National College of Ireland

Project Submission Sheet

Student Name: ……Mustafa Karaburun……

Student ID: ……x23216158……

Programme: ……MSC in Data Analytics…… **Year:** ……2024……

Module: …Scalable Systems Programming……

Lecturer: ……Noel Cosgrave……

Submission Due Date: …20/08/2024……

Project Title: …… **Clustering Analysis of Uber NYC For-Hire Vehicle Trip Data (2021)……..**

Word Count: ……2979……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: ……Mustafa Karaburun……

Date: ……20/082024……

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

SCALABLE SYSTEMS PROGRAMMING

Clustering Analysis of Uber NYC For-Hire Vehicle Trip Data (2021)

Your Name/Student Number	Course	Date
Mustafa Karaburun/x23216158	**MSC in Data Analytics**	**20/08/2024**

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI ACKNOWLEDGMENT

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
NA	NA	NA

DESCRIPTION OF AI USAGE

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

NA	
NA	
NA	NA

EVIDENCE OF AI USAGE

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

ADDITIONAL EVIDENCE:

NA

ADDITIONAL EVIDENCE:

NA

Clustering Analysis of Uber NYC For-Hire Vehicle Trip Data (2021)

MUSTAFA KARABURUN
X23216158
School of Computing
National College of Ireland
Dublin, Ireland

***Abstract*— This paper is an effort to derive meaningful insights with regard to the clustering analysis of the Uber NYC for-hire vehicle trip dataset in 2021, focusing on the integration of weather data. The present work uses two algorithms: K-Means and hierarchical clustering, together with a distributed processing environment implemented by Apache Spark. The study has the following main objectives: to find travel behavior patterns, explain how weather conditions might affect the characteristics of journeys, and benchmark the performance of different clustering algorithms. This will involve data acquisition, cleaning, transformation, and clustering analysis on a more than 4 GB dataset to prove the scalability and efficiency of the approach. Several of the important findings include the very high correlation of trip characteristics with weather, variability of different trip patterns across the five boroughs, and performance comparison of clustering algorithms. All these insights offer very useful information for urban transportation planning and optimization.**

Keywords—Clustering Analysis, Apache Spark, Uber NYC, K-Means, Hierarchical Clustering

I. INTRODUCTION

A great volume of data is generated in the transport sector daily, more so in metropolitan areas like New York City. It is very important to analyze data to understand travel patterns, optimize routes, and improve service efficiency. Urban transportation with data-driven insights can lead to better resource allocation that reduces congestion and results in enhanced passenger experiences. In the presence of big data tools such as Apache Spark, it has now become feasible to process and analyze large datasets in order to increase efficiency, thereby empowering stakeholders to make informed decisions based on empirical evidence.

.

II. OBJECTIVES

A. Objectives

The main goal of this study is the clustering analysis of trip data for-hire vehicles in the New York City Uber database for the entirety of 2021. The integration with weather data allows for patterns of trip behavior that are triggered by external variables to be discerned. It also targets the comparison of the performance between different clustering algorithms, mainly K-means and hierarchical clustering, within a distributed processing environment. The results will provide better insight into how environmental factors affect urban mobility and carry out benchmarking for assessing the efficiency of the clustering technique in large-scale data analysis.

B. Research Questions

1. How Do Weather Conditions Impact New York City Trip Characteristics for Uber For-Hire Vehicles?
2. What are the dominant trip behavior patterns across various boroughs, and how do they differ?
3. How nice are the K-Means and hierarchical clustering algorithms at detecting meaningful clusters within this dataset, and which of these provides more accurate and actionable insights?

III. RELATED WORK

Clustering algorithms have grown relevant to big data analytics through their implementation on large datasets. In this regard, a number of methodologies have been explored through various researches in handling the challenges of computational complexity and scalability associated with clustering in large-scale data environments. Alguliyev et al. [1] proposed a big-data-oriented batch algorithm for clustering, proving that it is efficient in the processing of large datasets. The approach is especially important when

considering Urban Mobility data that requires scalable solutions due to the volumes and velocities involved.

Another critical area of research is the incorporation of weather data into transportation analytics. Huang et al. [2] proposed an extreme-aware local-global attention mechanism for spatio-temporal urban mobility learning that well captured the influence brought by extreme weathers about transportation patterns. According to this paper, among other exogenous variables, weather has immense impacts on transportation patterns, thereby aligning with the goals of this project.

In distributed data processing, Apache Spark has recently come to the fore for big data analytics. Shangguan et al. [3] gave an example of the processing of big spatial data using Apache Spark, focusing on its efficiency in large volumes. The contribution from this research directly applies to the scope of the project in conducting clustering analysis on a large transportation dataset using the distributed computing features of Spark.

Chen et al. [4] further underline the relevance of data-driven approaches in urban transportation planning by applying such methods in planning plug-in hybrid electric taxi charging stations in Beijing. Their study demonstrated how these data-driven strategies could be used in optimizing urban infrastructure so critical to developing sustainable and efficient transportation systems.

Comparative studies on clustering techniques expose some strengths and limitations of various algorithms. Joshi et al. [8] carried out a performance analysis for the k-medoid algorithm, primarily for its scalability characteristics when dealing with large datasets containing only categorical data. Their work gives some insight into how to choose appropriate clustering techniques for large-scale data, which will become an important factor in this project's methodology.

Further, ranking and analysis of urban districts using taxi data, as done by Xie et al., go on to underscore the power of clustering algorithms in urban data analytics. Their work on district attraction ranking offers a model for understanding urban mobility patterns that can be expanded into other areas of transportation data analysis.

Finally, Leung et al. [7] put forward work on the scalable mining of Big Data that offers a broader context from which to understand the challenges and solutions in big data analytics. Their research stresses the importance of developing scalable algorithms dealing with complexity due to large datasets at the heart of this project.

IV. Methodology

A. Data Description

This analysis was done using data sourced from Kaggle, "Uber NYC For-Hire Vehicles Trip Data (2021)." Link: https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021 The data entail records of Uber for-hire vehicle trips in New York City over all of 2021. Each of the entries encompasses a variety of trip attributes, including request time, pickup and drop-off times, travel distances, fare details, and location identifiers all comprising a rich dataset for carrying out urban mobility analysis.

Besides the trip data, the dataset to be analyzed also contains the weather data of New York City. Weather can reveal a lot about environmental conditions while having the trip. The weather data contains temperature, humidity, and precipitation. The following variables are enlisted in the dataset for the explanation of the different trip patterns related to the weather. This dataset is in Parquet format to increase the efficiency in the storage of the dataset and the processing speed. The weather data will be saved in CSV format.

B. Data Processing

1. Data Acquisition

The trip data was sourced from the publicly available Kaggle dataset: "Uber NYC For-Hire Vehicles Trip Data (2021)." I have taken this dataset in Parquet format because it is a very efficient file format for large-scale data processing jobs. The weather data in relation to New York City was born out of a credible public source and hence was supplied in CSV format. The shapefile for the taxi zone was also sourced to map the trips against the specific geographic areas found in New York City. All datasets were uploaded to Google Colab for processing, using integration with

Google Drive that allows access to huge files without any hitch.

2. Data Cleaning and Transformation

In preparation of the dataset for analysis, there were some steps that were important in the cleaning process. First, all datetime fields in the trip data were standardized to be consistent across the dataset. Treatments of missing values for critical fields like trip distance and fare were either imputed with mean values or removed records with major data gaps. In the weather information dataset, datetime fields were transformed into a standard format as in the trip data, and unrequired columns have been dropped to be lean.

It was then merged with the weather dataset using the pickup date, so that each trip record would be matched to its relevant weather condition. This is treating a very important integration, as it analyzes how weather factors influence trip patterns. First, the taxi zone lookup data was used to append geographic information to the trip data so that it is then possible to conduct a spatial analysis by borough and zone. After merging, dropping redundant columns was done before creating a representative sample for clustering analysis.

3. Distributed Processing

Given the extent and complexity of this dataset, distributed processing was necessary. Apache Spark was used since it is a designed framework for supporting large data sets on multiple nodes. Here, configuration settings for the Spark session include increasing memory allocation and multiple cores to support efficient performance. In the Spark environment, around 4.1GB was used for balancing processing time and representativeness of data.

Within Spark, there were a number of analyses run using the MapReduce framework. Among them were calculation of average trip miles and fare by borough, counting per day the number of trips, and looking at the time distributions of trips by borough. The distributed nature of Spark enabled these computations to happen in parallel, significantly reducing processing time. The results were saved for further visualization and analysis.

4. Clustering Techniques

Two types of clustering techniques that described patterns within the data set were applied: K-Means and Hierarchical Clustering. K-Means clustering was conducted on a normalized data set in order to cluster trips with respect to features normalized on distance, time, fare, and weather conditions. Hierarchical clustering was also conducted on part of the data, generating a dendrogram reflecting inter-cluster relationships. We incorporated these clustering techniques to develop ideal urban patterns of mobility behavior, and their outcomes were compared to get meaningful information.

5. Ethical Considerations

The data used in this analysis is freely available and was retrieved from Kaggle and other credible public databases. There exists no PII within the data set, thus ensuring that analytic respect for privacy and unwritten rules of ethics are followed. Only aggregated data are dealt with, ensuring individual anonymity.

6. Scalability and Performance

The methodology is designed in a way to billet the large dataset efficiently; hence, it emphasizes scalability. With Apache Spark, this analysis will take advantage of distributed computing capabilities as its data would be parallelly processed across multiple nodes. This approach ensures that the analysis can handle datasets exceeding 10GB in size while maintaining performance.

The size of the dataset was kept in hand by processing only a representative 4.1GB sample in the Spark environment. This was to cut down the time of processing while still catching the dataset's general trends and patterns. Running K-Means, a very scalable algorithm for clustering, was effectively done on the whole sampled dataset in Spark. Hierarchical clustering is much more memory-intensive and thus was only run on a smaller subset to prevent memory issues.

In general, the scaling of the analysis with the size of the dataset or available computational resources, as shown by these chosen methodologies and tools, will help this approach be applicable to large-scale data analysis scenarios of many kinds.

V. Results

Clustering done on the Uber NYC for-hire vehicle trip data for the year 2021 produced some useful visualization and statistical outputs; these are quite important in giving clues concerning the distribution of the data and the outcome of clustering.

Controlling for these factors, the first visualization, KMeans Clustering of Trips Figure 1, creates a scatter plot of trip miles versus base passenger fare. In the clusters in different colors one may see how trips get aggregated on these two important features.

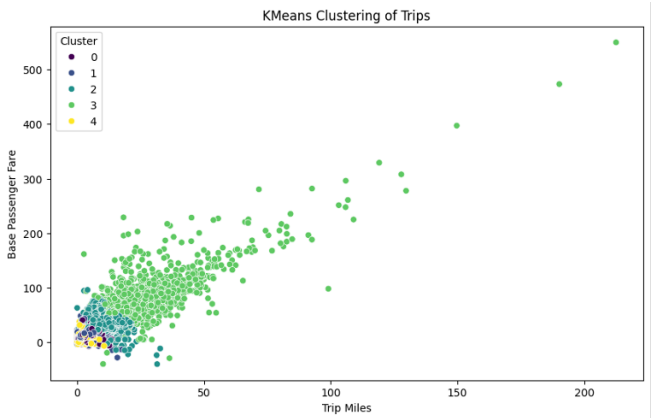

Figure 1: KMeans Clustering of Trips

The clustering dendrogram, as shown in Figure 2, is a hierarchy of structure in the data that illustrates how trips at different levels of similarity are iteratively grouped. This was later used to perform hierarchical clustering.

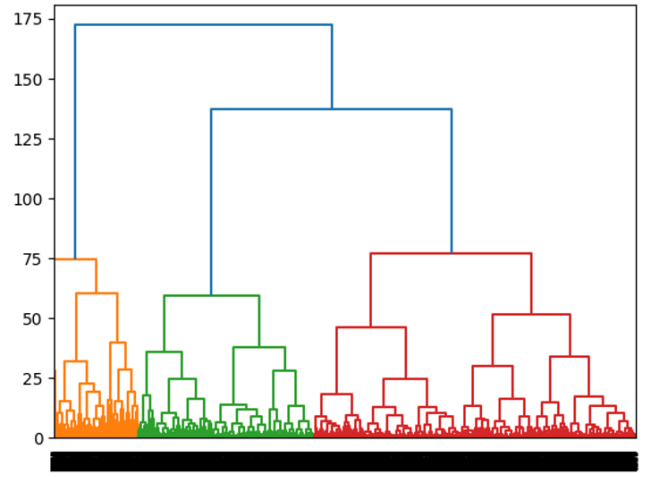

Figure 2: Clustering dendrogram

Hierarchical Clustering of Trips Subset Figure 3 shows the results of the hierarchical clustering algorithm on a smaller subset to prevent memory issues. The different clusters can be distinguished against trip miles and base passenger fare.

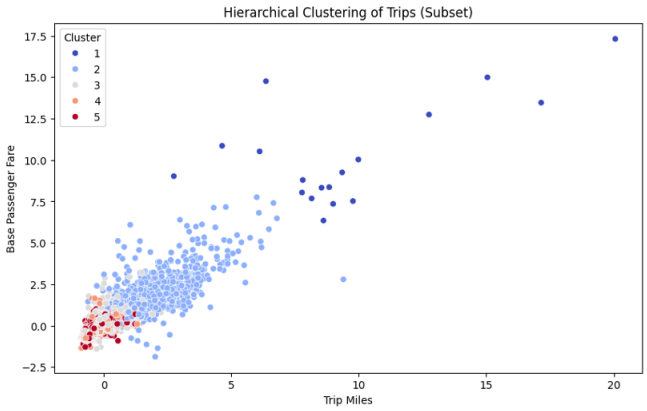

Figure 3: Hierarchical Clustering of Trips (Subset)

Other EDA visualizations include the distribution of trip miles, as shown in Figure 4, and that of base passenger fare, as shown in Figure 5. These two charts indicate how often and to what extent these variables exist within the dataset. Another graph is a scatter plot of trip miles versus base passenger fare, as shown in Figure 6. Figure 7 is the Distribution of Trip Time, that is, the length of time the trips take.

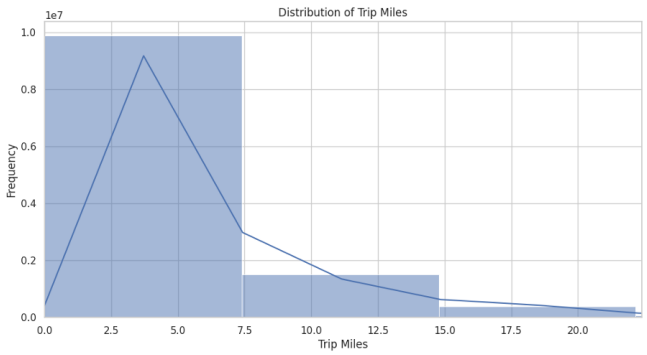

Figure 4: Distribution of Trip Miles

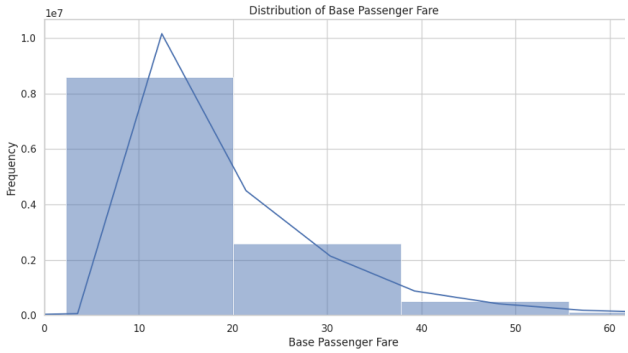

Figure 5:Distribution of Base Passenger Fare

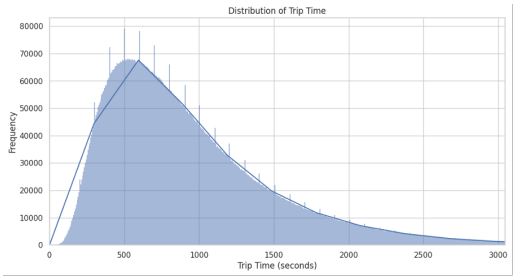

Figure 6: Distribution of Trip Time

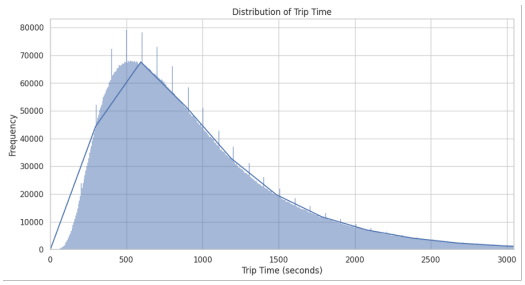

Figure 7: Distribution of Trip Time

Finally, box plots comparing the metrics across different boroughs are Trip Miles by Borough (Figure 8) and Base Passenger Fare by Borough (Figure 9), showing geographical variations. Scatter plots like Trip Miles vs Temperature (Figure 10) and Trip Miles vs Humidity (Figure 11) would define how trip miles vary with temperature and humidity levels respectively.

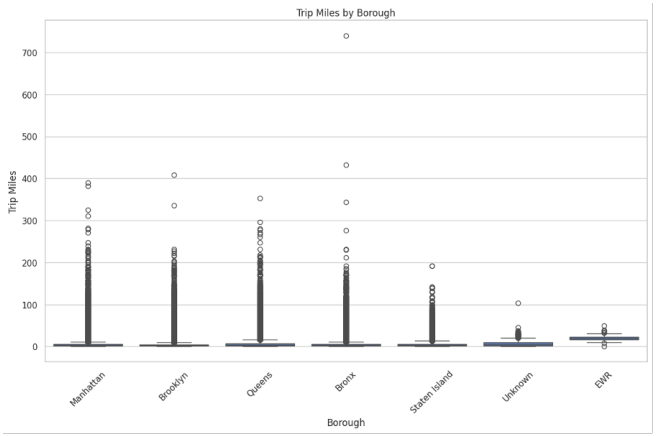

Figure 8: Trip Miles by Borough

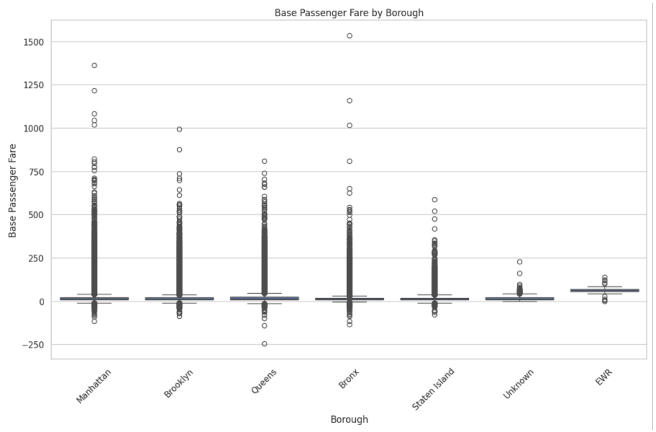

Figure 9: Base Passenger Fare by Borough

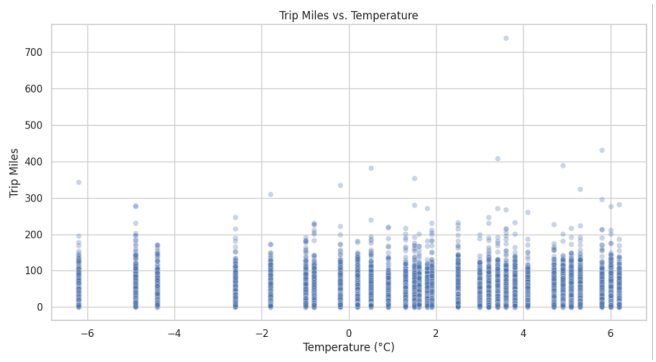

Figure 10: Trip Miles vs. Temperature

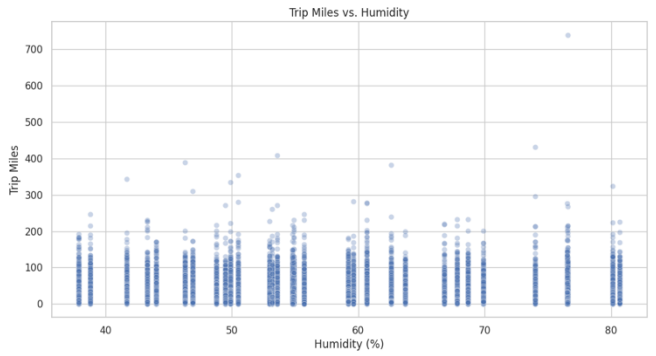

Figure 11: Trip Miles vs. Humidity

Insights

The clustering results based on Uber NYC trip data standalone, however, give several non-trivial insights in their own right. As one can see from Figure 1, KMeans Clustering rather strikingly reveals the fact that trips with lower miles have low base fares and form dense clusters at the bottom of the fare and distance spectrum. On the other hand, outliers at very high fares and distances, though fewer in number, form separate clusters indicative of different trip patterns probably long-distance airport trips or out-of-city travels.

Hierarchical Clustering, shown in Figures 2 and 3, supports these findings. The dendrogram indicates clear divisions from the dataset, matching the number of results expected from KMeans. It also provides a detailed view of how trips are grouped, including the sub-clusters within the larger clusters that KMeans has identified.

The EDA plots further enrich these insights. Figure 4: Distribution of Trip Miles; Figure 5: Base Passenger Fare. Both these distributions are skewed and indicate the typical urban transport pattern, where most trips cover short distances and short fares.

According to the plot of Trip Miles by Borough Figure 8, Manhattan has the highest variability in trip distances, probably due to the_LA central business district. Base Passenger Fare by Borough plots a similar pattern of fares.

The scatter plot for Trip Miles vs. Temperature does not support strong correlations, thus indicating that the temperature does not affect trip distances. On the other hand, Trip Miles vs. Humidity, as manifested in Figure 11, reveals that humidity is another factor with a non-significant effect on distances.

KMeans clustering (Figure 1) worked well with the dataset and did group the similar trips fast according to the features selected. However, in contrast, Hierarchical Clustering shown in Figures 2 and 3 gave detailed insight but was computationally intensive, especially when dealing with Large data sets. Therefore, for larger data sets, KMeans is preferred and for a detailed analysis on small subsets, Hierarchical methods are recommended.

VI. Conclusions & Future Work

In this research paper, an in-depth analysis of the Uber NYC for-hire vehicle trip data in 2021 was carried out, merging weather data and applying KMeans and hierarchical clustering. The results returned distinct patterns of trip distances and fares, where most of the trips were of short distance and low fare, mostly in Manhattan. Outliers of higher fare and longer distances formed separate clusters and indicated special trip categories. It also yielded information on the distribution of trip metrics across the different boroughs and how trip distances are not significantly correlated with climatic conditions such as temperature and humidity.

The insights that can be derived from this study have important implications for urban transportation planning, pricing strategies, and demand forecasting. Aggregations of trips can be beneficial in optimizing route plans, improving the fare structure, and enhancing service delivery within heavy-population locales like New York City.

The limitations of this study include a single city and short time period, probably missing seasonal or long-term trends. Besides, not all possible variables were included in the analysis. Further research may involve comparisons across cities and include more factors such as the road traffic condition, applying other more sophisticated machine learning techniques.

References

[1]. R. Alguliyev, R. Aliguliyev, A. Bagirov and R. Karimov, "Batch clustering algorithm for big data sets," *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, Baku, Azerbaijan, 2016, pp. 1-4, doi: 10.1109/ICAICT.2016.7991657.

[2]. H. Huang, S. He and M. Tabatabaie, "Extreme-Aware Local-Global Attention for Spatio-Temporal Urban Mobility Learning," *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, Anaheim, CA, USA, 2023, pp. 1059-1070, doi: 10.1109/ICDE55515.2023.00086.

[3]. B. Shangguan, P. Yue, Z. Wu and L. Jiang, "Big spatial data processing with Apache Spark," *2017 6th International Conference on Agro-Geoinformatics*, Fairfax, VA, USA, 2017, pp. 1-4, doi: 10.1109/Agro-Geoinformatics.2017.8047039.

[4]. H. Chen, Y. Jia, Z. Hu, G. Wu and Z. -J. M. Shen, "Data-driven planning of plug-in hybrid electric taxi charging stations in urban environments: A case in the central area of Beijing," *2017*

IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Turin, Italy, 2017, pp. 1-6, doi: 10.1109/ISGTEurope.2017.8260264.

[5]. G. Xie *et al.*, "AttractRank: District Attraction Ranking Analysis Based on Taxi Big Data," in *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 1679-1688, March 2021, doi: 10.1109/TII.2020.2994038.

[6]. P. Monmousseau, D. Delahaye, A. Marzuoli and E. Feron, "Door-to-Door Air Travel Time Analysis in the United States using Uber Data," *2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT)*, Singapore, 2020, pp. 1-7, doi: 10.1109/AIDA-AT48540.2020.9049179.

[7]. C. K. Leung, A. G. M. Pazdor and H. Zheng, "Scalable Mining of Big Data," *2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*, Atlanta, GA, USA, 2021, pp. 240-247, doi: 10.1109/SWC50871.2021.00041.

[8]. R. Joshi, A. Patidar and S. Mishra, "Scaling k-medoid algorithm for clustering large categorical dataset and its performance analysis," *2011 3rd International Conference on Electronics Computer Technology*, Kanyakumari, India, 2011, pp. 117-121, doi: 10.1109/ICECTECH.2011.5941667.