Assignment 2 Handover Report

Author: Chengyang Zhou (24873630)
Big Data Engineering

Introduction

Since 1971, the Taxi and Limousine Commission (TLC) has regulated New York City's iconic taxis, collecting comprehensive datasets from both yellow and green cabs. Yellow taxis, distinguished by their medallions, are a citywide fixture. In contrast, green taxis, rolled out in 2013, primarily cater to specific borough zones. The aim of this project is to use Spark's capabilities to analyze these datasets, culminating in the development of a predictive machine learning model.

Issues/Bugs Faced and Resolutions

Issue 1: Unrealistic Data in Part B
Description:
While working on Part B of the project, I noticed that the output consistently comprised unrealistic data.

Root Cause:
Upon investigation, I discovered that the root cause of the problem was my utilization of the original dataset to create the combined_df. I had mistakenly bypassed the cleaned data produced after the data cleaning phase.

Resolution:
To rectify this, I replaced the original dataset in combined_df with the cleaned dataset. After making this adjustment, the output data was consistent with expectations, indicating that the issue had been resolved.

Issue 2: Slow Model Training Due to Large Dataset
Description:
I experienced considerable delays during the model training phase, which negatively impacted the project's timeline and efficiency.

Root Cause:
The primary reason for this delay was the sheer volume of the dataset. The extensive data size significantly increased the computational time required for model training.

Resolution:
To overcome this challenge:

1. Data Sampling: I sampled 60% of the dataset to reduce its size while still maintaining a representative subset for model training.
2. Feature Selection: I employed feature selection techniques to reduce the number of input features, which not only sped up the training process but also potentially improved the model's generalization by eliminating irrelevant or redundant features.
3. Parallel Processing: I divided the workload by separating the process into two notebooks. This allowed me to simultaneously run two models, effectively halving the total time required for model training.

PART 1: Data Ingestion and Preparation

1. Integration:
A dedicated Spark session was established to ensure optimal processing and analysis of the voluminous datasets at hand.
Azure Blob Storage Integration: By leveraging the provided credentials, an integration with Azure Blob Storage was achieved, setting the stage for robust data ingestion processes.

2. Data Ingestion:
Azure Blob Storage was successfully anchored to Databricks, making the datasets readily accessible at the /mnt/bde-2/ directory.
Comprehensive datasets corresponding to Yellow and Green Taxis, covering the period from 2015 to 2022, were imported. This data serves as the foundational bedrock upon which subsequent operations would be performed.

3. Data Quantification:
Yellow Taxi Dataset: The dataset initially contained a staggering 663,055,251 records, reflecting the extensive operations of yellow taxis over the years.
Green Taxi Dataset: This dataset, though comparatively smaller, was still substantial with 66,200,401 records, showcasing the growing significance of green taxis in NYC's transportation ecosystem.

4. Data Transformation:
Format Conversion: Recognizing the need for format uniformity for specific analytical objectives, the 2015 Green Taxi dataset underwent a transformation from its native Parquet format to the more universally recognized CSV format.

5. Data Cleaning:
- Temporal Consistency:
  Removed records where the trip's drop-off time was recorded before the pick-up time.
- Date Range Filtering:
  Filtered out trips that fell outside the scope of the date range from January 1, 2015, to December 31, 2022.
- Speed Calculation and Filtering:
  Calculated the duration of each trip in hours.
- Derived the speed for each trip by dividing the trip distance by its duration.
  Excluded trips with negative speeds or speeds exceeding 65 mph, ensuring realistic data representations.
- Duration Constraints:
  Filtered out trips with extremely short durations (less than a minute) or overly long durations (greater than 3 hours).
- Distance Constraints:
  Filtered out trips with minimal distances (less than 0.1 miles) or excessively long distances (more than 100 miles).
- Fare Validity:

Removed trips where the fare amount was recorded as negative, ensuring the data's financial integrity.

- Handling Missing Values
  In the dataset, the following strategy was applied to handle NaN values:
  1. For trip_distance, NaN values were replaced with a default value of 0.0. This implies that trips with no recorded distance are considered to have not been taken or recorded erroneously.
  2. For passenger_count, NaN values were replaced with a default value of 1. This approach assumes that if the passenger count was not recorded, the trip likely had one passenger.
  3. RatecodeID missing values were replaced with 1, which is a standard rate code in the taxi fare system.
  4. For both PULocationID and DOLocationID, NaN values were replaced with 0, indicating an unknown or unrecorded location.

6. Data Association:

- Location Referential Data: Locational datasets, containing detailed information about various pick-up and drop-off points, were imported. This data played a pivotal role in enriching the primary datasets.
- Data Joins: The primary taxi datasets were then merged with the locational reference datasets. This was done based on the pick-up and drop-off coordinates, ensuring that each trip record was supplemented with comprehensive locational insights.

7. Conversion of "Green" 2015 Parquet to CSV and Storage on Azure Blob
File Size Comparison:
Upon uploading, it was observed that the CSV format resulted in three separate files on Azure Blob, with sizes of 685 MiB, 682 MiB, and 710 MiB, respectively. In contrast, the original Parquet file size was only 385 MiB.

Explanation:

- Compression and Columnar Storage:
  Parquet is a columnar storage file format. This means it stores data by columns, as opposed to row-based storage like CSV. Due to the nature of columnar storage, similar data types are stored together, making them highly compressible. As a result, Parquet files tend to be more compact than their CSV counterparts.
- Performance:
  Since Parquet is column oriented, it offers better performance for read-heavy operations, especially when only specific columns are needed for a query. This is because only the necessary columns' data blocks are read, not the entire row. In contrast, CSV files need to be read entirely, even if only a subset of columns is required.

8.. Data Archiving:
Parquet Preservation: For the dual objectives of long-term storage and efficient retrieval, the unified dataset was archived as a parquet file on DBFS.

View Creation: To further facilitate quick and efficient querying processes, a dedicated view named 'combined_view' was created from the parquet data.

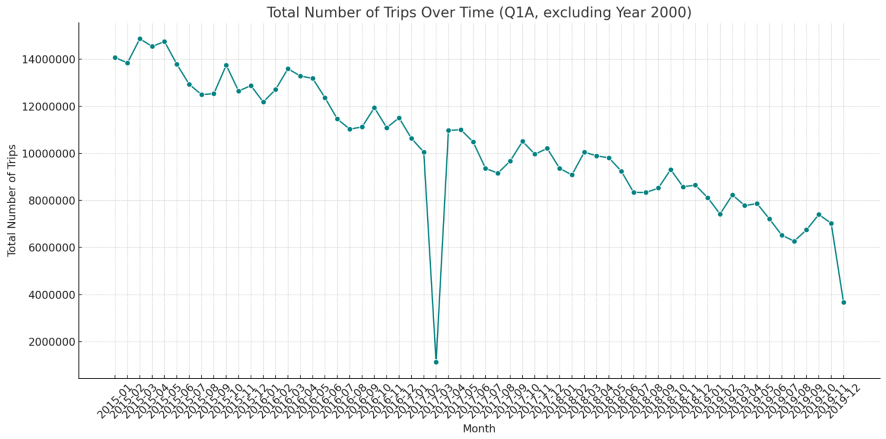
Part 2: Business Questions

Q1.For each year and month (e.g January 2020 => “2020-01-01” or “2020-01” or “Jan 2020”:

	Year	Month	DayOfWeek	HourOfDay	TotalTrips	AvgPassengers	AvgAmountPerTrip	AvgAmountPerPassenger
1	2015	1	1	0	118955	1.7439704089781851	14.681031482497353	8.41816547282999
2	2015	1	1	1	103781	1.7278788988350469	14.787082703001001	8.557939282070404
3	2015	1	1	2	85883	1.7301677864070888	14.925590396234943	8.626672229997885
4	2015	1	1	3	65125	1.724898272552783	15.38519232245324	8.919478074312027
5	2015	1	1	4	40240	1.7148359840954275	16.186308648108803	9.438983551915053
6	2015	1	1	5	18608	1.6527300085984522	18.97073463026676	11.47842329453092
7	2015	1	1	6	17767	1.539483311757753	20.283867844880795	13.17576338110548

10,000 rows | Truncated data | 1.74 minutes runtime — Refreshed 2 hours ago

Q1A: What was the total number of trips?

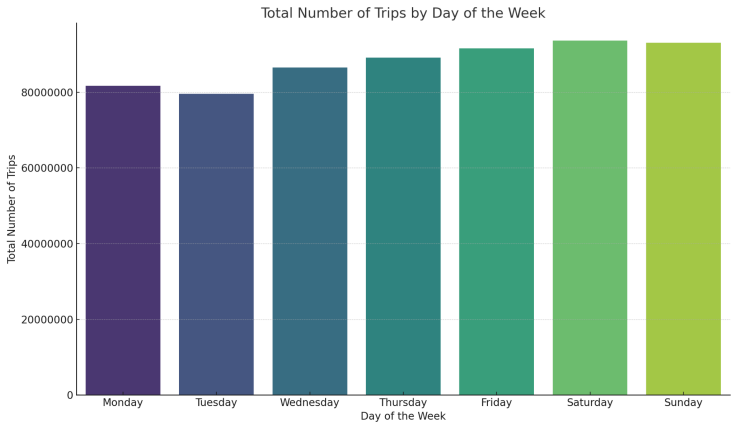


From the graph, it's evident that the total number of trips has been decreasing over time. Let's focus on this gradual decline:

Potential Causes:

- ➔ Market Saturation: The market might have reached a saturation point where growth slows down, and there's a general decrease in demand.
- ➔ Change in Consumer Behavior: People might be shifting to other modes of transportation or using alternative services.
- ➔ Increased Competition: New entrants in the market or alternative services might be drawing customers away.
- ➔ External Factors: Economic downturns, regulatory changes, or long-term external events might be affecting the industry.

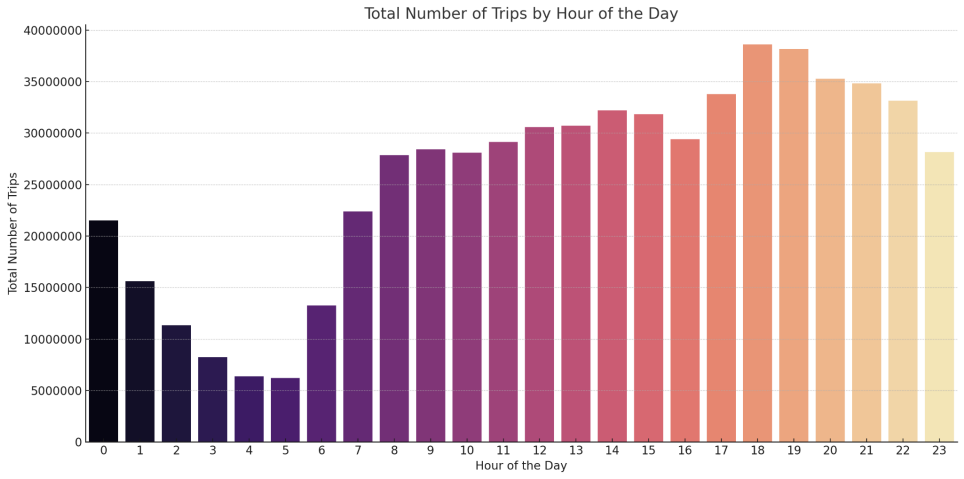
Q1B: Which day of week (e.g. monday, tuesday, etc..) had the most trips?



- ➔ Weekend Popularity: The bars for Saturday and Sunday are notably taller than those for the other days, confirming that weekends have the highest trip counts.
- ➔ Weekday Consistency: Among the weekdays, there's a fairly consistent pattern, with Monday through Friday having relatively similar numbers of trips.
- ➔ Saturday Dominance: Saturday stands out as the day with the most trips, followed closely by Sunday. This suggests that the service is particularly popular on weekends, possibly due to recreational activities, outings, or other weekend-specific factors.

Based on this pattern, businesses could strategize their operations, marketing, or promotions to capitalize on the weekend surge and potentially boost weekday usage.

Q1C: Which hour of the day had the most trips?

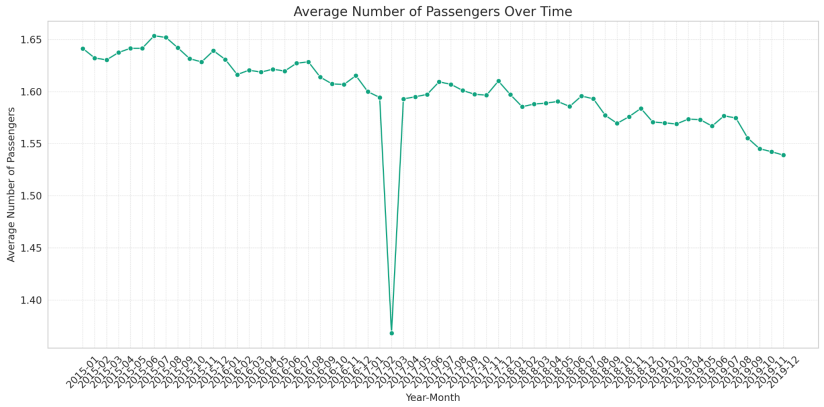


From the bar plot:

- Evening Peak: The bars for the hours from around 17:00 (5 PM) to 23:00 (11 PM) are significantly taller than those for other hours. This indicates that the evening to late-night hours are the busiest in terms of trips.
- Morning Lull: The early morning hours, particularly from 0:00 (midnight) to 5:00 AM, show the least activity with the lowest number of trips. This is expected, as these are typically non-active hours for most people.

Given the clear peak during evening hours, businesses could focus their operations, promotions, or marketing efforts during these hours to capitalize on the surge. Additionally, strategies could be devised to increase trips during the less active hours.

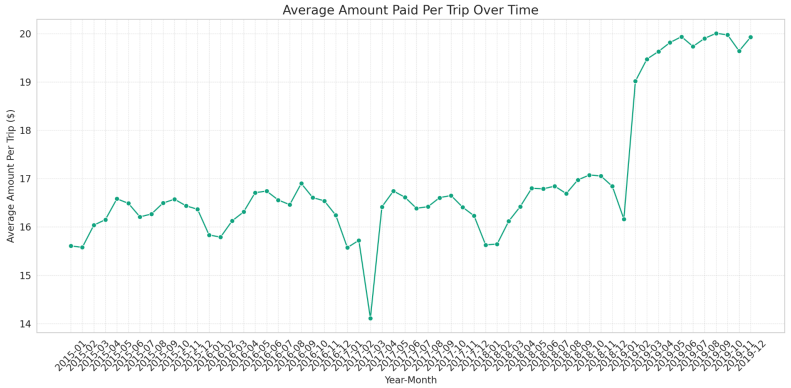
Q1D: What was the average number of passengers?



The line above showcases the average number of passengers per trip over the years:

1. There's a general downward trend in the average number of passengers per trip. This could be attributed to various factors such as changes in taxi fare structures, rise of ride-sharing apps, or special events such as Covid lockdown.
2. The average number of passengers seems to reach its lowest values towards the end of the year, possibly due to holiday seasons when people might prefer private or alternative modes of transportation.

Q1E: What was the average amount paid per trip (using total_amount)?



From the visualization:

1. We can observe that the average amount paid per trip tends to be higher in the middle of the year (around June to August) for most years.
2. There are some fluctuations, but the pattern remains fairly consistent across the years.

Q1F: What was the average amount paid per passenger (using total_amount)?

.

The graph shows the average amount paid per passenger for each month:

1. There seems to be a general increasing trend in the average amount paid per passenger over the years. This might be due to factors like inflation, changes in taxi rates, or changes in demand and supply.
2. Some months, especially in recent years, show higher average amounts, possibly indicating periods of higher demand or increased fare rates.

Q2a: What was the average, median, minimum and maximum trip duration in minutes (with 2 decimals, eg. 90 seconds = 1.50 min)?

TaxiColor	AverageDu	MedianDu	Minimuml	Maximum	AverageDi	MedianDis	Minimuml	Maximum	AverageSp	MedianSp	Minimum	MaximumSpeed
green	14.01	10.73333	1	180	4.93	3.138213	0.160934	160.8857	20.27	18.36008	0.058884	104.6071
yellow	14.43	11.28333	1	180	4.88	2.735878	0.160934	160.8696	18.73	16.44478	0.053734	104.6071

Green Taxis:

- ➔ The average trip duration for green taxis was approximately 14.01 minutes.
- ➔ The median (or the middle value) was approximately 10.73 minutes, suggesting that more than half of the trips were shorter than 14.01 minutes.
- ➔ The shortest trip lasted for just 1 minute, while the longest trip spanned 180 minutes or 3 hours.

Yellow Taxis:

- ➔ Yellow taxis had an average trip duration of about 14.43 minutes, slightly longer than green taxis.
- ➔ The median trip duration for yellow taxis was around 11.28 minutes.
- ➔ Like green taxis, the shortest trip duration was 1 minute, and the longest trip was 3 hours or 180 minutes.

Q2B: What was the average, median, minimum and maximum trip distance in km?

Green Taxis:

- ➔ On average, trips in green taxis covered a distance of approximately 4.93 km.
- ➔ The median distance was around 3.14 km, indicating that more than half of the trips were shorter than the average distance.

➔ The shortest recorded trip was just 0.16 km, while the longest journey spanned an impressive 160.89 km.

Yellow Taxis:

➔ Yellow taxis had an average trip distance of 4.88 km.
➔ The median distance was 2.74 km.
➔ The range of trip distances for yellow taxis was also between 0.16 km and 160.87 km, almost identical to green taxis.

Q2C: What was the average, median, minimum and maximum trip distance in km?

Green Taxis:

➔ The average speed for green taxis was 20.27 km/h.
➔ The median speed, which represents the typical speed for most trips, was about 18.36 km/h.
➔ Speeds varied from as low as 0.059 km/h to as high as 104.61 km/h.

Yellow Taxis:

➔ Yellow taxis had a slightly slower average speed of 18.73 km/h.
➔ The median speed was 16.44 km/h.
➔ The speed range for yellow taxis was between 0.054 km/h and 104.61 km/h
.

Q3. For each taxi colour (yellow and green), each pair of pick up and drop off locations (use boroughs not the id), each month, each day of week and each hours:

	TaxiColor	PickupBorough	DropoffBorough	Month	DayOfWeek	Hour	TotalTrips	AverageDistanceKM	AvgAmountPerTrip	TotalAmountPaid
2	green	Bronx	Bronx	2	6	18	3531	3.89	13.36	47183.83
3	green	Bronx	Bronx	3	2	8	3796	3.94	12.43	47167.05
4	green	Bronx	Bronx	3	6	8	3539	4.33	13.05	46169.35
5	green	Bronx	Bronx	3	6	17	3595	4.03	13.94	50096.61
6	green	Bronx	Bronx	3	6	18	3520	3.87	13.01	45800.27
7	green	Brooklyn	Brooklyn	1	1	0	21003	3.97	13.3	279289.17
8	green	Brooklyn	Brooklyn	1	1	1	18958	4.02	13.22	250584.88
9	green	Brooklyn	Brooklyn	1	1	2	16088	4.13	13.29	213877.36
10	green	Brooklyn	Brooklyn	1	1	3	12958	4.19	13.31	172492.62
11	green	Brooklyn	Brooklyn	1	1	4	8966	4.18	13.21	118451.77
12	green	Brooklyn	Brooklyn	1	1	5	3677	4.34	13.29	48884.36
13	green	Brooklyn	Brooklyn	1	1	9	4785	4.22	13.65	65334.36
14	green	Brooklyn	Brooklyn	1	1	10	7058	3.84	12.94	91301.98
15	green	Brooklyn	Brooklyn	1	1	11	8551	3.56	12.28	105010.84
16	green	Brooklyn	Brooklyn	1	1	12	9730	3.64	12.62	122769.29

Q3A. What was the total number of trips?

- Green Taxi:
  Peaks at 12 PM and is lowest around 2 AM.
- Yellow Taxi:
  Highest number of trips during early morning and evening commute hours, with a significant dip in the early morning hours.

Q3B. What was the average distance?

- Green Taxi:
  Longest average distance at 8 AM, shortest at 2 AM.
- Yellow Taxi:
  Peaks at 3.90 km at 4 AM and reaches its lowest at 2.93 km at 12 PM

Q3C. What was the average amount paid per trip (using total_amount)?

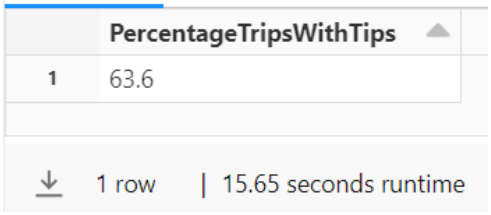
- Green Taxi:
  Highest fare at $12.49 at 8 AM and lowest at $11.15 at 1 AM.
- Yellow Taxi:
  Peaks at $12.94 at 10 PM and is at its lowest at $12.29 at 5 AM.

Q3D. What was the total amount paid (using total_amount)?

- Green Taxi:
  Earnings are highest at approximately $148,592 at 4 PM and lowest at about $47,394 at 2 AM.
- Yellow Taxi:
  Peaks at approximately $4.94 million at 12 AM and reaches its lowest at approximately $460,711 at 5 AM.

Both green and yellow taxis have clear demand patterns throughout the day. Peak demand for trips occurs around midday for green taxis and during early morning and evening commute hours for yellow taxis. The lowest demand is consistently observed during the early morning hours, specifically around 2 AM.
The average trip distance for green taxis is longest around 8 AM, which might indicate longer commutes during morning rush hours. For yellow taxis, the longest trips tend to be around 4 AM. This could be due to airport rides or long-distance commutes.

Q4. What was the percentage of trips where drivers received tips?



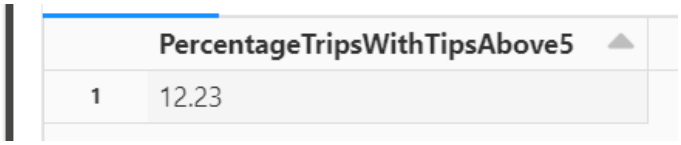
	PercentageTripsWithTips
1	63.6

1 row | 15.65 seconds runtime

- 63.6% of the trips resulted in drivers receiving tips.

This indicates that the majority of taxi rides resulted in drivers receiving a tip. It's a positive sign as it might suggest passenger satisfaction with the service provided or a cultural tendency to tip drivers in most scenarios. However, there could be multiple factors at play influencing tipping behavior, such as payment method, trip duration, service quality, or regional tipping norms.

Q5. For trips where the driver received tips, what was the percentage where the driver received tips of at least $5?



	PercentageTripsWithTipsAbove5
1	12.23

Out of the trips where drivers received tips, approximately 12.23% of those tips were of at least $5.
This suggests that while the majority of riders tend to tip their drivers, higher tip amounts (like $5 or more) are less common. Such a pattern could be influenced by the trip's total fare, the duration of the trip, the quality of service, or cultural norms around tipping. It might be that for shorter trips, passengers tip a smaller absolute amount but might still maintain a

similar percentage of the total fare. Alternatively, exceptional service or longer, more complex trips might result in passengers tipping $5 or more.

Q6. Classification of trips into duration bins and calculation of average speed and average distance per dollar for each bin

	DurationBin	AverageSpeedKMH	AvgDistancePerDollar
1	10 to 20 Mins	20.82	0.33
2	20 to 30 Mins	22.56	0.39
3	30 to 60 Mins	22.65	0.47
4	5 to 10 Mins	18.5	0.25
5	At least 60 Mins	19.73	0.24
6	Under 5 Mins	19.74	0.18

In summary, the analysis indicates that the value in terms of distance per dollar is best for trips lasting between 30 to 60 minutes. The average speed tends to increase with a trip duration of up to 30 minutes, after which it stabilizes or slightly decreases, likely due to varied road conditions or longer distances impacting average speed.

Q7. Duration bin recommendation for a taxi driver to maximize income

	DurationBin	AverageFare	TripCount	EstimatedIncome
1	At least 60 Mins	17.17	653729118	11222471121.24
2	10 to 20 Mins	15.63	21429098	334938989.19
3	20 to 30 Mins	25.36	7681865	194842034.95
4	5 to 10 Mins	9.35	19867030	185754995.63
5	30 to 60 Mins	38.64	4736223	182986284.38
6	Under 5 Mins	6.57	9889870	64986295.91

6 rows | 50.70 seconds runtime

- ➔ The trip duration bin "At least 60 Mins" has the highest estimated income, likely due to a significant number of long-distance trips that can charge higher fares.
- ➔ The "Under 5 Mins" bin has the lowest estimated income, which is expected as these are very short trips with lower fares.
- ➔ Interestingly, while the "30 to 60 Mins" bin has a high average fare, the estimated income is lower than that of "10 to 20 Mins". This suggests that while longer trips do earn more per trip, there are fewer of them compared to the more frequent shorter trips.

For a taxi driver looking to maximize income, targeting longer trips "At least 60 Mins" would be most beneficial based on the fare. However, the feasibility of consistently getting such long trips might be challenging. Therefore, it could be more practical to target the "10 to 20 Mins" duration, which has a significant number of trips and a good balance between trip frequency and fare amount.

PART 3: Machine Learning

In this section, we embarked on building machine learning models to predict the 'Total amount' of a taxi trip. Spark ML pipelines were utilized to craft, train, and validate multiple models. Additionally, a baseline model was developed for comparative evaluation.

Baseline Model:
The baseline model was established by leveraging the average fare calculated in Part 2 (Q3c). Essentially, for every trip in the validation set, the predicted total amount was set to this average value.

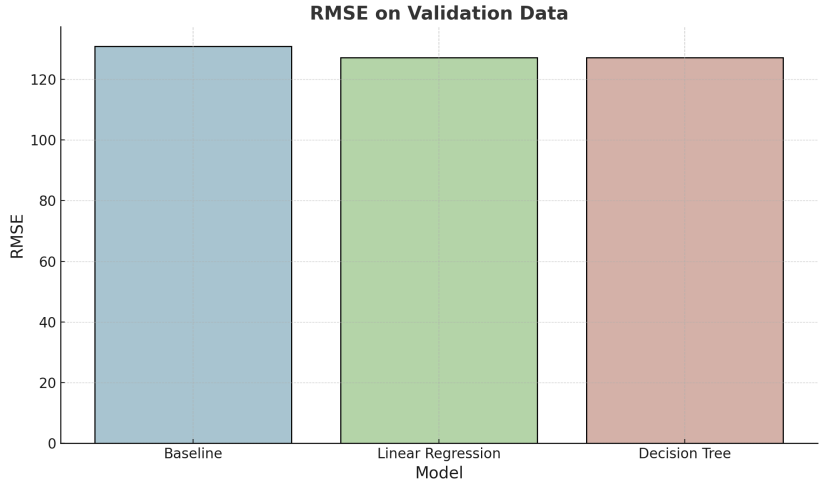
Results:
RMSE on Validation Data: 130.88
RMSE on Test Data: 28.24

Linear Regression Model:
The first ML model implemented was a linear regression model. Features such as trip distance, passenger count, rate code, and trip duration were utilized for predictions.

Results:
RMSE on Validation Data: 127.22
RMSE on Test Data: 7.19

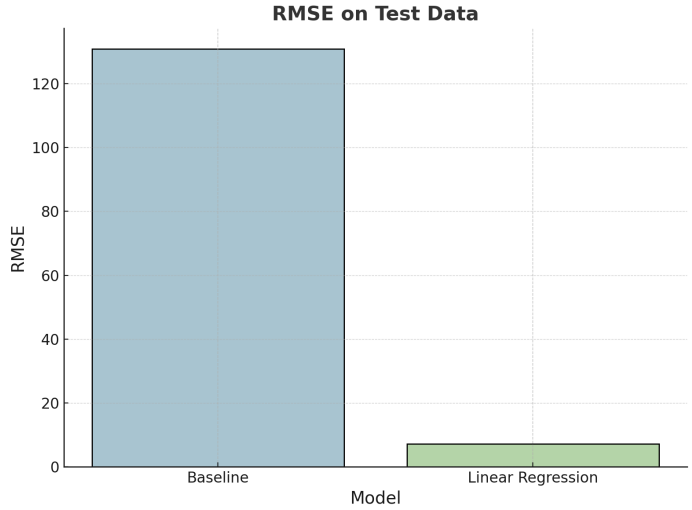
Decision Tree Regression Model:
The second ML model chosen was a decision tree regression. The same set of features as in the linear regression model was used.

Results:
RMSE on Validation Data: 127.22



Model Selection and Discussion:
Both the Linear Regression and Decision Tree Regression models yielded very close RMSE values on the validation data. Surprisingly, they had an almost identical performance on the

validation set. However, upon assessing the models on the test data, the Linear Regression model had a slightly lower RMSE than the Decision Tree model.



Reasons to choose the Linear Regression model over the Decision Tree:
Simplicity: Linear regression is a simpler model which makes it easier to interpret and understand compared to decision trees.
Processing Time: Typically, linear regression models tend to be computationally faster than tree based models.
Accuracy: Even though the difference is minor, the linear regression model had a slightly better RMSE on the test data
.
In the test scenario, both models substantially outperformed the baseline model. The linear regression model's RMSE on the test data was 7.19 compared to the baseline's 28.24, indicating its superior predictive accuracy.

Conclusions:
The linear regression model exhibited a much better predictive performance compared to the baseline model.
The usage of features such as trip distance, passenger count, rate code, and trip duration were instrumental in achieving this predictive performance.
For taxi fare prediction, the linear regression model proves to be both efficient and effective, making it a valuable tool for future fare predictions.