New York City TLC Project EDA Summary II

Executive Summary Report

Project Overview

In this part of the project, the TLC explored the data even further by analyzing, cleaning, and visualizing, to ensure that the data is ready for any modelling.

Key Insights

After running the second part of the EDA on a sample of the data provided by the TLC, it has came to the spotlight that:

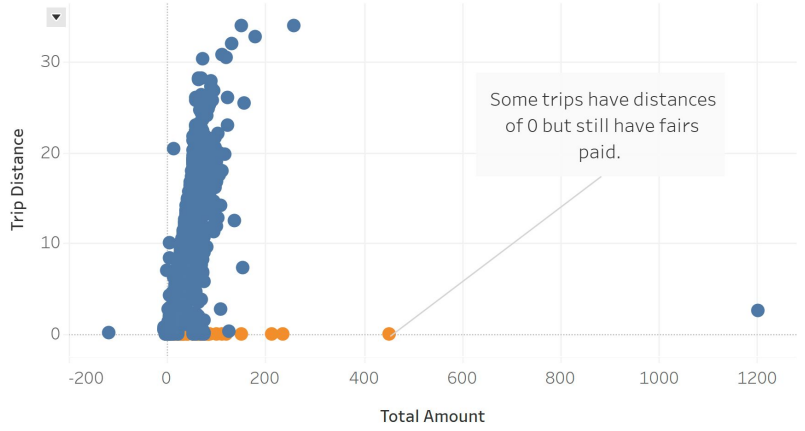
- there are several outliers that would have to be dealt with.
- there is something especially problematic: some trips that are entered with a distance value that is 0, yet they have a fare cost applied.

The latter part is especially problematic and not okay. Removal of such outlier is strongly advised.

Outliers would need to be handled properly, and each outliers would be determined and classified as different types, so each would be handled accordingly.

Details

Scatterplot (Total Amount vs Trip Distance)



This screenshot displays the mapping between the total amount and the trip distance.
Some are cut off due to limitations.

Next Steps

Next steps would include:
1. Determine unusual points that could still pose problems, such as locations that have long durations.
2. Determine variables that have the highest impact on fares.
3. Filter down to most relevant variables so that statistical analysis and modelling would be conducted smoothly.