NYC Trip data

Frances Lin

March 2021

Problem Statements

The objective of this project is to answer the questions of interest, which include

Questions 1

Questions 2

Questions 3

1. Obtain

TLC Trip Record Data is from The New York City Taxi & Limousine Commission's site. Overall, data is split across multiple **csv** files, is composed of more than one table of related data and is GB in total.

The yellow & green taxi trip records contain fields such as `VendorID`, `tpep_pickup_datetime`, `tpep_dropoff_datetime`, `PULocationID`, `DOLocationID`, `Trip_distance`, `Payment_type`, `Fare_amount`, `Tip_amount`, `Total_amount`, etc. The green taxi trip records has additional field `Trip_type` that is used to indicate whether it is `1=Stree-hail` or `2=Dispatch`.

On August 14, 2018, a Local Law of 2018 wassigned to create a new license category for TLC-licensed businesses that currently or plan to dispatch more than 10,0000 for-hire vehicle (fhv) trips in NYC per day. These businesses include Juno, Uber, Via, and Lyft.

The fhv & high volume for-hire vehicle (hfhv) trip records contain fields such as `Dispatching_base_num`, `Pickup_datetime`, `DropOff_datetime`, `PULocationID`, `DOLocationID`, and `SR_Flag`. The hfhv trip records has additional field `Hvfhs_license_num` that is used to identify Juno, Uber, Via, and Lyft.

There is also a taxi zone lookup record that contains fields such as `LocationID`, `Borough`, `Zone`, and `service_zone`.

** Initial Data**

a sample of *Yellow Taxi Trip Records 2020-01*

a sample of *Green Taxi Trip Records 2020-01*

a sample of *For-Hire Vehicle Trip Records 2020-01*

a sample of *High Volume For-Hire Vehicle Trip Records 2020-01*

a sample of *High Volume For-Hire Vehicle Trip Records 2020-01*

Feb 24 I wrote some **Python** script to obtain data through a list of **url**, merge dataframes, and write the resulting dataframe to a **zip** file using the `pandas` and `ZipFile` package. However, I don't think merging should take place prior to loading them to **Google Cloud**'s **BigQuery**. I also wondered if I should make API request if the data has already been uploaded somewhere, but to save time for processing for later section, I directly downloaded the **csv** files from the site and compressed them into multiple **zip** files since the maximum size for loading a **gzip** file to **Storage** is 4(?) GB.

The total uncompressed file is GB and the total compressed file is GB. Specifically, the uncompressed file of *Yellow Taxi Trip Records 2015-01 to 2020-06* is 67.77 GB and the compressed file is GB.

The uncompressed file of *Green Taxi Trip Records 2015-01 to 2020-06* is GB and the compressed file is GB.

The uncompressed file of *For-Hire Vehicle Trip Records 2015-01 to 2020-06* is 37.65 GB and the compressed file is 5.94 GB.

The uncompressed file of *High Volume For-Hire Vehicle Trip Records 2019-02 to 2020-06* is GB and the compressed file is GB.

Feb 25 I downloaded Yellow records.

Feb 26

2. Scrub

a sample of in BigQuery's Dataprep

a sample of

a sample of

a sample of

3. Explore

4. Model

a snip of Python script to run Dataproc job

5. Interpreting