Automatidata Exploratory Data Analysis (EDA) Project
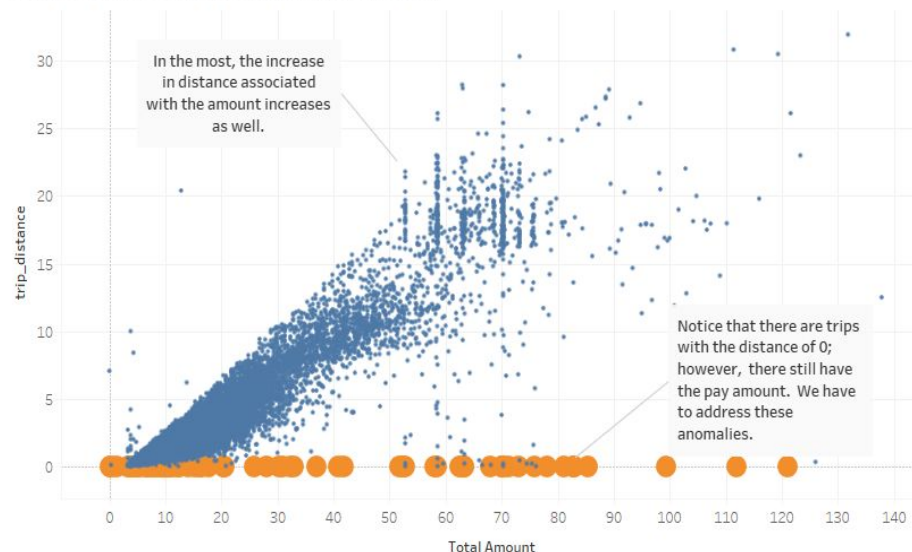
Project Overview

The project's purpose is to perform the EDA process after familiarized with the dataset from the previous section. By completing the EDA, we'll have deeper understanding on variables, the key insights and important findings, and recommend decisions to take for further steps in the project's course.

Details

Key Insights

- The variables (trip distance, total amount, tip amount) appear as right-skew distributions.
- VeriFone Inc. vendor accounts for higher shared tips amount compared to Creative Mobile Technologies, LLC.
- Most passengers count (exclude 0 passenger) have the average tip amount below <$2.
- The total rides and revenue are fairly high from March-June, while months in Q3 is pretty low.
- The same concepts above apply for weekday, which midweek accounts for the highest.
- Also the concepts apply for the hours (further analysis), which from 6-7pm account for highest rides and revenue. Conversely, 4-5am are the lowest.
- For drop off locations, further analysis shows that there are disproportionate number of locations receive the majority of the traffic, while the rest get relatively few trips. Trips with the highest traffic could be near to the popular/attractive tourisms, trains, and buses.

Total Distance and Total Amount TLC 2017

In the most, the increase in distance associated with the amount increases as well.

Notice that there are trips with the distance of 0; however, there still have the pay amount. We have to address these anomalies.

This scatterplot shows the relationship between trip distance and total amount. It's clear that the positive relationship appear between these two variables, and it's a good signal for building a predictive model. However, notice that there are trips with the distance of 0, and still have the total amount included.

Next Steps

- The Automatidata team will further analyze and address the anomaly between trip distance and total amount, which to determine whether there's a problem in collecting data process, or something else.
- With the client's perspective, we might want to account for any other variables that might be effective to include for building ride amount's predictive model.
- Having the next steps of statistical testing and model building when we'll completely solve the issues above.