Case Study: New York Taxi Trip Dataset

Muhammad Umair
mumair@uni-bremen.de
Universität Bremen
Bremen, Germany

Abdullah Al Noman
nom@biba.uni-bremen.de
Universität Bremen
Bremen, Germany

Abstract

Exploratory Data Analysis is the process of examining data and extracting insights from it in order to study its main features. EDA can be accomplished through the use of statistical and visualization techniques. If we don't study the data, we will never be able to make sense of such enormous databases.

Finding out how features affect the target variable, examining the nature of the features, discovering anomalies and outliers and handling them so they don't affect our model, and being able to perform data cleaning are all made feasible by exploring and analyzing the data. Without conducting exploratory data analysis, we won't be able to identify inconsistencies or gaps in the data that could lead our model to predict trends erroneously.

Business stakeholders frequently hold certain presumptions regarding data, from a business perspective. Exploratory data analysis enables us to dig deeper and determine whether our observations match the data. It aids in determining whether we are posing the proper queries. The foundation for responding to our business inquiries is also provided by EDA.

Keywords: Exploratory Data Analysis

Reference Format:
Muhammad Umair and Abdullah Al Noman. 2023. Case Study: New York Taxi Trip Dataset. In *Edge Computing Lab of Universität Bremen (Edge Computing SoSe'22).* Universität Bremen, Bremen, Germany, 2 pages.

1 Introduction

3 important steps to keep in mind:

1- Understand the data 2- Clean the data 3- Find a relationship between the data

2 Understanding the Data

refer to figure 1.

3 Cleaning and Filtering the Data

refer to figure 2.

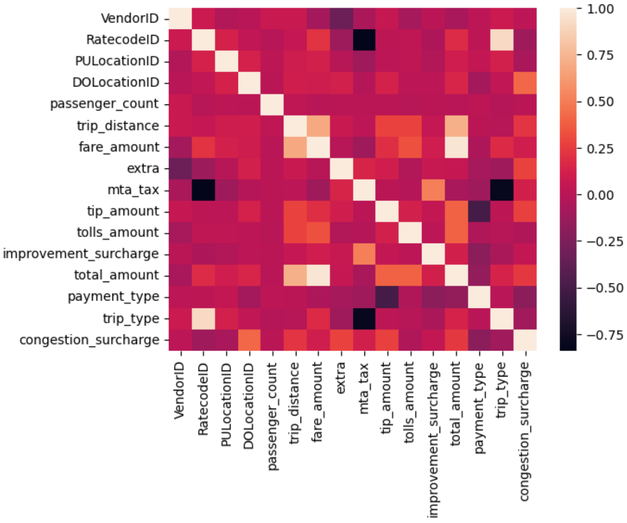

Figure 1. Heatmap of the correlation matrix

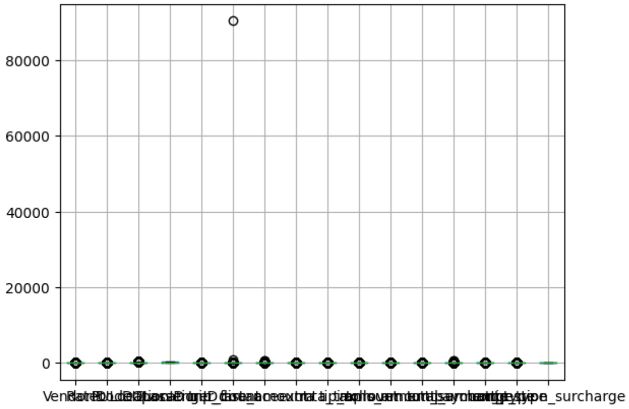

Figure 2. cheacking for outliers

4 Feature Creation

4.1 trips per hour/day/month

4.2 time duration minutes

5 Univariate Analysis

refer to figure 3

6 Bivariate Analysis

refer to figure 4

Edge Computing SoSe'22, 2022, Bremen, Germany

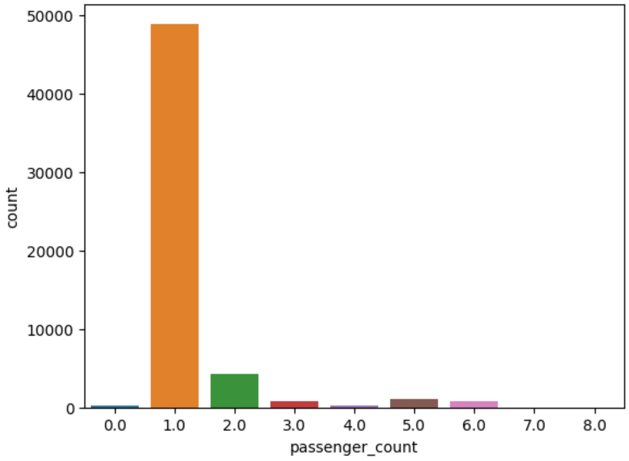

Figure 3. passenger count

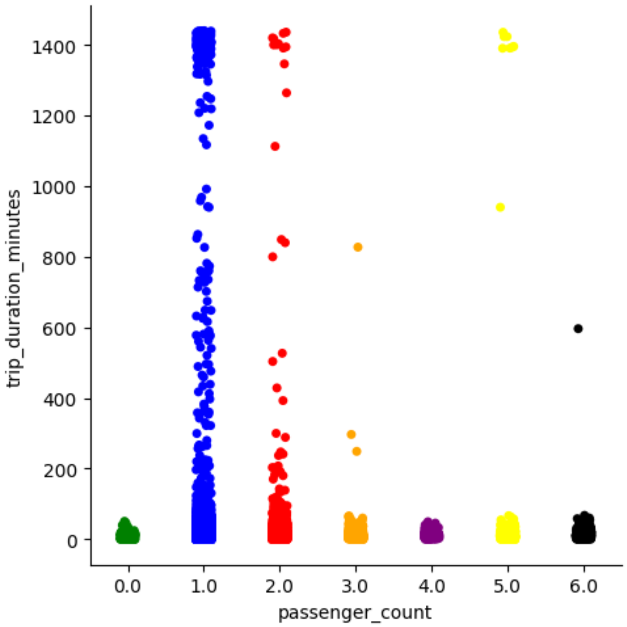

Figure 4. trip duration per Passenger counts

Bivariate Analysis involves finding relationships, patterns, and correlations between two variables

References

7 Appendix

A Source Code

The source code is published internally at: https://github.com/mumair97/EDA.git.