

**School of Computer Science Engineering and Information Systems (SCORE)**

**Winter Semester 2024-25**

**Course Code: SWE – 2009**

**Course Name: Data Mining Techniques**

**Faculty: Dr. DURAI RAJ VINCEMT P. M.**

**Digital Assignment**

*Performing EDA* on

*New York City’s Taxi Fare* Dataset

|  |  |
| --- | --- |
| **Student Name** | P. Mokshagna |
| **RegNo.** | 22MIS0377 |
| **Slot** | E2 + TE2 |

**Link to the Repository:** <https://github.com/PAGADALA-MOKSHAGNA/EDA_Taxi_Trip_Data>

# Introduction

New York City’s taxi system is **one of the largest and most complex transportation networks in the world.** With millions of trips recorded each year, analysing taxi fare data provides valuable insights into urban mobility, fare structures, and passenger demand patterns. **Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset, identifying trends, and detecting anomalies** that could impact fare predictions, policymaking, or ride-hailing business strategies.

# Statistical Insights

 NYC’s yellow taxis complete approximately **200,000–300,000 trips per day**.

 The average trip fare is around **$12–$15**, with a median trip distance of **1–3 miles**.

 Taxi fares are regulated and structured, consisting of **base fare, per-mile charges, surcharges, and tolls**.

 The introduction of ride-hailing services like Uber and Lyft has led to **shifts in fare distribution**, making EDA essential for understanding new trends.

# Description of the Dataset

Dataset is ideal for transport data analysis, predictive modeling, and fare optimization. Data scientists can use it to analyze traffic patterns, predict trip times, study passenger behavior, and evaluate taxi service efficiency. It’s a rich source for exploring New York City’s transportation dynamics and urban mobility trends.

* **VendorID**: A unique identifier for the taxi vendor or service provider.
* **tpep\_pickup\_datetime**: The date and time when the passenger was picked up.
* **tpep\_dropoff\_datetime**: The date and time when the passenger was dropped off.
* **passenger\_count**: The number of passengers in the taxi.
* **trip\_distance**: The total distance of the trip in miles or kilometers.
* **RatecodeID**: The rate code assigned to the trip, representing fare types.
* **store\_and\_fwd\_flag**: Indicates whether the trip data was stored locally and then forwarded later (Y/N).
* **PULocationID**: The unique identifier for the pickup location (zone or area).
* **DOLocationID**: The unique identifier for the drop-off location (zone or area).
* **payment\_type**: The method of payment used by the passenger (e.g., cash, card).
* **fare\_amount**: The base fare for the trip.
* **extra**: Additional charges applied during the trip (e.g., night surcharge).
* **mta\_tax**: The tax imposed by the Metropolitan Transportation Authority.
* **tip\_amount**: The tip given to the driver, if applicable.
* **tolls\_amount**: The total amount of tolls charged during the trip.
* **improvement\_surcharge**: A surcharge imposed for the improvement of services.
* **total\_amount**: The total fare amount, including all charges and surcharges.
* **congestion\_surcharge**: An additional charge for trips taken during high traffic congestion times.