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 ABSTRACT

The overall question our team is exploring is: What are the various factors that affect the frequency, fares and locations of Yellow Taxis in and around New York City. Our initial hypothesis was that the greatest contributing factors to New York City taxis would be the time of day, weather, and concentration of famous landmarks. In particular, we hypothesized that the highest traffic times would be early mornings and evenings, mainly associated with the start and end times of people's commutes. We also believed that the data would reflect higher taxi usage and fares in times of colder or more extreme weather conditions as a result of longer travel times. However, upon completion of our analysis the results we found were relatively surprising. While our hypothesis regarding the time of day showing the highest traffic was verified, our predictions of the resulting effects of weather were not. In plotting the average distance and resulting fares of taxi trips per month we found that the overall trend remained the same throughout. Even comparing the month with the highest average snowfall (February) to the month with the highest average temperature (June) showed no significant change. In addition to these findings, the most surprising result of our analysis was in regards to the pickup and drop-off locations. Our initial prediction was that the highest clusters of taxi rides revolved around the major landmarks of New York City. This holds true in the data as easily the highest concentrations of taxi rides revolved around Midtown Manhattan, a region home to the United Nations, Time Square, and Broadway. However what we didn't expect was that the next highest concentration shown occurs in the Upper East Side of Manhattan, a region that is not home to any significant landmarks or public service locations. In conducting further research we learned that the Upper East Side is one of the most financially prosperous neighborhoods in New York. This illuminates socioeconomic status as a factor that we did not anticipate in the frequency of New York City taxi rides. Overall our analysis revealed that there are a number of factors that contribute to the frequency of taxi rides in New York City. In addition to more obvious elements like rush hour times and famous landmarks, a significant contributor to NYC Yellow cab frequency is financial standings.

2 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 contributors 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 assorted attributes to determine and predict pick-up locations and the likelihood of an available taxi at any given time or location, as well as the viability of carpooling or pursuing alternative means of transportation 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 in large metropolitan areas like New York City. In order to optimize the taxi service business for both the company as well as the passenger, 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 extraneous conditions. With this information, not only will New York City taxi services be able to improve competition with Uber and Lyft on a cheaper and more stable fare amount, but ideally, will allow New York citizens to utilize the most cost and time efficient means of transportation available.

3 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].

4 PROPOSED WORK

4.1 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.

4.2 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. The main attributes that we will be using in our analysis will be Pick Up Location, Drop-Off Location, Trip Distance, Fare, Number of Passengers, and Time. 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. Unfortunately this particular data set isn't as normalized as we would like. Some months list pickup and drop off locations with latitude and longitude coordinates while others only list a three digit region code. 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 normalize the data between region codes and latitude and longitude coordinates to then plot the regions on a map for analysis and visualization.

4.3 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.

5 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. 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.

6 DATA SET

For our analysis we will be looking specifically at the New York City Taxi and Limousine Commission data set [1]. This data is divided into a separate CSV file for each month of a given year with each file containing information for over a million individual trips. The NYC TLC has also provided a database of corresponding zones and boroughs for each given location id. The attributes we will be examining are pickup/drop-off times and locations as well as distance travelled and payment amount. The data set contained numerous attributes but the main ones we focused on applying data mining techniques to were:

- tpep_pickup_datetime
- trip_distance
- tpep_drop-off_datetime
- fare_amount
- PULocationID
- DOLocationID
- passenger_count

7 MAIN TECHNIQUES APPLIED

7.1 Preprocessing

The data set that we obtained was from the New York City Limousine and Commissions website in which the data was mostly cleaned prior to working on it. The preprocessing techniques that were used mostly consisted of filling in any missing values through finding the mean and using WEKA to apply the filter to the data set. In addition to this, we removed any NULL values that would have corrupted our findings and dropped any attributes that weren't needed for analyzing and answering the questions we sought. The dropped values within our dataset included:

- VendorID
- RatecodeID
- payment_type
- extra
- mta_tax
- tip_amount
- tolls_amount
- improvement_surcharge

In addition to the basic preprocessing concepts filtered onto our dataset, we needed the timestamps provided tpep_pickup_datetime and tpep_drop-off_datetime to be in hour format. We used dateutil to apply the date parser across the dataset and convert the pickup and drop-off attributes to be single hour format to determine the frequency of day for pickup in the most frequent locations. Further preprocessing using the scikit allows us to normalize our attribute values to create a standardized column of values for statistical calculations.

7.2 Correlation Discovery

7.2.1 Trip Distance vs. Fare Amount

After preprocessing the dataset, we explored the attributes to determine if there are any unique correlations to be found within the dataset to help answer the questions we were after. One of the interesting questions we sought to answer was whether or not trip distance or weather factors changed the fare amount in addition to any other affects it would normally have.

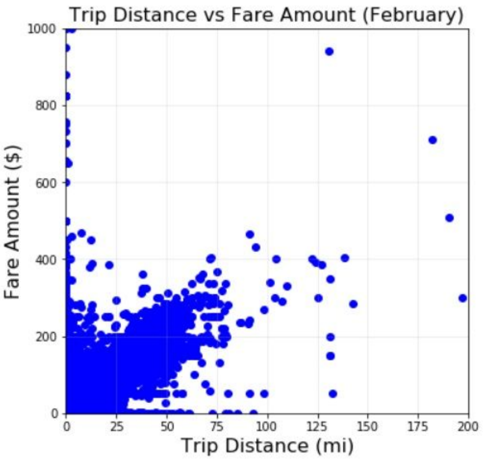

Figure 1. Correlation Scatterplot of Trip Distance vs. Fare Amount (February) with Linear Regression

As seen from the scatterplot of trip distance vs. fare amount, the correlation between the two attributes shows a linear regression model with a correlation coefficient of 0.96, meaning that even if weather contributed to the change in fare amount, the fare amount was fixed for the majority of trips. In order to determine if weather truly changed the fare amount rate we explored and analyzed the attribute correlation between trip distance and fare amount throughout the year.

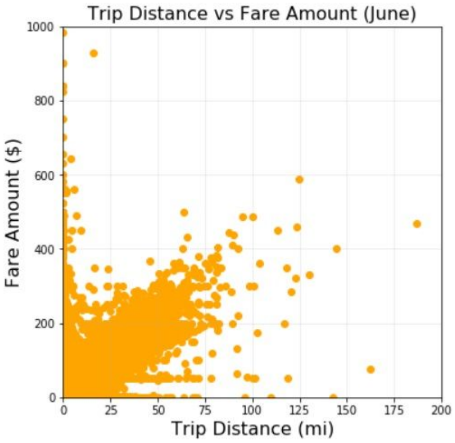

Figure 2. Correlation Scatterplot of Trip Distance vs. Fare Amount (June) with Linear Regression

Regardless of season throughout the year, with February being the snowiest month in New York

City and June being the hottest month in New York City, the linear regression model still holds and shows that the fare amount vs. trip distance doesn't change based on distance or factors such as weather, in which the fare amount remains stable and fixed. The outliers displayed within each scatterplot can be contributed to abnormally long travel distances of 75+ miles and incorrect records of fare meters being left on after the trip was already taken.

7.2.2 Trip Distance Analysis

Since there are multiple outliers within the trip distance vs. fare amount, in order to understand the general trip distances and relate this analysis to the pickup and drop-off frequency areas, we filtered our dataset to show the frequency of trip distances.

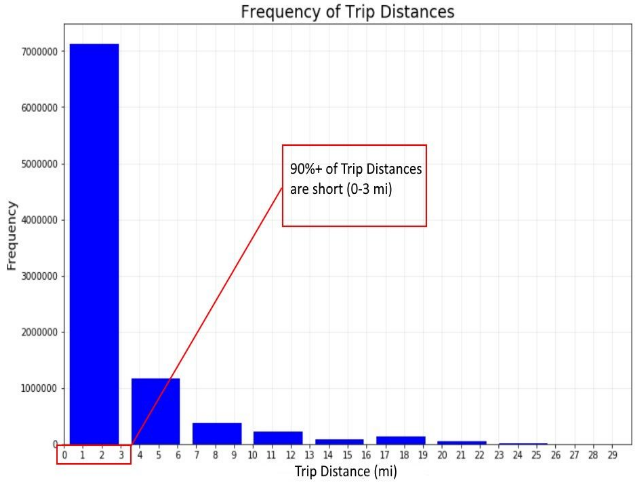

Figure 3. Frequency of Trip Distances

Our analysis shows that the majority of trip distances occurred within a short distance throughout New York City, The average trip distance was between 0 and 3 miles in length. Since the lower end of the trip distance vs. fare amount correlation displays that the linear regression model fits the trip distance vs. fare amount of 0.96 correlation coefficient, the general distance traveled within New York City tends to be particularly short. Therefore, fare amounts remain fixed except for cases in which the individual requests to go beyond central New York City limits, specifically to the John F. Kennedy International Airport.

7.2.3 Pickup and Drop-off Frequency Analysis

After discovering the average distance traveled correlation and that fares remain fixed, we wanted to explore the most frequent pickup and drop-off areas to determine the relationship between the locations due to such a short travel distance.

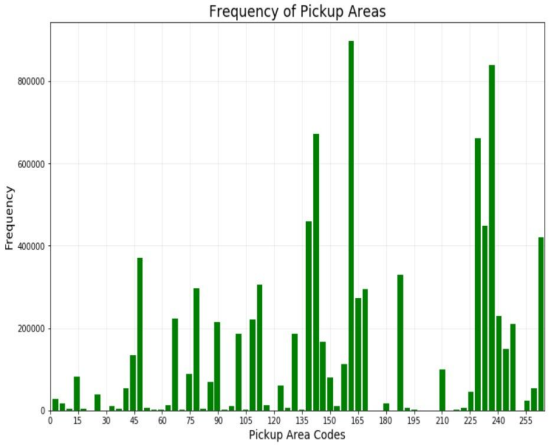

Figure 4. Frequency of Pickup Areas

As shown with Figure 4, the general frequency of pickup areas spikes within the central areas of New York City. The most common pickup locations are wealthy neighborhoods such as Upper East Side North/South and Lenox Hill as well as the central business sector of Midtown Manhattan. The least frequent pickup areas within New York City tend to be those with low income, specifically neighborhoods located in Harlem.

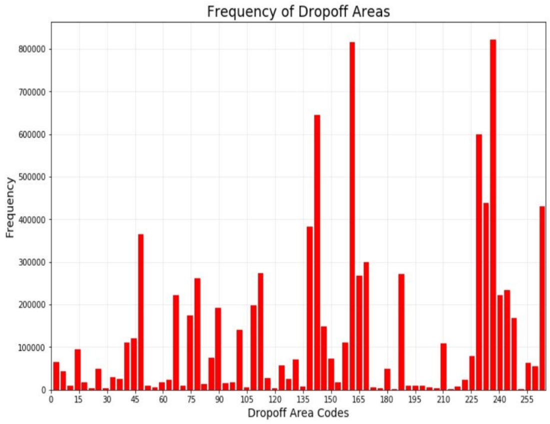

Figure 5. Frequency of Drop-off Areas

Similarly to the pickup area frequency, Figure 5, shows a general trend of drop-off areas being the same areas as pickup with Upper East Side North/South and Lenox Hill with an unusual spike in drop-offs within low income neighborhoods. As a result of the analysis of frequent pickup and drop-off locations we determined that the concentration of pickup and drop-off locations will most likely be within wealthy neighborhoods or central New York City near the most known businesses.

7.2.4 Trip Distance vs. Fare Amount

Hours	Count
9 AM - 10 AM	102,075
10 AM - 11 AM	105,134
11 AM - 12 PM	117,315
5 PM - 6 PM	142,690
6 PM - 7 PM	162,982
7 PM - 8 PM	154,134

Table 1. Frequent Hours for Pickup Locations

After running our preprocessing technique of date parsing with the dateutil parser library in Python, we discovered the most frequent pickup location hours within the day. Given the top 6 pickup areas within New York City being the wealthy neighborhoods and central New York City, the most frequent hours of pickup for taxi drivers in these areas is between 5 PM - 8 PM in the evening with less frequent pickup areas in the morning between 9 AM - 12 PM. With this information, the most concentrated hours appear to be in 3 hours intervals and correlate to rush hour times in New York City. Understanding that the distance traveled average per trip is between 0 - 3 miles in length, it appears that the most taxi usage is within the higher income neighborhoods who work a general 9 AM - 5 PM job. Since the neighborhoods of the higher income individuals are located near the central part of New York City, taxi users typically take the taxi to and back from work and live relatively close to the locations that they work at.

7.3 Supervised Learning: Classification

7.3.1 K-Nearest-Neighbor Classification

As we further analyzed our taxi data set, we hoped to find patterns among continuous variables from the locations and distances, as well as fares and passenger counts. We attempted to apply supervised learning using the K-Nearest-Neighbor Classification algorithm for the passenger count and a Naïve Bayes Classification algorithm for improved handling of nominal values, such as the pickup location and drop-off location. As the KNN Classifier is a lazy learner, it stores and holds the training portion of the inputted data for later use during the testing time. This classifier only learns from the test event using the stored training event when it reaches that particular test. In order to form predictions for our chosen attributes, this classification method requires learning from the nearest training points that are the closest to the unclassified test point. Beginning with the available statistical libraries in Python, we limited our pandas dataframe to the following attributes: trip distance, pick up location, drop off location, fare, and passenger count. In order to understand the viability of carpooling, we have applied a KNN classifier to predict the passenger count for a trip, given the input features of distance and fare. By specifying the passenger count attribute as the target, our model will learn from our training set to develop a trend to use for predictions. Using the SPSS classification tool, the learning model trained on 70% of the provided data and tested the prediction ability on the remaining 30% as shown in the Figure below of passenger predictions.

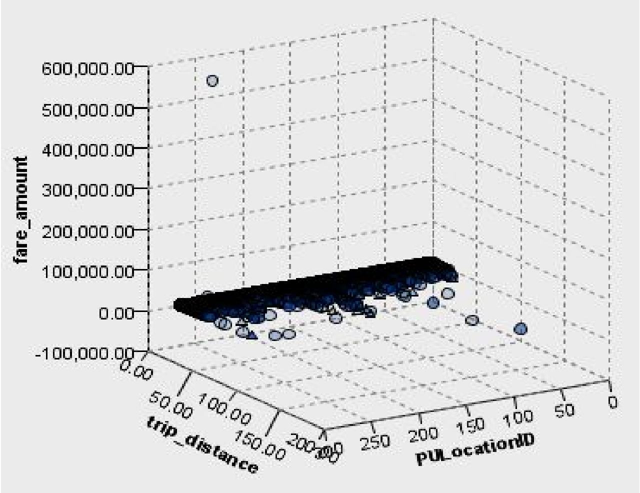

Figure 6. KNN Predictions for passenger counts as plotted points

Due to the limitations in the abilities of the KNN classifier to process nominal attributes, a Naïve Bayes Classifier would help us to learn from the location data. In order to predict the commonly occurring locations, we would need a model from the drop off location, distance, fare, and passenger count. Training a classifier on these attributes would require calculating the information gain to determine the most applicable attributes. By using this eager classifier, we would gain accuracy and speed compared to our KNN lazy classifier. In the case of this classifier, we would apply the concept of posterior probabilities, as each considered event is affected by a prior event (conditional probability). In the case of our taxi data, we could consider the probability of predicting the number of pickups in a location after learning that this particular location has experienced a drop in taxi activity. The Naive Bayes Classifier will form a prediction that an event belongs to a particular class if the posterior probability is the highest [4]. This classifier specifies that an event without a posterior probability will be equally likely to be placed within class A or class B (given two classes).

7.4 Unsupervised Learning: Clustering

7.4.1 K-Means Clustering

In order to group the most similar trips based on the pickup location and trip distance, we used a scikit learn library containing the K-Means clustering algorithm, which provides support for the nominal location values. To perform this clustering method, we created a training array and testing array from the January taxi data set. We then created a scatterplot of these arrays with the trip distance depending on the pickup location, in order to visualize the relationship before grouping these points. After the pickup locations and trip distances were combined into a two-dimensional array, we created four clusters with the centroid specified by the scikit learn library. Once our final array was fitted to our clusters, we plotted the points according to their proximity to the new centroids. This technique resulting in a closer concentration between similar points and improved the visualization of the relationship between the pickup location and trip distance. The figures 7 and 9 below show the original scatterplot followed by the results of iteratively recalculating the mean centroids (Figures 8 and 10). This unsupervised method of learning partitions our observations by choosing the closest centroids. The k-means clustering algorithm resulted in the following locations: Bushwick South (37), Flatbush(89), Maspeth(157), Van Cortlandt Park (240).

For future studies requiring the exclusion of outliers, a K-Medoids Clustering may be applied, as it is less sensitive to noise and outliers. Rather than calculating a mean value, which is used as a new point, a K-Medoid centroid is an existing point in the data, which is naturally “the most centrally located object in the cluster” [5].

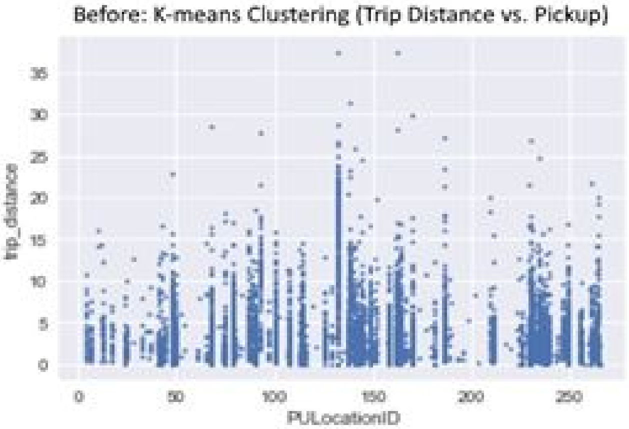

Figure 7. Before: K-means Clustering Scatterplot of Trip Distance vs. Pickup

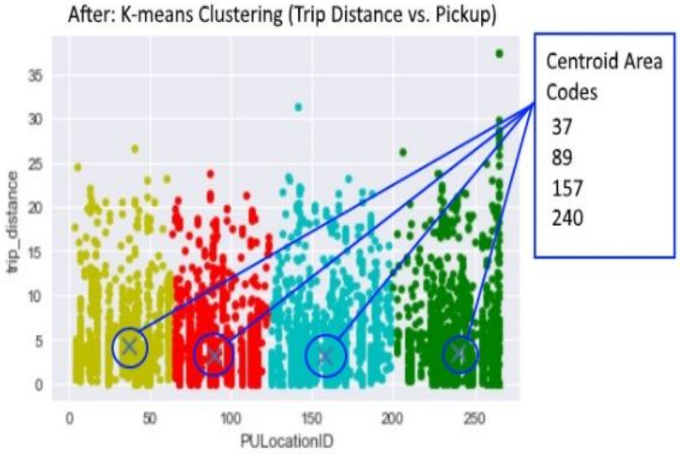

Figure 8. After: K-mean Clustering Scatterplot of Trip Distance vs. Pickup

After comparing the resulting scatterplot and clustered scatterplot for the drop-off locations, we also noticed that the pattern of clustering remained the highly similar to the patterns from the pickup data.

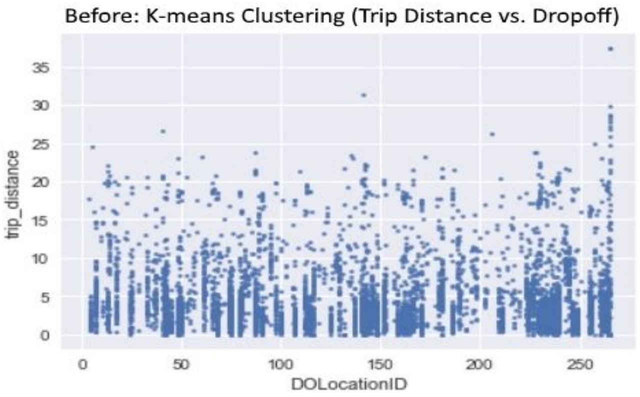

Figure 9. Before K-means Clustering Scatterplots of Trip Distance vs. Pickup

Although a comparison of the initial scatterplots display clear distinctions between the trip distances between pickup locations and drop-off locations, a strong similarity remains among the points and calculated centroids. A closer examination of our data revealed that a large portion of taxi passengers may be frequent customers and primarily rely on the yellow cab to commute or take other routine trips. The use of the K-Means clustering method, as well as other clustering methods, allows us to simply large data sets into manageable centroids. Performing a K-Means cluster carries the added advantage of greater efficiency compared to other clustering algorithms.

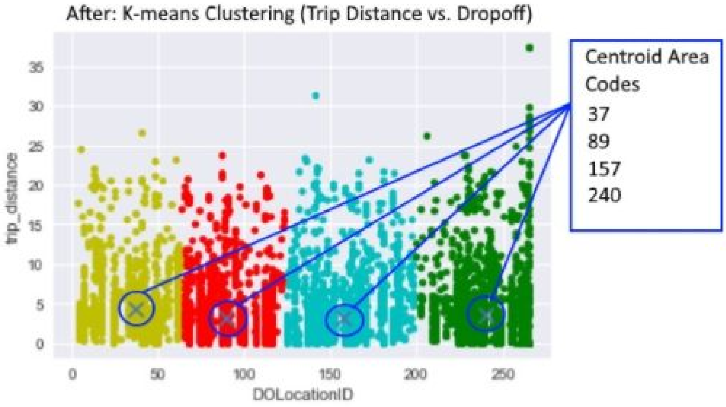

Figure 10. After K-means Clustering Scatterplots of Trip Distance vs. Pickup

8 TOOLS

In our investigation of the New York City Taxi Dataset we used many different tools. Initial scans of the data set was done using MySQL and MySQL Workbench for ease of use. In terms of our main data parsing and preprocessing we made use of Python-3 in conjunction with various libraries incorporated in Anaconda including numpy, scikit, and pandas. The last library listed was heavily utilized for additional data processing, classification, and data visualization. In addition to Python Pandas, we made use of WEKA for generating graphs and conducting supplemental data processing. The normalization option provided by WEKA also aided in our initial pre-processing to explore the effects of standardizing our trip_distances.

We also found the classification and clustering methods, in addition to the data visualization tools, provided by SPSS to be greatly helpful in quickly producing detailed graphs of our data. Our main medium for analytical data visualization came in the form of our various graphs. However, we also created a simple interactive web application to display the pickup and drop-off locations in dynamically sized distance clusters using HTML, Javascript, and the Google Maps API.

9 KEY RESULTS

Performing correlation calculations and visualizing the relationships among our selected attributes allowed us to identify the most common occurrences. From correlation of the trip distances and the fare amounts, it can be seen that the linear regression model was consistent throughout regardless of the time of year. This indicates that the fare amount remains consistent and fixed at a certain rate, especially since trip distance is quite short. This allows taxi drivers to determine the concentration of short distance trips being the most lucrative, especially within the main part of New York City area. Using an analysis of frequencies for our target attributes, we were able to identify the most popular time for taxi pickups, as well as the most frequent trip distances. A histogram of the frequencies for trip distances revealed that most taxi trips in New York City are relatively short and under three miles.

Our initial classification attempt to predict drop-off locations resulted in a low accuracy level, which may have been influenced by the test size for the split of our training and test sets. After improving the classification performance, a confusion matrix to view the specific false positives, false negatives, true positives, and true negatives would provide information to strengthen the classifier.

The use of the k-means clustering provided us with a simple and fast method of improving our ability to assess and process a large data set. In addition to choosing the appropriate clustering algorithm, we must also identify the appropriate distance formula (Euclidean, Manhattan, etc.) to measure the dissimilarity of the events. While the Euclidean distance provides the measure of a straight line between two points, the Manhattan distance provides the distance of moving horizontally or vertically along the grid of plotted points.

Visualizing the frequency of taxi pickup and drop-off locations both in the form of a histogram and in our web application revealed exactly the neighborhoods in which taxi traffic is the highest. The result that we found was that New York taxis are heavily concentrated in Midtown Manhattan. This in particular makes sense due to the presence of many high profile landmarks such as the Empire State Building, Times Square, and the United Nations Headquarters. The more surprising finding was the heavy concentration in neighborhoods like the Upper East Side and Lenox Hill in comparison to areas in direct proximity like Harlem. Upon further research we discovered that the Upper East Side and Lenox Hill make up two of the wealthiest neighborhoods in Manhattan with Harlem being one of the poorest. This indicates that rather than weather and rush hour times, the more significant contributors to NYC taxi traffic are population density and socioeconomic status.

10 APPLICATIONS

By predicting the passenger count for a given trip using the KNN Classification Algorithm, passengers in search of carpools will be more likely to find an available group. The knowledge gained from this supervised learner may be further expanded to help reduce the traffic congestion during peak periods in New York City and allow passengers to reach their destinations is less time.

The discovery of the correlations between the attributes we studied provided significant patterns and valuable information for passengers and drivers. During our analysis of the relationship between the trip distance and the fare amount for the winter and the summer, we were able to find that the distance is one of the strongest indicators of the price for a particular trip, as the addition of snow in the winter did not provide a significant change in the relationship between these attributes, due to the trips being short in distance (0 - 3 miles in length).

Further visualizations of the frequency of pickup locations and drop off locations revealed the zones in New York City with the greatest demands for taxi services. By identifying the most popular pickup locations and drop off locations, being centered around Midtown Manhattan and surrounding neighborhood areas of Upper East Side North/South and Lenox Hill, the taxi drivers can reduce the time needed to search for passengers. In addition, this opens the door for applications to be designed with the goal of providing tourists and normal citizens of New York City with recommendations on the most cost and time efficient methods of transportation based on their current location. With this information, users would be able to quickly evaluate the viability of taking a taxi versus pursuing alternate means of transportation.

11 VISUALIZATION

For an extra aspect of visualization we created a simple web application using HTML and Javascript. This heatmap of the frequent pickup and drop-off locations throughout the city helped analyze the relative location of the most frequent areas throughout New York City and understand that most of the trips are fairly short and go between the business sector of New York City centered around Manhattan and the surrounding neighborhoods of Upper East Side North/South and Lenox Hill. We also were able to notice the outlier that had frequent visits being the John F. Kennedy International Airport which is far from the central part of New York City where the concentration of short distance trips is prominently located. The primary piece of our application is the Google Maps API.

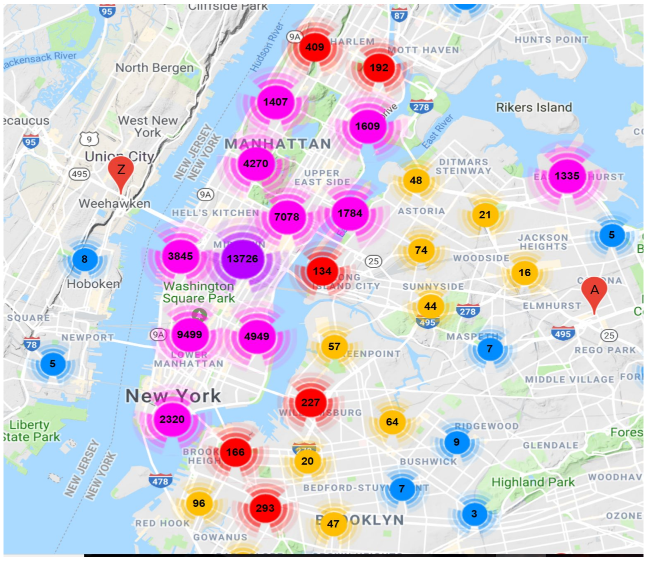

Figure 11. Web Application showing the clusters of taxi pickup and drop-off locations across New York City

We used the map to initially plot dynamically sized clusters of the pickup locations for a given month. We decided to use the clusters as it provides a fluid interactive element and presents the data in a much cleaner format than 9 million individual pins dropped on the map. After finalizing the visualization aspect we used Netlify to host the page which can be found here: https://upbeat-dijkstra-721484.netlify.com/

A HEADINGS IN APPENDICES

12 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.

[5] Jin X., Han J. "*K*-Medoids Clustering." In: Sammut C., Webb G.I. (eds) *Encyclopedia of Machine Learning*. Springer, Boston, MA 2011

[6] Yue, Yang, et al. "Mining Time-Dependent Attractive Areas and Movement Patterns from Taxi Trajectory Data." *17th International Conference on Geoinformatics*,2009.

13 DATA SET LINKS

[1] NYC Limousine and Commissions Data Set http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml