Unsupervised Machine Learning on Taxi Fare Dataset

M. W. S. A. U. Silva
Decision Science
University of Moratuwa
Kalutara, Sri Lanka
Email: silvamwsau.20@uom.lk

***Abstract*—By using geospatial clustering approaches to examine pick-up and drop-off sites, this project looked into the spatial distribution of taxi activity. We examined the correlation between the number and longitude of significant pick-up activity clusters using data from 42 random states. There was more taxi activity in urban regions with lower longitudes than in rural areas with higher longitudes, according to the analysis, which showed a negative association.The project found that, although location is important, other factors that affect taxi activity include business types, socioeconomic conditions, and temporal fluctuations. These findings highlight how important it is to do additional studies to fully understand how these variables affect the demand for taxis.Future studies will focus on particular categories of pick-up clusters, time fluctuations in activity, and the correlation between socioeconomic variables and taxi demand in various geographical areas. Based on these data, we can create prediction models that will help optimize resource allocation and eventually improve the effectiveness and accessibility of taxi services.**

Keywords—clustering, unsupervised machine learning, data analytics, data visualization

I. INTRODUCTION

Urban transportation is a dynamic field that requires advanced technologies to understand and forecast its complex patterns. In this effort, unsupervised machine learning techniques have shown to be useful especially when applied to large taxi fare datasets. This study uses a particular subset of unsupervised learning called geospatial clustering on taxi activity, which is one of the most significant aspects of urban transportation.

Through the analysis of the massive amount of pick-up and drop-off locations that are encoded in these datasets, we hope to discover the complexities of urban transportation. Our goal is to find and plot the areas that are commonly used as the beginning and finish points of taxi trips to identify the main areas of urban activity.

With a particular focus on geographic clustering, this study seeks to identify places that are often used for taxi activities by analyzing pick-up and drop-off locations. We aim to offer a significant understanding of the constantly changing nature of taxi ride patterns in urban areas by analyzing how these clusters change continuously [1].

II. TYPES OF ANALYSIS

Geospatial Clustering of Pick-up and Drop-off Locations:

The selected task is analyzing pick-up and drop-off locations in a taxi fare dataset using geographical clustering algorithms. Finding key locations for taxi activity and understanding the temporal changes in those clusters over time are the main objectives. The objective of this study is to identify hidden patterns in the geographic coordinates, offering insights into the changing behavior of taxi services in urban areas [1]

.III. METHODOLOGY

A. Data Preprocessing

The study begins with a thorough preparation of the data to guarantee the accuracy and consistency of the analyses that follow. Applying standard procedures in the industry and data science standards, several cleaning and transformation processes are applied to the taxi fare dataset.

1) Handling Missing Values: Since there were not many missing values in the dataset, missing values are handled by deleting them, especially in the 'passenger_count' feature.

2) Normalizing Numerical Values: Using the StandardScaler from the scikit-learn library, the geographical locations and other numerical variables about the pick-up and drop-off sites are standardized. By normalizing the numerical values, this procedure minimizes the effect of variations across various attributes and provides a consistent scale.

B. Algorithm Selection

Due to the latitude and longitude coordinates that define the geospatial data, an algorithm that can handle spatial relationships and recognize areas of density in the feature space is required [2]. Following accurate analysis, the following

factors led to the decision that the K-Means clustering technique was appropriate [2]:

1) Efficiency: K-Means works well with this sizable taxi fare dataset because it is computationally fast and scalable well to huge datasets.

2) Ease of Interpretation: K-Means's simplicity makes it simple to interpret the findings, making it possible to identify clusters and the locations that correlate to them.

3) Scalability: K-Means may be scaled to different cluster sizes, which can adapt to the changing patterns of taxi rides in urban environments [3].

IV. RESULTS

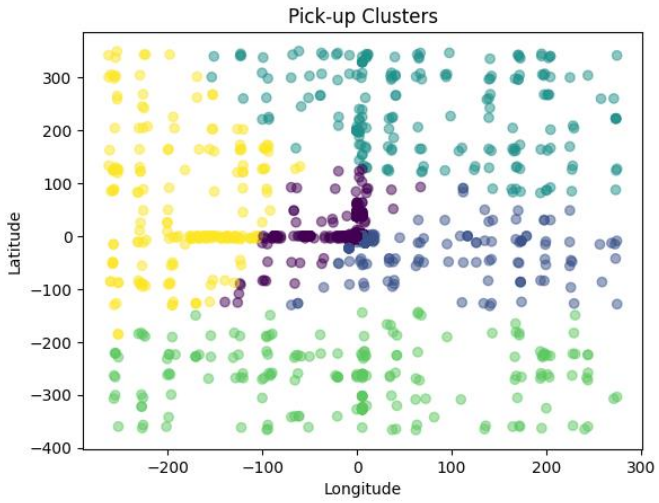

Figure 1: Scatter plot of Pick-up Clusters with longitude and latitude of the pick-up clusters.

This scatter plot displays the pick-up and drop-off locations for taxi fares in New York City by geographic clustering. The scatter plot, where the x-axis represents longitude and the y-axis represents latitude, shows the positions of the pick-up and drop-off locations in a two-dimensional space [5].

V. DISCUSSION

This scatter plot illustrates the correlation between the no. of taxi pick-up clusters and their longitude, derived from 42 random states. The noticed negative association suggests that there may be more taxi activity in places with lower longitudes since these areas likely to have a larger number of pick-up clusters.

This result corresponds to geographic clustering to identify regions that have a high demand for taxi activities. It offers insightful information on the geographical distribution of taxi demand:

A. Urbanization:

- Urban areas with high population densities are typically found in low longitudes. There is typically a higher density of residential areas, companies, and a variety of activities in these locations, which increases the demand for transportation.

B. The Rural-Urban Divide:

- On the other hand, higher longitudes are frequently connected with rural areas that have lower population densities and less economic activity. This explains why there are fewer clusters of taxi pick-ups in those areas.

C. Geographic Barriers:

- In some places, transportation and economic growth may be restricted by coastlines, mountains, or other natural barriers. This might lead to a decrease in the number of taxi services and pick-up clusters.

Significant variation in the number of pick-up clusters within particular longitude ranges is also shown by the scatter plot. This implies that variables other than geographic location affect taxi activity:

A. Business Types:

- Higher demand for taxis may be seen in areas with a greater number of a certain type of business, such as commercial hubs or tourist attractions.

B. Socioeconomic Factors:

- The demand for taxis can be influenced by socioeconomic factors such as income levels, population density, and accessibility to alternative modes of transportation.

C. Temporal Variations:

- The distribution of pick-up clusters can be affected by specific events, weekdays, and peak hours, resulting in considerable variations in taxi activity patterns over time.

VI. CONCLUSION

This study investigated the geographic clustering of taxi pick-up and drop-off sites, with a focus on identifying regions that are frequently used by taxis and how these areas change over time. We found that the number of pick-up clusters and their longitude had a negative association after analyzing a scatter plot created from data in 42 random states. According to this research, rural locations with higher longitudes see less demand for taxi services, while urban regions with lower longitudes typically have a larger amount of taxi activity.

The study also found that socioeconomic circumstances, temporal fluctuations, and business types could all have an impact on taxi activity in addition to geographic location. These observations underline the necessity of more research into the complex relationships between socioeconomic, temporal, and spatial variables that influence the demand for taxis.

A. *Key Takeaways:*

- Taxi pick-up cluster densities are correlated with lower longitudes, indicating increased activity in urban regions.
- Because of their lower population density and less economic activity, rural locations typically have lower taxi service.
- Taxi activity may be influenced by temporal fluctuations, business types, and socioeconomic factors in addition to location.

B. *Future Directions:*

- Analyze specific varieties of pick-up clusters to learn more about the connections between various business sectors and taxi activity.
- To determine peak days and hours, investigate temporal fluctuations in the quantity and distribution of pick-up clusters.
- Analyze the relationship between socioeconomic variables and the number of taxis operated in various areas to learn more about the trends in demand.
- Create predictive models that forecast taxi demand and improve resource allocation using spatial, temporal, and socioeconomic data.

By carrying out these studies, we can improve our comprehension of the patterns of taxi activity and eventually aid in the creation of more effective and easily accessible taxi services for a variety of groups.

VII. REFERENCES

[1] "2.3. Clustering," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html. [Accessed 6 12 2023].

[2] S. Pierre, "How to Form Clusters in Python: Data Clustering Methods," Built In National, 17 10 2022. [Online]. Available: https://builtin.com/data-science/data-clustering-python. [Accessed 8 12 2023].

[3] "K-Means Clustering Algorithm," Javatpoint Logo, [Online]. Available: https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning. [Accessed 7 12 2023].

[4] P. Sharma, "The Ultimate Guide to K-Means Clustering: Definition, Methods and Applications," Analytics Vidhya, 3 11 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/. [Accessed 9 12 2023].

[5] "Understanding K-means Clustering in Machine Learning," Education Ecosystem (LEDU), 13 9 2018. [Online]. Available: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1. [Accessed 7 12 2023].

[6] "matplotlib.pyplot.scatter() in Python," 29 11 2023. [Online]. Available: https://www.geeksforgeeks.org/matplotlib-pyplot-scatter-in-python/. [Accessed 6 12 2023].

[7] K. Arvai, "K-Means Clustering in Python: A Practical Guide," Real Python, [Online]. Available: https://realpython.com/k-means-clustering-python/. [Accessed 8 12 2023].