Yellow Taxi Trip Data (EDA)

Abstract

The purpose of this project is to look at the yellow taxi trip records, in order to investigate passenger's behavior. The project focus on the following things. First, at what time of the day passenger will give more tips. Second, see if the distance of the trip has an affection on the payment type. Last, know when the demand for taxi increase. This analysis will help the taxi drivers as well as the company to know when the best time to serve and when to reduce or avoid serving.

Design

The data gained from NTC taxi and limousine commission https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. They have recorded each passenger trip on 2021 January the first in New York city for the purpose of improving their services.

Data

The dataset contains 1027003 rows x 30 columns include pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and passenger counts reported by the driver. The data in the linked files was gathered and submitted to the NYC Taxi and Limousine Commission (TLC) by technology vendors authorized under the Taxicab and Livery Passenger Enhancement Programs (TPEP/LPEP).

Feature Enhancement

Make some new features out of the existing variables to acquire more insights from the data. Because both pickup datetime and dropoff datetime were of the object type. Then I've built a function that tells us what time of day the ride was taken. I've set up four time zones: 'Morning' (from 6:00 a.m. to 11:59 p.m.), 'Afternoon' (from 12 noon to 3:59 p.m.), 'Evening' (from 4:00 p.m. to 9:59 p.m.), and 'Late Night' (from 4:00 p.m. to 11:59 p.m.) (from 10:00 pm to 5:59 am).

Tools

Pandas, NumPy, matplotlib, seaborn, sqlite3, and datetime