iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71576

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Exploratory Data Analytics on Uber Transportation Patterns

Shanthini D
Department of MCA, Paavai Engineering College (Autonomous)

Abstract: ***This paper presents a comprehensive analysis of Uber trip data in New York City using data science techniques to improve ride-hailing efficiency. The project utilizes historical trip data from the NYC High Volume For-Hire Vehicle (HVFHV) dataset, incorporating clustering (K-Means) and prediction (Random Forest Regression) to forecast demand patterns. It also integrates Power BI for real-time visualization and insights. The aim is to optimize driver allocation, reduce passenger wait times, and enhance urban mobility.***
Keywords: Uber, Data Analysis, K-Means Clustering, Random Forest, Power BI, Ride-hailing Optimization

I. INTRODUCTION

Urban mobility faces increasing challenges in matching ride-hailing demand with efficient driver allocation. This study focuses on analyzing Uber trip data to understand spatial and temporal demand patterns, enabling better decision-making through data science and visualization.

II. METHODOLOGY

A. Data Collection and Preprocessing

Trip data in Parquet format was collected from the NYC TLC HVFHV dataset, merged with taxi zone and weather data. The dataset underwent preprocessing to clean nulls, convert types, and extract features.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) employed visualization tools like Matplotlib and Seaborn to uncover patterns such as peak travel hours, demand across zones, and seasonal variations.

C. Clustering with K-Means

K-Means was applied on features like pickup/drop zones and trip metrics to group locations with similar demand profiles.

D. Prediction using Random Forest Regression

Random Forest predicted trip volume and estimated wait times. Feature engineering and performance validation using R^2 score ensured accuracy (~84%).

E. Visualization with Power BI

Interactive dashboards showcased daily trends, zone distribution, and seasonal demand, enabling decision-makers to act on insights.

III. RESULTS AND DISCUSSION

The clustering identified high-demand pickup zones, while regression modeling yielded a high R^2 score, confirming strong predictive capability. Dashboards revealed consistent peak demand in zones like Midtown and Downtown during weekdays and evenings. Recommendations include real-time adjustments to vehicle distribution and dynamic pricing.

IV. CONCLUSION

The project successfully applied data analytics and machine learning to Uber trip data for proactive fleet management. By integrating visualization and prediction, it supports smarter, faster, and data-driven urban mobility solutions.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

TABLE I
TEST CASE REPORT

TEST CASE ID	TEST CASE TITLE	DESCRIPTION	EXPECTED RESULT	ACTUAL RESULT	STATUS
TC_01	LOAD PARQUET TRIP DATA	VERIFY CORRECT LOADING OF TRIP DATA	DATA LOADS WITH REQUIRED COLUMNS	AS EXPECTED	PASS
TC_02	MERGE WEATHER DATA	ENSURE WEATHER DATA INTEGRATES WITH TRIP RECORDS	WEATHER ALIGNED WITH TRIP DATES	AS EXPECTED	PASS
TC_03	APPLY K-MEANS CLUSTERING	CLUSTER TRIPS BY LOCATION AND TIME	CLUSTERS FORMED SHOWING DEMAND HOTSPOTS	AS EXPECTED	PASS
TC_04	TRAIN RANDOM FOREST MODEL	PREDICT TRIP TIME AND EXPORT RESULTS	HIGH R^2 SCORE AND EXCEL EXPORT	AS EXPECTED	PASS
TC_05	VISUALIZE TRENDS IN POWER BI	DISPLAY DAILY/MONTHLY TRENDS CORRECTLY	PROPER DATE FORMAT AND VISUALS	AS EXPECTED	PASS
TC_06	DEMAND ZONE CLUSTERING	IDENTIFY HIGH-DEMAND ZONES USING CLUSTERING	VISUAL DISPLAY OF TOP ZONES	AS EXPECTED	PASS



Fig .1 PowerBICharts

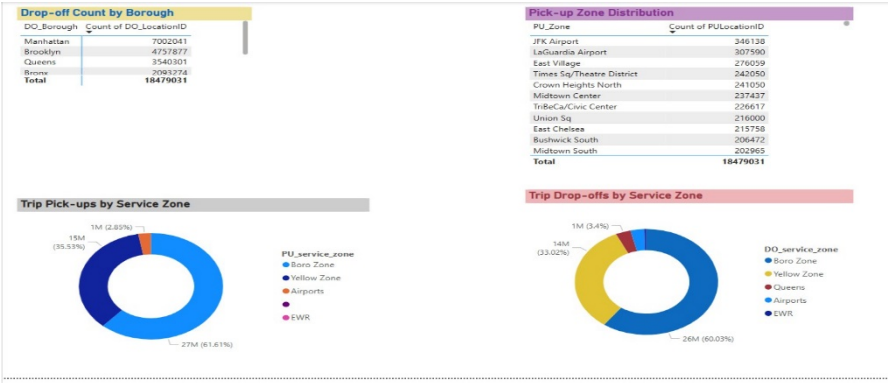
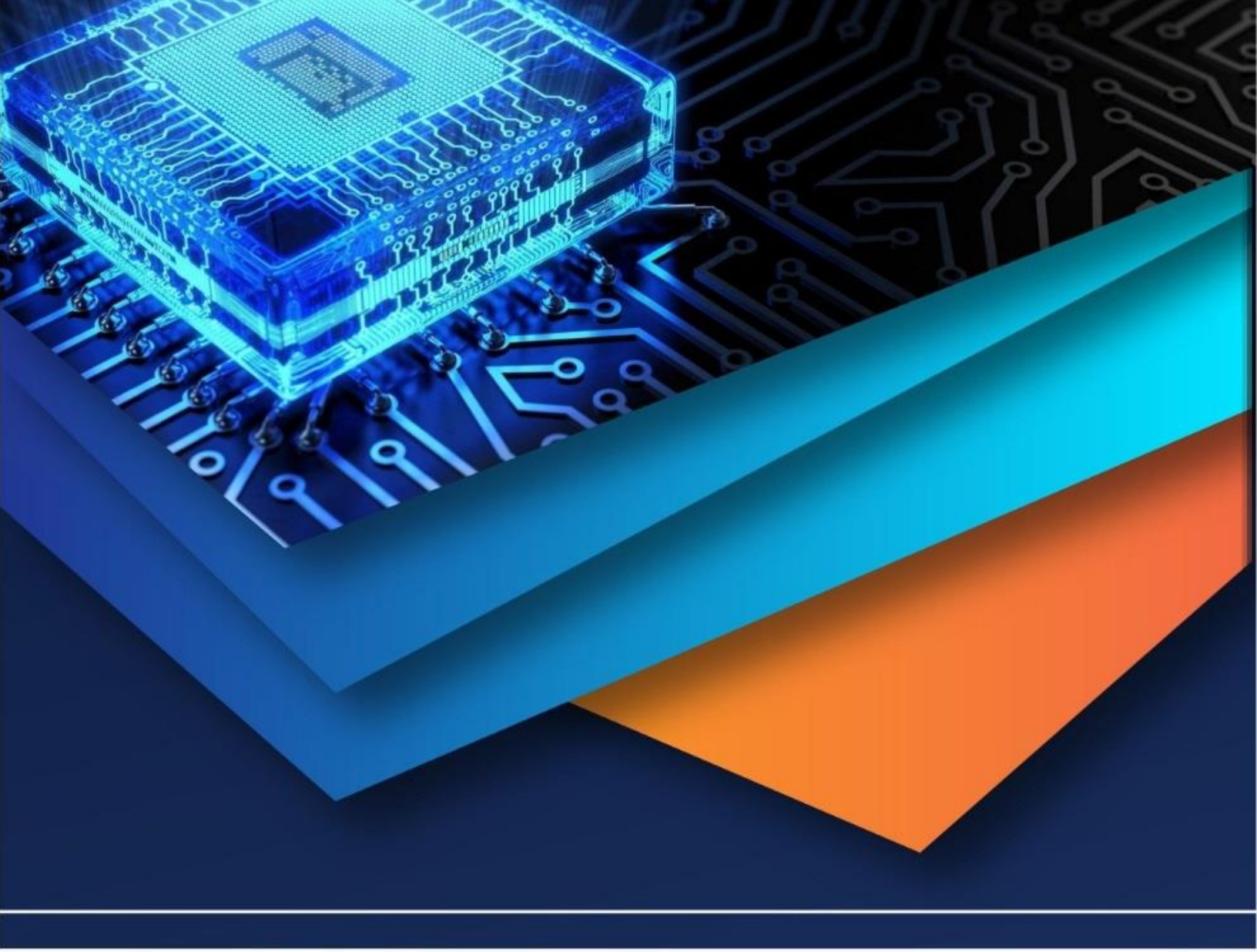


Fig.2ZonebasedVisualization

REFERENCES

[1] Poritigadda, L., et al. (2024). Spatial Data Analysis on On-Demand Cab Services Using Spark.
[2] Golshanrad, P., et al. (2024). Proposing a model for predicting passenger origin–destination in online taxi-hailing systems. Public Transport.
[3] Kokkiligadda, M. R., et al. (2023). Spatial Data Analysis on On-Demand Cab Services using Spark. IEEE ICIMI.
[4] Roy, B., & Rout, D. (2021). Predicting Taxi Travel Time Using ML Techniques. Springer.
[5] Pradhan, R., et al. (2021). Analysing Uber Trips using PySpark. IOP Conf. Ser.: Mater. Sci. Eng.
[6] Wang, H., et al. (2021). Applying deep learning to taxi demand forecasting: CNN-LSTM model. Transp. Res. Part C.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)