Clustering of Boarding Patterns in Public Transport*

*Note: Introduction to Data Mining Course Final Project Paper

Esra Kantarcı Çayır
Akdeniz University
Computer Science Engineering Dept.
Antalya, Turkey

***Abstract*—This research paper aims to analyze the patterns of public transport in Antalya, with the data given by the instructor from the date 18.12.2019, by using clustering algorithms applied in Python language.**

***Index Terms*—cluster analysis, boarding patterns, data mining, visualization, public transport, kmodes, kprototypes**

I. Introduction

Cluster analysis is an unsupervised machine learning technique for grouping and deducting the meaningful patterns. The data given is suitable for the cluster analysis because of the nature of the features of public transport statistics. In the data, columns are labeled as ”Line”, ”BoardingTime” and ”PassengerCount,” which are expected to have similar attributes in some specific grouping matches.The dataset consists of 1839 rows and 3 columns, Line and Boarding Time columns are in object datatype where Passenger Count is given as integer. The data preprocessing, analysis and visualization are done using Python and its machine learning and plotting libraries.

II. Objective

The objective of this paper is finding the meaningful patterns of the bus-boarding-trends out by using the given dataset and evaluating the results which are derived after using visualization and clustering methodologies.

The initial instinct one can have, as a part of the everyday public transport user in Antalya, is to add up some more columns for the transit route as well. This is due to fact that the grouping with line-codes or time-interval being obvious and natural; however, if we consider the everyday situation, there are popular and necessary routes (mainly) due to job, education or touristic reasons. Some bus stops in Antalya are more crowded than the others because of their centrality and location. Therefore, as a sub-task, after getting the unique line names, at least 3 bus-stop are to be added to our original dataset: beginning, mid, and the terminal stop of the transit route of the line.

The second approach is to check out the suitable ways to apply clustering algorithms to BoardingTime column values. When we check the data, it is an easier way to encode the date-time values rather than using time series approach because the data is from the same day and the time intervals are changed only by 30 minutes. Therefore, iterative mapping the integers with respective sequence is preferable and easy to recover the original values in the visualization.

III. Literature Review

There are many types of clustering analysis, therefore the first approach before the analysing the data is to choose the right tools for the learning. As the project suggests, checking out the documentary of the Python’s respective modules’ for the clustering is the first step to start choosing the right tools.

According to documentary and given use case details, K-Means Clustering and Agglomerative Clustering and Gassuian Mixtures are the suitable techniques for the given dataset. K-means clustering is pretty straightforward technique for the general approach using the distance between the points. Agglomerative clustering is a hierarchical clustering approach which gives transductive results due to connectivity and uses dendrogram for the visualization. Gaussian mixture is another approach which is suitable due to new parameters we had added up to our data, like coordinates of the stops and as we need to understand the density of the population’s choice by time and location, it is a powerful method.

As the values in the columns are partly categorical data, the another algorithms with the categorical data clustering techniques are also considered.

Henry Ralambondrainy proposed a conceptual clustering approach in November 1995, which is still functionally applicable in categorical variable cases. He mapped the categorical values into binary representations in order to create numeric cases. Since our dataset includes 51 unique values for lines and in total 40 different bus-stop names. Since ordinal encoding may end up affecting the model and one-hot-encoding enlarges the set unnecessarily big, Ralambondrainy’s approach is better and sustainable in this kind of task.

The solution he proposed was an alternated version of k-means clustering, which is called k-modes algorithm. In 1997, Zhexue Huang proposed another k-means clustering modification for the categorical data. In this case, he stated that using Gower’s similarity coefficient, standard hierarchical clustering techniques may overcome the issues with categorical and numeric values; however, k-means algorithm uses only numeric values. Since, k-means algorithm is pretty useful and powerful clustering technique among the unsupervised learning methodologies, he proposed a new variety of k-modes

Special thanks to Cebrail Çayır, for always supporting me.

algorithm: k-prototypes algorithm. It can be used in mixed-type objects, which is more preferable and useful in many cases.

In the good practices of clustering with categorical values, there is also sui-generis approach. You can create a mixture of techniques to encode or map for better accuracy. Using K-modes and K-protoyoes, Gaussian Mixture, Affinity Propogation did not give the expected results, therefore as a civilian who lives in Antalya and uses the given bus routes, the another approach is defining the ordinal categorical values mapping by heuristics and using the geographic coordinates of the routes.

Another issue to solve according to last feedback which was given during course was finding the relationship of the line choice by passenger behaviors. For such task, using Gaussian Distribution and Voigt Models (which were used for model-fitting for polynomial plots) is a way to model the data. Since, some of the bus routes' passenger count are different because of the population of the area, modelling with "line fit" using peaks is logical.

IV. Methodology

In order to apply CRISP-DM methodology for datascience purposes, first one should understand the task, then the data itself. As the task depends on purely on the author's evaluation of the data using clustering methods, the data understanding is extremely important.

A. Data Understanding

Data has 3 columns: Bus Lines with 51 unique values, the direction was not specified. BoardingTime, which includes date/time (generally) with 30 minutes interval and from the same day, which is from 18 December 2019, a normal Wednesday before COVID-19 pandemic hit Turkey. And lastly, PassengerCount, which is integer and column name is self-explained.

Line has categorical data type, BoardingTime should be processed to be split into only Time which is relevant since the date is from only 1 day (actually it is 2 days, however it is due to 00:00 value changes the day, but the turn of the second day can be ignored).

B. Hypothesis

As the first evaluation of the business understanding and data understanding phase of CRISP-DM, as a citizen of Antalya who used that bus routes, I had decided that there are mainly 3 important target pairs we can achieve:

- Time intervals - Passenger count: The relationship is the obvious one: when the sun is up, the passenger count should be higher.
- Line - Passenger count: Some lines have more popular stops and therefore crowded routes. For example KL08, LF09 should have more people than VS18.
- Line - Time intervals: Some routes have touristic sightseeing or cultural activity routes and some routes are used for going to work or educational areas such as universities, high schools etc. For example, VF01, DC15 should be more crowded at 07:00-09:00 and 17:00-19:00 intervals due to work/school starting and ending schedules.
- Line - Passenger Behavior: Some routes have busy times with different models according to passengers' behavior; some routes are used throughout the day, some are only busy during morning and some are busier at evening.

Another hypothesis is the importance of the geographic coordinates of the bus-stops. The Kundu, Döşemealtı, Aksu or Sarısu should be less crowded than the center places such as Markantalya, 100.Yıl, Meltem or Migros stops. Therefore, adding the start, middle and terminal stops to data would help in order to support this hypothesis.

Also, the longer the distance, the more people boarded into the bus can be evaluated with the coordinates and passenger count.

C. Data Preprocessing and Visualization

BoardingTime is split by "T", creating the columns Date and Time respectively. In order to change the object type of the Time column to numeric value, mapping is used for readability. Lines are enlarged with Start, Mid, and Terminal columns with the values retrieved from antalya-ulasim.com.tr for the respective Line values. After that, since One-Hot-Encoding or Label-Encoder might be problematic for the model, geographic coordinates with latitudes and longitudes of the stops are retrieved using Google Maps for the 40 unique bus stop values.

Mapping to the "Line" column had 3 different values: "Start," "Mid," and "Terminal" as stated. The middle points were the destinations which are more popular destinations and chosen heuristically. For example, Güllük, Meltem, Migros, 100.Yıl are stops which are strategically important for a citizen living in Antalya. Then these stops' latitudes and longitudes were retrieved from the Google Maps. Since the latitudes and longitudes are so close to each other, normalization and scaling were applied. In order to reduce the dimensions and ignore the noise, PCA was applied, but it was not helpful as it was in the graphical pattern recognition.

Using Matplotlib and Seaborn frameworks, after evaluation of the the visualizations of explanatory data analysis (EDA) of the original data-set, scatterplot of time and passenger count, and Pearson's correlation heatmap of the coordinates-time-passenger count (which are indicated below at the Results and Discussions section), appropriate clustering techniques were applied.

D. Clustering

1) Time Slices: Using scikit-learn version 0.24.2 and preprocessed dataframes, several clustering methodologies were applied for time based clustering. K-means clustering algorithm was applied several times with different approaches. In k-means clustering, the drawback is choosing the "k" by hand. However, using inertia chart, which is called "Elbow Method", we can use a better value for k. Even though the biggest difference was shown in k=2 due to 2 peaks(at morning and

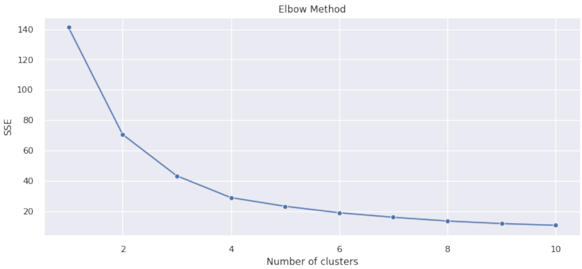


Fig. 1. Elbow Method

at evening, possible cause: work/education start and ending hours), the intuition says k=5 is better for group analysis. In order to apply k-means clustering, scaling was applied to stop coordinates as well.

K-prototype clustering is Huang's mixed variable (categorical and numeric) algorithm version of k-means clustering, as previously explained. As the cost function shows, the biggest difference was done at k=2, however the cost was still too high.

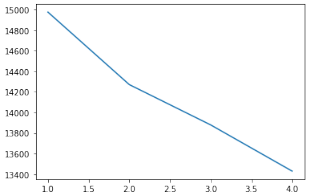


Fig. 2. Cost Function for K-Prototypes Clustering

Agglomerative clustering is a bottom-up approach and was not as efficient as it was expected. There are some outlying points that seem unnatural in the time-slice graph.

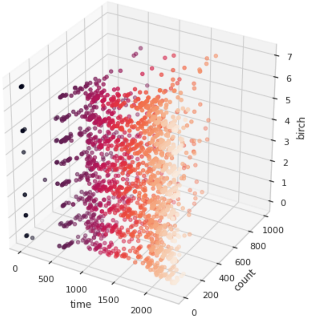


Fig. 3. BIRCH Clustering with Time-Count-BIRCH in 3D Chart

Also, we did not have that big data for sampling, therefore using k-means was more acceptable than using BIRCH for this task. As expected, the best results were done by GMM approach due to its good density evaluative algorithm and Mahalanobis distance.

The clear clusters by time-slices are pretty desirable and understandable. Still, k-means clustering results also similar to GMM clustering, therefore k-means with k=5 is chosen as main clustering method.

2) Popularity Behavior Based Clustering: As one can see from the visualization of the data by Time(x-axis) and PassengerCount(y-axis) regarding the Lines, there are some similar patterns with peaks during the day. These ”popular” time indicator peaks can be modeled by using Voigt and Gaussian Distribution models. In this case, density was analyzed, so the boarded passenger count was not a primary case - but the distribution during the day of the respective line modelling was the primary task.

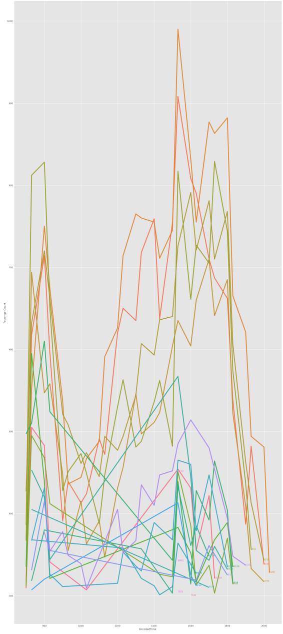


Fig. 4. Lines with higher than 300 Passenger Boarded

For example, VF01 bus line (Fig.4) has 2 big local maximums during the day. This indicates that the route is mainly used for business or education related issues. At noon, line seems stable and above median, so it is overall a popular line.

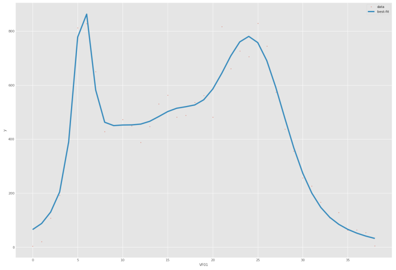
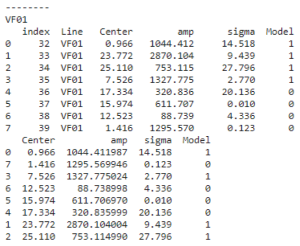


Fig. 5. Line fitting of VF01 using Voigt and Gaussian Distribution models

V. Results and Discussions

A. Visual Evaluations

1) EDA: As the initial investigation with explatory data analysis tools show, average count value of a line changes between 100-200 passengers. The density graph is right skewed.



```
--------
VF01
   index  Line   Center       amp      sigma  Model
0     32  VF01    0.966  1044.412     14.518      1
1     33  VF01   23.772  2870.104      9.439      1
2     34  VF01   25.110   753.115     27.796      1
3     35  VF01    7.526  1327.775      2.770      1
4     36  VF01   17.334   320.836     20.136      0
5     37  VF01   15.974   611.707      0.010      0
6     38  VF01   12.523    88.739      4.336      0
7     39  VF01    1.416  1295.570      0.123      0
   Center          amp   sigma  Model
0   0.966  1044.411987  14.518      1
7   1.416  1295.569946   0.123      0
3   7.526  1327.775024   2.770      1
6  12.523    88.738998   4.336      0
5  15.974   611.706970   0.010      0
4  17.334   320.835999  20.136      0
1  23.772  2870.104004   9.439      1
2  25.110   753.114990  27.796      1
```

Fig. 6. Example Model Outputs for VF01

The distribution of the bus boarding times are almost regular: there are less active buses between 24:00 - 06:00 time interval.

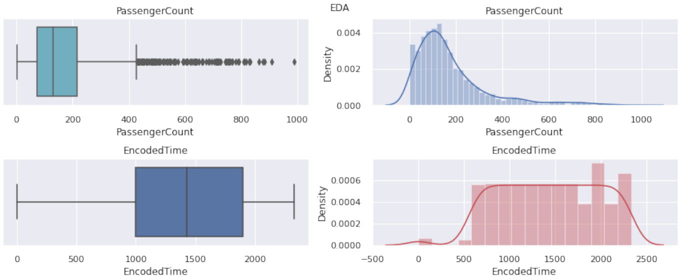


Fig. 7. EDA - PassengerCount

2) Bus Schedules: The buses are less frequent at night to morning, however the bus frequency is uniformly distributed between the time intervals of 06:00 - 23:30.

3) Popular Lines: As expected, specific lines, such as KC06, KL08, LF09, LF10, VF01, are more popular than the others, even though the bus schedules are (almost) uniformly distributed. This shows the route is important.

4) Coordinates and Popularity: As the 3d plots show that the bus routes with the specific center points of the Antalya has more popularity and the popular stops are mostly near to Konyaaltı rather than Muratpaşa part.

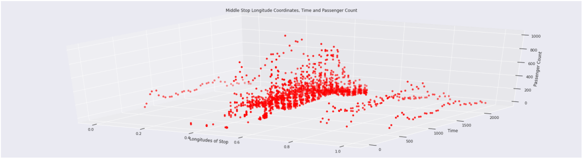


Fig. 8. Middle Stop Longtitude - Time - PassengerCount

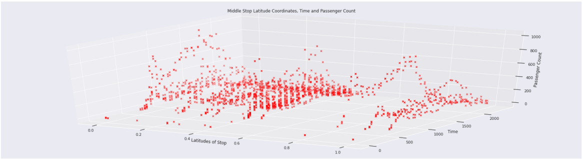


Fig. 9. Middle Stop Latitude - Time - PassengerCount

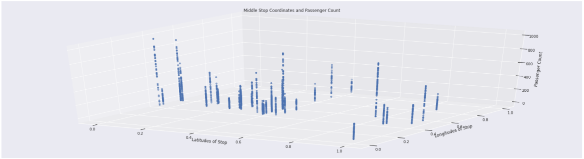


Fig. 10. Middle Stop Latitude - Longitude - PassengerCount

As the hierarchical clustering method, dendrogram shows the data has 3 main clusters and the Euclidian distance calculation using Ward's method. The linkage is not so human readable. The K-means clustering seems a better and more intuitive approach.

As the first visualization of the data, one can easily understand that there are peaks at certain time periods during the day, which are highly related to business and educational time.

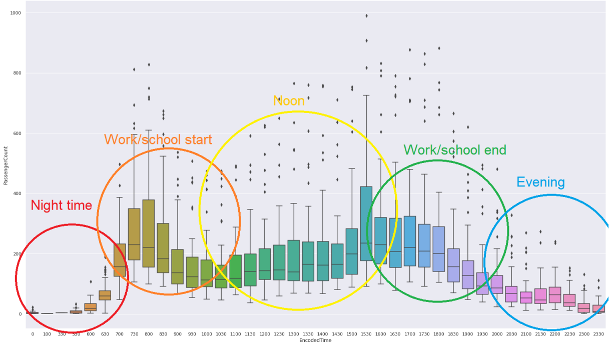


Fig. 11. Time - PassengerCount Distribution Bar Chart

K-means clustering gives the similar results as expected, and this clustering is logically acceptable.

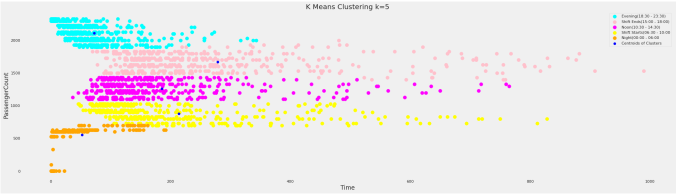


Fig. 12. K-Means Clustering with k=5

B. Correlations of the Pairs

1) Time and Line: The lines are almost uniformly distributed by the time, at night public transportation is rarely available, however the lines are distributed uniformly by the day time. The most frequent line has 39 boarding time where the least one has 33 boarding time count.

2) Time and PassengerCount: The most boarded time intervals are between 06:00 and 18:00, which is obviously for active social and business participation of the people. People should go to work or school at the morning between 06:00-10:00 and need to come back home at evening between 15:00

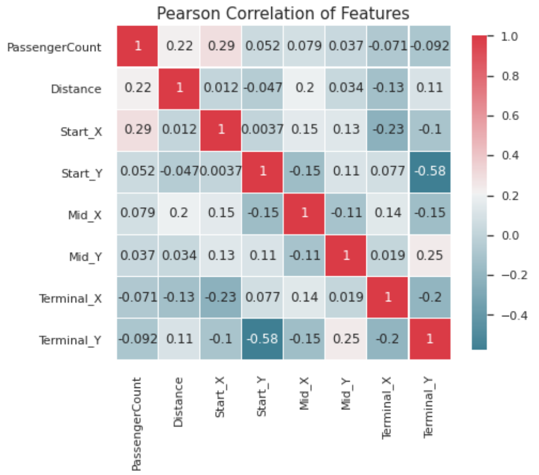


Fig. 13. Coordinates, Distance and PassengerCount Heatmap

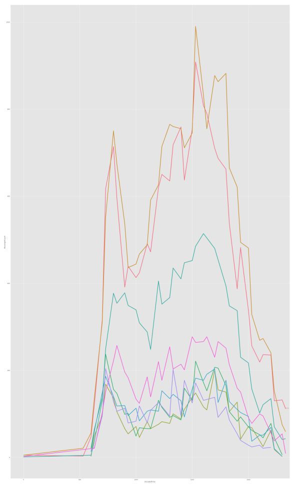


Fig. 14. Example Plot: Huang (K-Prototypes) Clustering, Cluster 3

and 19:00. If the date was not from winter time of Antalya -but from summer time-, for touristic reasons, we could see that the passenger count was even higher after 13:00 and busier evening than December's given data.

3) PassengerCount and Line: The lines which has more centeralized stops like Meltem, 100.Yıl, Migros or popular routes such that VF01, KL08, LF10 have more passenger counts, which is also because of the business. Also, the necessary routes such as DC15 and DC15A which has Döşemealtı stop or UC11 which has Uncalı and Güllük stops are also pretty popular due to Antalya having less alternative bus-line options of the same routes.

4) Stops and PassengerCount: Middle stops are generally more important than the start or terminal points according to crowded lines. However, this also due to choice of the author since the middle points were selected heuristically. This issue is still important to consider, because the ones which are near to west-side of Antalya is more popular destinations than the east-side of Antalya, even when we include Start and Terminal stops into the big picture. Therefore, we can say that, Konyaaltı is more popular destination than the Muratpaşa.

5) Distance and PassengerCount: Distance and passenger count was not correlated, which is against the initial hypothesis.

C. Line Fit Modelling and Clustering Results

All the models have some similar time-related high peaks due to shift starts and endings. However, some bus routes have different passenger density during the day time.

- Cluster 0: KC35, MF40, MZ78, TB72, TC93, TCD49A, TL94, VC57, VC59, VF66, VML55A - These have 4 sharp arisen peaks during the day. Respectively morning, noon(around 12:00 and 15:00) and evening peaks are highly popular. But during the day, and especially after noon, the popularity is still higher than the usual. At 22:00, there is also a local maximum which are not often seen in other clusters. These lines are not the busiest routes, the average of the boarded passengers count is around 200.
- Cluster 1: AC03, DC15A, TC16, TCD49, VF01, VL13, VL13A - The peaks are mainly on morning and evening, one can easily say that passengers use these lines for business or education related reasons. These lines have a stable and continuously increasing line during the noon, which indicates these lines are preferred by passengers.
- Cluster 2: 511, AF04, AF04A, CV67, KC33, KC35A, TCP45, UC11, UC32, VF63 - These lines have very frequent ridges and sharp edges during the day. Especially during the mid-day (10:00 - 15:00) they have increasing and multiple peaks. Also, at 22:00 we see another local maximum as well. The passenger count is lower than the other clusters' mean. Assuming these routes do not have alternative bus lines for their certain stops is logical.
- Cluster 3: CV48, KC06, KF52, KL08, LC07, LC07A, TK36, VC53 - These lines are the most popular ones during the day and night. One can say that these lines are the busiest and the most crowded ones after evening. There are multiple peaks during the day, as well. Therefore, assuming that these lines' bus stops are popular destinations is logical.
- Cluster 4: DC15, GM24, LF10, MC12, VF02, VML54, VS18 - This cluster have different behavior during the pre-evening session. Between 15:00 and 19:00 there are 3 drastic peaks and then there is a dramatic loss of popularity afterwards. The peaks at 15:00 and 16:00 makes one think that students are the one who mainly uses these lines.
- Cluster 5: CV14, CV47, KM61, LF09, ML22 - The morning and pre-evening sessions' popularity is almost the same in these lines. During the day, one can see decrease between 09:00 - 12:00, then the plot is followed by sharp and increasing ridges. The main difference among the other clusters are seen at the evening session. There are Voigt-modeled peaks after 16:00, which indicates that

these routes are preferred at evening and night session as well.
- Cluster 6: FL82, KPZ83, TC16A - The morning peaks of other clusters are often lower than the evening peaks; however in this cluster passengers preferred these lines for morning routes. One can assume that these lines have alternative bus routes at evening due to loss of popularity after 16:00.

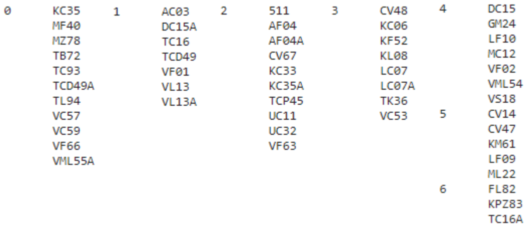
VI. Conclusion

The given dataset was preprocessed by adding line details such as start, middle and terminal stops for the given lines and their respective latitudes and longitudes. They are analyzed using clustering methods and necessary visualization forms. In conclusion, the most appropriate segmentation was popularity by time analysis. As the Fig.12 shows, the time slices are;

- Night: 00:00 - 06:00 (Unpopular Boarding)
- Start of the Day: 06:30 - 10:00 (Business and Educational Shift Start)
- Noon: 10:30 - 14:30 (Leisure Day Time)
- End of the Day: 15:00 - 18:00 (Business and Educational Shift End)
- Evening: 18:30 - 23:30 (Leisure Evening Time)

The day time has more passengers than the evening and night time, which also causes the biggest breaking point of k-means clustering algorithm into 2 slices at Elbow Method. But 5 slices are more suitable for the nature of the popularity peaks.

The distance and coordinates do not necessarily indicate strong relationship with passenger count. However, the routes with popular stops are more crowded than the others. Another inductive evaluation is the stops with less alternative bus-routes are more crowded than the average routes, such as Döşemealtı and Uncalı.



```
0    KC35      1    AC03     2    511      3    CV48     4    DC15
     MF40           DC15A         AF04          KC06          GM24
     MZ78           TC16          AF04A         KF52          LF10
     TB72           TCD49         CV67          KL08          MC12
     TC93           VF01          KC33          LC07          VF02
     TCD49A         VL13          KC35A         LC07A         VML54
     TL94           VL13A         TCP45         TK36          VS18
     VC57                         UC11          VC53     5    CV14
     VC59                         UC32                        CV47
     VF66                         VF63                        KM61
     VML55A                                                   LF09
                                                              ML22
                                                         6    FL82
                                                              KPZ83
                                                              TC16A
```

Fig. 15. Clusters using K-Prototypes Approach (k=7)

The popularity of the lines during the day have some patterns due to behavior and motive of the passengers, such as educational or business related transportation or due to options of the bus lines with same and desirable stops, or routes with popular destinations. As explained in the Line Fit Modelling and Clustering Results (subsection V.C.) part, there are 7 different clusters which are derived using K-Prototypes approach of Huang and line-fit-properties of the data, and the clustering results can be seen at Fig.15. For the further visualization and methodology details, the code can be seen at the Google Colab link which is given below in the bibliography section.

References

[1] E. Kantarcı Çayır, ”Boarding Patterns in Public Transport”, May.31, 2021. [Online] Available: https://colab.research.google.com/drive/18gTj4vV1sl7HoPlCmFQdpOlum_RpG3q6

[2] C. Ostrouchov, ”Peak fitting XRD data with Python”, Apr.13, 2018. [Online] Available: https://chrisostrouchov.com/post/peak_fit_xrd_python/

[3] H.Ralambondrainy, ”A conceptual version of the K-means algorithm”, Elsevier B.V., 1995

[4] Z.Huang, ”Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values”, Kluwer Academic Publishers, 1998

[5] J. Brownlee, ”10 Clustering Algorithms With Python”, Apr.6, 2020. [Online] Available: https://machinelearningmastery.com/clustering-algorithms-with-python/

[6] S. Prasad, ”Types of Clustering Algorithms”, Jul.5, 2020. [Online] Available: https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

[7] Explorium Data Science Team, ”Clustering — When You Should Use it and Avoid It”, Feb.3, 2020. [Online] Available: https://www.explorium.ai/blog/clustering-when-you-should-use-it-and-avoid-it/

[8] A. Ruberts, ”K-Prototypes - Customer Clustering with Mixed Data Types”, May.16, 2020. [Online] Available: https://antonsruberts.github.io/kproto-audience/

[9] NicoDV, ”K-Modes Github Repository”, [Online] Available: https://github.com/nicodv/kmodes/blob/master/kmodes/

[10] T.Beuzen, ”Unsupervised clustering with mixed categorical and continuous data”,May.20,2020. [Online] Available: https://www.tomasbeuzen.com/post/clustering-mixed-data/

[11] Lost Stats, ”Line Graph with Labels at the Beginning or End of Lines”, [Online] Available: https://lost-stats.github.io/Presentation/Figures/line_graph_with_labels_at_the_beginning_or_end.html

[12] T.Beuzen, ”Unsupervised clustering with mixed categorical and continuous data”, May.10,2020. [Online] Available: https://www.tomasbeuzen.com/post/clustering-mixed-data/