Group 7: Ruirui Mao,,Shangchen Li, Junyuan Xiang

NYC Yellow Taxi trip data Analysis from 2012-2021

Description of the Data:

This dataset was found on **NYC Taxi & Limousine Commission** and This data dictionary describes yellow taxi trip data from 2012-2021.

The trip records for yellow taxis typically contain various fields that capture important information about each trip. These fields can include the date and time when the passenger(s) entered and exited the taxi, the pickup and drop-off locations, the distance traveled during the trip, a detailed breakdown of the fares charged, the rate structure used for the trip, the payment method used to pay for the trip, and the number of passengers reported by the

https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Possible Additional Variables:

VendorID: A code indicating the taxi company that provided the trip.
tpep_pickup_datetime: The date and time when the passenger(s) entered the taxi.
tpep_dropoff_datetime: The date and time when the passenger(s) exited the taxi.
passenger_count: The number of passengers in the taxi during the trip.
trip_distance: The distance of the trip in miles.
payment_type: The payment method for the trip (e.g., credit card, cash).
fare_amount: The fare amount for the trip.
tip_amount: The tip amount for the trip.
total_amount: The total amount paid for the trip, including fare and tip.
improvement_surcharge: A surcharge fee that is added to the fare for trips that pass through certain areas during specified times.
mta_tax: A tax imposed by the Metropolitan Transportation Authority.
extra: An additional fee for trips that occur during specified times or under specified circumstances.

Reading Data Code:

```
import pyarrow.parquet as pq
table = pq.read_table('~/Desktop/405/yellow_tripdata_2017-01.parquet')
df = table.to_pandas()
print(df.head())
```

***In view of the large data sample size, we temporarily extracted the data in January 2017 for demonstration to ensure the whole dataset is readable.**

Procedure For Analysis:

1. Data Cleaning
2. EDA (Exploratory Data Analysis)
3. Data Visualization
4. Modeling {a . Logistic Regression (Classification) b. KNN (Classification) c.SVM (Classification) d. Decision Tree (Classification)}

Main Question:

1. How to maximize the total value of fare amount (number of passengers, date, operating time of each passenger, payment method, distance, etc.)

2. How will tip income be affected (operating time, payment method, distance for each,passenger ride)

---What is the total amount of tips received by each payment type (credit card, cash, etc.)?

---What is the average trip distance for each passenger count and payment type?

Sub questions:

- What is the average fare amount for each passenger count?
- What is the average fare amount for each day of the week?
- What is the average speed of each trip based on the pickup and dropoff time and distance?
- How many trips were taken on each day of the ten years?

Github repository: https://github.com/Rainieeie/Stats-405-group-7