Is Lyft Really Cheaper than Taxi in New York ?

Sikai Zhao[1], Hao Wang[1], Yaoxu Xiao[1], Yahui Ke[1], Xueyang Zhang[1],
[1]Project Group 25: Newbie, CSE6242 Data and Visual Analytics Fall 2019,
College of Computing, Georgia Institute of Technology

CSE6242A,Q Fall 2019
Data and Visual Analytics

Motivation/Introduction

The NYC Taxi and Limousine Commission (TLC) provides open data of yellow taxi trip records including pick-up and drop-off locations/times, trip distances, fare type, payment type and fare information. Behind those trip data, there must be a lot of interesting stories to tell. How much more expensive would it be to take a taxi during rush hours? Is it cheaper to take a taxi on weekends or on weekdays? Are there any particular periods of a year that is a costly or a cost-efficient phase for taking a taxi? And what potential factors might affect the trip price fluctuation? To find out the answers, the New York yellow taxi data of the year 2018 was acquired from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. A taxi fare prediction model was developed and trained using the data. A large number of trips categorized into different months, peek hours versus non-peak hours or workdays versus weekends were called and the corresponding taxi fare was rendered using our prediction model. Besides, a New York City taxi fare prediction web application was developed, allowing users to get the taxi fare prediction by selecting the pick-up and drop-off locations.

In addition, nowadays, peer-to-peer car hailing offered by ride-hailing companies like Uber and Lyft has become an increasingly popular approach of transportation services in cities. As alternative transportation choices, are they cheaper than Taxi? The trips fares of sample trips from Lyft versus taxi (with our taxi fare prediction model) in New York were acquired and compared.

This project is interesting and essential as we, all as customers, care about how much we need to pay for the daily transportation.

Methodology

Data Processing and Clustering

The New York yellow taxi data of the year 2018 was acquired from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Data cleaning and processing were performed with Spark/Scala on Databricks. Extra columns and invalid rows were removed. Trip duration was generated based on the pick-up and drop-off time. Spark is a unified engine for diverse big data processing and is suitable for our purpose of analyzing a whole year of yellow taxi trip data. Then, time cluster were created by extracting month, week number and days of a week. A time zone parameter was generated using a basic map function for model training purposes. This clustering method make it easier for our training.

Prediction Model (training and testing)

The gradient boosting LightBGM framework was used to train the model. LightBGM has faster training speed lower memory usage and higher efficiency with the capability of handling large-scale data. We are able to train the model in a few hours with less memory. 80% of the year 2018 taxi data was used for training and 20% of data was used for testing. Different training parameters such as num_leaves, max_depth, learning_rate, num_boost_round, early_stopping_rounds were explored in order to find the best training parameters. The testing dataset of the final model's mean percentage error is 10.05% and the mean square error is 3.17 USD.

Web Application

An interactive web application was designed and built for customers to get a price estimation of a trip using Flask. The front end is developed using HTML, css and Javascript. A New York map is generated by calling Google Map API using Javascript. The pick-up and drop-off locations (longitude and latitude) are passed from the user input, time is current time. Location addresses are acquired by implementing reverse geocoding via calling Google Geocoding API. The trip duration and trip distance are acquired by calling the Google Distance Matrix API. Location ID is determined using d3 function and the .json documentation provided by TLC. Necessary trip information are passed to the back end through jQuery and .ajax POST and Get. The back end includes the Taxi fare prediction model we developed, and the estimated price got from the model is returned to display. SQLite is used to store trip search history. Flask is a lightweight WSGI web application framework, providing developers with more freedom. It is much easier to get started and scale up to complex applications compared with other frameworks. Therefore, it is suitable to demand of our website. As for the front-end APIs, the Google Map APIs are chosen because Google map might be the most popular map App with user-friendly service and detailed documents and tutorial.

Plots Generation

Tableau is used for data visualization to create the line charts with the variables (e.g.: distance, time of a day and the average price of Lyft & Taxi). We can then get the significance relationship as the listed charts shown in the results part.

Experiments

Experiment-1: Evaluate the taxi fare difference between rush hour versus non-peak hour and workday versus weekend.

Taxi fare comparison among different time of a day: time period of sample trips are categorized as 4:00-6:00 (early morning), 8:00-10:00 (morning rush hour), 16:00-19:00 (night rush hour) and 21:00-23:00 (midnight). Taxi fare comparison between workdays and weekends: date of sample trips is categorized as workdays and weekends.

Experiment-2: Taxi Fare comparison: Lyft versus Taxi in New York City.

Sample trips with different distance, time (peek/non_peek hours), workdays versus weekends are used to get the Taxi fare prediction from our model and Lyft fare estimation from Lyft app. Time point are set to 9:00AM, 12:00AM, 3:00PM, 7:00PM, 11:00PM and distance are set to short (< 2 location zone), medium and long (> 5 location zone).

Approach Evaluation and Comparison

To evaluate our model training, 20% of data was used for testing. Different training parameters such as num_leaves, max_depth, learning_rate, num_boost_round, early_stopping_rounds were explored in order to find the best training parameters. The testing dataset of the final model's mean percentage error is 10.05% and the mean square error is 3.17 USD.

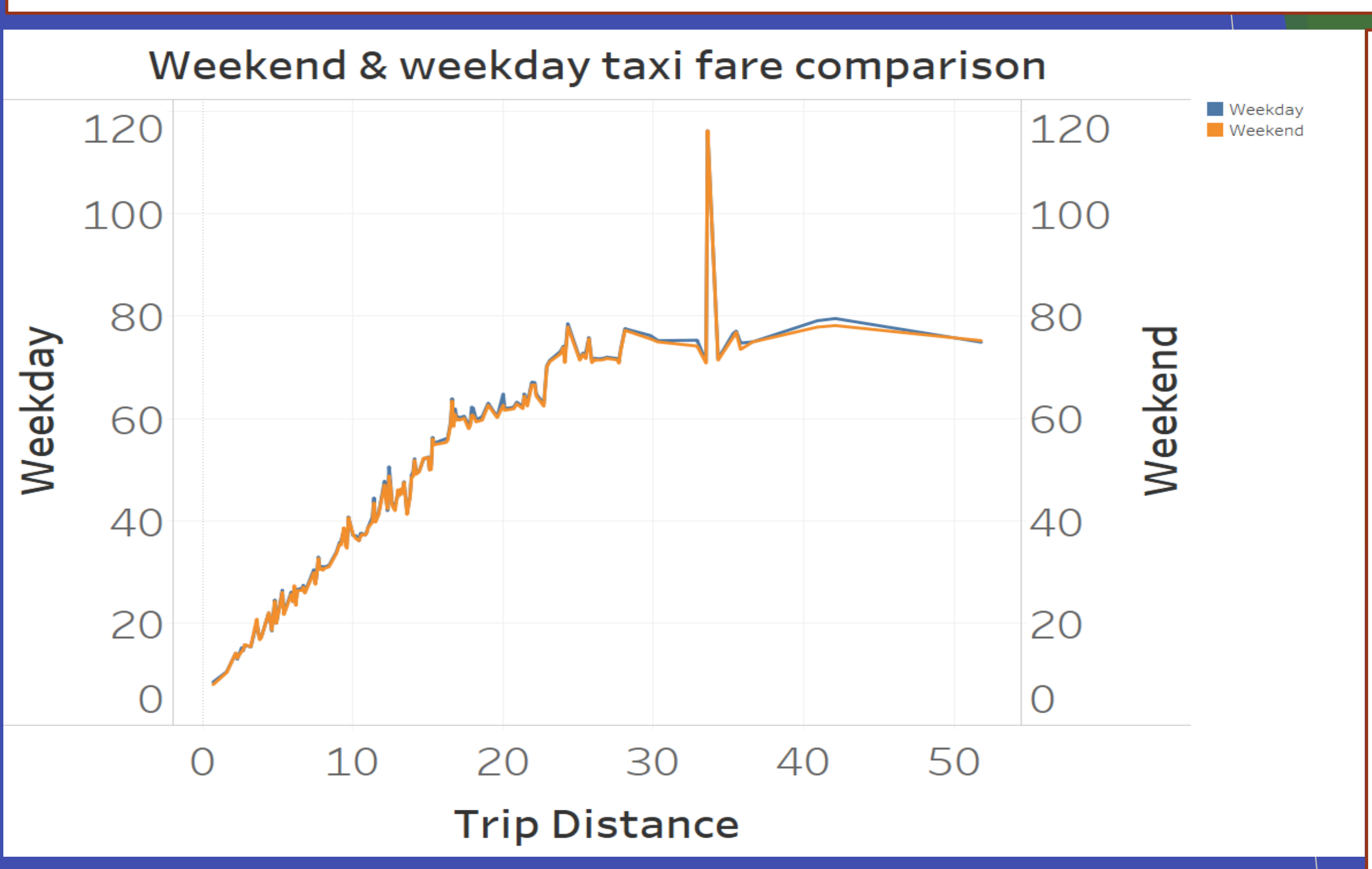
Flask framework is used for web application development. Flask is a lightweight WSGI web application framework, providing developers with more freedom. It is much easier to get started and scale up to complex applications compared with other frameworks. As for the other popular web frameworks, take Django as an example, it is extremely all-inclusive, making it difficult to change or implement. And Django offers many functions that we do not need, causing it over-weighted. Therefore, flask is suitable to demand of our website.

As for the front-end APIs, the Google Map APIs are chosen because Google map might be the most popular map App with user-friendly service and detailed documents and tutorial. In contrast, another popular app, Here Map is not so popular among people as google map, and not so widely used by apps or webs nowadays. Hence, Google map APIs are chosen.

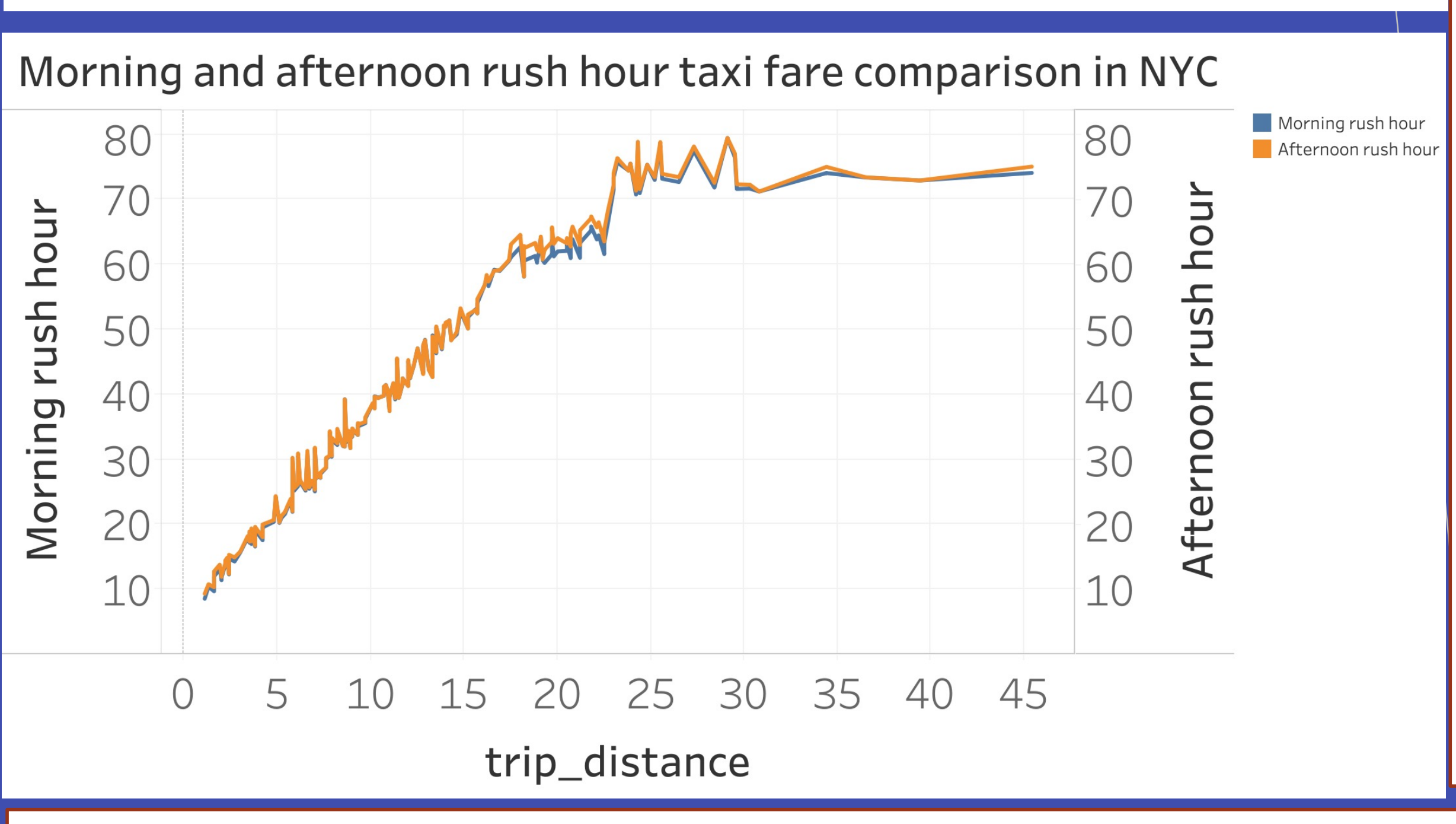
Data

The New York yellow taxi data of the year 2018 was downloaded from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. The data size is around 7.85G. It contains 112M rows of trip data records with 17 columns including pick-up datetime, drop-off datetime, trip distance, pick-up location, drop-off location, fare rate, and fare, etc. The data is historical data and it is not temporal.

Results: Experiment-1

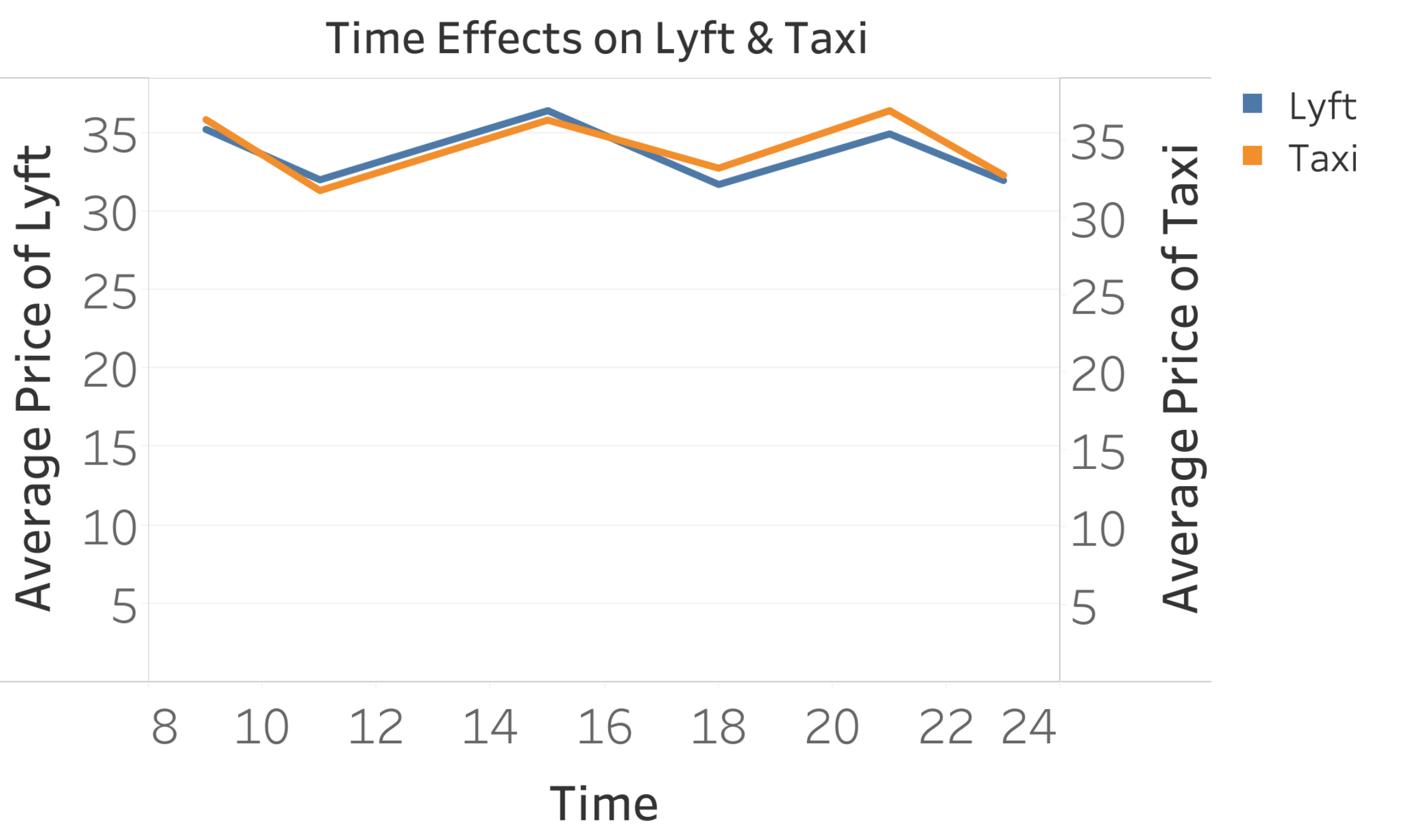


Weekend & Weekday comparison:
As the figure shows weekday generally has more uniform fare prices for the same distance. The reason is that weekday traffic has the same trip patterns as people go to work and back home. Weekend traffic is more diverse as people have different trip purposes with different destinations. Weekend, weekday along has no big impaction to taxi fare prices for the same trip distance.

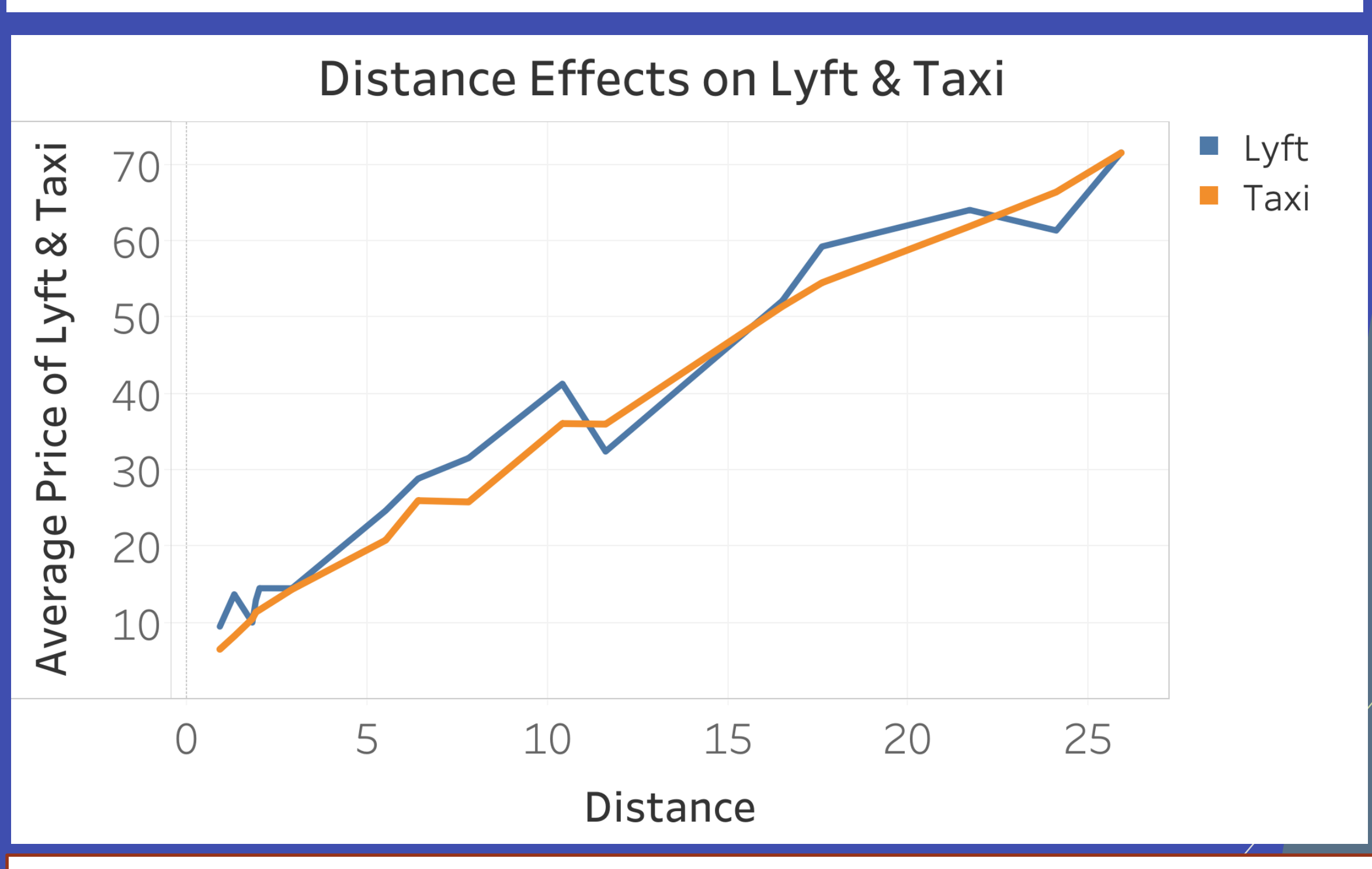


Morning & afternoon rush hour taxi fare comparison:
The figure shows that during the morning rush hour, the traffic has the similar trip patterns as people go to work. Afternoon rush hour traffic is more diverse as people have different trip purposes with different destinations. Morning, afternoon along has no big impaction to taxi fare prices except for the mid range distance.

Results: Experiment-2

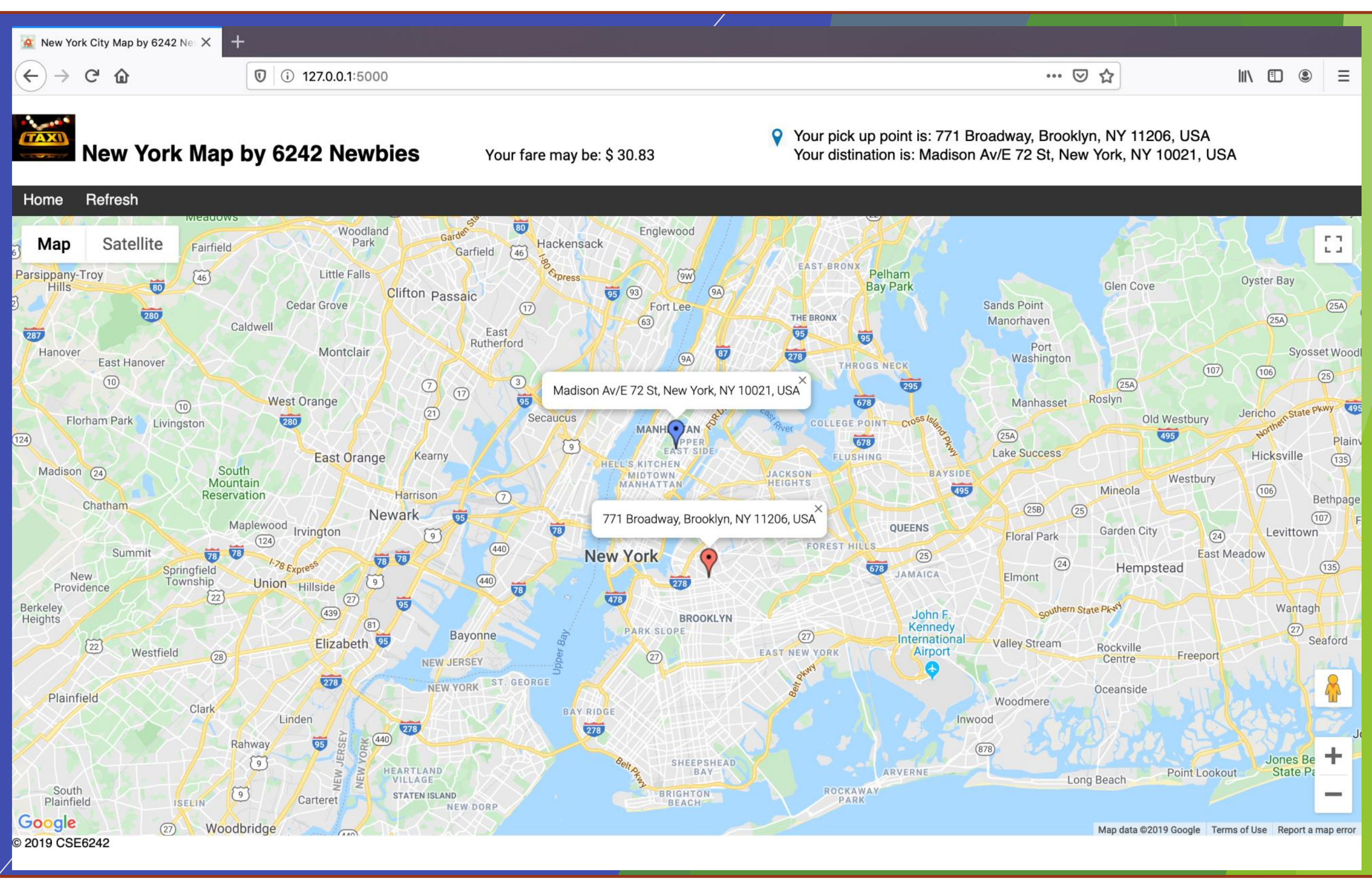


Time effects on Lyft and Taxi:
As shown in the upper panel, the overall trends of price changes according to different time period of a day are pretty similar between Lyft and Taxi. However, Lyft seems to be slightly cheaper than Taxi during rush hours in the morning before 11 am and in the afternoon after 3 pm. Lyft might be a more cost-efficient approach for transportation during rush hours.



Distance effects on Lyft and Taxi:
As shown in the lower panel, when distance is less than 10 miles, Taxi is cheaper than Lyft; when distance is longer, Lyft might be cheaper than Taxi; when distance is even longer to exceed ~17 miles but smaller than ~23 mile, Taxi is cheaper; when distance is between 23-25 miles, Lyft is cheaper. As the distance increases, the Taxi fare seems to continually and smoothly increase; while the Lyft fare shows some fluctuation.

Web Application Demo



New York City Map by 6242 Ne
127.0.0.1:5000

New York Map by 6242 Newbies

Your fare may be: $ 30.83

Your pick up point is: 771 Broadway, Brooklyn, NY 11206, USA
Your distination is: Madison Av/E 72 St, New York, NY 10021, USA

Home Refresh

Map Satellite

Madison Av/E 72 St, New York, NY 10021, USA

771 Broadway, Brooklyn, NY 11206, USA

© 2019 CSE6242

Acknowledgement

CSE6242 A,Q, Fall 2019, Georgia Tech, College of Computing