From Anomalies to Patterns and Predictions : A Comprehensive Data Analytics Approach on Transport for London using Unsupervised and Supervised Learning

Umar Farooq Khan (22179780)
Data Analytics for Artificial Intelligence – H9DAI
MSCAI_SEP23
School of Computing
National College of Ireland

1 Background Research

Our research focuses on studying how people in London used buses, the metro, and trains from 2018 to 2020. We want to understand how changes in the city, like more people and different jobs, affected transportation. Our goal is to see if improvements are needed in these services to match the city's changes and make travel better for everyone.

We have a dataset that tells us how people used buses, the metro, and trains in London each week. This helps us see if there are quick changes week by week or bigger trends over time. By looking at this data, we can understand how people move around in London. This helps us figure out if anything needs to be improved to make buses, the metro, and trains work better for the people who use them. We're using a dataset from data.gov.ie that has information about buses, the metro, and trains in London annualized. Our research helps us see patterns and changes over time throughout the year.

Our aim is to ensure that buses, the metro, and trains in London provide the best service possible. If there are any issues, we want to suggest ways to make transportation smoother and more convenient for everyone. Other countries can take help from this research as well to improve their transport infrastructure because London is proved to be the best in Transportation.

2 Data Analytics

This research focuses on analyzing Transport for London's ridership data from 2018 to 2020, encompassing modes like the Underground, Overground, DLR, and TfL Rail. The data, recorded at quarter-hourly intervals but without daily records, provides a detailed view of passenger flow in London's transport system.

The initial dataset for 2018, consisting of 2,584 rows and 111 columns, revealed peak commuting hours through a histogram of reporting times, showing a bimodal distribution.

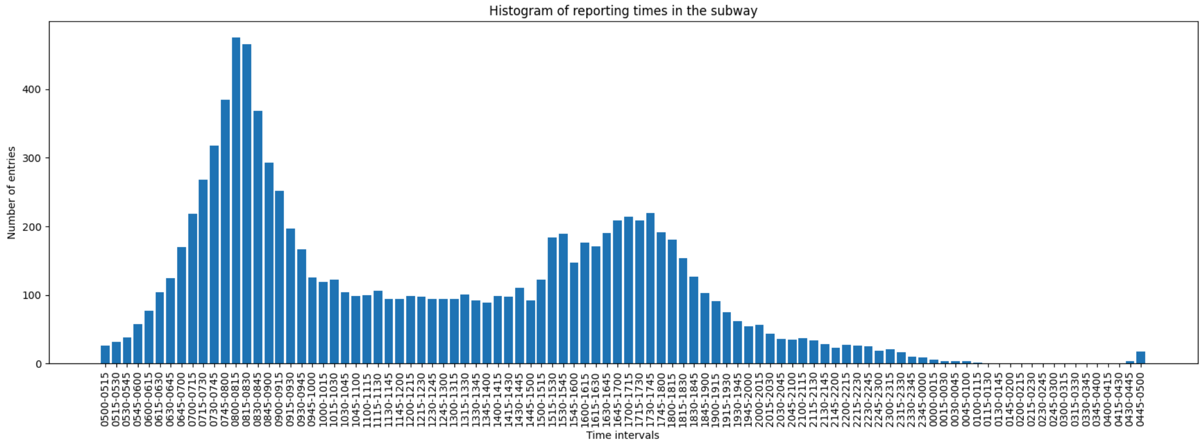


Figure histogram of reporting times in subway

The study used supplementary data to create maps showing ridership distribution in London. It merged datasets like 'weekends_entries.csv' for weekend entries, 'londontemp.csv' for temperature, and station coordinates, along with UK public holiday data. The dataset was transformed from wide to long format with 344,064 rows using pandas melting function for better analysis. It was also divided by travel direction (entries and exits) for focused analysis.

The correlation matrix heatmap shows more correlation between PM Peak, Evening and Late.

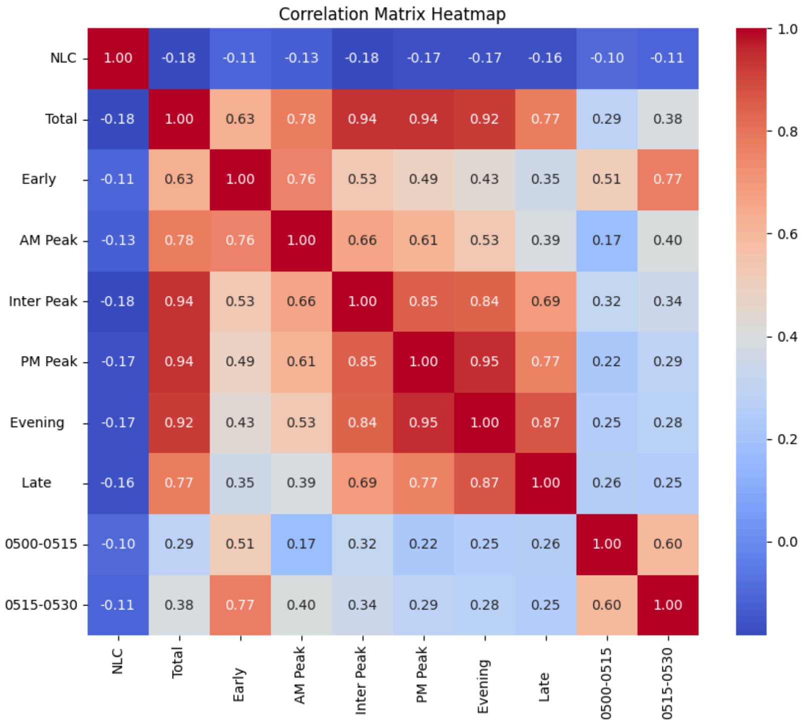


Figure correlation matrix heatmap

2.1 Geospatial Visualization and Hotspot Identification

For geospatial visualization techniques, we analyze data, with a focus on identifying hotspots and anomalies in urban mobility patterns in the London Transport Infrastructure.

A map of London was created displaying tube stations as markers, scaled according to ridership discrepancies.

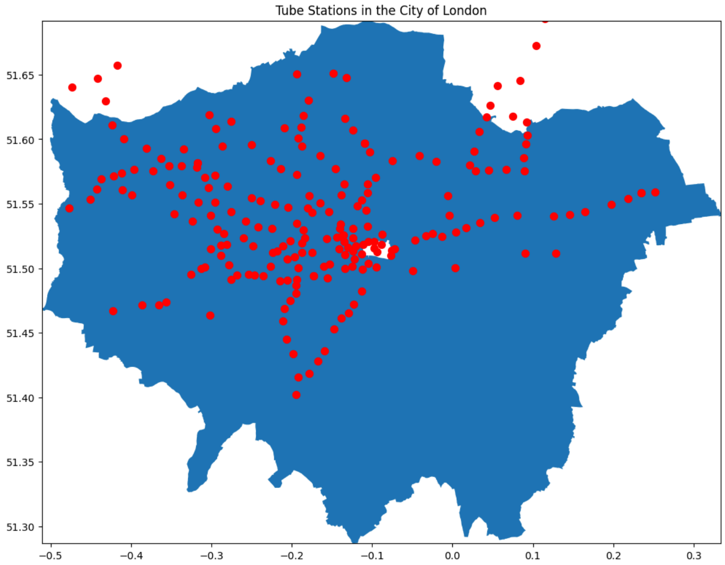


Figure tube stations in the city of london

The top 30 busiest stations were identified, providing a clear visualization including interactive heatmap to show hotspots. This approach highlighted the most frequented stations, offering insights into the city's busiest areas.

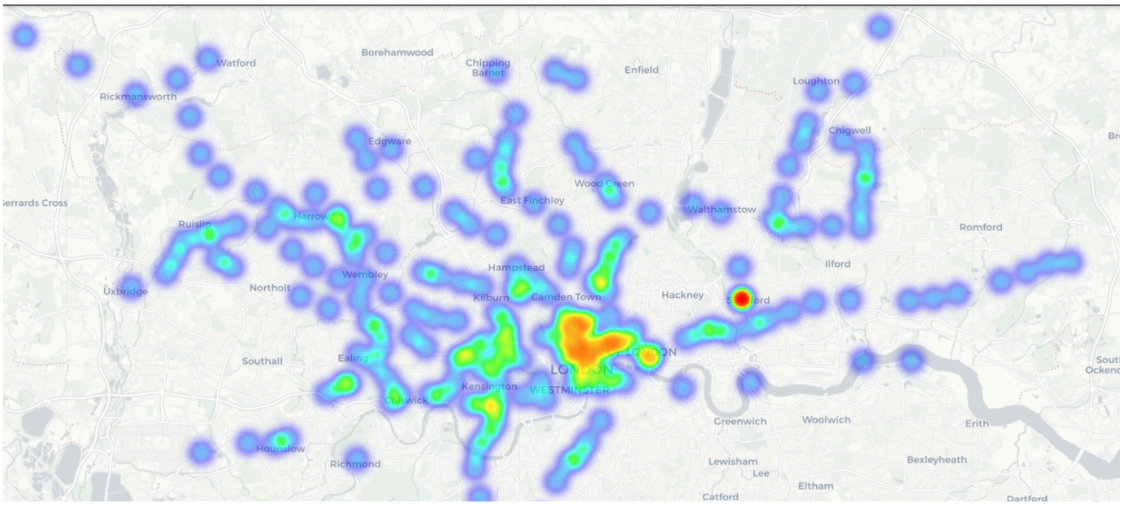


Figure Zoomable heatmap of frequently used stations

Another simple geographical map showing London's busiest stations.

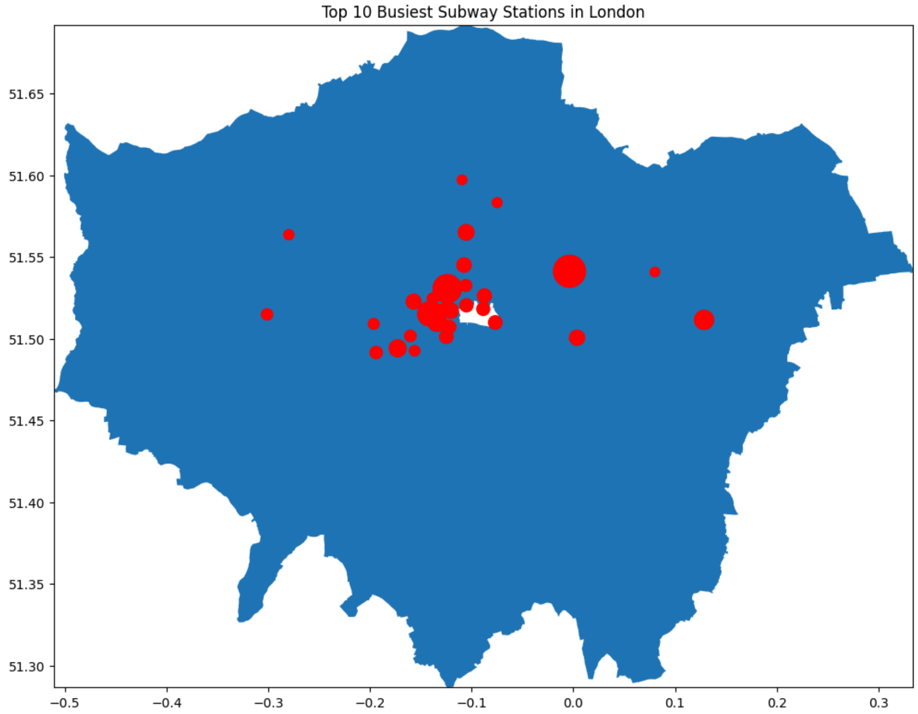


Figure top 10 busiest subway stations in london

2.2 Anomaly Detection in Early Morning Ridership

We focused on detecting anomalies in early morning ridership from Monday to Thursday. An Isolation Forest model was used, tailored with specific parameters to isolate atypical patterns during these time periods. After identifying anomalies, the data was visualized geospatially, merging station location information with anomaly data. This facilitated the mapping of the top ten stations with the most significant early morning anomalies, illustrating potential geographical factors affecting ridership, such as proximity to educational institutions.

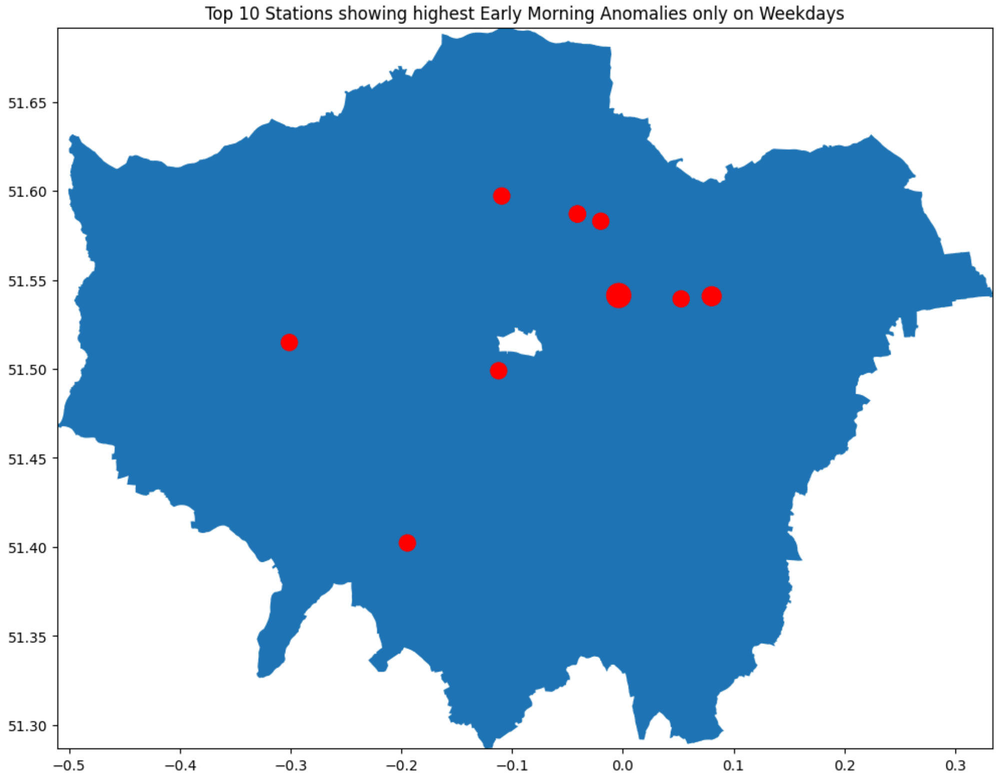

Top 10 Stations showing highest Early Morning Anomalies only on Weekdays
51.65
51.60
51.55
51.50
51.45
51.40
51.35
51.30
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3

Figure top 10 stations highest anomalies in morning at weekdays

A tree map showing stations exhibiting most anomalies in which Stratford exhibiting most anomaly

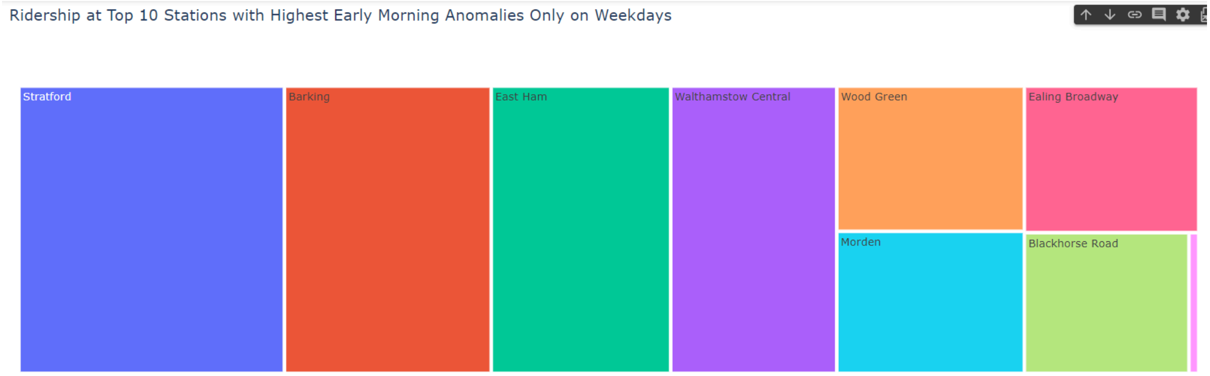


Figure ridership tree map of 10 stations highest anomalies in morning at weekdays

2.3 Ridership and Temperature Trends Visualization

Focusing on only one station 'Acton Town', the period before the COVID-19 pandemic (April 2018 - December 2019). The ridership looks like this on line graph:

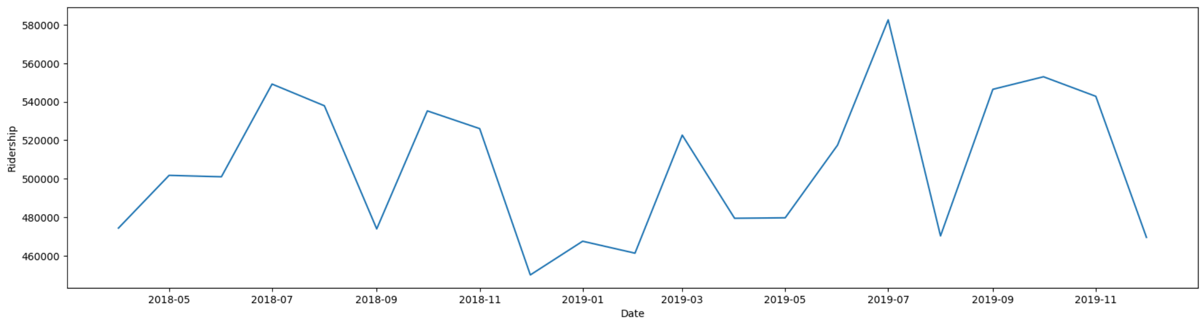


Figure of ridership line graph

The dataset was enhanced by merging it with temperature data, Line graphs were used to visualize ridership trends alongside temperature changes, shedding light on the relationship between these two variables.

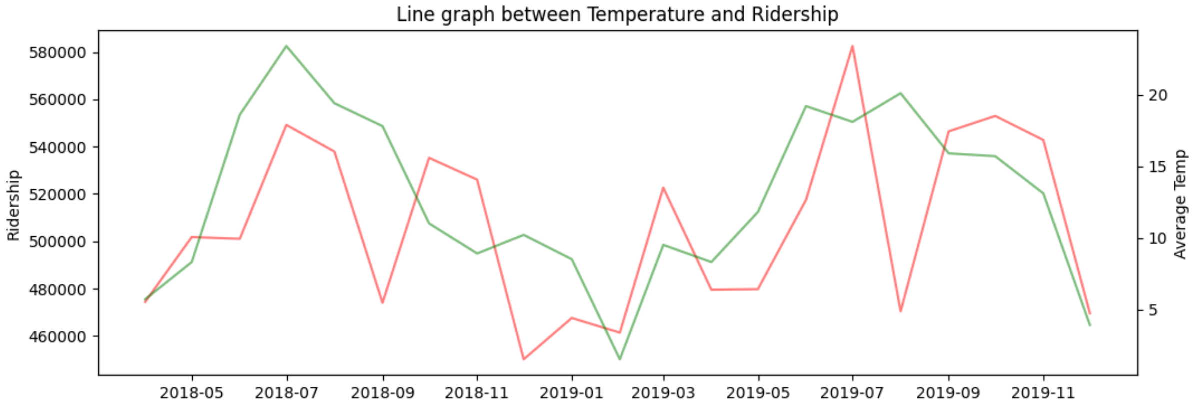


Figure of temperature and ridership line graph

The data preparation process was initiated to facilitate the computation of ridership figures across weekends and weekdays, encompassing all modes of transportation. It was observed that the ridership on weekends, encompassing only three days, exceeded the cumulative ridership during the entire span of weekdays.

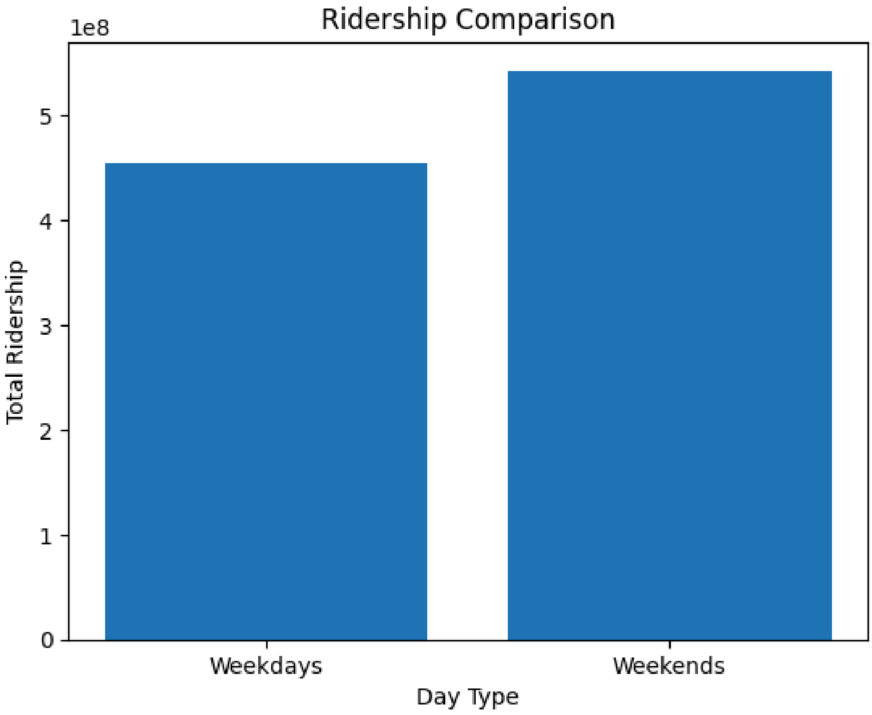


Figure of ridership comparison

A comprehensive histogram encompassing entries and exits during both weekdays and weekends was presented. The analysis revealed a substantial disparity, where weekend entries significantly surpassed not only the exits during both weekdays and weekends but also the entries recorded on weekdays.

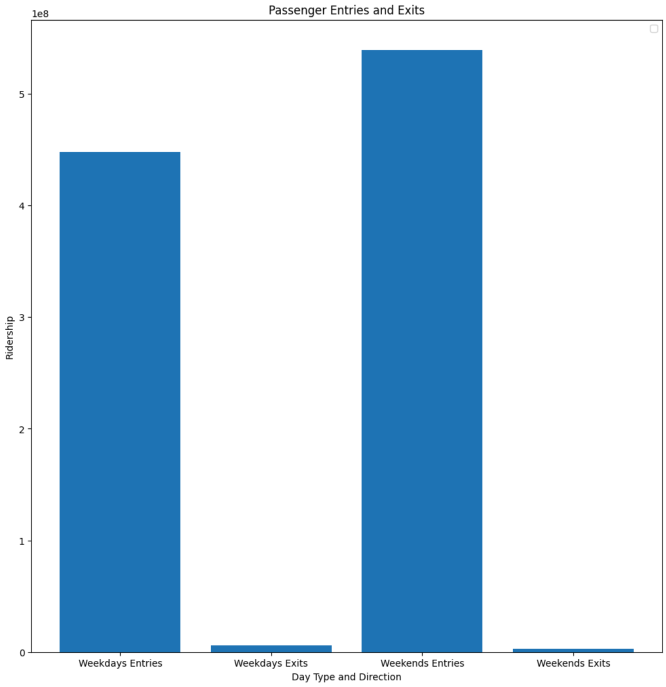


Figure of passengers entries and exits

We showed the relationship between entries and exits was conducted. The findings underscored a notable contrast, with exits representing a mere fraction of the total entries. This discrepancy is attributed to the fact that entries at a given station are considerably higher, as passengers disembark at various disparate stops along their journeys.

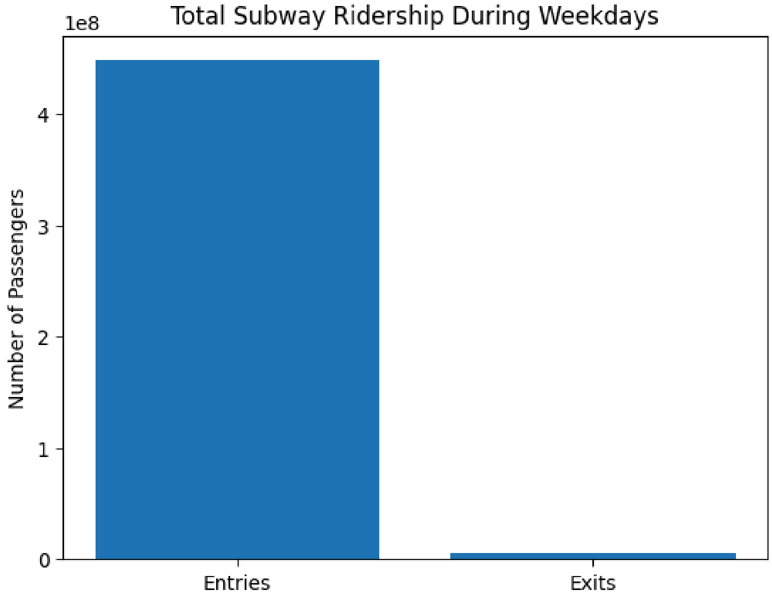


Figure of ridership entries and exits during weekdays

Another comparison between weekdays and weekends with weekly ridership.

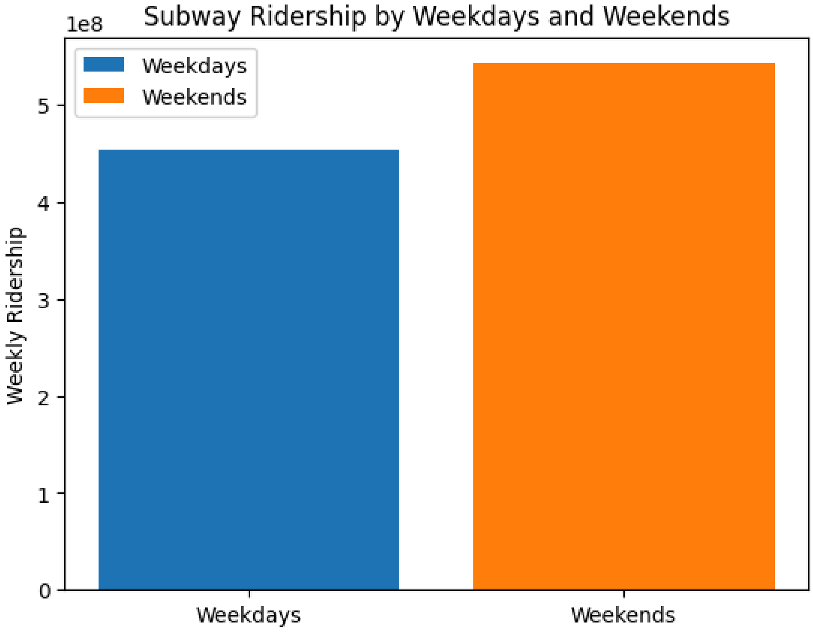


Figure of ridership by weekdays and weekends

We made a histogram between AM Peaks, Inter Peak and PM peaks depicting that the AM peak is very less as compared to their counterparts.

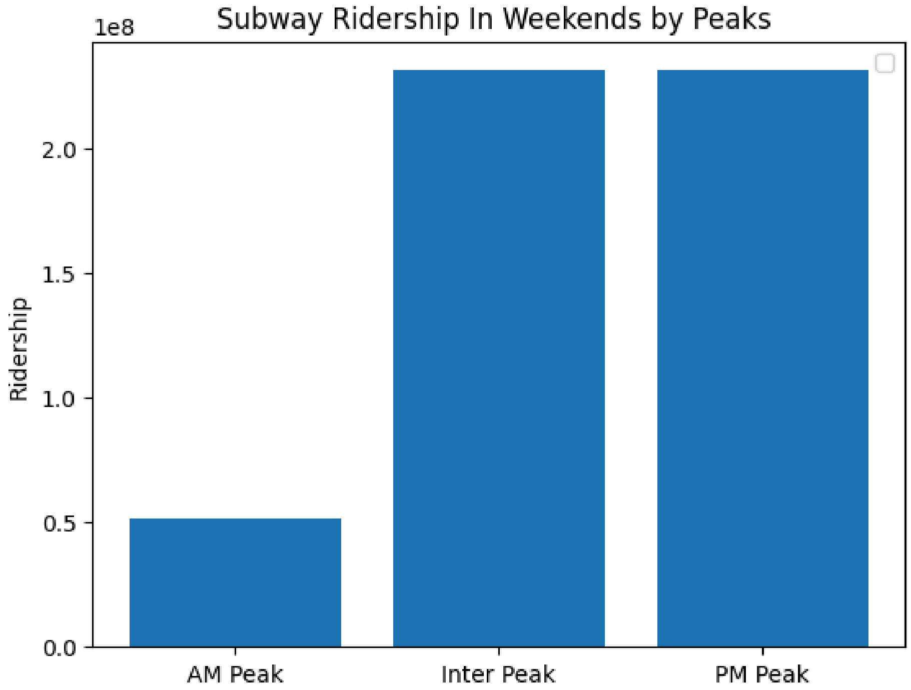


Figure of ridership in weekends by peaks

The box plot shows total entries on weekends versus weekdays, showing that weekends have a greater range and variability in entries, with more outliers indicating extreme values, whereas weekdays exhibit a tighter distribution with fewer extreme values.

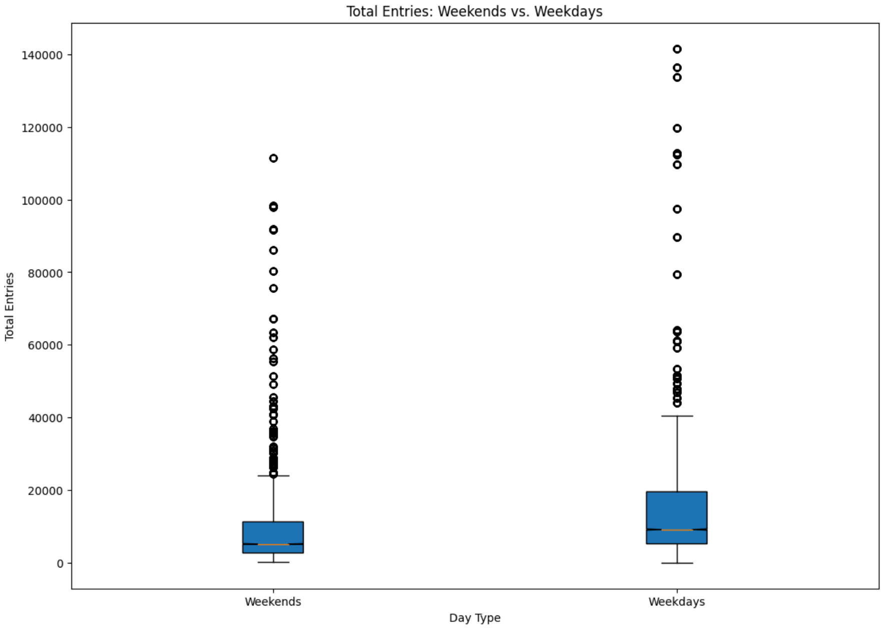


Figure of weekdays vs weekends

3 Machine Learning Algorithms

As our dataset does not show a time series, has no dates, it is annualized and is grouped into weekdays and weekends meaning weekends of the whole year, so we applied these models specifically which could be best for our dataset which helps in finding patterns ,clusters and predictions.

3.1 Linear Regression (Impact of UK Holidays on Ridership including Temperature)

UK holidays and London Temperature data was processed, aggregated by month, and then merged with the ridership dataset to examine the impact of holidays on ridership patterns. This integration provided a comprehensive view, enabling an analysis of seasonal variations and the specific impact of holidays on ridership.

A predictive model using Linear Regression was developed to assess the relationship between ridership and the combined features, including weather conditions, holiday occurrences, and other relevant variables. The model's predictions were visualized alongside actual ridership data, offering a comparative perspective of the model's accuracy and practical applicability.

We employed line graphs, to juxtapose actual ridership against predicted values. This visualization highlighted the model's performance over time, offering insights into its predictive capabilities. Additionally, an analysis of the model's coefficients provided an understanding of the influence exerted by each feature on ridership predictions, shedding light on the factors most impactful in determining ridership levels.

3.2 SVM:

Then we developed a Multiclass Support Vector Machine (SVM) model for the prediction of transportation modes based on various features. The primary goal of this research is to create a predictive model that accurately classifies transportation modes, and this report discusses the methodology employed to achieve this objective.

The study presumably involves using SVM algorithms to classify data into distinct categories, based on patterns learned from the training dataset.

The dataset used in this research contains information about transportation activities, including features related to 'day,' 'Station,' 'dir,' and time intervals. The 'Mode' variable represents the target variable that we aim to predict.

Categorical variables such as 'day,' 'Station,' and 'dir' were transformed using one-hot encoding.

Then we choose a Linear kernel to transform the data into a higher-dimensional space where it can be more easily separated. Then we try to tune the model parameters, such as the regularization parameter and the kernel coefficient, to optimize the performance of the SVM.

3.3 Clustering (Hierarchy Dendrogram clustering, K Mean, DBSCAN, Elbow Method)

In this part of analysis, a clustering approach was applied to a dataset of station entries to uncover patterns and groupings inherent within the data. The motivation for employing different clustering approaches across various techniques stems from the absence of a precise and universally accepted definition for the concept of a "cluster."[1]. The process began with feature selection, focusing on key variables indicative of station usage patterns across various times of day.

The dataset was standardized using a Standard Scaler to ensure the clustering algorithms. An initial exploration for the optimal number of clusters (k) for KMeans clustering was performed using the elbow method to know the best number of clusters.

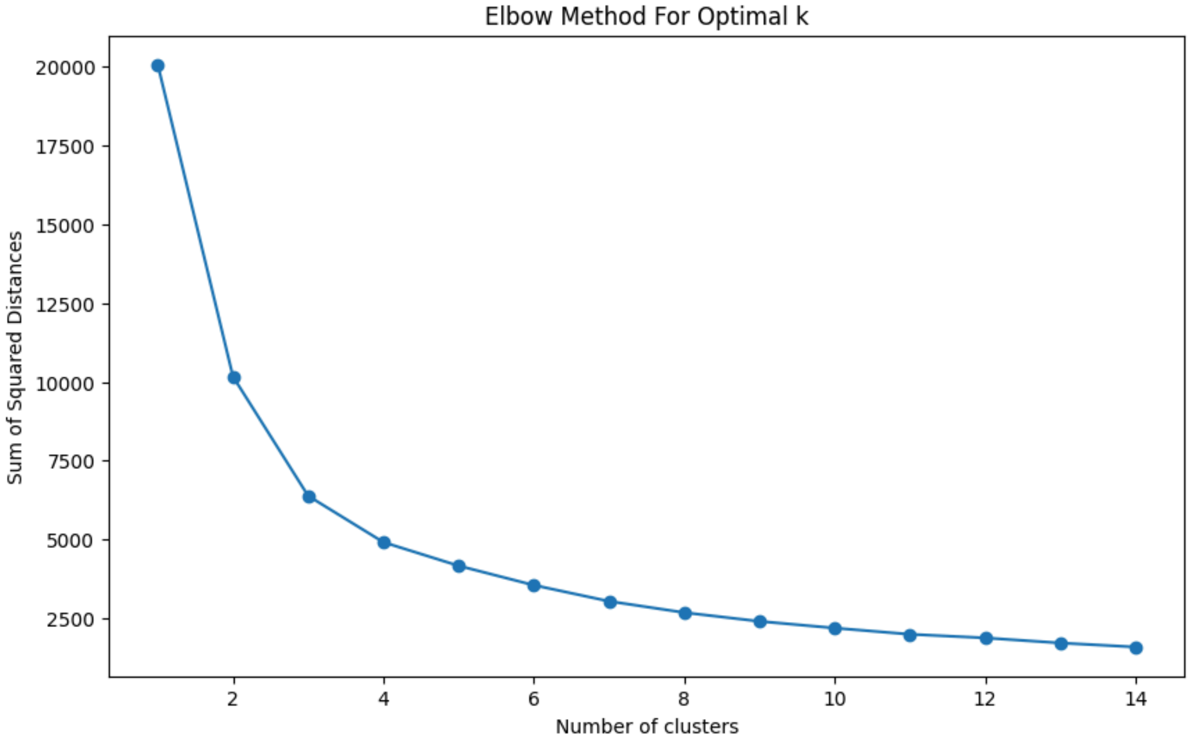


Figure of elbow methods for optimal K

Subsequently, KMeans clustering was executed with a chosen number of clusters, inferred from the elbow plot to be three, to segment the stations based on total entries and AM peak entries. This provided a clear visual division of stations into three distinct groups.

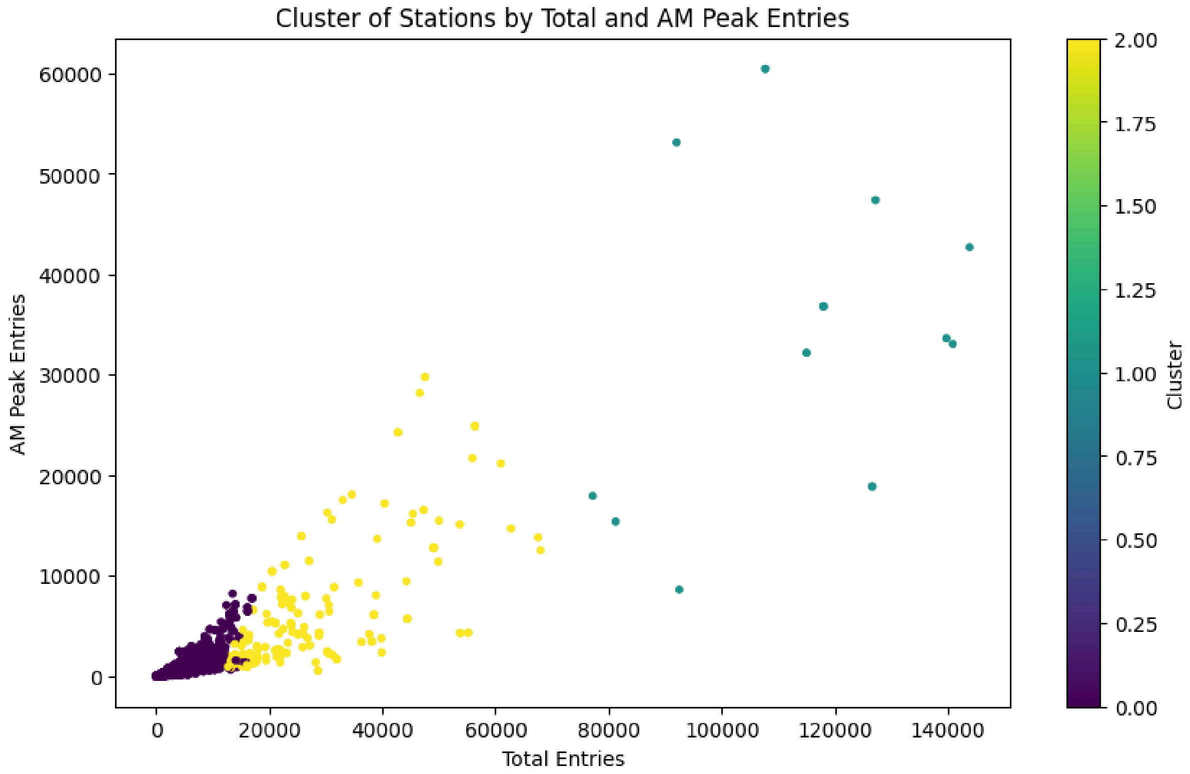


Figure of Clusters on the basis of AM peak entries by station

Further segmentation was conducted, this time considering early and late entries with an increased number of clusters (seven), to capture more nuanced patterns of station usage during the beginning and end of the day.

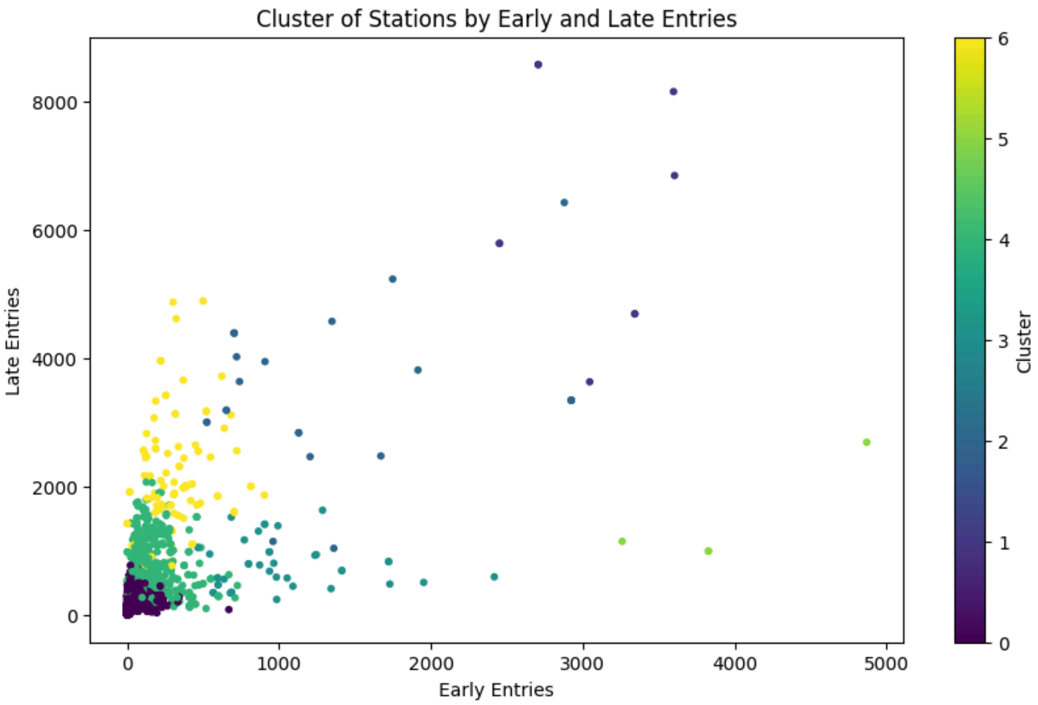


Figure of Clusters based onLate peak entries by stations

3.4 DBSCAN:

A different clustering algorithm, DBSCAN, was employed to account for the density-based spatial distribution of the data. DBSCAN is particularly useful for identifying outliers and for data that do not necessarily form globular clusters. The DBSCAN results were visualized, illustrating clusters along with noise points (denoted by -1), which do not fit into any cluster.

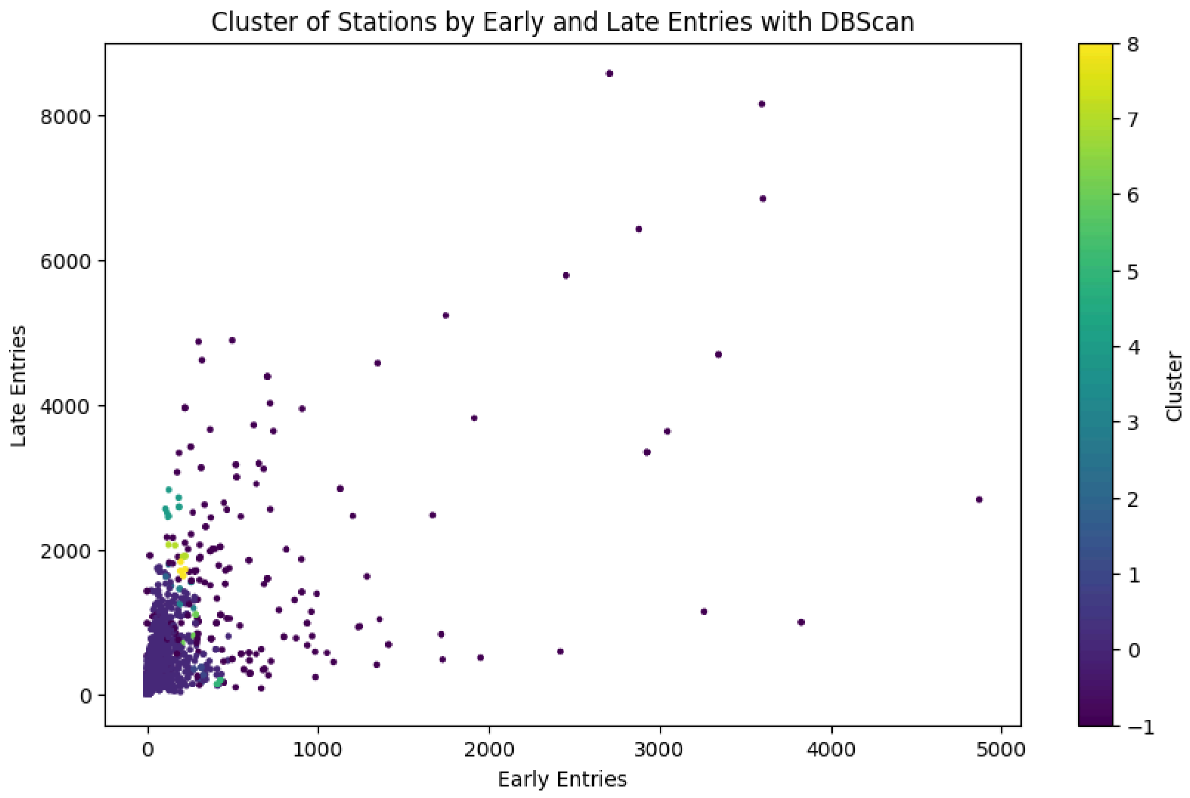


Figure of cluster for early and late entries by DBScan

For a geospatial perspective, an agglomerative hierarchical clustering was applied to a subset of the top stations based on the Haversine distance, which calculates distances between points on a sphere, such as Earth.

Within hierarchical clustering techniques, clusters are generated through a step-by-step process that involves either top-down or bottom-up approaches. Hierarchical clustering encompasses two primary forms: agglomerative clustering and divisive clustering.[2]

A dendrogram was generated to aid in deciding the number of clusters, and the final clustering was visualized on a map using Folium, providing a geographical representation of station groupings.

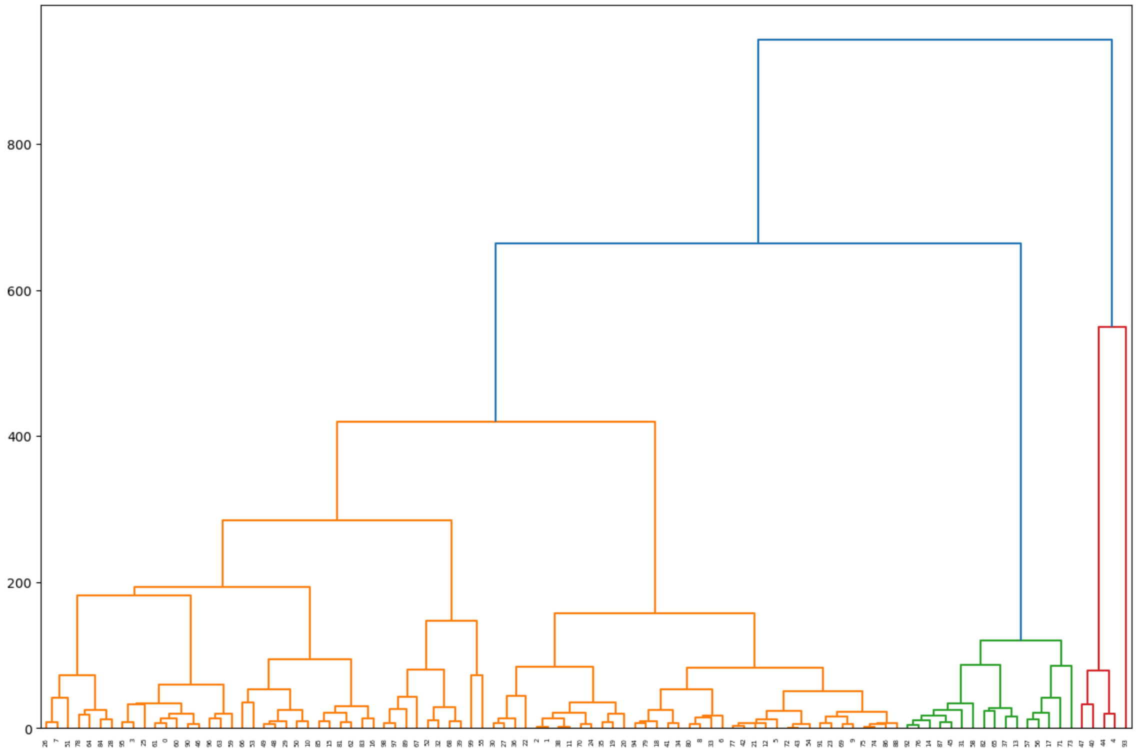


Figure of geographical representation of station groupings

We showed these clusters based on the distance on London Map using Folium, which seems very logical to cluster nearby stations into one cluster:

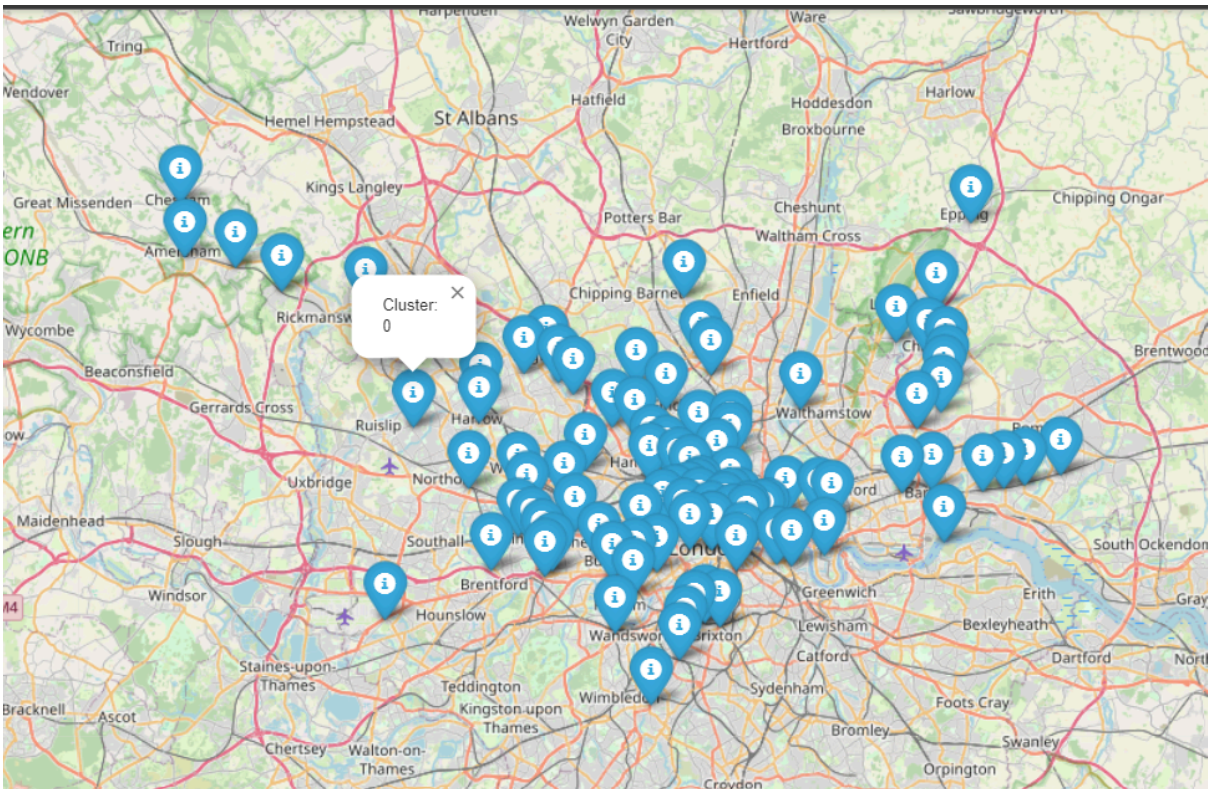


Figure of stations clustering in london map

4 Evaluation and Discussion

4.1 Linear Regression

The regression analysis was employed to forecast ridership, and the resulting line graph illustrates the correlation between observed and predicted ridership.

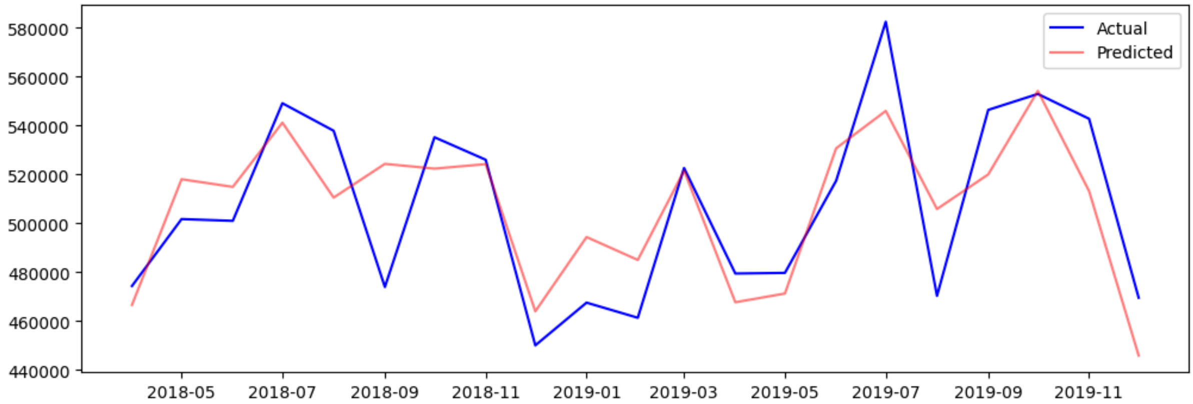
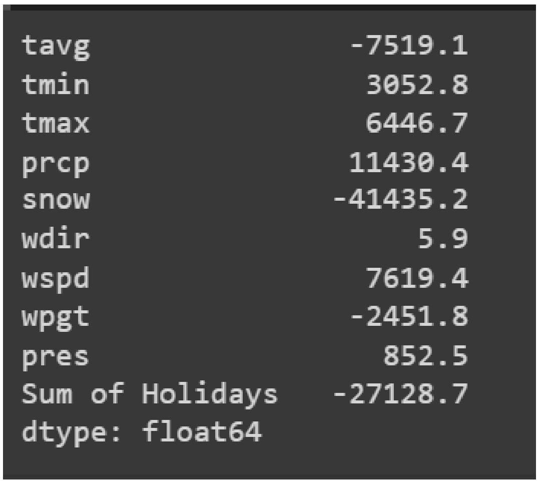


Figure of actual and predicted ridership levels

Mean Squared Error	505787675.77
Mean Absolute Error	18576.85
R squared R2 Score	0.62

Then we calculate coefficients of the features in the model:



```
tavg                -7519.1
tmin                 3052.8
tmax                 6446.7
prcp                11430.4
snow               -41435.2
wdir                    5.9
wspd                 7619.4
wpgt                -2451.8
pres                  852.5
Sum of Holidays    -27128.7
dtype: float64
```

Figure of coefficients of model

By looking at the coefficients of our model, we can gain some specific insight into the contribution of each feature. For example, each degree increase in average temperature seems to increase in 9035 ridership which is true. with each inch of snow, there is significantly decrease in subway rides i.e 342704 which again seems logically correct.

4.2 SVM:

This model tries to predict mode of transport from various ridership features which helps them to predict ridership of specific mode of transport on some specific time.. The confusion matrix for SVM seems very prominent, showing an accuracy of 93%. All scores and F1 scores are above 90 showing a good machine learning model performance.

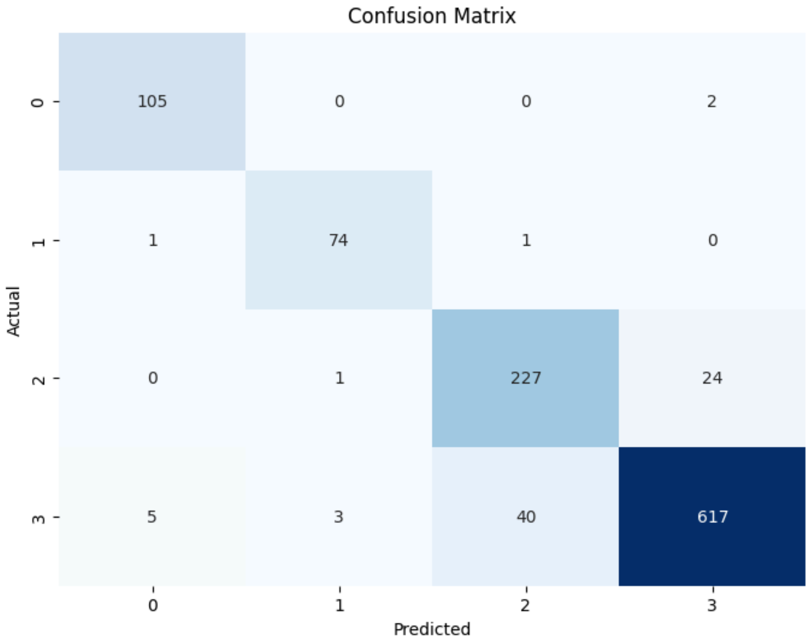
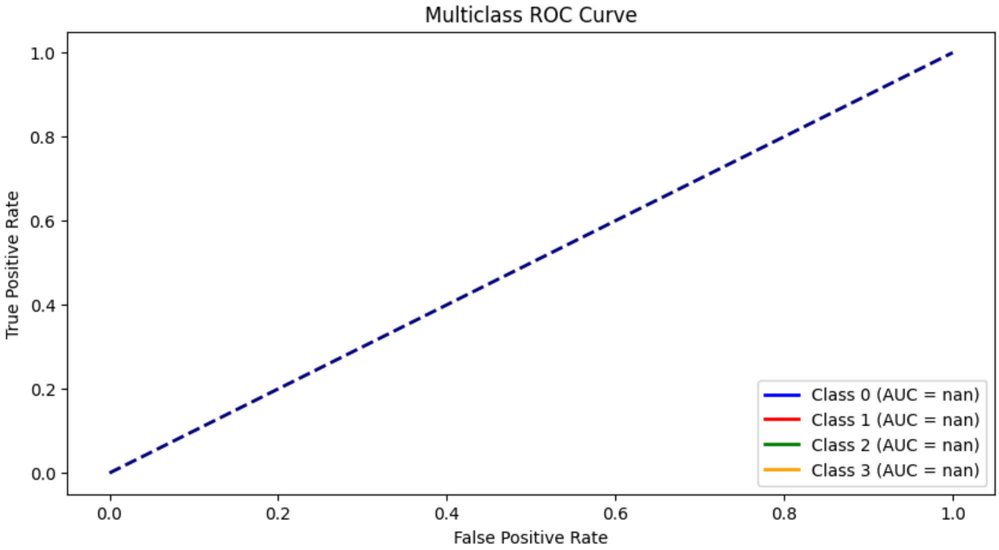


Figure of SVM confusion matrix

Metric	**Value**
Accuracy	0.93
Precision (Weighted)	0.9317
Recall (Weighted)	0.93
F1-Score (Weighted)	0.9305
Specificity (True Negative Rate)	0.9459
Micro Sensitivity (True Positive Rate)	0.93
Macro Sensitivity (True Positive Rate)	0.9459
Weighted Sensitivity (True Positive Rate)	0.93
Precision (Positive Predictive Value)	0.9253
Micro Precision	0.93
Micro Recall	0.93
Micro F1-Score	0.93
Macro Precision	0.9253
Macro Recall	0.9459
Macro F1-Score	0.9352



4.3 Clustering:

The clustering analysis conducted on the dataset of station entries has provided valuable insights into the patterns and groupings inherent in London's transport system. The objective was to uncover distinct station usage patterns across various times of the day and understand the spatial distribution of stations.

4.3.1 K-Means Clustering:

The K-Means clustering algorithm was applied to segment the stations into groups based on total entries and entries during the AM peak hours. Initially, an exploration to determine the optimal number of clusters (k) was conducted using the elbow method. The elbow plot suggested that three clusters would be appropriate, resulting in a clear division of stations into three distinct groups. This segmentation helped identify stations with similar usage patterns and provided a broad overview of station categorization.

4.3.2 DBSCAN Clustering:

DBSCAN, one of the best algorithms for a density-based clustering algorithm.[3] was employed to account for the spatial distribution of station entries. DBSCAN is effective in identifying outliers and clusters in data that may not form clear spherical clusters. The results of DBSCAN clustering were visualized, showing distinct clusters along with noise points, which represent stations that do not fit into any cluster. This approach allowed for the identification of spatially dense station groups and highlighted stations that deviated significantly from these clusters.

4.3.3 Hierarchical Clustering:

Agglomerative hierarchical clustering was applied to a subset of the top stations, focusing on the Haversine distance to calculate distances between stations. The need for geometrical presentation of the spherical earth becomes very relevant when we take into consideration an ever increasing junctions inside a city.[4]

A dendrogram was generated to assist in determining the appropriate number of clusters. The final clustering results were visualized on a map using Folium, providing a geographical representation of station groupings. This approach allowed for a geospatial perspective, showcasing how stations clustered together based on their proximity.

The clustering analysis has successfully grouped stations based on their usage patterns and spatial distribution. The results provide a comprehensive understanding of how stations are categorized in terms of passenger entries and geographical proximity. These insights can be valuable for transportation planning, station management, and resource allocation within London's transport system, ultimately enhancing the efficiency and effectiveness of public transportation in the city.

5 Conclusion

Our study used data analytics and Machine Learning to understand London's public transport usage and patterns and try to forecast the ridership. By examining holiday dates and weather, it found that warmer temperatures increase ridership, while snow decreases it with the help of Linear Regression.

A machine learning model (SVM) was created to predict transport modes. The SVM relied on a linear kernel for data transformation, and its effectiveness was fine-tuned using parameters like the regularization constant, which balances the trade-off between achieving a low error on the training data and minimizing model complexity.

Clustering methods like KMeans and DBSCAN grouped stations based on entry times and location, showing distinct travel patterns. Clustering techniques such as KMeans, guided by the elbow method for optimal cluster numbers, and DBSCAN, showing outliers, helped identify patterns in station usage.

DBSCAN's performance was optimized by tuning the 'eps' and 'min_samples' parameters, crucial for determining the density threshold required to form clusters, thus identifying core points, border points, and outliers effectively. The research highlights how combining environmental data with AI can reveal important trends in city transport and help in planning.

Conducting all of this analysis and Prediction helps the Government to ease the public more easily and help them in taking decisions. Other countries can also take help from this analysis if they are planning to improve their transportation system like Ireland and Pakistan, India.

References

[1] L. Rokach, "Clustering methods," in Data Mining and Knowledge Discovery Handbook, Springer, 2005, pp. 331–352.

[2] Chen, Y., Ruys, W., and Biros, G. (2020). "KNN-DBSCAN: A DBSCAN in High Dimensions." In Proceedings of the arXiv preprint arXiv:2009.04552.

[3] A. Saxena, N. R. Pal, and M. Vora, "Evolutionary methods for unsupervised feature selection using Sammon's stress function," Fuzzy Inf. Eng., vol. 2, no. 3, pp. 229–247, 2010.

[4] N. R. Chopde and M. Nichat, "Landmark-based shortest path detection by using A* and Haversine formula," International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, no. 2, pp. 298-302, 2013.

[5] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms which use cluster centers," Comput. J., vol. 26, no. 4, pp. 354–359, 1984.