Introduction to Big Data and Analytics - Fall 2024 - GWU

Project 3: NYC Taxi Trip Insights

Introduction

The NYC Taxi Trip Dataset contains detailed trip records from taxis in New York City, dating back to 2009. This dataset, provided by the NYC Taxi and Limousine Commission (TLC), captures billions of taxi rides, making it one of the largest publicly available datasets for urban transportation research.

Key Features of the Dataset

Each row in the dataset represents a single taxi trip and contains the following key features:

- **Pickup and Drop-off Timestamps:** The time when a trip starts and ends
- **Pickup and Drop-off Locations:** Latitude and longitude coordinates of where the trip begins and ends
- **Trip Distance:** The distance traveled during the trip (in miles)
- **Fare Amount:** Total cost of the trip, including base fare, taxes, and surcharges
- **Payment Type:** The method of payment used (e.g., cash or card)

These features provide crucial insights into trip duration, travel patterns, pricing, and customer behavior.

Collection Process

The dataset is collected by the NYC Taxi and Limousine Commission (TLC) from yellow and green taxis operating in the city. Each taxi is equipped with a GPS-enabled meter that records key details of every trip, including pickup/drop-off locations, fare, and payment method. The data is collected in real-time and stored in the TLC's central database, then anonymized and made available to the public for analysis.

Structure

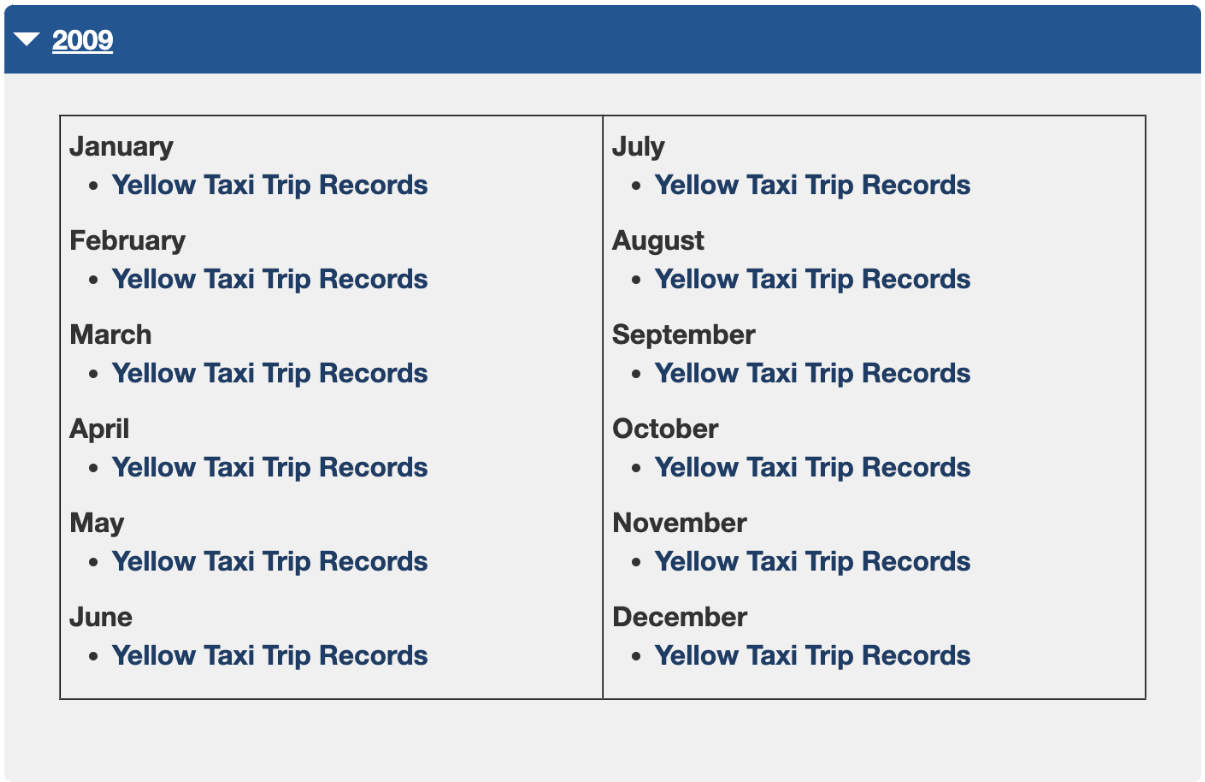
We will be focusing on Yellow Taxi trips. The dataset is already available in Parquet format.

Data Coverage

Start Year: 2009
End Year: Present (the dataset continues to be updated monthly)
Frequency of Updates: Monthly

Data Access

Details of the dataset are present on the TLC website.
The dataset is publicly available. The data for each month starting 2009 can be downloaded as shown below:



▼ 2009

January
- Yellow Taxi Trip Records

February
- Yellow Taxi Trip Records

March
- Yellow Taxi Trip Records

April
- Yellow Taxi Trip Records

May
- Yellow Taxi Trip Records

June
- Yellow Taxi Trip Records

July
- Yellow Taxi Trip Records

August
- Yellow Taxi Trip Records

September
- Yellow Taxi Trip Records

October
- Yellow Taxi Trip Records

November
- Yellow Taxi Trip Records

December
- Yellow Taxi Trip Records

This project will be undertaken in groups of one, i.e. ***individually***.

We will focus on data from 2012 to 2023. Each individual will be assigned 6 months of non-overlapping data; for example, Individual 1 will work with data from Jan-June 2012, Individual 2 with July-Dec 2012, and so on. The dataset allocation will be assigned by the course TA.

Tooling

You are expected to use a combination of pyspark or pandas for processing and MLlib or `sklearn` for analytics in Google Colabs for this project. As always, my recommendation is to upload the dataset to your GWU Google Drive and then process it using pyspark. This dataset is already in Parquet format so no format conversion is necessary.

You can use any Python library, such as `matplotlib`, etc. of your choosing to create the plots. In addition, all documentation/comments should be in the form of Markdown within the notebook itself.

Analysis

For each piece of analysis, provide an explanation of the process as well as an interpretation of the results. In addition you should aim to improve the performance of your underlying models. The final performance along with the process followed (feature engineering, etc.) should be documented as well.

1. Identify regions in the city with high demand for taxis. For instance, you can create a heatmap of NYC showing demand hotspots on a weekly or monthly basis
2. Predict the duration of a taxi trip based on trip start time and location (pick up and drop off)
3. Segment passengers on the basis of trip characteristics, such as trip distance, fare, time of day, and payment method
4. Predict whether a passenger will pay by card or cash
5. Predict the fare of a taxi trip based on borough-specific factors such as pickup and drop-off boroughs, and analyze how fares vary across different NYC boroughs
6. Predict the tip amount given by passengers based on trip characteristics
7. Predict whether a trip will result in a high or low fare based on early trip data
8. Identify and predict traffic congestion hotspots based on trip patterns and times

Due Date

12 November, 2024 - 11:59AM

You can submit the link to your Google Colab via Blackboard in the Assignments section.