

Geospatial Clustering of Pick-up and Drop-off Location

Aabidh musthaq
Decision science department
University of Moratuwa206001F
Moratuwa, Sri Lanka
206001F

I. INTRODUCTION

In the world of taxi services, meeting customer needs and operating efficiently depends on knowing where customers are dropped off and picked up. The dataset I'm working with includes important details like these location coordinates, the total number of passengers, and the associated fares in New York City. The main goal is to use K-Means clustering to divide these coordinates into different groups so that the temporal and spatial distribution of taxi activities can be revealed.

K-means clustering is an unsupervised learning method that is widely used in data clustering tasks. The basic idea behind it is to divide a dataset into K distinct, non-overlapping clusters, with each observation falling into the cluster that has the closest mean. [1] When this taxi fare dataset is used, K-Means clustering is an essential technique for identifying geographic patterns in the pickup and drop-off coordinates. It reveals spatial concentrations or areas of concern suggestive of well-liked taxi pick-up and drop-off locations by grouping these locations. This division makes it possible to pinpoint locations where there is a significant need for both pick-up and drop-off taxi services, providing insight into the habits and preferences of commuters. Furthermore, adding characteristics to these clusters, such as the number of passengers and the fare amounts, might reveal correlations or trends that point to usage patterns or economic factors in various geographical areas.

This study is motivated by the growing reliance on taxi services amid the changing dynamics of urban living. Taxi services play a critical role in encouraging efficient transportation systems, so understanding and optimizing

them is critical. By improving taxi services, we hope to reduce traffic congestion, increase accessibility, and promote overall transportation sustainability in urban areas. This research seeks to dig deeper into the complexities of taxi service dynamics to meet current demands and anticipate and adapt to the evolving needs of urban commuters and the transportation ecosystem.

A. Research Objectives

Finding High-Demand Taxi areas

Using advanced geospatial clustering techniques to identify key zones with increased taxi activity, identifying clusters based on pick-up and drop-off data to reveal concentrated demand areas.

Examining Dynamic Temporal Patterns

Examine the daily, weekly, and seasonal fluctuations in demand to grasp the changing patterns over time and reveal the dynamics of taxi usage trends.

II. TYPE OF ANALYSIS

The chosen analysis for this research centers on the geospatial clustering of pick-up and drop-off locations within the context of New York City's taxi activities. This strategic approach involves employing clustering techniques to group geographical locations into distinct clusters based on shared characteristics, emphasizing the identification of high-traffic and popular areas for taxi services. The analysis extends beyond static spatial patterns, integrating the temporal dimension to capture fluctuations and variations in these popular areas over time. This holistic approach provides a comprehensive understanding of the dynamic nature of taxi service demand within the city, enabling insights not only into spatial concentrations but also into the temporal evolution of these key service areas. [2]

III. METHODOLOGY

A. Data Preprocessing

First, the data must be prepared for analysis. The dataset, which may have contained a large amount of data, was streamlined by randomly choosing a subset of 500,000 rows to work with using specialized libraries like Pandas and NumPy. Rows containing null or empty values or missing data were eliminated to ensure quality. Furthermore, the column labeled 'key' was excluded from the analysis as it might not have been relevant to the analysis.

The dataset was cleaned in the following stage. The 'pickup_datetime' column was transformed into a consistent date and time format that could be processed and understood easily because dates and times frequently come in different formats. This led to the creation of two new columns called "pickup date" and "pickup time," which are used to store the date and time details for each trip independently

Geographical information, including coordinates for latitude and longitude, was carefully checked for accuracy. Rows with inaccurate or unfeasible coordinates (such as latitude values exceeding 90 or longitude values below -180) were recognized and eliminated. Furthermore, occurrences in which all the coordinates for the pickup and drop-off locations were zero were regarded as incorrect and were removed from the dataset as a result. To preserve the integrity of the data falls with zero or negative fare amounts were likewise removed.

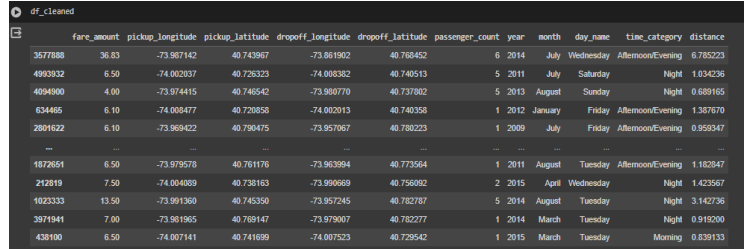
The 'pickup_date' was reformatted into a more explicit date/time format to improve the dataset and extract more meaningful information. As a result, more information could be extracted, including the year, month, and day names, which were then separated into different columns for improved organization and analysis.

Additionally, a thorough analysis of the time of day for every trip was made. Time ranges were defined, classifying trips according to their pickup times into groups like Morning, Afternoon/Evening, Night, and Early Morning.

'time_category' was the new column where this classification was kept.

In addition, the geodesic distance formula was used to determine the distance between the pickup and drop-off locations. This made it easier to spot trips with a zero distance, which could be an error distance traveled. To ensure the accuracy and reliability of the subsequent analysis, these instances, as well as data points corresponding to locations over water bodies, were systematically removed from the dataset.

Following these careful stages of cleaning and improvement, the dataset was eventually produced in a refined form free of mistakes such as missing values, inaccurate coordinates, journeys with zero distances, or fare amounts that were either zero or negative. The final cleaned data set looks like the below image.



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	year	month	day_name	time_category	distance
3677888	36.83	-73.987142	40.743967	-73.961902	40.768452	6	2014	July	Wednesday	Afternoon/Evening	6.785223
4893832	6.50	-74.002037	40.726323	-74.008382	40.740513	5	2011	July	Saturday	Night	1.034236
4094900	4.00	-73.974415	40.746542	-73.980770	40.737802	5	2013	August	Sunday	Night	0.689165
634486	6.10	-74.008477	40.730858	-74.002013	40.740358	1	2012	January	Friday	Afternoon/Evening	1.387670
2801622	6.10	-73.969422	40.786475	-73.957067	40.780223	1	2009	July	Friday	Afternoon/Evening	0.958347
...
1872651	6.50	-73.979578	40.761176	-73.963994	40.773564	1	2011	August	Tuesday	Afternoon/Evening	1.180347
212819	7.50	-74.004089	40.738183	-73.990669	40.756092	2	2015	April	Wednesday	Night	1.423587
1823333	13.50	-73.991360	40.745350	-73.957245	40.782787	5	2014	August	Tuesday	Night	3.142736
3971941	7.00	-73.981965	40.769147	-73.979007	40.782277	1	2014	March	Tuesday	Night	0.919280
436100	6.50	-74.007141	40.741899	-74.007523	40.729542	1	2015	March	Tuesday	Morning	0.839133

Table 1- cleaned data.

B. Algorithm Selection

K-means clustering is selected because it works well with spatial data and is easy to understand. The places where taxis pick people up and drop them off are geographically defined by latitude and longitude. K-means, which attempt to cluster similar points together based on their proximity, work well with numerical, continuous data, such as coordinates. [1] Because K-means is computationally efficient, it can be applied to large datasets where efficiency is important, such as taxi location data. K-means can effectively cluster pick-up and drop-off locations with potentially millions of data points without imposing a significant computational burden. [3] K-means are simple to understand. Since they are defined by centroids, it is easy to distinguish between areas that are less frequently visited and areas with high taxi activity. K-means can adjust to evolving data patterns. The utilization of clustering analysis across various time intervals facilitates the comprehension of the temporal evolution of popular taxi

areas. By repeating the clustering algorithm for distinct time segments, K-means can capture these temporal changes in taxi demand, which may vary by hour, day, or season. [4]

Finally, K-means clustering appears to be a suitable option when considering the intrinsic spatial nature of taxi pick-up and drop-off locations defined by latitude and longitude, as well as the requirement for a computationally efficient method capable of handling large datasets.

IV. RESULTS

When it comes to the K-means cluster algorithm first we need to select an appropriate number of clusters. To do this we used the Optimal 'k' Value using the Elbow Method which is based on the within-cluster sum of squares (WSS) or distortion across different values for 'k' using Python. When looking at the graph that shows how the number of clusters ('k') and the WSS relate to each other, among can see a pattern. As 'k' increases, the graph first shows a sharp decline in the WSS, which is followed by a more gradual decline. The optimal 'k' value is indicated by the point on the graph where this decrease abruptly shifts and resembles an 'elbow'. In the provided graph, this inflection point occurs at 'k=4', signifying a substantial reduction in WSS. Consequently, 'k=4' is suggested as the most suitable number of clusters for this dataset. [5]

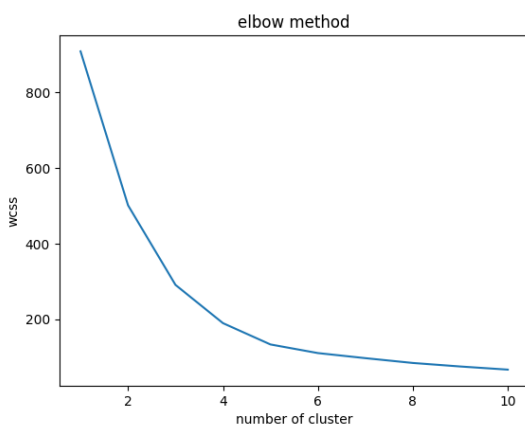


Figure 1- pick-up elbow

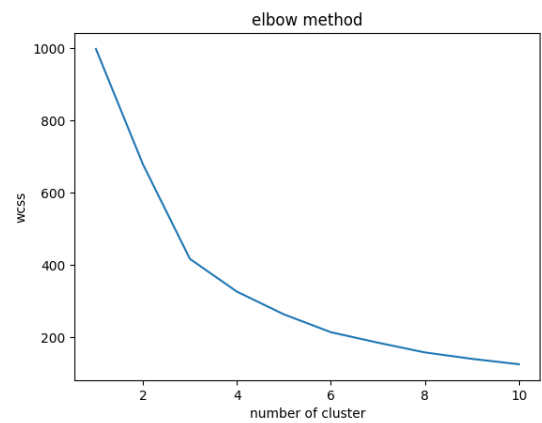


Figure 2- drop-off elbow

Based on the elbow method analysis, 'k=4' emerges as the optimal number of clusters for this dataset, offering a balance between capturing meaningful patterns within the data and avoiding overfitting or excessive complexity in the clustering model.

The subsequent analysis revealed intriguing patterns, as shows in the following graphs using QGIS.

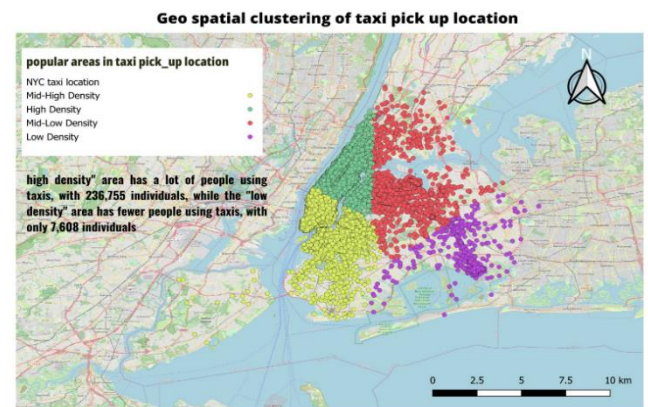


Figure 3- Taxi pick-up cluster

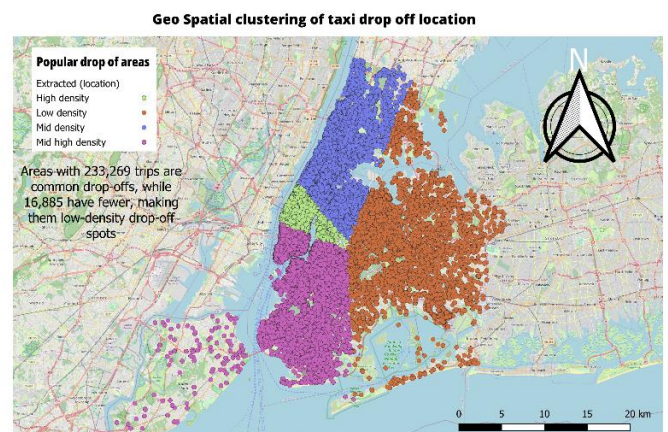


Figure 4- Taxi drop off cluster.

Figure 3 and 4 shows a clustering analysis based on pickup location & drop off-location data in New York City, with the clusters classified into four density levels: high density, mid-high density, mid-low density, and low density. The figures above represent the number of pickup points in each cluster.

The analysis of taxi service data in New York City reveals distinct density clusters for both pick-up and drop-off points. The high-density cluster represents areas with a significant concentration of activity, comprising 236,755 pick-up points and 233,269 drop-off points. In contrast, the mid-high-density category, with 224,991 pick-up points and 121,382 drop-off points, signifies locations with moderately high taxi service activity. Moving to the mid-low-density cluster, characterized by 13,330 pick-up points and 112,224 drop-off points, these areas exhibit a moderate level of taxi service engagement. Finally, the low-density cluster stands out with 7,608 pick-up points and 16,885 drop-off points, signifying areas with relatively lower taxi service activity. This clustering approach helps identify and categorize areas based on their varying levels of demand and usage, providing valuable insights for optimizing service delivery and resource allocation within the city.

High-density clusters are likely to represent areas with a high frequency of pickups within New York City. This cluster represents areas with a significantly high concentration of taxi service pick-up locations. Mid-high-density areas are likely to be crowded and busy, indicating a high demand for taxi services. The mid-low-density cluster represents the areas with a moderate density of taxi service pick-ups. These areas, which are located between high-density and low-density areas, may have a more balanced demand for taxi services. The low-density cluster includes areas with a low concentration of taxi service pick-up points. When compared to other clusters, these areas may be less populated or have a lower demand for taxi services.



Figure 4- Pick-up heatmap

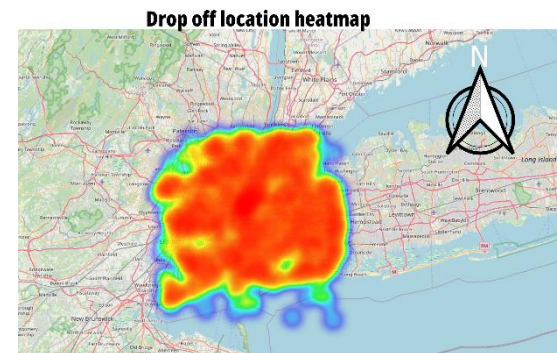


Figure 5- Drop off heatmap.

The heatmaps reveal hotspots where taxi service activities are most concentrated, assisting in identifying areas where transportation services are in high demand. To visualize the spatial distribution of these activities across New York City, a heatmap was created using QGIS, shown in Figures 4 and

5.

The drop-off heatmap visualization shows the intensity and concentration of taxi service drop-offs across various areas of New York City. Areas shown in warmer colors (e.g., red, orange) on the heatmap represent higher-density clusters of drop-off points, indicating regions with a high number of taxi service destinations. Cooler colors (e.g., blue, and green) represent areas with lower drop-off concentrations, highlighting locations with fewer destinations.

The heatmap for pick-up locations, on the other hand, depicts the distribution and density of taxi service pick-up points throughout New York City. Warmer colors on the pick-up heatmap, like the drop-off heatmap, indicate high-density clusters of pick-up points, indicating areas with high demand for taxi services. Cooler colors denote regions with fewer

pick-up locations, indicating areas with lower demand for taxi services.

V. DISCUSSION

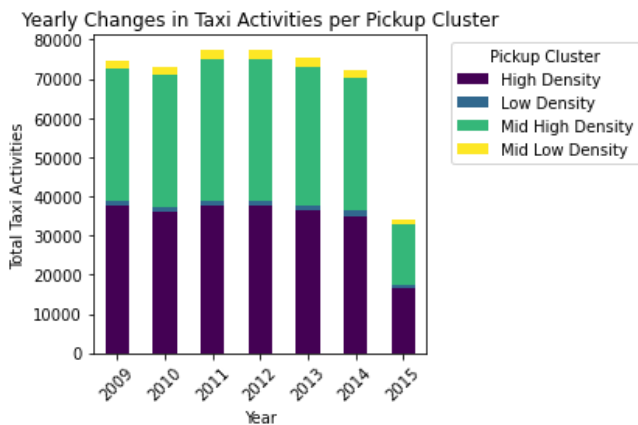


Figure 6- taxi activities based on year and cluster.

The graph depicts taxi activity from 2009 to 2015, revealing a consistent trend until 2015 when there was a significant downturn. Industrial production has been negating for 12 months in a row, while retail sales growth has slowed, indicating economic fragility [6]. Global weaknesses exacerbated this economic vulnerability. The apparent decline in taxi activity in 2015 corresponds to the economic slowdown, implying potential implications for consumer spending habits and transportation preferences [7]. The relationship between reduced taxi utilization and the economic slowdown highlights the interplay between economic conditions and consumer behavior, implying that economic uncertainty may have an impact on transportation choices.

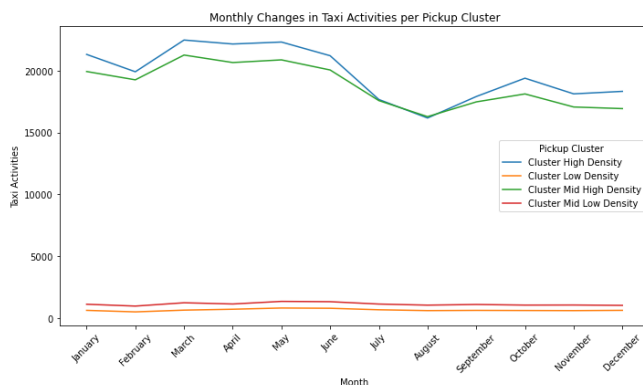


Figure 7- monthly taxi pattern changes

There was a slight decrease in high-density and mid-high-density clusters from January to February, possibly indicating reduced taxi activity during the winter months. The period from March to June, on the other hand, saw a resurgence in these clusters, reflecting an increase in taxi rides during the spring and early summer. However, it fell in July, possibly due to seasonal shifts or changes in transportation preferences associated with the summer season. August saw a larger drop than July, indicating a significant shift in taxi demand during the late summer. Following that, a gradual increase in high-density clusters was observed from September to October, indicating a potential rebound in taxi rides during the transition to the fall season. The pattern changed in November with another drop, which could have been influenced by external factors such as holidays or changing weather conditions. Finally, minor fluctuations were observed in December, indicating stability or minimal changes in clusters as the year ended. [2]

This graph provides critical insights into the temporal dynamics of taxi services in New York City, laying the groundwork for future research into the factors influencing these observed trends.

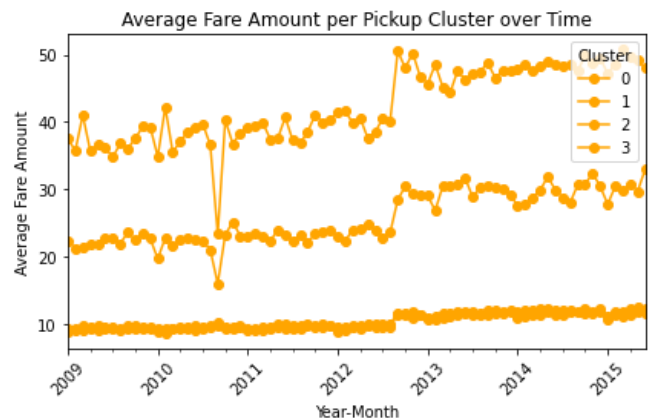


Figure 8- Fare amount over time

From 2009 to 2015, the visual representation shows a consistent increase in taxi fare prices across various cluster areas (high density, low, low middle, high middle) in New York City. Fare prices increased significantly each year in comparison to the previous year.

Several underlying factors are contributing to this upward trend. Inflation-increased operational costs, and fuel price fluctuations all have a significant impact on fare increases. Regulatory changes in the taxi industry, technological advancements, and changes in demand-supply dynamics all play important roles in shaping these fare increases. [8]

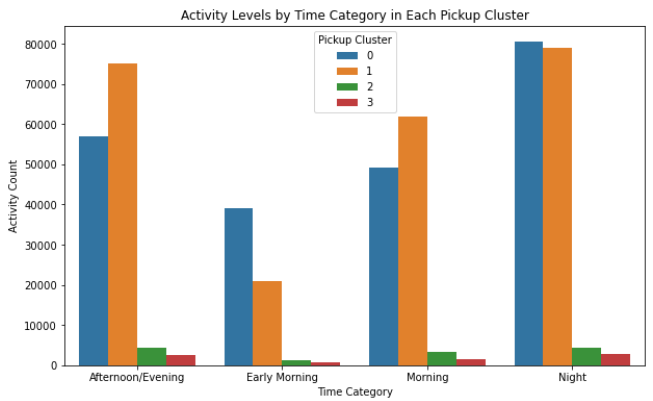


Figure 9- Activity Levels by Time Category

Taxi ride patterns in New York City vary greatly throughout the day. The night has a spike in taxi rides, which is likely due to late-night activities, events, and people returning home from various engagements. Due to bustling social activities and shifts ending at unusual hours, demand is high during this period. [9]

A significant number of taxi rides continue into the evening and afternoon. These times correspond to rush hours when people leave work or engage in recreational activities, which increases the demand for taxi services. The confluence of commuters and those engaged in recreational activities significantly contributes to this increased demand. [10]

In the morning When compared to the peaks, this graph depicts moderate taxi rides. This could be attributed to people commuting to work, but because of staggered work schedules or different transportation preferences in the morning, it is typically less than the evening rush. [11]

The early morning indicates a decrease in taxi rides. Fewer people are traveling during this time, which could be due to fewer activities, reduced commuting needs, or people who

have already arrived at their destinations from the previous night's activities. [12]

These patterns reflect typical urban commuting behavior and social dynamics that influence taxi demand in New York City throughout the day. Nighttime and evening peaks correspond to social activities and commuting patterns, whereas morning and early morning periods have lower demand due to different routines and decreased activity.

This in-depth understanding of taxi activity has far-reaching implications for city transportation planning. High-density clusters denote areas in need of strong transportation infrastructure to handle heavy traffic and demand. Low-density areas, on the other hand, may benefit from optimized service allocation and route planning to ensure efficient resource utilization.

The observed temporal patterns in taxi activity, particularly the variations across months, provide important insights for transportation policymakers. Understanding these trends aids in the efficient deployment of resources, ensuring adequate service coverage during peak demand periods, and optimizing operational strategies during low demand periods.

Furthermore, the consistent increase in taxi fare prices across various cluster areas from 2009 to 2015 indicates economic implications. Rising operational costs, because of inflation and fluctuating fuel prices, may have an impact on consumer transportation choices. Fare dynamics are also influenced by regulatory changes and technological advancements, which may affect travel behavior. Recognizing the ups and downs of taxi demand enables city planners and ride-sharing services to improve user convenience by ensuring adequate availability during peak periods while optimizing operations during low-demand periods.

The comprehensive analysis of taxi activity trends, combined with observed temporal patterns and fare dynamics, highlights the complex relationship between economic conditions, consumer behavior, and transportation

preferences. These findings provide a solid foundation for making informed decisions about urban transportation planning.

VI. CONCLUSION

The application of K-means clustering techniques to the dataset of New York City taxi services has yielded critical insights into the spatial, temporal, and economic dynamics of urban transportation. The clustering analysis successfully identified distinct density clusters for both pick-up and drop-off locations, delineating high-demand areas and allowing for a better understanding of taxi service usage patterns throughout the city. Temporal analyses revealed nuanced patterns in taxi activity throughout the year, highlighting fluctuations over months and years. These trends showed correlations with economic indicators, indicating that economic conditions have an impact on transportation preferences and consumer behavior. Notably, the consistent rise in fare prices across cluster areas highlights the impact of inflation, operational costs, regulatory changes, and technological advancements on transportation options. The findings provide useful guidance for urban transportation planning. High-density clusters indicate the need for strong infrastructure, whereas low-density areas highlight opportunities for better resource allocation and route planning. The temporal insights serve as a foundation for resource allocation strategies during peak and off-peak demand periods.

VI. FUTURE RESEARCH DIRECTION

- Technological Integration: Examine the effects of emerging technologies on taxi service dynamics, such as ride-sharing apps or electric vehicles, with a focus on their impact on user preferences and service demand.
- Consumer Behavior Analysis: Examine the relationships between economic indicators and

transportation choices to predict shifts in consumer behavior during economic cycles.

VII. REFERENCES

- [1] I. Dabbura, "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks," 17 sep 2018. [Online]. Available: <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>.
- [2] Y. W. ., J. X. ., X. Z. ., J. S. Xiao-Jian Chen, "Urban hotspots detection of taxi stops with local maximum density," *Science direct*, vol. 89, 2021.
- [3] Rustam Mussabayev, Nenad Mladenovic , Bassem Jarboui , Ravil Mussabayev , "How to Use K-means for Big Data Clustering?," *elsevier*, vol. 137, 2023 .
- [4] Huimin Luo , Jianming Cai , Kunpeng Zhang , Ruihang Xie , Liang Zheng , "A multi-task deep learning model for short-term taxi demand forecasting considering spatiotemporal dependences," *Science direct*, vol. 8, pp. 83-94, 2021.
- [5] B. Saji, "Elbow Method for Finding the Optimal Number of Clusters in K-Means," 21 september 2023. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/>. [Accessed 10 10 2023].
- [6] M. T. Kiley, "Financial and Macroeconomic Indicators of Recession Risk," 2022.
- [7] O. McEvoy, "The Great Recession in the U.S. - Statistics & Facts," *statista*, 2023.
- [8] S. Woodhouse, "Taxi Fares Are Going Up 23% in New York City," 2022.
- [9] J. M. F. M. Joya Deri, "New York city taxi analysis with graph signal processing," *ResearchGate* , 2016.
- [10] C. Y. Eric J. Gonzales, "Modeling Taxi Trip Demand by Time of Day in New York City," *ResearchGate*.
- [11] S. Faghih, "Understanding and Modeling Taxi Demand Using Time Series," 2019.
- [12] C. .. W. ., L. W. Ruijie (Rebecca) Bian, "Estimating spatio-temporal variations of taxi ridership caused by Hurricanes Irene and Sandy: A case study of New York City," *ScienceDirect*, vol. 77, 2019.