Rapido Pune Data Analysis Project

Overview

This project involves analyzing a synthetic dataset for Rapido Pune rides from January 1 to January 31, 2024. The primary goal was to explore the types of questions that should be asked in such analyses and identify key metrics to consider when studying visualizations.

The project includes:

1. Five Power BI dashboard snippets showcasing key insights (e.g., Overview, Vehicle Type, Revenue, Cancellations and Ratings).
2. A PDF containing SQL query snippets used for the analysis.
3. A Python file used for preprocessing and exploratory analysis.

Dataset Details

- **Source**: The dataset was synthetically generated using a Large Language Model (ChatGPT) to ensure dynamic entries across various columns.
- **Prompt**: The LLM prompt ensured that all necessary metrics were included. Without this, the data would have been overly organized and lacked realistic variations.

The dataset consists of the following 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. `Reason for Cancelling by Customer`
12. `Cancelled Rides by Driver`
13. `Reason for Cancelling by Driver`
14. `Incomplete Rides`
15. `Incomplete Rides Reason`
16. `Booking Value`
17. `Ride Distance`
18. `Driver Ratings`
19. `Customer Rating`
20. `Payment Method`

Data Preprocessing

Upon examining the dataset, I identified several null values in specific columns. These null values were addressed as follows:The dataset required some preprocessing in Python using Jupyter Notebook. Key steps included:

- Addressing null values in the following columns:
    - `Avg CTAT`: Filled with `0`.
    - `Avg VTAT`: Filled with `0`.
    - `Reason for Cancelling by Customer`: Filled with `Not Applicable`.
    - `Reason for Cancelling by Driver`: Filled with `Not Applicable`.
    - `Incomplete Rides Reason`: Filled with `Not Applicable`.
    - `Booking Value`: Filled with `0`.
    - `Ride Distance`: Filled with `0`.
    - `Driver Ratings`: Filled with `0`.
    - `Customer Rating`: Filled with `0`.

Visualizations

Python Visualizations

Visualizations in Python were used to verify the dataset's alignment with the LLM prompt. These were non-interactive and primarily served to check data distribution and consistency.

Power BI Dashboards

Power BI dashboards were created as the final deliverable. These dashboards are interactive, intuitive, and suitable for presentation to non-technical stakeholders.

Why both Python and Power BI visualizations?

- Python visualizations were used for data validation and exploratory analysis.
- Power BI dashboards provide an interactive and user-friendly interface for presenting insights.

Inspiration and Variations

This project is inspired by the **OLA Data Analyst Project by Top Varsity**. However, I introduced several variations to make the project unique and more aligned with my goals.

This document, along with the dashboards, SQL snippets, and Python files, provides a comprehensive overview of the project. While the dataset is synthetic, it serves as a robust foundation for understanding key metrics and visualization techniques for ride-hailing companies.