IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Decoding Cab Dynamics: A Comprehensive Analysis Of Ride-Hailing Data

Ashish Kumawat[1], Ratnesh Litoriya[2], Vansh Chourasia[3], Vanshika Yadav[4], Vishal Dangi[5]

[1]Assistant Professor, Department of CSE, Medi-Caps University, Indore, MP, India

[2] Professor & HOD, Department of CSE, Medi-Caps University, Indore, MP, India

[3] Student, Department of CSE, Medi-Caps University, Indore, MP, India

[4] Student, Department of CSE, Medi-Caps University, Indore, MP, India

[5] Student, Department of CSE, Medi-Caps University, Indore, MP, India

Abstract : Businesses have long relied on data analytics in ways to improve operational efficiency. To fully benefit from these advantages, a thorough comprehension of the subject is required. The study analyses taxi travel data in a city using exploratory data analysis and advanced statistical approaches. It aims to identify patterns, trends, and insights in ride-hailing data, which could be useful for urban mobility planning, optimizing taxi service operations, and understanding user behaviour. The project uses Mage for ETL operations, analysing ride-hailing data, and storing it in Google BigQuery for efficient management and scalability, with SQL queries for trip details and passenger counts. The study uses Looker Studio to transform analysed data into interactive dashboards, enabling stakeholders to discover trends and improve services across platforms, integrating ETL techniques, cloud data warehousing, SQL analysis, and visualization tools.

Keywords – Data Analysis, Data Visualization, ETL, Mage, Cloud Storage, GCP, BigQuery, SQL, Looker

I. INTRODUCTION

The business strategy of many ride-hailing services focuses on connecting supply and demand, transforming transport services. Analyzing massive amounts of data is crucial for decision-making in various industries, including the transportation sector. Ride-hailing services can benefit greatly from the analysis of massive datasets, which offer significant information for operational efficiency, customer experiences, and strategic decision-making. A project analyzing taxi trip data can apply various analytical techniques and tools, such as SQL analysis to query and extract insights, visualisation tools to display findings, and Extract, Transform, and Load (ETL) methods to prepare and transform raw data for analysis. Combining these methodologies allows organisations to gain a greater understanding of ride-hailing data, uncover relevant insights, and optimize their operations in the transportation sector.

The project aims to develop a robust analytical framework using ETL methods, utilising the Mage tool to efficiently handle trip records. This involves converting raw data into a structured format and loading it into Google BigQuery. This infrastructure is crucial for managing and analysing vast trip data, enabling in-depth research. SQL analysis is employed to investigate popular routes, rider behaviour, and patterns in cab trip data. This data is crucial for improving service quality, optimizing operations, and informing strategic decision-making within the service ecosystem. The project aims to extract valuable insights from trip data, resulting in improvements across all the ride-hailing operations.

The project utilizes SQL analysis, ETL, and data visualization tools to extract meaningful insights from massive datasets like Cab Travels. Looker Studio is used to convert the analysed data into user-friendly dashboards and dynamic displays. This comprehensive approach to transportation data analysis aims to improve operational efficiencies and user experiences on ride-hailing platforms by pushing the boundaries of data-driven methodologies. The visualizations offer a thorough overview of key performance factors, trends, and patterns, enabling stakeholders to make informed decisions.

II. LITERATURE REVIEW

2.1 Mrunal Patil, Vidya Kumari, Adarsh Patil, Laxmikant Ahire, and Assistant Professor Umakant Mandawkar undertook a study in 2021 to analyse an Uber dataset in New York City utilising data analytics and visualisation approaches. They used the R language and utilities such as tidy,lubricate, dplyr, and ggplot2 to process and visualise the data. The study aimed to understand how time and location affect Uber customer visits in New York City. Visualisations, including a Geo plot, were used to show patterns of visits from different regions. The findings could reveal insights into Uber usage dynamics, peak hours, popular pickup/drop-off areas, and potential trends. This study showcases the use of data analytics and visualization tools to understand real-world phenomena, particularly Uber customer behaviour in metropolitan areas like New York City.

2.2 In 2019**, P. Devika, Y. Prasanna, P. Swetha, and G. Akhilesh Babu** suggested a MapReduce-based system for analysing data on journeys and active cars in basements. The solution includes data transfer from local sources to a Hadoop cluster for processing, as well as a MapReduce programme that identifies the days with the most journeys and active vehicles in each basement. The program runs on the Hadoop cluster, processing data in parallel across numerous nodes while exploiting Hadoop's distributed computing capabilities. After the job is finished, the results are pulled from the Hadoop cluster, including information about the busiest days and active vehicles in each basement. The programme is compiled into a JAR file, containing its logic and dependencies.

2.3 Yog Patil, Aryan Raskar, Sonal Singh, Ayush Shukla, and Prof. Rajendra Pawar (2023) presented a project on Uber Data Analysis, utilizing regression models and random forest algorithms to create an accurate fare prediction model and geospatial analysis.

2.4 Abel Brodeurand and Kerry Nield's 2018 study examines the impact of rain on Uber rides in New York City after 2011. They found that despite a 5% increase in taxi rides per hour, Uber journeys increased by 22% due to increased rainfall.

2.5 In their 2019 presentation**, A Bharathi and S Surya Prakash** used data analytics to analyse Uber's transaction dataset and projected potential consequences and changes. They visualised the results with heat maps and histograms, which provided a straightforward depiction of the dataset.

2.6 MK Saravana and K Harish (2017) utilised Hadoop to analyse Uber data for Business Intelligence, substituting traditional RDBMS such as MySQL ,Libre Office and Oracle .

III.METHODOLOGY

3.1 Cloud Storage

The project begins with storing raw data in a bucket instance using Google Cloud Platform's Cloud Storage service (GCP Cloud Storage). This phase involves uploading the cab trip data files, which may come in various formats such as CSV or JSON, to a specific bucket under Cloud Storage.

Cloud Storage offers a secure, long-lasting, and scalable method for storing massive amounts of data in the cloud. By utilising Cloud Storage, the project ensures that raw data is easily available and securely saved, providing the groundwork for subsequent data processing and analysis activities.

Once the raw data is securely retained in the GCP Cloud Storage bucket, it can be accessed and processed using ETL techniques, converted to a structured format, and loaded into Google BigQuery for further analysis.

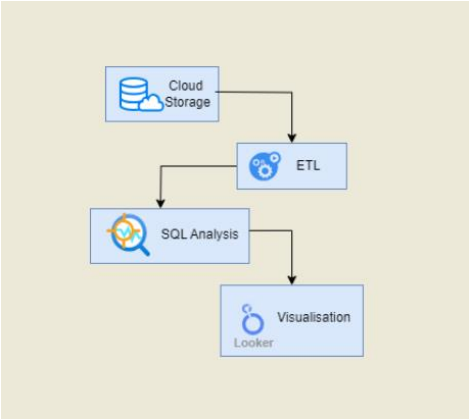

fig 1.flowchart

3.2 Data Extraction and Pre-processing

After storing the raw cab trip data in a GCP Cloud Storage bucket, the project extracts it using the Mage tool. Mage simplifies the process by efficiently gathering data from many sources and preparing it for future processing. After extraction, the raw data goes through a thorough transformation process to ensure consistency and cleanliness. This involves meticulously finding and correcting any anomalies or discrepancies found in the data. Cleaning, standardisation, enrichment, and deduplication are some of the procedures utilised to standardise and improve the format and quality of data, making it appropriate for analysis.

The cleansed and standardised data is then transferred to Google BigQuery for storage and management. Google BigQuery, a fully managed data warehouse service, offers a scalable platform for efficiently storing, searching, and analysing massive datasets. By exploiting its features, the project ensures that data is safely kept and easily accessible for study.

The use of the Mage tool for extraction and data transformation techniques assures that the data saved in Google BigQuery is high-quality and ready for analysis. This lays the groundwork for SQL analysis, visualisation, and gaining important insights from cab trip data. Stakeholders can receive meaningful information from these analytical procedures, enabling them to make better judgements and enhance transportation operations.

3.3 SQL Data Analysis

After the data is securely stored in Google BigQuery, SQL analytic techniques are used to extract useful insights from the ride-hailing data. SQL queries are written to investigate various aspects of the dataset, allowing analysts to uncover patterns and trends in passenger behaviour, identify high-traffic hours, find popular routes, and evaluate driver performance metrics. These queries are intended to achieve certain analytical goals and extract useful information from a dataset.

For instance, SQL queries can be used to:

3.3.1 Analyse Passenger Behaviours: By querying the dataset, analysts can uncover common passenger behaviours such as preferred pickup and drop-off locations, average trip durations, and ride frequency at various times of day or week.

3.3.2 Discover High Traffic Hours: SQL queries can be used to aggregate trip data based on timestamps, allowing analysts to discover peak activity times and understand demand patterns throughout the day or week.

3.3.3 Explore Popular Routes: Using SQL queries, analysts can analyse trip data to find commonly travelled routes, popular destinations, and common travel patterns in the dataset. These SQL analysis approaches provide useful insights into many elements of cab trip data, allowing stakeholders to make more informed decisions, optimise operations, and improve the overall user experience within the transportation ecosystem. Iterative querying and analysis can reveal actionable insights that drive strategic decision-making and increase the efficiency and efficacy of ride-hailing services.

3.4 Data Visualisation

Looker Studio is a platform for creating interactive dashboards and visualisation. These dashboards give the stakeholders a comprehensive picture of important performance indicators, trends and patterns discovered by SQL analysis.

Looker Studio provides us with choices from a variety of visualization options such as charts, graphs and maps to effectively represent the data. These visualisations are designed to be customised to display relevant dimensions, metrics and filters based on the SQL analysis results.

After our visualisations are created we organize their layout logically and intuitively, add annotations to highlight key findings and publish and share dashboards with the stakeholders.

IV. ANALYSIS AND DISCUSSION

In the execution of this project to understand customer behaviour and different trends we performed SQL analysis on our dataset. We utilized SQL to perform comprehensive analysis and to develop and execute SQL queries to investigate trends and patterns.

4.1 Total Number of Trips per Day:

Considering our dataset when we performed SQL analysis for a given period we came across several results and from the result of our analysis we were able to conclude that the average number of journeys per day was approximately 76780. By knowing the average number of trips a day we can utilize this data in deciding what number of vehicles we would require on average to fulfil the task of completing these number of trips without any hassle.

4.2 Total Fare Amount by Payment Type:

In transportation services like taxi or ride sharing customers use various modes of payment whether it be cash or online or through credit cards. While performing analysis on our data we came to recognize that a fairly large amount of people prefer to use credit cards as their mode of payment.

4.3 Number of Passengers Picked Up at Each Location per Hour:

In Ride sharing services people opt for these services at various locations such as street corners, designated addresses or some other location. Analyzing the number of passengers picked up at each location per hour provides valuable insights into the demand patterns for transportation services throughout the day and across different locations.

By tracking passenger pickups by location and hour, transportation companies can forecast demand trends and allocate resources more efficiently, such as deploying more vehicles during peak demand periods or in areas with high pickup activity.

4.4 Most Popular Pickup Locations:

There are various locations in a town that are crowded or most popular and we can say these places can be the hot spot for the maximum of the pickup or drop location for any taxi services. Analyzing these popular pickup locations provides valuable insights into passenger behaviour, travel patterns, and demand distribution across different areas.

4.5 Most Frequent Travel Times:

It refers to the peak hours a transportation service experiences the most amount of bookings. It can be during the morning when people leave for work or any specific time. It indicates when the transportation system is the busiest and in need of great services. By identifying peak travel times, transportation companies can allocate their resources more efficiently, such as scheduling more vehicles or drivers during these high-demand periods to reduce passenger wait times and meet increased demand.

4.6 Average Fare Amount by Rate Code:

In the transportation system, the fare amount may change depending on different rate codes which can be any special circumstances during a specific time of the day or due to weather conditions etc. Analyzing the average fare amount by rate code provides valuable insights into the pricing structure and revenue generated from different types of trips.

4.7 Total Number of Trips by Payment Type:

Analysing the total number of trips according to the mode of payment helps transportation companies optimize their operations, enhance customer experience, and develop strategies to attract and retain passengers based on their preferred payment methods. As for our earlier analysis, we were able to tell that a large number of customers like to use credit cards as their mode of payment, and hence we can conclude here that the maximum number of trips taken by the payment type was the one where the mode of payment was through credit card and then cash.

4.8 Average Number of Passengers per Trip :

After analysing the average number of passengers per trip we were able to conclude that people mostly like to travel with a partner rather than travelling alone. Analysing the number of passengers per trip helps in maintaining an optimal average number of passengers per trip. It contributes to a better passenger experience by reducing overcrowding, improving comfort levels, and minimizing wait times for passengers.

V. RESULT

The objective of this analysis was to gain insight into the usage of different patterns, revenue streams, and the operational efficiency of the taxi transportation services using Data Visualization, GCP, and Big Query. We were able to identify different patterns and decode various insights with our analysis.

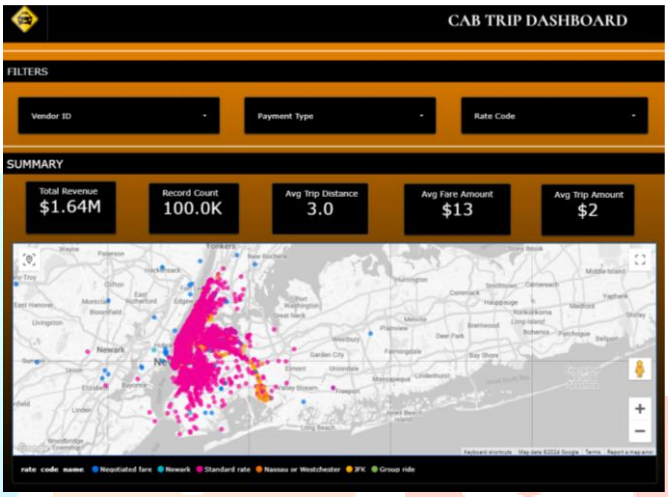

fig 2. Dashboard 1

Based on the results we obtained from the SQL analysis we were able to create this dashboard using the Looker Studio. The visualisation shows us filters like the Vendor ID, Payment type and Rate Code we can apply these filters to obtain selected insights. Next are the KPIs (Key performance indicators) i.e. total payment, record count, Avg trip distance, average fare Amount, and trip Amount that change based on the filters we use. After identifying the patterns, relationships and trends performed a Geospatial analysis. Geospatial analysis is a technique for analysing, interpreting, and visualising data with a geographic or spatial component.

We plotted a geospatial map using rate codes such as negotiated fare, standard rate, group ride etc.

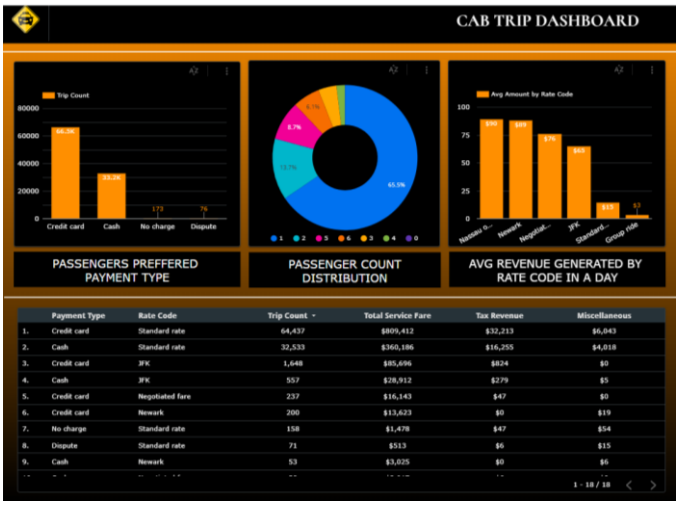

fig 3. Dashboard 2

In our second visualisation, we generated an OLAP Table, OLAP is used in business intelligence (BI), decision support, and a variety of business forecasting and reporting applications. The first table tells us about the payment method preferred by the passengers and we found that a maximum of customers preferred credit cards as the payment method that were around 66.5k and 33.2k preferred cash as the payment method. The next chart tells us the passenger count distribution 66.5% of customers travelled alone,13.7% of customers travelled in couples and the remaining in groups. The last chart gives us an insight into the average revenue generated by rate code in a day. What factors affected the revenue the most and what factors helped us make the maximum revenue?

VI. CONCLUSION

Storing and processing data is a demanding work; with increasing data, we don't have appropriate solutions to store and handle this data. In our project, we aim to highlight the importance of data analytics for enhancing the efficiency of businesses related to ride-hailing services. The main aim was to improve mobility planning optimise taxi operations and understand their behaviour.

The key tools we used in this project included Mage for ETL operations, Google Big Query for data storage, SQL queries for trip details and data analysis, and Looker Studio for transforming the analyzed data into interactive dashboards. These tools helped us to discover various trends and improve services across platforms, integrating ETL techniques, and helping with cloud warehousing, SQL analysis and visualisation.

Further, it emphasized the importance of data analysis in ride-hailing services, developing an analytical framework using ETL methods, Mage Tool, Google BigQuery and SQL analysis. The project aims to enhance the operational efficiency of ride-hailing platforms and user experience by leveraging data-driven methodologies and providing stakeholders with informed decision-making capabilities through comprehensive data visualisation.

VII. FUTURE SCOPE

7.1 Data Integration

Integrating additional datasets can help provide valuable context and enhance accuracy and performance. Integrating weather data such as temperature, humidity, and precipitation helps identify weather-related patterns and their impact on transportation services. By incorporating event data, traffic data, holiday and seasonal data, transportation companies can acquire a more complete picture of the elements that drive taxi demand and other operations.

7.2 Real-Time Monitoring

In our research, we made use of already available data but if we incorporate real-time monitoring we'll be able to make predictions in real-time, real time monitoring includes continuously collecting, analysing, processing and visualizing data of the taxi trips as they occur. It can help companies to make informed decisions and respond promptly to changing conditions.

7.3 Advanced analytics using ML, AI models

AI and ML models can provide deeper insights and predictive capabilities when analysing taxi-driven data. AI/ML models can predict future cab demand based on factors like time, weather and events and also historical data. Adjusting fare in real-time based on the demand, traffic conditions and other variables to maximize revenues.

VIII. CONFLICT OF INTERESTS/COMPETING INTERESTS

As a researcher, I declare that there are no conflicts of interest regarding the findings presented in this study and no funds were raised.

IX. DATA AVAILABILITY

The data that support the findings of this study are available on -

1. https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf
2. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

X. REFERENCES

[1] Mrunal Patil1 , Vidya Kumari2 , Adarsh Patil3 , Laxmikant Ahire4 and Asst.Prof. Umakant Mandawkar5 1,2,3,4 B.Tech, Computer Science and Engineering, Sandip University, Nashik, India 5 Asst Prof., Computer Science and Engineering, Sandip University, Nashik, India, UBER DATA ANALYSIS USING GGPLOT, JES Vol 12, Issue 7, July/2021 ISSN NO:0377-9254

[2] P. Devika, Y. Prasanna , P. Swetha ,G. Akhilesh Babu, Uber Data Analysis using Map Reduce, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019

[3] Yog Patil, Aryan Raskar, Sonal Singh, Ayush Shukla, Prof. Rajendra Pawar, Dept. of CSE, MIT-WPU, Maharashtra, India, Uber Data Analysis, International Journal of Scientific Research & Engineering, Trends, Volume 9, Issue 3, May-Jun-2023, ISSN (Online): 2395-566X

[4] Abel Brodeur and Kerry Nield, An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC, Journal of Economic Behavior & Organization, 2018, vol. 152, issue C, 1-16

[5] A Bharathi, S Surya Prakash, Bannari Sathyamangalam, India, an approach to predict taxi-passenger demand using the quantitative histogram on Uber data, 2019 International Conference on Advances in Computing and Communication Engineering(ICACCE), IEEE,10.1109/ICACCE46606.2019.9079980

[6] MK Saravana, K Harish, Dept. of Computer Science & Engineering, Jyothy 1Institute of Technology, Bengaluru, India, a case study on analyzing Uber datasets using Hadoop framework IEEE, 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 10.1109/ICECDS.2017.838966

[7] Vsevolod Salnikov, 1Renaud Lambiotte, 1Anastasios Noulas, 2 and Cecilia Mascolo21naXys, University of Namur, Belgium2ComputerLab, University of Cambridge, UK, OpenStreetCab: Exploiting Taxi Mobility Patterns in New York City to Reduce Commuter Costs, arXiv:1503.03021v1[cs.SI]2015

[8] https://docs.mage.ai/guides/load-api-data

[9]https://www.techrepublic.com/article/google-data-studio-guide/

[10] https://medium.com/@muhammadaris10/nyc-taxi-trip-data-analysis-45ecfdcb6f91

[11] https://pandas.pydata.org/docs/

[12]https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

[13]https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page (March, 2016)