DSC680 - Project 01
NYC Taxi Cab Industry

David Suffolk

Overview

- Introduction of Problem
- Data Source
- Data Preparation
- Data Exploration
- Machine Learning Models
- Recommendations and Summary

Introduction of Problem

- The taxi cab industry of New York City has seen a significant negative financial impact since 2014.
- Taxi Medallions
  - Once valued at $1 million and difficult to obtain
  - Value has dropped and many are now sitting in inventory
- Two reasons for why the taxi cab industry is struggling
  - Increase in ride-sharing services (ex: Uber, Lyft)
  - Financiers providing bad loans for businesses to buy medallions
    - Task force created in January 2020 to investigate potential bad practices in the lending

Introduction of Problem

- Goals of this project
  - The decline in revenue is already thoroughly researched and documented
  - Approach from a business perspective
    - What are the opportunities to maximize revenue?
      - NYC residents and visitors are still using the yellow taxi cab service
      - What strategy is best for taxi cab industries to be profitable?
    - How can data help answer the questions of keeping a taxicab business profitable while the industry sees decline?
      - What does statistical analysis show?
      - Can machine learning algorithms assist?

Data Source

- The New York City Taxi and Limousine Commission provides monthly reports on Taxi Cab rides.
- The monthly reports range from January 2009 to December 2019
- The variables include pick-up and drop-off date and time, number of passengers, trip distance, location IDs for pick-up and drop-off, and total fare (which also includes a breakdown of tip, surcharges, taxes, and payment method).
- This project focused on the data for the Yellow Taxi Cab rides

Data Preparation

- Due to the size of the files, the first 10,000 rows of each month from January 2014 to December 2019 were imported
  - Unable to see standard revenue and trip numbers over time
  - However, this still provides data for 720,000 trips
- Variables changed from year to year
  - Inconsistent variables were removed (Vendor ID, Rate Code ID, longitude and latitude, Store and Forward Flag)

Data Preparation

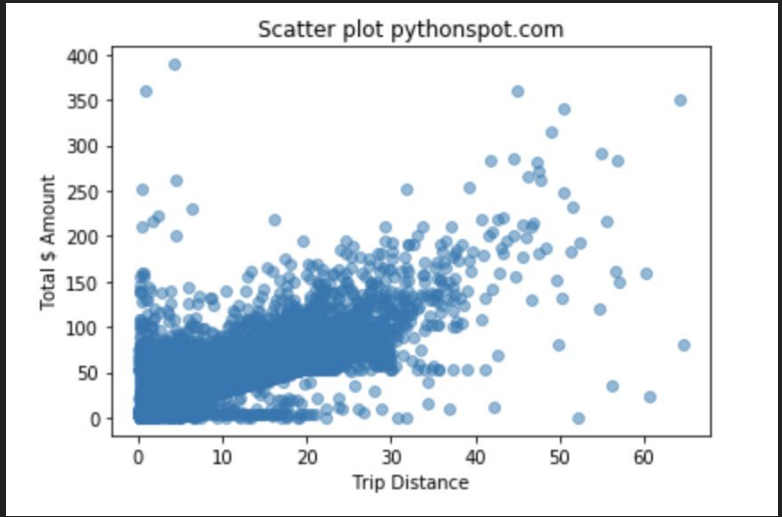
- The following variables were kept in the Master Dataset:
  - Pickup_datetime
  - Dropoff_datetime
  - Passenger_count
  - Trip_Distance
  - Fare_Amount
  - Surcharge
  - MTA_tax
  - Tip_amount
  - Tolls_amount
  - Total_amount

Data Preparation

- Pickup and Dropoff Datetime variables were separated into new columns for month, year, date, and time using Python Datetime library
- A column for Trip Duration was then created (format in seconds)
- Outliers
  - Removed rows with NA values
  - Incorrect years
  - Extremely high distances

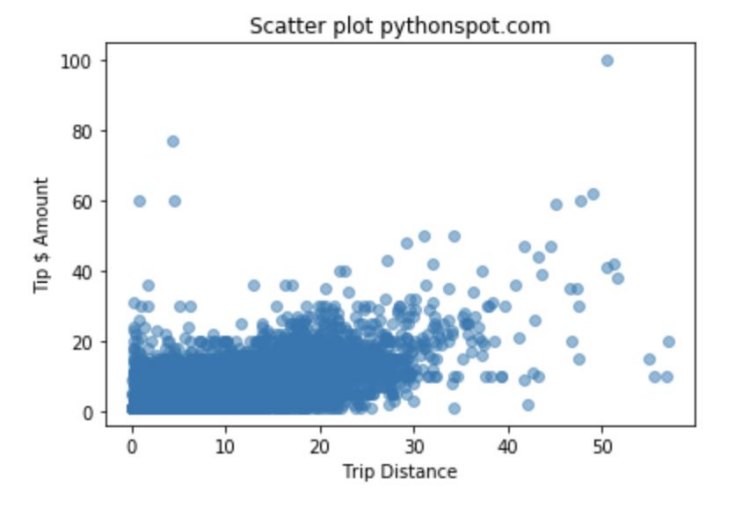
Data Preparation

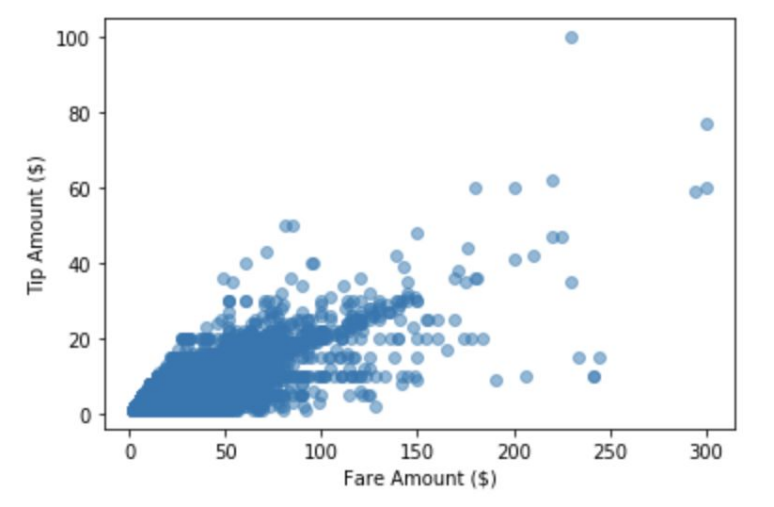
- Values of zero were researched
  - Total amount, fare amount and trip distance all needed to be greater than zero
- Total amount needed to be under $400



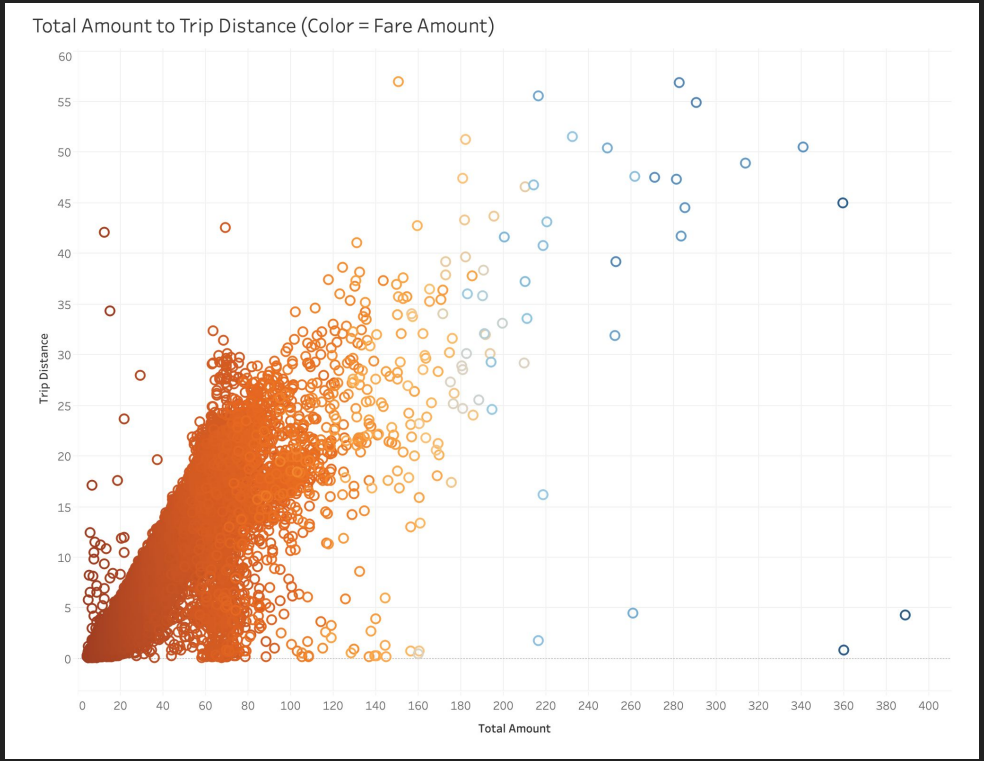
Data Preparation

- Trips with high tips but low fare amounts
  - Some records had $0 fares with tips over $50
- Created Tip Percentage column
  - Any row with a tip percentage over 75% was removed





Data Preparation



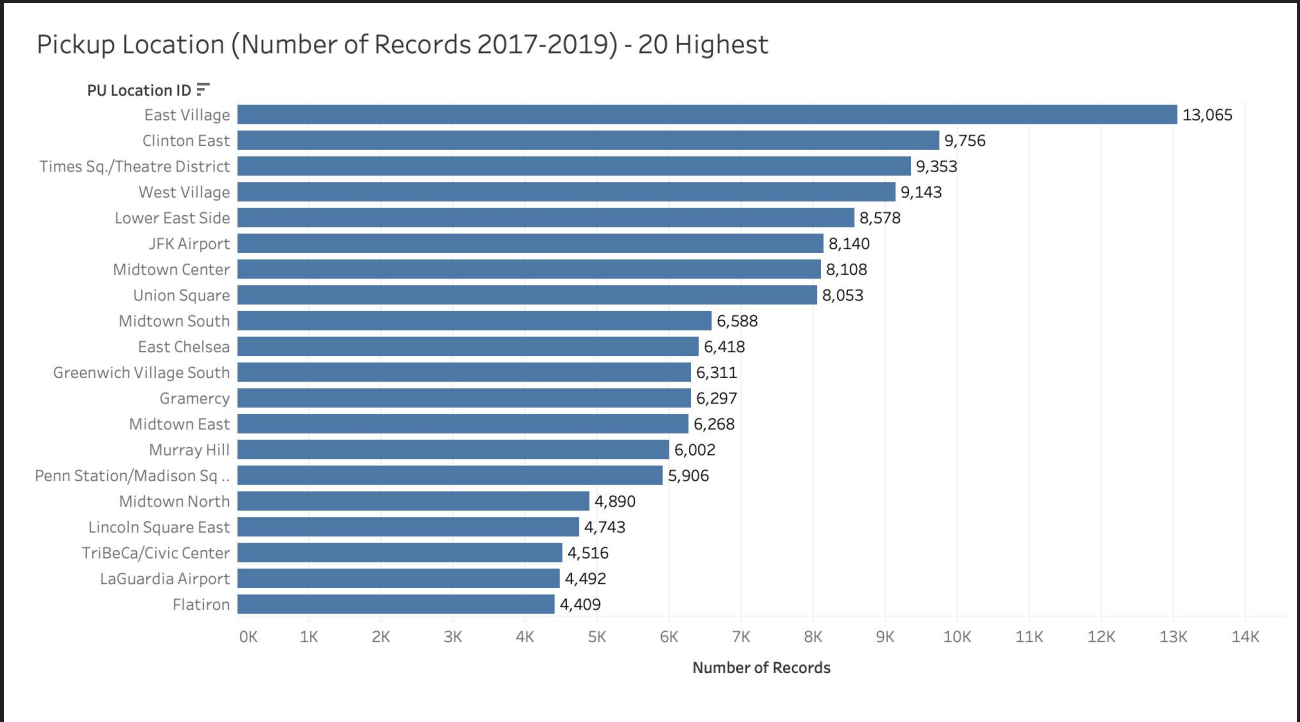
Data Exploration

- Linear Regression Models in R
  - All variables against the following three variables:
    - Total Amount
    - Pickup Month
    - Pickup Year
  - The first model based on Total Amount saw the most positive p-values compared to the other two models.
  - Any positive values were very small values and there did not appear to be any strong correlations that could be explored.
  - This still provides some possible insight. It likely means that there have been few circumstances that have changed fares, rates, taxes throughout the years that are being explored. If there were significant increases or decreases, we would likely see some correlations here.

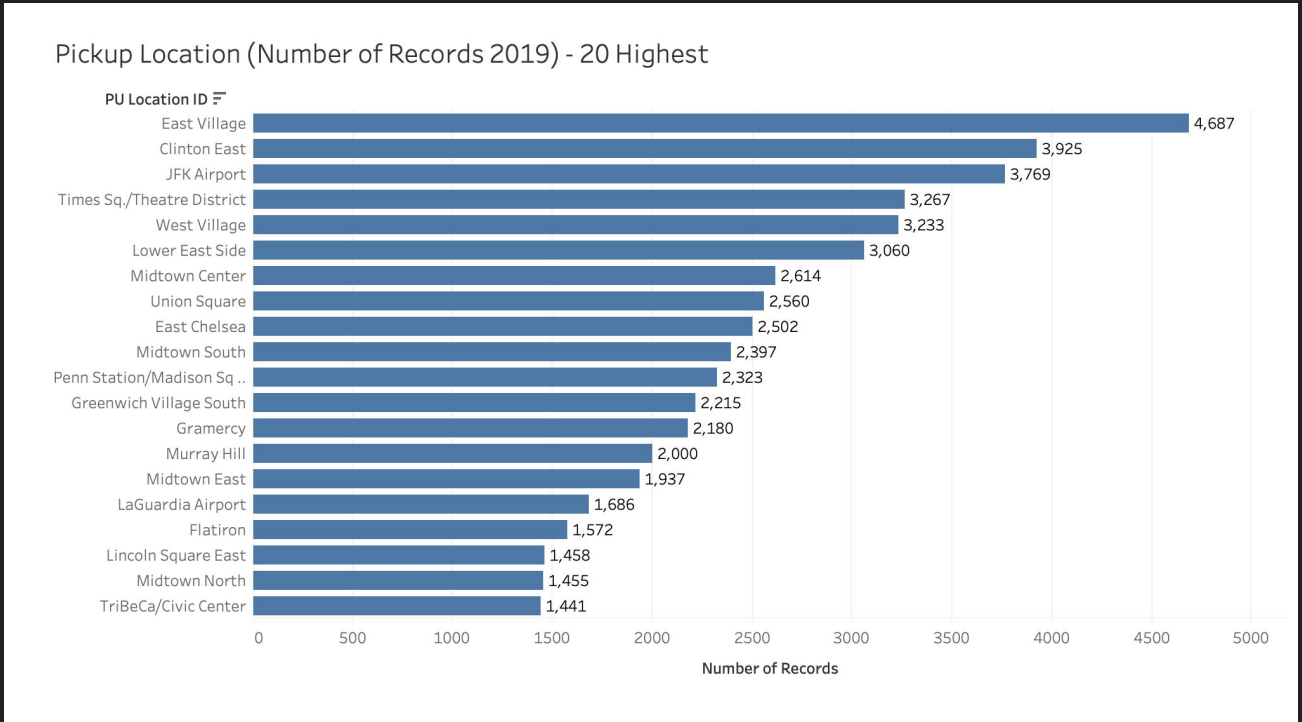
Data Exploration

- Since there appears to be no significant correlations or patterns between fares, fees, and time, the focus shifts to location
- If patterns or trends can be found in pickup location, taxi cab companies can focus on where to be ready for new rides in order to maximize revenue
- Note: Because of changes in data in how pickup location was tracked, the years explored were 2017-2019

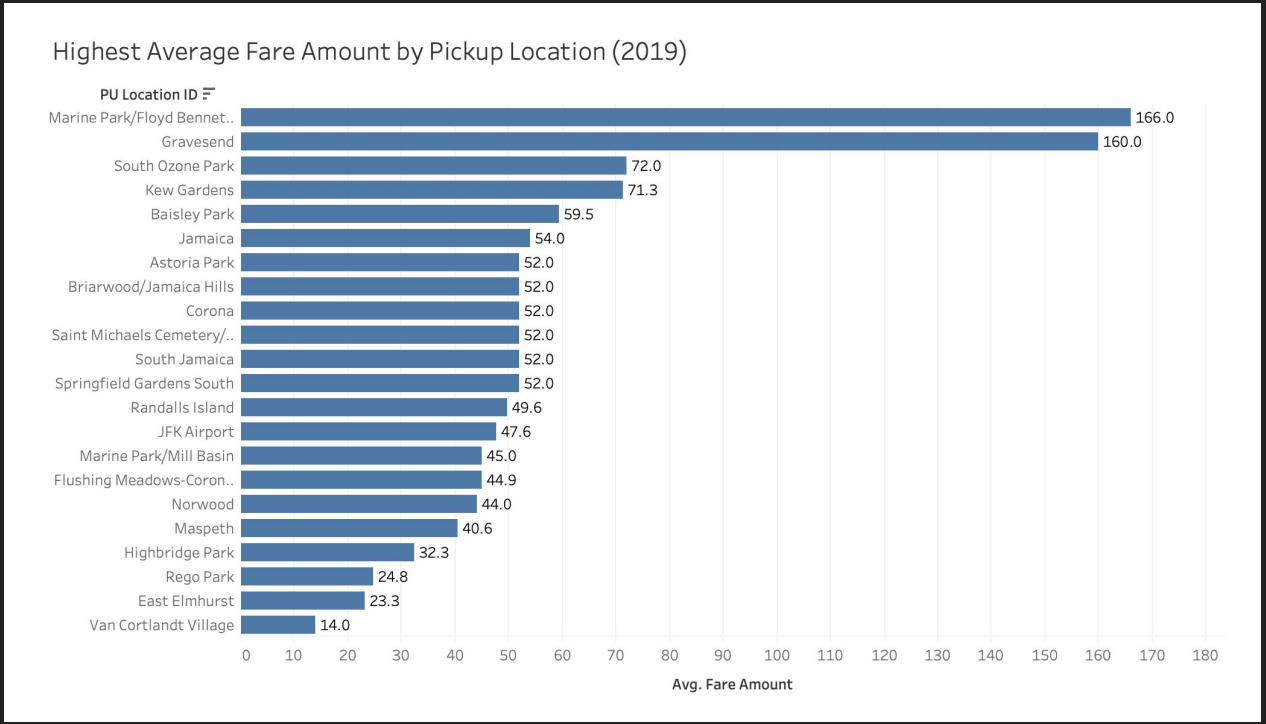
Data Exploration



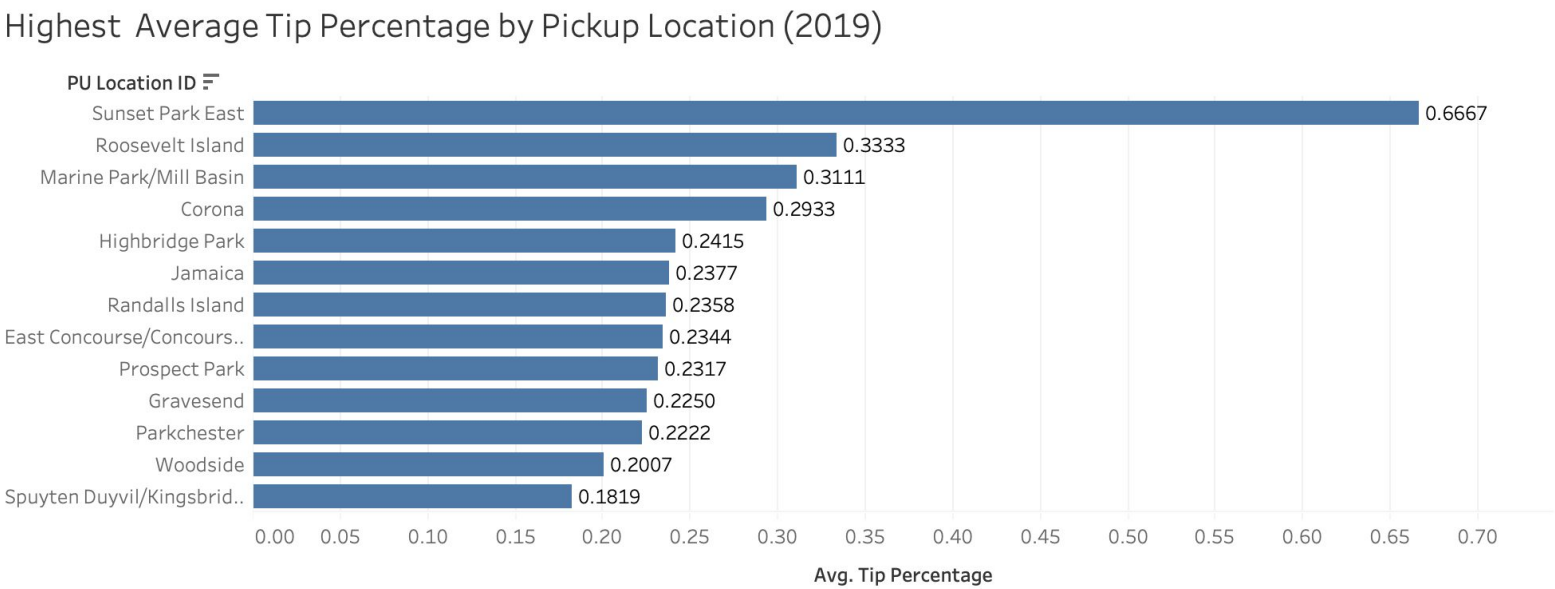
Data Exploration



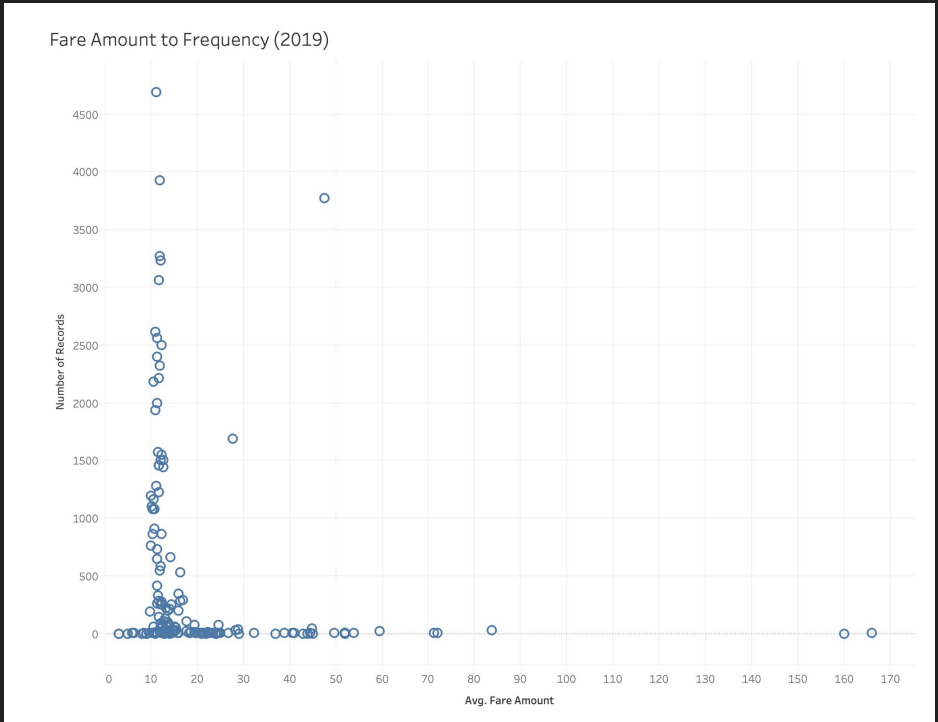
Data Exploration



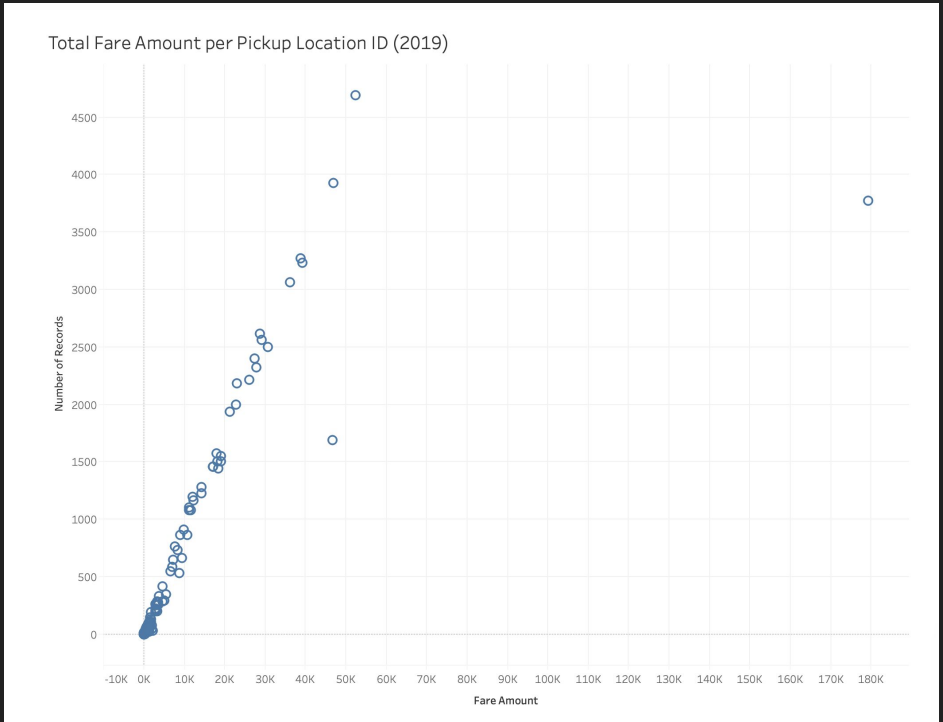
Data Exploration



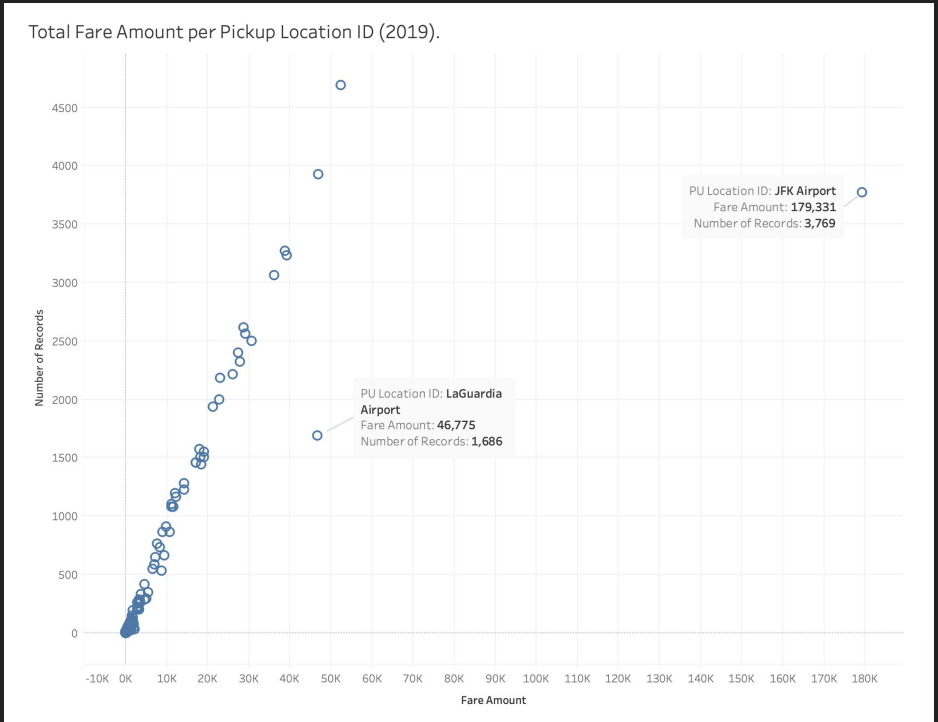
Data Exploration



Data Exploration



Data Exploration



Machine Learning Models

- Neural Networks and Ensemble Models
- Continued with the initial path of data exploration and attempted to predict the following:
  - Passenger Count
  - Pickup Year
  - Pickup Month
- The models had very low accuracy
- Pickup Year had a higher accuracy than Pickup Month
  - Possibly signals that fares and fees changed more year to year but no significant results to back this up

Machine Learning Models

- More patterns and insights when exploring Pickup Location
- 21 Target Variables
  - 9 locations with high number of records in 2019
  - 10 locations with highest average fare in 2019
  - 2 airports (JFK and LaGuardia)
- Results from initial Machine Learning Models
  - Neural Network: 34.31%
  - Random Forest (10): 32.4%
  - Random Forest (50): 35.2%
  - Random Forest (100): 35.7%
  - Bagging: 43.79%
  - Boosting: 38.6%

Machine Learning Models

- Bagging had the highest accuracy for testing and the lowest for training
- All other models had close to 100% for training models
- Indication of overfitting
  - Adjust size of test and training sets
  - Reduce the number of feature variables
- Attempted to change the size of the test/train set
- No significant changes to the results

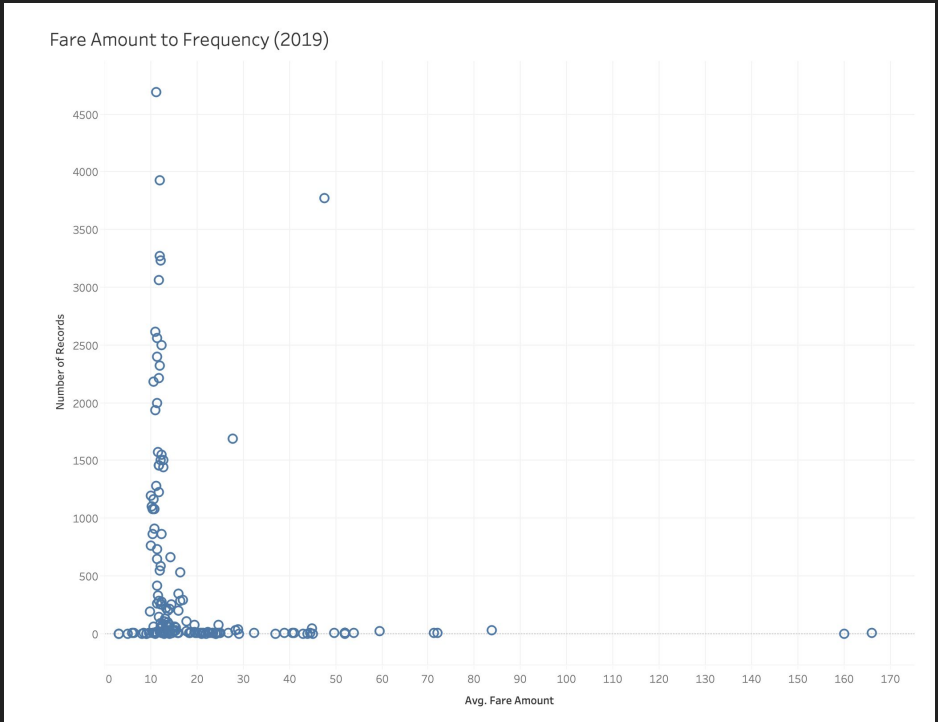
Machine Learning Models

- Performed Pearson Correlation, Spearman Correlation, and Covariance analysis on all variables to Pickup Location ID
- No significant correlations (Pearson or Spearman)
  - Results were all very close to zero
- There were some high Covariance scores from the following variables
  - Trip Distance
  - Dropoff Location ID
  - Fare Amount
  - Tip Amount
  - Total Amount
- Created dataset using only those five variables as the feature variables

Machine Learning Models

- Results of models with fewer feature variables:
  - Neural Network: 36.58%
  - Random Forest (10): 39.98%
  - Random Forest (50): 41.46%
  - Random Forest (100): 41.52%
  - Bagging: 44.16%
  - Boosting: 40.49%
- Training scores remain close to 100% indicating that overfitting may continue to be a problem

Machine Learning Models



Machine Learning Models

- Reduce the number of target variables
- 3 Pickup Locations:
  - JFK Airport
  - East Village (High frequency, low average fare)
  - Astoria Park (Low frequency, high average fare)
- Results:
  - Neural Network: 98.68%
  - Random Forest (10): 98.84%
  - Random Forest (50): 98.82%
  - Random Forest (100): 98.7%
  - Bagging: 98.77%
  - Boosting: 98.28%

Machine Learning Models

- Takeaways
  - Possibility: Pickup locations that drive revenue can be broken up into three categories
    - Distinctions between the specific pickup locations are not significant
  - Decision Trees (part of ensemble models) will almost always perform better with fewer target variables
  - Training models continue to be close to 100% but overfitting may be less of a problem since the Testing models are also close to 100%
- Next Steps
  - Add additional locations in each of the three categories and run models again

Machine Learning Models

- 6 Target Variables
  - Airports: JFK and LaGuardia
  - High Frequency, Low Average Fare: East Village and Clinton East
  - Low Frequency, High Average Fare: Astoria Park and Jamaica
- Results:
  - Neural Network: 80.41%
  - Random Forest (10): 86.79%
  - Random Forest (50): 87.98%
  - Random Forest (100): 88.16%
  - Bagging: 90.29%
  - Boosting: 89.42%

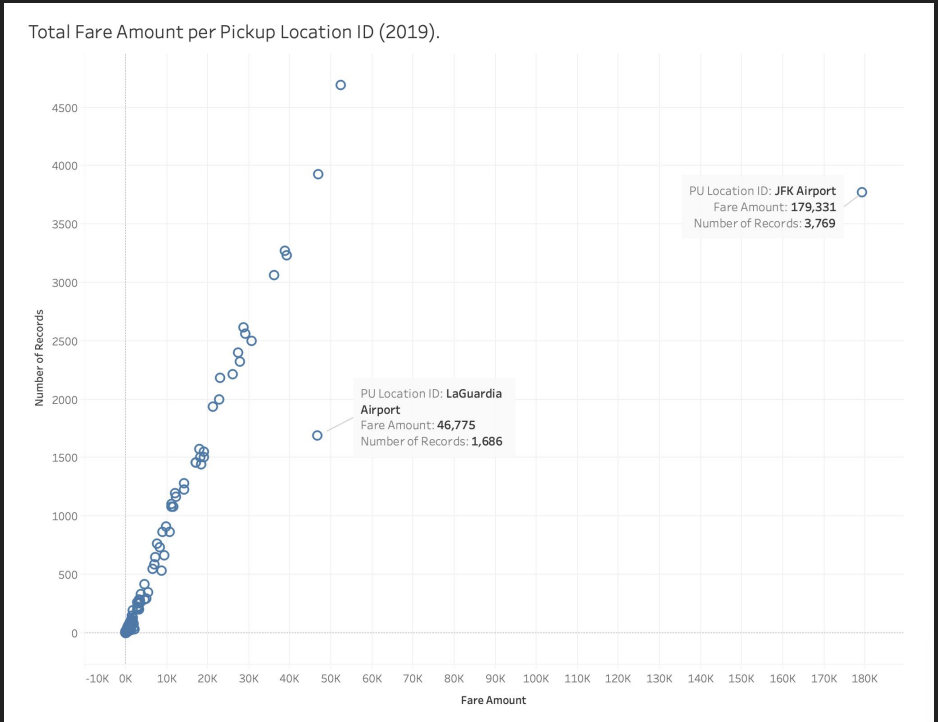
Machine Learning Models

Model	All Feature and Target Variables	Reduced Feature Variables	3 Target Variables (with Reduced Feature Variables)	6 Target Variables (with Reduced Feature Variables)
Neural Network	34.31%	36.58%	98.68%	80.41%
Random Forest (10)	32.4%	39.98%	98.84%	86.79%
Random Forest (50)	35.2%	41.46%	98.82%	87.98%
Random Forest (100)	35.7%	41.52%	98.7%	88.16%
Bagging	43.79%	44.16%	98.77%	90.29%
Boosting	38.6%	40.49%	98.28%	89.42%

Recommendations and Summary

- Goal: Increase opportunity for revenue
- Drivers should be ready to pick-up at airports and Manhattan
  - JFK consistently busy with high fares. LaGuardia is a similar trend.
  - Locations with high frequency/low average fare are in Manhattan
- While locations in the boroughs have a higher average fare, their frequency is much lower
- The lower fare but higher frequency locations are frequent enough to drive total fare amount

Recommendations and Summary



DSC680 - Project 01

David Suffolk