Data Analyst Project

By Parth Shethia

OLA DATA ANALYST PROJECT REPORT

Prepared By: Parth Shethia
Role: Data Analyst / Business Analyst
Tools Used: Python (Pandas), SQL, Power BI

ChatGPT Prompt to Create Data

Please create a spreadsheet with 1 lac rows, for Bengaluru city. Give the following columns. The data will be for 1 month. use the following column -

1. Date
2. Time
3. Booking ID
4. Booking Status
5. Customer ID
6. Vehicle Type
   - Auto
   - Prime Plus
   - Prime Sedan
   - Mini
   - Bike
   - eBike
   - Prime SUV
7. Pickup Location (Create dummy location points Take any 50 areas from Bangalore)
8. Drop Location (Take from dummy pickup locations)
9. Avg VTAT (Time taken to arrive at the vehicle)
10. Avg CTAT (Time taken to arrive the Customer)
11. Cancelled Rides by Customer
12. Reason for cancelling by Customer

- Driver is not moving towards pickup location
- Driver asked to cancel
- AC is not working (Only for 4-wheelers)
- Change of plans

OLA Data Analyst Project

- Wrong Address

13. Cancelled Rides by Driver
- Personal & Car related issues
- Customer related issue
- The customer was coughing/sick
- More than permitted people in there

14. Incomplete Rides
15. Incomplete Rides Reason
- Customer Demand
- Vehicle Breakdown
- Other Issue
16. Booking Value
17. Ride Distance
18. Driver Ratings
19. Customer Rating

Keep the overall booking status success for this data at 62%. If the booking status is successful, then only fare charge ratings, average VTAT, average CTAT, and other data will be there.
Make sure orders cancelled by customers should not be more than 7%
Make sure orders cancelled drivers should not be more than 18%

Also, increase the number of orders on weekends and match days. Keep match day by using the following dates.
keep incomplete rides less than 6%
Keep order value high on weekends

in Food Category keep around 67 Indian keep order ID with
10 digits starting with CNR and then digits keep orders
under 500 value 70% keep orders above 500 value 28%
keep remaining orders above 1000

SQL Questions:

1. Retrieve all successful bookings:
2. Find the average ride distance for each vehicle type:
3. Get the total number of cancelled rides by customers:
4. List the top 5 customers who booked the highest number of rides:
5. Get the number of rides cancelled by drivers due to personal and car-related issues:
6. Find the maximum and minimum driver ratings for Prime Sedan bookings:
7. Retrieve all rides where payment was made using UPI:
8. Find the average customer rating per vehicle type:
9. Calculate the total booking value of rides completed successfully:
10. List all incomplete rides along with the reason:

OLA Data Analyst Project

Power BI Questions:

1. Ride Volume Over Time
2. Booking Status Breakdown
3. Top 5 Vehicle Types by Ride Distance
4. Average Customer Ratings by Vehicle Type
5. cancelled Rides Reasons
6. Revenue by Payment Method
7. Top 5 Customers by Total Booking Value
8. Ride Distance Distribution Per Day
9. Driver Ratings Distribution
10. Customer vs. Driver Ratings

Data Columns

1. Date
2. Time
3. Booking_ID
4. Booking_Status
5. Customer_ID
6. Vehicle_Type
7. Pickup_Location
8. Drop_Location
9. V_TAT
10. C_TAT
11. cancelled_Rides_by_Customer
12. cancelled_Rides_by_Driver
13. Incomplete_Rides
14. Incomplete_Rides_Reason
15. Booking_Value
16. Payment_Method
17. Ride_Distance
18. Driver_Ratings
19. Customer_Rating

Data Characteristics

- **Total Rows:** 100,000 (1 lakh)
- **City:** Bengaluru
- **Duration:** 1 Month
- **Data Nature:** Synthetic (realistic business constraints applied)

Business Constraints Applied

- 62% successful bookings
- Customer cancellations < 7%
- Driver cancellations < 18%
- Incomplete rides < 6%
- Higher booking volume & value on weekends and match days
- 70% of orders under ₹500

OLA Data Analyst Project

- 28% orders between ₹500–₹1000
- Remaining orders above ₹1000
- Booking IDs start with **CNR** followed by digits

Data Cleaning & Preparation (Python – Pandas)

BEFORE CLEANING:-

```
[4]: df.head()
     df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103024 entries, 0 to 103023
Data columns (total 20 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Date                        103024 non-null  datetime64[ns]
 1   Time                        103024 non-null  object
 2   Booking_ID                  103024 non-null  object
 3   Booking_Status              103024 non-null  object
 4   Customer_ID                 103024 non-null  object
 5   Vehicle_Type                103024 non-null  object
 6   Pickup_Location             103024 non-null  object
 7   Drop_Location               103024 non-null  object
 8   V_TAT                       63967 non-null   float64
 9   C_TAT                       63967 non-null   float64
 10  Canceled_Rides_by_Customer  10499 non-null   object
 11  Canceled_Rides_by_Driver    18434 non-null   object
 12  Incomplete_Rides            63967 non-null   object
 13  Incomplete_Rides_Reason     3926 non-null    object
 14  Booking_Value               103024 non-null  int64
 15  Payment_Method              63967 non-null   object
 16  Ride_Distance               103024 non-null  int64
 17  Driver_Ratings              63967 non-null   float64
 18  Customer_Rating             63967 non-null   float64
 19  Vehicle Images              103024 non-null  object
dtypes: datetime64[ns](1), float64(4), int64(2), object(13)
memory usage: 15.7+ MB
```

AFTER CLEANING:-

```
[12]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103024 entries, 0 to 103023
Data columns (total 20 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Booking_ID                  103024 non-null  object
 1   Booking_Status              103024 non-null  object
 2   Customer_ID                 103024 non-null  object
 3   Vehicle_Type                103024 non-null  object
 4   Pickup_Location             103024 non-null  object
 5   Drop_Location               103024 non-null  object
 6   V_TAT                       63967 non-null   float64
 7   C_TAT                       63967 non-null   float64
 8   Canceled_Rides_by_Customer  10499 non-null   object
 9   Canceled_Rides_by_Driver    18434 non-null   object
 10  Incomplete_Rides            63967 non-null   object
 11  Incomplete_Rides_Reason     3926 non-null    object
 12  Booking_Value               103024 non-null  int64
 13  Payment_Method              63967 non-null   object
 14  Ride_Distance               103024 non-null  int64
 15  Driver_Ratings              63967 non-null   float64
 16  Customer_Rating             63967 non-null   float64
 17  booking_datetime            103024 non-null  datetime64[ns]
 18  ride_outcome                103024 non-null  object
 19  cancelled_by                28933 non-null   object
dtypes: datetime64[ns](1), float64(4), int64(2), object(13)
memory usage: 15.7+ MB
```

OLA Data Analyst Project

Cleaning Steps Performed

- **Loaded Excel data** and inspected structure, datatypes, and missing values.
- **Removed irrelevant columns** (`Vehicle Images`).
- **Converted text nulls** (`"null"`, `#NAME?`) into real missing values.
- **Merged Date and Time** into a single `booking_datetime` column.
- **Standardized booking outcomes** (`Success`, `Canceled`, `Failed`) and identified who canceled.
- **Unified cancellation reasons** into one column and dropped redundant fields.
- **Converted numeric columns** to proper numeric types.
- **Applied business rules**: ratings kept only for successful rides.
- **Standardized categorical fields** (case, spacing).
- **Renamed columns** to MySQL-friendly format.
- **Exported cleaned data** as CSV and imported into MySQL.
- **Removed hidden carriage-return characters (`CHAR(13)`)** and whitespace from text fields.
- 📌 **Outcome:**

A clean, structured dataset ready for SQL analysis and Power BI visualization.

SQL Analysis & Data Modeling

Database Setup

```
CREATE DATABASE Ola;
USE Ola;
```

SQL Views Created

- `Successful_Bookings`
- `ride_distance_for_each_vehicle`
- `cancelled_rides_by_customers`
- `Top_5_Customers`
- `Rides_cancelled_by_Drivers_P_C_Issues`

OLA Data Analyst Project

- `Max_Min_Driver_Rating`
- `UPI_Payment`
- `AVG_Cust_Rating`
- `total_successful_ride_value`
- `Incomplete_Rides_Reason`

📌 **Business Benefit:**
SQL views improve query performance and enable seamless integration with Power BI.

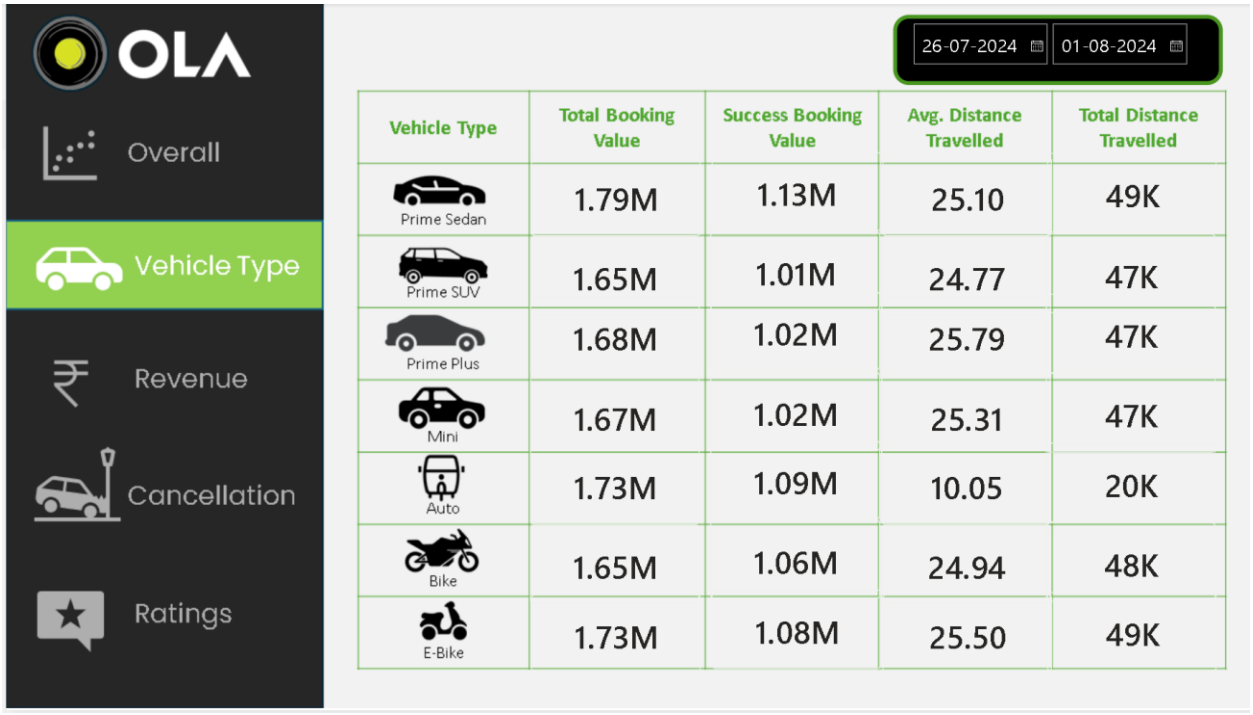
SQL Questions & Insights

Key Queries Answered

1. Successful bookings list
2. Average ride distance per vehicle type
3. Total customer cancellations
4. Top 5 customers by number of rides
5. Driver cancellations due to personal/car issues
6. Max & min driver ratings for Prime Sedan
7. UPI-based payments
8. Average customer ratings by vehicle
9. Total revenue from successful rides
10. Incomplete rides with reasons

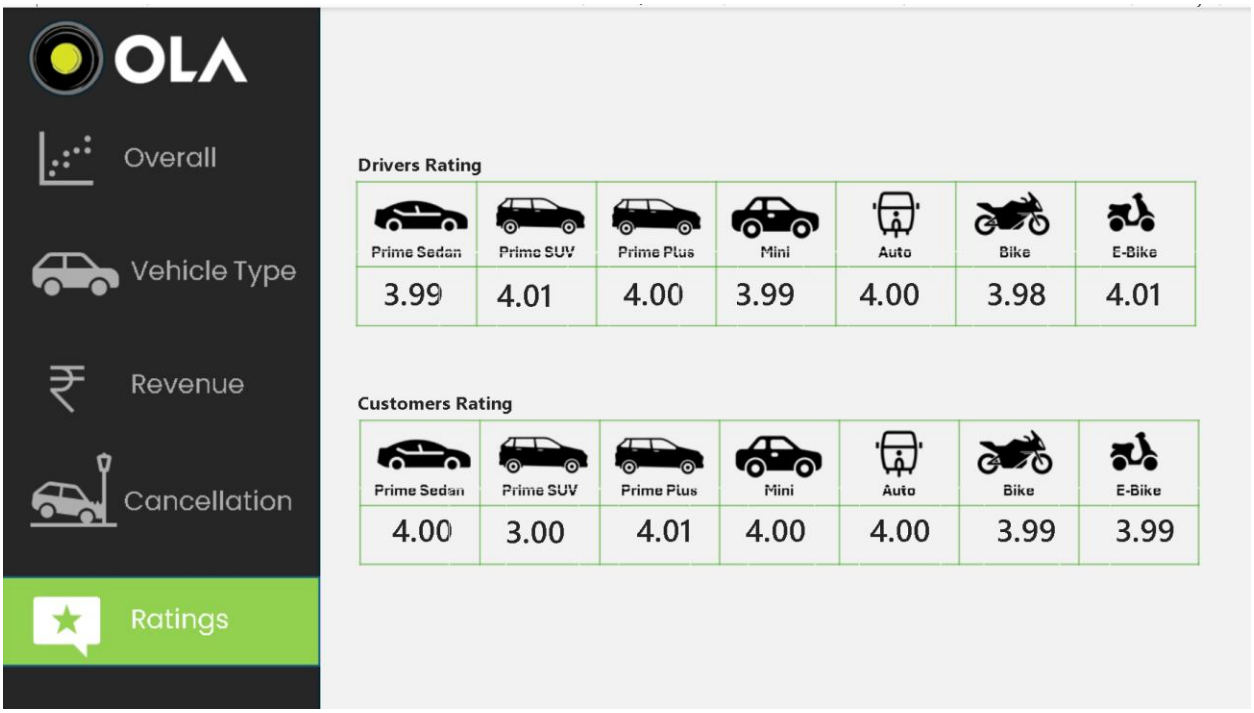
OLA Data Analyst Project

Power BI Dashboard Design

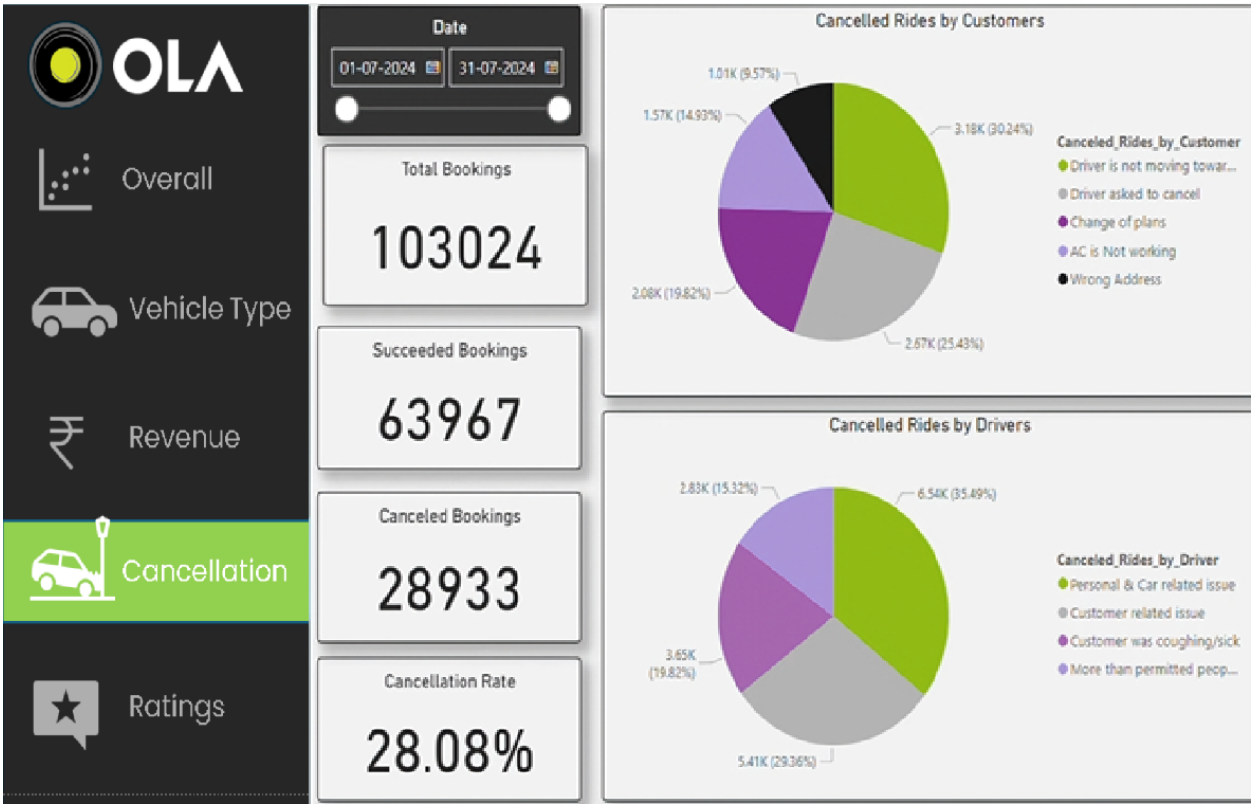


Vehicle Type	Total Booking Value	Success Booking Value	Avg. Distance Travelled	Total Distance Travelled
Prime Sedan	1.79M	1.13M	25.10	49K
Prime SUV	1.65M	1.01M	24.77	47K
Prime Plus	1.68M	1.02M	25.79	47K
Mini	1.67M	1.02M	25.31	47K
Auto	1.73M	1.09M	10.05	20K
Bike	1.65M	1.06M	24.94	48K
E-Bike	1.73M	1.08M	25.50	49K

OLA Data Analyst Project



OLA Data Analyst Project



The Power BI report is divided into **five analytical sections** for clear business storytelling.

Key Business Insights

1. Overall Booking Performance

- Total bookings during the period: **103,024**
- Successful bookings: **63,967**
- Cancelled bookings: **28,933**
- Overall cancellation rate: **28.08%**
- Nearly **1 out of every 3 bookings** does not convert into a completed ride, directly impacting revenue and customer experience.
- Ride volume across days remains relatively stable with minor peaks, indicating predictable demand patterns.

OLA Data Analyst Project

2. Cancellation Analysis

Customer-Side Cancellations

- Major reasons include:
  - Driver not moving towards pickup location
  - Driver requesting customer to cancel
  - Change of travel plans
  - Wrong pickup address
- These reasons indicate **operational execution issues** rather than lack of customer intent.

Driver-Side Cancellations

- Primary reasons include:
  - Personal and vehicle-related issues
  - Customer-related issues
  - Passenger policy violations (exceeding permitted passengers)
- Driver reliability and availability remain critical contributors to cancellations.

3. Revenue Insights

- Total booking value generated: **₹35 million**
- Cash is the dominant payment method, followed by UPI.
- Credit and debit card usage is minimal.
- Heavy cash reliance increases payment handling risk and reduces operational efficiency.

4. Vehicle Type Performance

- Total booking value by vehicle type:
  - Prime Sedan: **₹1.79M**
  - Prime SUV: **₹1.65M**
  - Prime Plus: **₹1.68M**
  - Mini: **₹1.67M**
  - Auto: **₹1.73M**
  - Bike: **₹1.65M**
  - E-Bike: **₹1.73M**
- Average distance travelled:
  - Auto: **~10 km** (short-distance trips)
  - Other vehicle types: **24–26 km** (medium to long-distance trips)
- Different vehicle types cater to distinct travel needs and pricing strategies should reflect this difference.

OLA Data Analyst Project

5. Customer and Driver Ratings

- Average driver ratings across vehicle categories: **3.98 – 4.01**
- Average customer ratings across vehicle categories: **3.99 – 4.01**
- Ratings indicate acceptable service quality but highlight scope for improvement to achieve a premium experience.

6. Customer Behavior and Ride Distance

- A small group of customers contributes a disproportionately high share of total revenue.
- Revenue concentration is higher in short and mid-distance rides, with fewer long-distance bookings.
- Retention of high-value customers is critical for long-term growth.

7. Key Business Recommendations

- Reduce cancellation rate by:
  - Tracking driver movement immediately after ride acceptance
  - Penalizing drivers who request customers to cancel
  - Automatically reallocating drivers when delays are detected
- Promote digital payments by incentivizing UPI and reducing cash dependency.
- Optimize vehicle-type strategy by:
  - Promoting autos for short urban trips
  - Offering premium pricing and bundles for longer rides in Prime categories
- Improve service quality by focusing on driver training and pickup punctuality to raise average ratings beyond **4.5**.
- Introduce loyalty programs to retain high-value repeat customers and stabilize revenue.

8. Summary

- The platform demonstrates **strong demand and revenue potential**, but operational inefficiencies—especially cancellations—limit overall performance.
- Addressing driver reliability, payment digitization, and customer retention can significantly improve profitability and customer satisfaction without increasing acquisition costs.

OLA Data Analyst Project