NCR Ride Bookings

SQL Analysis Report

Compiled by Sesethu M. Bango

On: 28 September 2025

Background

The modern world is one of constant motion. Human beings and goods need to move from one location to another on a continuous basis. E-hailing has become a very convenient and cost-effective part of the overall transportation industry.

To gain some insight into the e-hailing industry we will be analysing the NCR Ride bookings dataset. Through this analysis we will be able to better understand the ride patterns, cancellations, fares and operational metrics of this e-hailing service.

Summary of Findings

- In cleaning, more than 32% of all rows were removed because of null values. That could affect insights and findings.
- This loss of rows was mitigated by re-introducing those rows when they were necessary for analysis.
- 66% of bookings are valued at 300 Rupees or more.
- The Auto vehicle type is the most booked vehicle type.
- The most prevalent values in the *Booking Status* field are *Completed* at 62% followed by *Cancelled by driver* at 18%.
- A Pearson correlation of 0.4161, for *Ride Distance* vs *Booking Value.* Indicates a modest positive correlation between the two.
- Completion rates range between 61% and 62.5%. With Uber XL, marginally having the highest value.
- From table 13, we see that the most profitable route is the *New Delhi Railway Station -> Rajouri Garden* route.
- Hours 17, 18 and 19 make up the 3 highest ride counts, in comparison to all the other hours of the day.
- Monthly customer retention counts are very low compared to cohort size. This is the case with all the Cohort months; it's certainly a cause for concern.

Reflection

The duration of query executions was at times concerning. For more complex queries with window functions and nested SELECT statements it would sometimes take more than 5 minutes to complete the execution. Perhaps it's a hardware issue.

1. Data Collection

The file *NCR_Ride_bookings - NCR_Ride_bookings.csv* was downloaded from the prescribed website. The table *bookings_raw* was then imported into the newly created Database *Ride_booking* into the created schema *RIDE*.

Table 1 shows the missing value percentage per column in *bookings_raw.* The columns with the most missing values, as a percentage of total number of rows, are highlighted in darker blue.

Table 1

Missing value percentage per column			
Date_null_percent 0.00000000	**Time_null_percent** 0.00000000	**Booking_ID_null_percent** 0.00000000	**Booking_Status_null_percent** 0.00000000
Customer_ID_null_percent 0.00000000	**Vehicle_Type_null_percent** 0.00000000	**Pickup_Location_null_percent** 0.00000000	**Drop_Location_null_percent** 0.00000000
Avg_VTAT_null_percent 7.00000000	**Avg_CTAT_null_percent** 32.00000000	**Cancelled_Rides_C_null_percent** 93.00000000	**Reason_C_null_percent** 93.00000000
Cancelled_Ride_D_null_percent 82.00000000	**Reason_D_null_percent** 82.00000000	**Incomplete_Rides_null_percent** 94.00000000	**Incomplete_Rides_Reason_null_percent** 94.00000000
Booking_Value_null_percent 32.00000000	**Ride_Distance_null_percent** 32.00000000	**Driver_Ratings_null_percent** 38.00000000	**Customer_Rating_null_percent** 38.00000000
Payment_Method_null_percent 32.00000000			

From here a decision must be made as to what to do with the null values. The advised approach is to create and populate a new table, *bookings* with rows where critical fields are not NULL. The question then arises, what constitutes a critical field? Certainly *Booking*_Value is a critical field.

Once the NULL values have been removed, missing value percentages are once again queried. This time, the missing values are mostly limited to the *cancelled* and *incomplete* fields, thus making the table cleaner.
At a later stage when analysing cancelled and incomplete rides, the table *bookings_raw* will be used. This is because most, if not all the removed rows are vital to the incomplete and cancelled rides fields.

Row count before cleaning: 150 000,
Row count after cleaning: 102 000.

2. Data Preparation

A query to remove duplicates in the *Booking_ID* field was executed. 548 rows were found to be duplicates and were removed.

Some feature engineering was then done, where certain columns were consolidated and others edited. The following columns were the result of such action:
Day_Of_Week, Hour_Of_Day, pickup_ts, Route and *Payment_Method_Norm*

The resulting table after this data preparation process is *bookings_clean*, which is placed in the schema *CLEAN*.

3. Data Exploration

In this section, 4 exploratory queries were created.

The first, a **distribution analysis**. Table 2 shows the booking value distribution, where the data is organised into bins. Here we see the overwhelming distribution of booking values is in the bin >=300. Meaning 66% of bookings are valued at 300 Rupees or more. The city New Delhi is mentioned as destination later in this document; hence the assumption of currency is the Indian Rupee.

Table 2

Booking_Value_Distribution	CountValues	Percentages
>=300	66903	65.95000000
100 - 199.99	14156	13.95000000
200 - 299.99	14102	13.90000000
<100	6291	6.20000000

Secondly, a **categorical analysis**. A count of the top 10 **vehicle types** and a count of top 10 **booking statuses**. Each of these should have row wise and column wise percentages.
Table 3 shows each vehicle type, its count and its percentage of the total, from highest to lowest. The Auto vehicle type is the most booked vehicle type.

Table 3

Vehicle_Type	Count_VehicleType	Percent_VehicleType
Auto	25277	24.92000000
Go Mini	20245	19.96000000
Go Sedan	18236	17.98000000
Bike	15271	15.05000000
Premier Sedan	12253	12.08000000
eBike	7141	7.04000000
Uber XL	3029	2.99000000

Table 4 shows **booking status**, its count and its percentage of the total, also from highest to lowest. Here we see that there are only two entries that result. This doesn't tell the whole story. A lot of data was lost during cleaning. That data is vital to understanding this specific analysis.
The table RIDE.bookings_raw should be used here. After duplicates have been removed, table 5 is what results.
Table 5 shows a more complete analysis in terms of the *booking_Status* field. The most prevalent statuses being, *Completed* which is the largest by far, followed by *Cancelled by Driver*.

Table 4

Booking_Status	Count_Booking_Status	Percent_VehicleType
Completed	92504	91.18000000
Incomplete	8948	8.82000000

Table 5

Booking_Status	Count_Booking_Status	Percent_VehicleType
Completed	92246	62.01000000
Cancelled by Driver	26786	18.01000000
Cancelled by Customer	10409	7.00000000
No Driver Found	10407	7.00000000
Incomplete	8919	6.00000000

Next is a **relationship analysis** of *Ride Distance* vs *Booking Value.* A Pearson correlation calculation was done, using SQL aggregates. Table 6 shows the result. A Pearson correlation of 0.4161 is neither strong nor weak, it's somewhere in the middle. It means there is a positive correlation between ride distance and booking value, but it's not a strong correlation.
Short but expensive trips, and visa versa, will be anomalies.

Table 6

Pearson_correlation_RideDistance_vs_BookingValue
0,4161

Finally, we have a comparative analysis of *Booking Value* by *Payment Method.* This is done in the form of a 5 number summary.
Table 7 shows a 5 number summary of Booking Value by each type of payment method. The type of payment method used has no bearing on booking value. Regardless of payment method, the corresponding number summaries are all essentially equal.

Table 7

Payment_Method_Norm	Min_BookingVal	Q1_BookingVal	Q2_BookingVal	Q3_BookingVal	Max_BookingVal
CASH	50	234	417	686	4133
DEBITCARD	50	234,75	413	685	4228
CREDITCARD	50	241	415	683	3985
UBERWALLET	50	230	413	685	4202
UPI	50	233	413	692	4277

4. Applied Statistical Analysis

Some important insights have been gleaned thus far. Attention will now turn to an applied statistical analysis, starting with **Descriptive statistics**.
For the fields *Booking_Value* and *Ride_Distance*, the mean, median and standard deviation were found, results are summarised in tables 8 and 9.

Table 8

Median_Booking_Value	Mean_Booking_Value	StdDev_Booking_Value
414	508	395,9

Table 9

Median_Ride_Distance_Value	Mean_Ride_Distance	StdDev_Ride_Distance
23,72	24,64	14

Next is a **correlation analysis** which involves the exact same fields as in the relationship analysis in section 3. Namely, *Ride Distance* vs *Booking Value.* We use the same approach as we did there. The result is the same as in section 3. **Pearson Correlation = 0.4161**.

Outlier detection is another vital analysis point. Table 11 shows the top 20 outlier rows based on *Booking Value,* where *Booking_Value > Q3 + 1.5*IQR,* ordered from highest to lowest.
Beyond this top 20, there are 3418 total outliers based on this booking value calculation.

Table 11

Booking_ID	Route	Booking_Value
CNR7954315	Saidulajab -> Netaji Subhash Place	4277
CNR1798489	Ashram -> Patel Chowk	4228
CNR8487909	Welcome -> Jama Masjid	4220
CNR5182516	Subhash Nagar -> Laxmi Nagar	4202
CNR7356012	IMT Manesar -> Sarojini Nagar	4133
CNR8849175	Ashok Vihar -> Basai Dhankot	4109
CNR5553074	GTB Nagar -> Narsinghpur	4093
CNR8715944	Karol Bagh -> Pitampura	4088
CNR7652202	AIIMS -> Bhikaji Cama Place	4060
CNR8875064	Dwarka Mor -> Seelampur	4044
CNR1017046	Paschim Vihar -> Malviya Nagar	4032
CNR7877701	Ghaziabad -> Samaypur Badli	4026
CNR5928940	Badarpur -> Chirag Delhi	4008
CNR5245395	Noida Sector 18 -> Indirapuram	3985
CNR3507687	Rajouri Garden -> Punjabi Bagh	3984
CNR1704323	DLF City Court -> Hauz Rani	3978
CNR4802931	Jama Masjid -> IFFCO Chowk	3962
CNR6309807	Udyog Vihar Phase 4 -> Gwal Pahari	3942
CNR3578487	Pitampura -> Vishwavidyalaya	3921
CNR1018014	Central Secretariat -> Anand Vihar	3917

5. Advanced Analysis

The following section is for a more advanced analysis approach, where deeper insights will be gained.
First up is an analysis of **operational performance**, where the ride completion rate will be interrogated. **Completion rate** = *Completed / (Completed + Cancelled + Incomplete)*.
In table 12 we observe that the completion rates range between 61% and 62.5%. With Uber XL, marginally having the highest value. We cannot say however that any one vehicle type has a

significantly larger or smaller ride completion rate than any other. They all have essentially the same completion rate.

Table 12

Vehicle_Type	Completion_Rate
Uber XL	62.460000000000
Bike	62.360000000000
Go Mini	62.260000000000
Premier Sedan	62.140000000000
eBike	62.060000000000
Auto	61.850000000000
Go Sedan	61.470000000000

Looking now at route profitability, we identify the **top 10 Routes** by **total Booking value**, and the average **Booking value** and **ride count** per route.
We see in table 13 that the most profitable route is the *New Delhi Railway Station -> Rajouri Garden* route. The route with the highest ride count is Ambience Noida -> Vaishali with 11 rides.

Table 13

Route	Sum_Booking_Value	Avg_Booking_Value	Ride_Count
New Delhi Railway Station -> Rajouri Garden	9559	1593	6
Cyber Hub -> Gurgaon Railway Station	9348	934	10
Nirman Vihar -> Vatika Chowk	9284	1856	5
Ashok Vihar -> Basai Dhankot	9280	1031	9
Anand Vihar ISBT -> Noida Film City	8960	1280	7
Mayur Vihar -> Samaypur Badli	8588	954	9
Model Town -> Jahangirpuri	8540	1067	8
Ambience Mall -> Akshardham	8518	774	11
Greater Noida -> Jor Bagh	8252	1031	8
Noida Extension -> Vaishali	8202	1025	8

In terms of **cancellation forensics**, within the cancelled bookings – the **top 5 cancellation reasons** are shown in table 14 and 15. Shown in these tables are cancel reasons for customer and driver. We are using the table *bookings_raw* for this problem as we are dealing with cancelled and incomplete rides.
Within each query we intentionally leave out 'Unspecified' rows. These are scenarios where there is either no reason given for cancellation, or it's a completed ride.
In both Customer and Driver tables, there is no one reason that stands out among the rest. Each reason seems to be just a likely as any other to be the reason for cancellation.

Table 14

Reason_for_cancelling_by_Customer	Count_Value	Percentage
Wrong Address	2347	22,55
Change of plans	2329	22,37
Driver is not moving towards pickup location	2315	22,24
Driver asked to cancel	2276	21,87
AC is not working	1142	10,97

Table 15

Driver_Cancellation_Reason	Count_Value	Percentage
Customer related issue	6779	25,31
The customer was coughing/sick	6694	24,99
Personal & Car related issues	6673	24,91
More than permitted people in there	6640	24,79

Homing in on service levels (time windows), we see in table 16 the **average Avg_VTAT** and **average Avg_CTAT** grouped by **Hour of Day,** and Ride count ordered in descending order**.**
The hour 18 has the highest Ride count of any hour. In fact, hours 17, 18 and 19 make up the 3 highest ride counts. These times fall within Rush-hour, or just after it, so it makes sense that these would be the busiest times of day.
The *average Avg_VTAT* and *average Avg_CTAT* remains the same no matter the time of day. There's no definitive information that can be gained from that, except to know the value of these averages.

Table 16

Hour_Of_Day	Avg_Avg_VTAT	Avg_Avg_CTAT	Ride_Count
18	8,4	29,3	12296
17	8,5	29,1	10960
19	8,4	29,1	10950

Lastly, in this advanced analysis section, the **customer cohorts** and **churn** were analysed.
Table 17 shows the monthly cohort breakdown in the year 2024. *Cohort_Month* represents the first booking by each customer in the *Cohort_Size* column. *Retained_Count* is the number of customers who booked in the month proceeding their initial booking.
We see here that retention counts are very low compared to cohort size. This is the case with all the Cohort months; it's certainly a cause for concern.

With regard to churn risk, comparing the years 2023 and 2024, all the data in the table *bookings_raw* is for the year 2024. Thus, it's impossible to determine churn risk for the scenario in question.

Table 17

Cohort_Month	Cohort_Size	Retained_Count
Jan-24	8707	9
Feb-24	7990	6
Mar-24	8627	8
Apr-24	8283	11
May-24	8578	8
Jun-24	8416	5
Jul-24	8646	6
Aug-24	8404	4
Sep-24	8167	5
Oct-24	8557	7
Nov-24	8325	8
Dec-24	8210	0

Conclusion

The analysis done on and insights gained from the NCR Ride Bookings dataset will aid in understanding the business better. Areas of weakness have been identified. Plans can then be put in place to try improve, knowing exactly where the issues lie.
The better performing parts of the business can also be identified, lessons can be learnt, and improvements can be made across the board.
From these insights one can the home in on a certain subset of data, to gain even more actionable intel.