Northeastern University
College *of* Engineering

Data Warehousing &
Integration IE 6750
FALL 2024

NYC TLC Analysis

Report

Group 13

Dheeraj Kumar Goli
Saathvika Kethineni

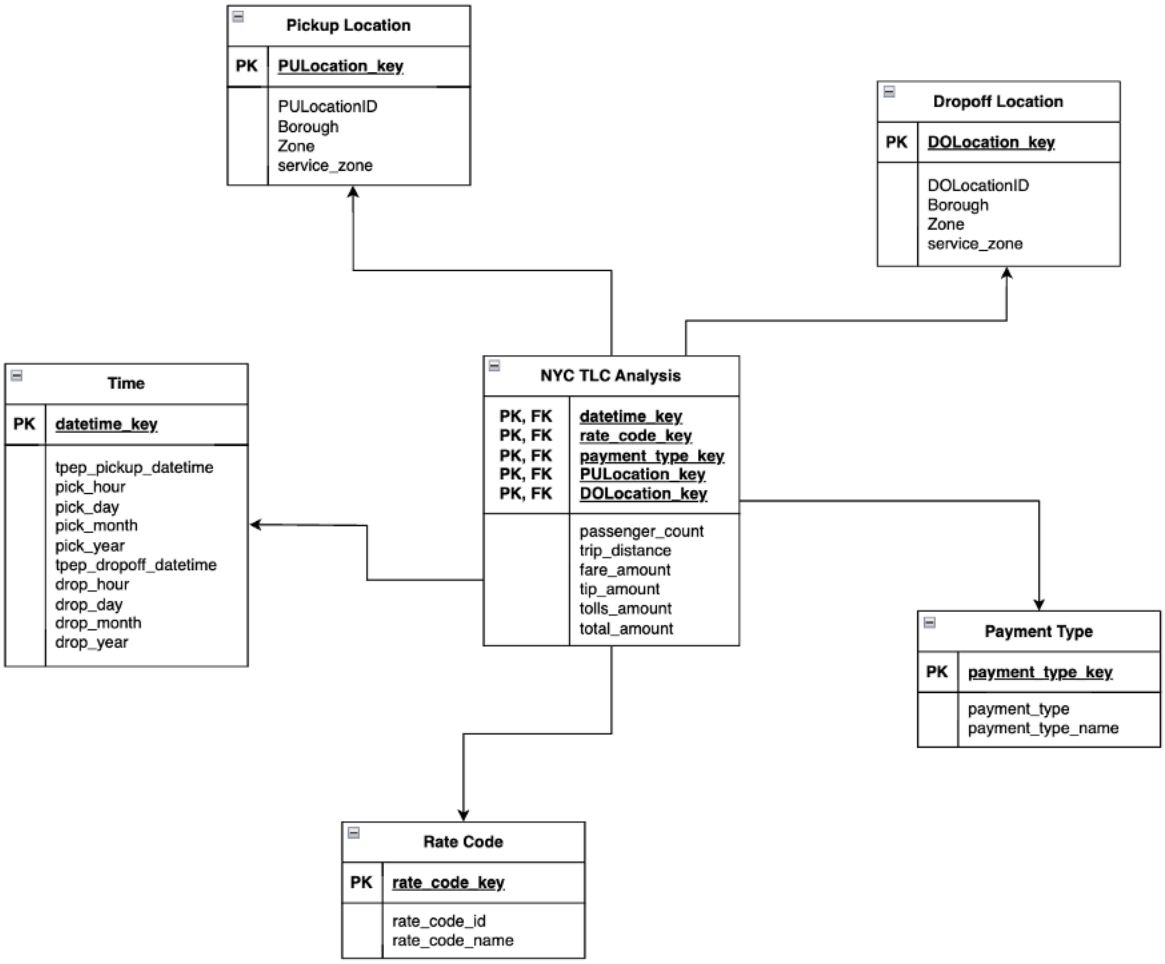
goli.dh@northeastern.edu
kethineni.s@northeastern.edu

Submission Date: 12/07/24

Problem Statement:

This project aims to analyze New York City taxi trip data to derive actionable insights into urban mobility trends and optimize transportation services. Leveraging the comprehensive trip data provided by the NYC Taxi and Limousine Commission (TLC), the study focuses on key metrics such as trip distances, fare amounts, pickup and drop-off hotspots, passenger counts, payment methods, and hourly revenue trends. The goal is to uncover patterns in taxi demand across different boroughs, identify major trip hotspots, and evaluate how factors such as time, location, and payment preferences influence travel behavior and pricing. By answering these questions, this analysis aims to support stakeholders in making data-driven decisions for improving taxi operations, enhancing customer experience, and addressing urban transportation challenges.

LOGICAL MODEL:



FILES & DATA SOURCES:

Files:

1. yellow_tripdata_2022-01.parquet
2. Taxi_zone_lookup.csv

Data Sources:

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

PIPELINE TOOLS:

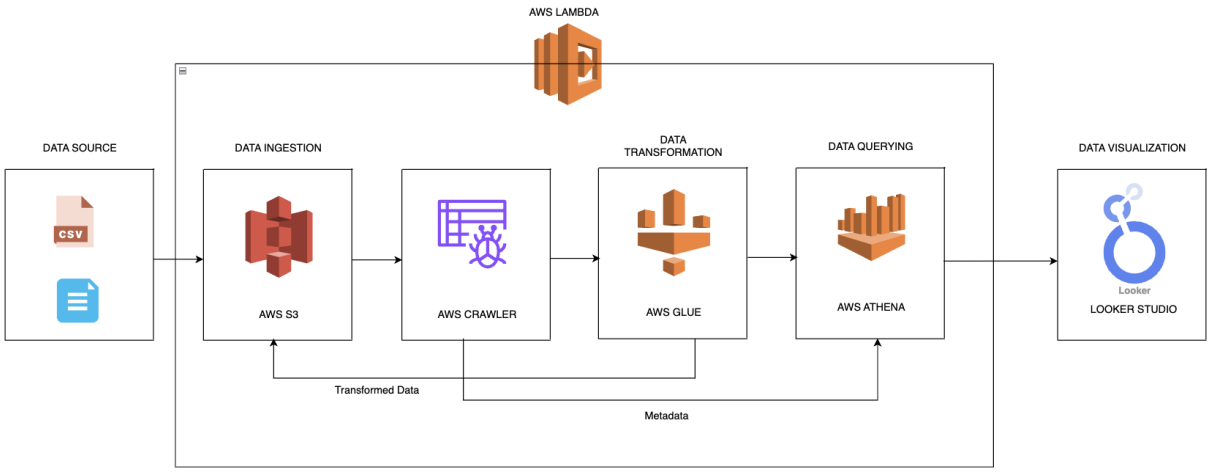
1. S3
2. Athena
3. Glue
4. Lamda

PROGRAMMING CHOICE:

Python was the primary programming language used for:

- Writing Glue jobs for ETL processes.
- Implementing the Lambda function for automation and integration.
- Transforming raw data into dimensions and fact tables.

DOCUMENTATION:



1. Data Source

- Input Files:
    - yellow_tripdata.parquet: Contains raw transactional data of taxi trips, including timestamps, pickup/dropoff locations, fare amounts, and payment details.
    - taxi_zone_lookup.csv: A reference file providing metadata for taxi zones, including zone names, boroughs, and service zones.
- Purpose:
    - These files serve as the raw data source for the ETL pipeline.
    - They are critical for creating dimensions (e.g., location, time) and the fact table used for analytical queries.

2. Data Ingestion

- AWS S3:
    - Raw data files are uploaded to an S3 bucket (dwcloudbucket).
    - S3 acts as the OLTP storage layer, where unstructured and structured data are stored for processing.
    - S3 ensures durability and scalability for storing large datasets.
- Key Benefits:
    - Centralized storage accessible to other AWS services like Glue, Lambda, and Athena.
    - Supports event-driven workflows triggered by new uploads.

3. Schema Discovery

- AWS Glue Crawler:
    - A Glue Crawler scans the raw data in the S3 bucket and automatically detects the schema of the files.
    - Outputs metadata tables in AWS Glue Data Catalog, which integrates seamlessly with Athena for querying.
    - Example Outputs:
        - A table for yellow_tripdata.parquet with fields such as pickup_datetime, dropoff_datetime, and fare_amount.
        - A table for taxi_zone_lookup.csv with fields such as LocationID, Borough, and Zone.

4. Data Transformation

- AWS Glue:
    - Glue jobs written in Python perform ETL operations to transform raw data into an OLAP-ready format.
    - Steps involved:
        1. Dimension Creation:

        - Pickup Location Dimension: Extracts information about pickup locations, including boroughs and zones.
        - Dropoff Location Dimension: Similar to the pickup dimension but for dropoff locations.
        - Rate Code Dimension: Maps rate codes to descriptive labels (e.g., Standard Rate, JFK, Newark).
        - Payment Type Dimension: Maps payment type IDs to names (e.g., Credit Card, Cash).
        - Time Dimension: Extracts granular time details (e.g., hour, day, month) from timestamps.
    2. Fact Table Creation:
        - Combines data from dimensions and includes metrics like trip distance, passenger count, fare amount, and tips.
    3. Output:
        - Transformed data is saved back to a specific prefix in the S3 bucket (e.g., transformed_final/).

5. Data Querying

- AWS Athena:
    - Athena is used to query the transformed data stored in the S3 bucket.
    - Tables representing dimensions and the fact table are queried for analytical insights.
    - Example Queries:
        - Find the top 10 pickup locations with the highest trip counts.
        - Analyze fare and tip amounts by borough and zone.
    - Benefits:
        - Serverless query execution with a pay-per-query model.
        - Efficient OLAP-style querying over large datasets.

6. Pipeline Automation

- AWS Lambda:
    - Lambda orchestrates the ETL pipeline by responding to new file uploads in the S3 bucket.
    - Functions:
        - Trigger Glue Job:
            - When a new file is uploaded, Lambda starts the Glue job to process and transform the data.
        - Execute Athena Queries:
            - After the Glue job completes, Lambda executes predefined Athena queries to analyze the transformed data.
        - Monitor Workflow:
            - Tracks the status of Glue jobs and Athena queries, ensuring successful completion or handling errors.

- Example Use Case:
  - A new yellow_tripdata_2022-02.parquet file is uploaded.
  - Lambda detects the file, triggers the Glue job, and executes queries to analyze February's data.

7. Data Visualization

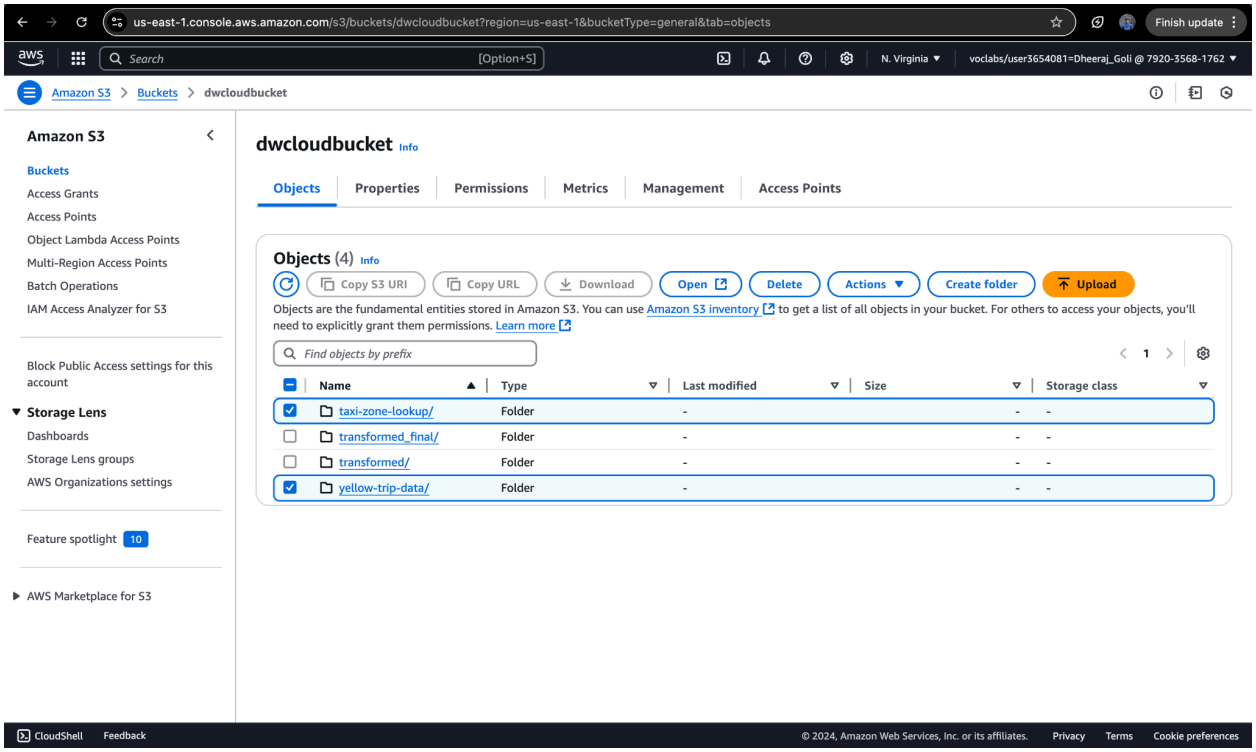
- **Looker Studio (formerly Google Data Studio)**:
  The final stage of the pipeline involves data visualization and reporting. Looker Studio is integrated with AWS Athena to create insightful dashboards and visualizations. Features include:
  - **Dynamic Dashboards**: Real-time visuals such as charts, graphs, and tables that update based on query results.
  - **Customizable Reports**: Users can build tailored reports to suit specific business requirements.
  - **Collaboration**: Reports can be shared and accessed by multiple stakeholders for decision-making.

Looker Studio provides a user-friendly interface for business users to derive insights from complex datasets.

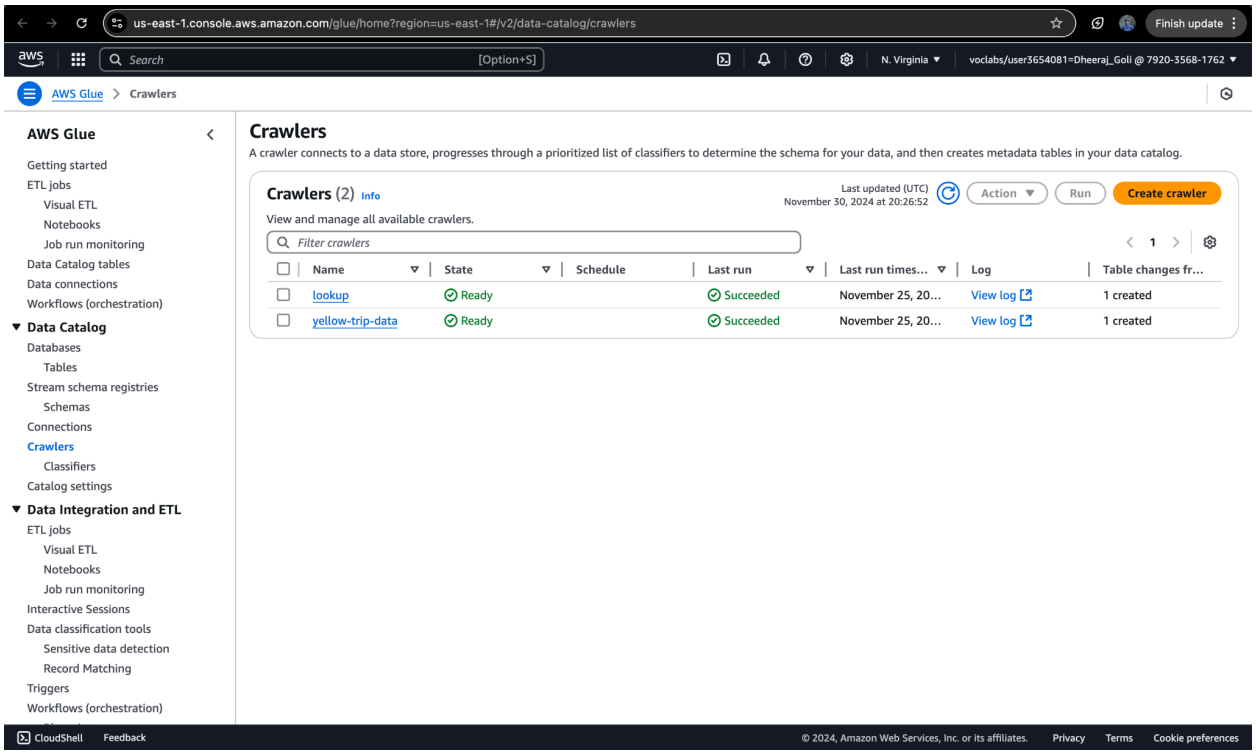
PIPELINE EXECUTION:

Workflow:

1. Data Upload:
   - yellow_tripdata.parquet and taxi_zone_lookup.csv were uploaded to the S3 bucket.

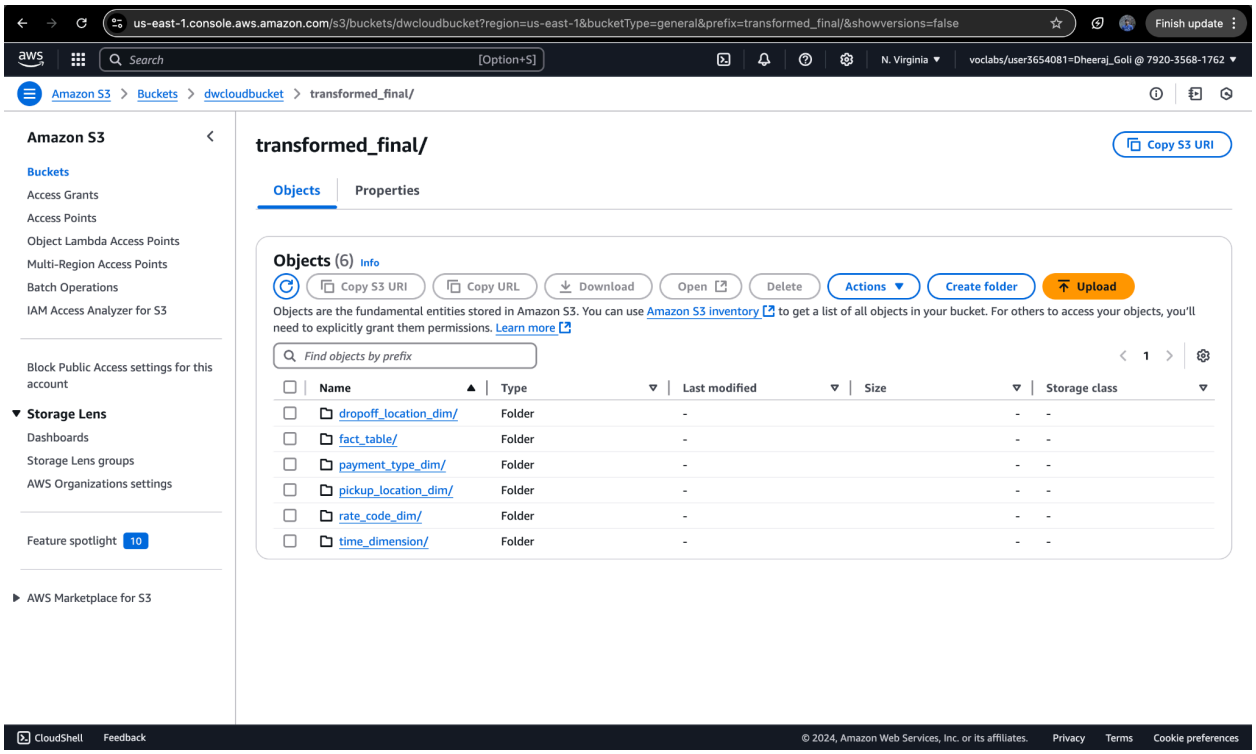


2. Data Crawling:
    - AWS Glue Crawler created tables in Athena based on the S3 data.



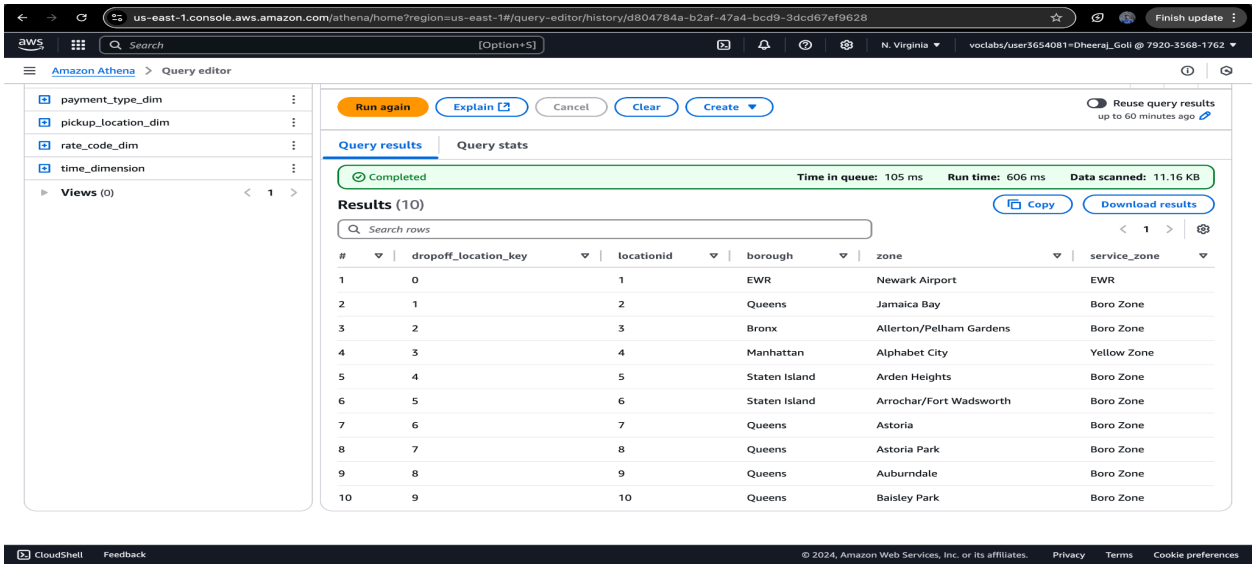
3. ETL Process:
    - AWS Glue job processed the raw data to generate:
        - Dimensions: pickup_location_dim, dropoff_location_dim, rate_code_dim, payment_type_dim, and time_dimension.
        - Fact Table: fact_table.

    - Transformed data was saved back to S3.

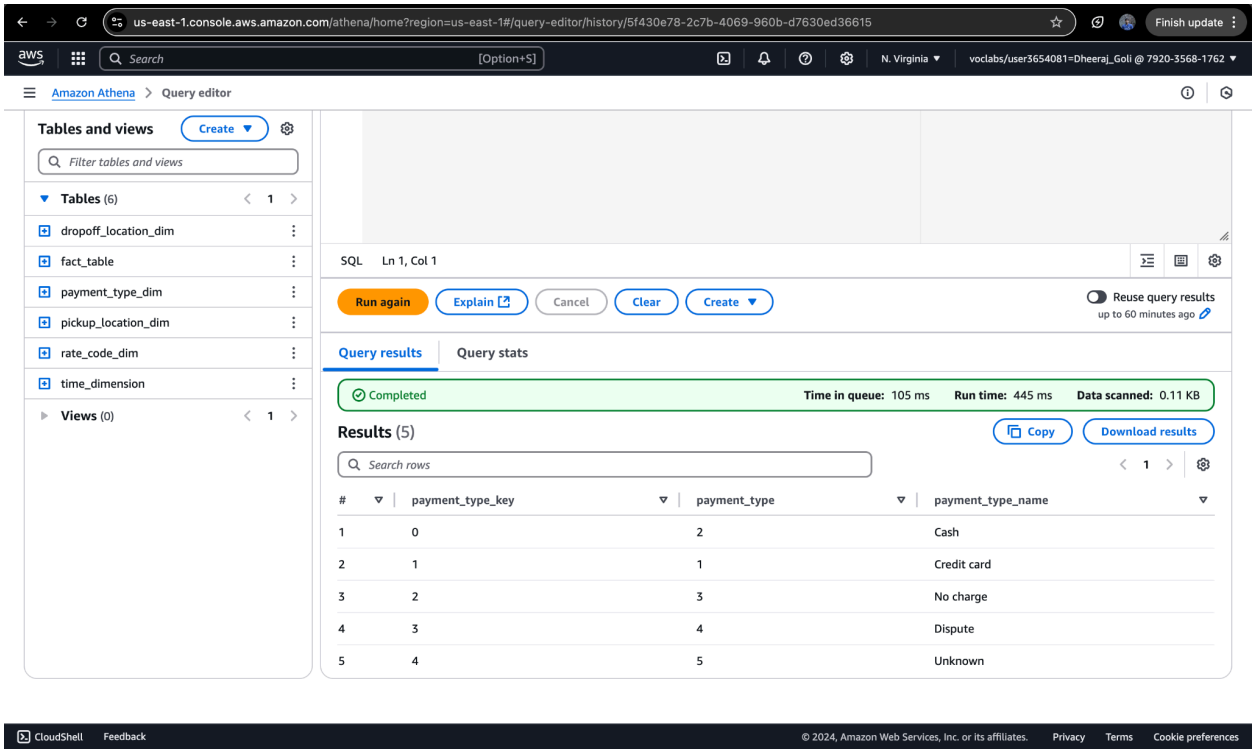


4. OLAP Database:
    - Fact and dimension tables were used for aggregation and querying via Athena.

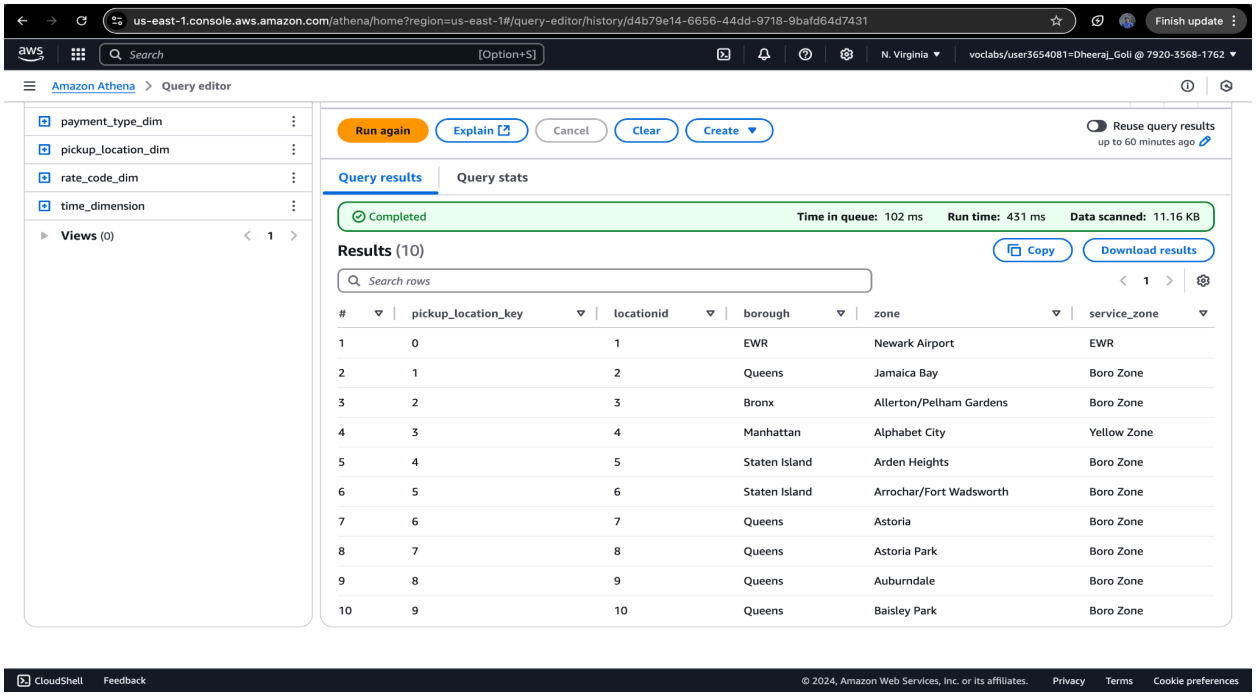
dropoff_location_dim



payment_type_dim

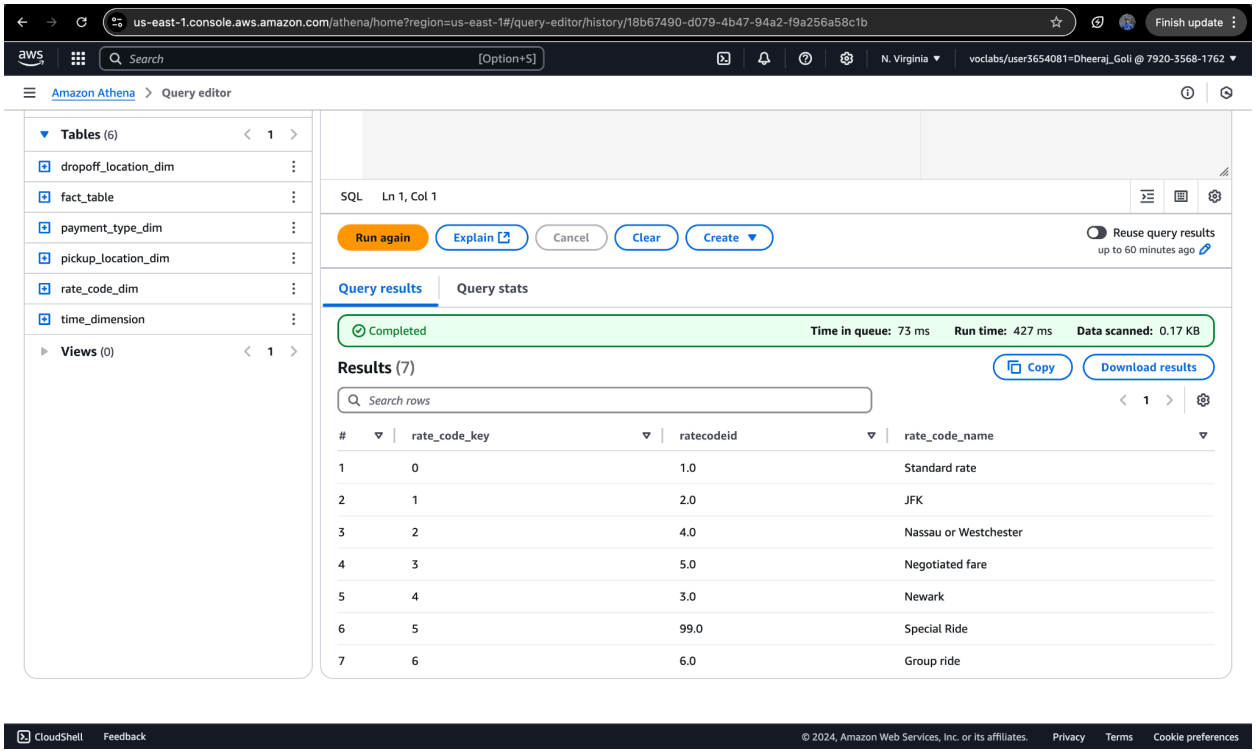


pickup_location_dim



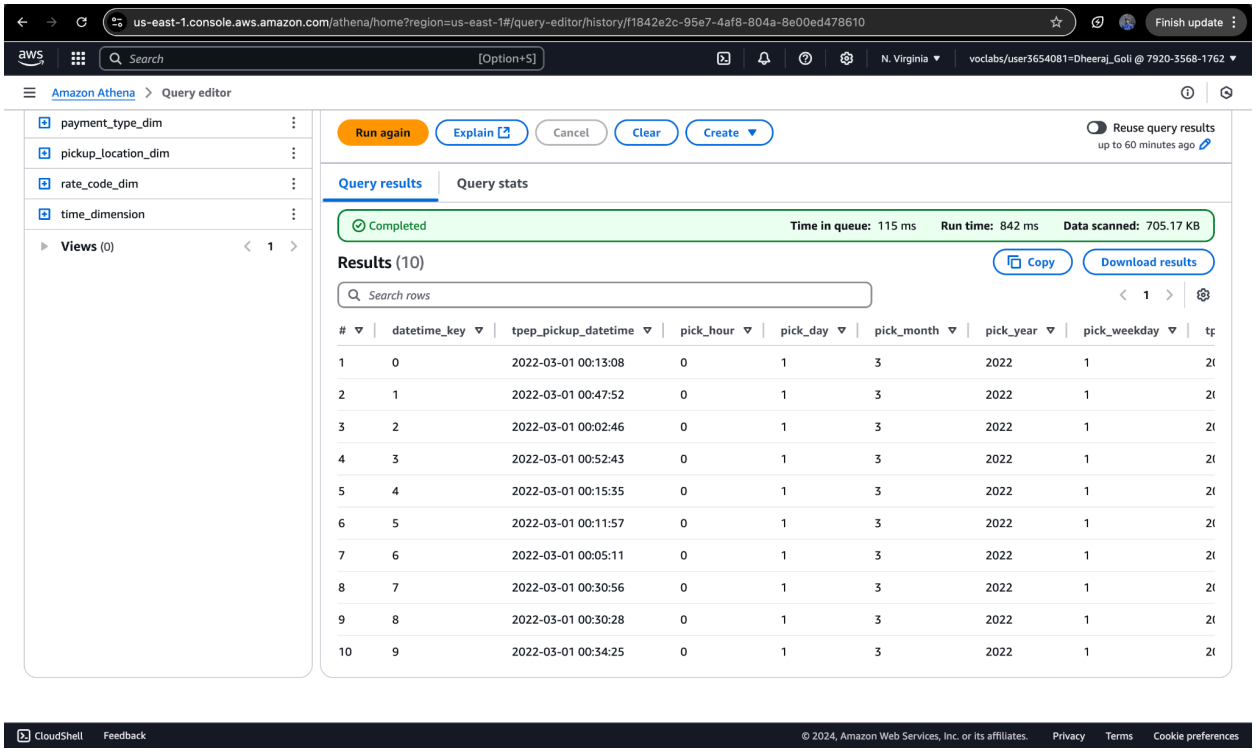
#	pickup_location_key	locationid	borough	zone	service_zone
1	0	1	EWR	Newark Airport	EWR
2	1	2	Queens	Jamaica Bay	Boro Zone
3	2	3	Bronx	Allerton/Pelham Gardens	Boro Zone
4	3	4	Manhattan	Alphabet City	Yellow Zone
5	4	5	Staten Island	Arden Heights	Boro Zone
6	5	6	Staten Island	Arrochar/Fort Wadsworth	Boro Zone
7	6	7	Queens	Astoria	Boro Zone
8	7	8	Queens	Astoria Park	Boro Zone
9	8	9	Queens	Auburndale	Boro Zone
10	9	10	Queens	Baisley Park	Boro Zone

rate_code_dim

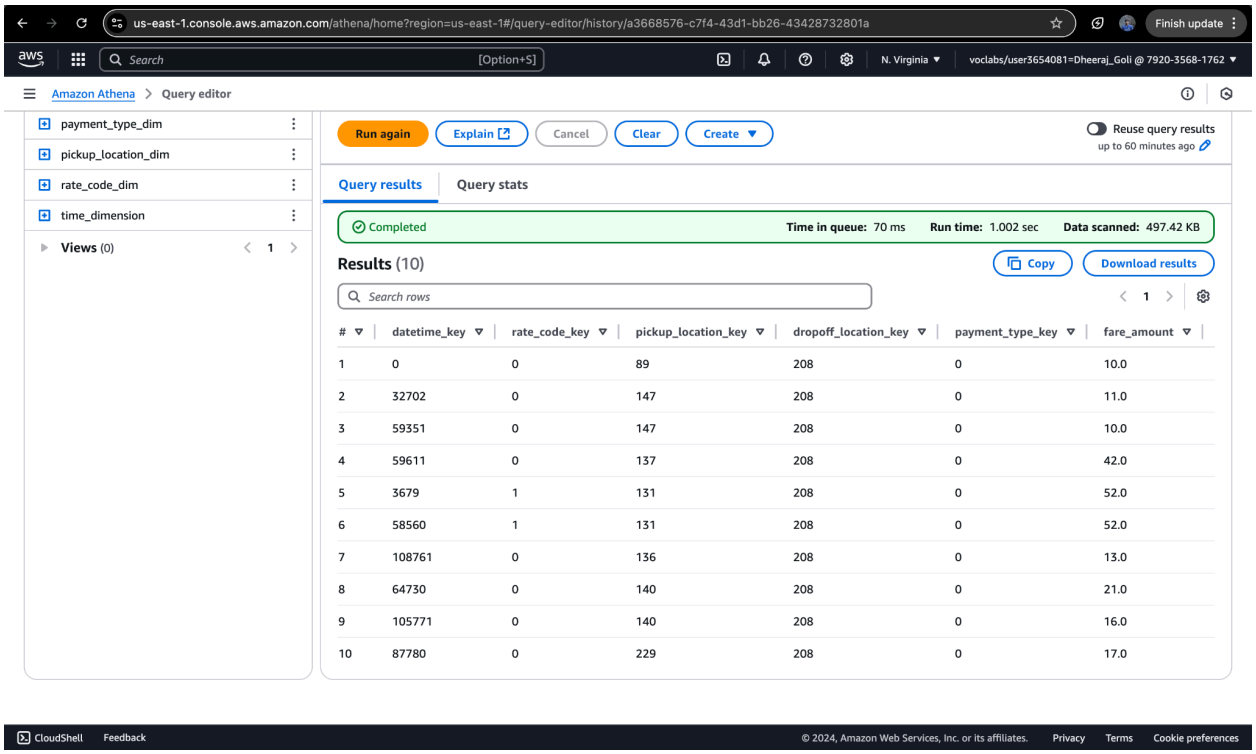


#	rate_code_key	ratecodeid	rate_code_name
1	0	1.0	Standard rate
2	1	2.0	JFK
3	2	4.0	Nassau or Westchester
4	3	5.0	Negotiated fare
5	4	3.0	Newark
6	5	99.0	Special Ride
7	6	6.0	Group ride

time_dimension

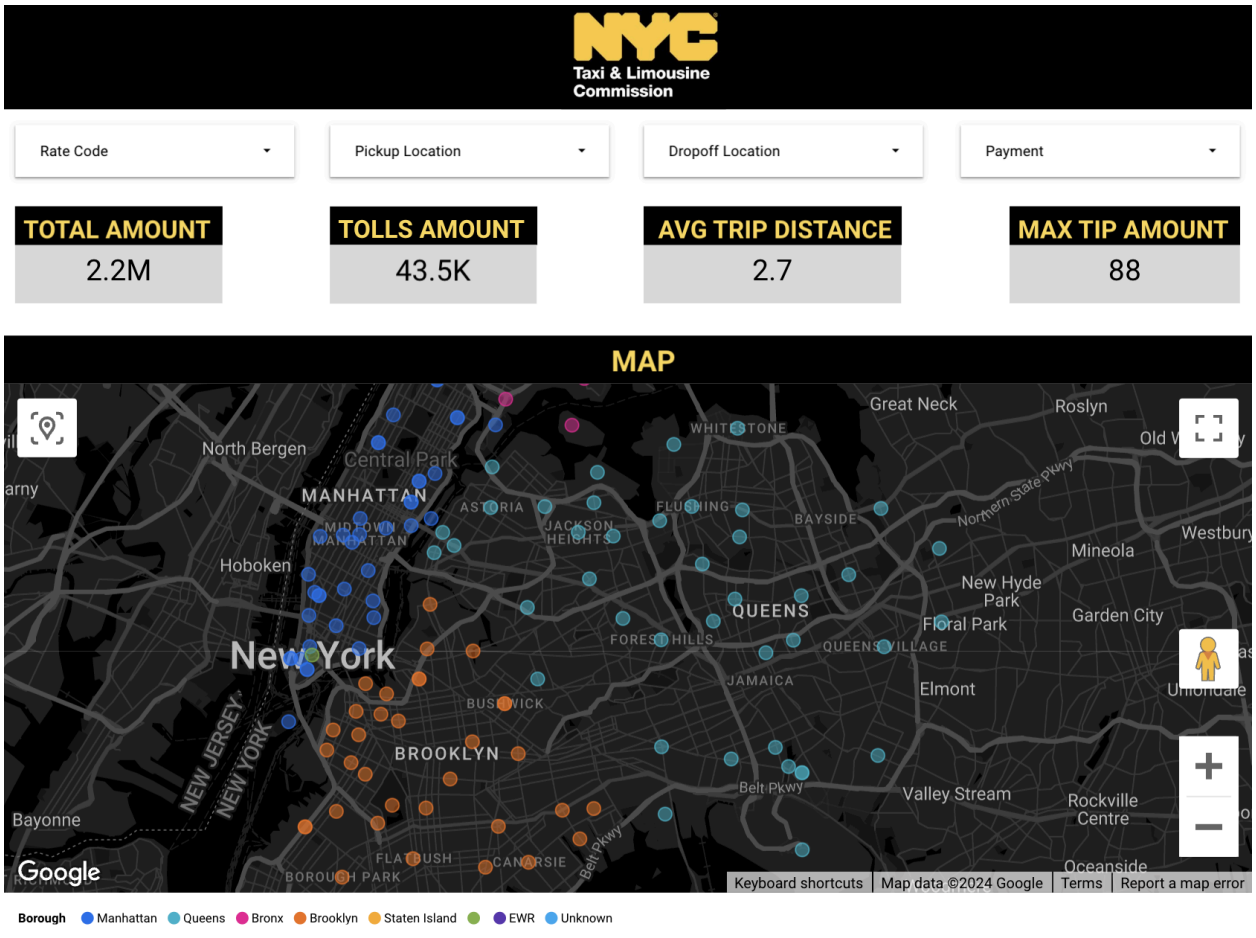

us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/f1842e2c-95e7-4af8-804a-8e00ed478610
Finish update
Search
[Option+S]
N. Virginia
voclabs/user3654081=Dheeraj_Goli @ 7920-3568-1762
Amazon Athena
Query editor
payment_type_dim
pickup_location_dim
rate_code_dim
time_dimension
Views (0)
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 115 ms
Run time: 842 ms
Data scanned: 705.17 KB
Results (10)
Copy
Download results
Search rows
| datetime_key | tpep_pickup_datetime | pick_hour | pick_day | pick_month | pick_year | pick_weekday
1 | 0 | 2022-03-01 00:13:08 | 0 | 1 | 3 | 2022 | 1
2 | 1 | 2022-03-01 00:47:52 | 0 | 1 | 3 | 2022 | 1
3 | 2 | 2022-03-01 00:02:46 | 0 | 1 | 3 | 2022 | 1
4 | 3 | 2022-03-01 00:52:43 | 0 | 1 | 3 | 2022 | 1
5 | 4 | 2022-03-01 00:15:35 | 0 | 1 | 3 | 2022 | 1
6 | 5 | 2022-03-01 00:11:57 | 0 | 1 | 3 | 2022 | 1
7 | 6 | 2022-03-01 00:05:11 | 0 | 1 | 3 | 2022 | 1
8 | 7 | 2022-03-01 00:30:56 | 0 | 1 | 3 | 2022 | 1
9 | 8 | 2022-03-01 00:30:28 | 0 | 1 | 3 | 2022 | 1
10 | 9 | 2022-03-01 00:34:25 | 0 | 1 | 3 | 2022 | 1
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Fact_table

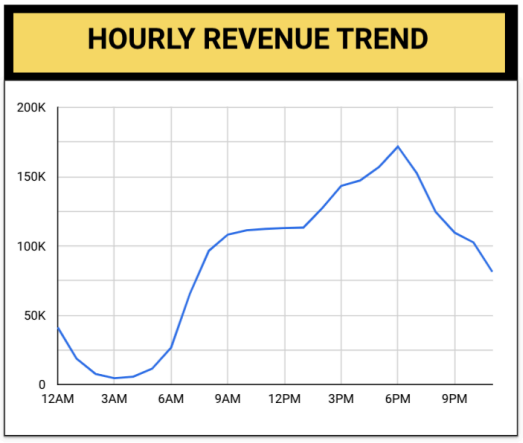

us-east-1.console.aws.amazon.com/athena/home?region=us-east-1#/query-editor/history/a3668576-c7f4-43d1-bb26-43428732801a
Finish update
Search
[Option+S]
N. Virginia
voclabs/user3654081=Dheeraj_Goli @ 7920-3568-1762
Amazon Athena
Query editor
payment_type_dim
pickup_location_dim
rate_code_dim
time_dimension
Views (0)
Run again
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query results
Query stats
Completed
Time in queue: 70 ms
Run time: 1.002 sec
Data scanned: 497.42 KB
Results (10)
Copy
Download results
Search rows
| datetime_key | rate_code_key | pickup_location_key | dropoff_location_key | payment_type_key | fare_amount
1 | 0 | 0 | 89 | 208 | 0 | 10.0
2 | 32702 | 0 | 147 | 208 | 0 | 11.0
3 | 59351 | 0 | 147 | 208 | 0 | 10.0
4 | 59611 | 0 | 137 | 208 | 0 | 42.0
5 | 3679 | 1 | 131 | 208 | 0 | 52.0
6 | 58560 | 1 | 131 | 208 | 0 | 52.0
7 | 108761 | 0 | 136 | 208 | 0 | 13.0
8 | 64730 | 0 | 140 | 208 | 0 | 21.0
9 | 105771 | 0 | 140 | 208 | 0 | 16.0
10 | 87780 | 0 | 229 | 208 | 0 | 17.0
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

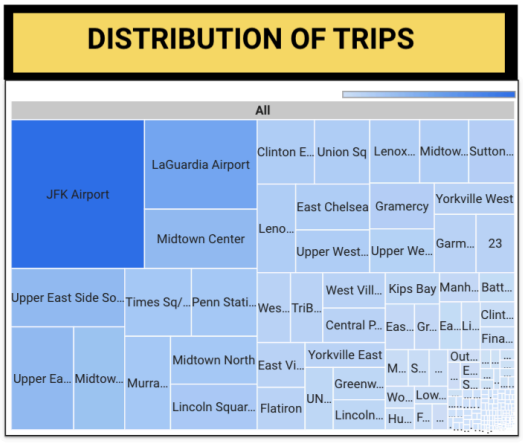
DASHBOARD:

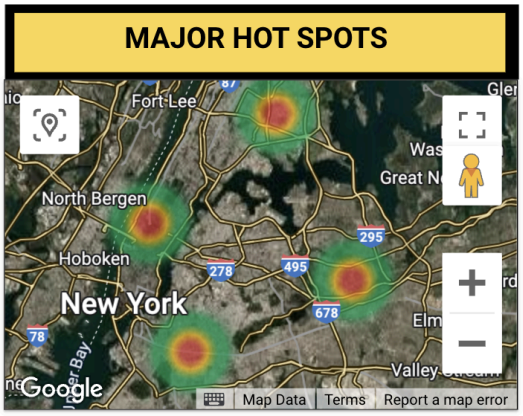
Page 1:

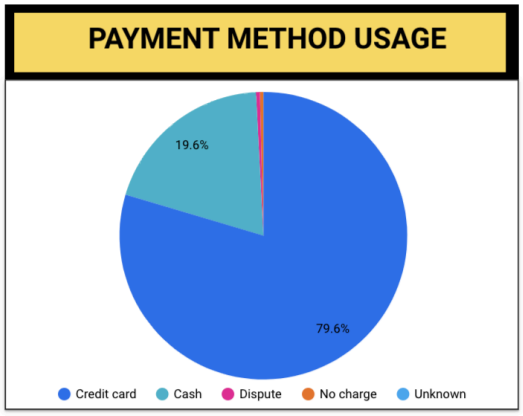


Page 2:









This section highlights insights derived from the analysis of NYC Taxi & Limousine Commission (TLC) trip data. Various visualizations are used to explore revenue trends, trip distributions, hotspot locations, and payment behaviors, offering a comprehensive understanding of the dataset.

1. Hourly Revenue Trend

- **Description**: The line chart visualizes the total revenue generated by taxi rides at different times of the day.
- **Key Insights**:
    - Revenue starts to increase significantly from early morning hours (around 6 AM).
    - A peak in revenue is observed during the late afternoon and early evening (around 6 PM), coinciding with commute hours.
    - Revenue decreases sharply after 9 PM, indicating reduced activity during late-night hours.
- **Interpretation**: The trend suggests that the highest demand for taxi services occurs during traditional working hours and early evening, likely driven by commuting patterns.

2. Distribution of Trips

- **Description**: The treemap represents the distribution of trips across different pickup and drop-off locations within NYC.
- **Key Insights**:
    - Major locations like **JFK Airport**, **LaGuardia Airport**, and **Midtown Center** dominate the trip counts, reflecting high passenger movement in these areas.
    - Dense urban neighborhoods such as **Union Square**, **Times Square**, and **Upper East Side** also contribute significantly to trip volumes.
- **Interpretation**: Airports and business hubs act as key trip generators, highlighting the role of taxis in facilitating intercity travel and commerce.

3. Major Hotspots

- **Description**: The heatmap identifies high-demand areas for taxi pickups and drop-offs.
- **Key Insights**:
    - Hotspots are concentrated in Manhattan, particularly around **Times Square**, **Midtown**, and **Financial District**.
    - Other boroughs like Brooklyn and Queens show localized hotspots near major transit hubs and residential areas.
- **Interpretation**: These hotspots align with areas of high population density, tourism, and economic activity, suggesting these regions as focal points for taxi operations.

4. Payment Method Usage

- **Description**: The pie chart depicts the distribution of payment methods used by taxi passengers.
- **Key Insights**:
    - **Credit card payments** account for the majority (79.6%), followed by **cash payments** (19.6%).
    - Other methods, such as disputes or "no charge," make up a negligible portion.
- **Interpretation**: The high reliance on credit card payments indicates a shift toward digital transactions, reflecting passenger preferences for cashless payments.

5. Summary Statistics

- **Total Amount**: The cumulative revenue from taxi rides is $2.2M, indicating significant financial activity within the dataset.
- **Tolls Amount**: Approximately $43.5K of the total amount corresponds to toll charges, emphasizing the frequent crossing of tolled bridges and tunnels.
- **Average Trip Distance**: An average trip spans 2.7 miles, showing the prevalence of short to medium-distance travel.
- **Max Tip Amount**: The highest recorded tip is $88, reflecting occasional high-value gratuities, possibly from long-distance trips or generous passengers.

6. Borough-Level Trip Mapping

- **Description**: The map provides a geographical overview of trips categorized by borough.
- **Key Insights**:
    - **Manhattan** dominates in trip density, particularly in business and tourist hubs.
    - **Brooklyn** and **Queens** show moderate activity, with trips often originating or ending near bridges, airports, and major neighborhoods.
    - Other boroughs like Staten Island and the Bronx exhibit minimal trip density.
- **Interpretation**: Manhattan remains the core operational zone for taxis, with peripheral boroughs serving as secondary markets.

7. Overall Observations

The analysis provides actionable insights into passenger behavior, operational hotspots, and financial metrics. The visualizations suggest:

- Targeted deployment of taxis in high-demand areas like Manhattan and major airports can maximize revenue.
- Enhanced digital payment options can further cater to passenger preferences.
- Additional analysis can be performed to explore the impact of factors like weather, events, or seasonality on taxi operations.

CONCLUSION:

The analysis of NYC Taxi and Limousine Commission (TLC) data provides valuable insights into passenger behavior, revenue generation, and operational patterns across New York City. The data reveals significant trends such as peak demand during commuting hours, high trip density in Manhattan, and a dominant reliance on digital payment methods. Key hotspots like airports, business districts, and tourist destinations play a critical role in taxi operations, emphasizing the importance of these regions in optimizing service delivery. Additionally, the summary statistics highlight the economic scale of the taxi industry, with millions in revenue generated from short-to-medium-distance trips.

This report demonstrates the utility of data-driven decision-making in understanding and optimizing urban transportation systems. The findings can aid stakeholders such as taxi operators, city planners, and policymakers in making informed decisions to enhance operational efficiency, customer satisfaction, and resource allocation.

FUTURE SCOPING:

The current analysis lays the foundation for further exploration and improvements in the following areas:

1. Sustainability Measures:
    - Analyze fuel consumption patterns and assess the potential benefits of transitioning to electric or hybrid vehicles.
    - Study the environmental impact of taxi operations and propose strategies for reducing carbon footprints.

2. Customer Segmentation:
    - Conduct an in-depth study of customer preferences and behaviors based on trip characteristics, payment methods, and locations.
    - Develop tailored services or loyalty programs for frequent riders.

3. Anomaly Detection and Fraud Prevention:
    - Use advanced analytics to identify anomalies in fare calculations, tip amounts, and payment disputes.
    - Enhance fraud prevention mechanisms for more secure transactions.

4. Event-Driven Insights:
    - Analyze the impact of citywide events (e.g., parades, concerts, or sports games) on taxi operations to better prepare for demand surges.
    - Provide real-time alerts and recommendations to operators during such events.