Data Analysis Report on Taxi Rides

April 23, 2025

1 Introduction

This report details the steps and logic applied to analyze taxi ride data. The dataset used is a '.parquet' file of NYC taxi trips in January 2025. The goal was to clean the data and prepare it for further modeling and analysis, such as agent assignment and routing.

2 Dataset

2.1 Dataset Description

The dataset represents trip records from High Volume For-Hire Services (HVFHS) in New York City. Each row is a unique ride entry dispatched by a major licensed FHV base (Uber, Lyft, Via, or Juno). The core fields included are:

- **hvfhs_license_num**: Identifier for the dispatching company (e.g., HV0003 = Uber).
- **pickup_datetime**, **dropoff_datetime**, **request_datetime**, **on_scene_datetime**: Timestamps related to the ride request and pickup/dropoff events.
- **PULocationID**, **DOLocationID**: Location IDs referencing TLC taxi zones.
- **trip_miles**, **trip_time**: Distance (in miles) and duration (in seconds) of the trip.
- **base_passenger_fare**: Fare before any taxes or surcharges.
- **airport_fee**, **tolls**, **sales_tax**, **congestion_surcharge**, **bcf**: Breakdown of additional ride-related fees.
- **tips**, **driver_pay**: Monetary values related to tipping and net driver income.
- **shared_request_flag**, **shared_match_flag**: Flags indicating whether the ride was requested/shared.
- **wav_request_flag**, **wav_match_flag**: Indicates accessibility needs and accommodations.

- **cbd_congestion_fee**: Additional congestion fee effective from Jan 2025 for central business district rides.

subsection*Initial Dataset Observations Prior to modeling, a number of data cleaning operations were applied:

- Removed trips with missing datetime values and fares.
- Filtered out entries with invalid or unknown taxi zone identifiers.
- Converted date/time fields to Pandas `datetime` objects for temporal analysis.

Data Coverage: The dataset reflects rides dispatched by major FHV platforms in January 2025, with hundreds of thousands of records. Each record captures logistical, financial, and accessibility-related ride metadata.

3 Exploration Data Analysis

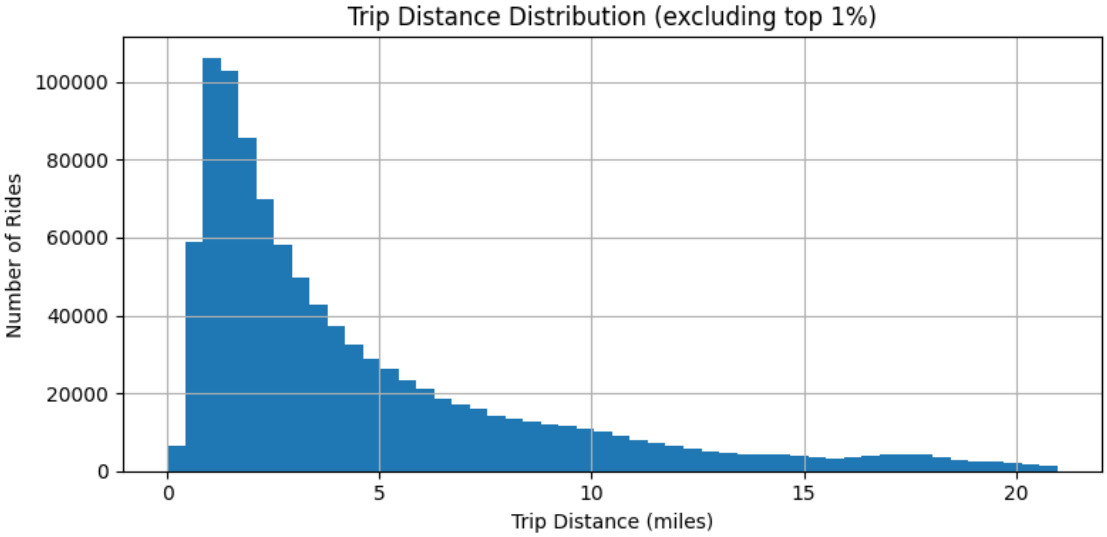

Figure 1: Trip Distance Distribution

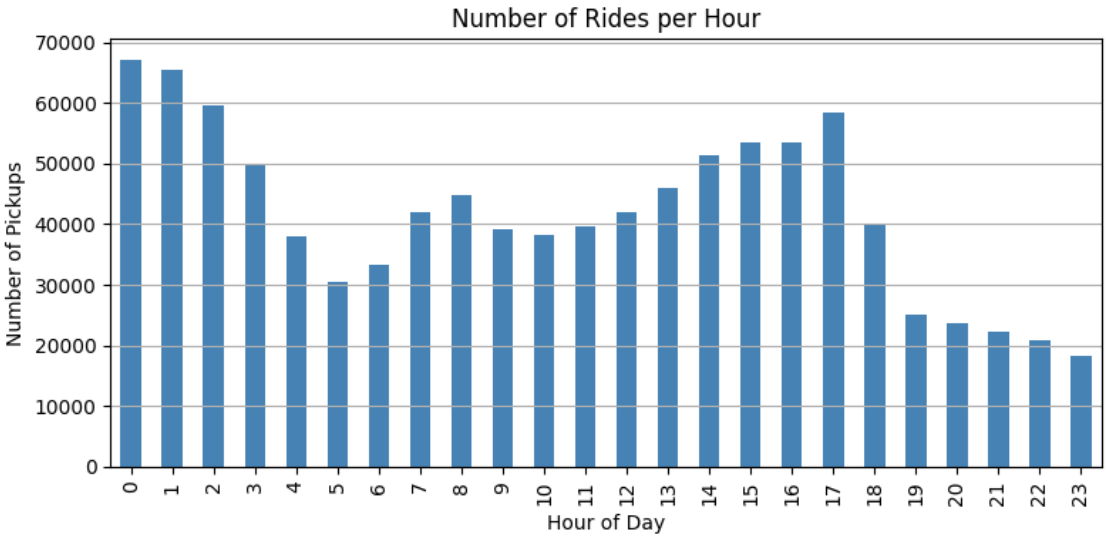

Figure 2: Number of Rides per Hour

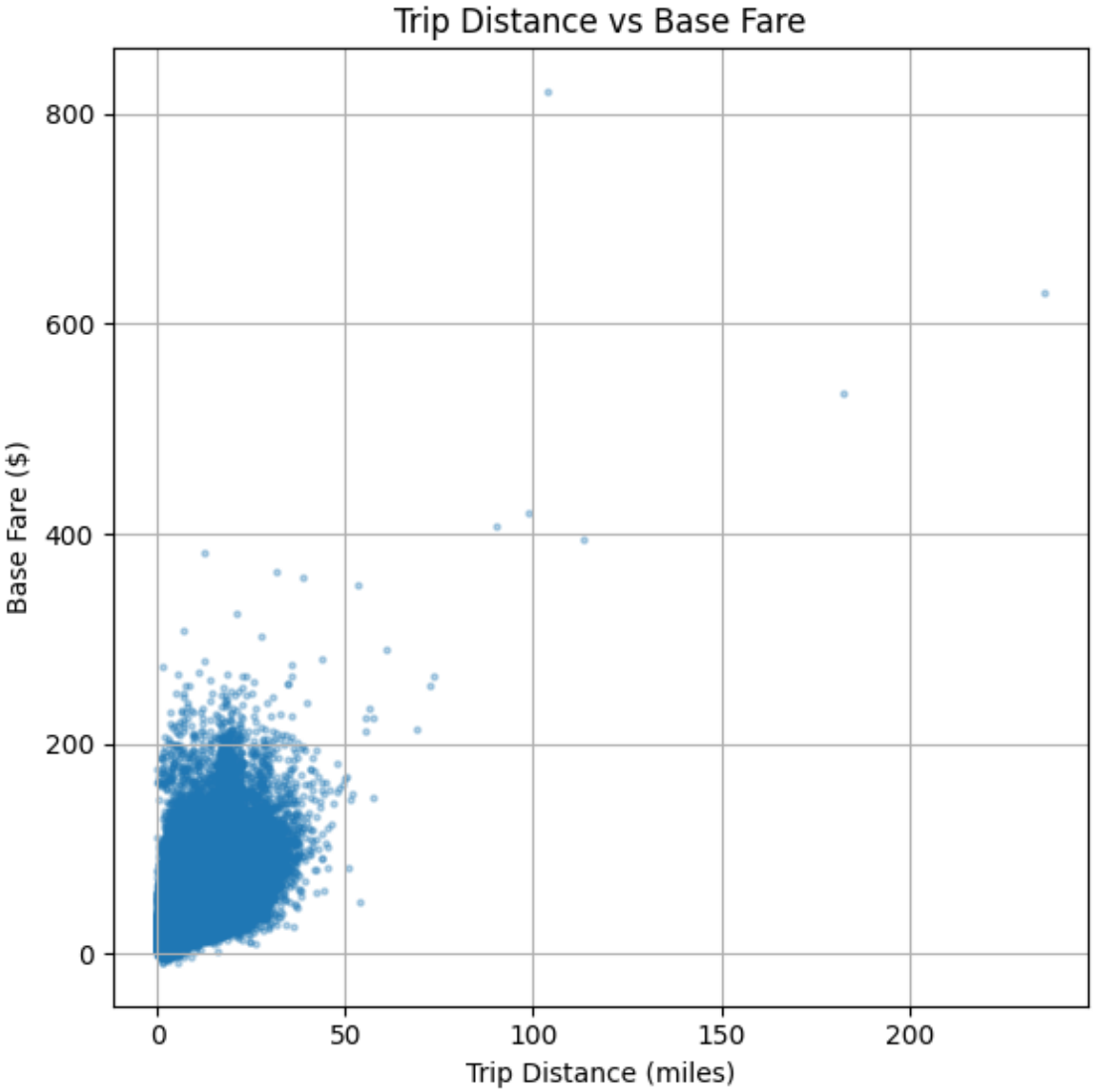

Figure 3: Enter Caption

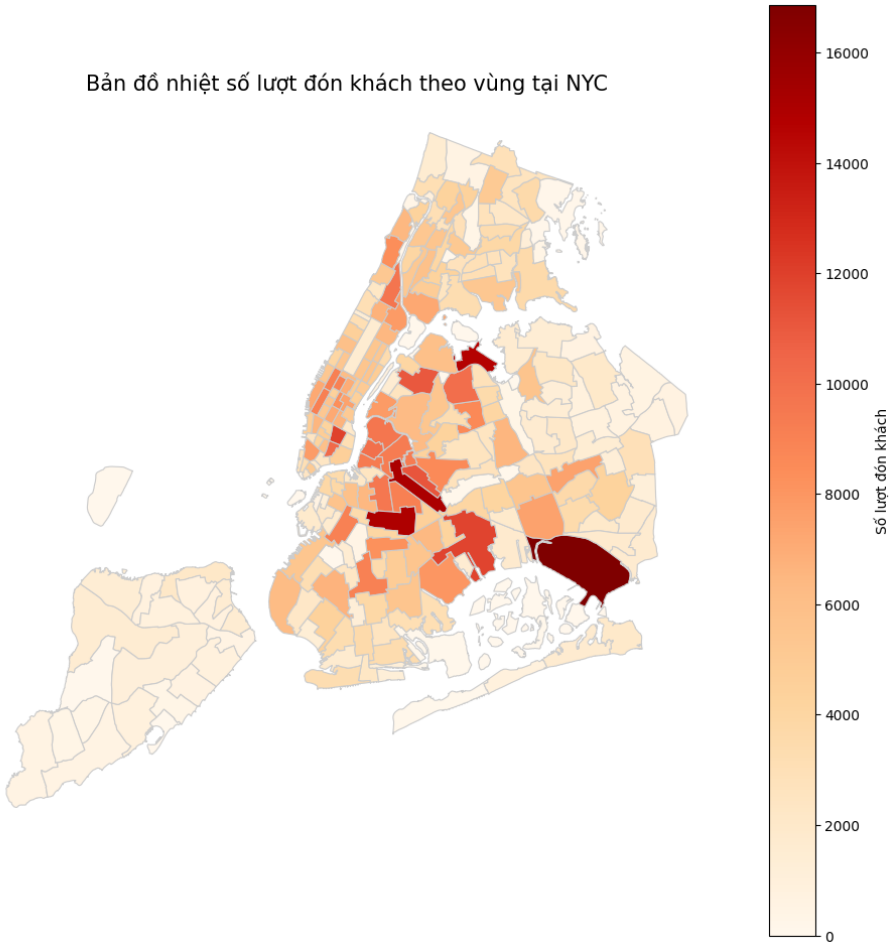

Figure 4: Pick up location heatmap