New York City Taxi and Limousine Commission project
Preliminary Data Summary

Project overview

In this part of the project, a preliminary inspection of the data supplied by the NYC Taxi and Limousine Commission was performed in order to understand key data variables, and ensure the information provided is suitable for generating clear and meaningful insights.

Key Insights

1. The dataset consists of 22699 entries. Eighteen columns provide information about pickup and dropoff locations and times, ride distances and durations, payment types, fare amounts, and tip amounts.
2. Data types insights:

   - Datetime information is represented as string, which is inconvenient for working with dates.

   - store_and_fwd_flag is also string, but it would be more comfortable to work with this variable as boolean.

   - There are not null values.
3. For our project we can use variables trip_distance and total_amount.
4. Maximum and minimum values of total amount differ greatly from mean and median values.

Details

	ID	trip_distance	total_amount
8476	11157412	2.60	1200.29
20312	107558404	0.00	450.30
13861	40523668	33.92	258.21
12511	107108848	0.00	233.74
15474	55538852	0.00	211.80
6064	49894023	32.72	179.06
16379	101198443	25.50	157.06
3582	111653084	7.30	152.30
11269	51920669	0.00	151.82
9280	51810714	33.96	150.30

Data sorted by total_amount

Next Steps

- Change type of datetime information and store_and_fwd_flag.
- Investigate dependence between trip_distance and total_amount.
- Provide EDA

New York City Taxi and Limousine Commission project
Exploratory Data Analysis

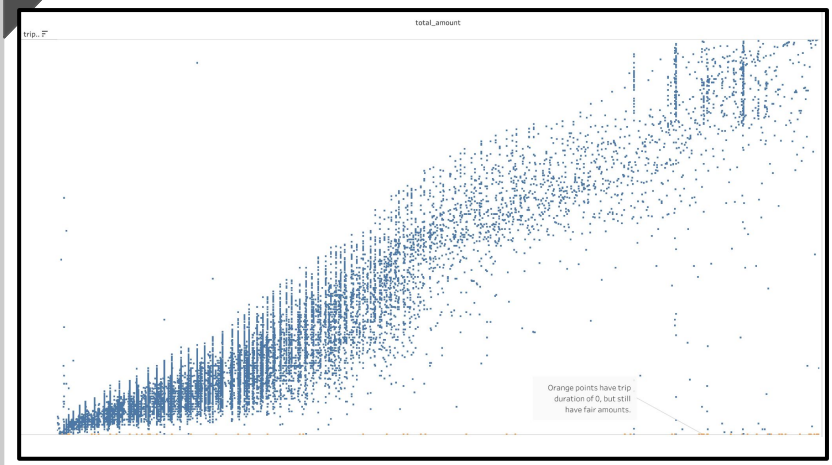
Project Overview

An exploratory analysis of the data supplied by the NYC Taxi and Limousine Commission was performed.

Key Insights

1. Although we can observe a strict linear dependency between trip distance and total amount, there is a certain number of rides with no distance but fare amounts.
2. For trip distance, the median is 1.61 miles. The distribution of trip distance is skewed right (there can be potential outliers with long ride distance).
3. The median total amount is $11.8, and notably, there are instances of negative values.
4. Interestingly, the mean tip amount is highest with no passengers on a ride.
5. The total number of monthly rides shows us the lowest numbers in July, August, and September. The same picture is for monthly revenue.
6. The highest numbers of rides are in March and October. The highest revenue is in March, May, and October.

Details



Total distance and total amount
New York City Taxi & Limousine Commission 2017

Next Steps

- Determine any unusual data points that could pose a problem for future analysis in predicting trip fares.
- Determine the variables that have the largest impact on trip fares.

New York City Taxi and Limousine Commission project
Hypothesis testing

Project Overview

An A/B test was conducted in order to analyze whether there is a relationship between payment type and fare amount.

Key Insights

1. There is a statistically significant difference in the average fare amount between customers who use credit cards and customers who use cash. The average fare amount is larger for customers, who use credit cards as a payment method.

2. This A/B test project might not be realistic because it is difficult to force customers of two groups to pay with cash or card. There is also a possibility that larger fare amounts are more suitable to pay with a credit card. In other words, it's far more likely that fare amount determines payment type, rather than vice versa.

Details

$H0$: There is no difference in the average fare amount between customers who use credit cards and customers who use cash.

HA: There is a difference in the average fare amount between customers who use credit cards and customers who use cash.

```
card_subset = taxi_data[taxi_data['payment_type'] == 1]
cash_subset = taxi_data[taxi_data['payment_type'] == 2]

stats.ttest_ind(a=card_subset['fare_amount'], b=cash_subset['fare_amount'], equal_var = False)
```

```
Ttest_indResult(statistic=6.866800855655372, pvalue=6.797387473030518e-12)
```

The P-value is much less than 5%, so the null hypothesis should be rejected.

Next Steps

- New York City TLC should consider encouraging customers to pay with credit cards.

New York City Taxi and Limousine Commission project
Regression Analysis

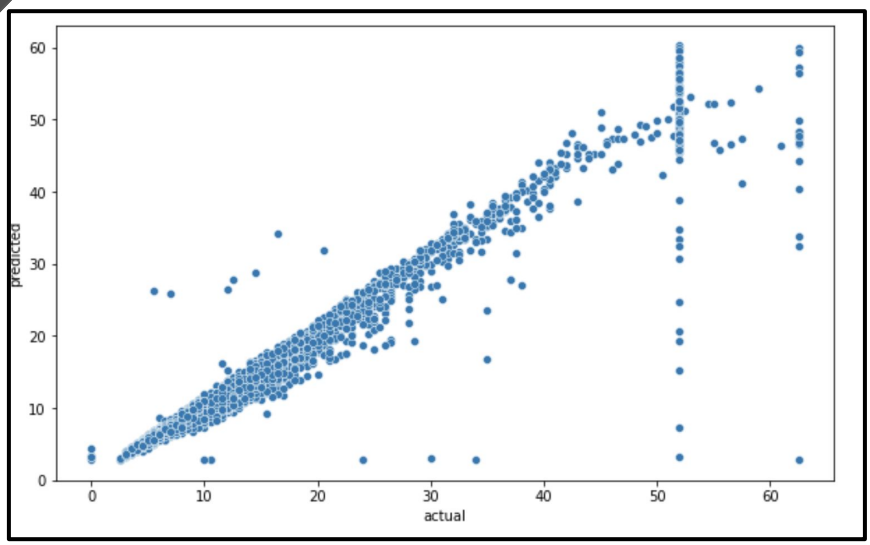
Project Overview

A multiple linear regression model was built to help estimate taxi fares before the ride for the New York City Taxi and Limousine Commission.

Key Insights

- As independent variables were chosen trip distance and trip duration (correlation coefficients with fare amount 0.91 and 0.86 respectively). Since the model will not know the duration of a trip until after the trip occurs, it was trained on a statistics that capture the mean distance and mean duration for each group of trips that share pickup and dropoff points.
- The model explains 84% of the variation in fare amount. This makes the model an effective predictor of fare amount.
- According to the model, when trip distance and duration equal 0, the fare amount will be 2.8106 dollars.
- An increase of one mile for the trip distance will result in an estimated 2.3389 dollars more in fare amount. There is a 95% chance the interval [2.297, 2.381] contains the actual parameter value of the slope.
- An increase of one minute for the trip duration will result in an estimated 0.2414 dollars more in fare amount. There is a 95% chance the interval [2.228,2.255] contains the actual parameter value of the slope.

Details



Scatterplot visualizes the relationship between actual and predicted values using a testing dataset.

Next Steps

- Consider the possibility and practicability of adding another independent variable to increase the percentage of fare amount variability explained by the model. For example, whether there is rush hour or not.
- Since distance and duration are crucial for the fare amount, it would be helpful to consider some loyalty programs for customers whose rides are long.