DATA ANALYSIS PROJECT

By. Aryan Dave

UBER Data Analyst Project

ChatGPT Prompt to Create Data

Please create a spreadsheet with 90000 rows, for Delhi. Give the following columns.

The data will be for 2 Years. use the following column -

1. Date

2. Time

3. Booking ID

4. Booking Status

5. Customer ID

6. Vehicle Type

- Auto
- UberXL (SUV's)
- Premier (Sedan)
- Uber Go
- Moto
- Courier Delivery

7. Pickup Location (Create dummy location points Take any 50 Real Areas from Delhi)

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
- Waiting Time is Too long
- Change of plans
- Wrong Address

13. Cancelled Rides by Driver

    - Personal & Car related issues
    - Customer related issue
    - Pickup Too far

- More than permitted people in there

14. Incomplete Rides

15. Incomplete Rides Reason

- Customer Demand
- Vehicle Breakdown
- Other Issue

16. Booking Value

17. Payment Method

- UPI
- Credit Card
- Uber Wallet
- Cash
- Debit Card

18. Ride Distance

19. Driver Ratings

20. Customer Rating

Keep the overall booking status success for this data at 65%. 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 20%

Make sure orders cancelled drivers should not be more than 10%

Also, increase the number of orders on weekends and match days. Keep match day by using the following dates.

keep incomplete rides less than 7%

Keep order value high on weekends.

Fill the "Payment Method" column only where the "Booking Status" is "Successful". For all other booking statuses, the Payment Method should remain blank or null.

Fill the "Payment Method" with the following probabilities:

UPI → 40%

Credit Card → 5%

Debit Card → 8%

Cash → 29%

Uber Wallet → 18%.

In Rental Category keep around 67 Indian

keep order ID with 10 digits starting with UBR and then digits

keep orders under 500 value 60%

keep orders above 500 value 28%

keep remaining orders above 1000

UBER Data Analyst Project

Requirements-

SQL Questions-

1. Retrieve all bookings where the Booking Value is greater than ₹1000.
2. Find the total number of bookings for each Vehicle Type.
3. Get the number of successful rides that started and ended in the same location.
4. List the top 10 customers with the highest average Booking Value.
5. Retrieve the total number of bookings cancelled by customers on weekends.
6. Find the most frequently occurring cancellation reason by drivers.
7. Calculate the average Booking Value for Successful Rides.
8. What is the percentage of bookings made using each Payment Method out of all successful bookings?
9. Find rides longer than 20 km where Avg CTAT was below 5 minutes.
10. Find the Pickup Locations with the highest number of incomplete rides due to 'Vehicle Breakdown'.
11. Retrieve the earliest and latest Booking Date in the dataset.
12. Calculate the total Booking Value of successful rides grouped by month and Payment Method.
13. List the customers who have cancelled more than 3 rides.
14. Find the average VTAT and CTAT for each Vehicle Type on match days.
15. Get the Pickup and Drop combinations (route pairs) that have been used more than 30 times.

Power Bi Questions-

1. What trends are observed in ride volumes over the years 2022 and 2023?
2. How are bookings distributed across different statuses?
3. Which vehicle types have the highest total distance travelled?
4. Which vehicle types show the most booking volume?
5. Which vehicle types generate the highest successful booking value?
6. How has revenue changed over time during 2022 and 2023?
7. Which payment methods generate the most revenue?

8. What is the revenue contribution of high-value vs normal rides?
9. Which vehicle types generate the most revenue?
10. Who are the top customers by total booking value?
11. What are the most common reasons for customer cancellations?
12. What are the most common reasons for incomplete rides?
13. How are cancellations and failures distributed overall?
14. What are the average customer ratings by vehicle type?
15. How do driver ratings vary across vehicle types?

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. Avg VTAT	10. Avg CTAT
11. Cancelled by Customer Reason	12. Cancelled by Driver Reason
13. Incomplete Ride Reason	14. Booking Value
15. Payment Method	16. Ride Distance
17. Driver Ratings	18. Customer Rating

UBER Data Analyst Project

Data Cleaning

Tool- MS Excel

Replaced blank cells with NULL
Used Find and Replace to convert all empty cells to "NULL" to ensure compatibility with MySQL during import.

Removed duplicate records
Applied Excel's Remove Duplicates feature across all columns, removing **1,086** exact duplicate rows from the dataset.

Standardized values in Vehicle Type column

- Corrected inconsistent values using Find and Replace:
    - "AUTO" and "auto" → "Auto"
    - "moto" and "motoo" → "Moto"
    - "uber go", "Uber go", "Ubr Go" → "Uber Go"
- Used the TRIM function to eliminate leading/trailing white spaces in this column.

Fixed invalid values in Payment Method column

- Replaced incorrect entries with standardized values:
    - "Cashh" → "Cash"
    - "Upii" → "UPI"
    - "creditcard" → "Credit Card"
    - "UberWallet" → "Uber Wallet"

Standardized cancellation reasons in Cancelled by Customer Reason

- Unified inconsistent entries to "Driver Asked to cancel"
    - Fixed variations like "Driver ask to cancel", "Driver askd to cancel", and "driver asked to cancel".

Filtered out invalid booking dates

- Retained only rows with dates in **2022** and **2023**.
- Removed all rows with dates from **2021**, **2025**, and **2026**.

n

- Verified the cleaned dataset contains exactly **90,000 rows.**

Imported the cleaned CSV into MySQL

Used the MySQL Workbench Data Import Wizard to load the dataset into a new table.

New Table – Uber Data

Further Data Cleaning

Removed any leftover whitespace using TRIM () within SQL queries.

Converted data types where needed (e.g., fixing numeric columns stored as text). how to do it

Date & Time Formatting

- **Formatted the Date column:**
    - Converted string dates to SQL DATE format using STR_TO_DATE ().
    - Changed the column's data type to DATE.
- **Formatted the Time column:**
    - Transformed time strings (e.g., 01:25) into SQL TIME format.
    - Updated the column's data type to TIME.
- **Modified Booking ID:**
    - Set to VARCHAR (12) to store alphanumeric booking codes.
- **Adjusted Customer ID:**
    - Set to VARCHAR (10) for customer identifiers.
- **Updated Pickup Location and Drop Location:**
    - Set both columns as VARCHAR (50) to accommodate location names.
- **Cleaned Booking Status:**
    - Initially set as VARCHAR (25).
    - Then converted to an ENUM to only allow:
        - Successful, Incomplete, Cancelled by Customer, Cancelled by Driver.
- **Standardized Vehicle Type:**

- Converted to an ENUM with allowed values:
    - Premier, Uber Go, Courier Delivery, Auto, UberXL, Moto.

- **Validated Payment Method:**
    - Converted to an ENUM to accept only:
        - Cash, UPI, Credit Card, Debit Card, Uber Wallet, Moto.
- **Formatted Booking Value:**
    - Changed to DECIMAL (10,1) for accurate monetary representation.
- **Formatted Ride Distance:**
    - Set to DECIMAL (10,1) for precise distance values.
- **Formatted Rating Columns:**
    - Driver Ratings and Customer Rating both set to DECIMAL (5,1) to store one decimal precision (e.g., 4.3, 3.5).

The dataset has been cleaned using both Excel and MySQL and is now ready for analysis.

MySQL ANSWERS

1. Retrieve all bookings where the Booking Value is greater than ₹1000.

   Answers- SELECT * FROM Booking_Value_greater_than_₹1000;

   Syntax - CREATE VIEW Booking_Value_greater_than_₹1000 AS

   SELECT * FROM Uber_data

   WHERE `Booking Value` > 1000;

2. Find the total number of bookings for each Vehicle Type.

   Answers- SELECT * FROM

   Total_number_of_bookings_for_each_Vehicle_Type;

   Syntax - CREATE VIEW Total_number_of_bookings_for_each_Vehicle_Type AS

   SELECT `Vehicle Type`, COUNT (`Booking ID`) As Total_num_of_Booking

   FROM Uber_data

   GROUP BY `Vehicle Type';

3. Get the number of successful rides that started and ended in the same location.

   Answers- SELECT * FROM Successful_rides_Ended_in_the_Same_Location;

   Syntax - CREATE VIEW Successful_rides_Ended_in_the_Same_Location AS

   SELECT * FROM uber_data

   WHERE `Booking Status` = 'Successful'

   AND `Pickup Location` = `Drop Location`;

4. List the top 10 customers with the highest average Booking Value.

   Answers- SELECT * FROM

   Top_10_Customer_With_Highest_Average_Booking_Value;

   Syntax - CREATE VIEW

   Top_10_Customer_With_Highest_Average_Booking_Value AS

   SELECT `Customer ID`, AVG (`Booking Value`) AS Avg_Booking_Value

   FROM uber_data

   GROUP BY `Customer ID`

   ORDER BY Avg_Booking_Value DESC

   LIMIT 10;

5. Retrieve the total number of bookings cancelled by customers on weekends.

   Answers- SELECT * FROM

   Total_bookings_Cancelled_by_customer_on_weekends;

   Syntax - CREATE VIEW Total_bookings_Cancelled_by_customer_on_weekends

   AS

   SELECT COUNT (`Booking ID`)

   FROM uber_data

   WHERE `Booking Status` = 'Cancelled by Customer' AND

   DAYOFWEEK(`Date`) IN (1,7);

6. Find the most frequently occurring cancellation reason by drivers.

   Answers- SELECT * FROM

   Most_frequently_occurring_cancellation_reason_by_drivers;

   Syntax - CREATE VIEW Most_frequently_occurring_cancellation_reason_by_drivers

   AS

```
SELECT `Cancelled by Driver Reason`, COUNT (*) AS total Cancellation
FROM uber_data
WHERE `Booking Status` = 'Cancelled by Driver'
GROUP BY `Cancelled by Driver Reason`
ORDER BY total Cancellation DESC;
```

7. Calculate the average Booking Value for Successful Rides.

   Answers- SELECT * FROM Average_Booking_Value_for_Successful_Rides;

   Syntax-
```
CREATE VIEW Average_Booking_Value_for_Successful_Rides AS
SELECT AVG (`Booking Value`)
FROM uber_data
WHERE `Booking Status` = 'successful';
```

8. What is the percentage of bookings made using each Payment Method out of all successful bookings?

   Answers- SELECT * FROM Percentage_of_booking_made_using_each_payment_method;

   Syntax-
```
CREATE VIEW Percentage_of_booking_made_using_each_payment_method
AS
SELECT `Payment Method`, COUNT (*) / (SELECT COUNT (*) FROM uber_data
WHERE `Booking Status` = 'Successful') * 100 AS Percentage_Booking_Made
FROM uber_data
WHERE `Booking Status` = 'Successful'
GROUP BY `Payment Method`;
```

9. Find rides longer than 20 km where Avg CTAT was below 5 minutes.

   Answers- SELECT * FROM Rides_longer_than_20km_where_AvgCTAT_was_below_5_minutes;

   Syntax-
```
CREATE VIEW
Rides_longer_than_20km_where_AvgCTAT_was_below_5_minutes AS
SELECT *
FROM uber_data
WHERE `Booking Status` = 'Successful'
AND `Ride Distance` > 20
```

AND `Avg CTAT` < 5;

10. Find the Pickup Locations with the highest number of incomplete rides due to 'Vehicle Breakdown'.

    Answers- SELECT * FROM Incomplete_Ride_due_to_Vehicle_Breakdown;

    Syntax- CREATE VIEW Incomplete_Ride_due_to_Vehicle_Breakdown AS

    SELECT `Pickup Location`, COUNT (*) AS Incomplete_Rides

    FROM uber_data

    WHERE `Incomplete Ride Reason` = 'Vehicle Breakdown'

    AND `Booking Status` = 'Incomplete'

    GROUP BY `Pickup Location`

    ORDER BY Incomplete_Rides DESC;

11. Retrieve the earliest and latest Booking Date in the dataset.

    Answers- SELECT * FROM Earliest_and_latest_Booking_Date;

    Syntax- CREATE VIEW Earliest_and_latest_Booking_Date AS

    SELECT MIN(`Date`) AS Earliest_Date, MAX(`Date`) AS Latest_Date

    FROM uber_data;

12. Calculate the total Booking Value of successful rides grouped by month and Payment Method.

    Answers- SELECT * FROM Booking_Value_Grouped_by_Month_and_Payment_Method;

    Syntax- CREATE VIEW Booking_Value_Grouped_by_Month_and_Payment_Method AS

    SELECT date format (`Date`, '%Y-%m') AS Month, `Payment Method`, SUM (`Booking Value`) AS Total_Booking_Value

    FROM uber_data

    WHERE `Booking Status` = 'Successful'

    GROUP BY Month, `Payment Method`

    ORDER BY Month DESC;

13. List the customers who have cancelled more than 3 rides.

    Answers- SELECT * FROM Customers_who_have_cancelled_more_than_3_rides;

    Syntax- CREATE VIEW Customers_who_have_cancelled_more_than_3_rides AS

    SELECT `Customer ID`, COUNT (*) AS Cancelled Ride

    FROM uber_data

    WHERE `Booking Status` = 'Cancelled by Customer'

```
GROUP BY `Customer ID`
HAVING COUNT (*) >2
ORDER BY Cancelled_Ride DESC;
```

14. Find the average VTAT and CTAT for each Vehicle Type on match days.

    Answers- SELECT * FROM Average_VTAT_and_CTAT_for_each_Vehicle_Type_on_match_days;

    Syntax-
```
CREATE VIEW Average_VTAT_and_CTAT_for_each_Vehicle_Type_on_match_days AS
SELECT `Vehicle Type`, AVG (`Avg VTAT`) AS Avg_VTAT, AVG (`Avg CTAT`) AS Avg_CTAT
FROM uber_data
WHERE `Booking Status` = 'Successful'
AND `Date` IN (‘2022-03-27’, ‘2022-05-15’, ‘2022-08-21’, ‘2023-03-19', '2023-06-11’,'2023-10-22')
GROUP BY `Vehicle Type`;
```

15. Get the Pickup and Drop combinations (route pairs) that have been used more than 30 times.

    Answers- SELECT * FROM Pickup_Drop_combinations_been_used_more_than_30_times;

    Syntax-
```
CREATE VIEW Pickup_Drop_combinations_been_used_more_than_30_times AS
SELECT `Pickup Location`, `Drop Location`, COUNT (*) AS Total_Rides
FROM uber_data
WHERE `Booking Status` = 'Successful'
GROUP BY `Pickup Location`, `Drop Location`
HAVING COUNT(*) >= 30
ORDER BY Total_Rides DESC;
```

Power Bi Questions Segregation

Overall Dashboard (Ride Volume & Booking Status)

Q1. What trends are observed in ride volumes over the years 2022 and 2023?

Q2. How are bookings distributed across different statuses?

Vehicle Type Dashboard

Q3. Which vehicle types have the highest total distance travelled?

Q4. Which vehicle types show the most booking volume?

Q5. Which vehicle types generate the highest successful booking value?

Revenue Dashboard

Q6. How has revenue changed over time during 2022 and 2023?

Q7. Which payment methods generate the most revenue?

Q8. What is the revenue contribution of high-value vs normal rides?

Q9. Which vehicle types generate the most revenue?

Q10. Who are the top customers by total booking value?

Cancellation Dashboard

Q11. What are the most common reasons for customer cancellations?

Q12. What are the most common reasons for incomplete rides?

Q13. How are cancellations and failures distributed overall?

Ratings Dashboard

Q14. What are the average customer ratings by vehicle type?

Q15. How do driver ratings vary across vehicle types?

Power Bi Answers-

Overall Dashboard (Ride Volume & Booking Status)

Q1. What trends are observed in ride volumes over the years 2022 and 2023?

The ride volume line charts show that bookings slightly increased in 2023 (45,086) compared to 2022 (44,914). Both years experienced dips in February and September, and peaks in October, indicating a recurring seasonal trend.

Q2. How are bookings distributed across different statuses?

According to the booking status doughnut chart:

- 65.02% of bookings were successful
- 20.11% were cancelled by customers
- 10.01% were cancelled by drivers
  This reveals that over one-third of all bookings fail due to cancellations.

Vehicle Type Dashboard

Q3. Which vehicle types have the highest total distance travelled?

Courier Delivery leads with 155.9K km, followed by Premier (151.7K km) and Uber XL (151.6K km), showing couriers typically cover longer distances.

Q4. Which vehicle types show the most booking volume?

Courier has the highest booking count (9,971), followed by Uber XL (9,783) and Premier (9,730), indicating high demand for these services.

Q5. Which vehicle types generate the highest successful booking value?

Courier tops the chart with ₹6.87M, followed by Premier (₹6.72M) and Uber XL (₹6.61M), again highlighting the revenue power of courier services.

Revenue Dashboard

Q6. How has revenue changed over time during 2022 and 2023?

Revenue trends follow a seasonal pattern, peaking in April, July, and October, while dipping in February and September. These trends align closely with ride volumes.

Q7. Which payment methods generate the most revenue?

UPI is the top payment method by revenue, followed by Cash and Uber Wallet. Debit and Credit Cards contribute the least, showing a strong user preference for digital/mobile payments.

Q8. What is the revenue contribution of high-value vs normal rides?

62.53% of revenue comes from normal rides, while 37.47% is from high-value rides. This shows that premium services make up a significant chunk of earnings.

Q9. Which vehicle types generate the most revenue?

Courier Delivery contributes the most revenue, followed by Premier, Auto, and Uber Go. This is consistent with their ride distance and booking volume data.

Q10. Who are the top customers by total booking value?

The top customer is CUST97411 (₹6,746), followed by CUST44748 (₹5,257) and CUST97228 (₹5,092). These high-value customers are ideal for loyalty or premium engagement strategies.

Cancellation Dashboard

Q11. What are the most common reasons for customer cancellations?

The main reasons are:

- More than permitted wait time (25.38%)
- Pickup too far (25.09%)
- Customer-related issues (24.98%)
  These reflect user frustration with delays or logistics.

Q12. What are the most common reasons for incomplete rides?

The major causes of incomplete rides are:

- Other Issues: 1,507
- Vehicle Breakdown: 1,471
- Customer Demand: 1,400
  Technical failures and ride interruptions due to customer changes are key contributors.

Q13. How are cancellations and failures distributed overall?

Overall ride failures are broken down as:

- 57.48% cancelled by customers
- 28.61% cancelled by drivers

- 13.91% incomplete rides
  Customer cancellations are the most frequent failure reason.

Ratings Dashboard

Q14. What are the average customer ratings by vehicle type?

Customer ratings range from 2.58 to 2.61. Auto and Moto have the highest ratings (2.61), while Premier is the lowest at 2.58. Ratings are generally consistent across services.

Q15. How do driver ratings vary across vehicle types?

Driver ratings also stay close to 2.60. Auto and Moto again lead with 2.61, while Premier is slightly lower at 2.57. Overall, driver performance is perceived similarly across categories.