Analyzing the New York City Taxi Dataset

Spencer Milbrandt
University of Colorado
Boulder, Colorado
Spencer.Milbrandt@colorado.edu

Rasheeq Jahan
University of Colorado
Boulder, Colorado
Rasheeq.Jahan@colorado.edu

Peter Wang
University of Colorado
Boulder, Colorado
Peter.Wang@colorado.edu

1 PROBLEM STATEMENT AND MOTIVATION

The purpose for analyzing the New York City taxi data is to recognize patterns in common pick-up and drop-off hotspots and the various factors that affect them. This includes weather data, distance traveled, fare amounts, and locations. With these various datasets, the goal is to discover any correlation between these various attributes to determine and predict pick-up locations and the likelihood of an available taxi at any given time, as well as the viability of carpooling based on average distance and weather factors. The inspiration for pursuing this project can be attributed to the rise in popularity of Uber and Lyft driving services and their impact on taxi services, specifically New York City. In order to optimize taxi service business, considering non-obvious factors when discovering patterns in distance traveled and fare amounts based on weather can give insight into improving taxi routes and providing more stable taxi rates regardless of weather. With this information, New York City taxi services will be able to improve competition with Uber and Lyft on a cheaper and more stable fare amount.

2 LITERATURE REVIEW

Prior work on this topic includes research focused improving the organization of taxi schedules through the development of an index that indicates the degree of passenger availability for taxi drivers [2]. The analysis conducted in this study, which was set in Shenzhen, China, uncovers patterns for trips over various categories, including time and location, as well as reasons for the differences in distribution of pick-up and drop-off locations. Floating-car technology is a key factor that formed the basis for this study and allowed researchers to perform large scale collection of live data. Through the analysis of the distribution of pick-up and drop-off locations, researchers discovered that at times passengers prefer greater use of taxi services for pick-ups from certain subareas, than drop-offs to those subareas and vice versa. A reason for this uneven demand for taxi services was "due to land use, major establishments, and urban layout" [2].

A similar study on "mining the best passenger-finding strategies," focused on improving the efficiency of taxi companies by informing drivers of locations with the greatest possibility of finding passengers in need of rides. Using spatial data patterns and forecasting methods for time-series data, researchers worked to improve the "unbalanced relationship between the passenger demand and the number of running taxis" [1]. A step of preprocessing in this study included developing "a time series of taxi demand services … aggregated for a period of P minutes" [1]. In order to predict demand, researchers divided the taxi service area into "clusters based on historical passenger demand" and applied the time-varying Poisson model, as well as the ARIMA (autoregressive integrated moving average) model. Two advantages of the ARIMA algorithm covered in this study is its ability to allow for various types of time series, as well as continuous modifications that "update itself to changes in the model" [1].

3 PROPOSED WORK

3.1 Data Collection

The first step in our analysis is of course collecting the data. Our data set is provided by the Taxi and Limousine Commission as a subsection of the New York City Official website. At the moment the information is divided into two main sections: Yellow and Green. The data is then further split into a single file for each month of a given year. Therefore our first task is to download a specific sample of the available data and consolidate it into a single document. The temporal range of this sample will be determined by the information available from our second data set: The NYC Weather data set. Fortunately, the Weather data and our third dataset: NYC Uber Trips both come in single csv documents which makes collection relatively simple.

3.2 Preprocessing

Before any work can be done in analyzing the data, there are a few preliminary tasks in preprocessing that need to be addressed. First and foremost, the three separate data sets need to be normalized in terms of distance units and time format. The NYC Taxi data also contains a large amount of extra attributes that are irrelevant and need to be cleaned out. In addition to this, we are planning to remove any abnormally long trips as outliers. The major portion of the preprocessing work also comes in the NYC Taxi data set. The main issue in this data comes in the format of the pickup and drop-off locations. The NYC Taxi and Limousine Commission divides the city into separate numbered regions and lists these in the data. In order to properly understand the information, we will have to plot the regions on a map for analysis and visualization.

3.3 Design

Our data comes conveniently clustered by pickup and drop-off locations. At the moment our current plan is to strictly categorize by pickup location and potentially explore further categorization by drop-off location. This is particularly relevant to our analysis as it will simplify the process for filtering out irrelevant data for any given user. Using these built in clusters, we hope to discover patterns for high traffic areas, particularly slow travel times, and possibly variations in fare. As we learn about classification models used to form predictions, we hope to learn about and apply one in particular found in related studies, which is the classification based on the k-nearest-neighbor algorithm. An advantage of using a lazy-learner classification model is that it "naturally supports incremental learning" [4]. This classification model searches the training space of patterns for the instances or nearest neighbors that are closest to a given unknown instance. As a lazy-learner, this model is resource intensive to run and may require the use of the partial distance method and the pruning of training tuples. While prior studies studied the effects of weather on taxi data, we hope to also consider the impact of recent factors, such as the growth of ride-sharing services, on our forecasting model.

4 EVALUATION

Our analysis of the data will involve establishing averages for travel time and fare for a given distance and under specific weather conditions. After these are established we can then compare individual trips to these averages and determine times and paths that are particularly slow or expensive compared to Uber. The last step of our evaluation is the visualization. For this we plan on plotting the various taxi zones over an interactive map with specific filters for time and weather.

5 DATA SET

For our analysis we will be looking specifically at 3 data sets. Our first set is the NYC Taxi Data set[1]. This data is divided into a separate CSV file for each month of a given year with each file containing

[1] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

information for over a million individual trips. The attributes we will be examining are pickup/drop-off times and locations as well as distance travelled and payment amount. For our second data set: the NYC Weather Data set[2], we have downloaded a single CSV file that lists the high, low, and average temperature for any given day as well as amount of precipitation and depth of snow. In our last data set, the NYC Uber data set[3], we have information on individual trips for specific Uber units.

6 TOOLS

The tools that will be used to complete and assess an in-depth analysis of the taxi data will include but not be limited to: Python3, WEKA data visualization, SQL, and Python3 Libraries included in Anaconda to help analyze and visualize the data through numpy, scikit, and pandas. Due to prior work, there are certain tools that would be better for data visualization of pick-up and drop-off location hotspots but further research is required to discover these tools and adjust based on our findings.

7 MILESTONES

We plan to follow the proposed work described above and complete these milestones by the following dates (2018) to achieve steady progress while also finalizing and submitting the final report by the assigned due date: 5/1
Data Collection: 3/9
Preprocessing: 3/30
Design and Correlation: 4/13
Visualization and Finalizing: 4/23

7.1 Milestones Completed

The completed work we have planned out to accomplish within the designated time-frame listed out in the previous section, Milestones, is that we have gathered the required data-sets needed for our project and have completed the preprocessing stage of our design plan and have started on using tools such as WEKA and Python to find any correlation between different attributes and start to cluster pick-up and drop-off locations to gain a better understanding of distance vs. fare rates. We have decided to utilize the Yellow (2016) cab data from the NYC Taxi and Limousine Commission. Preprocessing allowed us to clean the data of any NULL values and to negate any attributes in the data-set that are unnecessary for our project. However, the project team has decided to attempt to merge the datasets of NYC Taxi (2016) data with the NYC (2016) weather data. During the cleaning phase of processing the taxi data, WEKA allowed us to improve the completeness of the data through the use of a cleaning filter, which replaced the missing values with the mean or the median of the particular attribute based on the distribution. We later hope to consider potential improvements in our data cleaning process through the application of a decision tree learning algorithm that would provide us with predictions for the missing values.

7.2 Milestones To-Do

Comparing our milestones progress and the proposed work schedule listed in the Milestones section, we plan to have done most of our design process and correlation work by 4/13 to allow us enough time before the final deadline to visualize, finalize, and discover any patterns that would help us confirm our initial hypothesis and project proposal. We perceive that the most difficult work that will need to be achieved within the deadline is correctly clustering the pick-up and drop-off locations and then finding a correlation between fare rates along with any weather conditions that could have impacted the price fluctuations at any given time. We have yet to decide whether or not we need to utilize the NYC Uber (2016) data-set to find any impact that Uber has had on the average distances or fare rates within NYC.

[2] https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016

[3] https://data.cityofnewyork.us/Transportation/Uber-Trips-NYC-2016/gt3n-7ri6/data

8 RESULTS

The results of our Milestones thus far have been successfully cleaning the data-sets we will be using for this project [NYC Taxi (2016) and NYC Weather (2016) data-sets]. We have started to utilize the WEKA tool to initially find any clustering patterns within our data for pick-up and drop-off locations and expect to correlate the average fare rates based on the distance traveled, ultimately expecting to correlate any weather conditions within our NYC Weather (2016) data-set to the fluctuation of pick-up and drop-off locations as well as the fare rates for that given time. We hope to further explore the correlation between the pickup locations and trip distances to determine if trips originated from specific parts of New York City are more likely to lead to longer or shorter trips. A scatterplot of the pickup location and trip distance is shown in Figure 1.

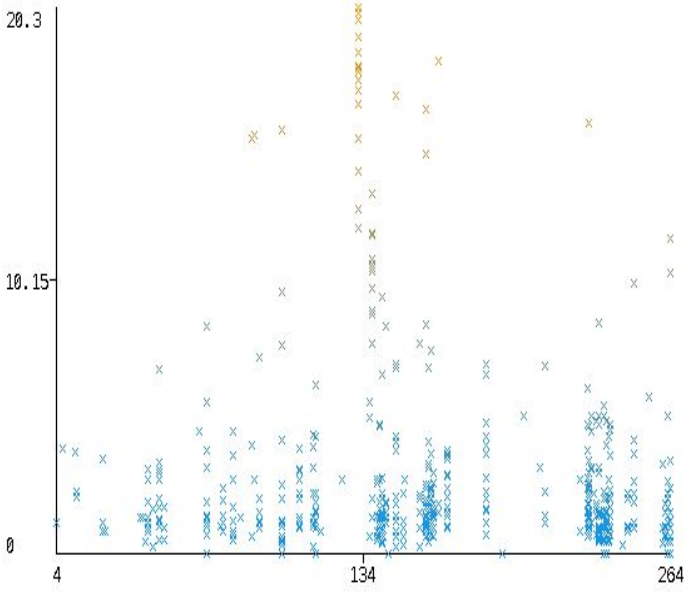

Figure 1: Trip distances throughout New York City by pickup locations

A HEADINGS IN APPENDICES

A.1 Problem Statement and Motivation

A.2 Literature Review

A.3 Proposed Work

A.3.1 Data Collection

A.3.2 Preprocessing

A.3.3 Design

A.4 Evaluation

A.5 Data Set

A.6 Tools

A.7 Milestones

A.7.2 Milestones Completed

A.7.1 Milestones To-Do

A.8 Results

A.9 References

9 REFERENCES

[1] Moreira-Matias, Luis, et al. "Predicting Taxi-Passenger Demand using Streaming Data." IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 3, 2013, pp. 1393-1402.

[2] Zhang, Zheng, and Xiang He. "Analysis and Application of Spatial Distribution of Taxi Service in City Subareas Based on Taxi GPS Data.".

[3] Tang, LL, et al. "Uncovering Distribution Patterns of High Performance Taxis from Big Trace Data." *Isprs International Journal of Geo-Information*, vol. 6, no. 5, 2017, pp. 134.

[4] Han, Jiawei, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Elsevier Science, Burlington, 2011.