Ride Booking Analysis - Python Project Report

# Contents

1. Executive Summary
2. Project Background & Objectives
3. Data & Methods
4. Descriptive Findings
5. Customer Behaviour Insights
6. Driver Performance Evaluation
7. Operational Metrics
8. Correlation & Statistical Notes
9. Recommendations & Next Steps
10. Limitations
11. Appendix: Code snippets & Glossary

# 1. Executive Summary

This Python-based analysis covers 150000 ride booking records and provides operational insights, cancellation patterns, demand peaks, and recommendations for improving platform reliability and customer satisfaction. Key highlights are:

- Most popular vehicle type: auto (37419 bookings).

- Average booking value: 478.12

- Average ride distance: 24.64

- Driver cancellations (dataset): 27000

- Customer cancellations (dataset): 10500

- Peak cancellation windows: around 10:00 AM and 6:00 PM.

# 2. Project Background & Objectives

Objective: Use Python to replicate and expand upon the Power BI analysis, providing reproducible code, statistics, and exportable visual artifacts. Specific goals:

- Clean and standardize the dataset for reproducible analysis.

- Produce descriptive stats for vehicle types, distances, booking values, and ratings.

- Surface customer and driver cancellation patterns and reasons.

- Quantify VTAT and CTAT and their variation by vehicle type and location.

- Provide actionable recommendations and code snippets to replicate analyses.

# 3. Data & Methods

Source: Cleaned Project dataset.csv provided by the user. Analysis performed in Python using pandas and matplotlib. Key preprocessing steps:

- Merged Date and Time to create DateTime; extracted hour and day-of-week.

- Converted numeric fields and coerced non-numeric values to NaN.

- Filled or handled missing values based on column context (mean/median/NA/default flags).

- Standardized categorical strings (trim, lower/upper as needed).

- Aggregated and grouped by vehicle type, pickup location, hour and day for summaries.

# 4. Descriptive Findings

Total records: 150000

Completed rides: 93000

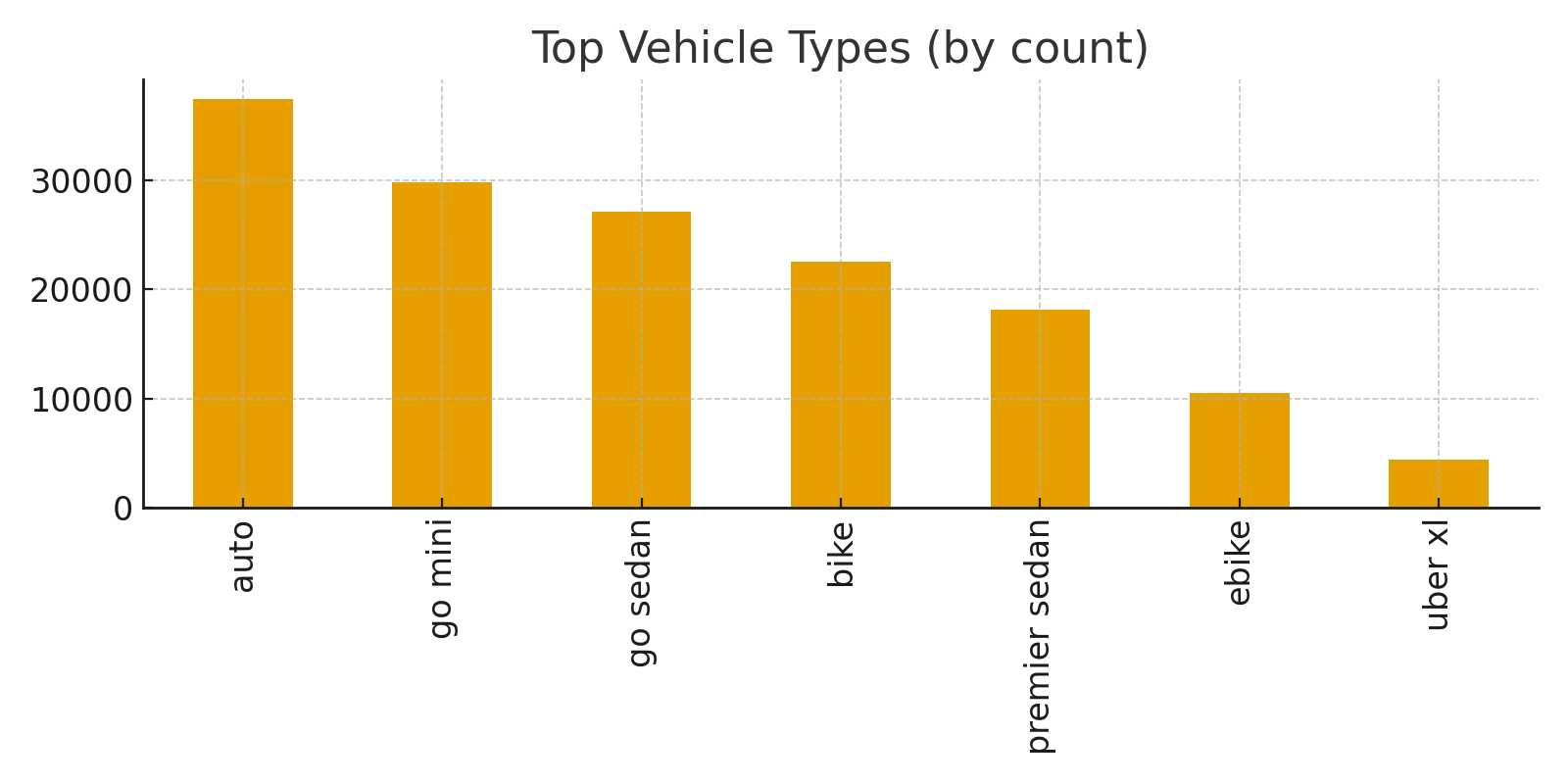
Cancelled by customers: 10500

Cancelled by drivers: 27000

## 4.1 Vehicle type popularity

|  |  |
| --- | --- |
| Vehicle Type | Count |
| auto | 37419 |
| go mini | 29806 |
| go sedan | 27141 |
| bike | 22517 |
| premier sedan | 18111 |
| ebike | 10557 |
| uber xl | 4449 |

Figure: Top Vehicle Types



## 4.2 Average ride metrics

- Average booking value: 478.12

- Average ride distance: 24.64

## 4.3 Ratings distribution

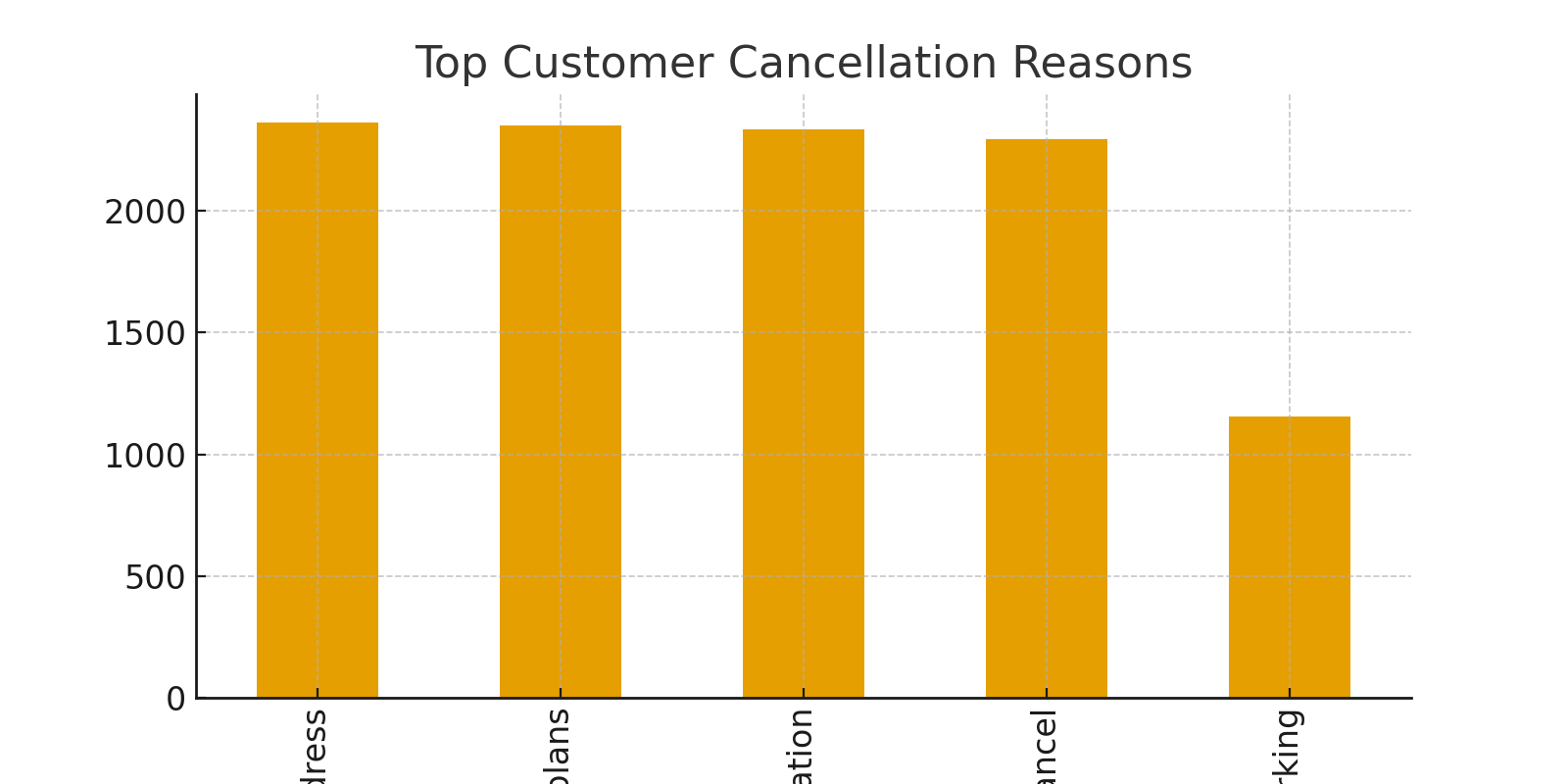
- Customer rating — count: 150000, mean: 4.44, median: 4.50

# 5. Customer Behaviour Insights

## 5.1 Cancellation reasons and frequency

* - wrong address: 2362
* - change of plans: 2353
* - driver is not moving towards pickup location: 2335
* - driver asked to cancel: 2295
* - ac is not working: 1155

Figure: Customer cancellation reasons



**5.2 Frequent cancellers (sample)**

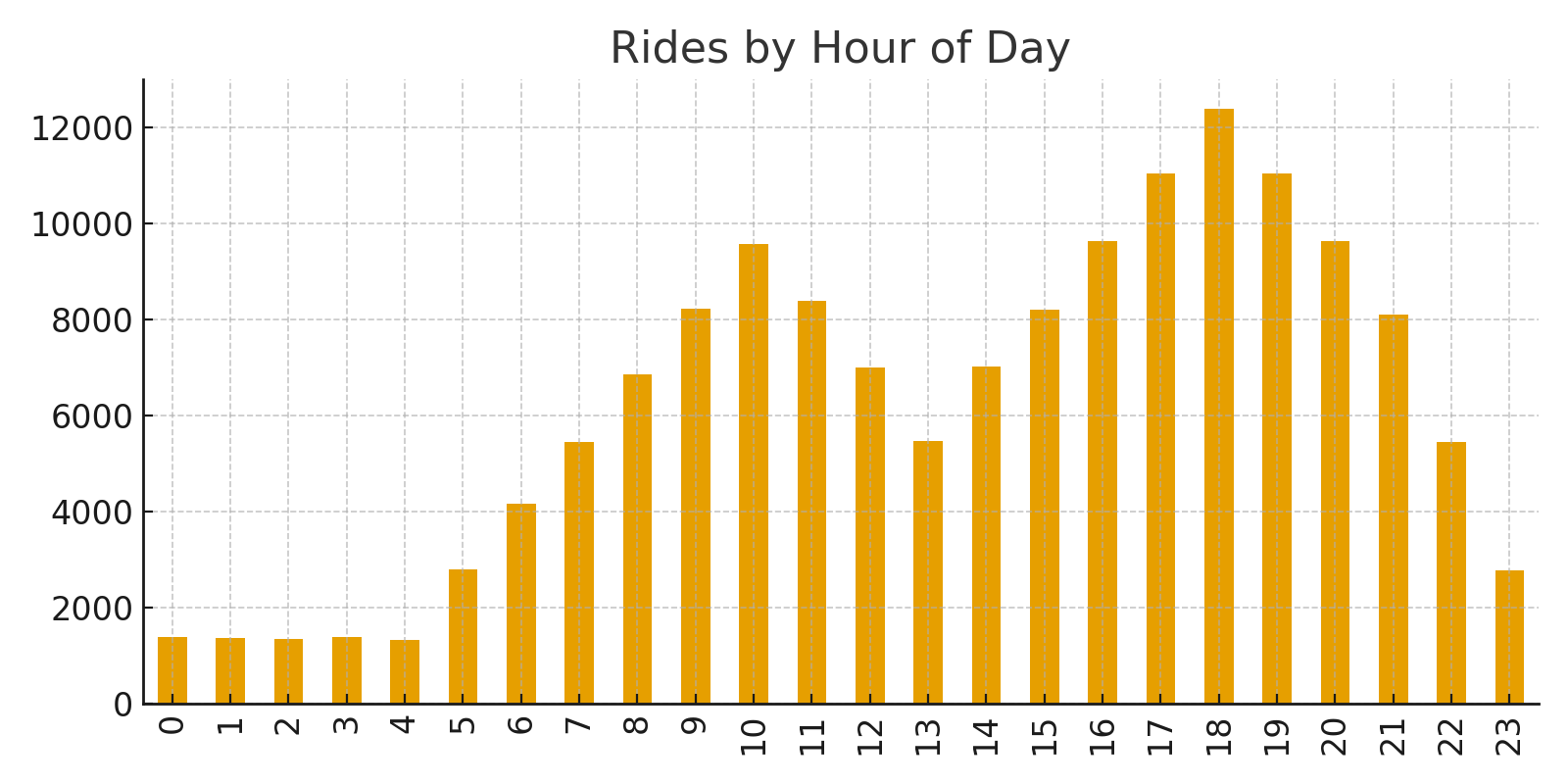
* - Customer "CID5505307": 2 cancellations
* - Customer "CID5251785": 2 cancellations
* - Customer "CID7939117": 2 cancellations
* - Customer "CID8519067": 2 cancellations
* - Customer "CID8416905": 2 cancellations
* - Customer "CID8887672": 2 cancellations
* - Customer "CID5488244": 2 cancellations
* - Customer "CID2358909": 2 cancellations
* - Customer "CID4028966": 1 cancellations
* - Customer "CID5962105": 1 cancellations

**5.3 Time-of-day and day-of-week patterns**

- Peak ride hours (top 6):

* - Hour 18: 12397 rides
* - Hour 19: 11047 rides
* - Hour 17: 11044 rides
* - Hour 16: 9633 rides
* - Hour 20: 9630 rides
* - Hour 10: 9577 rides

Figure: Rides by hour distribution



# 6. Driver Performance Evaluation

**6.1 Driver cancellations and reasons**

* - customer related issue: 6837
* - the customer was coughing/sick: 6751
* - personal & car related issues: 6726
* - more than permitted people in there: 6686

**6.2 VTAT & CTAT summary**

- VTAT: mean=8.46, median=8.46, count=150000

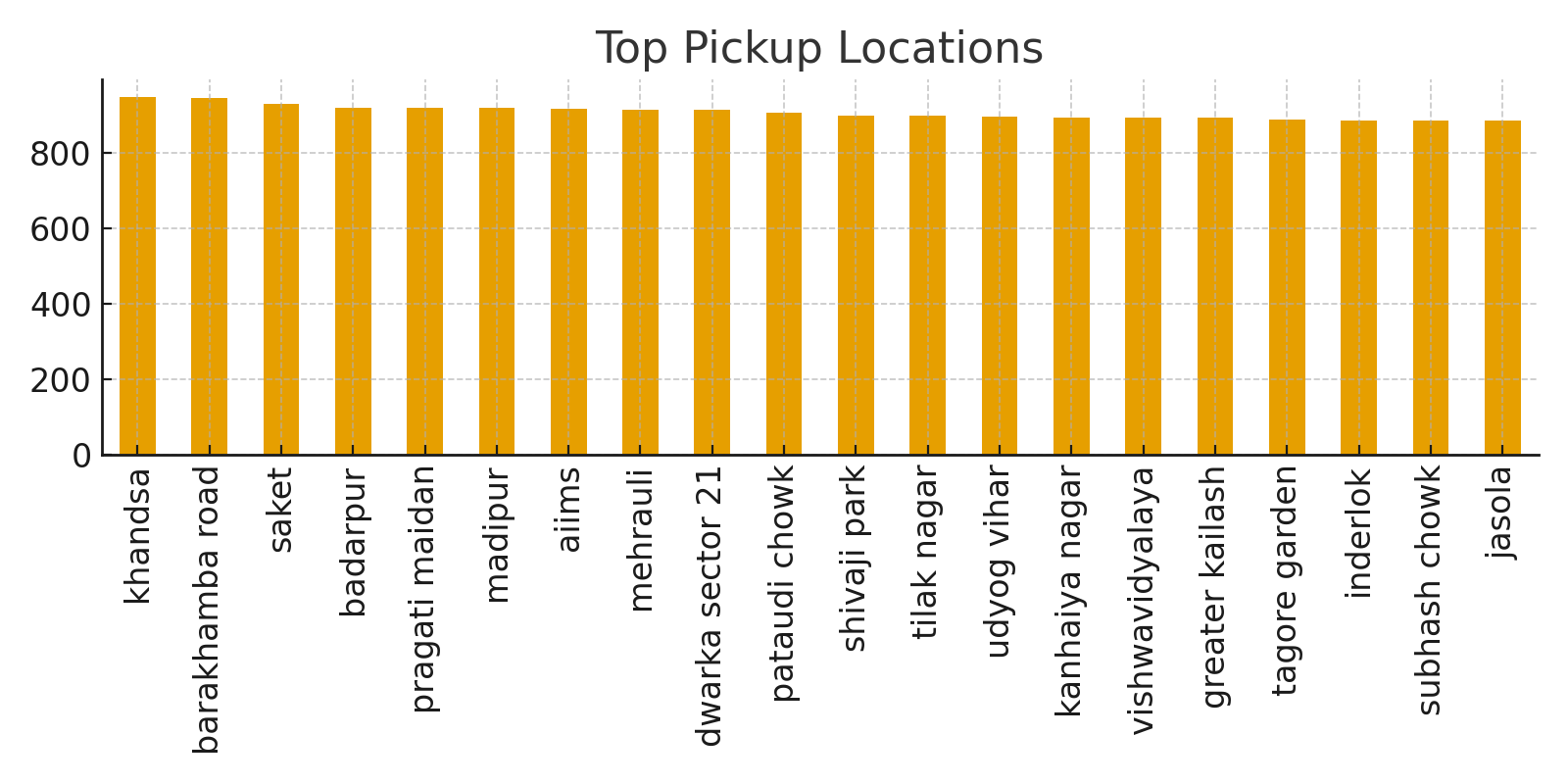
- CTAT: mean=29.15, median=29.15, count=150000

# 7. Operational Metrics

**7.1 Top pickup locations (sample)**

|  |  |
| --- | --- |
| Pickup Location | Count |
| khandsa | 949 |
| barakhamba road | 946 |
| saket | 931 |
| badarpur | 921 |
| pragati maidan | 920 |
| madipur | 919 |
| aiims | 918 |
| mehrauli | 915 |
| dwarka sector 21 | 914 |
| pataudi chowk | 907 |
| shivaji park | 900 |
| tilak nagar | 900 |
| udyog vihar | 897 |
| kanhaiya nagar | 895 |
| vishwavidyalaya | 895 |
| greater kailash | 895 |
| tagore garden | 889 |
| inderlok | 887 |
| subhash chowk | 887 |
| jasola | 887 |

Figure: Top Pickup Locations



**7.2 Peak hours and capacity planning**

- Peak demand occurs during morning commute and evening commute windows; increase driver supply and consider surge incentives.

# 8. Correlation & Statistical Notes

- Correlation matrix (sample):



- Observed correlations are generally weak between ride distance, booking value and ratings.   
**So Basically There is no Correlation between Booking value, Customer satisfaction and Ride distance.**

# 9. Recommendations & Next Steps

- Standardize pickup/drop location values using fuzzy matching or a reference table.

- Add structured cancellation reasons (dropdowns) to reduce free-text noise.

- Instrument unique driver IDs for driver-level KPIs and dashboards.

- Implement pre-ride confirmations and small friction for repeat cancellers (A/B test first).

- Collect geolocation (lat/long) to enable spatial hotspot analysis.

# 10. Limitations

- Driver ID missing: prevents per-driver metrics.

- Free-text reason fields: noisy and need cleaning/standardization.

- Location naming inconsistencies: impacts hotspot identification.

- Potential timezone ambiguity in Date/Time fields.

# 11. Appendix: Code snippets & Glossary

**11.1 Example Python snippets used in this analysis (pandas):**

- df['DateTime'] = pd.to\_datetime(df['Date'].astype(str) + ' ' + df['Time'].astype(str))

- df['hour'] = df['DateTime'].dt.hour

- vehicle\_counts = df['Vehicle Type'].value\_counts()

- corr = df[['Ride Distance','Booking Value','Customer Rating']].corr()

**11.2 Glossary:**

xssVTAT = Vehicle Turnaround Time; CTAT = Customer Turnaround Time.