Data Analysis Project Report

By – Arijit Das

1. ChatGPT prompt to create data -

Create a spreadsheet with 25 thousand rows, for Kolkata city. The data will be for 1 month, consider any month from the year 2024.
Use the following columns -

1. Date (Follow DD-MM-YYYY HH:MM:SS format)

2. Time (Follow HH:MM:SS format)

3. Booking_ID

4. Booking_Status
   - Cancelled by Driver
   - Cancelled by Customer
   - Driver not found
   - Success

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 60 areas from Kolkata)

8. Drop_Location (Take from dummy pickup locations)

9. V_TAT (Time taken to arrive at the vehicle)

10. C_TAT (Time taken to arrive at the customer)

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

12. Cancelled_Rides_by_Driver
    - Personal & Car related issues
    - Customer related issue
    - The customer was coughing/sick
    - More than permitted people in there

13. Incomplete_Rides
    - Yes
    - No
14. Incomplete_Rides_Reason
    - Customer Demand
    - Vehicle Breakdown
    - Other Issue

15. Booking_Value

16. Payment_Method
    - Cash
    - Credit Card
    - Debit Card
    - UPI

17. Ride_Distance

18. Driver_Ratings (ratings should range from 1 to 5 with 1 being the lowest and 5 being the highest)

19. Customer_Ratings (ratings should range from 1 to 5 with 1 being the lowest and 5 being the highest)

- Keep the overall Booking_Status Success for this data at 70%. If the Booking_Status is Success, then, ratings, VTAT, CTAT, and other relevant data must be there.
- For those records where there is no particular value which is there to be entered w.r.t the column, assign the values to be null.
- The Booking_value should never contain null value.
- Make sure orders cancelled by customers should not be more than 8%.
- Make sure orders cancelled drivers should not be more than 12%.
- Keep incomplete rides less than 8%.
- Also, increase the number of bookings on weekends.
- Keep Booking_Value high on weekends.
- Keep Booking_ID with 10 digits starting with CNR followed by the digits.
- Keep Customer_ID with 6 digits starting with CID followed by the digits.
- Make sure that 70% of the data is having Booking_value below 500.
- Make sure that 25% of the data is having Booking_value above 500.
- Make sure that remaining data has Booking_Value above 1000.

2. 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_Ratings**

3. Business Problems based on Problem Statement –

Data Transformation 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.

Data Visualization Charts:
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

4. Initial Pre-Processing of data in Excel (Data Cleaning) –

After doing initial data cleaning of our data through excel, the cleaned dataset looks like :

Ola-Cab-Bookings-Kolkata.xlsx

5. Importing data to MS SQL Server for further data transformation –

```
Create database [Ola Bookings];
Use [Ola Bookings];

create view bookings as
select * from [Ola-Cab-Bookings-Kolkata];
select * from bookings;
```

1. **Retrieve all successful bookings.**

```
create view Successful_Bookings as
select * from bookings where Booking_Status = 'success';
```

Thus query which can be used to fetch the desired records directly is –

```
select * from Successful_Bookings;
```

2. **Find the average ride distance for each vehicle type.**

```
create view Avg_Ride_Distance_by_VehicleType as
select vehicle_type, avg(ride_distance) [Avg Ride Distance] from bookings
group by vehicle_type;
```

Thus query which can be used to fetch the desired records directly is -

```
select * from Avg_Ride_Distance_by_VehicleType;
```

3. **Get the total number of cancelled rides by customers.**

```
create view Cancelled_Rides_by_Customers as
select count(*) [No. of cancelled rides by customer] from bookings
where booking_status = 'cancelled by customer';
```

Thus query which can be used to fetch the desired records directly is –

```
select * from Cancelled_Rides_by_Customers;
```

4. **List the top 5 customers who booked the highest number of rides.**

```
create view Top_5_Customers as
select top(5) customer_id, count(booking_id) [No. of rides booked] from bookings
group by customer_id
order by [No. of rides booked] desc;
```

Thus query which can be used to fetch the desired records directly is –

```
select * from Top_5_Customers;
```

5. **Get the number of rides cancelled by drivers due to personal and car-related issues.**

```
create view [No. of rides cancelled by drivers due to Personal & Car related issues] as
select count(*) [No. of rides cancelled by drivers] from bookings
where Cancelled_Rides_by_Driver = 'Personal & Car related issues';
```

Thus query which can be used to fetch the desired records directly is –

```
select * from [No. of rides cancelled by drivers due to Personal & Car related issues];
```

6. **Find the maximum and minimum driver ratings for Prime Sedan bookings.**

```
create view Max_Min_Driver_Ratings as
select max(Driver_Ratings) [Max. Driver Ratings],
min(Driver_Ratings) [Min. Driver Ratings]
from bookings
where vehicle_type = 'Prime Sedan';
```

Thus query which can be used to fetch the desired records directly is –

```
select * from Max_Min_Driver_Ratings;
```

7. **Retrieve all rides where payment was made using UPI.**

```
create view UPI_Payment as
select * from bookings where payment_method = 'UPI';
```

Thus query which can be used to fetch the desired records directly is –

```
select * from UPI_Payment;
```

8. **Find the average customer rating per vehicle type.**

```
create view avg_CR_per_vehicle_type as
select vehicle_type, avg(customer_ratings) [Avg customer rating] from bookings
group by vehicle_type;
```

Thus query which can be used to fetch the desired records directly is –

```
select * from avg_CR_per_vehicle_type;
```

9. **Calculate the total booking value of rides completed successfully.**

```
create view [Total Successful Booking Value] as
select sum(booking_value) [Total Successful Booking Value] from bookings
where booking_status = 'Success' and Incomplete_Rides = 'No';
```

Thus query which can be used to fetch the desired records directly is –

```
select * from [Total Successful Booking Value];
```

10. **List all incomplete rides along with the reason.**

```
create view All_Incomplete_Rides as
select booking_id, incomplete_rides_reason from bookings
where incomplete_rides = 'Yes';
```

Thus query which can be used to fetch the desired records directly is –

```
select * from All_Incomplete_Rides;
```

7. Data Visualization in Power BI -

1. Overall
   - Ride Volume Over Time
   - Booking Status Breakdown

2. Vehicle Type
   - Top 5 Vehicle Types by Ride Distance

3. Revenue
   - Total Booking Value by Month
   - Ride Distance Distribution by Month
   - Top 5 Customers by Total Booking Value

4. Cancellation
   - Cancelled Rides Reasons (Customer)
   - cancelled Rides Reasons (Drivers)

5. Ratings
   - Driver Ratings
   - Customer Ratings

8. Business Problems that could be addressed after data analysis -

- Calculating average ride distance for each vehicle type and the average customer rating per vehicle type is a good measure of segregating various vehicle types to identify popularity.

- Segmenting customers using ride history, location and ratings, and identifying top customers with highest booking value to design personalized offers and discounts can prove to be a great measure to increase customer loyalty.

- Analysing incomplete rides with a list of rides with incomplete rides reason can help in taking steps to reduce those ride cancellations.

- A visual representation of Total Booking value per month can provide an important indicator that which month of the year aids to maximum revenue generation.