SQL ANALYSIS

NCR RIDE BOOKINGS

Prepared For :

zAIO iNSTITUTE OF TECHNOLOGY

BY

Lee-Anne van der Merwe

Data Collection

For the NCR Ride bookings analysis, the data was collected, and the flat file was loaded onto Microsoft SQL Server with appropriate datatypes chosen. The missing values per column were then calculated as a percentage using the SQL query in Figure 1A. Followed by this, all critical fields which excluded rows with NULL values were placed in a table called ‘Bookings’ and the row count before and after were recorded in Figure 1B.

Figure 1A

Results | Messages

	MV_DATE	MV_TIME	BOOKING_ID	MV_BOOKING_STATUS	MV_CUSTOMER_ID	MV_VEHICLE_TYPE	MV_PICKUP_LOCATION	MV_DROP_LOCATION	MV_AVG_VTAT	MV_CANCELLED_BY_CUSTOMER	MV_CANCEL_REASON_CUSTOMER	MV_CANCELLED_BY_DRIVER	MV_CANCEL_REASON_DRIVER	MV_INCOMPLETE_RIDES	MV_INCOMPLETE_RIDES
1	0	0	0	0	0	0	0	0	7	93	93	82	82	94	94

	DROP_LOCATION	MV_AVG_VTAT	MV_CANCELLED_BY_CUSTOMER	MV_CANCEL_REASON_CUSTOMER	MV_CANCELLED_BY_DRIVER	MV_CANCEL_REASON_DRIVER	MV_INCOMPLETE_RIDES	MV_INCOMPLETE_RIDES_REASON	MV_BOOKINGS_VALUE	MV_RIDE_DISTANCE	MV_DRIVER_RATINGS	MV_CUSTOMER_RATING	MV_PAYMENT_METHOD
1		7	93	93	82	82	94	94	32	32	38	38	32

Figure 1B

Before

20	2024-05-24	19:53:57.0000000	CNR9465840	Cancelled by Driver	CID9046501	eBike	Pitampura	Rajiv Nagar	10.30	NULL	NULL	NULL	1	Customer related issue	NULL	NULL	NULL
21	2024-12-14	21:27:04.0000000	CNR9843229	Completed	CID4929303	Go Sedan	Punjabi Bagh	Mansarovar Park	6.10	33.70	NULL	NULL	NULL	NULL	NULL	NULL	264.00
22	2024-07-29	06:28:22.0000000	CNR7911434	Completed	CID8465004	Auto	Greater Noida	Botanical Garden	7.00	21.30	NULL	NULL	NULL	NULL	NULL	NULL	183.00

Query executed successfully. DELL-LEE22\SQLEXPRESS (16.0... DELL-LEE22\Dell (51) BOOKINGS_PROJECT 00:00:01 150 000 rows

150 000 rows

After

21	2024-11-03	12:24:59.0000000	CNR8779845	Completed	CID7444351	Auto	Faridabad Sector 15	Malviya Nagar	8.90	31.10	NULL	NULL	NULL	NULL	NULL	NULL	2847.00	39.21
22	2024-11-15	05:53:55.0000000	CNR3871541	Completed	CID9849259	Go Mini	Mundka	Preet Vihar	8.90	41.00	NULL	NULL	NULL	NULL	NULL	NULL	282.00	8.36

Query executed successfully. DELL-LEE22\SQLEXPRESS (16.0... DELL-LEE22\Dell (51) BOOKINGS_PROJECT 00:00:01 102 000 rows

102 000 rows

SQL Query for Data Collection

```sql
--MISSING VALUE PERCENTAGE PER COLUMN
USE BOOKINGS_PROJECT
SELECT
(COUNT(*)- COUNT(Date))*100/COUNT(*) AS MV_DATE,
(COUNT(*)- COUNT(Time))*100/COUNT(*) AS MV_TIME,
(COUNT(*)- COUNT(Booking_ID))*100/COUNT(*) AS BOOKING_ID,
(COUNT(*)- COUNT(Booking_Status))*100/COUNT(*) AS MV_BOOKING_STATUS,
(COUNT(*)- COUNT(Customer_ID))*100/COUNT(*) AS MV_CUSTOMER_ID,
(COUNT(*)- COUNT(Vehicle_Type))*100/COUNT(*) AS MV_VEHICLE_TYPE,
(COUNT(*)- COUNT(Pickup_Location))*100/COUNT(*) AS MV_PICKUP_LOCATION,
(COUNT(*)- COUNT(Drop_Location))*100/COUNT(*) AS MV_DROP_LOCATION,
(COUNT(*)- COUNT(Avg_VTAT))*100/COUNT(*) AS MV_AVG_VTAT,
(COUNT(*)- COUNT(Cancelled_Rides_by_Customer))*100/COUNT(*) AS MV_CANCELLED_BY_CUSTOMER,
(COUNT(*)- COUNT(Reason_for_cancelling_by_Customer))*100/COUNT(*) AS MV_CANCEL_REASON_CUSTOMER,
(COUNT(*)- COUNT(Cancelled_Rides_by_Driver))*100/COUNT(*) AS MV_CANCELLED_BY_DRIVER,
(COUNT(*)- COUNT(Driver_Cancellation_Reason))*100/COUNT(*) AS MV_CANCEL_REASON_DRIVER,
(COUNT(*)- COUNT(Incomplete_Rides))*100/COUNT(*) AS MV_INCOMPLETE_RIDES,
(COUNT(*)- COUNT(Incomplete_Rides_Reason))*100/COUNT(*) AS MV_INCOMPLETE_RIDES_REASON,
(COUNT(*)- COUNT(Booking_Value))*100/COUNT(*) AS MV_BOOKINGS_VALUE,
(COUNT(*)- COUNT(Ride_Distance))*100/COUNT(*) AS MV_RIDE_DISTANCE,
(COUNT(*)- COUNT(Driver_Ratings))*100/COUNT(*) AS MV_DRIVER_RATINGS,
(COUNT(*)- COUNT(Customer_Rating))*100/COUNT(*) AS MV_CUSTOMER_RATING,
(COUNT(*)- COUNT(Payment_Method))*100/COUNT(*) AS MV_PAYMENT_METHOD
FROM BOOKINGS_RAW;
--Row count before cleaning
SELECT * FROM BOOKINGS_RAW;
--ROW_COUNT=150 000 ROWS
```

```sql
--CREATING THE BOOKINGS COLUMN WITH ONLY VALID ROW DATA
SELECT
    [Date], [Time], [Booking_ID], [Booking_Status], [Customer_ID], [Vehicle_Type],
    [Pickup_Location], [Drop_Location], [AVG_VTAT], [Avg_CTAT],
    [Cancelled_Rides_by_Customer], [Reason_for_cancelling_by_Customer],
    [Cancelled_Rides_by_Driver], [Driver_Cancellation_Reason],
    [Incomplete_Rides], [Incomplete_Rides_Reason], [Booking_Value],
    [Ride_Distance], [Driver_Ratings], [Customer_Rating], [Payment_Method]
INTO BOOKINGS_PROJECT.dbo.BOOKINGS
FROM BOOKINGS_PROJECT.dbo.BOOKINGS_RAW
WHERE
    [Date] IS NOT NULL
    AND [Time] IS NOT NULL
    AND [Booking_ID] IS NOT NULL
    AND [Booking_Status] IS NOT NULL
    AND [Customer_ID] IS NOT NULL
    AND [Vehicle_Type] IS NOT NULL
    AND [Pickup_Location] IS NOT NULL
    AND [Drop_Location] IS NOT NULL
    AND[Booking_Value] IS NOT NULL
    AND [Ride_Distance] IS NOT NULL AND [Payment_Method] IS NOT NULL;
--ROW COUNT AFTER CLEANING
SELECT * FROM BOOKINGS
-- ROW COUNT= 102 000
```

Data Preparation

Data preparation focused on removing duplicates, engineering features from the pickup timestamp (extracting day of week and hour), creating a route column, and normalising payment methods. The SQL queries used are shown in Figure 2A, and the resulting engineered columns were added to the ‘Bookings_clean’ table .

Reflection: Found the data engineering interesting and did learned a lot about Datepart etc.

Figure 2A

```sql
USE BOOKINGS_PROJECT;
GO
--DUPLICATE ANALYSIS
WITH RankedBookings AS (
    SELECT *,
           ROW_NUMBER() OVER (
               PARTITION BY Booking_ID
               ORDER BY [Date], [Time]
           ) AS rn
    FROM BOOKINGS
)
SELECT COUNT(*) AS DuplicatesRemoved
FROM RankedBookings
WHERE rn > 1;

SELECT * FROM BOOKINGS;

--CREATING A PICKUP TIMESTAMP FROM DATE AND TIME
ALTER TABLE BOOKINGS
ADD pickup_ts DATETIME;
UPDATE BOOKINGS
SET pickup_ts= CAST([Date] AS DATETIME) + CAST([Time] AS DATETIME);

--DERIVE DAY OF THE WEEK AND HOUR OF THE DAY FROM TIMESATMP
ALTER TABLE BOOKINGS
ADD Day_of_Week AS DATENAME (WEEKDAY, [pickup_ts]);
ALTER TABLE BOOKINGS
ADD Hour_of_Day AS DATEPART (HH, [pickup_ts]);

--CREATE A ROUTE COLUMN SHOWING PICKUP AND DROP-OFF ROUTE
ALTER TABLE BOOKINGS
ADD Route AS Pickup_Location + ' -> ' + Drop_Location;

--CREATING A PAYMENT METHOD COLUMN WHERE ITS NORMALISED TO UPPERCASE LETTERS AND REMOVING ANY SPACES
ALTER TABLE BOOKINGS
ADD Payment_Method_Norm AS UPPER(REPLACE (TRIM([Payment_Method]),'',''));

--CREATING A DUPLICATE TABLE CALL BOOKINGS_CLEAN
SELECT
[PICKUP_TS],
[DAY_OF_WEEK],
[HOUR_OF_DAY],
[ROUTE],
[PAYMENT_METHOD_NORM]
INTO BOOKINGS_CLEAN
FROM BOOKINGS;
```

Table: Bookings_Clean

165 % | 1 | 0

Results | Messages

	PICKUP_TS	DAY_OF_WEEK	HOUR_OF_DAY	ROUTE	PAYMENT_METHOD_NORM
1	2024-11-29 18:01:39.000	Friday	18	Shastri Nagar -> Gurgaon Sector 56	UPI
2	2024-08-23 08:56:10.000	Friday	8	Khandsa -> Malviya Nagar	DEBIT CARD
3	2024-10-21 17:17:25.000	Monday	17	Central Secretariat -> Inderlok	UPI
4	2024-09-16 22:08:00.000	Monday	22	Ghitorni Village -> Khan Market	UPI
5	2024-02-06 09:44:56.000	Tuesday	9	AIIMS -> Narsinghpur	UPI
6	2024-06-17 15:45:58.000	Monday	15	Vaishali -> Punjabi Bagh	UPI
7	2024-03-19 17:37:37.000	Tuesday	17	Mayur Vihar -> Cyber Hub	UPI
8	2024-12-16 19:06:48.000	Monday	19	Rohini -> Adarsh Nagar	CASH
9	2024-06-14 16:24:12.000	Friday	16	Udyog Bhawan -> Dwarka Sector 21	CASH
10	2024-09-11 19:29:39.000	Wednesday	19	Malviya Nagar -> Ghitorni Village	UPI
11	2024-10-18 18:28:53.000	Friday	18	Madipur -> GTB Nagar	UPI

Data Exploration

Next, data exploration was tackled by doing the following analysis: 1) Distribution, 2) Categorical, 3) Relationship, and 4) Comparative.
Starting with the distribution analysis, Figure 3A provides a fare distribution table. The results demonstrated the spread and range of fares across all bookings, highlighting key trends in the data.

Figure 3A

Results | Messages

	FARE_DISTRIBUTION_BUCKET	FARE_DISTRIBUTION_Count	Percentage
1	200-299.99	14173	13.90000000000000
2	500-599.99	7352	7.21000000000000
3	400-499.99	13958	13.68000000000000
4	<100	6323	6.20000000000000
5	100-199.99	14231	13.95000000000000
6	>=600	31752	31.13000000000000
7	300-399.99	14211	13.93000000000000

Transitioning to categorical analysis, the top 10 vehicle types were identified based on booking status. In Figure 3B, only the top seven vehicles were listed, with ‘Automatic’ vehicles at the top. Automatic vehicles were found to be the vehicle of choice among 24.98% of bookings, followed by ‘Go Mini’ cars at 19.96%. However, the vehicle of choice may be more of a driver's decision than a customer-based decision and would need to be investigated.

Figure 3B

	VEHICLE_TYPE	NUMBER_OF_BOOKINGS	PercentageWithinCategory
1	Auto	25415	24.9200000000000
2	Go Mini	20364	19.9600000000000
3	Go Sedan	18318	17.9600000000000
4	Bike	15362	15.0600000000000
5	Premier Sedan	12315	12.0700000000000
6	eBike	7181	7.0400000000000
7	Uber XL	3045	2.9900000000000

Moving on to relationship analysis, the association between a ride’s distance and the booking value was calculated using the Pearson correlation in Figure 3C. The Pearson correlation coefficient is 0.005, indicating a weak positive relationship; therefore, ride distance does not highly influence booking value.

Figure 3C

Results | Messages

	N	SUM_X	SUM_Y	SUM_XY	SUM_X_SQUARED	SUM_Y_SQUARED	correlation_coefficient
1	102000	2512975.19	51846183.00	1280259820.3800	81910110.5145	42332591695.0000	0.00517397712510193

Reflection: SQL Server does not have a corr. function, found out I can add an R package while researching how to do a correlation manually. Will definitely insert the R package in future.

Finally, for comparative analysis, the booking value was examined across different payment methods. The SQL query is shown in Figure 3D, which displays the minimum and maximum values. The 25th percentile, 50th percentile (median), and 75th percentile values can also be observed. While all payment methods had a minimum of R50, the UPI method had a maximum of R4277, which was the highest among the different payment methods. However, cash payments still had the highest median value of R417. Hence, most high-value rides were paid for with UPI, while low-value rides were paid with Uber Wallet.

Figure 3D

	Payment_Method_Norm	Min_Value	Max_Value	Q1	Q2_Median	Q3
1	CASH	50.00	4133.00	234.00	417.00	686.00
2	CREDIT CARD	50.00	3985.00	242.00	416.00	683.00
3	DEBIT CARD	50.00	4228.00	235.00	413.00	685.00
4	UBER WALLET	50.00	4202.00	230.00	413.00	685.00
5	UPI	50.00	4277.00	233.00	413.00	692.00

Applied Statistical Analysis

Next, an applied statistical analysis was done on the booking values and ride distances. The mean, median, and standard deviation for both categories are in Figure 4A. The average booking value was higher than the median at R508.30, indicating that a few outliers may be distorting the data into a positively skewed distribution. The same was observed for the ride distance, and the standard deviation for the booking value and ride distance are R395.81 and 14km. During an outlier query, the top 20 outliers were identified among the route and booking values categories - Figure 4B. The number one outlier was on the route between Saidulajab and Netaji Subhash Place, valued at R4277. The relationship between the booking values and ride distance had a very weak positive correlation, explaining why expensive trips could be anomalies.

Figure 4A

Results | Messages

	AVG_Booking_Value	AVG_RIDE_DIS
1	508.300000	24.640000

	Booking_Value_Median
1	414.000000

	Ride_Distance_Median
1	23.720000

	STD_Booking_Value	STD_RIDE_DIS
1	395.81	14

	Pearson_Correlation
1	0.00517397712581381

Figure 4B

165 % | No issues found

Results | Messages

	Booking_ID	Route	Booking_Value
1	CNR7954315	Saidulajab -> Netaji Subhash Place	4277.00
2	CNR1798489	Ashram -> Patel Chowk	4228.00
3	CNR8487909	Welcome -> Jama Masjid	4220.00
4	CNR5182516	Subhash Nagar -> Laxmi Nagar	4202.00
5	CNR7356012	IMT Manesar -> Sarojini Nagar	4133.00
6	CNR8849175	Ashok Vihar -> Basai Dhankot	4109.00
7	CNR5553074	GTB Nagar -> Narsinghpur	4093.00
8	CNR8715944	Karol Bagh -> Pitampura	4088.00
9	CNR7652202	AIIMS -> Bhikaji Cama Place	4060.00
10	CNR8875064	Dwarka Mor -> Seelampur	4044.00
11	CNR1017046	Paschim Vihar -> Malviya Nagar	4032.00
12	CNR7877701	Ghaziabad -> Samaypur Badli	4026.00
13	CNR5928940	Badarpur -> Chirag Delhi	4008.00
14	CNR5245395	Noida Sector 18 -> Indirapuram	3985.00
15	CNR3507687	Rajouri Garden -> Punjabi Bagh	3984.00
16	CNR1704323	DLF City Court -> Hauz Rani	3978.00
17	CNR4802931	Jama Masjid -> IFFCO Chowk	3962.00
18	CNR6309807	Udyog Vihar Phase 4 -> Gwal Pahari	3942.00
19	CNR3578487	Pitampura -> Vishwavidyalaya	3921.00
20	CNR1018014	Central Secretariat -> Anand Vihar	3917.00

Advanced Analysis

Finally, the following advanced analyses were queried: 1) The completion rate for each vehicle type, 2) The top 10 routes by total booking value, 3) Top 5 reasons for cancellation, 4) The three busiest hours in a day, and 5) Number of customers who booked in 2023 but not in 2024.

1)The Completion Rate for Each Vehicle Type:

The completion rate is highest for Uber XL Vehicles which could be due to the car size being popular.

Results | Messages

	Vehicle_Type	Completion_Rate	Completion_Rank
1	Uber XL	0.913957307060	1
2	Premier Sedan	0.913682501015	2
3	Bike	0.913552922796	3
4	eBike	0.912268486283	4
5	Auto	0.911076136140	5
6	Go Mini	0.910872127283	6
7	Go Sedan	0.910361393165	7

SQL Query for Completion Rate

```
USE BOOKINGS_PROJECT
--OPERATIONAL PERFORMANCE: FIND COMPLETION RATE FOR VEHICLE TYPES AND RANK THEM
ALTER TABLE BOOKINGS
DROP [Completion_Rate] ;
-- Compute completion rate per Vehicle_Type
SELECT
    Vehicle_Type,
    COUNT(CASE WHEN Booking_Status = 'Completed' THEN 1 END) * 1.0 /
    NULLIF(COUNT(CASE WHEN Booking_Status IN ('Completed', 'Cancelled', 'Incomplete') THEN 1 END), 0) AS Completion_Rate,
    RANK() OVER (
        ORDER BY
            COUNT(CASE WHEN Booking_Status = 'Completed' THEN 1 END) * 1.0 /
            NULLIF(COUNT(CASE WHEN Booking_Status IN ('Completed', 'Cancelled', 'Incomplete') THEN 1 END), 0) DESC
    ) AS Completion_Rank
FROM
    Bookings
GROUP BY
    Vehicle_Type;
```

Advanced Analysis

2) The Top 10 Routes by Total Booking Value:

The top route was between New Delhi Railway Station and Rajouri Garden. This could be because the people who travel between these routes are tourists.

165 % No issues found

Results | Messages

	Route	Ride_Count	Total_Booking_Value	Avg_Booking_Value
1	New Delhi Railway Station -> Rajouri Garden	6	9559.00	1593.170000
2	Cyber Hub -> Gurgaon Railway Station	10	9348.00	934.800000
3	Nirman Vihar -> Vatika Chowk	5	9284.00	1856.800000
4	Ashok Vihar -> Basai Dhankot	9	9280.00	1031.110000
5	Anand Vihar ISBT -> Noida Film City	7	8960.00	1280.000000
6	Mayur Vihar -> Samaypur Badli	9	8588.00	954.220000
7	Model Town -> Jahangirpuri	8	8540.00	1067.500000
8	Ambience Mall -> Akshardham	11	8518.00	774.360000
9	Greater Noida -> Jor Bagh	8	8252.00	1031.500000
10	Noida Extension -> Vaishali	8	8202.00	1025.250000

SQL Query for Top 10 Routes, AVG Booking Value and AVG Ride Count per Route

```sql
--ROUTE PROFITABILITY:IDENTIFY TOP 10 ROUTES VIA BOOKING VALUE
SELECT TOP 10
    Route,
    COUNT(*) AS Ride_Count,  -- Calculates number of rides per route
    SUM(Booking_Value) AS Total_Booking_Value,
    ROUND(AVG(Booking_Value),2) AS Avg_Booking_Value
FROM
    BOOKINGS
GROUP BY
    Route
ORDER BY
    SUM(Booking_Value) DESC;
```

3) Top 5 Reasons for Cancellation:

No data was found for both queries.

SQL Query for Top 5 Cancellations by a Customer

```
--TOP 5 CANCELLATION REASONS BY CUSTOMER
SELECT TOP 5
    ISNULL(NULLIF(LTRIM(RTRIM([Reason_for_cancelling_by_Customer])), ''), 'Unspecified') AS Cancel_Reason,
    COUNT(*) AS Reason_Count,
    ROUND(100.0 * COUNT(*) / SUM(COUNT(*)) OVER (), 2) AS Percentage
FROM
    BOOKINGS
WHERE
    [Cancelled_Rides_by_Customer] > 0
GROUP BY
    ISNULL(NULLIF(LTRIM(RTRIM([Reason_for_cancelling_by_Customer])), ''), 'Unspecified')
ORDER BY
    Reason_Count DESC;
```

SQL Query for Top 5 Cancellations by a Driver

```
--TOP 5 CANCELLATION REASONS BY Driver
SELECT TOP 5
    ISNULL(NULLIF(LTRIM(RTRIM([Driver_Cancellation_Reason])), ''), 'Unspecified') AS Cancel_Reason,
    COUNT(*) AS Reason_Count,
    ROUND(100.0 * COUNT(*) / SUM(COUNT(*)) OVER (), 2) AS Percentage
FROM
    BOOKINGS
WHERE
    [Cancelled_Rides_by_Driver] > 0
GROUP BY
    ISNULL(NULLIF(LTRIM(RTRIM([Driver_Cancellation_Reason])), ''), 'Unspecified')
ORDER BY
    Reason_Count DESC;
```

4) Service Levels- The Three Busiest Hours in a Day:

The busiest hours were during the time that most people finish work- 5pm, 6pm and 7pm.

165 % No issues found

Results Messages

	Hour_Of_Day	Ride_Count	Avg_VTAT	Avg_CTAT
1	18	7617	8.440000	30.200000
2	17	6860	8.550000	30.010000
3	19	6798	8.490000	30.030000

SQL Query for Three Busiest Days

```
-- SERVICE LEVELS: AVG VTAT AND AVG CTAT BY HOUR OF DAY + LIST 3 BUSIEST HOURS
SELECT TOP 3
    DATEPART(HOUR, CAST([Time] AS TIME)) AS Hour_Of_Day,
    COUNT(*) AS Ride_Count,
    ROUND(AVG([AVG_VTAT]), 2) AS Avg_VTAT,
    ROUND(AVG([Avg_CTAT]), 2) AS Avg_CTAT
FROM
    BOOKINGS
WHERE
    [Booking_Status] = 'Completed'
GROUP BY
    DATEPART(HOUR, CAST([Time] AS TIME))
ORDER BY
    Ride_Count DESC;
```

5) Define cohorts by first month booking:

Reflection: Could not complete the query, found it a bit hard.

6) Number of customers who booked in 2023 but not in 2024:

Found no data for any bookings in 2023. The earliest booking was made in 2024.

SQL Query for Number of Customers Booked in 2023 but not 2024

```
---LIST CUSTOMERS WHO BOOKED 2023 BUT HAD 0 BOOKINGS IN 2024
SELECT [Customer_ID]
FROM BOOKINGS
WHERE YEAR([Date]) = 2023
EXCEPT
SELECT [Customer_ID]
FROM BOOKINGS
WHERE YEAR([Date]) = 2024;
```

Reflection: Found no data in 2023, I might have removed it or it was never in the data.

Conclusion: The NCR bookings data showed low correlation between the booking value and ride distance overall. The busiest times for rides were during rush hour when most people leave work and the most popular vehicles used were automatic. The most popular route were also from the railway station which could be due to tourism or working class people with even longer travel routes. The vehicle with the highest completion rate was also the Uber XL, which is usually used among ride share occupants. In conclusion, most customers are people going home from work and the price of the ride could be due to ride sharing and not the vehicle type.