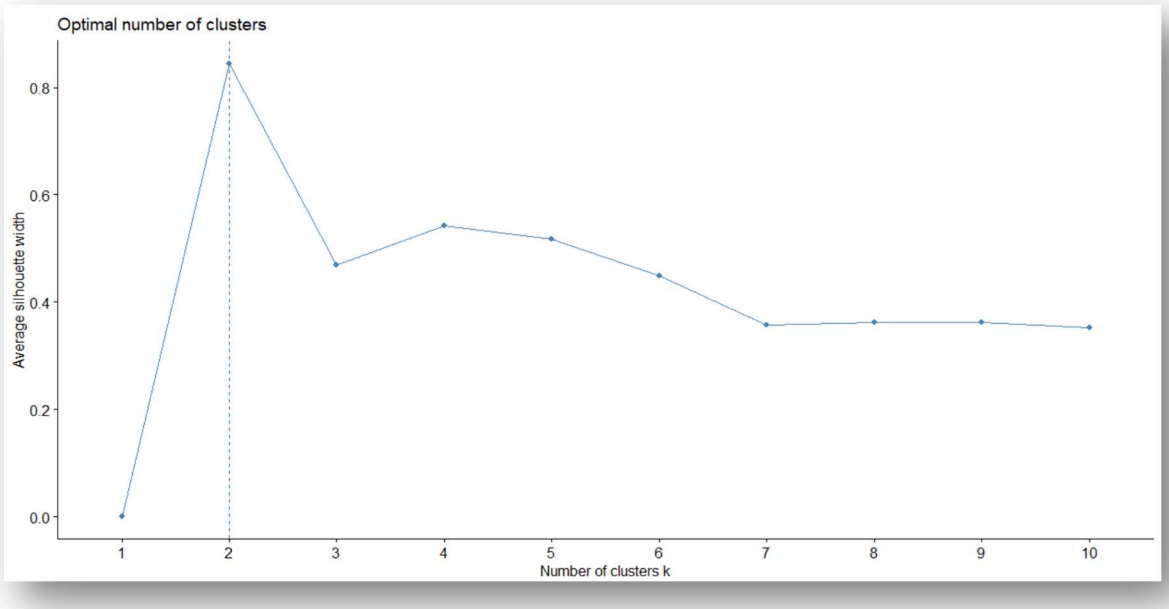
ELBOW METHOD

The elbow technique performs k-means clustering on the dataset for a range of k values (say, 1-10) and then computes an average score for all clusters for each value of k.



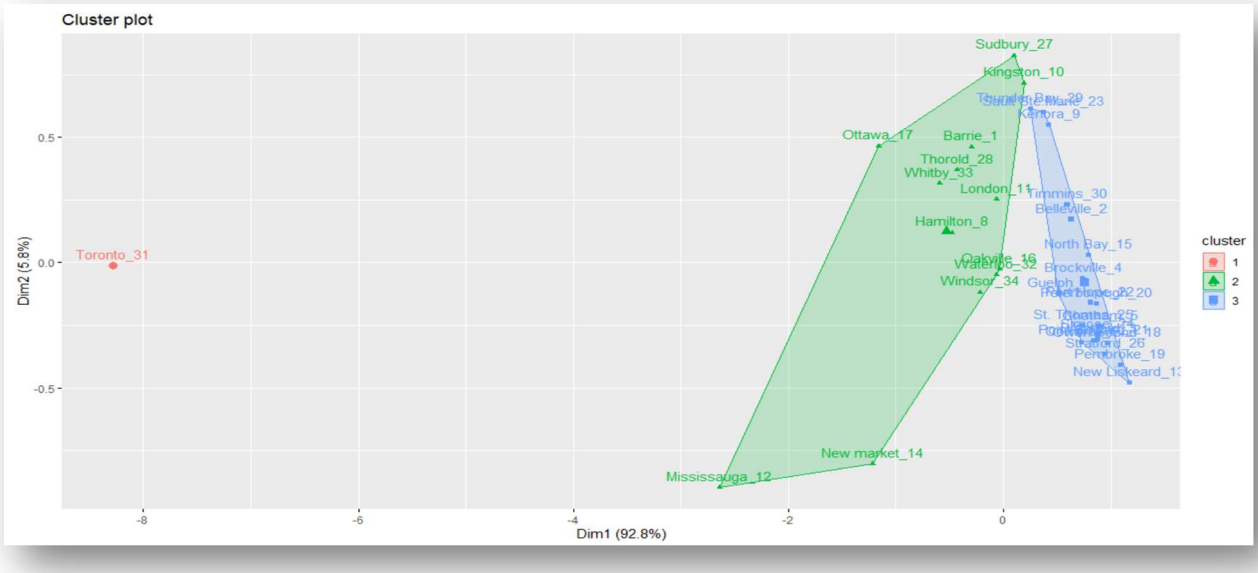
SILHOUETTE METHOD

The silhouette plot shows how near each point in one cluster is to points in surrounding clusters and so gives a visual way to examine factors such as cluster number.



From the above graphs the optimal number of clusters to be chosen is 3. Let's proceed by plotting the three clusters.

CLUSTER PLOT



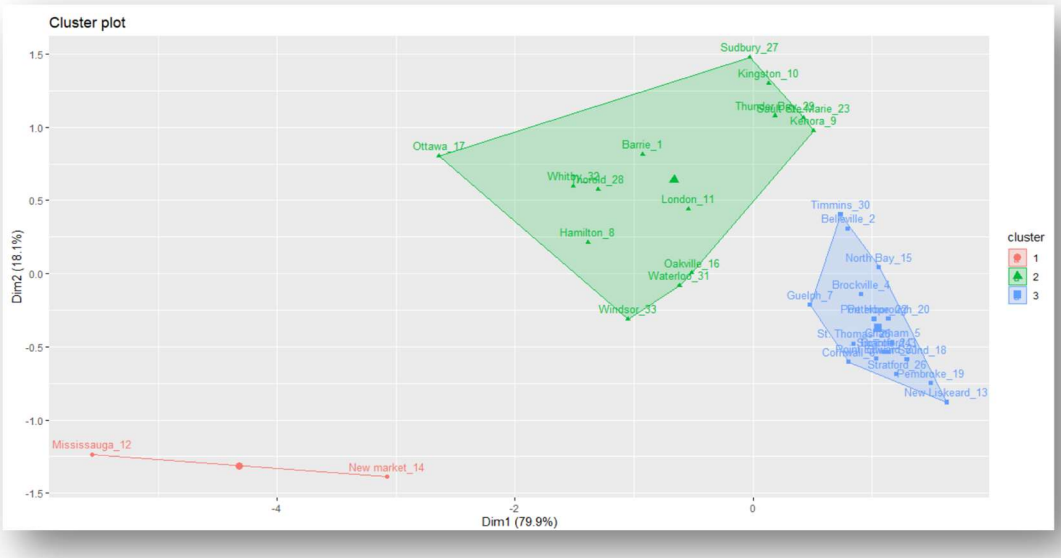
CLUSTER 1	CLUSTER 2	CLUSTER 3
Toronto	Barrie, New Market, Hamilton, Kingston, London, Mississauga, Oakville, Ottawa, Windsor, Sudbury, Thorold, Waterloo, Whitby	Timmins, Thunder Bay, Stratford, St. Thomas, Simcoe St., Sault Ste.Marie, Port Hope, Point Edward, Peterborough, Pembroke, Owen Sound, North Bay, Kenora, New Liskeard, Guelph, Cornwall, Chatham, Brockville, Brantford, Belleville

PLOT ANALYSIS

Toronto has the most cases and requires more laws and regulations to be implemented. It's also worth noting that Toronto is an anomaly aka outlier. Although Cluster 2 has fewer instances than Toronto, it would still need the government's strict enforcement of the rules. Cluster 3 has the fewest instances; therefore, the limits can be looser, but the tighter the limitations, the higher the chances of reducing cases. To obtain interesting results let's remove the outlier Toronto and perform the same technique again.

WITHOUT OUTLIER TORONTO

Number of clusters (k) = 3

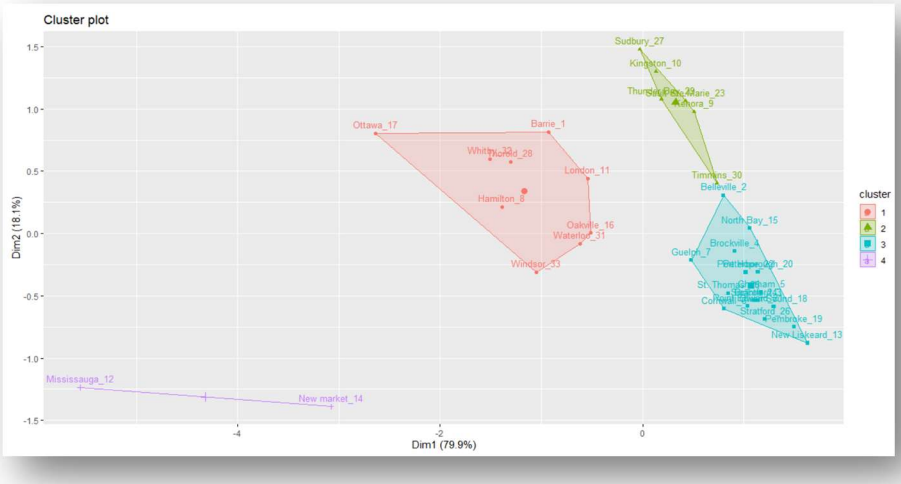


CLUSTER 1	CLUSTER 2	CLUSTER 3
Mississauga, New Market	Barrie, Hamilton, Kenora, Kingston, London, Oakville, Ottawa, Sault Ste.Marie, Windsor, Sudbury, Thorold, Waterloo, Whitby, Thunder Bay	Timmins, Stratford, St. Thomas, Simcoe St., Port Hope, Point Edward, Peterborough, Pembroke, Owen Sound, North Bay, New Liskeard, Guelph, Cornwall, Chatham, Brockville, Brantford, Belleville

PLOT ANALYSIS

We see some interesting results as removing Toronto makes Mississauga and New Market the next big Target. Which means tighter implementation of rules is required. Also, we can see that Mississauga and New market have similar number of cases which was not possible to see before due to the outlier. This why getting rid of the outlier is very important to extract accurate information. Had we not removed Toronto we would have assumed that Mississauga and New market have the similar case ratios as the other cities in Cluster 2. But they have slightly more cases than the rest.

Number of clusters (k) = 4



CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
Barrie, Hamilton, London, Oakville, Ottawa, Windsor, Thorold, Waterloo, Whitby	Kenora, Kingston, Sault Ste.Marie, Sudbury, Thunder Bay, Timmins	Stratford, St. Thomas, Simcoe St., Port Hope, Point Edward, Peterborough, Pembroke, Owen Sound, North Bay, New Liskeard, Guelph, Cornwall, Chatham, Brockville, Brantford, Belleville	Mississauga, New Market

PLOT ANALYSIS

An interesting plot to analyze! We can see here that a few cities from the previous clusters are pulled out and clustered in a new cluster. This plot gives us an insight to more cites which have similarity in comparison to the other plots. The cities from cluster 2 of the previous plot are broken down and clustered more deeply showing us which cities are more closely related in terms of Covid cases.

HIERARCHICAL CLUSTER

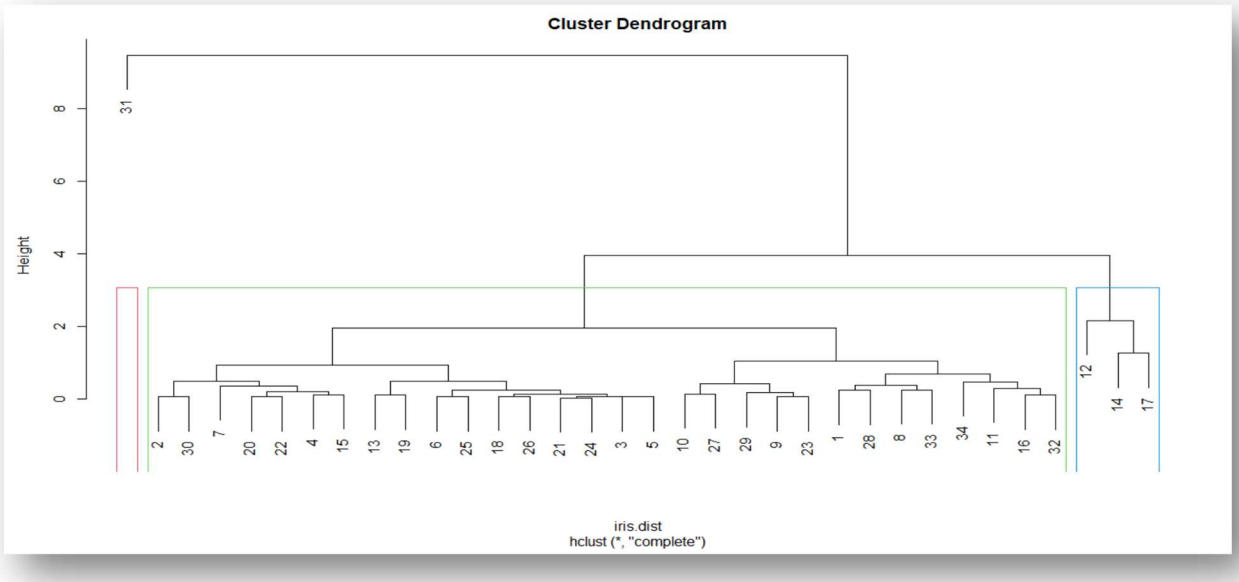
There are two forms of hierarchical clustering: agglomerative and divisive. I will be using agglomerative hierarchical clustering.

- Hierarchical cluster analysis (or hierarchical clustering) is a broad method in which the goal is to group objects or records that are "near" to one another together.
- It is a method of grouping related items into clusters. The endpoint is a collection of clusters, each of which is different from the others yet the items inside each cluster are roughly similar.

- The frequent calculation of distance measurements between items and between clusters, after objects begin to be sorted into clusters, is a critical component of the analysis.
- Initially, each data point is treated as a separate cluster in this method. Similar clusters merge with other clusters in each iteration until one cluster or K clusters are produced.
- Agglomerative core algorithm is straightforward.
- Make a proximity matrix.
- Each data point should be considered a cluster.
- Repeat: Update the proximity matrix by merging the two closest clusters.
- Until there is only one cluster left
- The calculation of the distance between two clusters is a crucial procedure.
- The result is visually displayed as a dendrogram.

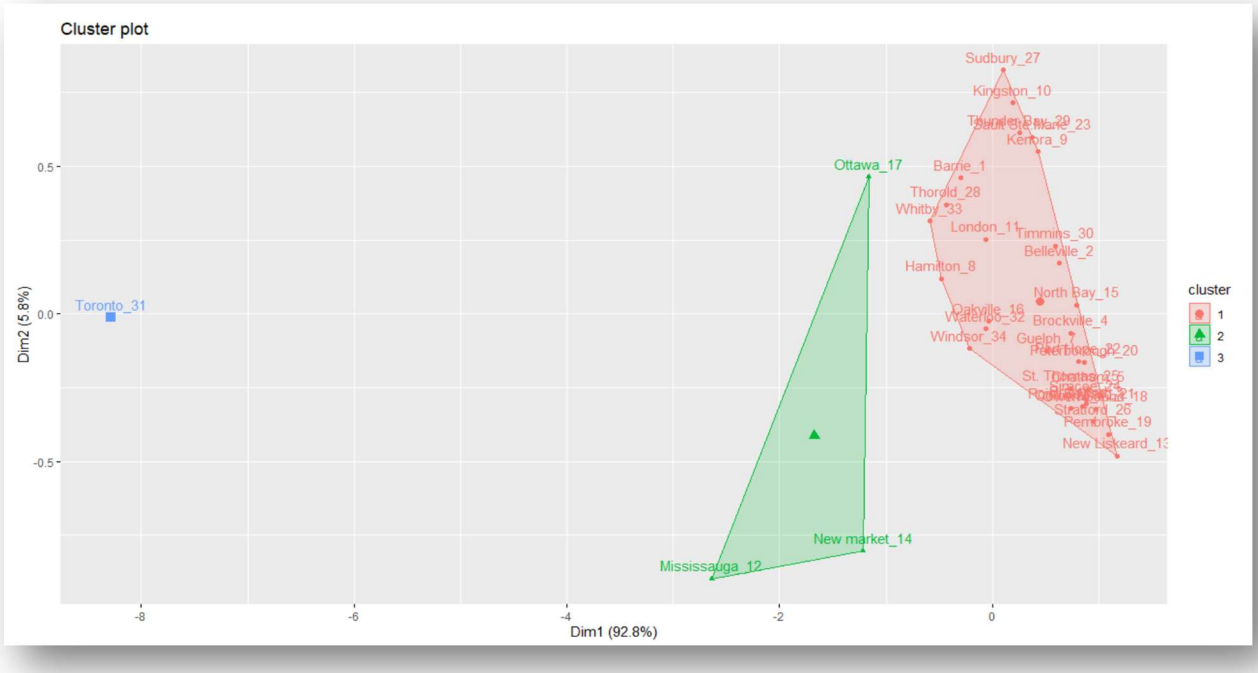
DENDOGRAM

- A dendrogram is a diagram that depicts the object's hierarchical connection. It is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data.
- A dendrogram is only correct when the data meets the ultrametric tree inequality, which is unlikely to be the case for any real-world data.
- For the dataset first the assumed number of clusters is 3



The rectangles in the dendrogram show us the clusters. Once we have the optimal number of clusters again, we will go ahead and plot the clusters.

CLUSTER PLOT 1



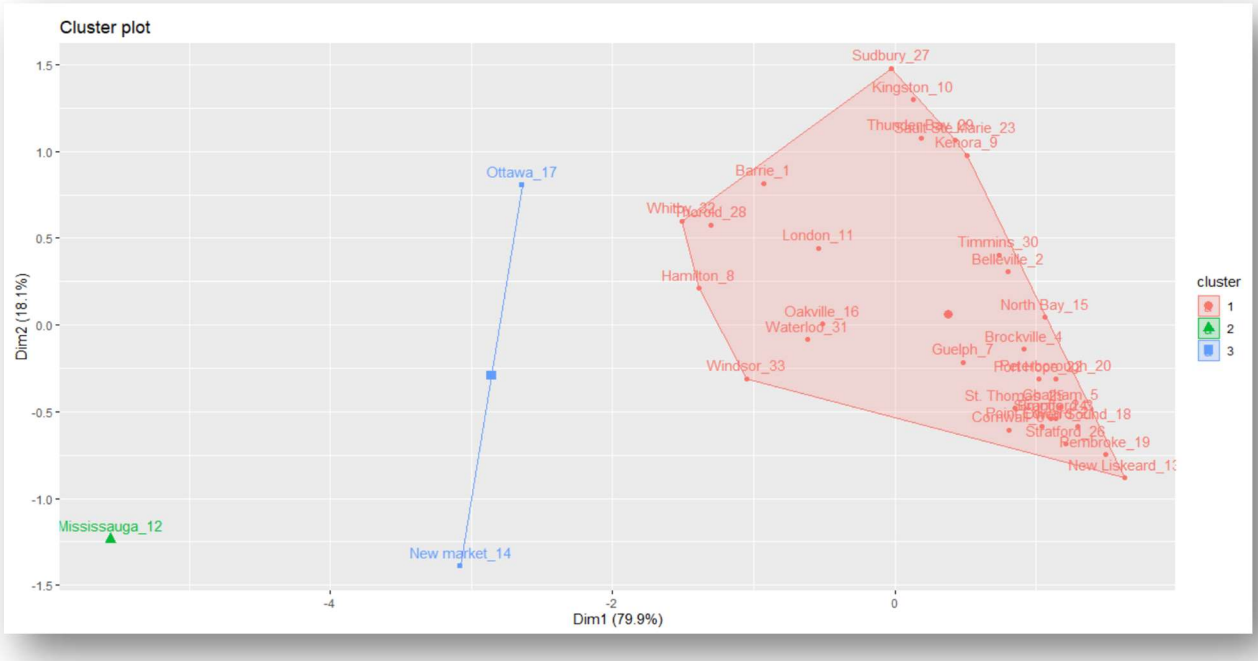
CLUSTER 1	CLUSTER 2	CLUSTER 3
Barrie, Hamilton, Kingston, London, Oakville, Windsor, Sudbury, Thorold, Waterloo, Whitby, Thunder Bay, Sault Ste.Marie, Kenora, Timmins, Stratford, St. Thomas, Simcoe St., Port Hope, Point Edward, Peterborough, Pembroke, Owen Sound, North Bay, New Liskeard, Guelph, Cornwall, Chatham, Brockville, Brantford, Belleville	New Market, Mississauga, Ottawa	Toronto

PLOT ANALYSIS

Toronto has the highest cases. The remaining two clusters are not accurate since the main attempt is to determine similarity in the cites as per the three categories, but when the data from the original set is matched with this cluster plot it is very obvious to notice that some cities are clustered in the wrong plot, giving us an inaccurate result. Thunder Bay, Sault Ste.Marie, Kenora are the additional cities in cluster 2 which have relatively less cases. Looking at the previous plots it is obvious that this cluster plot is not accurate enough. Again, Toronto is the outlier here.

WITHOUT OUTLIER TORONTO

Number of clusters (k) = 3

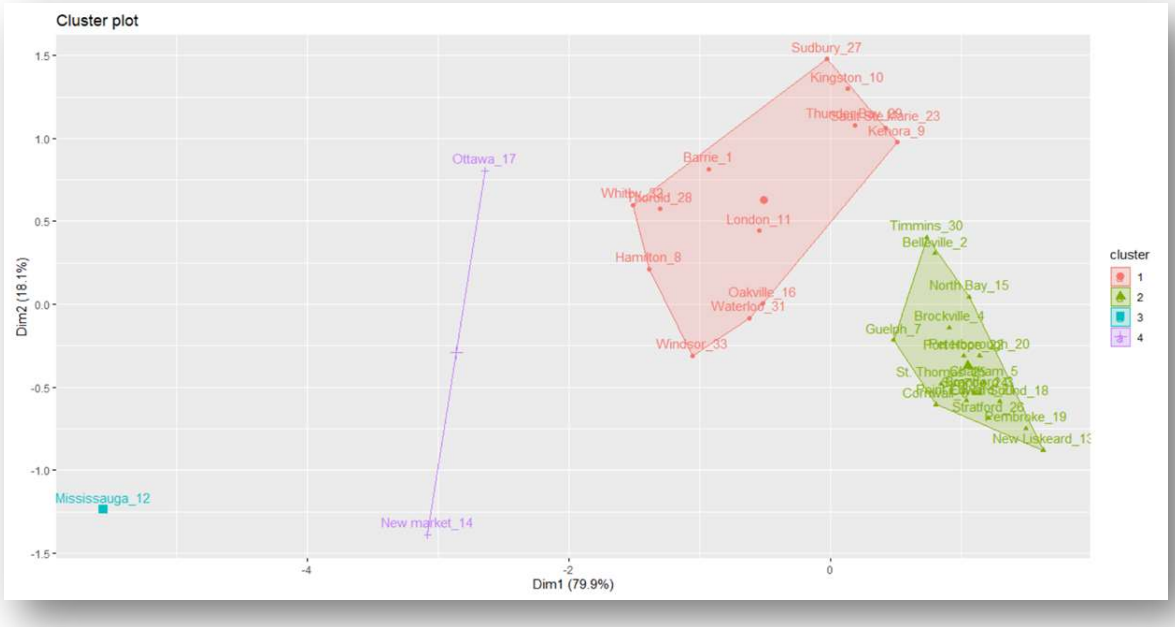


CLUSTER 1	CLUSTER 2	CLUSTER 3
Barrie, Hamilton, Kingston, London, Oakville, Windsor, Sudbury, Thorold, Waterloo, Whitby, Timmins, Thunder Bay, Stratford, St. Thomas, Simcoe St., Sault Ste.Marie, Port Hope, Point Edward, Peterborough, Pembroke, Owen Sound, North Bay, Kenora, New Liskeard, Guelph, Cornwall, Chatham, Brockville, Brantford, Belleville	Mississauga	Ottawa, New Market

PLOT ANALYSIS

After removing the outlier, we notice that the highest covid struck city is Mississauga and then followed by New market and Ottawa. But if we do look at the dataset and run the numbers, we notice that new market and Mississauga have almost the same cases. Therefore, we can conclude that this plot is not very accurate but good enough to work with.

Number of clusters (k) = 4



CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4
Barrie, Hamilton, London, Oakville, Ottawa, Windsor, Thorold, Waterloo, Whitby	Kenora, Kingston, Sault Ste.Marie, Sudbury, Thunder Bay, Timmins	Stratford, St. Thomas, Simcoe St., Port Hope, Point Edward, Peterborough, Pembroke, Owen Sound, North Bay, New Liskeard, Guelph, Cornwall, Chatham, Brockville, Brantford, Belleville	Mississauga, New Market

CONCLUSION

The findings of the provincial cluster are anticipated to assist the government in developing policies relating to limits on community activities or other measures aimed at combating the spread of COVID-19. All the cluster plots result in helping us understand how the cities in Ontario are similar to each other in terms of three indicators, resolved, Not resolved and Fatal cases. The main observation is that Toronto is the highest struck city by the pandemic and has the highest number of cases. After we remove Toronto from the dataset, we observe that two cities Mississauga and New market are the next cities with the highest number of cases. But to note that Ottawa had also been clustered along with the former two cities so, I will assume that the three cities Mississauga, Newmarket and Ottawa are the next red zones after Toronto. This means tighter implementations of rules and regulation for the 4 cities i.e., Toronto, Mississauga, Ottawa and New market.

The cities which got clustered in the same cluster every simulation are: Stratford, St. Thomas, Simcoe St., Port Hope, Point Edward, Peterborough, Pembroke, Owen Sound, North Bay, New Liskeard, Guelph, Cornwall, Chatham, Brockville, Brantford, Belleville. From this we understand that these cities have similar Covid-struck situation on the dependence of three parameters Not resolved, Resolved and Fatal cases. So, the set of rule implementations for these cities should be similar.

After analyzing the cluster plots of K-means and Hierarchical clustering, in my opinion, it is fair to deduce that the cluster plots of hierarchical clustering technique are quite inaccurate, as few

cities with high covid cases have been clustered along with the cities with a low Covid struck rate. Hierarchical clustering rarely provides the best solution as it involves lots of arbitrary decisions and it does not work with missing data. In addition, it works poorly with mixed data types and it does not work well on very large data sets Its main output, the dendrogram, is commonly misinterpreted. If there is a specific number of clusters in the dataset, but the group they belong to is unknown, choose K-means. With large number of variables K-means computes faster, which is why here I prefer K-means as a better approach than Hierarchical clustering.

Mathematical computations along with technological world can help us deal with real life scenarios which would have been hard to deal with as large sets of un-processed data are of no use if we do not have the computational capability to work with it. In this project we saw how we could help reduce the effects of the pandemic by processing large datasets with the help of Mathematics and Computer Science.

REFERENCES

- Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2021). The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. *Quality & Quantity*. https://doi.org/10.1007/s11135-021-01176-w
- https://www.facebook.com/displayr. (2018, March 21). *What is a Dendrogram?* Displayr. https://www.displayr.com/what-is-dendrogram/#:~:text=A%20dendrogram%20is%20a%20diagram,to%20allocate%20objects%20to%20clusters.
- Maugeri, A., Barchitta, M., Basile, G., & Agodi, A. (2021). Applying a hierarchical clustering on principal components approach to identify different patterns of the SARS-CoV-2 epidemic across Italian regions. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-86703-3
- *K-Means Clustering vs Hierarchical Clustering*. (2020, October 11). Globaltechcouncil.org. https://www.globaltechcouncil.org/clustering/k-means-clustering-vs-hierarchical-clustering/

Research Clustering Techniques

Hidaya Patel

3/2/2022

K-Means Clustering for the data

```
#Using package factoextra
library(factoextra)
```

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

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

```
# Importing Dataset
library(readr)
Covid_19_cities <- read_csv("C:/Users/Hidaya/Desktop/Winter 2022/research/Covid-19-cities.csv")
```

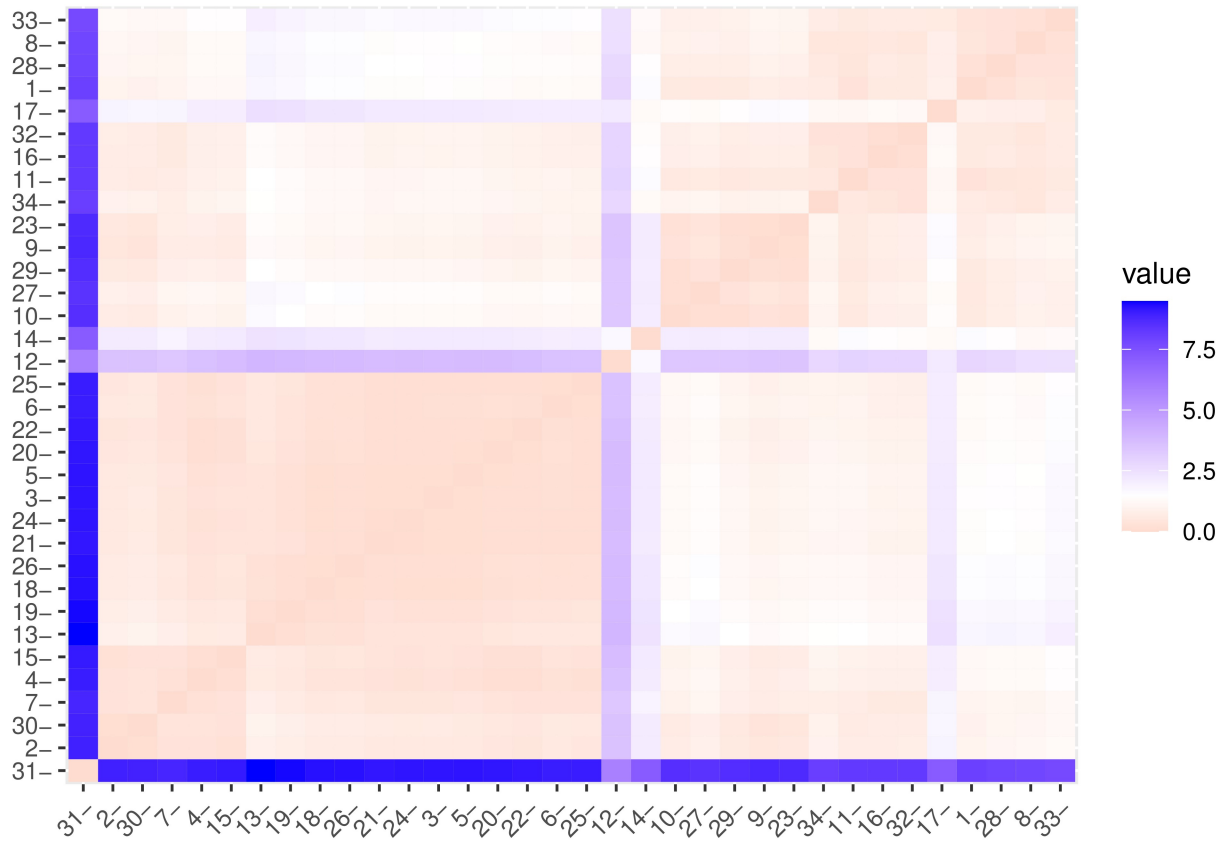
```
## Rows: 34 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Cities
## dbl (3): Not resolved, Fatal, Resolved
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

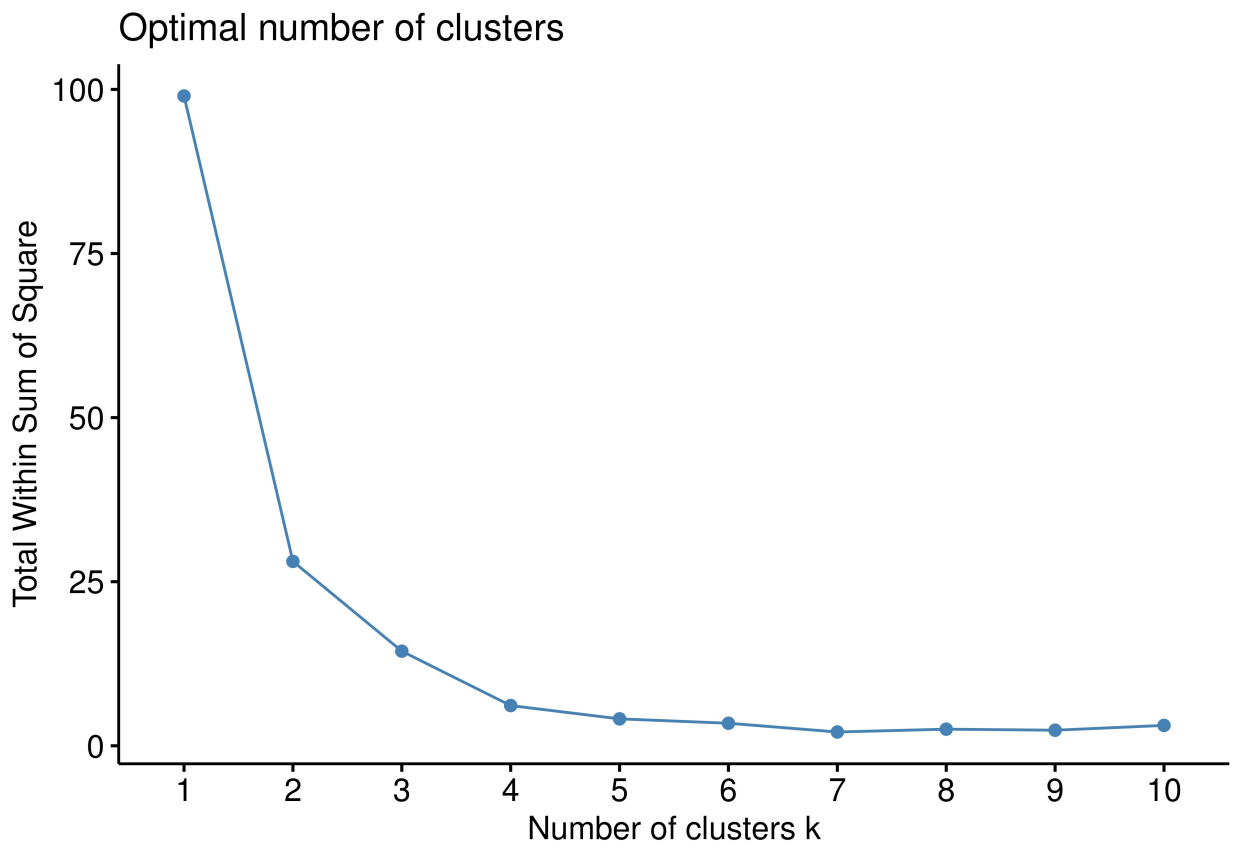
```
cities <- Covid_19_cities

# Storing columns 2,3,4 in data
data <- cities[2:4]

# Scaling the data
data <- scale(data)

# Finding the euclidean distance
euclidean <- get_dist(data)
summary(euclidean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03784 0.48929 0.97668 1.49710 1.45740 9.47273
```

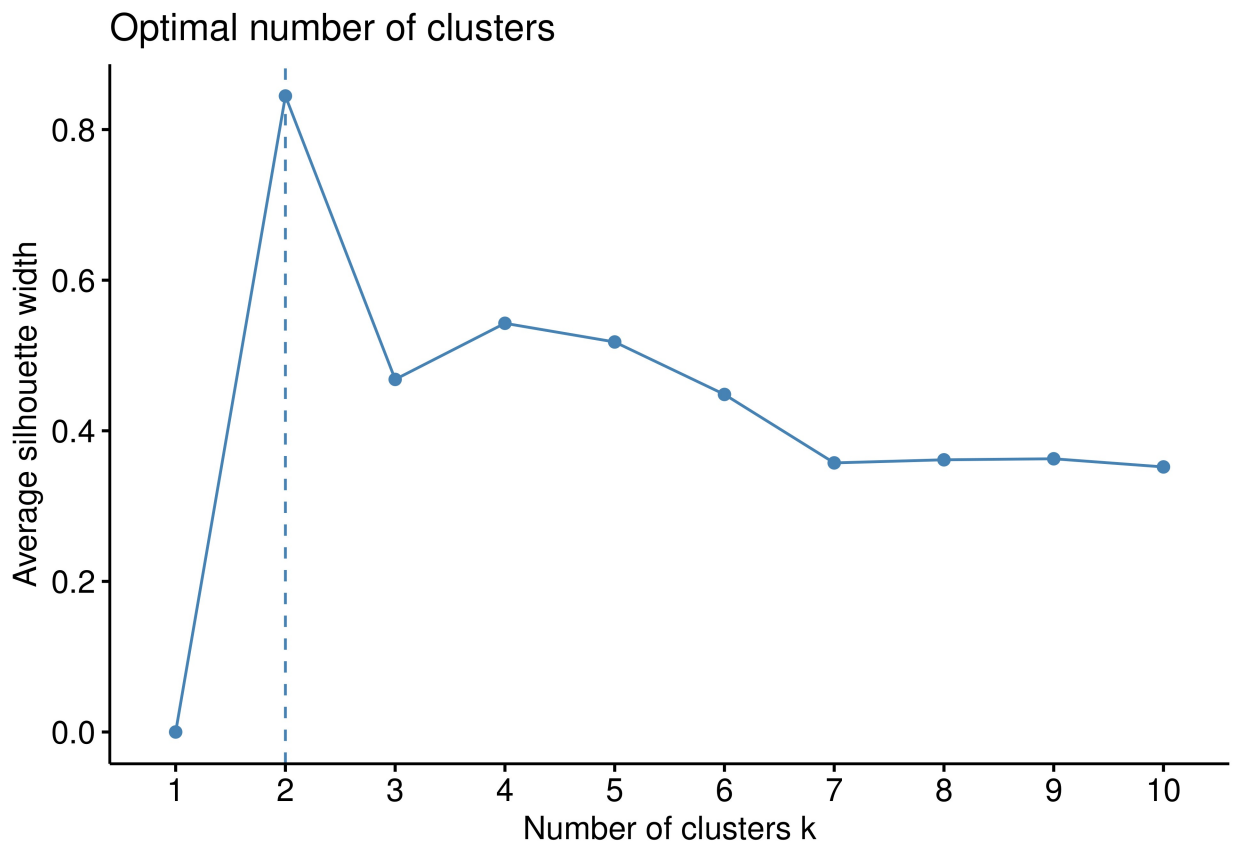
```
# Plotting the distance matrix
fviz_dist(euclidean)
```




```
set.seed(123)

# Finding the optimal number of clusters
fviz_nbclust(data, kmeans, method = "wss")
```



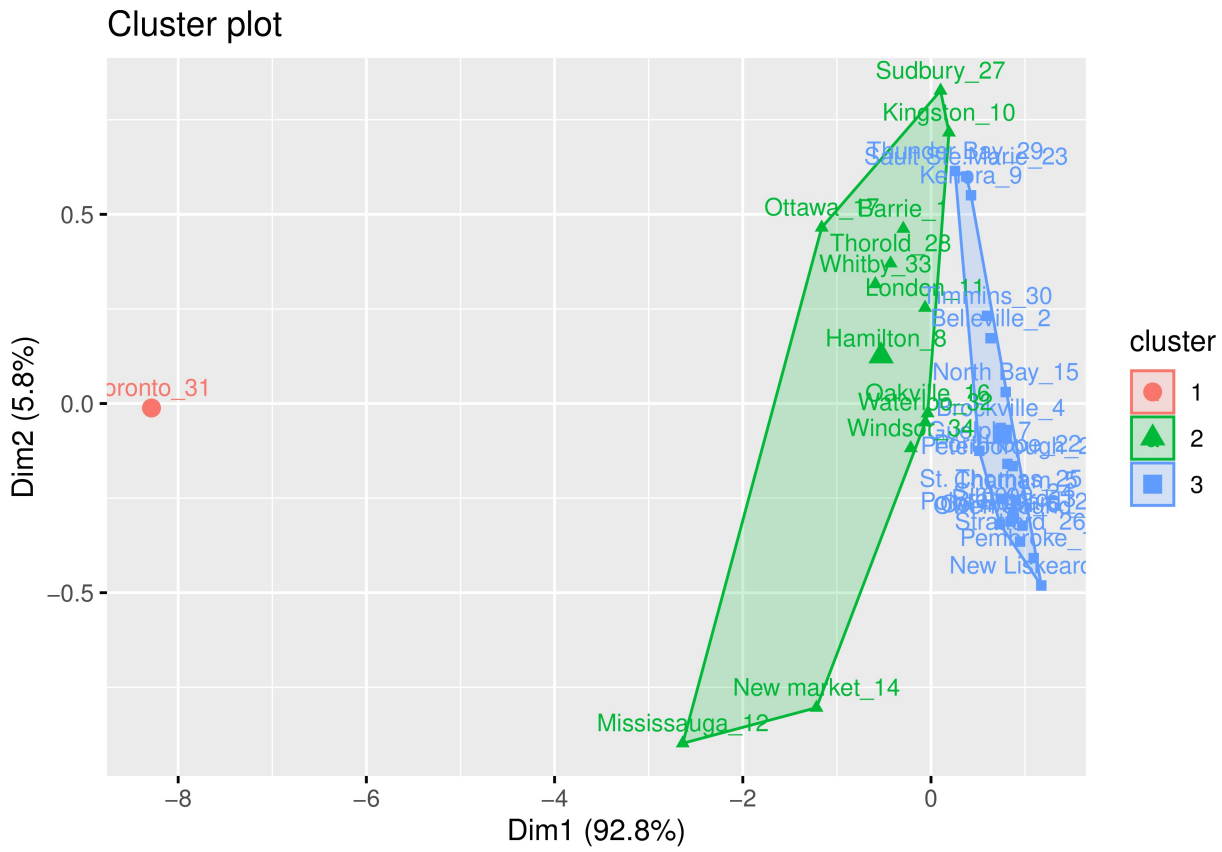

```
set.seed(123)
fviz_nbclust(data, kmeans, method = "silhouette")
```




```r
set.seed(123)

# Applying K-means clustering
endkmeans <- kmeans(data, 3, nstart = 100)

print(endkmeans)
```

```
## K-means clustering with 3 clusters of sizes 1, 13, 20
##
## Cluster means:
##   Not resolved      Fatal   Resolved
## 1    4.5904815  5.2202015  4.5276077
## 2    0.4188369  0.1841023  0.3200466
## 3   -0.5017681 -0.3806766 -0.4344107
##
## Clustering vector:
##  [1] 2 3 3 3 3 3 3 2 3 2 2 2 3 2 3 2 2 3 3 3 3 3 3 3 3 3 2 2 3 3 1 2 2 2
##
## Within cluster sum of squares by cluster:
## [1]  0.000000 11.069158  3.347895
##  (between_SS / total_SS =  85.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
# Plotting the clustered cities
rownames(data) <- paste(cities$Cities, 1:dim(cities)[1], sep = "_")

fviz_cluster(endkmeans, data = data, labelsize = 10)
```




```r
# Just a visual of which city belongs to which cluster
table(endkmeans$cluster,cities$Cities)
```

```
##
##     Barrie Belleville Brantford Brockville Chatham Cornwall Guelph Hamilton
##   1      0          0         0          0       0        0      0        0
##   2      1          0         0          0       0        0      0        1
##   3      0          1         1          1       1        1      1        0
##
##     Kenora Kingston London Mississauga New Liskeard New market North Bay
##   1      0        0      0           0            0          0         0
##   2      0        1      1           1            0          1         0
##   3      1        0      0           0            1          0         1
##
##     Oakville Ottawa Owen Sound Pembroke Peterborough Point Edward Port Hope
##   1        0      0          0        0             0             0         0
##   2        1      1          0        0             0             0         0
##   3        0      0          1        1             1             1         1
##
##     Sault Ste.Marie Simcoe St. Thomas Stratford Sudbury Thorold Thunder Bay
```

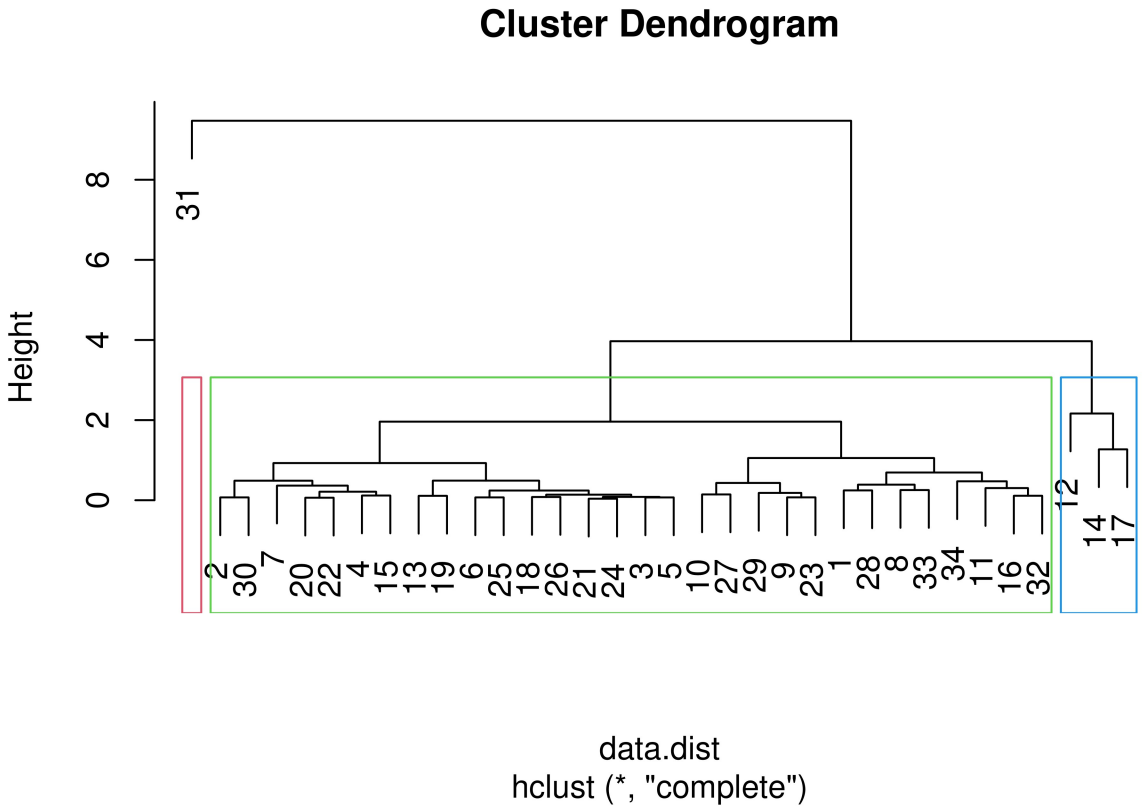
```
##   1              0      0          0         0       0       0           0
##   2              0      0          0         0       1       1           0
##   3              1      1          1         1       0       0           1
##
##     Timmins Toronto Waterloo Whitby Windsor
##   1       0       1        0      0       0
##   2       0       0        1      1       1
##   3       1       0        0      0       0
```

Heirarchical clustering for the data

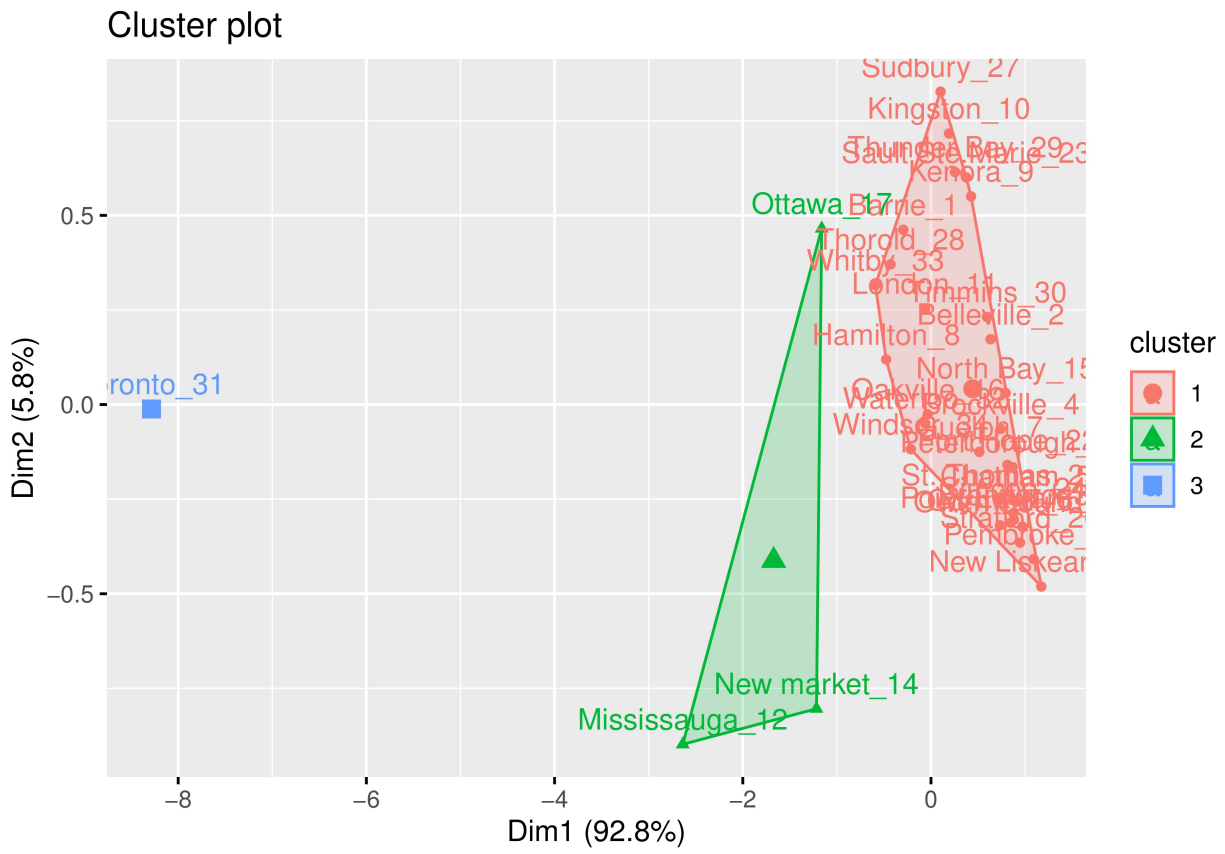
```
# Extracting Data
Covid_Cities <-Covid_19_cities$Cities

# Loading required data and scaling it
data <- cities[2:4]
data <- scale(data)

# Distance Matrix Computation
data.dist <- dist(data)

# Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.
hc.cities <- hclust(data.dist, method = "complete")
hc.cities
```

```
##
## Call:
## hclust(d = data.dist, method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 34
```

```
# Plotting the enhanced dendogram with rectangles around the clusters
plot(hc.cities)
rect.hclust(hc.cities, k = 3, border = 2:5)
```




```
COvid.clusters <- cutree(hc.cities, k = 3)

rownames(data) <- paste(cities$Cities, 1:dim(cities)[1], sep = "_")

fviz_cluster(list(data = data, cluster = COvid.clusters))
```




```
table(COvid.clusters, cities$Cities)
```

```
##
## COvid.clusters Barrie Belleville Brantford Brockville Chatham Cornwall Guelph
##              1      1          1         1          1       1        1      1
##              2      0          0         0          0       0        0      0
##              3      0          0         0          0       0        0      0
##
## COvid.clusters Hamilton Kenora Kingston London Mississauga New Liskeard
##              1        1      1        1      1            0            1
##              2        0      0        0      0            1            0
##              3        0      0        0      0            0            0
##
## COvid.clusters New market North Bay Oakville Ottawa Owen Sound Pembroke
##              1          0         1        1      0          1        1
##              2          1         0        0      1          0        0
##              3          0         0        0      0          0        0
##
## COvid.clusters Peterborough Point Edward Port Hope Sault Ste.Marie Simcoe
##              1            1            1         1               1      1
##              2            0            0         0               0      0
##              3            0            0         0               0      0
##
## COvid.clusters St. Thomas Stratford Sudbury Thorold Thunder Bay Timmins Toronto
##              1          1         1       1       1           1       1       0
```

```
##              2          0         0       0       0           0       0       0
##              3          0         0       0       0           0       0       1
##
## COvid.clusters Waterloo Whitby Windsor
##              1        1      1       1
##              2        0      0       0
##              3        0      0       0
```

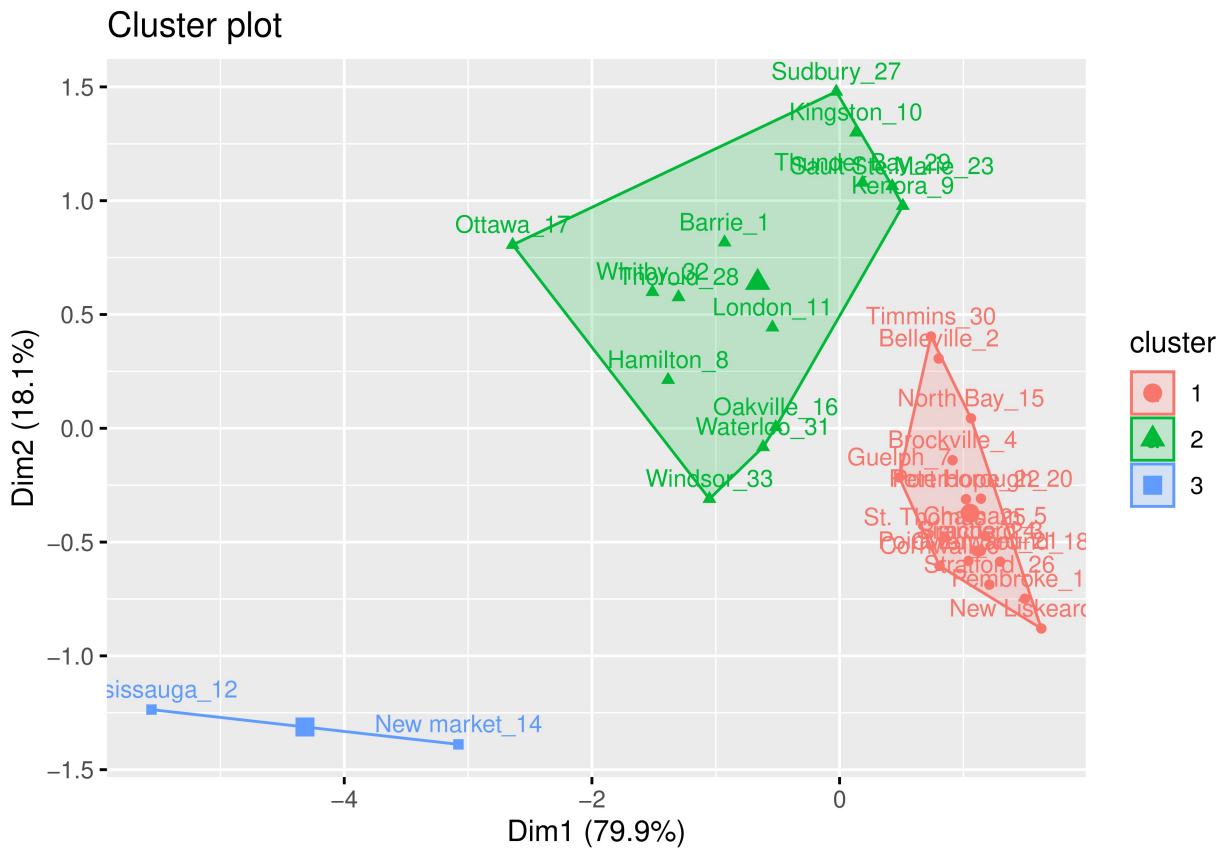
K-means without outlier Toronto (outlier), cluster numbers 3 and 4 for each technique

3 clusters

```
library(factoextra)
library(readr)
modified <- read_csv("C:/Users/Hidaya/Desktop/Winter 2022/research/covid-cities-2.csv")
```

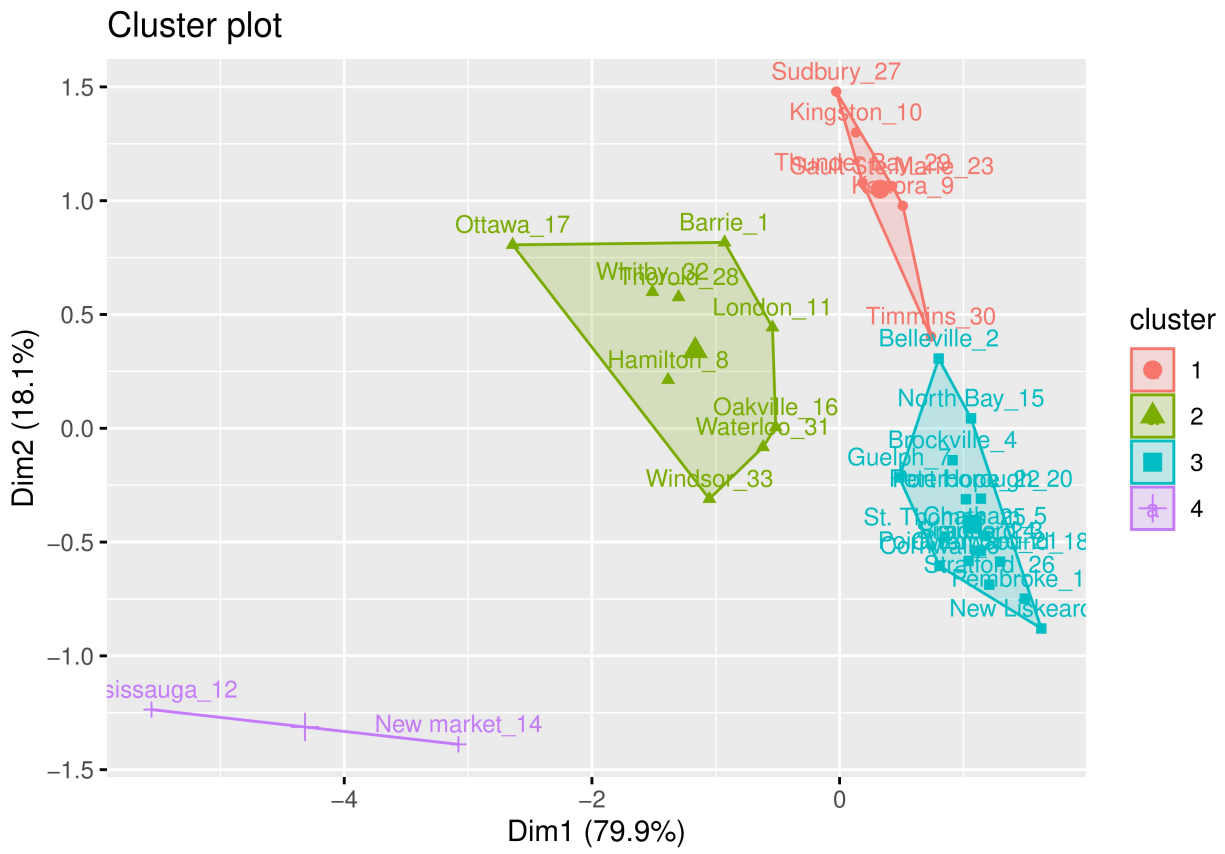
```
## Rows: 33 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Cities
## dbl (3): Not resolved, Fatal, Resolved
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
cities_new <- modified

# Storing columns 2,3,4 in data
data1 <- cities_new[2:4]

# Scaling the data
data1 <- scale(data1)

# Finding the euclidean distance
euclidean1 <- get_dist(data1)

# Applying K-means clustering
endkmeans1 <- kmeans(data1, 3, nstart = 100)

rownames(data1) <- paste(cities_new$Cities, 1:dim(cities_new)[1], sep = "_")

fviz_cluster(endkmeans1, data = data1, labelsize = 10)
```



4 clusters

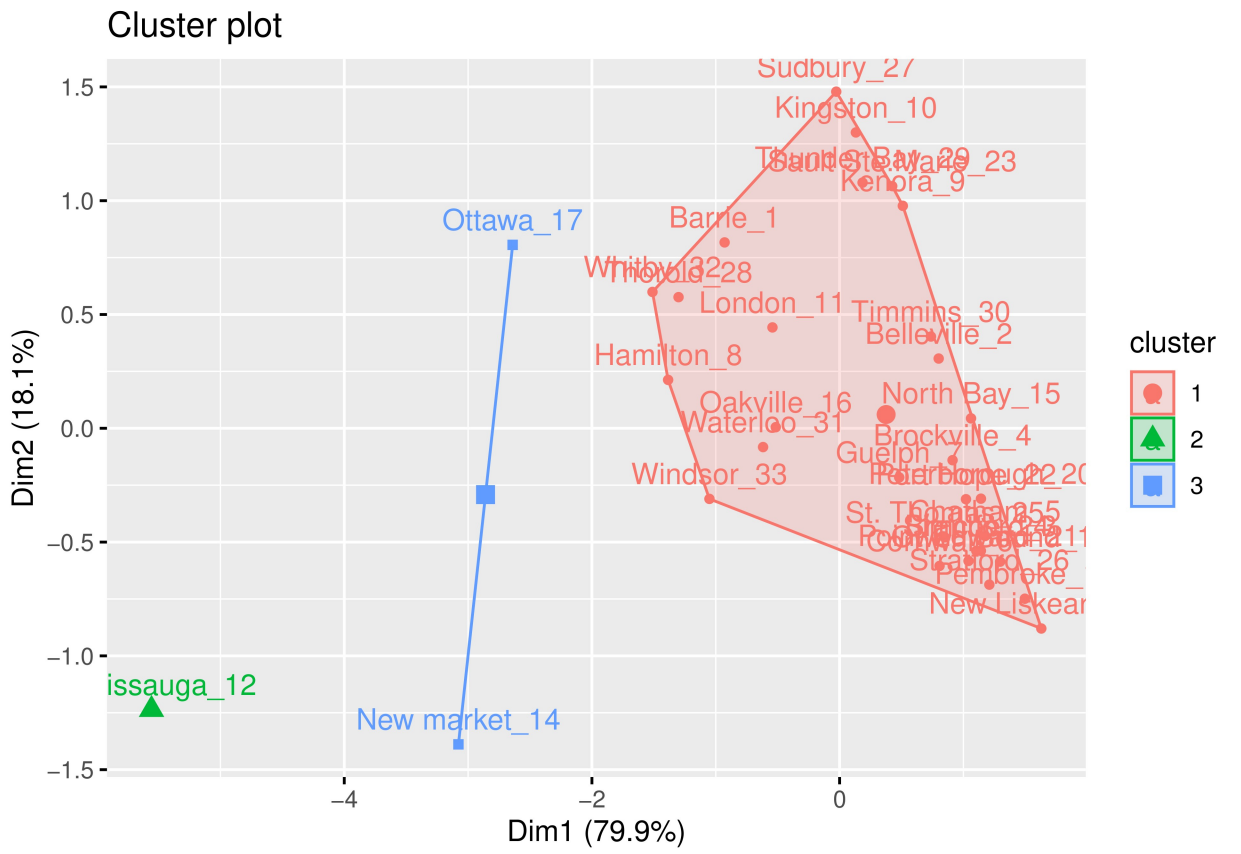
```
# Applying K-means clustering
endkmeans1 <- kmeans(data1, 4, nstart = 100)

rownames(data1) <- paste(cities_new$Cities, 1:dim(cities_new)[1], sep = "_")

fviz_cluster(endkmeans1, data = data1, labelsize = 10)
```



Heirarchical clustering without Toronto (outlier)

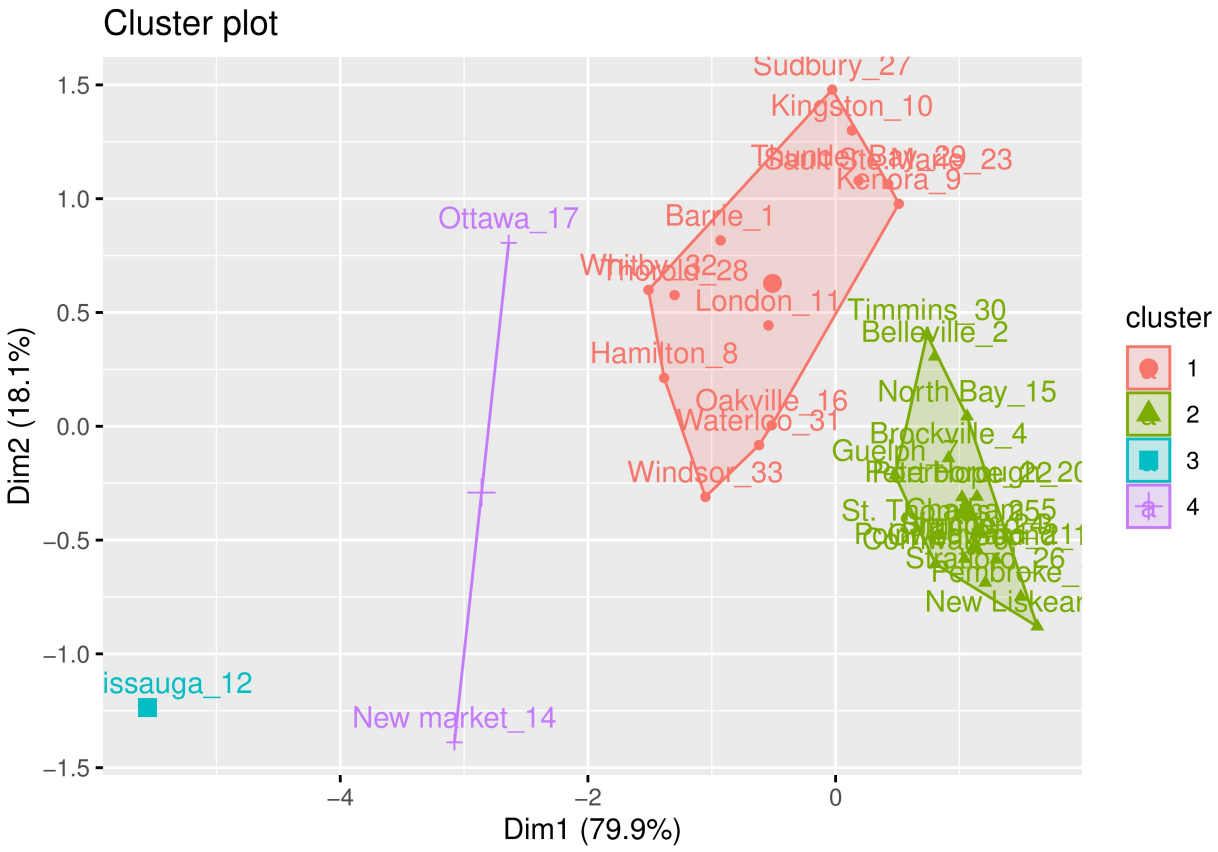
3 clusters

```
# Distance Matrix Computation
data1.dist <- dist(data1)

# Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.
hc.cities1 <- hclust(data1.dist, method = "complete")

COvid.clusters1 <- cutree(hc.cities1, k = 3)

rownames(data1) <- paste(cities_new$Cities, 1:dim(cities_new)[1], sep = "_")

fviz_cluster(list(data = data1, cluster = COvid.clusters1))
```



4 clusters

```
# Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.
hc.cities1 <- hclust(data1.dist, method = "complete")

COvid.clusters1 <- cutree(hc.cities1, k = 4)

rownames(data1) <- paste(cities_new$Cities, 1:dim(cities_new)[1], sep = "_")

fviz_cluster(list(data = data1, cluster = COvid.clusters1))
```


Cluster plot
Sudbury_27
Kingston_10
Kenora_9
Ottawa_17
Barrie_1
London_11
Timmins_30
Belleville_2
Hamilton_8
North Bay_15
Oakville_16
Waterloo_31
Brockville_4
Windsor_33
issauga_12
New market_14
cluster
1
2
3
4
Dim2 (18.1%)
Dim1 (79.9%)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−4
−2
0