Project Title:	Data Analysis of Dublin Bus System
Student ID:	18210747
Student name:	Jasmeet Kaur
Student email	jasmeetkaur2@mail.dcu.ie
Chosen major:	Data Analytics
Supervisor	Mark Roantree
Date of Submission	11-08-2019

DISCLAIMER

An report submitted to Dublin City University, School of Computing for module CA685 Data Analytics Practicum , 2018/2019.

I understand that the University regards breaches of academic integrity and plagiarism as grave and serious.

I have read and understood the DCU Academic Integrity and Plagiarism Policy. I accept the penalties that may be imposed should I engage in practice or practices that breach this policy.

I have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the sources cited are identified in the practicum paper references.

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

By signing this form or by submitting this material online I confirm that this paper, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online I confirm that I have read and understood DCU Academic Integrity and Plagiarism Policy.

Name: Jasmeet Kaur

Date: 11-08-2019

Data Analysis of Dublin Bus System

Jasmeet Kaur
School of Computing
Dublin City University
Dublin 9, Ireland
jasmeetkaur2@mail.dcu.ie

***Abstract*— An analysis of the features and behavior of bus stops can help to find clusters of similar stops. In this paper, exploratory analysis and three clustering methods i.e. K-means, DBSCAN and Hierarchical are performed to extract hidden information from large database provided by Bus Átha Cliath (BAC) of Dublin, Ireland. The database records the actual and aimed time for buses, stop number, route number and difference between actual and aimed time. In this paper, first of all data is prepared for the clustering process by applying some pre-processing methods in which new attributes are derived to find distance between stops and delay between aimed and actual time. After this, all the three clustering methods are applied and their performance is measured by silhouette score, according to which K-means is better among all as it is having highest silhouette score. Clusters obtained by K-means show three different zones on the map. Stops with red flags need attention as there is severe delay, stops with yellow flags have normal delay and stops with green flags are operating in perfect manner.**

Keywords—Transportation; Open Source Routing Machine (OSRM); Principal Component Analysis (PCA); Clustering; Clustering methods; Silhouette score.

I. INTRODUCTION

Public transport in Dublin underwent a major expansion in recent years. As per records of Dublin City Council (2015), bus share among public transport was highest that was 28.9%. The number of leap card users were reported to have increased by 500,000 to 2.5 million users in September 2017. It is very important to understand how public use the transportation as it would be helpful to plan new routes to make public transit more convenient, effective and efficient [12]. Effective public transit would encourage people to use it more often. In transit system, it is necessary to find common patterns between different stops and areas, so that required attention can be paid to those locations which is facing more delays, large number of boarding and alighting as there could be number of reasons for these patterns. Once the reasons are known, required facilities can be provided at such locations.

As transit system is recording data every second for each route and it is difficult to handle such large data to extract useful information from it. Number of methods are there to tackle such problems in the field of data analytics. Data mining is very useful process in terms of extracting desirable information from large data and clustering is a critical tool for data mining and analyzing data. In transport systems, it is desirable to find commonalities and patterns between different stops, time and weekly patterns.

In this paper, various data mining techniques are used such as deriving new attributes from given attributes, PCA, standardization and clustering. Three new attributes are derived, first is the distance between each stop, and it is calculated from given latitude and longitude. Second to find delay between aimed and actual time and last is delay per kilometer which is calculated from distance and delay. Then data is prepared for clustering by standardize it and by applying dimensionality reduction method i.e. PCA.

Clustering is used to group similar data to classify data into particular group. In clustering, some details are overlooked in exchange of data simplification. Clustering provides concise summaries of data therefore related to many disciplines. So, it can be used in wide range of applications. Large datasets, data with many attributes and attributes of different types usually use clustering techniques to find suitable results. This complex data tends to impose severe computational requirements that present real challenges to classic clustering algorithms which further led to the development of powerful data mining clustering methods [11]. In this paper three clustering techniques are used namely K-means, DBSCAN and Hierarchical. In data mining, clustering is an important technique in which similar data is put into one group or cluster and the remaining data in the other group. The clustering is an unsupervised learning technique. In terms of identifying different patterns in number of business applications, clustering is very beneficial technique.

Some briefing about clustering algorithms that are used in this paper are as follows:

- *K-Means Clustering:* It is most well-known clustering algorithm. This algorithm is quite fast and has a linear complexity *O(n)* [9].
- *Density-Based Spatial Clustering of Applications with Noise (DBSCAN):* It is density-based technique having similar features as mean-shift algorithm but with more advantages such as identifying outliers as noises [9].
- *Agglomerative Hierarchical Clustering*: It follows two approaches that are top-down and bottom-up. It has many advantages at the cost of lower efficiency [9].

Bus traveler system is one of such systems which involves large amount of data with number of attributes and to understand such data is a difficult task. For such type of data, clustering techniques can be useful to extract required information. The main issue with today's bus traveler system

is that these do not show the accurate arrival time to the passengers which can make public transportation less prominent among people.

The transit networks'arrival time is random because of fluctuations in travel times on links, dwell times at stops and delays at intersections which are spatial and temporal. Due to these fluctuations, passengers have to suffer as waiting time keeps varying which of course deteriorates the service quality that discourage people to use public transportation [2]. Therefore bus transit system is needed to analyse regularly to find any patterns, so that if any particular location or route is facing severe delays, then it must be addressed to keep the public transportation prominent among public.

First of all, the collection and analysis of bus data is required to detect any patterns in the data. The detection of delays is a very complex task that involves the analysis of large amounts of data. By finding the anomalies in the dataset, the search space is reduced and attention can be given to required places only. To detect anomalies there are number of data mining techniques that have been proven useful. To do this effectively experts' abilities as well as powerful interactive tools work together [8].

The overall aim is to implement three clustering methods and to compare them in terms of their performance by silhouette score to find out which clustering method would be suitable for given domain. This paper analyses various patterns in given data at various levels i.e. stop, hourly and weekly. Description of related work in this domain and clustering field is given in Section II which helped in guiding this analysis. Section III gives the description of data to understand the domain. In Section IV, all the methods and techniques applied on the data are explained with their outcomes. Finally, Section V gives the conclusion and future scope.

II. Related Work

Steven et al. (2002), discussed about providing accurate arrival time about vehicles in traveller system by predicting via two Artificial Neural Networks (ANNs) that are the stop-based ANN and the link-based ANN. Both the networks were integrated with an adaptive algorithm to adapt the prediction error. However, for multiple intersections between stops, the stop-based ANN is preferable and in contrast, for stops with few intersections, the link-based ANN is preferred. The hybrid ANN is developed as an extension but still further improvements are possible [2].

Michael et al., compared between agglomerative hierarchical clustering and bisecting K-means on the basis of many evaluation factors like entropy, F-measure and overall similarity. Sometimes both of these approaches are applied together to obtain the best results because K-means provides efficiency and agglomerative hierarchical clustering provides quality. In this paper, bisecting K-means is proved to be the good clustering technique in terms of processing time [1].

Jayakrishna et al. (2004), discussed number of factors that influence the estimated arrival time of buses such as geometric conditions, route length, number of intermediate stops and intersections, turning movements, incidents, ridership variation at stops, weather conditions time of the day and day of the week. The purpose is to build a model that estimates reliable arrival times. In this research, collected data is processed, analysed and refined; and independent variables are examined for their behaviour and impact. It is mentioned that while dealing with large databases, it becomes important to carefully inspect the data for flawed entries and inconsistencies. The reasons behind the incorrect data could be the machinery malfunction, human errors, software errors and other causes. Suggestion is given to develop a dynamic algorithm and to integrate with developed model to enhance the accuracy of prediction of transit models [3].

Daniel (2011) explained the hierarchical agglomerative clustering which is useful in the field of machine learning. In this scheme, first of all data set is partitioned into singleton nodes and then consequently merging of mutually closet nodes is done into a new node and this process keeps on going until the last node left that holds the whole data set. In this article number of other clustering schemes are mentioned with their algorithms like primitive clustering algorithm, generic clustering algorithm, nearest -neighbour-chain algorithm and many more. All these algorithms are efficient in different schemes according to their functionality and performance [4].

Saurabh and Inderveer (2014) explained the concept of clustering and types of clustering. Clustering is classified into three main categories named hierarchical, partition and density-based clustering, and further sub-categorized into number of types. A framework is presented to compare different clustering methods on the basis of type of dataset, execution time, cluster quality, merits and demerits. It is observed that different methods are suitable for find some particular outcomes such as partitioning clustering performs well for big datasets. On contrary, for large datasets with real and synthetic data, hierarchical clustering is good in addition to speedup execution time. Density-based clustering algorithms are good at filtering noise and in providing good quality clusters. Generic algorithms are domain-specific which are used for only specific fields like bio-medical or census as it provides good execution time as well as good quality clusters [7].

Ali et al. (2014) discussed about the clustering algorithms from the very first till the time of their research. Clustering is defined as method in which data are divided into groups in such a way that objects in each group share more similarity than with other objects in other groups. Clustering algorithms help to deal with complexity and computational cost that subsequently enhance the scalability and speed. Different clustering techniques are explained which are classified into two major types: single-machine clustering and multiple-machine clustering. The trend of improvements in clustering algorithms is discussed in the paper [5].

Adil et al. (2014) explained the different types of clustering algorithms which are used to handle large datasets. A framework is created to classify clustering algorithms in order to compare them on the three aspects of dataset named volume, velocity and variety which are further subdivided on related factors. Most representative algorithms from different types are analysed over datasets in order to compare their outputs [6].

Laura et al. (2016), discussed about the transit system of Rochester, New York. Exploratory analysis of ridership has been presented and clustering individual stops into desired groups so that these results can be further passed on to policymakers and city planners. In this analysis, similarity and dissimilarity between stops is found out and clustering is done on the basis of stops. Here, also boarding and alighting patterns are examined on different days of week and hours of the day. Analysis is done on both of these characters that is usage and location [13].

III. DATA

Database for Dublin Bus System was provided by Bus Átha Cliath (BAC) of Dublin, Ireland. This database consists of eight tables with different information in all of them. Among these eight tables, data in two tables is missing. First table named 'getdestination' consists of stop ids and their names. Second table named 'realtimedata' consists of real time data of buses which consists aimed time, actual time, route numbers, stop numbers, stop names, service provider, direction of bus (inbound or outbound), aimed and expected due time. Third table named 'route' consists of route numbers and their destination names. Fourth table named 'similarity' consists of similarity between two routes. Fifth table named 'stop' consists of latitude and longitude for each stop. Last table named 'stopdatabyroutes' consists of information about the sequence of stops followed by different routes in both directions i.e. inbound and outbound.

IV. METHODS

As inputs are in the SQL dump or database dump which is usually used for backing up a database so its content can be restored in case of data loss. Therefore, it is needed to parse the database to get names and values of each table. A file in python is created which makes a connection between database and python to read SQL file line by line and to create necessary csv file in correct format. Main function of this file to take the SQL database as an input and to create csv files inside a folder that is ingested for the subsequent analysis.

The very first step is to understand the data that is about bus traveler system of Dublin, Ireland. Data cleaning is important pre-processing step to understand the exact behavior of data as data may contain irrelevant and incorrect data. Number of efficient tools are present to clean the data by which data will become more useful as cleaning will remove duplicate entries, filling up missing values, identifying or removing outliers and resolving inconsistencies [10].

First file to analyze which is named 'stopbyroute'. This contains route number, order of stops to follow by any particular route in a particular direction i.e. inbound (I) and outbound (O), latitude and longitude for each stop, stop numbers and their addresses. Main idea is to create new column which contains distance between two stops on a route. To do so, first of all, another column is needed to create for the last stop by looking up the last order number with same direction and route for every observation. To calculate distance between stops, Open Source Routing Machine (OSRM) is used which is an open-source router designed for use with data from the OpenStreetMap Project. It gives the routing distance between two points. The reason to use this method to get the real distance between two points rather than straight line distance.

Next file named 'realtimedata' which is used for processing contains real time data about buses like at what time these are scheduled to leave a stop and at what time these left the stop. Date and time are mentioned in same column. Other than this, table 'realtimedata' contains name of service provider, day, time, month, stop name, direction etc. Seven columns in this table are empty, hence these are removed. Four columns have same information, so three of them are eliminated. Similarly, aimed time and expected time is mentioned in two columns, so only one from each is kept for the analysis. In analysis, we considered route number, stop number, direction, aimed and expected time. A new column is added which records time difference in seconds between aimed and expected time. In this column, negative values show that bus is at stop before the expected time and positive values represent, there is delay. Zero shows that bus is on time. Values in direction column are stored as 'Inbound' and 'Outbound', however, to save storage space, values are renamed as 'I' and 'O' respectively.

After this, real-time data is merged with distance data that already contains the details of last stop. Here the data is getting prepared for clustering. There are some routes which are in 'stopbyroute' table but not in 'realtimedata' table, as a result, these routes are removed from the first table and joining can be performed. Then, column names are cleaned as their names vary in both tables but containing the same information. A new column is derived which contains delay per kilometer.

Next step is to aggregate data on stop and direction level for clustering. To aggregate, there is need to choose a measure like mean or median. To aggregate distance, median is chosen as one stop can be part of different routes, but it will lead to different distances. To average out a stop, median is a good measure. On the other hand, for delay and delay per kilometer, median came up with value zero and after checking it showed all the mid values are zero, that is why it is returning median as zero. So here mean is chosen.

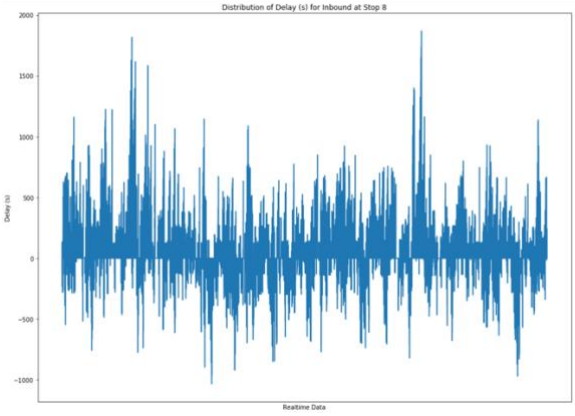

Figure 1 Delay in Realtime Data for Stop 8

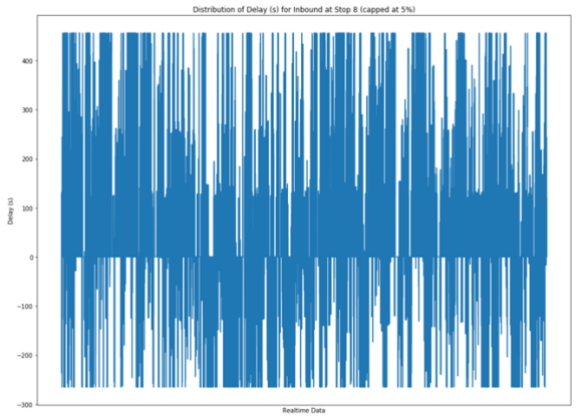

Figure 2 Delay in Realtime Data for stop 8 capped at 5%

Similarly, delay for each stop is plotted which shows the average delay in seconds for each stop. There are some peak values in plots which shows the outliers and reasons behind these outliers could be road accidents or some unusual events on the roads.

Next step is to perform clustering. Before that to an insight into data is taken.

	StopNumber	DistanceTraveled_km	Delay_s	EarlyCount	DelayCount	OnTimeCount	Delay_s_per_km	Delay_m
count	749.000000	749.000000	749.000000	749.000000	749.000000	749.000000	749.000000	749.000000
mean	2293.436582	0.814913	35.028986	3136.160214	6735.244326	13373.662216	88.774506	0.583816
std	1840.875408	1.026998	45.183535	3191.390441	6234.056134	10532.647385	148.372124	0.753059
min	8.000000	0.105900	-135.543103	0.000000	0.000000	10.000000	-221.910778	-2.259052
25%	847.000000	0.269900	0.000000	518.000000	1586.000000	4577.000000	0.000000	0.000000
50%	1656.000000	0.399800	19.443862	1977.000000	4334.000000	10533.000000	21.228423	0.324064
75%	3584.000000	0.986100	60.293801	5031.000000	10787.000000	19652.000000	108.844556	1.004897
max	7671.000000	9.085000	197.450820	14966.000000	34055.000000	46252.000000	1063.535257	3.290847

Figure 3 Insight into Data through Mean, Standard Deviation

In figure 3 values from standard deviation show that data is skewed. To check skewness in data, plots are made for each concerned attribute. As it can be noticed that data is skewed and to address skewness, standardization is done because it makes the data scaled with mean 0 and standard deviation 1.

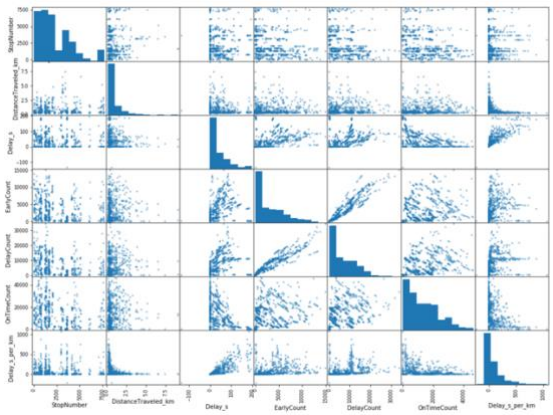

Figure 4 Plots between Attributes

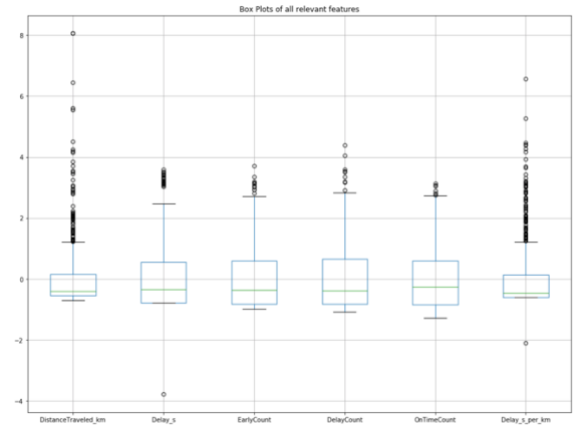

Figure 5 Boxplots for each Attribute

After transformation, boxplots for each column are created to validate the outcome. In figure 5, data seems to be standardized. However, it still has some outliers, but there is no need to perform any outlier treatment like capping as it may negate its effects in clustering. Because here outliers represent occasional delays that are also needed to identify.

Before diving into clustering, dimensionality reduction is performed to find optimal Principal Component Analysis (PCA) components. PCA is performed on scaled data and is resulted into three components which are explaining the variation without any bias. Results of clustering can be compared by applying it on transformed data as well as on reduced data.

Next step is to perform three types of clustering i.e. K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Agglomerative Hierarchical clustering on PCA transformed data. To validate consistency within the cluster, silhouette measure is used. The silhouette score measures how similar an object is to its own cluster compared to other clusters. Its values range from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

First of all, K-means clustering is performed on PCA transformed data as well as original data, however, clustering better performed on original values as silhouette score is higher for original values. Clustering is performed for different values of k range from 1 to 50 to check standard squared error and it is found that squared error is least for value of k equals to 3.

As it is shown in figure 7 that sum of squared error is least for k=5, however it is not efficient from business point of view, that is why value 3 is chosen for k. In figure 8, distribution for cluster levels is shown.

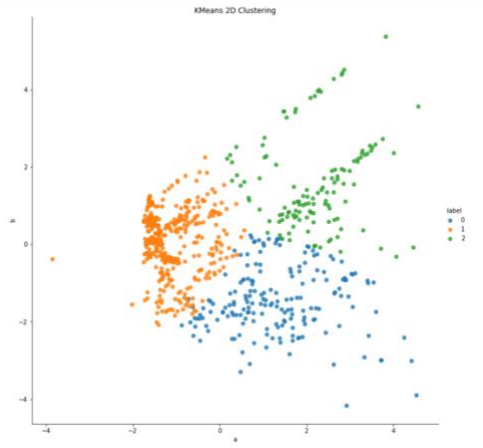

Figure 6 K-Means Clustering

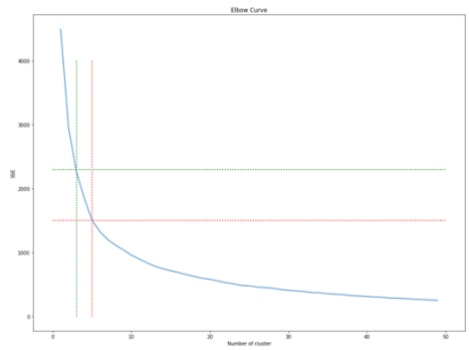

Figure 7 Number of Clusters vs Sum of Squared Errors

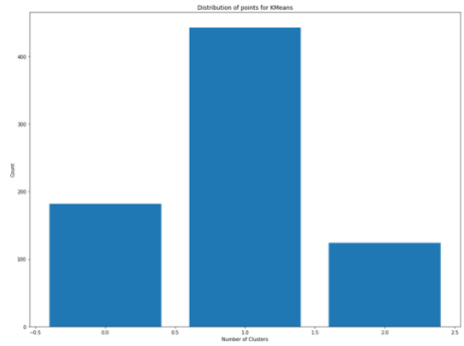

Figure 8 Distribution of Points for K-Means

Next clustering which is implemented is DBSCAN which gives the following outputs. As it can be noticed that in figure 9, data is not distinguished properly as there are more data points associated with only one point. Also, in figure 10, it can be noticed that most of the points are under one level only. Results of this clustering does not seem to be promising.

Figure 9 DBSCAN Clustering

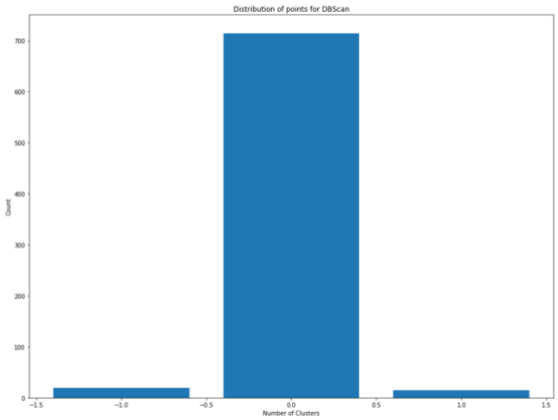

Figure 10 Distribution of Points for DBSCAN

After this, Agglomerative Hierarchical clustering is applied and here represent the following results. In figure 12, it is clearly shown that with increasing number of clusters, sum of squared error is also increasing. Figure 13 and 14 shows the distribution of data into clusters.

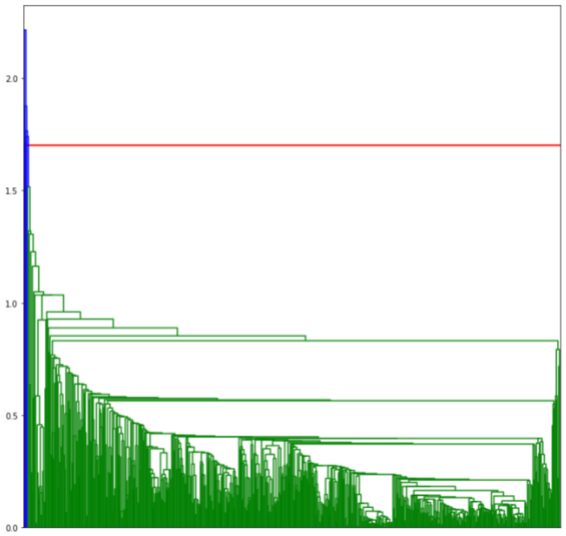

Figure 11 Dendrogram for Hierarchical Clustering

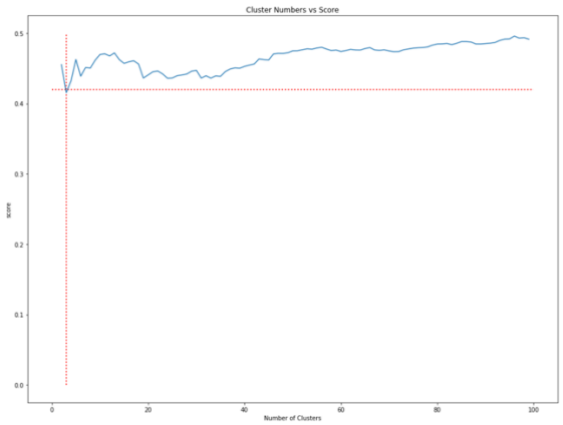

Figure 12 Number of Clusters vs Sum of Squared Errors

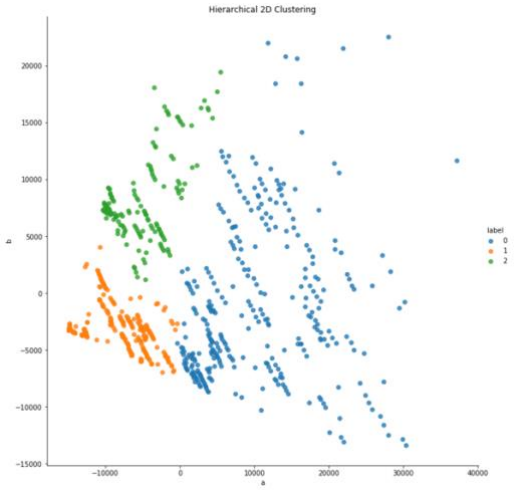

Figure 13 Hierarchical Clustering

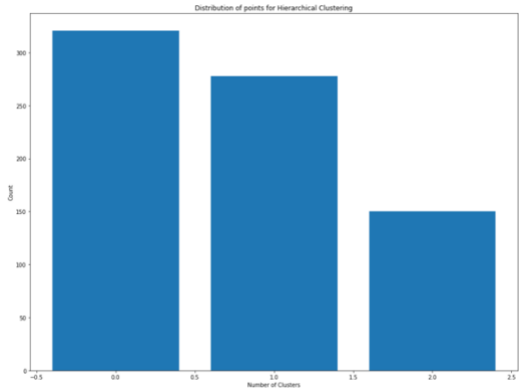

Figure 14 Distribution of Points

As K-means gave the highest silhouette score, so next step is to visualize clusters that were represented with 0,1 and 2 in figure 6 and distributions of these clusters are shown in figures 15, 16 and 17 respectively. These distributions show number of counts for low, average and high delays per kilometer. Clusters obtained from K-means clustering are now represented on map in figure 18 by flags of different colors. To represent cluster 0, 1 and 2 flags of yellow, green and red are used respectively. Stops with yellow flags are experiencing moderate delays, so there is no need to pay immediate attention. On the other hand, stops with red flags are facing severe delays, it can be suggested to pay important attention to these stops so that these delays can be reduced to some point as long delays would discourage public to use public transport. Stops that are marked with green flags are operating in good manner as there is no or very low delays.

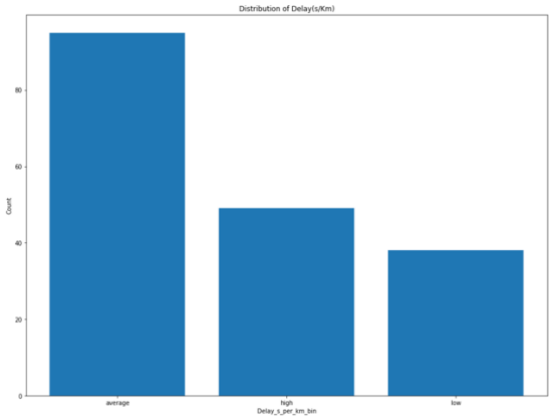

Figure 15 Cluster 0 from K-means Clustering

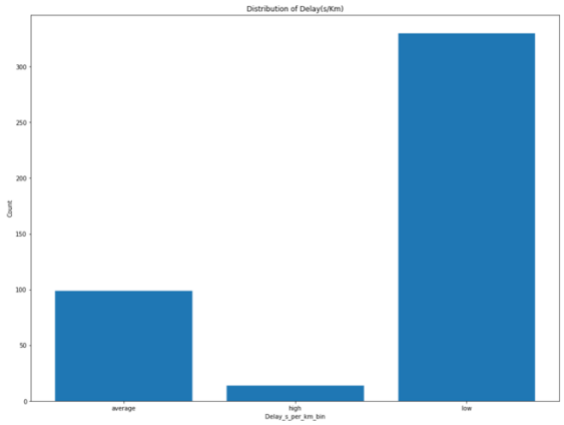

Figure 16 Cluster 1 from K-means Clustering

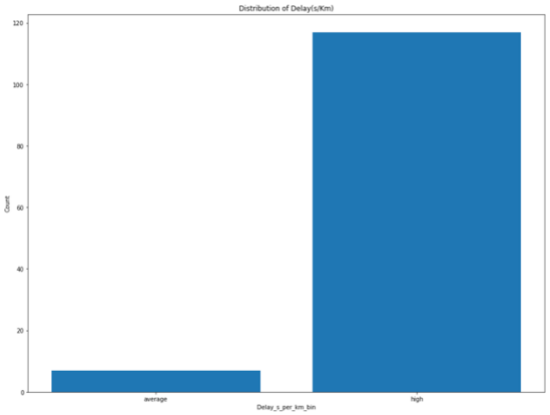

Figure 17 Cluster 2 from K-means Clustering

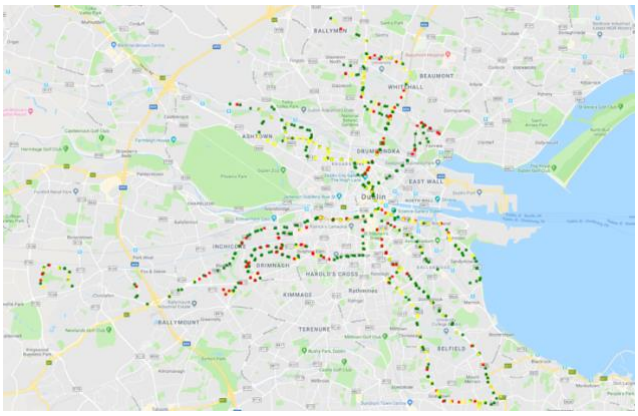

Figure 18 Results of K-Means Clustering shown on Map

Clusters are now obtained; next step is to map these clusters with 'realtimedata' to observe a clear pattern of the data. To perform cluster analysis, three new variables are created named hour, day and month. To see traffic pattern for buses across various times of the day, over the weekdays and weekends; and over the month to see if these patterns exhibit any seasonality.

In figure 19, hourly delay is plotted. As it can be noticed that the peak delays occurred in time period of 10:00 to 11:00. Other than this time period, 7:00 to 8:00 as well as at 15:00 and 20:00, some peaks can be noticed. These hours are mostly school hours and office hours which experience most of the boarding and alighting.

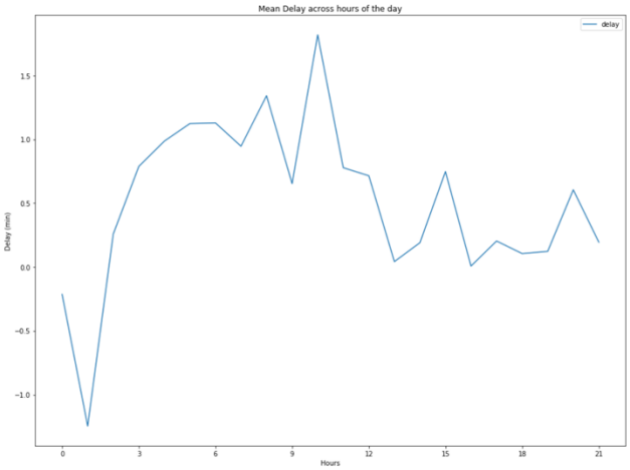

Figure 19 Mean Delays Across Hours of the Day

Similarly, weekly analysis is also performed, however, it does not exhibit much as data is limited to few days only. As per plot, Sunday is experiencing the highest delay and least delay is experienced on Tuesday. To make weekly analysis more informative, much more data records are needed. Same analysis can be repeated for months, however, there would be requirement of data for months and years.

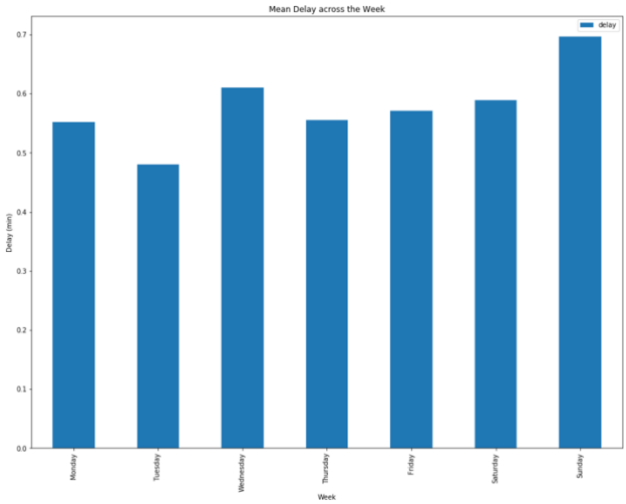

Figure 20 Mean Delay Across the Week

V. CONCLUSION

This paper has presented an analysis of bus transit system of Dublin, Ireland. To do so, clustering has been performed on the given data to achieve some patterns at stop-level so that required action can be taken to improve the transit system wherever possible. Three clustering methods have been applied on the data to achieve clusters on the basis of delay. Among all, K-means performed best as its silhouette score was the highest. The clusters achieved by k-means clustering are represented on map with three different flags. These flags represent stops with different delays. Stops with red flags are experiencing highest delays, hence these stops need some action to cope with these delays by planning routes systematically or by providing new bus routes for specific hours only which are facing peak delays. On the other hand, green flagged stops are operating in good manner, there is no need to take any action for these stops. Yellow flagged regions are dealing with average delays, these stops do not need immediate action, however, should be consider in coming future to make Dublin Bus System operate in a perfect manner. By taking these results into consideration, new plans can be made for bus routes, so that public do not get disappointed with severe delays and they would prefer public transport most of the times for their journeys.

There is scope to perform more deep analysis on this data as data considered in this research is limited to few weeks. More data will give more idea about weekly delays and monthly delays which would helpful in long-term planning of routes. In given data, there is no information about the number of boarding at each stop, if there would be knowledge about boarding from stops, it will also help in providing new public facilities around those stops from where more boarding reported. This is the shortcoming of this data as it is not recording the taps of leap cards. With the help of boarding data, it is also possible to eliminate some stops from routes where there is no boarding at any time.

Other than this implementation, classification can also be done on this data. As this data has no labels, unsupervised machine learning can be applied on this data. With

classification, prediction for delays can be calculated and this would also help in making decisions about new plans and routes regarding bus routes and city planning.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude towards my supervisor Mark Roantree for providing support throughout the project. His guidance and suggestions helped me to complete my project successfully.

REFERENCES

[1] M. Steinbach, G. Karypis and V. Kumar, "A Comparison of Document Clustering Techniques".

[2] M. Steven I-Jy Chien, Y. Ding and a. C. Wei, "Dynamic Bus Arrival Time Prediction with Artificial Neural Networks," *JOURNAL OF TRANSPORTATION ENGINEERING*, vol. 128, no. 5, September 2002.

[3] J. Patnaik, S. Chien and A. Bladikas, "Estimation of Bus Arrival Times Using APC Data," *Journal of Public Transportation*, vol. 7, no. 1, 2004.

[4] D. Mullner, "Modern hierarchical, agglomerative clustering algorithms," 12 September 2011.

[5] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah and T. Herawan, "Big Data Clustering: A Review," *Springer*, vol. 8583, 2014.

[6] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou and A. Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," *IEEE*, vol. 2, no. 3, pp. 267-279, September 2014.

[7] S. Arora and I. Chana, "A Survey of Clustering Techniques for Big Data Analysis," *IEEE*, pp. 59-65, 10 November 2014.

[8] M. Riveiro, M. Lebram and M. Elmer, "Anomaly Detection for Road Traffic: A Visual Analytics Framework," *Institute of Electrical and Electronics Engineers*, vol. 18, no. 8, pp. 2260-2270, August 2017.

[9] G. Seif, "The 5 Clustering Algorithms Data Scientists Need to Know," 5 February 2018.

[Online]Available:https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.

[10] J. Han and M. Kamber, Data Mining: Concepts and Techniques, vol. second, Morgan Kaufmann Publishers, 2006, p. 48.

[11] P. Berkhin, "A Survey of Clustering Data Mining Techniques," in *Grouping Multidimensional Data*, Springer, 2006, pp. 25-71.

[12] D. S. M. J. C. H. Laura L. Tupper, "Mixed Data and Classification of Transit Stops," 16 November 2016.