NYC Taxi and Limousine - High Volume Dashboard

John Cruz

2023-04-01

Introduction

The NYC Taxi and Limousine Commission publishes through their NYC Open Data program trip information on different designations. One such category are *High Volume For-Hire Vehicles* (HVFHV) which comprise of app based ride requests such as Uber and Lyft. The data being used was collected from their website and will be used for this analysis.

Source:

TLC Trip Record Data Taxi Zones

The NYC Taxi and Limousine Commission also provides data dictionaries for their datasets.

HVFHV Data Dictionary

Objective

Create a Tableau Public dashboard that provides descriptive information based on the HVFHV industry for the month of August 2022.

Required Libraries

To be able to read the parquet type files, you must have the *arrow* library installed.

```
library(tidyverse)
library(arrow)
```

Import Trips Data

Read in high volume trip data from August 2022.

```
trips_df <- read_parquet('fhvhv_tripdata_2022-08.parquet')
```

Alter Columns

The following columns will be created for the dashboard.

1. Define each app company:

- HV0003: Uber
- HV0005: Lyft

2. How long did passengers wait until the driver shows up to their location?
3. How long does the driver waits for the passenger before the trip begins?

```
trips_df <-
  trips_df |>
  mutate(app = case_when(hvfhs_license_num == 'HV0003' ~ 'Uber',
                         hvfhs_license_num == 'HV0005' ~ 'Lyft'),
         passenger_wait = on_scene_datetime - request_datetime,
         driver_wait = pickup_datetime - on_scene_datetime) |>
  select(-c(dispatching_base_num, originating_base_num, access_a_ride_flag, hvfhs_license_num))
```

Filter Data

The original dataset contains close to 17 million trip records. The data will be cleaned and filtered with the following conditions:

1. Trip travel time that is non-negative.
2. Trip travel time that is at most five hours long.
3. Trip travel distance that is non-negative.
4. Driver pay is more than $0.01.
5. Base passenger fare is more than $0.01.
6. Pickup and dropoff locations are within the NYC region. Based on the taxi zone data, this excludes zones 264 and 265.

```
trips_df <-
  trips_df |>
    filter(trip_time >= 0,
           trip_time < 18000,
           trip_miles >= 0,
           driver_pay > 0.01,
           base_passenger_fare > 0.01,
           !PULocationID %in% c(264, 265),
           !DOLocationID %in% c(264, 265),
           dropoff_datetime >= '2022-08-01')
```

Import Taxi Zone Data

```
zone <- read_csv('taxi_zone.csv')
```

Merge Trips and Zone Data

Given the size of this data containing almost 17 millions trip records, it will be broken into two chunks. One set will be trips that originated in Manhattan regardless of their dropoff zone and the other set will be trips that originated within the outer boroughs (Brooklyn, Queens, Staten Island, Bronx) regardless of their dropoff zone.

```r
manhattan_trips_df <-
  trips_df |>
  left_join(zone, join_by(PULocationID == LocationID)) |>
  filter(Borough == 'Manhattan') |>
  rename(pu_borough = Borough, pu_zone = Zone) |>
  select(-service_zone) |>
  left_join(zone, join_by(DOLocationID == LocationID)) |>
  rename(do_borough = Borough, do_zone = Zone) |>
  select(-service_zone)

outer_trips_df <-
  trips_df |>
  left_join(zone, join_by(PULocationID == LocationID)) |>
  filter(Borough != 'Manhattan') |>
  rename(pu_borough = Borough, pu_zone = Zone) |>
  select(-service_zone) |>
  left_join(zone, join_by(DOLocationID == LocationID)) |>
  rename(do_borough = Borough, do_zone = Zone) |>
  select(-service_zone)
```

Export Files

```r
# write.table(manhattan_trips_df, file = "manhattan_trips.txt", sep = "\t",
#             row.names = FALSE)
#
# write.table(outer_trips_df, file = "outer_trips.txt", sep = "\t",
#             row.names = FALSE)

write_parquet(manhattan_trips_df, "manhattan_trips.parquet")
write_parquet(outer_trips_df, "outer_trips.parquet")
```