

DASHBOARD TO QUICKLY ESTIMATE THE COST AND DURATION OF A NYC GREEN TAXI TRIP



Author: Isaac Braun | Coauthor: Germán H. Alférez, Ph.D.
{isaacbraun, harveya}@southern.edu

School of Computing, Southern Adventist University, Collegedale, TN 37310



ABSTRACT

Before hailing a New York City (NYC) taxi, residents and tourists do not easily know how much the trip will cost them or how long it may take. Taxis are still heavily used, even with the increase of ride-hailing services like Uber, and a new system has yet to be built to provide customers with these two metrics before taking a trip. This project aims to give riders a quick way to estimate a ride's cost and duration through an interactive dashboard that allows filtering by pickup and drop-off neighborhoods. This is accomplished by analyzing three years of public data made available by the city of New York and presenting it to users through an interactive dashboard.

METHODOLOGY

1. Import Data

- Four data files were downloaded from Google BigQuery [1] and data.gov [2]. One file for each year within 2020-2022, and one with the location data of taxi zones.
- Uploaded separate files into Google Cloud Storage [3]
- Imported into BigQuery as a new set of tables

2. Reduce + Export

- Due to the extreme size of the initial data, queries were run in BigQuery to obtain only the desired fields, ending with 3,588,894 records.
- Downloaded the tables that resulted from the queries as individual files

3. Prepare

- Imported all files into Tableau Prep Builder [4]
- Focusing on Fare, Date/Time, and Pickup/Dropoff Location, unnecessary fields were removed, outliers were cleaned up, and extra fields were computed.
- Merged all three years into one file, which was made available online [5]

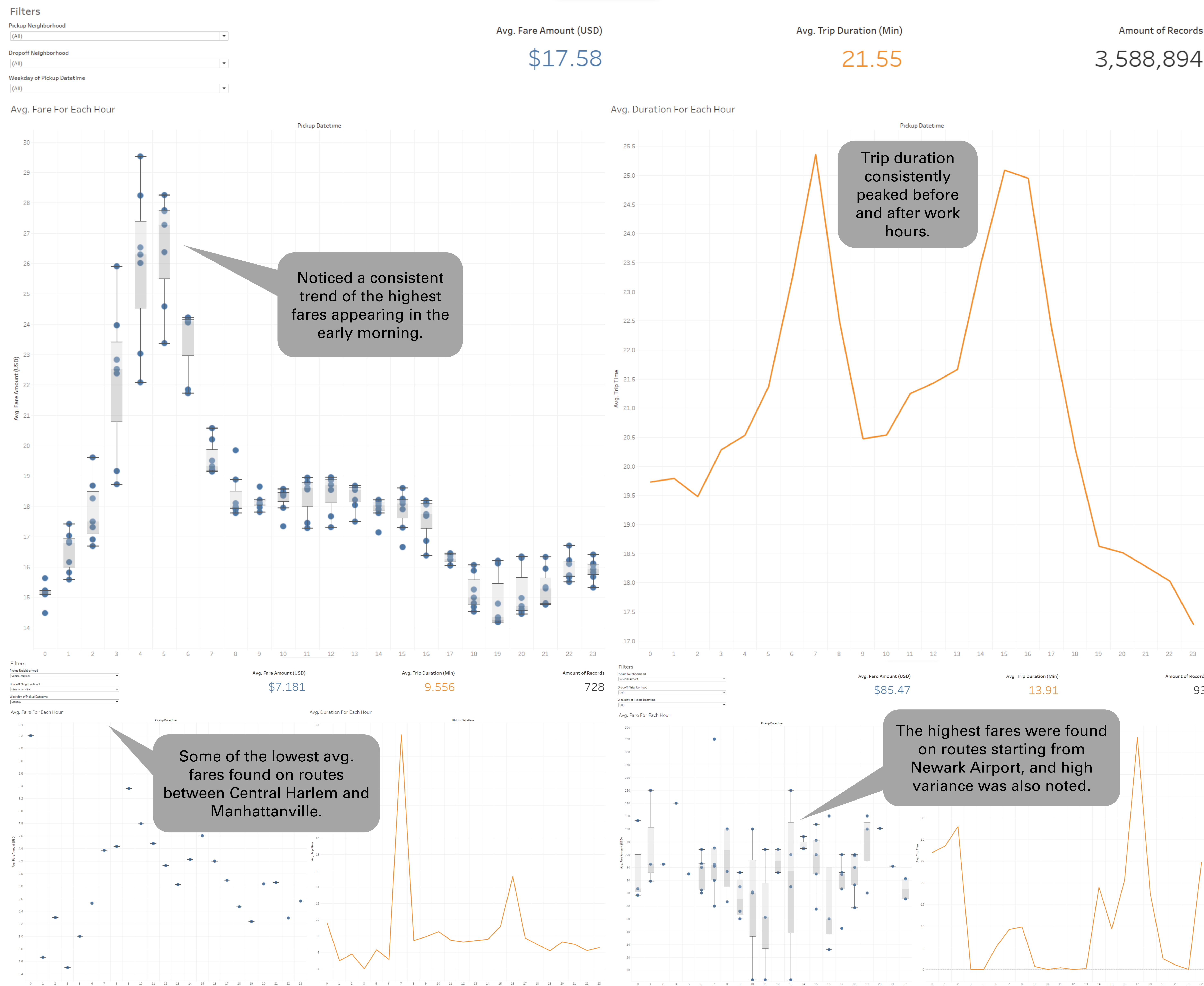
4. Visualize

- Used Tableau Desktop [6] to visualize the data and create an interactive dashboard

RESULTS

An interactive dashboard was created, allowing users to find real-world data about the cost and duration of NYC Green Taxi trips to and from specific neighborhoods. Inputs in the top left corner filter the two charts and three figures based on pick-up location, drop-off location, and day of the week. This allows users to quickly see the average fare amount, average trip duration, amount of records, and the trend charts for based on their inputs.

In the process, some observations of trends in the data were made. Regarding fares, a consistent trend of prices spiking during the early hours, 3 to 7 a.m., and a dip around 6 to 7 p.m. were noticed. Concerning trip duration, there were typical peaks just before and after work hours, with the highest peaks commonly in the evening.



FUTURE WORK

Future work may involve adding more data from other years or the NYC Yellow Taxi service to increase the accuracy of the estimates. The dashboard may be built as a web or mobile application to improve the system's ease of use. Other trip and locational data could be added that would allow users to get estimates by entering in a specific address rather than just a neighborhood.

REFERENCES

- <https://cloud.google.com/bigquery/public-data>
- <https://data.gov/>
- <https://cloud.google.com/storage>
- <https://www.tableau.com/products/prep>
- https://bauen-public-resources.s3.amazonaws.com/nyc-taxi/green_trips_20-22_final.csv
- <https://www.tableau.com/products/desktop>