EDX CYO Final project

Report and findings

Fernando A. Mitidieri

03/06/2024

1. Introduction

This is the report of the final project for the HarvardX Data Science Professional Certificate. The objective of the project is to develop a framework aimed at helping maintenance service companies optimize their resources to enhance the quality of services they provide, specifically for companies managing Automatic Teller Machines, also called ATMs.

Professional service companies that maintain equipment such as ATMs encounter significant challenges in resource management and service quality. This project addresses two primary issues: determining the the best locations for each regional center of maintenance, the bases were the field engineer are related, and maximum distances and time travels from these branches to the ATM locations, to allow the SLA companies achieve their contracted Service Level Agreements (SLAs).

Companies in the ATM service industry, also referred as Service Level Management (SLM) providers, face challenges in achieving cost savings and fulfilling the maintenance services as per the SLAs. The SLAs typically define the time frames within which the SLM providers must respond and resolve any arising issues.

For this analysis, the dataset utilized was downloaded from the data.nyc.gov website, which provides data transparency related to the Government of the State of New York. The dataset includes information about ATMs installed across New York State, owned by banks. This dataset forms the basis for modeling the number of field engineers needed and their optimal locations relative to ATM sites to minimize travel time and response delays.

ATM dataset is a large dataset with 5041 different ATM locations, from 236 different Financial Institutions.

The ATM dataset is composed by 6 columns stated below:

names	description	type
Name of Institution	Name of the institution that owns the ATM	Plain Text
Street Address	Street address of the ATM location	Plain Text
City	City in which the ATM is located	Plain Text
ZIP Code	ZIP Code	Number
County	County	Plain text
Georeference	Georeference	Point

The project's source code and related files are managed through GitHub, ensuring version control and public accessibility. The project repository can be accessed at:

GitHub - https://github.com/fmitidieri/EDX_CYO/

Within the project directory, the following files are available:

EDX_CYO_Final.R: Script in R containing all development aspects of the model;
EDX_CYO_Final.RMD: R Markdown file to generate the report in PDF format;
A PDF format report.

Additionally, two sub-directories are included under the main directory:

figures/ - for storing images used in the report.
data/ - for storing datasets used in the analysis.

This report sets the stage for further discussion and presentation in subsequent sections, detailing the steps required to achieve the defined project goal. By addressing the identified challenges, the developed framework aims to significantly improve the operational efficiency of SLM providers, ensuring better compliance with SLAs and overall service quality.

To accomplish the goal defined above, will be necessary several steps that follow:

A. Explonatory Analysis
a) Download complete dataset.
b) Pre-processing data.
c) Plot ATM locations.

B. Model
a) Hyper parameters.
b) Develop models.

2. Methods and Analysis

A. Exploratory Analysis

a) Download complete dataset.

Was used a script to download the dataset directly from the source located in the following webpage https://data.ny.gov/api/views/ndex-ad5r/rows.csv./

Doing this way is guaranteed that the analysis always will run witl the latest data available.

b) Pre-processing data.

Pre-process the data to separate the latitude and Longitude coordinates in individual columns to be used by developed functions that has the intent to calculate the centyriod of the clusters and also to define geographical distance among the center of the cluster and the ATM locations.

This is the unique pre-processing needed, once we advance in the analysis and in the cluster definition, might would be necessary more work on the data.

After pre-processed the dataset, it was saved as a Rdata file to be used during the development stage of analysis.

c) Plot ATM locations.

The first approach is plot the map of the State of New York and plot also the location of each ATM included in this dataset.

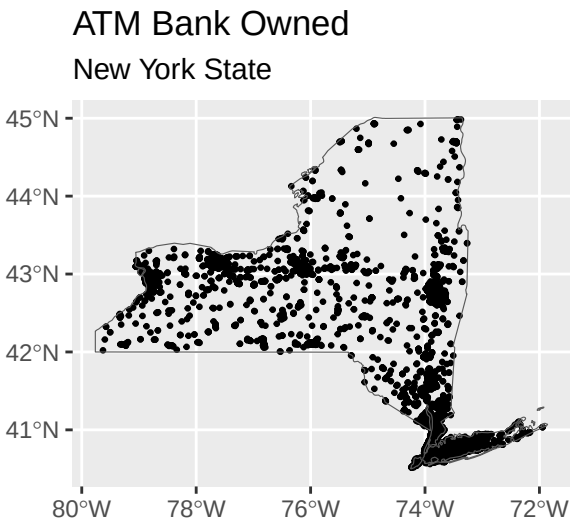

Fig. 1 ATM Geographical Distribution

By inspection one can note that looks like there are some clusters that can be identified. Basically there are 4 cluster along the parallel 43o and two other in the parallel 41o and longitude 73o nd 74o. And of course the Long island has the major density of ATMs in the State.

Not surprising the cluster are related some how with the biggest cities in the state as it can be vosualized below.

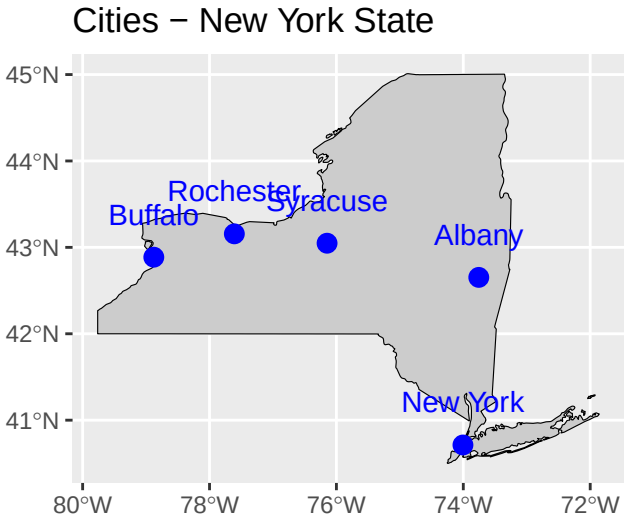

Fig 2. Main cities of New York State

But also there are several ATMs that are spread over the State. The key point here is the drivable distance between them and the time to go from one place to another and the field engineer bases.

This work will test several clusters algorithms and also will use some hyper parameters, like travel time, average of ATMs per field engineer, average of calls per ATM per month, and average time for each call to support the experimentation and improve results.

If it is considered the drivable distances inside the New York State from the Western city of Buffalo and the Easten city of Albany, will be noticed that it is need to account this parameter to define the finals clusters and number of field engineer.

Just to give an exemple, the drivable distance between these two cities is 417.01 km and in a business day it takes approximately 03:47 hours to drive from one city to another.

And if we consider the distance and time between the city of Plattsburgh and New York are 443.11 km and 04:02 hour respectively.

From the data above, it is clear that distance and time to travel between the ATMs should be a concern to produce a good performance.

From our preliminary analysis also is clear that this is a problem of unsupervised learning and clustering, and this will be the approach during the modeling phase.

B. Model

a) Hyper parameters.

For this specific problem there are some aspects that we must consider to support the decision regarding the number of clusters.

A field engineer's working day is typically 11 hours long, which includes a one-hour lunch break. Typically, an ATM call takes 45 minutes and a field engineer can make 4 calls per day.

Considering these hyper parameters, then, in a normal work day, a field engineer will spend 3 hours in front of an ATM, 1 hour having lunch and 7 hours traveling from one ATM to another and on his commute. Which means he has approximately 1h30 on average to travel from one point to another.

This will be one of the criteria to evaluate the results: an average of 1:30 hour travel time between the center of the clusters and the ATMs.
Other criteria will be defined as a maximum distance of 250 km between the ATM and the center of the respective cluster.

These hyper parameters will be used to define how many clusters to consider as optimal for the algorithms.

b) Developing the models.

Considering the clusters algorithms, this work will use the K-means and the DBscan.

K-means is an algorithm that cluster all elements based on distances of each element to centroids of the clusters. It is an easy and commonly algorithm used to solve cluster problems.

DBScan is a density-based algorithm for clustering and allows create clusters with any shape, which is pretty useful for data with different geographical distributions.

b.1) K-means.

The first approach will be with K-means algorithm and the first step will be try to figure out the best K for the dataset in use. One of the most common techniques is use the Elbow curve to identify the best number of clusters.

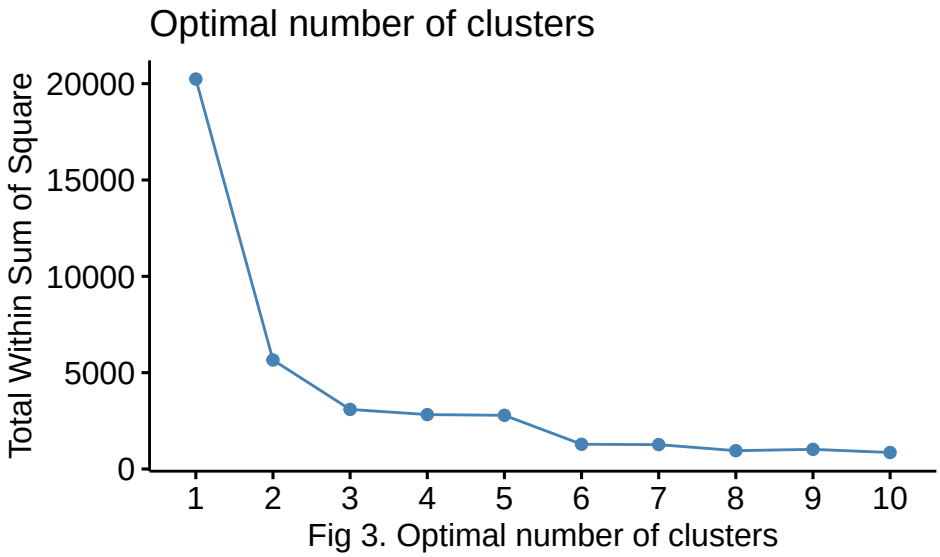

Fig 3. Optimal number of clusters

Considering the elbow curve calculated, the ideal K is 2, so let's run the K-means algorithm considering this quantity of clusters and analyze the results.

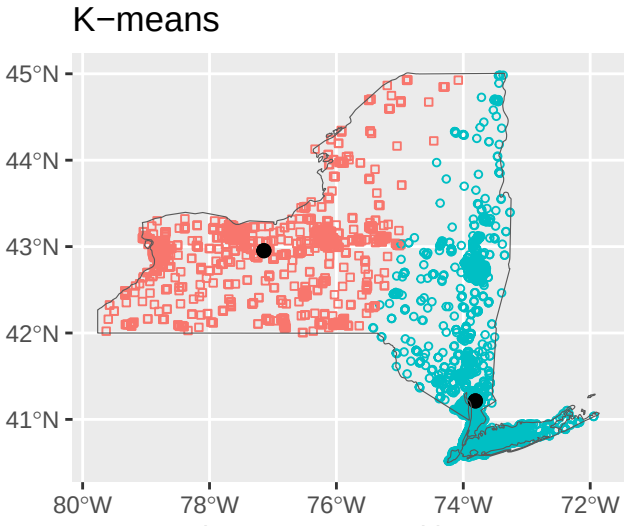

Fig 4. K−means with K=2

According with which is possible to visualize in the map plotted below, only 2 clusters will generated a very spread group of ATMs and this will probably result in a higher distance and time travel between the ATMs and the centroid of each cluster.

Let's try validate this conclusion analyzing some analytical data from the results. In the table below is showed that there are maximum distances between 450 and 550 km and also maximun time travels between 2:59 and 3:49 hours. So, according with these criteria, we will need a higher number of clusters in order to reduce the maximum distance and maximum time travel.

	Max Distance (km)	Min Distance (km)	Mean Distance (km)	Max Time (hh:mm)	Min Time (hh:mm)	Mean Time (hh:mm)
Cluster1	329.18	0.97	102.84	02:59	00:00	00:56
Cluster2	420.60	3.00	78.09	03:49	00:01	00:42

Clearly 2 clusters are not the best solution, considering the distances and times to travel between an ATM and the centroid of its cluster. We want cluster that have travel time on average of 1:30 hour and also maximum distance of 250 km, as explained above.

Additionally, the cluster 2 has a huge density in the Long island, and this helped to reduce the average time to almost 1 hour.

For this reason, it will be tested other values for K to find out that one capable to outcome maximum distance less than 250 km and maximum travel time less than 90 minutes.

b.2) Best K for K-means according with the distance and travel time.

Now the idea is try several different values for K to identify the right one that will generate the best maximum distance and travel time between each ATM location and the centroid of the clusters.

For this, was developed a loop that will calculate the cluster for a sequence of K, used from 3 to 20, and will compare the maximum distance between the ATMs and the centroid of each cluster, when this maximum distance were less than 250 km and the maximum trave time were less than 90 minutes, than this will be the lowest K that accomplish these goal.

K	Cluster	Max Distance (km)	Min Distance (km)	Mean Distance (km)	Max Time (hh:mm)	Min Time (hh:mm)	Mean Time (hh:mm)
3	1	233.65	5.96	83.72	02:07	00:03	00:45
3	2	203.67	2.57	90.17	01:51	00:01	00:49
3	3	155.31	1.95	33.79	01:24	00:01	00:18
4	1	232.84	6.21	62.52	02:07	00:03	00:34
4	2	148.86	4.11	58.62	01:21	00:02	00:31
4	3	155.55	1.85	33.78	01:24	00:01	00:18
4	4	226.98	4.13	65.38	02:03	00:02	00:35
5	1	107.39	0.32	24.11	00:58	00:00	00:13
5	2	148.86	4.11	58.62	01:21	00:02	00:31
5	3	137.62	0.71	28.20	01:15	00:00	00:15

K	Cluster	Max Distance (km)	Min Distance (km)	Mean Distance (km)	Max Time (hh:mm)	Min Time (hh:mm)	Mean Time (hh:mm)
5	4	226.77	4.35	65.04	02:03	00:02	00:35
5	5	232.48	6.12	62.47	02:06	00:03	00:34
6	1	107.51	0.43	24.12	00:58	00:00	00:13
6	2	110.83	2.78	32.75	01:00	00:01	00:17
6	3	143.99	0.70	28.21	01:18	00:00	00:15
6	4	108.06	1.14	41.72	00:58	00:00	00:22
6	5	231.58	3.23	55.69	02:06	00:01	00:30
6	6	227.84	4.07	65.58	02:04	00:02	00:35

Despite the result showed in the table above, it clear by the following result with K=6 that this numer of clusters is much better than the K=2, and it is easy to verify it visually in the graph below:

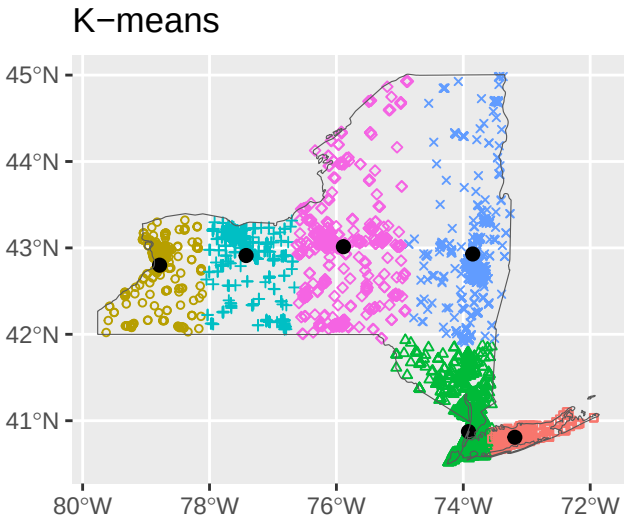

Fig 5. K−means with K=6

b.3) Including an extra dimension for K-means algorithm with K=6.

The next experiment is include an extra dimension to be used by the K-means. Was defined two reference points in the State of New York and then calculated the distance of these point to all ATM locations and included these data in the oroiginal dataset before calculate the new cluster distribution.

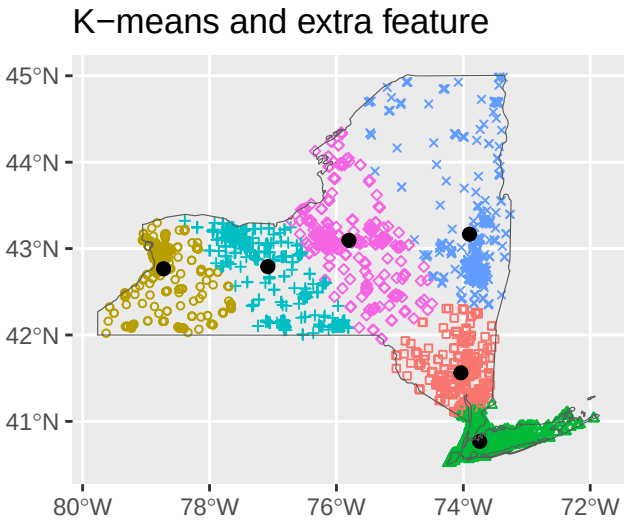

Fig 6. K−means with K=6

The result presents a slightly difference in the centroids of clusters and in the ATM distribution respectively. Besides the inclusion of this new feature previously could not make sense (also is based on distances), once were added these neew features generate a quite different outcomes.

The numerical result will be compared in the Results session.

b.4) DBScan.

Now let's try another algorithm, the idea is test DBscan to verify if it can produce a better result. DBScan is an algorithm based on the density of data, and differently of K-means, the clusters can assume any shape, which can have more effectiveness in this business case of the SLM for ATMs.

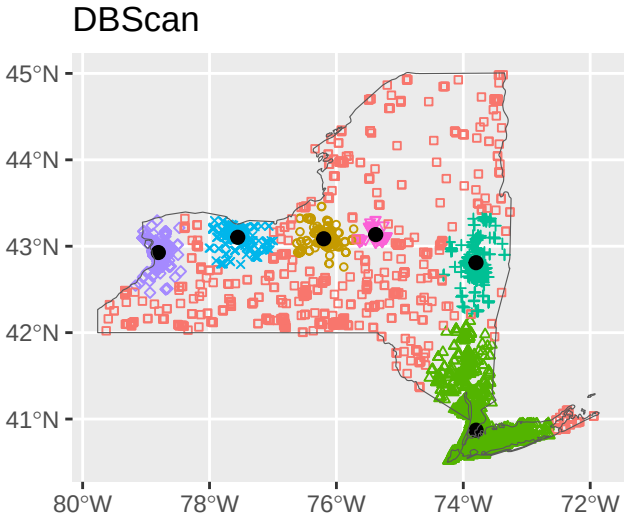

Fig 7. Clustering with DBScan

The clear problem is that it results in several ourliers that aren't associated with any cluster calculated by the algorithm. Several values for epsilon and Minimal Points were tested, and finally the eps = 10 and MinPts = 50 were the ones that performed better in terms of number of cluster and quantity of ATMs associated with clusters.

For comparison with the Kmeans algorithm, the criteria used to define the eps and MinPts was the minimun quantity of outliers generated for a number of cluster equal 6.

b.5) Combining DBScan with Kmeans.

As the ATMs locations not assigned by the DBScan algorithm were a issue to obtain a usable results, they will be distributed using the Kmeans algorithm among the clusters generated by the DBScan.

To perform this combination of algorithms, the centroids of the DBScan will be calculated as the center point of all originally assigned ATM location for a particular cluster and kept them fixed to Kmenas.

The result is presented below.

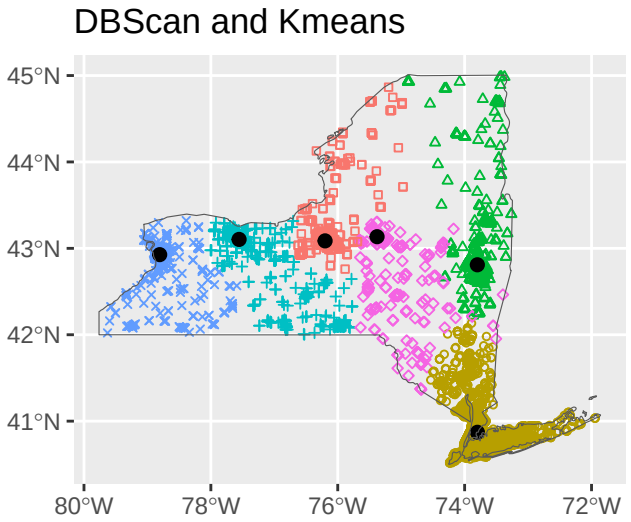

Fig 8. Clustering with DBScan and Kmeans

Now that all ATM locations were assigned for one cluster, it will be calculated some numerical results to be compared in the Results session.

b.6) Including an extra dimension for DBscan algorithm.

Using the same idea of the experiment of item b.3 it will be added two extra features in the original dataset and then applied the Dbscan algorithm in this dataset transformed.

The result shows that applying only the Dbscan generated plenty of ATM locations not assigned to any cluster again. Once again, Dbscan will be combined with Kmeans to test if it is possible reach some improvment on the outcome.

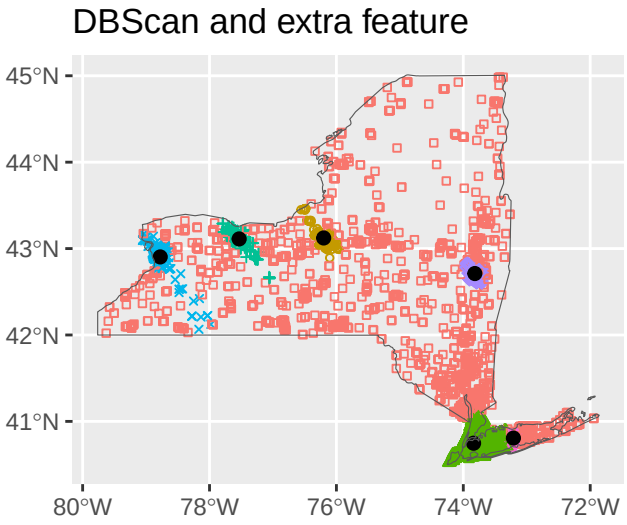

Fig 9. Clustering with DBScan

b.7) Combination of DBScan and K-mean.

As expected, now all ATM locations were assigned for a cluster.

Now will be possible to calculate some numerical results to be compared in the Results session.

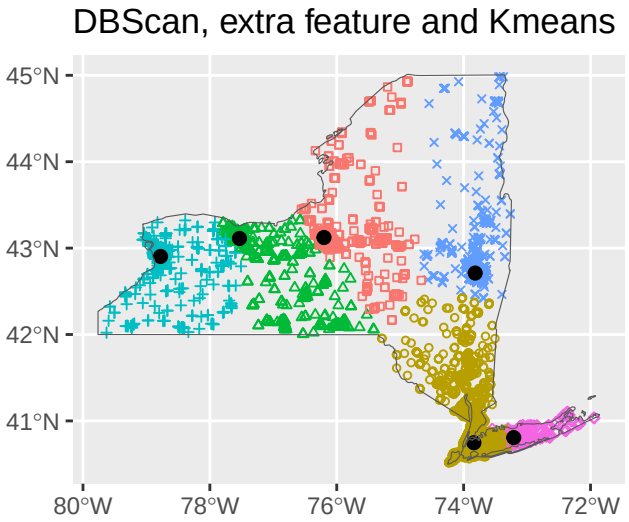

Fig 10. Clustering with DBScan and Kmeans

3. Results

The results of this work is presented in the table below. After several experiments with different configurations and parameters used with algorithms, it is clear that once has the number of cluster defined it is possible to reach the goals defined in the begining: distances less than 250km and also less than 90 minutes of travel time.

Is possible to note that there isn't any difference in the number of engineers according with the used algorithm, what was actually something expected, since it is basically a constant rate of 60 ATM per engineer. It could have difference, based on the distribution of ATM among the clusters caused by some sort of rounding, but with this particular dataset and clusters defined it not happened.

Also it is possible to conclude that had an high variability among the methods, regarding the distance and time travel, from an ATM and the centroid of the cluster it is part of. By the way, the centroids as very good candidates to be optimal localization as team bases, and also an optimal location for the warehouses.

	Max distance	Max time	Variance of # ATMs per Cluster	Amplitude of # ATMs per Cluster	# of Field Engineers
Kmeans best K=2	420.60	03:49	2309100.50	2149	84
Kmeans calculated K=6	233.65	02:07	606409.77	2024	84
Kmeans Extra Feature K=6	215.42	01:57	718151.77	2150	84
Dbscan and Kmeans K=6	222.88	02:01	4781.90	200	84
Dbscan, Extra Feature and Kmeans k=6	214.55	01:57	17173.87	374	84

Analyzing the table above, two solutions had better performance and are the winners, they are DBScan with extra features combined with Kmeans and K=6, and Kmeans with extra features and K=6.

But if we consider the distribution of how many ATMs each cluster would have and considering that a homogeneous distribution is desirable, then we should choose the last method: DBSCan with extra feature combined with Kmeans. In this option, in addition to good performance, a more equitable number of ATMs distributed among the clusters will also contribute to have smaller warehouses, distributed inventories and better logistics in terms of time and cost.

4. Conclusion

This project was quite challenging and has been the best opportunity to apply all the knowledge acquired in the previous modules of this course.

Basically we were able to apply what was learned in each module, but it was still necessary to research and look for other sources of information. The internet and R's online documentation were a valuable source of reference and helped with some details of the language.

This work could be continued implementing other cluster algorithms and/or combinations and compare results.

One immediate improvement would be also create some loop to explore and find the best eps and MinPoints for DBScan algorithm in this particular domain. Similar has been developed a loop to find the best K for the Kmeans algorithm considering some external parameter as a success criteria: 250km and 90 minutes.

Also should be great if had a real case to compare with these results, it means test the solution generated by this work with a real life SLM company.

Finally, this is just the beginning of the Data Science journey. Projects like this motivate us to continue learning and improve our knowledge.

Even when we achieve a reasonable result, we know we can always do better.

References

[1] Rafael A. Irizarry, Introduction to Data Science.
[2] Hadley W. and Garret G. 4th Edition. R for Data Science.
[3] Michael Hashler et al, Jornal of Statistical Software, dbsacn: Fast Density-Based CLustering with R