IST 718 SPRING 2024

BIG DATA ANALYTICS

Navya Kiran V B (SUID: 291769531)
Jovita Andrews (SUID: 522110149)

1. PROBLEM STATEMENT

The main objective of the project was to develop a model to predict the yellow taxi fare amount in New York City. We considered a range of factors within the dataset as well as weather related factors. We utilized PySpark, Spark SQL, Google BigQuery and Python for the process of analyzing the data and building predictive models.

2. DATA DESCRIPTION

We obtained the data from two different sources. Our first source was the nyc.gov website for the yellow taxi trip data and taxi zones data. To enable efficient processing of the dataset, we made the decision to utilize January and February trip data. The other source was Google BigQuery's open dataset. We obtained the weather and weather station data from BigQuery.
We utilized queries written in Spark SQL to combine the data from these tables resulting in a dataset with over 500k rows and 43 columns.

3. PREDICTION AND INFERENCE GOAL

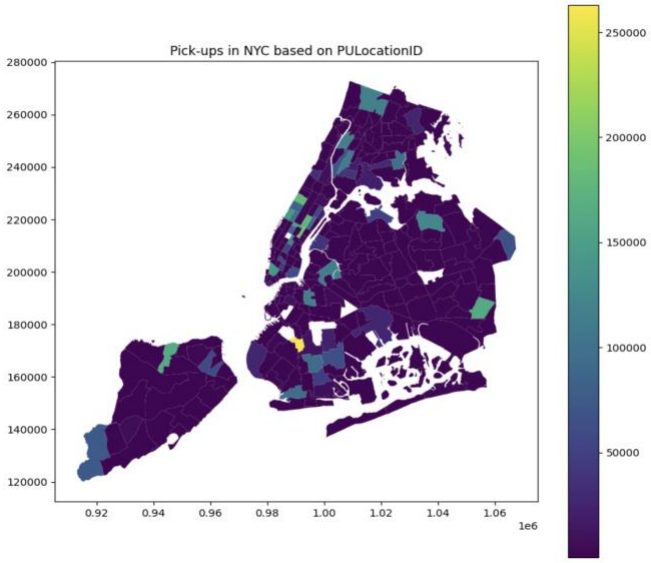
Conduct exploratory data analysis in order to identify factors that influence the total fare amount of a yellow taxi in New York City and provide actionable insights to stakeholders to enable better resource optimization.
Utilize Apache Spark's MLLib library in order to build predictive models using linear regression, decision trees and random forest with fare amount as the dependent variable.

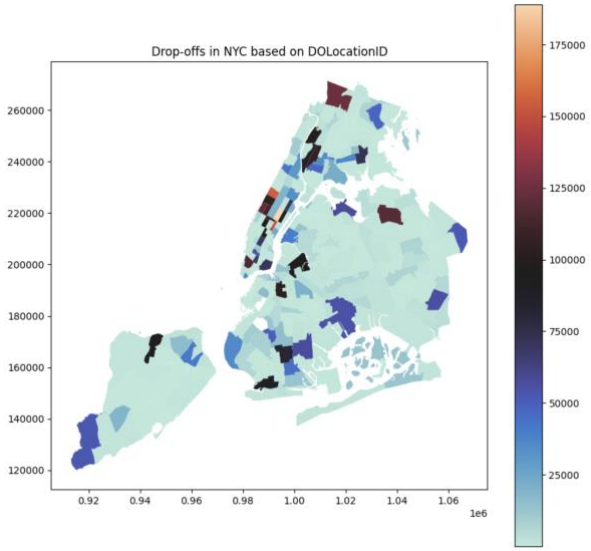
4. DATA CLEANING

The steps performed in the cleaning of the dataset are as follows:

- Removal of Null Values:
    - Null values, indicative of cancelled rides or no-shows, were removed from the dataset to focus on completed trips only.
- Handling Outliers:
    - The trip distance was more than 1 mile and less than 100 miles.
    - The total amount charged was between $3 (the minimum fare) and $200.
    - The passenger count was less than 5 passengers per vehicle.
    - The pickup and drop-off dates were between January 1, 2023, and February 28, 2023.
- Creating new columns to represent day name, week number and day of week of the pickup date.
- Converting pickup and drop off date to datetime columns.
- Some weather-related columns, despite containing integer data were represented as strings. So, we casted these columns to integer type.

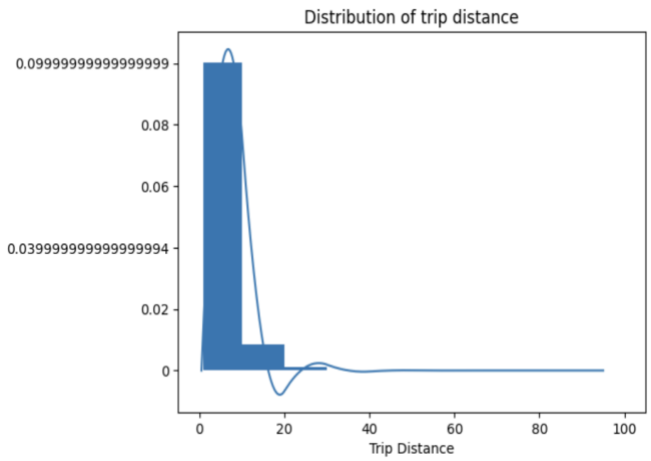
5. EXPLORATORY DATA ANALYSIS

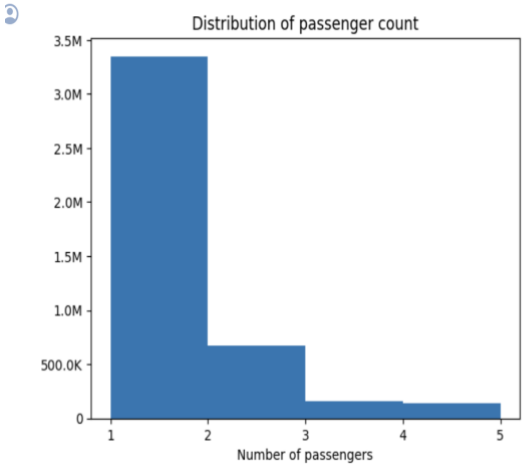




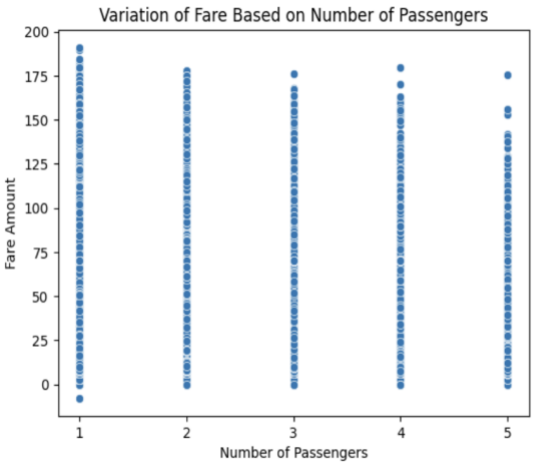
The first visualization represents the number of trips based on the pick-up location. As we can see, most number of pick-ups are from Manhattan. Within Manhattan, the most number of pickups are from Midtown Center.

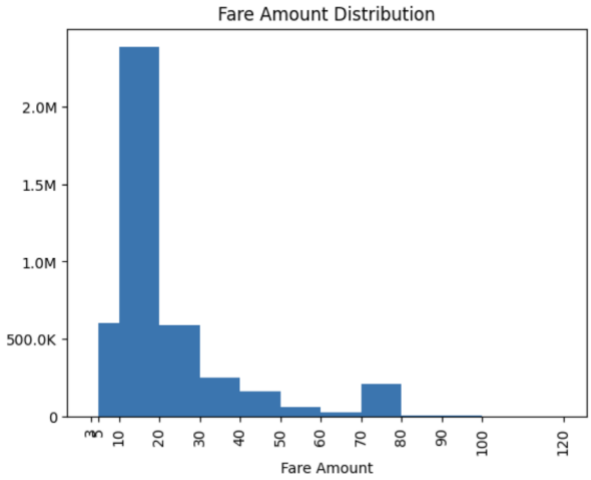
The second visualization represents the number of trips based on the drop-off location. As per the graph, the greatest number of drop-offs are in Manhattan. Within Manhattan, the most number of drop-offs are to Upper East Side N.





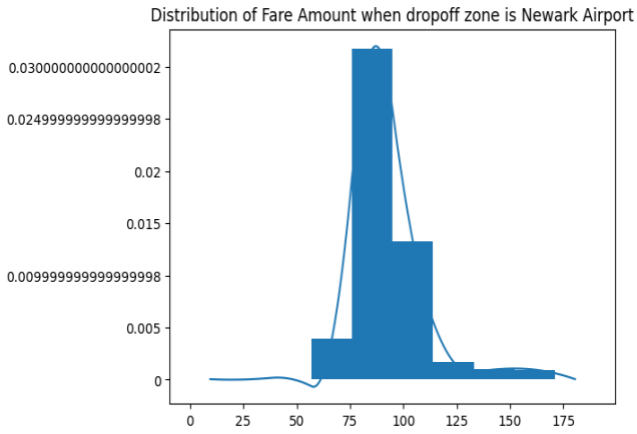
As per the above visualizations, we can see that most of the trips are short distance trips with a single passenger count.

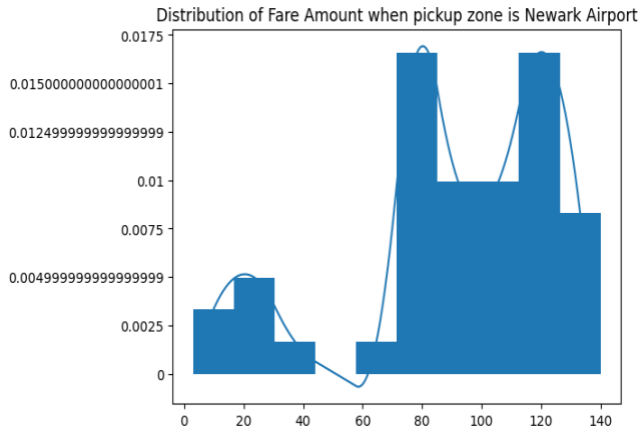




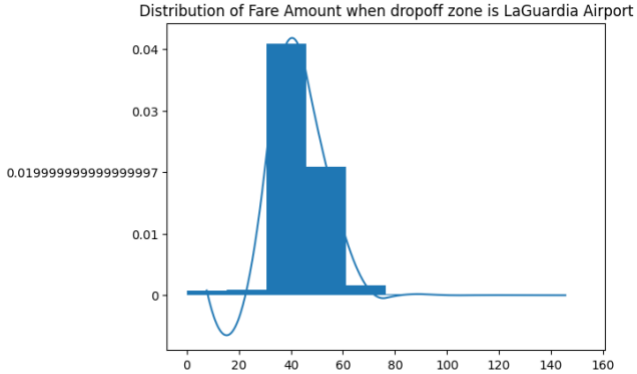
The number of passengers does not influence the fare amount, since the fare amount is distributed over a large range regardless of the number of passengers. Also, most of the trips have a fare amount between $10 and $20.

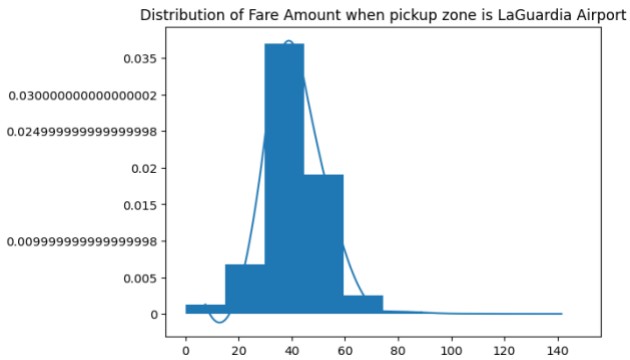
The highest average fare amount based on pick up and drop off location is EWR, indicating that people to/from EWR airport have pickups/drop-offs within the limits of the city.



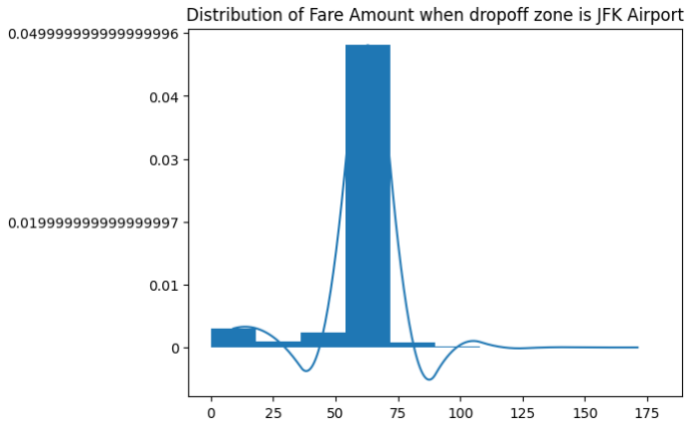


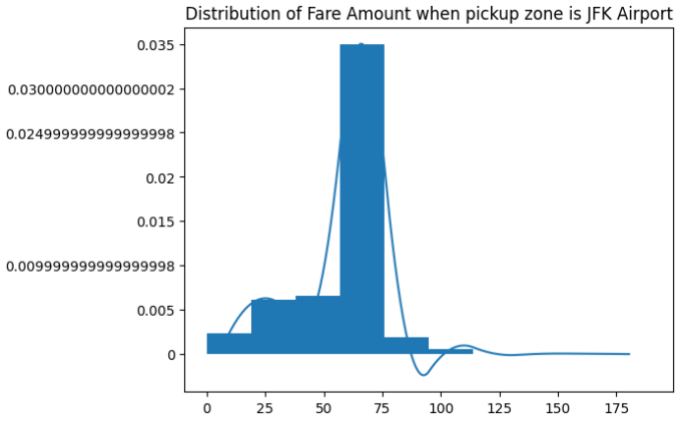
When the drop off zone is EWR, then the fare amount ranges from $75-$100 indicating that people come to EWR from far off locations. When the pick up zone is EWR, then there are 3 peaks, once at $20 indicating short distance trips and the other two are at $80 and $120 indicating far off pick up zones.



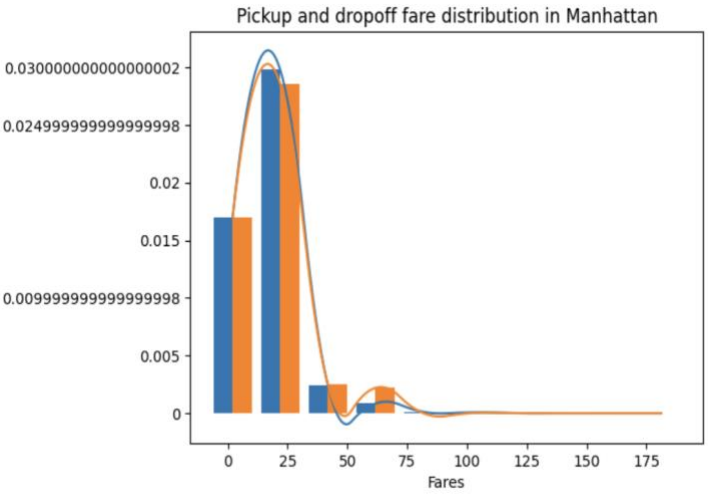


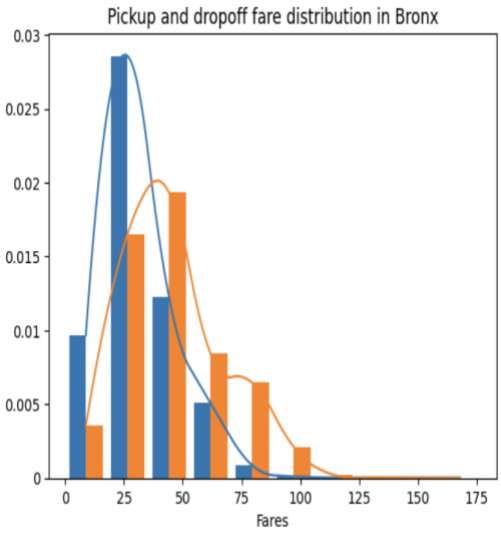
When the drop off as well as pick up zone is LaGuardia airport, then most of the trips have a price between $40-$60.



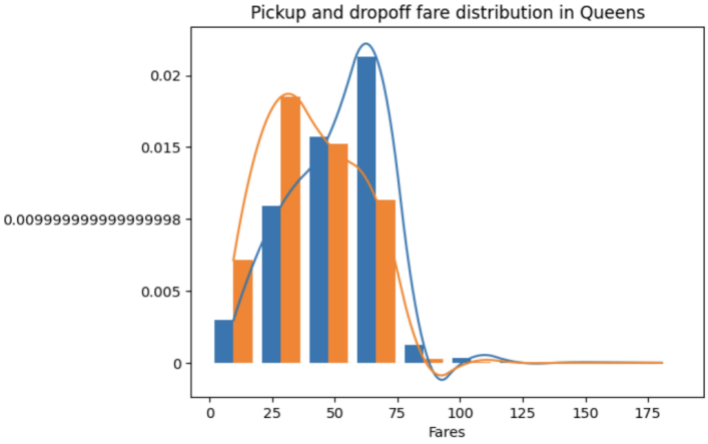


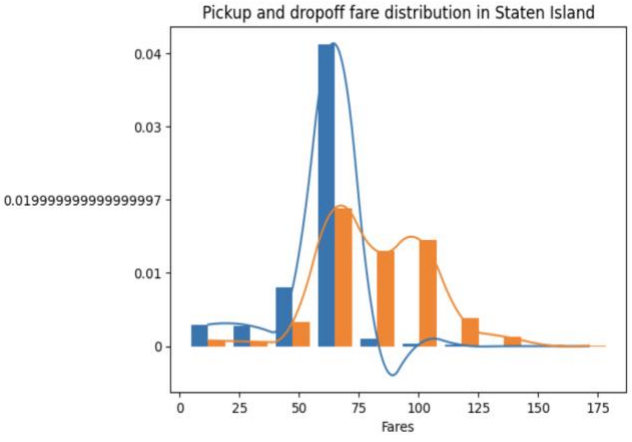
When drop-off zone is JFK, then most of the trips have fare amount between $55-$75 and when the pickup zone is JFK, most of the trips have a fare amount between $60 - $75.



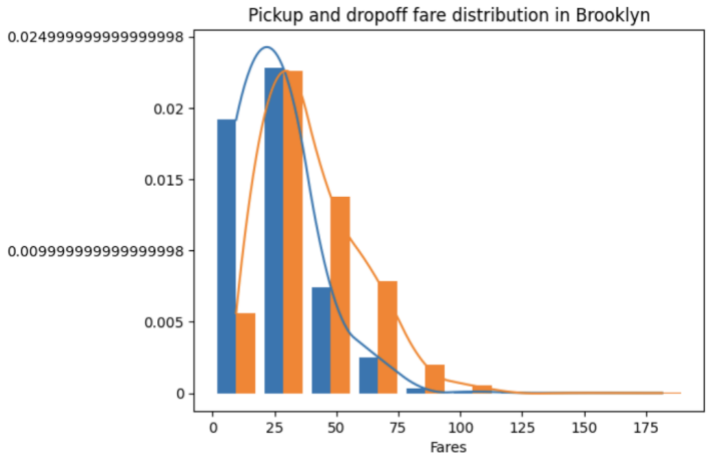


The pick up and drop off fare in Manhattan are similar to each other with a peak at $25. The pickup and dropoff fares in Bronx are different from each other, with pickup fares being greater than drop off fares. The pickup fare has a a peak at $25 and the drop off fare has a peak at $50.

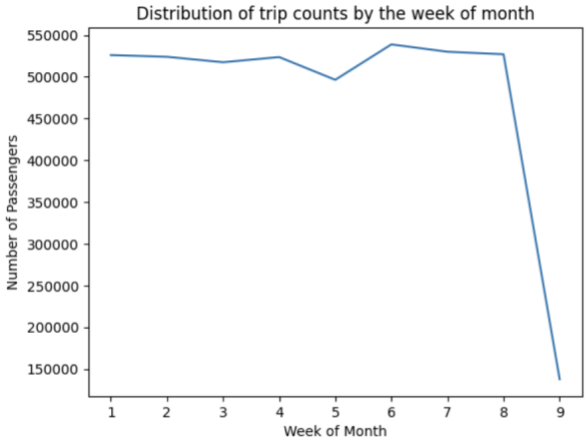


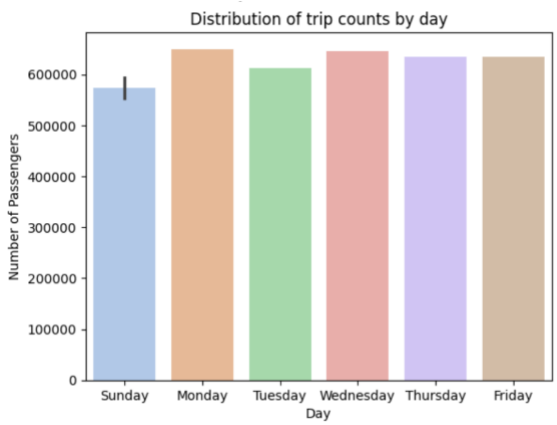


The pick up and drop off fares for most rides to/from Queens is relatively expensive with most pick-ups having a fare amount of 65$ and most drop-offs having a fare amount of $30. When the pickup is Staten Island, most of the rides including pick-ups and drop-offs have a fare greater than $50.

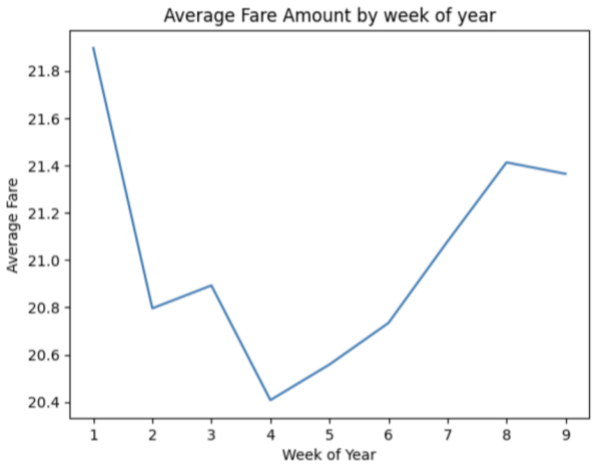


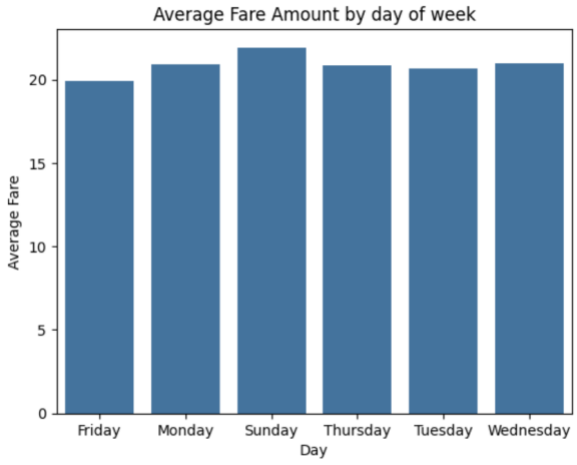
The fare amount for most rides with pickup and drop off in Brooklyn have similar fares with a peak at $25.



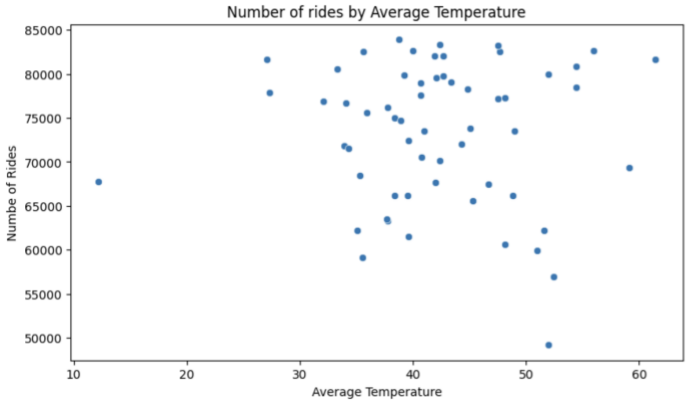


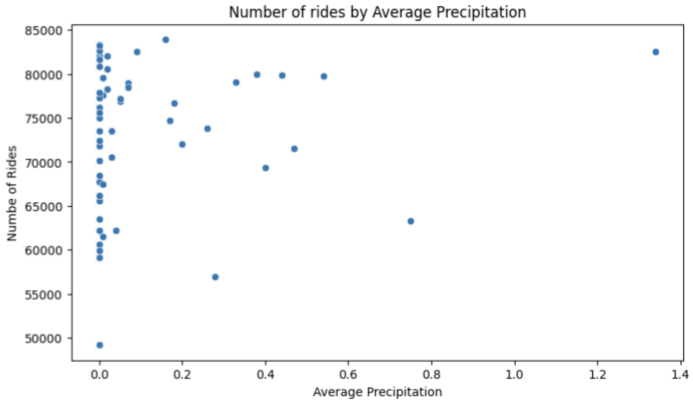
Since, we utilized data from the month of January and February, there are only 9 weeks in the dataset. The number of rides over the weeks stay fairly consistent with a slight dip in week 5 of the year. Across the days of the week, the number of rides stay the same with a slight dip on Sunday.



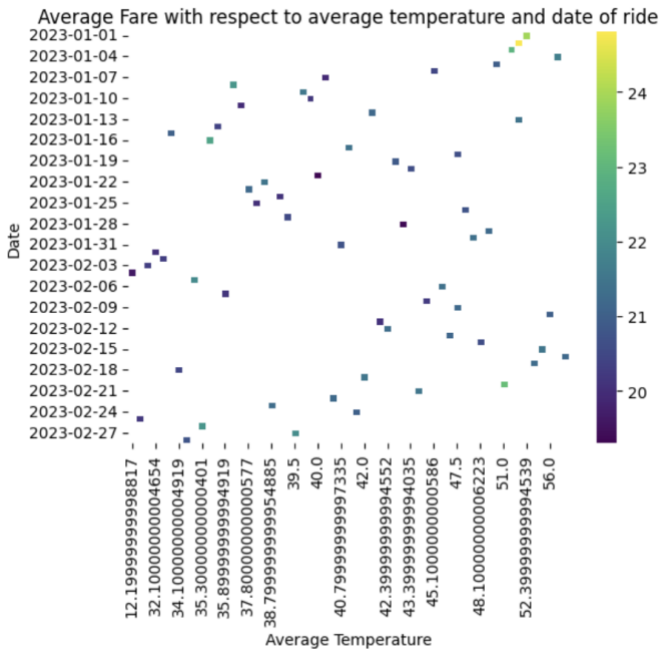


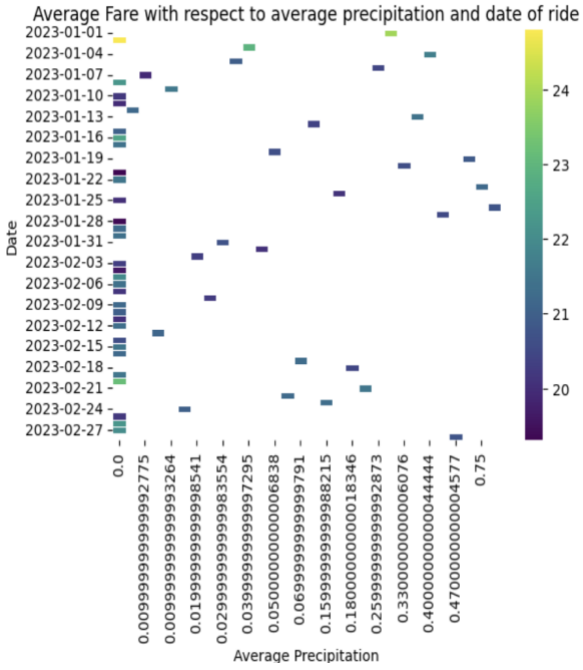
When we look at the distribution of average fare amount throughout the weeks, we see that there is a higher average fare amount in the first week of the year, indicating higher movement then followed by a dip in the fourth week of the year. Also, the average fare amount stays consistent throughout the days with it being slightly higher on Sunday.





The above visualizations show the correlation between fare amount and the average temperature/average precipitation. With the increase in average temperature, there is an increase in number of rides, indicating that people move around in better weather conditions. Also, when the precipitation is zero, then there are more number of rides when compared to higher precipitation. This indicates that people prefer staying inside when its raining.





The above plots show the variance in fare amount across the dates and average temperature/average precipitation. In January, with the increase in temperature, there is an increase in the average fare amount. Zero precipitation shows a wide range of average fare amounts with an increase in precipitation showing an increase in the average fare amount.

Something that was surprising during the analysis process was that we expected the number of rides to increase with the increase in amount of precipitation but that did not seem to be the case. Also, we expected lower average temperatures to have higher number of rides. But with the proliferation of work-from-home, people prefer staying inside during harsh weather conditions.

Based on the above analysis process, we identified certain key drivers and have included them as the predictor variables for our machine learning models. We chose to go with three predictive models, namely, linear regression, random forest regressor and decision tree regressor.

6. MACHINE LEARNING METHODS USED TO SOLVE THE PROBLEM

Before applying the predictive models, the following steps were performed:

1. **Feature Selection**: We selected features relevant to fare prediction including trip distance, pickup and dropoff boroughs, fare amount, payment type, and various weather conditions like temperature, visibility, wind gust, precipitation, and occurrences of fog, rain, snow, hail, and thunder.
2. **Categorical Feature Transformation:** Categorical features such as 'pickup_borough' and 'dropoff_borough' were first indexed using the StringIndexer. This process assigns an index to each category from 0 to (number of categories – 1). Then, we applied OneHotEncoder to the indexed categories to encode them.
3. Some variables were integers being represented as strings, so we casted these variables into integer type before using the vector assembler.
4. **Feature Vectorization:** The VectorAssembler was utilized to combine all feature columns into a single vector column. This was done since MLLib requires the input to be presented in the form of a single vector rather than a number of predictor columns.
5. **Feature Scaling:** The StandardScaler was used to normalize the features by scaling them to have a mean of zero and a standard deviation of one, to ensure all features contribute equally to the result. This was done since each of the predictor columns were of a different scale.
6. **Train and Test set:** The dataset was split into training and test data with training data being 80% of the input dataset and test data being 20% of the input dataset.
7. **Model Training:** We trained three different regression models:
    - Linear Regression: A 'LinearRegression' model was trained on the feature vector.
    - Random Forest Regressor: A 'RandomForestRegressor' model was trained to capture non-linear relationships and feature interactions.
    - Decision Tree Regressor: A 'DecisionTreeRegressor' was used to model the fare amount prediction as a hierarchical decision-making process.
8. **Model Evaluation:** The root mean squared error (RMSE) was the basis of our evaluation. The results obtained are as follows:
    - The Linear Regression model achieved a test RMSE of 3.75, indicating the average error margin of the fare predictions from the actual values.
    - The Random Forest model resulted in a slightly higher test RMSE of 4.22 compared to Linear Regression, which might suggest overfitting or a need for hyperparameter tuning.
    - The Decision Tree model performed comparably to Linear Regression, with a test RMSE of 3.71, suggesting good generalization with a simple model

7. RESULTS SUMMARY

- In Queens, Bronx, Manhattan, Brooklyn, and Staten Island, the fare distributions show distinct peaks suggesting common fare ranges within each borough. Manhattan has the highest frequency of fares in the middle range, which might indicate a combination of trip lengths and traffic conditions unique to Manhattan.

- Fares in the outer boroughs like Queens and Staten Island show wider distributions, suggesting more variability in trip distances and possibly less consistency in pricing or travel patterns compared to Manhattan.
- The overall fare amount distribution across NYC indicates a strong peak at the lower end, which suggests that short to medium-range trips are most common.
- The steep drop-off past the initial peak implies that high-cost rides are relatively rare, which could be due to long-distance trips being less frequent or alternative transportation options for such distances.
- The scatter plot does not show a clear correlation between the number of passengers and the fare amount, implying that the number of passengers may not significantly impact the fare, considering that fares are often flat rates irrespective of passenger count, up to the vehicle's capacity.
- Single-passenger rides dominate the distribution, indicating that most taxi trips are taken solo. There's a sharp decline in frequency as the number of passengers increases, suggesting that group rides are less common.
- Trip distance distribution shows that short trips are more common than longer ones, with a steep decrease as distance increases.
- The density of pickups and dropoffs in specific areas highlight the hotspots for taxi activities. Central and Downtown Manhattan have the highest concentration of taxi activity, while airports also show significant activity levels.
- Drop-off heatmaps show that some areas are more popular as destinations than as pickup locations, which could be indicative of commuting patterns or popular venues/attractions.
- There's a notable trend in trip counts varying by day of the week, with a significant drop on Sundays, which may reflect lower demand for taxis on the traditional "rest day" or non-workday for many people.
- Week-of-month trip distribution displays a dip towards the end of the month. This might be due to monthly financial cycles affecting transportation spending.
- Histograms for fares to and from JFK, LaGuardia, and Newark airports show peaks at certain fare amounts, which could indicate standard airport rates or common distances from key areas in the city to these airports.
- Scatter plots examining fare against weather conditions suggest a weak correlation between precipitation or temperature and the average fare, though ride frequency seems to decrease slightly on days with higher precipitation, which might be due to people preferring not to travel in adverse weather conditions.
- Temperature appears to have a more complex relationship with the number of rides, where both very low and high temperatures might be associated with increased taxi usage, possibly due to discomfort from walking in such conditions.

- The average fare seems to fluctuate slightly over the weeks, but there isn't a clear upward or downward trend, indicating that fare prices are relatively stable over time. However, there's a slight increase in average fares during mid-weekdays, which could be related to higher business activity.

8. CHALLENGES

The main challenge we encountered during this project was related to the volume of the dataset. Since our systems have limited processing capacity, we were able to only use data for January and February. Expanding the dataset to include data for the whole year would give us a better model with a lower RMSE.

9. SUMMARY

The main goal of our project was to build a predictive model to determine the fare amount of a yellow taxi trip in New York City. We were able to achieve a low RMSE which could further be reduced by the utilization of a larger dataset.