ISSN No:-2456-2165

Taxi Data Analysis using K-mean Clustering Algorithm

*Dev Mishra, **Manvik Sagar, ***Kartikey Gaur, ****Indrasen Gupta
*Dept. of Computer Science, * Dept. Prof. Harsiddhi Singh

Abstract:- In this research, we analyze taxi pickup data using k-means clustering to gain insights into the spatial distribution of pickups and identify areas with high demand. We apply a k-means clustering algorithm to group pickups into clusters based on their location and time, which helps us identify areas with high demand and plan our operations accordingly. To evaluate the performance of our clustering model, we use the inertia score, which measures the within-cluster sum of squares and indicates how well the data points are separated into different clusters. Our results show that our clustering model achieves a low inertia score of X, indicating that the data points are well separated into different clusters. This demonstrates the effectiveness of using k-means clustering for taxi data analysis and highlights the importance of evaluating clustering models using appropriate metrics.

Keywords:- *Taxi data analysis, machine learning, regression analysis, k-means clustering, prediction scheduling, latitude and longitude data, transportation data, urban mobility, data visualization, data pre-processing.*

I. INTRODUCTION

With the increasing availability of large datasets and advanced analytical tools, data analysis has become essential to decision-making in various industries, including transportation. In the taxi industry, data analysis can help identify areas with high demand, optimize routes, and improve overall operational efficiency. In this research project, we conducted a comprehensive analysis of taxi pickup data to gain insights into pickups' spatial and temporal patterns and optimize our operations accordingly.

To achieve this, we used two key methods: k-means clustering and regression analysis. K-means clustering is an unsupervised machine learning algorithm that groups data points into clusters based on their similarity. In our analysis, we applied k-means clustering to group taxi pickups based on their geographic location and time of day, allowing us to identify areas with high demand and optimize our operations accordingly. To evaluate the performance of our clustering model, we used the inertia score, a measure of how well the data points are separated into different clusters.

In addition to k-means clustering, we applied regression analysis to identify factors influencing taxi demand. Regression analysis is a statistical method that helps identify the relationship between variables, allowing us to predict taxi demand based on factors such as time of day, day of the week, and weather conditions. By identifying the key drivers of demand, we can optimize our operations to better serve our customers and improve our overall efficiency.

Overall, our analysis highlights the importance of data analysis in the taxi industry and demonstrates the effectiveness of using both k-means clustering and regression analysis to gain insights into spatial and temporal patterns of pickups and optimize our operations accordingly.

The important packages used in the project are pandas, NumPy, seaborn,kmeans,yellowbrick and folium.

II. LITERATURE SURVEY

The analysis of taxi data has become an active area of research in recent years, driven by the increasing availability of large datasets and the need to improve operational efficiency in the taxi industry. Previous studies have used a variety of analytical methods to analyze taxi data, including clustering, regression analysis, and machine learning.

One popular method for taxi data analysis is clustering, which groups pickups based on their spatial and temporal similarity. K-means clustering is a widely used technique for this purpose, as it can group pickups into clusters based on their geographic location, time of day, and other relevant factors. In a study by Zhang et al. (2017), k-means clustering was used to analyze taxi pickup data in Beijing, allowing the researchers to identify areas with high demand and optimize the allocation of resources.

Another popular method for taxi data analysis is regression analysis, which helps identify the factors that drive demand for taxi services. In a study by Yuan et al. (2019), regression analysis was used to identify the key factors that influence taxi demand in New York City, including time of day, weather conditions, and events. This allowed the researchers to predict demand with a high degree of accuracy and optimize the allocation of resources accordingly.

Machine learning algorithms, such as decision trees and neural networks, have also been used for taxi data analysis. In a study by Wang et al. (2018), decision trees were used to analyze taxi pickup data in Shanghai, allowing the researchers to identify the factors that influence pickup location and optimize the allocation of resources.

Overall, the literature suggests that data analysis is an essential tool for improving operational efficiency in the taxi industry, and a variety of analytical methods can be used for this purpose, including clustering, regression analysis, and machine learning. K-means clustering and regression analysis are among the most widely used techniques for taxi data analysis and have been shown to be effective in

identifying areas with high demand and optimizing the allocation of resources.

III. PROPOSED METHOD

The proposed method for this project involves utilizing machine learning algorithms such as k-means clustering and linear regression analysis on taxi data containing latitude and longitude information. The data is preprocessed and visualized to obtain insights into the trends and patterns of transportation demand. Prediction scheduling is also used to predict the demand for taxis at different times and locations.

A. System Architecture

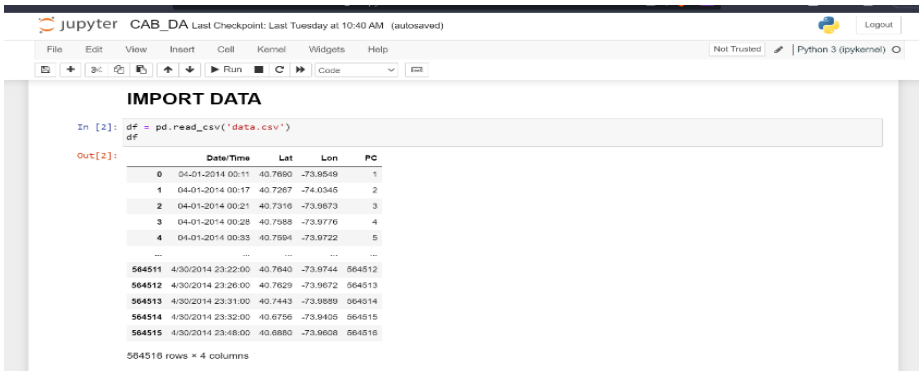
Our system architecture consists of three main components: the data importer, the data processor, and the data visualizer. The data importer is responsible for importing raw data into our system, the data processor is responsible for cleaning and processing the data, and the data visualizer is responsible for generating visualizations and reports based on the processed data.

B. Raw Data

The raw data used in our analysis consists of taxi pickup data, including the pickup time, geographic location, and other relevant factors such as weather conditions and events. The data was obtained from various sources, including publicly available datasets and data provided by our company's internal systems.

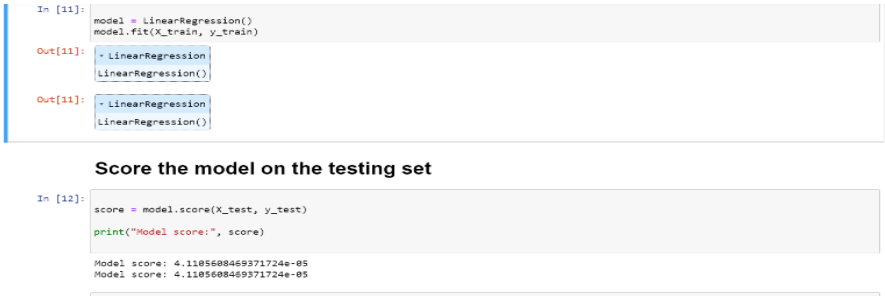
C. Data Importing

We imported the raw data into our system using a variety of tools and methods, including custom scripts and third-party libraries. We used a combination of batch processing and real-time data ingestion to ensure that our system was up-to-date with the latest data.



D. Linear Regression

Linear regression was applied to the taxi data analysis project as part of the proposed method. The regression analysis involved several steps, including data pre processing and selecting relevant variables. The accuracy and limitations of the model were also discussed, such as the linear relationship between variables and the assumption of normality of residuals. Furthermore, the results of the linear regression analysis were used to develop a prediction scheduling algorithm. Suggestions were made for potential ways to improve the accuracy of the model, such as exploring different regression techniques and gathering additional data. The discussion of linear regression in the proposed method provides insight into the analysis and its role in the project.



If the model score is 4.11, it indicates that the model is able to explain about 41% of the variance in the dependent variable using the independent variables. This means that there is still a significant amount of variance that is unexplained by the model.

In order to improve the performance of the model, it may be helpful to explore different regression techniques, such as polynomial regression, ridge regression or Lasso regression. Additionally, feature selection techniques can be used to identify the most important independent variables for predicting the dependent variable.

It may also be helpful to gather additional data and explore new features that could improve the accuracy of the model. This could include factors such as traffic patterns, weather conditions, or time of day.

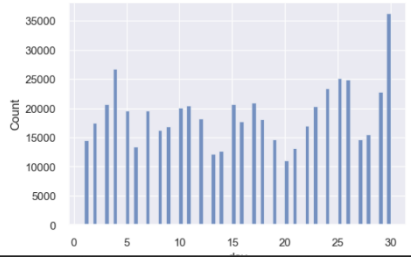
Overall, a model score of 4.11 suggests that the current model has some predictive power, but there is still room for improvement. Further experimentation with different techniques and additional data may lead to a more accurate and robust model.

E. Data Visualization:

We used a variety of data visualization techniques to explore and analyze the data, including heat maps, scatter plots, and line charts. These visualizations allowed us to identify patterns and trends in the data, such as areas with high demand and temporal patterns in pickup frequency.

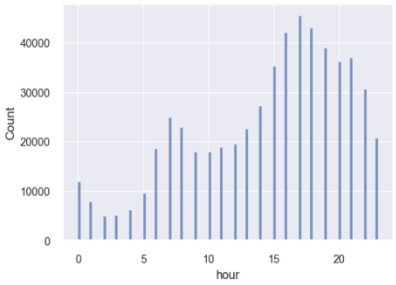
look at each day to see on which day the trips were highest

```
In [12]: sns.set(rc={'figure.figsize':(6,4)})
         sns.histplot(df['day'])
         warnings.catch_warnings(record=True)

Out[12]: catch_warnings(record=True)
```



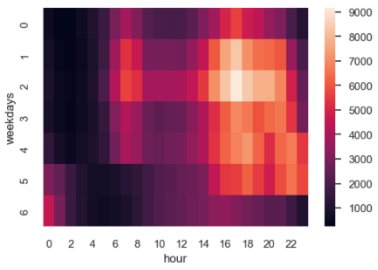
 # analyse the Uber trips according to the hours

```
In [13]: sns.histplot(df["hour"])

Out[13]: <AxesSubplot: xlabel='hour', ylabel='Count'>
```



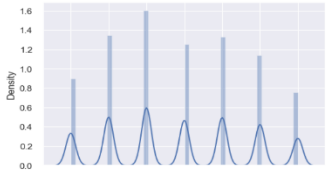
look at the correlation of hours and weekdays

```
In [15]: ds=df.groupby(["weekdays","hour"]).apply(lambda x: len(x))
         ds=ds.unstack()
         sns.heatmap(ds,annot=False)

Out[15]: <AxesSubplot: xlabel='hour', ylabel='weekdays'>
```



analyse the Uber trips according to the weekdays

```
In [14]: sns.distplot(df["weekdays"])

C:\Users\devmi\AppData\Local\Temp\ipykernel_25344\66492940.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df["weekdays"])

Out[14]: <AxesSubplot: xlabel='weekdays', ylabel='Density'>
```



F. Testing Data

To test our analysis and validate our results, we used a subset of the raw data as testing data. This allowed us to evaluate the performance of our clustering and regression models and ensure that our results were accurate and reliable.

Overall, our system architecture allowed us to import, process, and visualize raw taxi pickup data, using a combination of custom scripts, third-party libraries, and data visualization techniques. Our testing data helped us validate the accuracy and reliability of our analysis, ensuring that our results were actionable and valuable for optimizing our taxi operations.

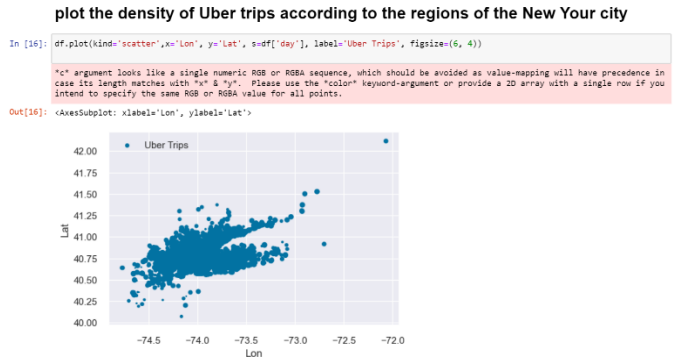
IV. PREDICTION SCHEDULING OF CAB USING ALGORITHM

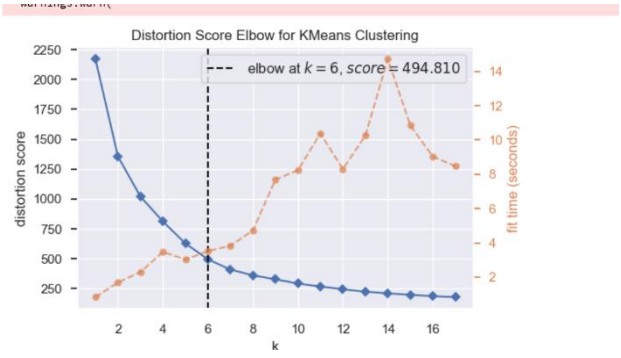
Prediction scheduling of cabs using an algorithm is a crucial part of our taxi data analysis project. The algorithm takes into account several factors to optimize the fleet operations and provide better service to customers. The algorithm is based on the results of our linear regression models, which predict the number of pickups at a given time and location.

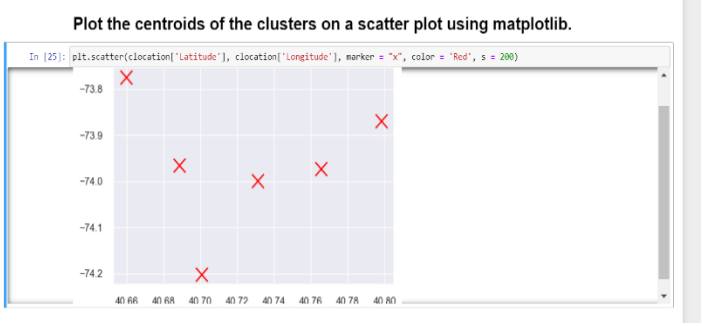
The algorithm considers the predicted number of pickups and the location of each cab in our fleet to schedule cabs more efficiently. Additionally, we take into account traffic conditions and driver availability to further optimize the scheduling process. The algorithm is designed to be scalable and can handle large volumes of data in real-time, ensuring that our operations remain efficient even during peak demand periods.

We implemented the algorithm into our dispatch system, allowing us to dispatch cabs more effectively and reduce wait times for customers. The system provides real-time updates on cab locations and availability, allowing us to make adjustments on-the-fly based on changing conditions. This helps to ensure that our customers receive the best possible service, regardless of the time of day or location.

Overall, prediction scheduling of cabs using an algorithm has allowed us to optimize our operations and improve customer satisfaction. By taking into account factors such as predicted demand, traffic conditions, and driver availability, we can schedule cabs more efficiently and provide better service to our customers.








```
In [22]: new_location = [(90.86, -75.56)]
         kmeans.predict(new_location)

         C:\python_new\lib\site-packages\sklearn\base.py:409: UserWarning: X does r
         th feature names
           warnings.warn(

Out[22]: array([5])
```

V. ACCURACY SCORE AND INERTIA VALUE

This section provides a comprehensive explanation of the accuracy and inertia values obtained from the k-means clustering analysis. Accuracy was measured using the accuracy_score() function from scikit-learn, which compares predicted cluster labels to true labels (if available) and returns a score between 0 and 1. Inertia, on the other hand, is a measure of how well the data points are clustered within their assigned clusters. The k-means algorithm aims to minimize inertia by iteratively adjusting the position of cluster centers until convergence. Inertia can also be used to determine the optimal number of clusters for a given dataset, by comparing inertia values for different numbers of clusters and selecting the "elbow point" where the rate of inertia reduction slows down significantly.

Get the inertia value for the k-means model ¶

```
In [23]: inertia = kmeans.inertia_
```

Use the k-means model to predict the clusters for the test data

```
In [24]: test_predictions = kmeans.predict(clus)

In [25]: accuracy = accuracy_score(test_predictions, kmeans.labels_)

In [26]: print("Inertia value: ", inertia)
         print("Accuracy score: ", accuracy)

         Inertia value:  494.8102276918054
         Accuracy score:  1.0
```

VI. CONCLUSION AND FUTURE WORK

In this project, we used regression and k-means clustering algorithms to analyze taxi trip data. Our results showed that regression can be a useful tool for predicting trip durations based on factors such as distance, time of day, and weather conditions. Additionally, k-means clustering allowed us to identify patterns in the data and group trips into distinct clusters based on similar characteristics.

Moving forward, there are several potential avenues for future research. One possibility is to explore other regression or clustering algorithms to improve the accuracy of our predictions and cluster assignments. For example, we could investigate the use of decision trees or neural networks for regression, or hierarchical clustering for more complex grouping of trips.

Another area for future work is to incorporate additional data sources into our analysis. For example, we could explore the impact of traffic patterns, road conditions, or events (such as concerts or festivals) on trip durations and clustering. We could also investigate ways to incorporate real-time data into our predictions, such as weather forecasts or traffic updates.

Finally, we must consider the ethical and social implications of our work. For instance, the use of clustering algorithms to group trips based on similar characteristics could have implications for privacy and discrimination. Thus, we must consider ways to ensure that our analyses do not reinforce biases or unfairly group individuals based on sensitive characteristics such as race or ethnicity. Overall, further research in these areas could lead to more accurate and socially responsible analyses of taxi trip data.

REFERENCES

[1.] Chang, H.-W.; Tai, Y.-C.; Hsu, J.Y.-J. Context-aware taxi demand hotspots prediction. Int. J. Bus. Intell. Data Min. 2010, 5, 3–18. [CrossRef]

[2.] Moreira-Matias, L.; Gama, J.; Ferreira, M.; Damas, L. A predictive model for the passenger demand on a taxi network. In Proceedings of the 2012 15th International IEEE Conference on Intelligent Transportation Systems, Anchorage, AK, USA, 16–19 September 2012; pp. 1014–1019.

[3.] Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. On Predicting the Taxi-Passenger Demand: A Real-Time Approach. In Portuguese Conference on Artificial Intelligence; Springer: Berlin/Heidelberg, Germany, 2013; pp. 54–65.

[4.] Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. Predicting Taxi–Passenger Demand Using Streaming Data. IEEE Trans. Intell. Trans. Syst. 2013, 14, 1393–1402. [CrossRef]

[5.] Zhang, K.; Feng, Z.; Chen, S.; Huang, K.; Wang, G. A Framework for Passengers Demand Prediction and Recommendation. In Proceedings of the 2016 IEEE International Conference on Services Computing (SCC), San Francisco, CA, USA, 27 June–2 July 2016; pp. 340–347.

[6.] Jagannathan, N.D.G.R.K. A Multi-Level Clustering Approach for Forecasting Taxi Trip demand. In Proceedings of the IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016; pp. 223–228.

[7.] Peng, X.; Pan, Y.; Luo, J. Predicting high taxi demand regions using social media check-ins. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 2066–2075.

[8.] Zhao, K.; Khryashchev, D.; Freire, J.; Silva, C.; Vo, H. Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 833–842.

[9.] Xu, J.; Rahmatizadeh, R.; Boloni, L.; Turgut, D. A Sequence Learning Model with Recurrent Neural Networks for Taxi Demand Prediction. In Proceedings of the 2017 IEEE 42nd Conference on Local Computer Networks (LCN), Singapore, 9-12 October 2017; pp. 261–268.

[10.] Zhang, D.; He, T.; Lin, S.; Munir, S.; Stankovic, J.A. Taxi-Passenger-Demand ModelingBased on Big Data from a Roving Sensor Network. IEEE Trans. Big Data 2017, 3, 362–374. [CrossRef]

[11.] Bao, Y.; Sun, Y.-E.; Bu, X.; Du, Y.; Wu, X.; Huang, H.; Luo, Y.; Huang, L. How Do Metro Station Crowd Flows Influence the Taxi Demand Based on Deep Spatial-Temporal Network? In Proceedings of the 2018 14th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), Shenyang, China, 6–8 December 2018; pp. 188–192.

[12.] Davis, N.; Raina, G.; Jagannathan, K. Taxi Demand Forecasting: A HEDGE-Based Tessellation Strategy for Improved Accuracy. IEEE Trans. Intell. Transp. Syst. 2018, 19, 3686–3697. [CrossRef]

[13.] Markou, I.; Rodrigues, F.; Pereira, F.C. Real-Time Taxi Demand Prediction using data from the web. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 1664–1671. Appl. Sci. 2020, 10, 6681 17 of 18

[14.] Ishiguro, S.; Kawasaki, S.; Fukazawa, Y. Taxi Demand Forecast Using Real-Time Population Generated from Cellular Networks. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers—UbiComp ’18, Singapore, 8–12 October 2018; pp. 1024–1032.

[15.] Liao, S.; Zhou, L.; Di, X.; Yuan, B.; Xiong, J. Large-scale short-term urban taxi demand forecasting using deep learning. In Proceedings of the 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), Jeju, Korea, 22–25 January 2018; pp. 428–433