**CS 105 Final Report**

**Group 9**: Samuel Ha, Jonathan Kim, Sathya Rajesh, Ian Oh

**Link to Presentation:** <https://drive.google.com/file/d/1DLoJxg3UXNmzIM7HljTawyzXCeoj7ET0/view?usp=sharing>

**Project Description / Proposal**:

We want to predict whether a taxi trip occurs during day or night depending on several trip features given to us by the dataset (Day is defined as 6AM to 6PM while night is anytime outside of that). The Taxi & Limousine Commission (TLC) records all trips provided by yellow and green taxis. We chose to use the dataset from January 2018. Our goal is to utilize the given features (e.g. trip\_duration, total\_amount, trip\_distance, passenger\_count) in order to make predictions of when they rode the taxi. The predictions will be made based on models from KNN Classification and Decision Tree Classification.

<https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>.

**Data Collection**:

We wanted to pick a year that was pre-Covid for our dataset. Covid has impacted the way people work and commute so we wanted to avoid using years starting from 2020. As a group we chose 2018.

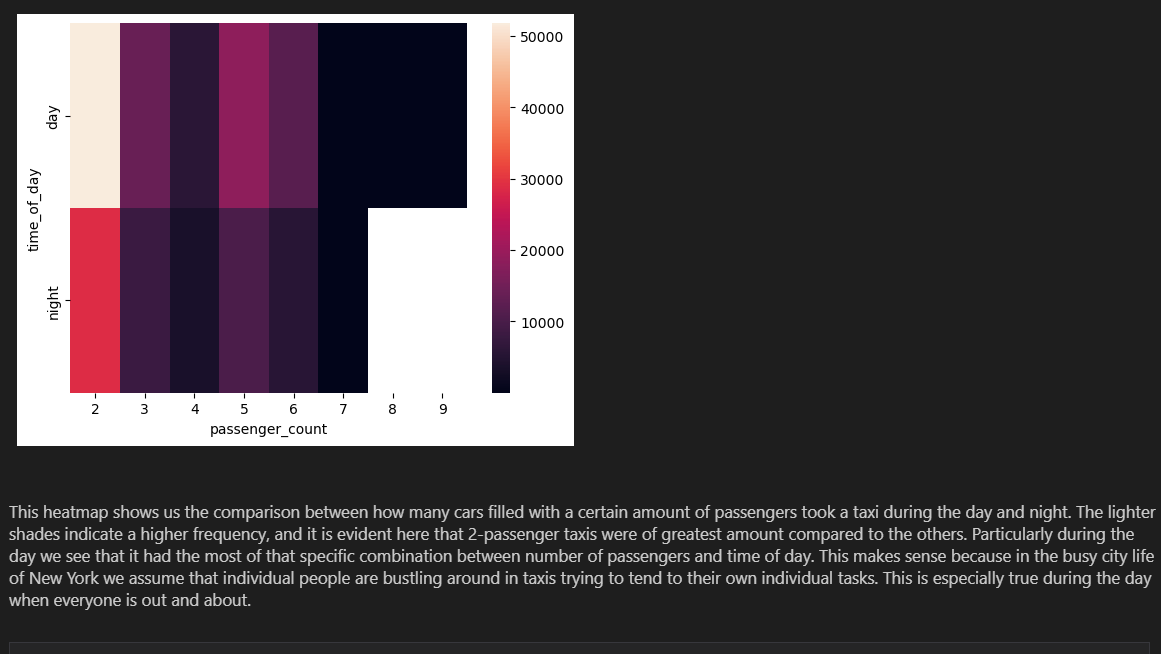
**Data Cleaning:**

The data, provided by the TLC, comes as a PARQUET file. Parquet is optimized to work with complex data in bulk and features different ways for efficient data compression and encoding types. In the month of January of 2018, there were 8 million entries alone. This was before any data cleaning.

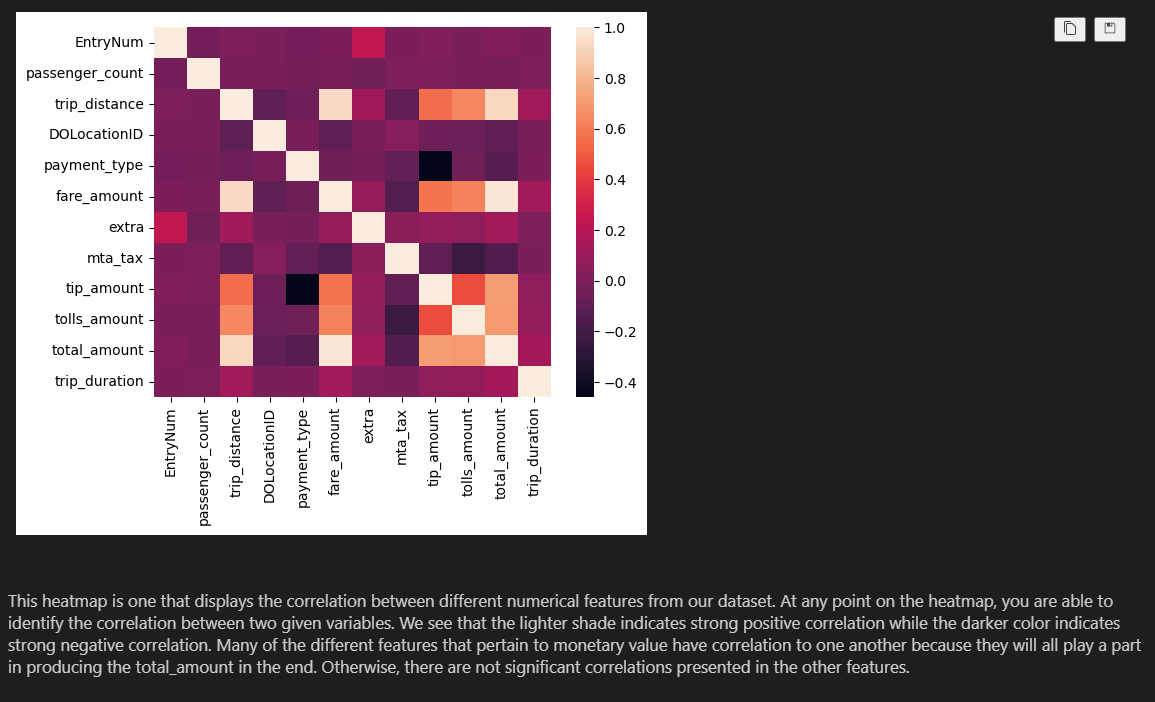
The most important part of data cleaning was to cut down the size of the file. No one in the team could run the KNN classifications or perform EDA without taking a considerable amount of time or space. We would have liked to use the entire file for the month of January. However, due to the size we had to decide on a cutoff. Initially, we tried cutting to two weeks in January and then three days. The same problem persisted. We had to settle with using one day’s worth of data. We added more columns to the dataset. Columns needed for analysis and regression included day and night times and duration of trips. We defined day time as 6 am to 6 pm. Anything out of the range was considered night time. The duration of trips was calculated by subtracting the date times of dropoff and pickup. After we cleaned the data, we had 618,034 entries to work with.

**EDA**

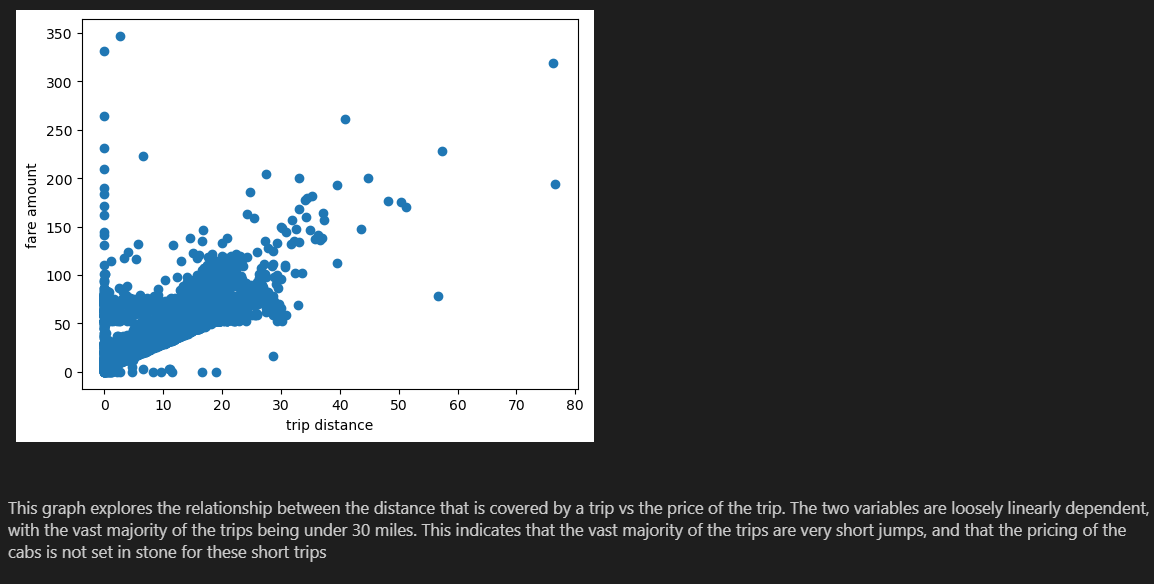
Our EDA primarily focused on the relationship between the various parts of our model. One heat map demonstrated that as the passenger count increased, the intensity of the color increased to a darker shade. Since a darker shade indicates a lower frequency, we were able to determine that the 2-car taxi rides had the highest frequency. We also used a heat map to determine the correlation between different numerical features and found that monetary features specifically have correlation to one another as they affect the output of total\_amount. We used a scatter plot to determine how trip distance affects the fare amount and a trip distance between zero and 30 were the most popular, which makes sense as customers will most likely be traveling inside the city and not a long distance. From our bar chart visualization, most people travel during the day while the number of people traveling at night are fewer. It was also found in another bar chart visualization that the average fare was slightly higher at night, which could be attributed to a longer travel distance at night. In a separate bar chart, we found that although travel distance was longer at night, day time travel took a larger amount of time, which was attributed to heavier traffic that is more prevalent during the day.

****

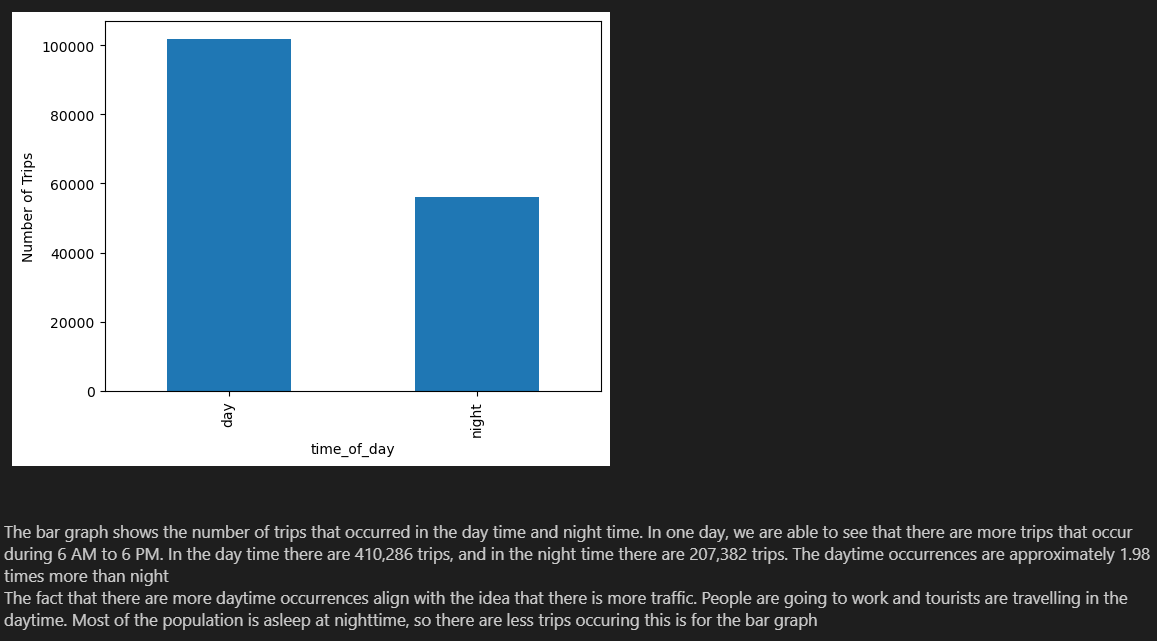
This heatmap shows us the comparison between how many cars filled with a certain number of passengers took a taxi during the day and night. The lighter shades indicate a higher frequency, and it is evident here that 2-passenger taxis were of greatest amount compared to the others. Particularly during the day we see that it had the most of that specific combination between number of passengers and time of day. This makes sense because in the busy city life of New York we assume that individual people are bustling around in taxis trying to tend to their own individual tasks. This is especially true during the day when everyone is out and about.

****

This heatmap is one that displays the correlation between different numerical features from our dataset. At any point on the heatmap, you are able to identify the correlation between two given variables. We see that the lighter shade indicates strong positive correlation while the darker color indicates strong negative correlation. Many of the different features that pertain to monetary value have correlation to one another because they will all play a part in producing the total\_amount in the end. Otherwise, there are not significant correlations presented in the other features.

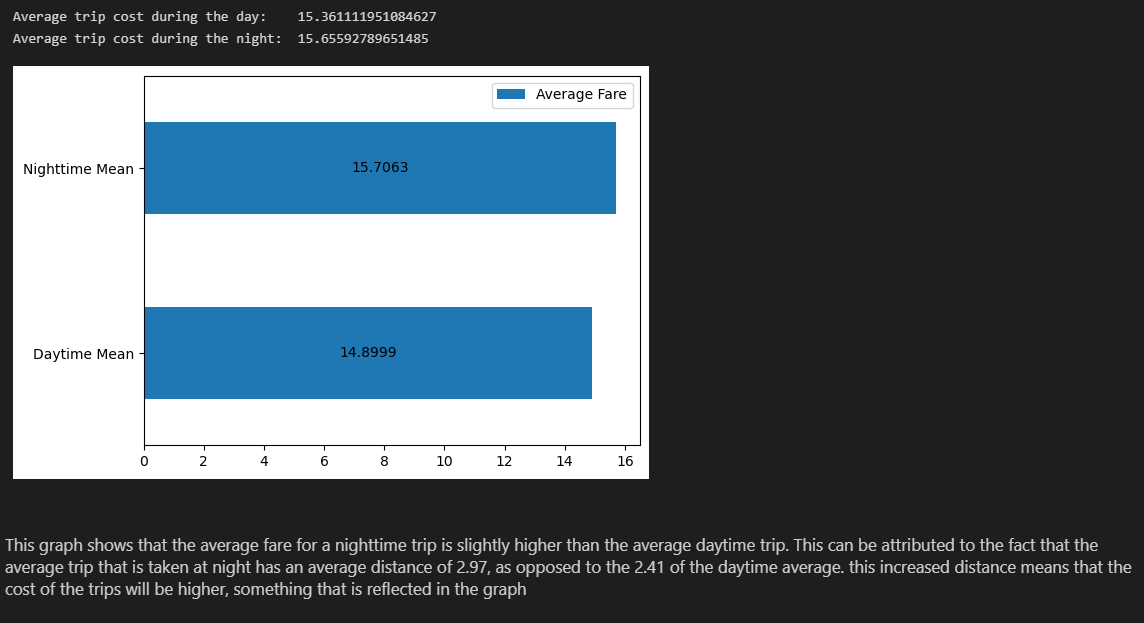
****

This scatter plot shows the overall distribution of the different taxi trips when comparing fare amount to the trip's distance. The trend of the points seems to show a general trend with positive correlation. On the left-hand side we see several stray points climbing the side of the graph that serve to be outliers and we suspect these are outliers that came about as a result of human error due to the fact that these values are inputted by the taxi drivers and sometimes are not cleaned out. The reason why we didn’t omit these values was due to the possibilities of a taxi driver being insanely greedy, receiving an abnormal tip as a gift, or other unforeseen circumstances. The data provided does not give us context of circumstances, so we kept these data points to be safe. The graph shows a general positive correlation between fare amount and trip distance. Typically you pay more, the more you travel.

****

The bar graph shows the number of trips that occurred in the daytime and nighttime. In one day, we are able to see that there are more trips that occur during 6 AM to 6 PM. In the day time there are 410,286 trips, and in the night time there are 207,382 trips. The daytime occurrences are approximately 1.98 times more than night.

The fact that there are more daytime occurrences align with the idea that there is more traffic. People are going to work and tourists are traveling in the daytime. Most of the population is asleep at nighttime, so there are less trips occuring this is for the bar graph.

****

This graph shows that the average fare for a nighttime trip is slightly higher than the average daytime trip. This can be attributed to the fact that the average trip that is taken at night has an average distance of 2.97, as opposed to the 2.41 of the daytime average. This increased distance means that the cost of the trips will be higher, something that is reflected in the graph.

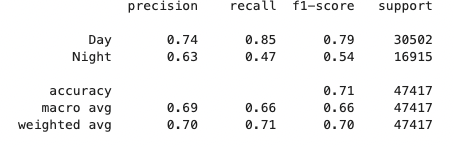
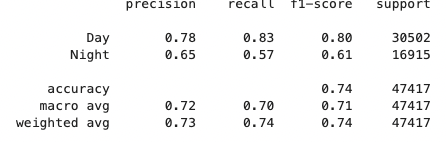
****

When looking at this graph, there is a slight, but noticeable difference in the time that is spent during a trip. Previous analysis notes that the average distance of the ride is actually shorter during the day, but the graph shows that it actually takes longer for the average trip to complete during the day. This can be attributed to heavier traffic, which is most prevalent during the day. This means that despite the distance of the trip being shorter on average than at night, it actually takes longer to get to the destination.

**KNN**

**k = 750 (top)**

**k = 22 (bottom)**

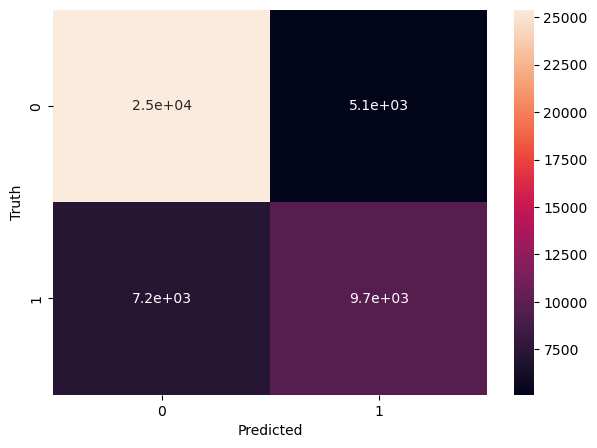


When doing the KNN, the ideal value that we strive to achieve is a value of 1.0, which indicates that the model follows the trend of data and has a high degree of accuracy. We took the data and partitioned it by 30% and 70% in order to make the testing set and training set of the data so that we could go ahead and create the model that would predict it.

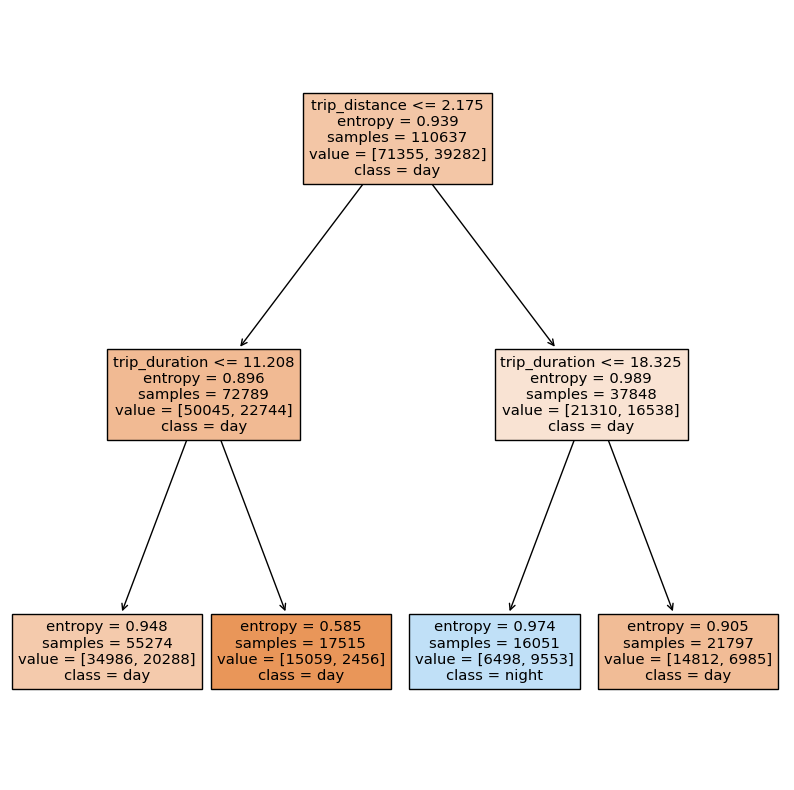
The k-value is typically the most optimal when k is the square root of n, the number of items. We found that the square root of 600,000 is approximately 750. When we trained our model using this k-value, the accuracy was around 71%. We used a trial-and-error method to find an optimal-k. Our main candidates for the k-values were low integers and halves of the original square roots, such as 375,187, 93, 46, 23, 10, 5. By training our model to these values, we eventually came to the conclusion that 22 is the best value for k. We were able to get 74% accuracy with this model.

As seen in the classification reports above, our precision, recall, and f1-score came out to be weaker than we would like. Precision is a good measure for when false positives are high. Recall is a good measure when we are determining how many actual positives we can secure. The F1-score is the balance between precision and recall and is used in a conjoined manner to calculate how well your model works. Unfortunately, as seen in the k = 22 example, our f1-scores are 0.79 and 0.54 for Day and Night respectively. The model is a bit weak in predictive power but we believe it is still moderately statistically significant.

Finally , in order to be able to visualize our data, we made a confusion matrix. The top left and bottom right squares indicate the correct predictions while the top right and bottom left squares indicate when we were incorrect. The value 1 indicates day and 0 indicates night. The matrix indicated that the model was able to predict daytime rides to a moderate extent, but was much less accurate when it came to nighttime predictions. This was further accentuated by the use of coloration in the matrix.



**Decision Tree Classification**

****

We see with this Decision Tree Classifier that the accuracy does not prove to be higher than that of the value we found from the KNN Classification. Our accuracy here is around ~67% when running the accuracy\_score(y\_test, y\_pred). This is still slightly lower than the KNN value of ~74%. In order to read this tree, we see at the top of the block is the feature trip\_distance. Depending on whether it is less than or equal to the given value in the box, it will go down to the left child if it is true, or to the right if it is false. The entropy is an indication of the purity of each node and any sub-split. Value gives us the distribution of how many entries took place in the day and how many were at night.

**Conclusion / Results / Reflection:**

In the end, from the data we collected, cleaned, and performed EDA on, there are many points of analysis that we can draw from. Although the accuracy of our KNN Classification and Decision Tree Classification were not significantly strong (74% and 67% respectively), there was a lot that we were able to learn from the process of calculating these measures.

Had we obtained a greater variety of features for this dataset, we would potentially see an improvement in the calculations we made. The original dataset had over 8 million trip records so being able to utilize more of that data may have been able to help. We chose January 31, 2018 and were able to reduce it to six hundred thousand records. Throughout the year there are potential holidays and abnormalities such as strikes. All of these potential variables would provide data and context to potentially increase the accuracy of the model. Unfortunately, it was too much for our poor laptops to handle when trying to compute all that data.

**Contributions:**

Jonathan: Calculated the KNN Classification Model and Decision Tree Classification Model, Data Cleaning, Report

Sathya Rajesh: converted the parquet to a csv, removed extreme outliers from the data, made some of the visualizations, report

Samuel Ha: Data cleaning, visualization, EDA, assisted in KNN k-value determination, report

Ian Oh: Data cleaning, visualization, slide presentation, report