URBAN MOBILITY IN FOCUS: ANALYSING NYC TAXI TRIPS

CHLOY COSTA-2447116 JOSAIAH M DKHAR-2447125 PRANAB RAI-2447137

CHRIST (DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

INTRODUCTION

Millions of passengers board yellow taxis daily in New York City. The dataset analyzed was downloaded from the NYC Taxi and Limousine Commission. The dataset size is over 1 million observations for a single month of September 2024.

- How do taxi trip patterns vary across different hours of the day, and what trends can be observed over the month of September 2024?
- What is the relationship between trip variables and are there specific patterns or anomalies trips that could provide insights into passenger behavior?
- How do various factors influence the overall fare amounts and are there noticeable inconsistencies in cost distribution?

DATA CLEANING AND PRE-PROCESSING

- **Raw Dataset:** 1 million + observations.
- **File Conversion:** Converted Parquet to CSV for simplicity and broader compatibility.
- **Feature Extraction:**

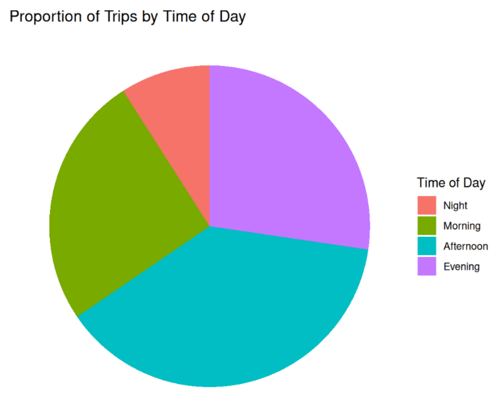
1. **Hour:** Extracted hourly data to represent trips for each hour over 30 days.
2. **Distance:** Filtered out zero/negative values and excluded extreme outliers beyond the 99.5th percentile for reliability.
3. **Fare Amount:** Retained as a key variable for analysis.

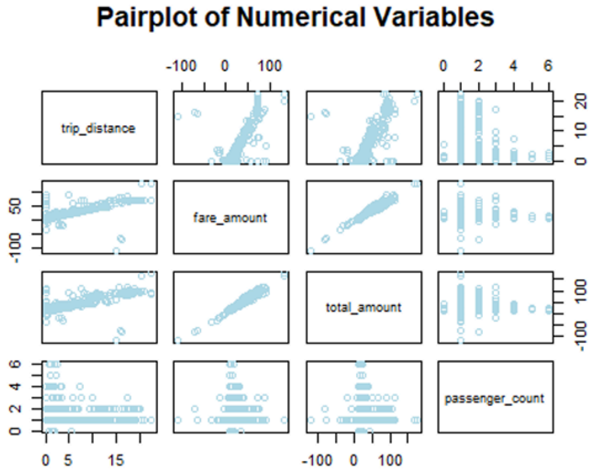
- **Data Cleaning:**
  - Handled missing data in R, ensuring incomplete entries were excluded to maintain data quality.
  - Removed rows with invalid values, such as non-positive fare amounts, zero passenger counts, and zero/negative trip distances.
- Outliers were identified but retained to enable interpretation during analysis.

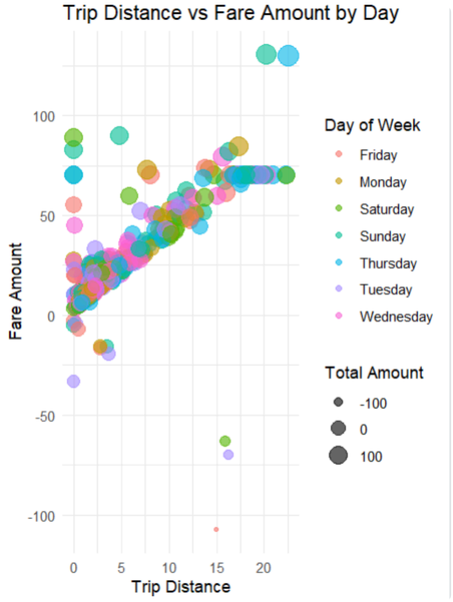
RESULT ANALYSIS

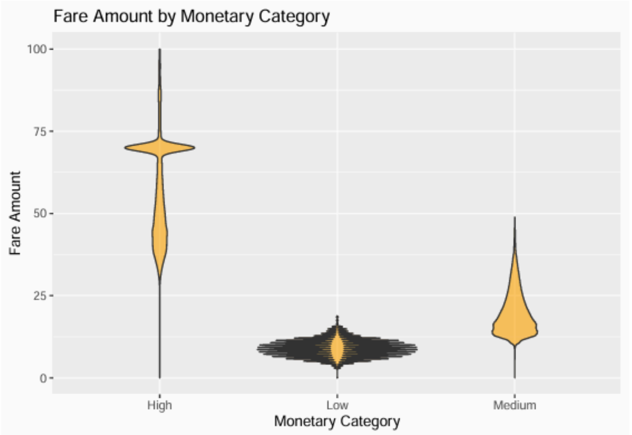
Research Questions	Test Type	Test Score
Weekday vs Weekend Trip Distance	T-test	$p < 0.05$
Trip Distance by Passenger Count	ANOVA	Significant
Fare Variation by Time of Day	Descriptive	Evening Peak
Pickup Location Density	Descriptive	High at Hubs
Correlation between Fare and Trip Distance	Correlation	Strong
Outlier Analysis (Fare vs Distance)	Descriptive	Notable

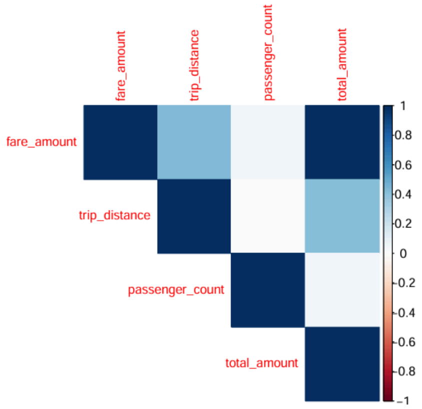
DATA VISUALISATION

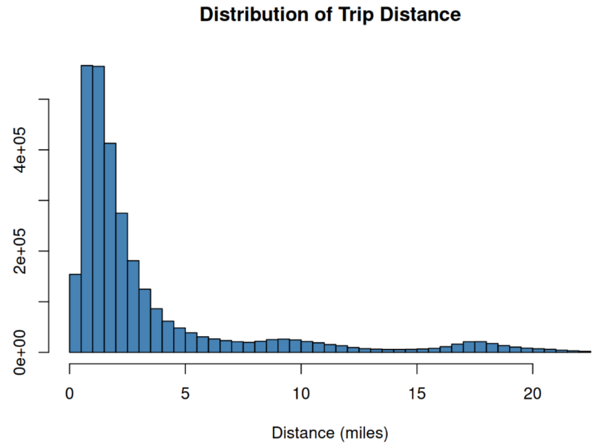












FINDINGS

- Most trips are short-distance, with trip frequency declining as distance increases.
- Evening hours reflect the highest demand, while night trips are least common.
- Fare increases with trip distance, though outliers exist.
- Passenger count influences fare, but variability remains within single-passenger trips.
- Weekend trips are longer and more expensive on average compared to weekdays.
- Popular pickup locations indicate high-traffic hubs and significant urban hotspots.

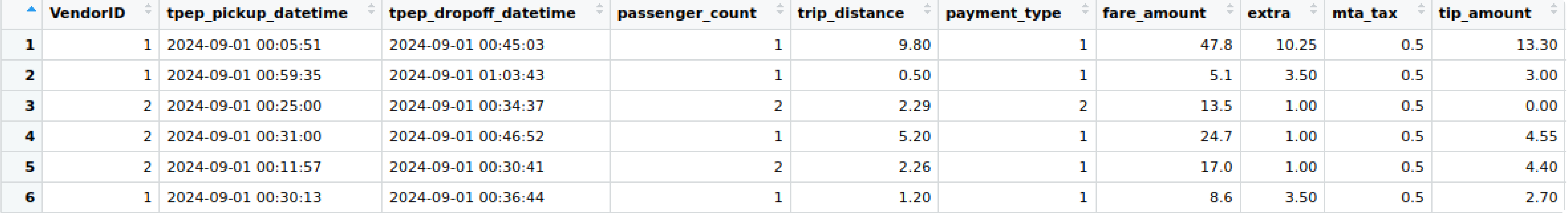
LIMITATIONS

- The dataset was reduced to around 1000 observations for simplicity, which may have led to the loss of some detailed insights.
- Outliers were retained, which could have influenced the analysis outcomes.
- The analysis focused on time-related features, excluding other factors like weather or traffic conditions.
- One observation per hour was assumed to represent the hour, potentially missing within-hour variations.
- The scope was limited to one month, making it difficult to generalize findings to other periods.

CONCLUSION

- This study effectively reduced a large dataset into a manageable size, maintaining key time-based characteristics.
- The approach highlights hourly trends in taxi trip data while providing a scalable method for handling large datasets.
- Future analyses could include additional factors and broaden the scope to ensure a more comprehensive understanding of trip patterns.

DATASET



	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	payment_type	fare_amount	extra	mta_tax	tip_amount
1	1	2024-09-01 00:05:51	2024-09-01 00:45:03	1	9.80	1	47.8	10.25	0.5	13.30
2	1	2024-09-01 00:59:35	2024-09-01 01:03:43	1	0.50	1	5.1	3.50	0.5	3.00
3	2	2024-09-01 00:25:00	2024-09-01 00:34:37	2	2.29	2	13.5	1.00	0.5	0.00
4	2	2024-09-01 00:31:00	2024-09-01 00:46:52	1	5.20	1	24.7	1.00	0.5	4.55
5	2	2024-09-01 00:11:57	2024-09-01 00:30:41	2	2.26	1	17.0	1.00	0.5	4.40
6	1	2024-09-01 00:30:13	2024-09-01 00:36:44	1	1.20	1	8.6	3.50	0.5	2.70

FEEDBACK

Thank you for your patience...