PRINCIPLES OF BIG DATA MANAGEMENT

COMP-SCI 5540 [Spring Semester 2024]

Research Project (40 Marks)

Submission Date: April 18th, 2024 (Thursday)

Instructions:

- Research project is a group-based activity.
- Submit your work in a single document with the answers to the questions and the code. (Preferably it can be the pdf document)
- Late and copied work won't be graded and will result as ZERO credit.

Project Title: Taxi Service Analyzing System using NLP

Group Members:

Name	Student-ID
Olena Khrystenko	16277259
Laxmi Prasanna Vollala	16358583
Rodda Mamo	16339770
Mounika Sanaboyina	16354670
Rishitha Girra	16359336
Naveen Kuchi	16353887
Chris Lee	16292115

The New York Yellow Taxi Trip Dataset Analysis

Contents

Abstract

In the bustling thoroughfares of New York City, the iconic yellow taxi cabs serve as vital conduits for millions of urban dwellers navigating the bustling streetscape. Within this vibrant metropolis lies a wealth of data captured by the comprehensive "NYC Yellow Taxi Trip Data" dataset that includes data from 2003 to 2021, offering a rich tapestry of insights into the intricate dynamics of urban mobility. Our study embarks on a rigorous exploration of this dataset, aiming to elucidate the underlying patterns and trends that define the essence of city life on wheels. Through a methodical analysis, we unveil the multifaceted nuances of taxi demand, travel behavior, and the economic forces shaping the taxi industry landscape. Leveraging analytical techniques, we meticulously examine factors such as trip durations, passenger counts, fare structures, and spatial-temporal dynamics to glean actionable insights into the urban commuting experience. Our investigation traverses the urban landscape, from peak hours to popular pickup and drop-off locations, revealing the fabric of mobility within the city. By discerning the ebbs and flows of taxi travel, we provide valuable intelligence to policymakers, transportation planners, and taxi service providers, empowering them to optimize service efficiency, refine fare structures, and alleviate urban congestion. As stewards of data-driven decision-making, we recognize the transformative potential of our findings in shaping the future trajectory of urban mobility, aiming to create a more inclusive, equitable, and vibrant urban environment that reflects the spirit and dynamism of the city that never sleeps.

1. Introduction

Urban mobility is a complex and ever-evolving landscape shaped by various factors such as demographics, infrastructure, and economic influences. In the pursuit of enhancing transportation efficiency and addressing urban congestion, a thorough examination of taxi trip data can offer valuable insights. This study delves into the "NYC Yellow Taxi Trip Data" dataset, including data from 2003 to 2021, employing advanced analytical techniques to uncover intricate patterns and trends in urban mobility. By analyzing factors like trip durations, passenger counts, and fare structures, we aim to reveal essential insights into taxi demand, travel behavior, and economic influences on the industry. These findings hold significant implications for policymakers, transportation planners, and taxi service providers, offering actionable insights to improve service efficiency and mitigate urban congestion. Understanding commuter preferences and travel patterns can guide strategic decisions aimed at fostering sustainable transportation solutions in densely populated urban areas.

a. Project Scope

This project entails a comprehensive analysis of the "NYC Yellow Taxi Trip Data" dataset to uncover insights into urban mobility patterns. It encompasses data collection, preprocessing, and advanced analytical techniques to examine trip durations, passenger counts, and fare structures. The scope extends to identifying factors influencing taxi demand, travel behavior, and economic influences on the industry. Findings will inform recommendations for policymakers, transportation planners, and taxi service providers to enhance efficiency and alleviate urban congestion.

- **Data extraction:** The first step is to collect data ("NYC Yellow Taxi Trip Data" CSV) from Kaggle and load them.
- **Cleaning the data:** The extracted data will be analyzed and cleans the missing values, duplication, and any other irregularities.
- **Additional errors:** The data cleaning procedure will entail deleting any invalid data, formatting as well as removing outliers.
- **Data transformation:** converting data into a format that can be easily read and analyzed. This will entail aggregating the data depending on several dimensions such as location, time periods, and so on.
- **Data Visualization:** The converted data will be shown via tools such as Matplotlib, seaborn, Word Cloud, and Folium. The visualizations will aid in the identification of patterns, trends, and insights in the data.

The project aims to foster sustainable transportation solutions by understanding commuter preferences and guiding strategic decisions for densely populated urban areas, ensuring relevance and applicability to stakeholders.

b. Project Limitations & Constraints

The project faces several limitations and constraints that may impact the comprehensiveness and validity of its findings. Firstly, it relies exclusively on the "NYC Yellow Taxi Trip Data" dataset from 2003 to 2021, potentially overlooking broader aspects of urban mobility that could influence transportation efficiency and congestion. Data quality issues, including inaccuracies and missing values, may introduce biases into the analysis, affecting the reliability of insights. The temporal scope of the analysis is confined to a single year, limiting the exploration of long-term trends and seasonal variations in urban mobility patterns. Geographical constraints restrict the study to a specific area, potentially reducing the generalizability of findings to other regions with different transportation networks and demographics. Ethical considerations related to privacy and data protection regulations may limit the depth of analysis or access to certain data points. Resource constraints such as time, budget, and computational resources may also restrict the complexity of analysis techniques and the depth of exploration. Additionally, external factors such as regulatory changes or unforeseen events could influence the validity and applicability of recommendations. Limited stakeholder engagement may further constrain the relevance and adoption of proposed solutions.

c. Feasibility Study

The feasibility study for analyzing the "NYC Yellow Taxi Trip Data" data indicates a promising endeavor. Technical feasibility is assured with the dataset's availability and suitable analytical tools. Economic viability is supported by manageable costs and anticipated benefits like improved transportation efficiency. Legal and ethical compliance is ensured through adherence to data protection regulations. Operationally, adequate resources and stakeholder collaboration facilitate project execution within the proposed timeline. Overall, the study demonstrates feasibility across technical, economic, legal, operational, and scheduling aspects, promising valuable insights into urban mobility patterns and informing strategic decision-making for stakeholders.

2. System Requirement Specifications (SRS) [MDRE | Bespoke]

I. Introduction

The Taxi Trip Analytic System is a software solution designed to analyze and derive insights from taxi trip data. This document outlines the detailed requirements for the development of the Taxi Trip Analytic System.

II. Functional Requirements

- **Data Loading**

The system shall provide functionality to load taxi trip data from a CSV file into a PySpark.pandas DataFrame.

Users shall be able to specify the file path and format of the dataset to be loaded.

- **Data Preprocessing**

  The system shall support various data preprocessing tasks, including cleaning, filtering, and transforming the dataset.

  Users shall have the option to handle missing values, outliers, and inconsistencies in the data.

- **Exploratory Data Analysis (EDA)**

  The system shall generate descriptive statistics, visualizations, and insights to explore the characteristics of the taxi trip data.
  Users shall be able to analyze key metrics such as trip distances, fare amounts, passenger counts, etc.

- **Feature Engineering**

  The system shall facilitate the creation of new features or variables from existing data attributes to improve model performance.
  Users shall have access to feature engineering techniques such as encoding categorical variables, scaling numerical features, etc.

- **Machine Learning Modeling**

  The system shall leverage machine learning algorithms implemented in "sklearn" for predictive analytics tasks.

  Users shall be able to train and evaluate machine learning models for tasks such as fare prediction, trip duration estimation, etc.

III. Non-Functional Requirements

- **Performance**

  The system shall be capable of handling large volumes of taxi trip data efficiently, with minimal processing delays.
  Response times for data loading, preprocessing, and model training shall be optimized for user experience.

- **Scalability**

  The system architecture shall be designed to accommodate future growth in data volume and user demand.

Scalability features such as parallel processing and distributed computing may be implemented to handle increasing workloads.

IV. External Interfaces

- **Kaggle Notebook Integration**

  The system shall integrate with the Kaggle Notebook environment for data analysis and model development.
  Kaggle Notebooks shall be utilized for loading datasets, conducting exploratory data analysis, and training machine learning models.

V. Data Requirements

- **Dataset**

  The system shall utilize a dataset containing information about taxi trips, including pickup/drop-off locations, trip distances, fare amounts, etc.
  The dataset shall be stored in CSV format and loaded into the system for analysis and modeling using pandas.

VI. Constraints

- **Technological Constraints**

  The system shall be developed using Python programming language and relevant libraries/frameworks, including pandas, seaborn, numpy, MatplotLib, sklearn, and mpl_toolkits.- Computational resources such as memory and processing power may impose limitations on the size and complexity of datasets and models.

VII. Assumptions and Dependencies

- **Assumptions**

  It is assumed that users have basic knowledge of data analysis and machine learning concepts.- It is assumed that the dataset used for analysis and modeling is representative of real-world taxi trip data and is of sufficient quality for analysis.

- **Dependencies**
  The successful implementation of the system depends on the availability of necessary resources, including development tools, libraries, and computational infrastructure.

3. System Design

a. Use-Case Diagram

The use-case diagram below represents a taxi service system.
Actors:

- **Rider**: Represents individuals who use taxi services in New York City.
- **Driver**: Represents individuals who provide taxi services in New York City.
- **Administrator**: Represents intermediary between drivers and customers.

- **Data Analyst**: Represents individuals or systems responsible for analyzing the dataset.
- **System Administrator**: Represents individuals responsible for maintaining and managing the dataset.
- **The system** serves for data collection

Use Cases:

- **Request Taxi Service**: The Taxi Service User requests a taxi service by providing pickup and dropoff locations, along with other details such as pickup time and passenger count.
    - **View Driver Details**
- **Cancel A Ride**: This may be done by both a rider and a driver.
- **Provide Taxi Service:** The Taxi Driver provides taxi service to the user by picking them up at the specified location and dropping them off at the destination.
    - **Accept a Ride Request**
    - **View Rider Details**
- **Data Collection**: The system records trip data, including pickup time, dropoff time, passenger count, trip distance, fare amount, etc.
- **Analyze Trip Data**: The Data Analyst analyzes the dataset to gain insights into taxi trip patterns, fare trends, passenger demographics, etc.
- **Manage Dataset**: The System Administrator manages the dataset by performing tasks such as data cleaning, storage, backup, and access control.
- **Oversee Systsem's Operation And Management**: The taxi service administrator plays a crucial role in managing the taxi service system. This includes:
    - **Manage Rider Complaints**
    - **View Reports**
    - **Manage Drivers**

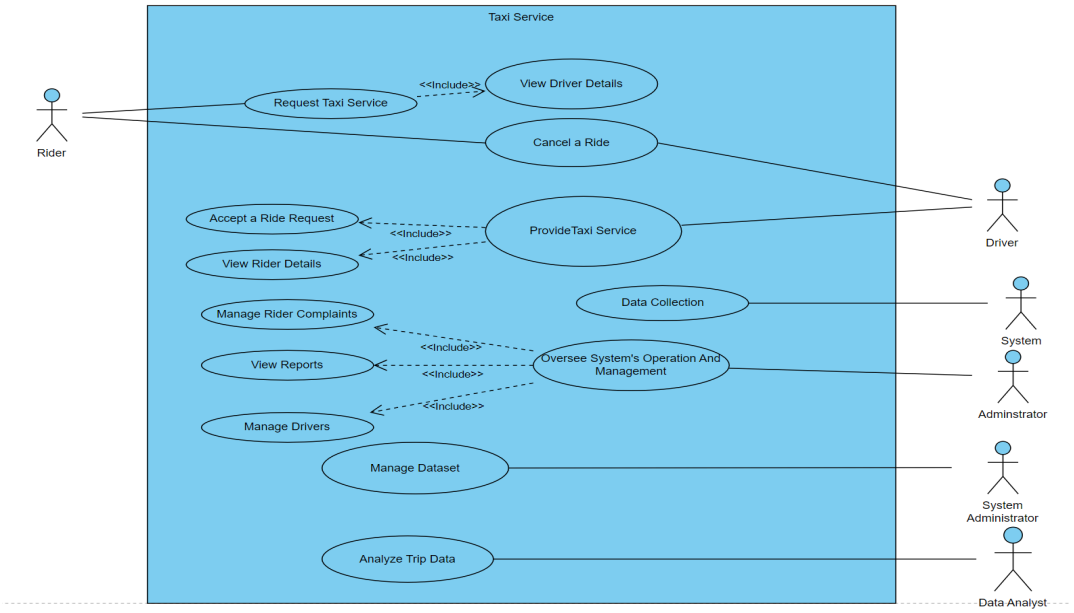


Figure 1 - Use-case diagram of a taxi service

b. Sequence Diagram

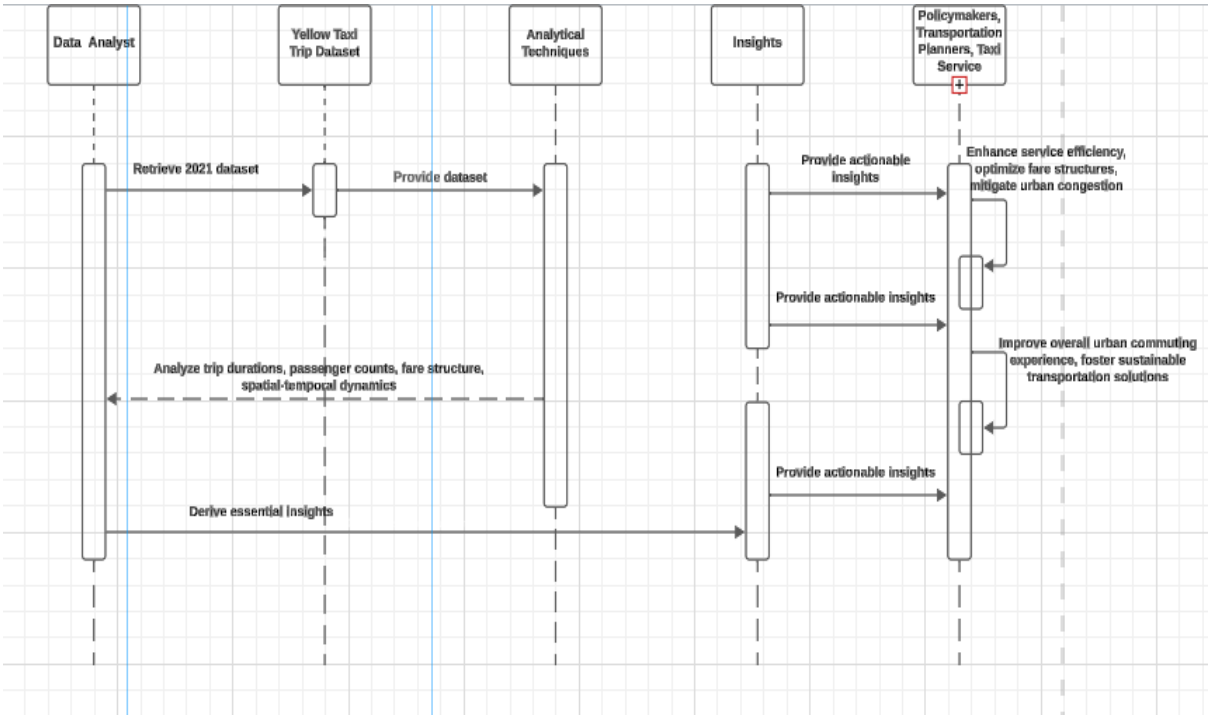


Figure 2 - Sequence diagram

Data Analysis and Model Training Process

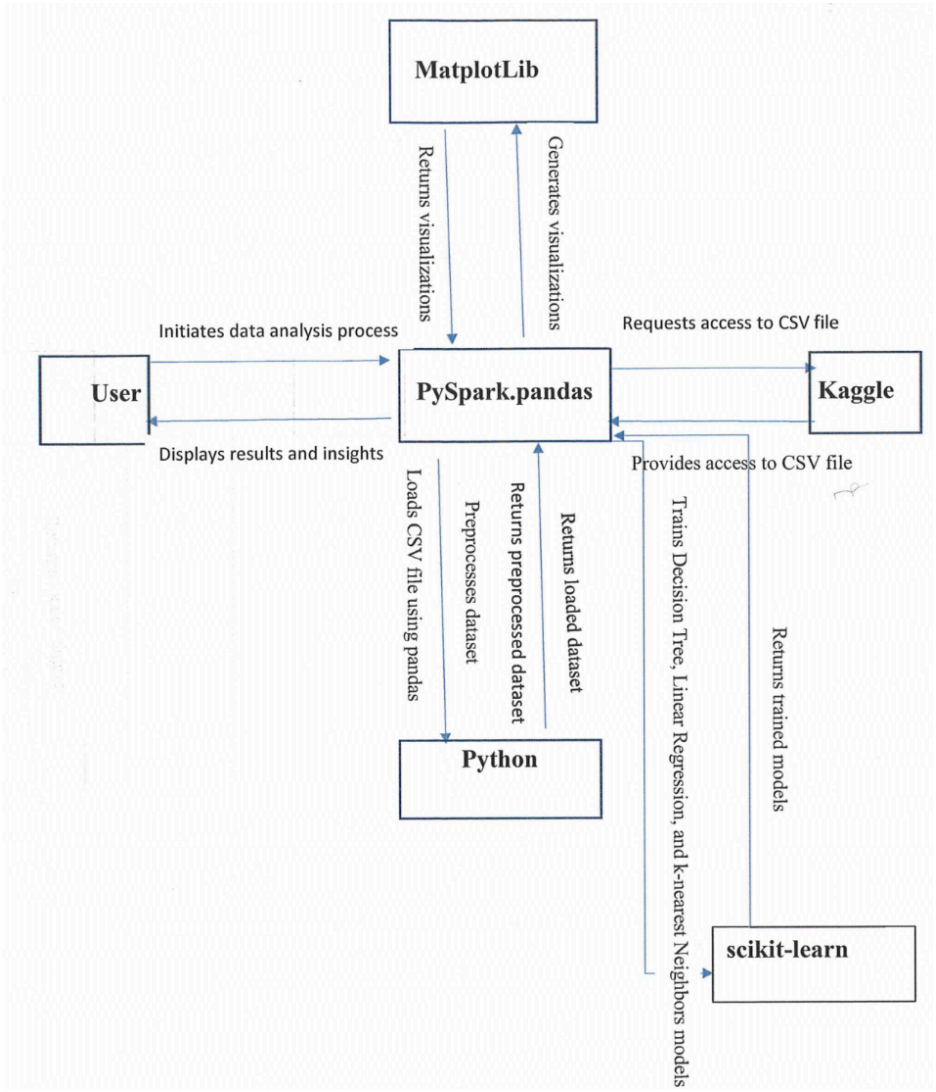


Figure 3 - Data analysis and model training process diagram

4. Data Design

a. ETL Process

Extract the "NYC Yellow Taxi Trip Data" dataset, ensuring integrity. Transform data by cleaning, standardizing, and engineering features like trip distance. Analyze data using advanced techniques, identifying insights on trip durations, passenger counts, and fares. Generate reports and visualizations for stakeholders, offering actionable recommendations. Iterate on the process, optimizing workflows for efficiency and scalability.

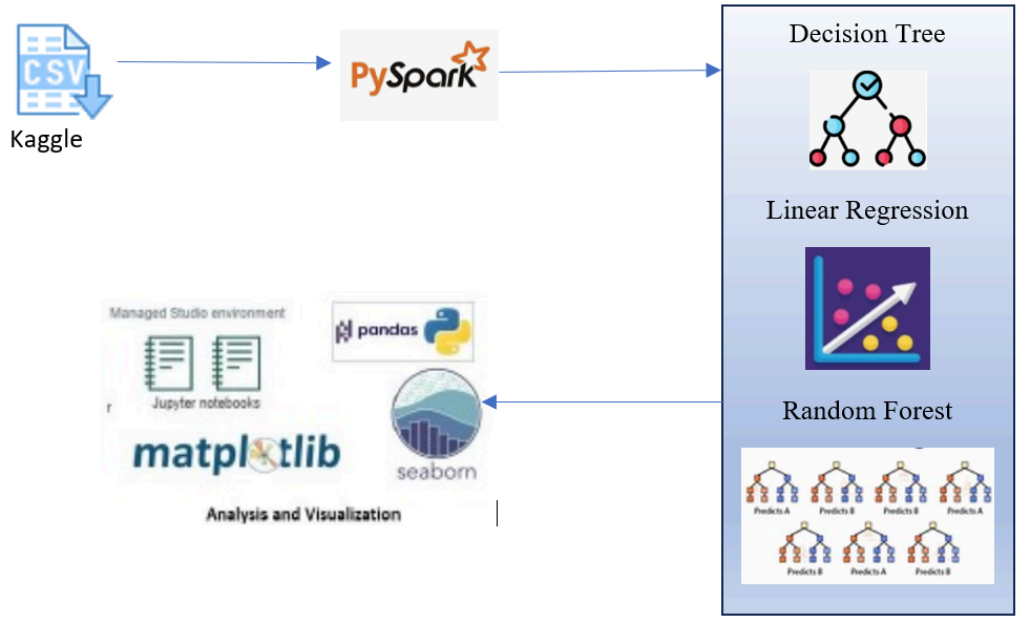


Figure 4 - ETL process diagram

b. Data Management

The project focuses on analyzing the "NYC Yellow Taxi Trip Data" dataset to unveil urban mobility patterns. Data management involves collecting, cleaning, and organizing trip data, including durations, passenger counts, and fare structures. Utilizing advanced analytics, we extract insights into taxi demand, travel behavior, and economic impacts. Peak hours, popular locations, and fare fluctuations are identified. Implications extend to policymakers, planners, and taxi providers, enabling service enhancement, fare optimization, and congestion mitigation. Understanding commuter preferences guides strategic decisions for improving urban commuting and fostering sustainable transportation solutions in dense urban areas.

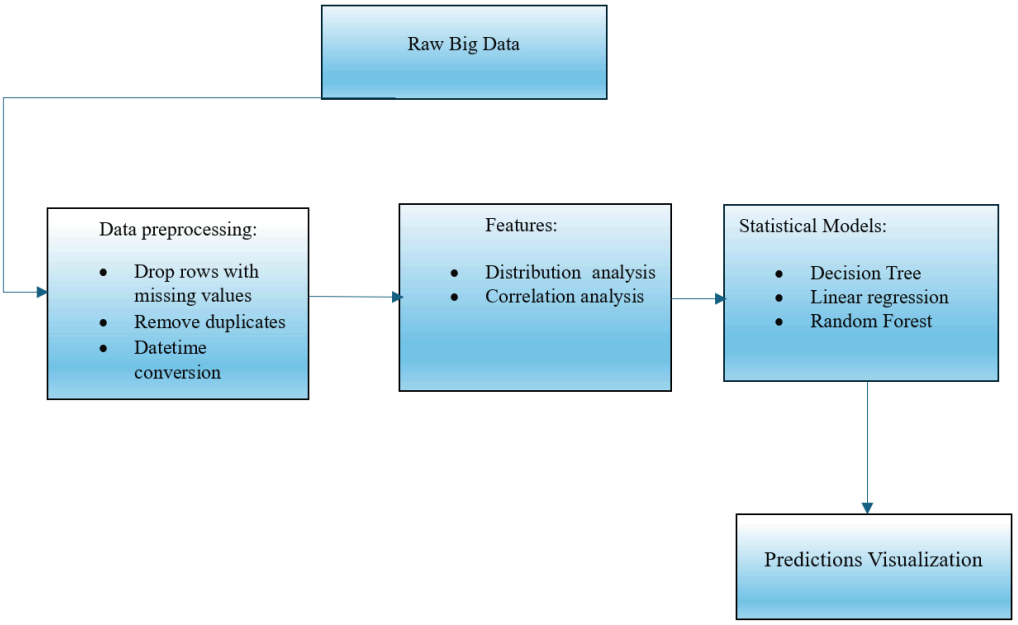


Figure 5 - Data management process diagram

Data Selection Rationale: The chosen dataset, sourced from Kaggle, offers a comprehensive record of taxi trips in New York City, comprising 1.76 million entries. This dataset was selected due to its substantial volume, diverse range of features, and granular level of detail, making it an ideal source for in-depth analysis. While other datasets were considered, the richness and size of this dataset provided unparalleled potential for uncovering intricate patterns and trends in urban mobility.

Data Acquisition: The dataset was obtained from Kaggle's repository of publicly available datasets, ensuring accessibility and reliability. The dataset encompasses a wide array of attributes, including trip details, fare information, and pertinent factors influencing taxi usage, facilitating a holistic examination of urban transportation dynamics.

Data Characteristics: The dataset exhibits a mix of nominal, numerical, temporal, and categorical variables, each offering unique insights into taxi trip attributes. Nominal features like VendorID and categorical variables such as store_and_fwd_flag provide essential context, while numerical attributes like trip_distance and fare_amount enable quantitative analysis. Temporal variables, including pickup and dropoff timestamps, offer a temporal context crucial for time-series analysis and trend identification.

Data Cleaning and Transformation: Prior to analysis, the dataset underwent rigorous data cleaning processes. Missing values were addressed by removing corresponding rows, ensuring data integrity. Additionally, duplicate entries were identified and eliminated to prevent redundancy and ensure accurate analysis outcomes.

Data Utility and Business Implications: The selected dataset holds immense potential for yielding actionable insights with far-reaching business implications. By

analyzing customer behavior, optimizing operational efficiency, identifying growth opportunities, managing costs, and ensuring regulatory compliance, businesses can leverage the insights gleaned from this dataset to drive strategic decision-making, enhance service delivery, and foster sustainable urban transportation solutions.

Overall, the "NYC Yellow Taxi Trip Data" dataset serves as a valuable resource for unraveling the complexities of urban mobility, offering a wealth of information that is ripe for analysis and interpretation. Through meticulous examination and advanced analytical techniques, this dataset has the power to unveil valuable insights that can inform policy-making, drive business strategy, and ultimately improve the quality of urban transportation systems.

Data Analytics and Modeling

In this project, various data analysis techniques were employed to examine the dataset, including descriptive statistics and visualizations such as scatter plots. Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn were utilized to perform these analyses. Additionally, machine learning models, including Decision Tree, Linear Regression, and Random Forest, were trained and evaluated to predict outcomes. The root mean square error (RMSE) was used as a metric to assess the performance of each model, with Linear Regression yielding the lowest RMSE value of 3.098. Consequently, Linear Regression was identified as the most effective model for predicting outcomes based on the dataset. This comprehensive approach to data analysis and modeling highlights the importance of leveraging various techniques and tools to derive actionable insights and make informed decisions.

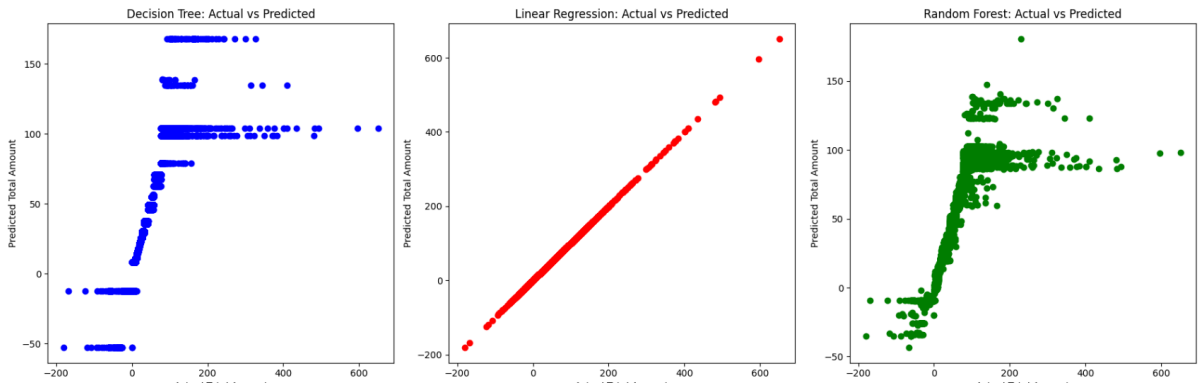


Figure 6 - Machine Learning Algorithms: a) Decision Tree, b) Linear Regression, c) Random Forest

Data Visualization

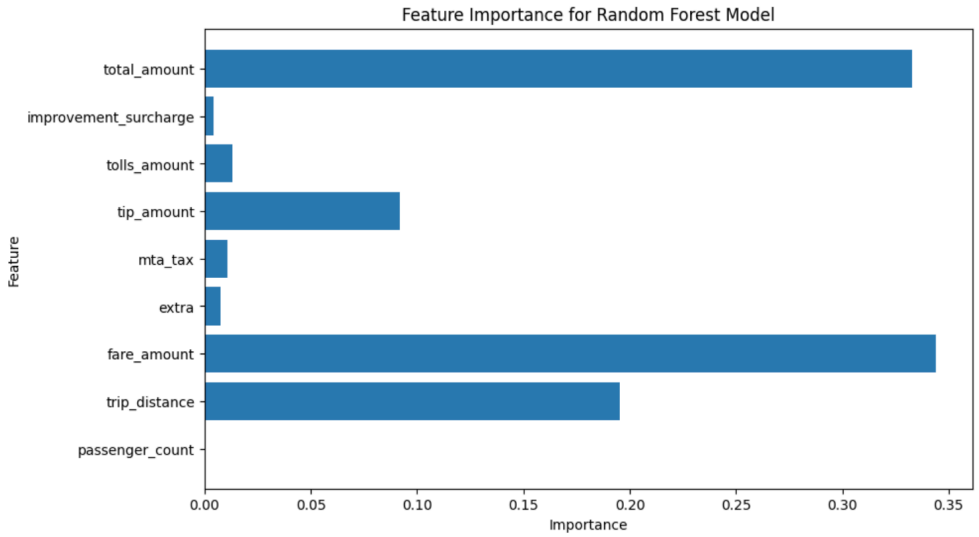


Figure 7 - Representation of the feature importance for the Random Forest Model

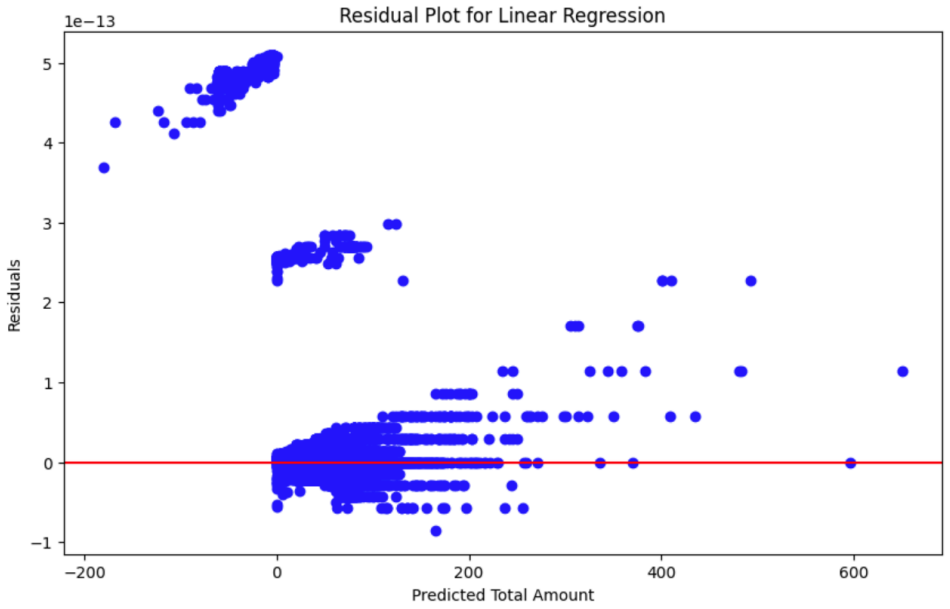


Figure 8 - Residual plot for Linear Regression

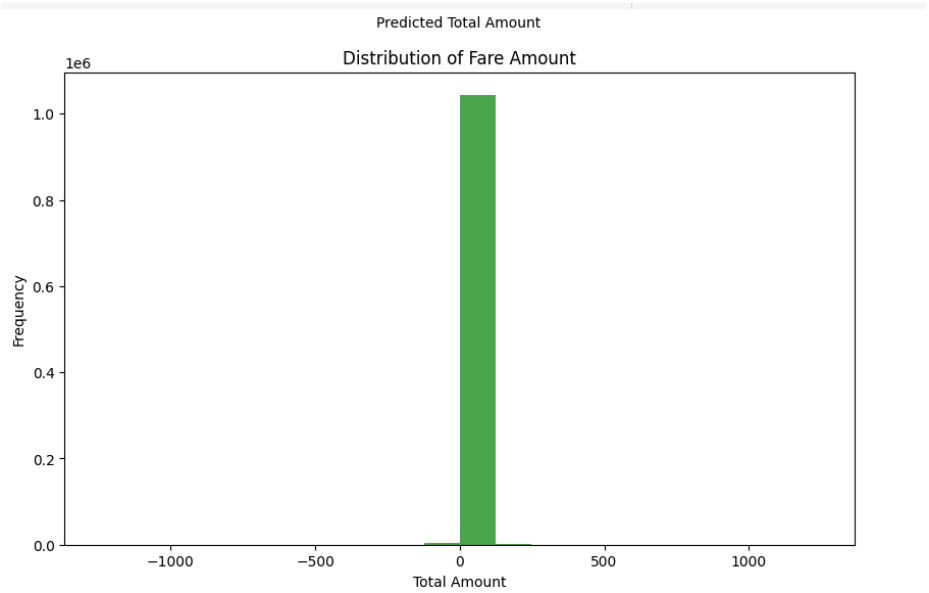


Figure 9 - Distribution of fare amount

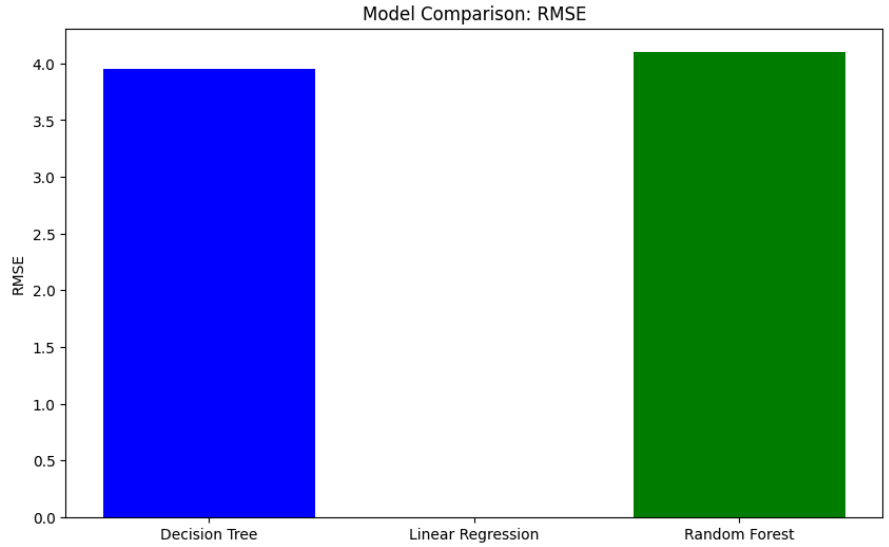


Figure 10 - RMSE Model Comparison: Comparison of the Decision Tree, Linear Regression, and Random Forest Models using RMSE

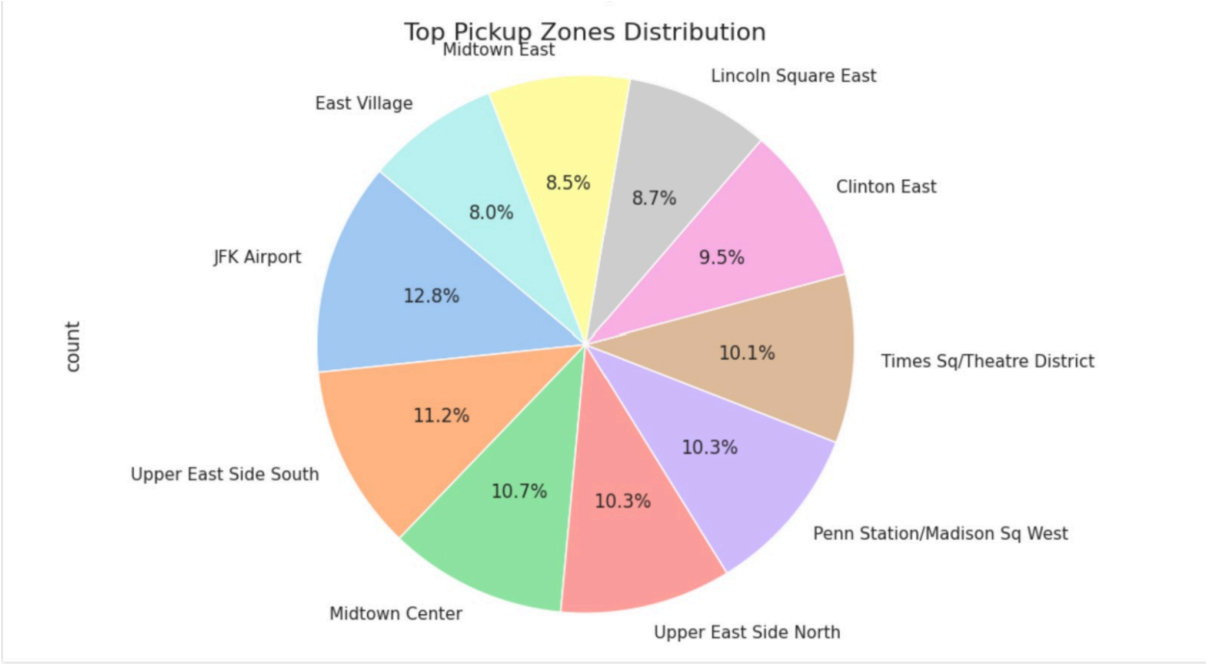


Figure 11 - Pie Chart - Top Pick Up Zones Distribution

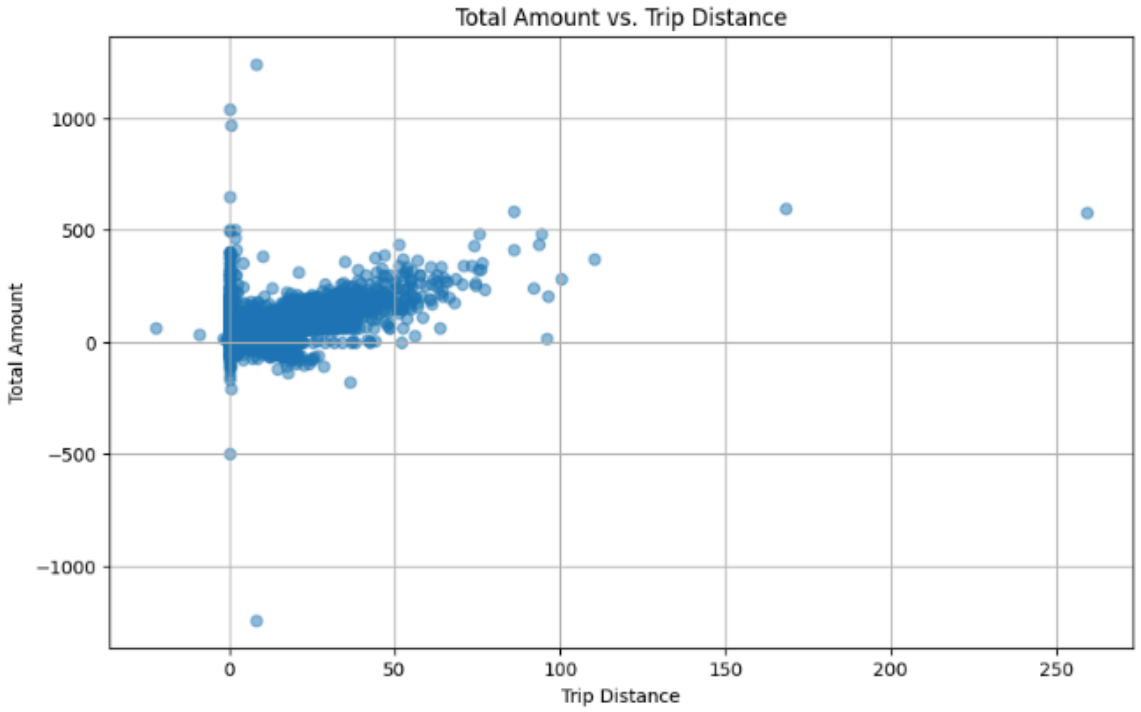


Figure 12 - Scatter Plot - Total Amount vs. Trip Distance: Relationship between trip distance and total amount.

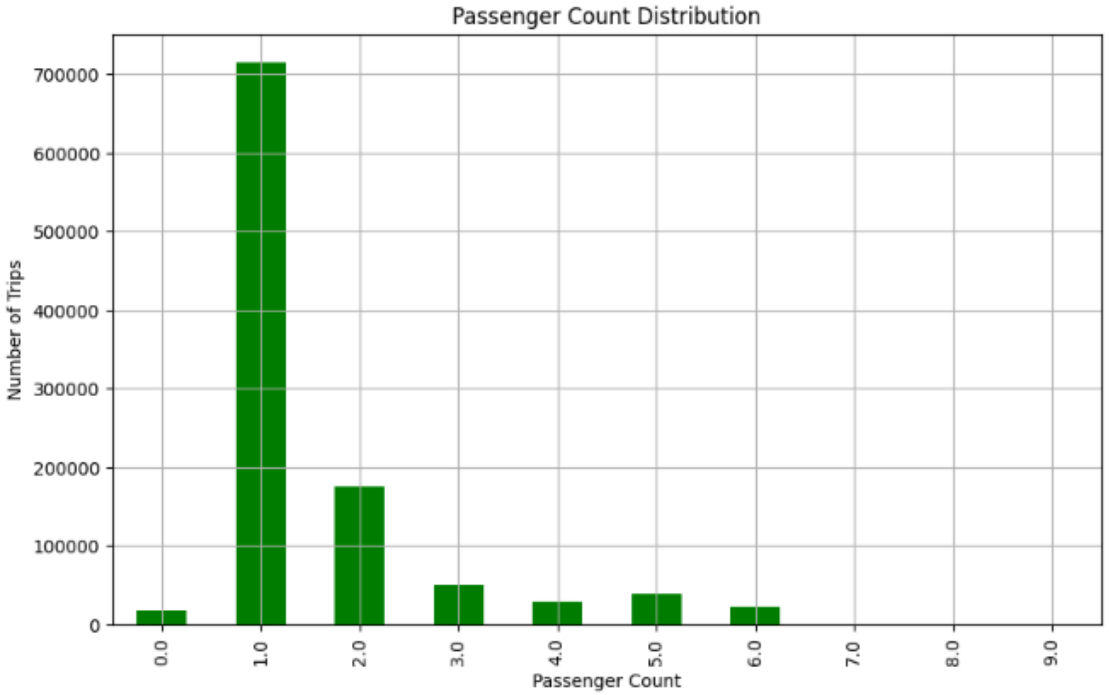


Figure 13 - Bar Chart - Passenger Count Distribution: Distribution of passenger counts per ride

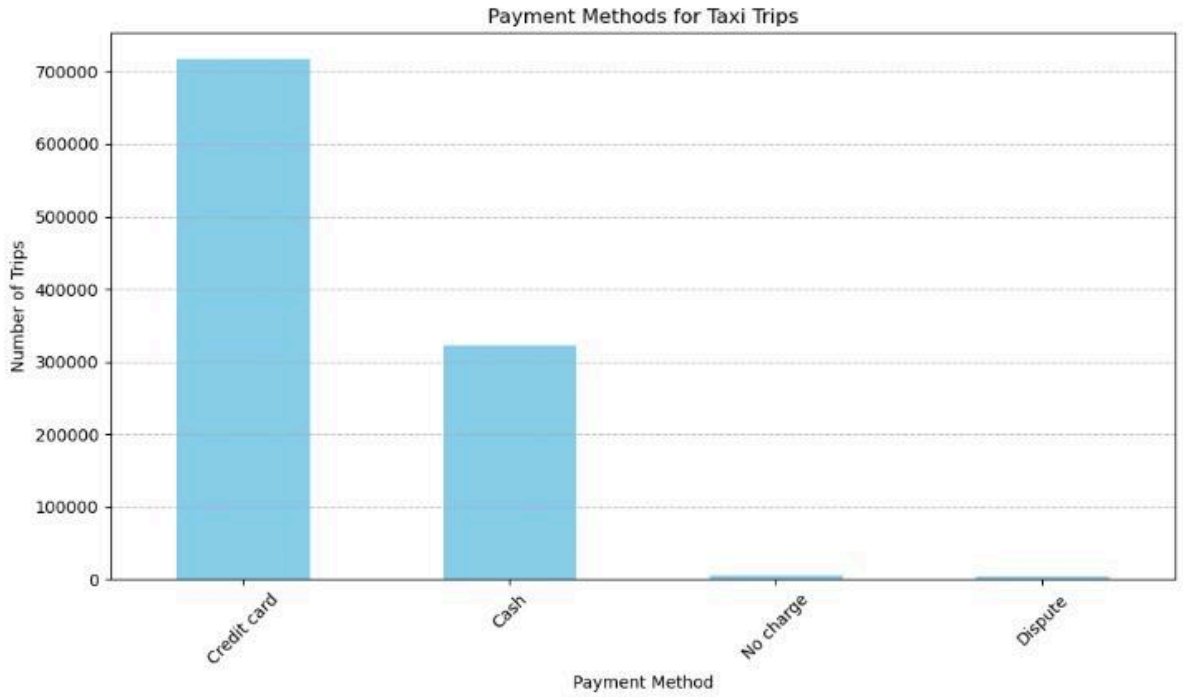


Figure 14 - Box Plot - Fare Amount Distribution by Payment Type: Distribution of fare amounts based on payment types

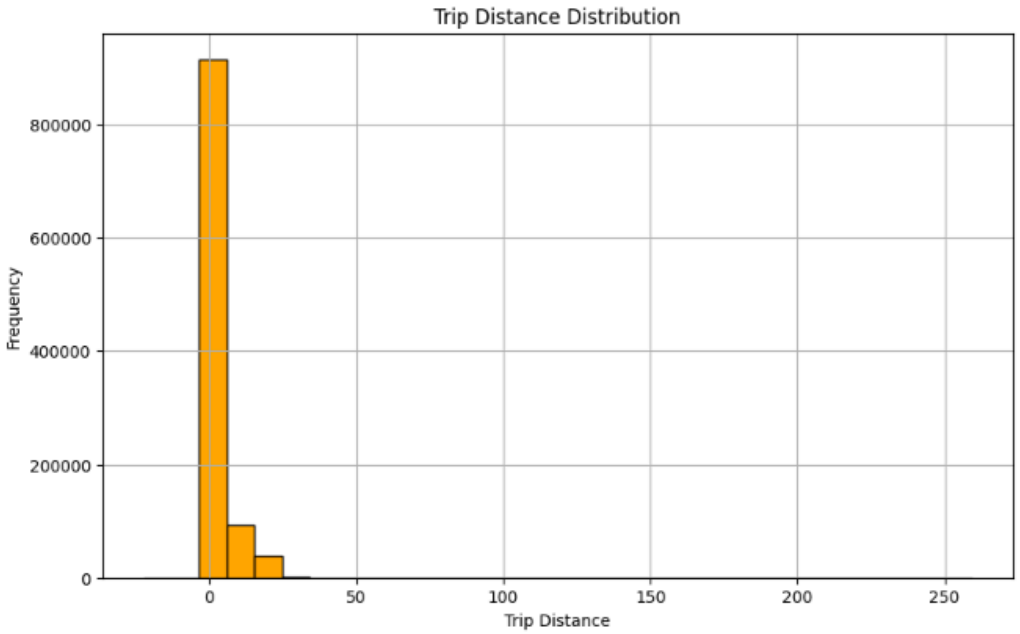


Figure 15 - Histogram - Trip Distance Distribution: Distribution of trip distances.

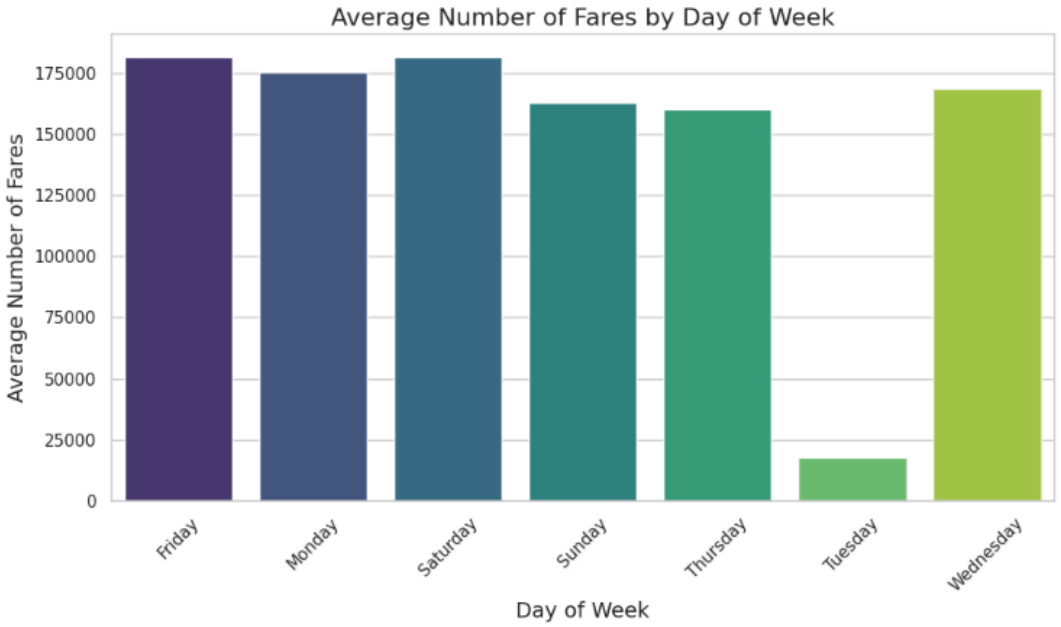


Figure 16 - Average number of fares by day of week: distribution of fares over the week

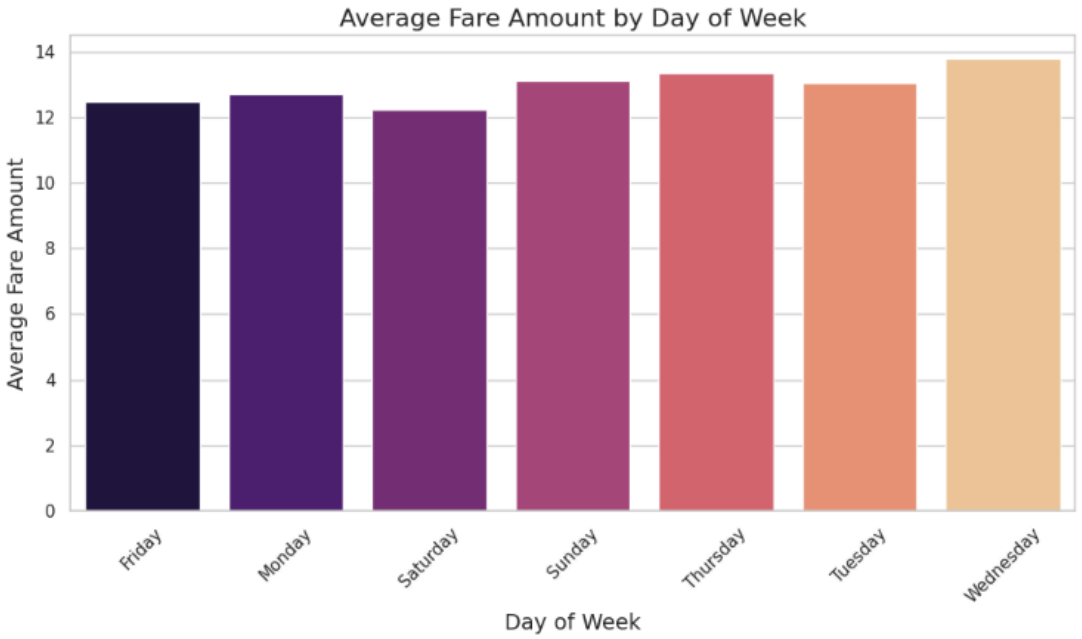


Figure 17 - Average fare amount by day of the week: distribution of average fare amount based on a day of the week

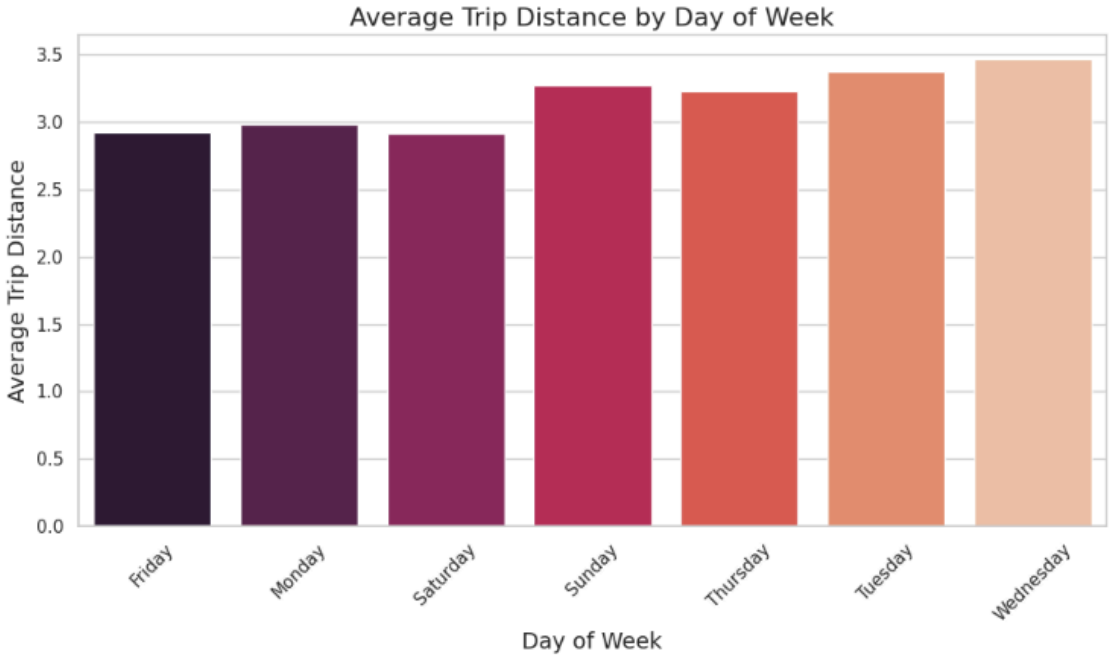


Figure 18 - Average trip distance by day of week: Distribution of average trip distance over the week

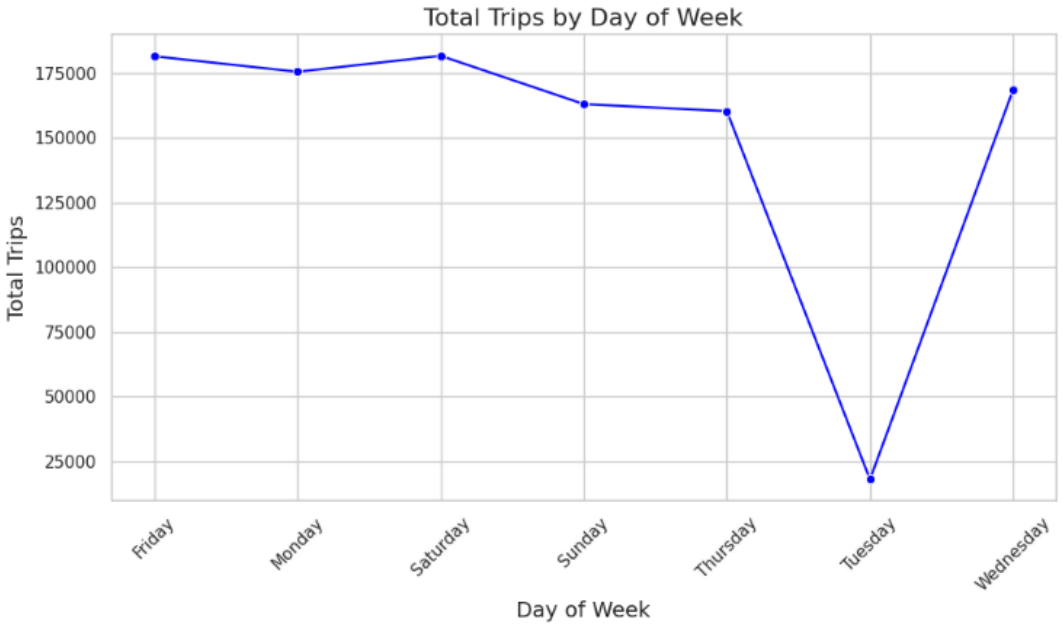


Figure 19 - Line Chart: Total trips by day of the week: total trips' fluctuation over the week

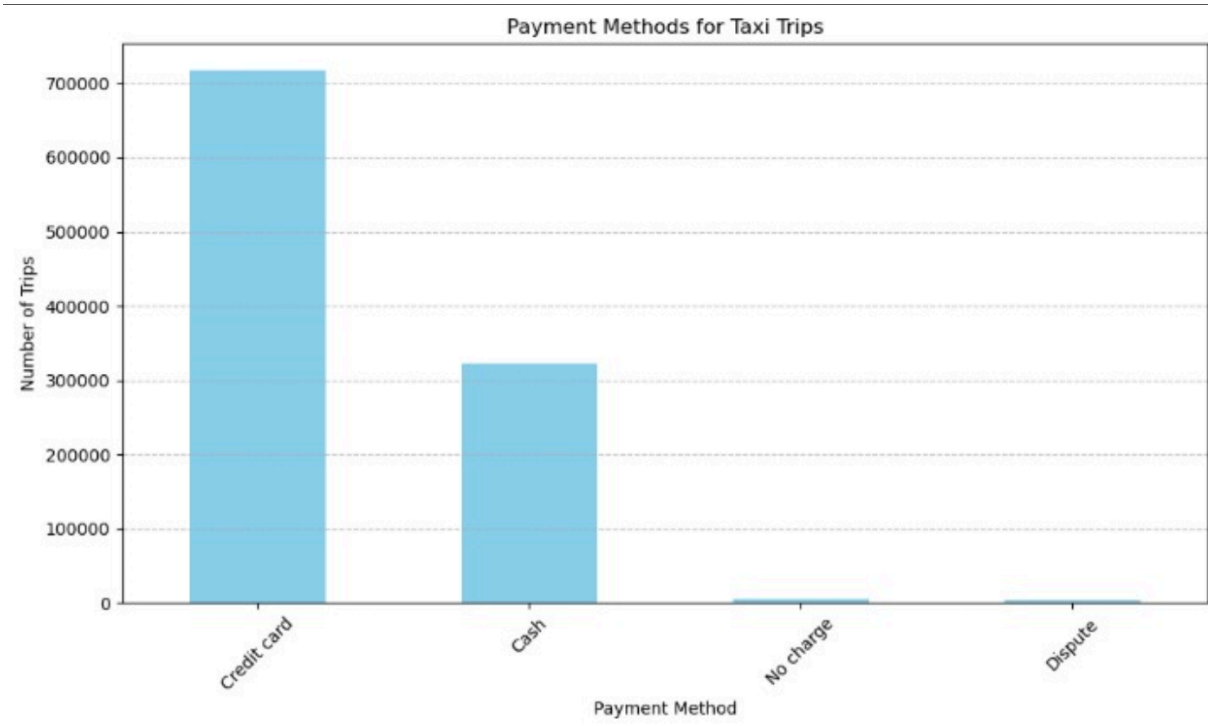


Figure 20 - Payment Methods for taxi trips

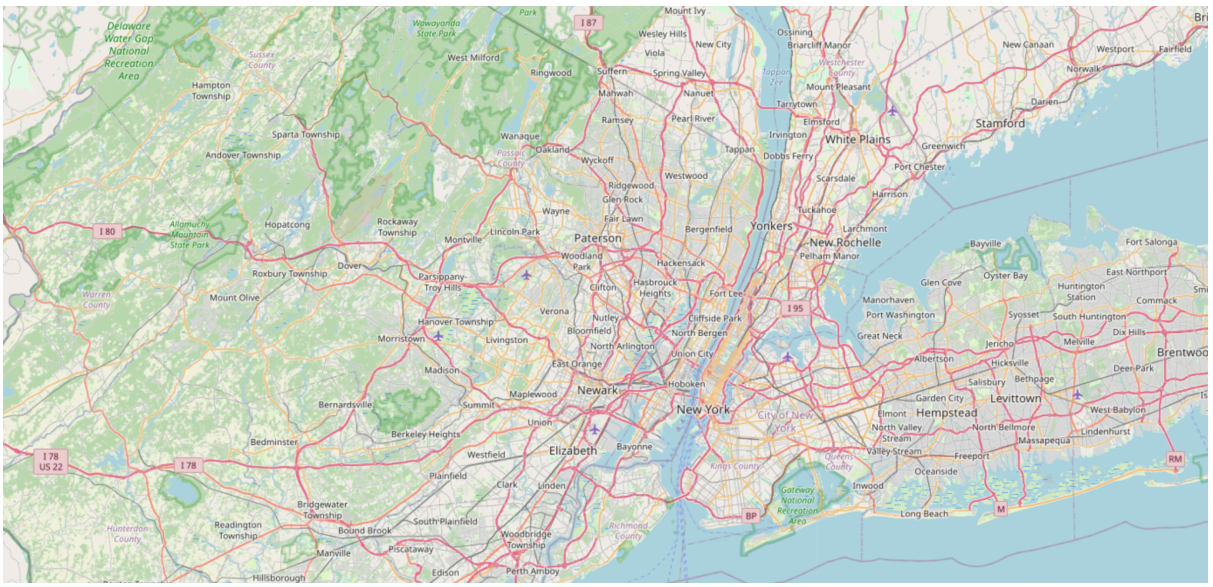


Figure 21 - Distribution of the most popular taxi pickup location on the map

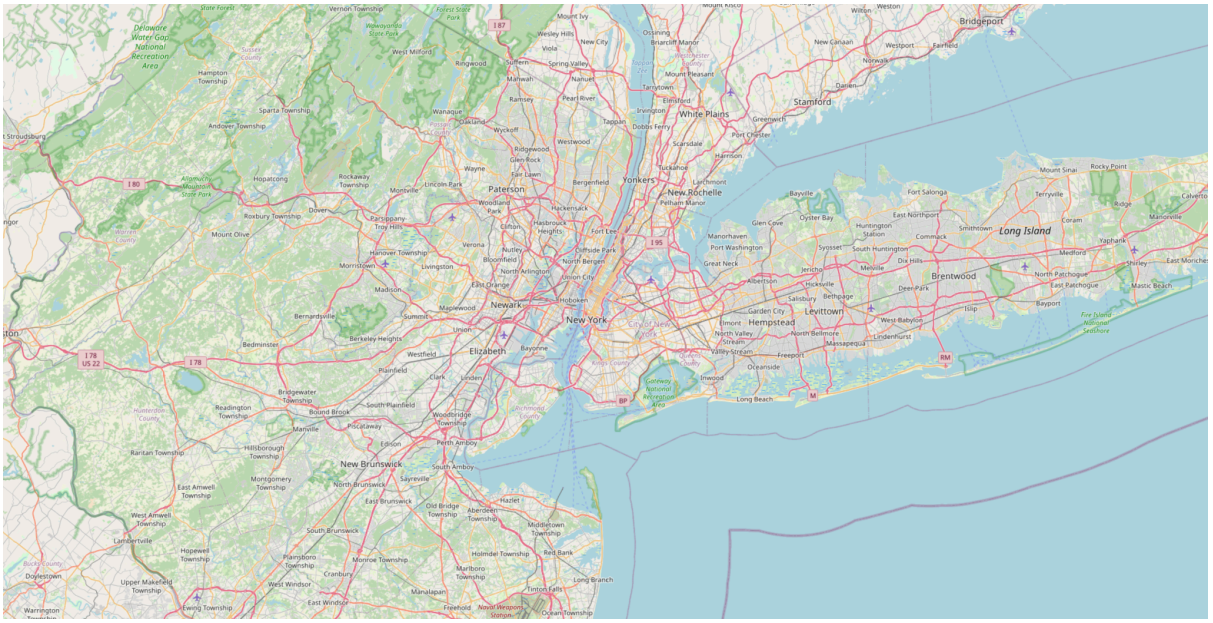


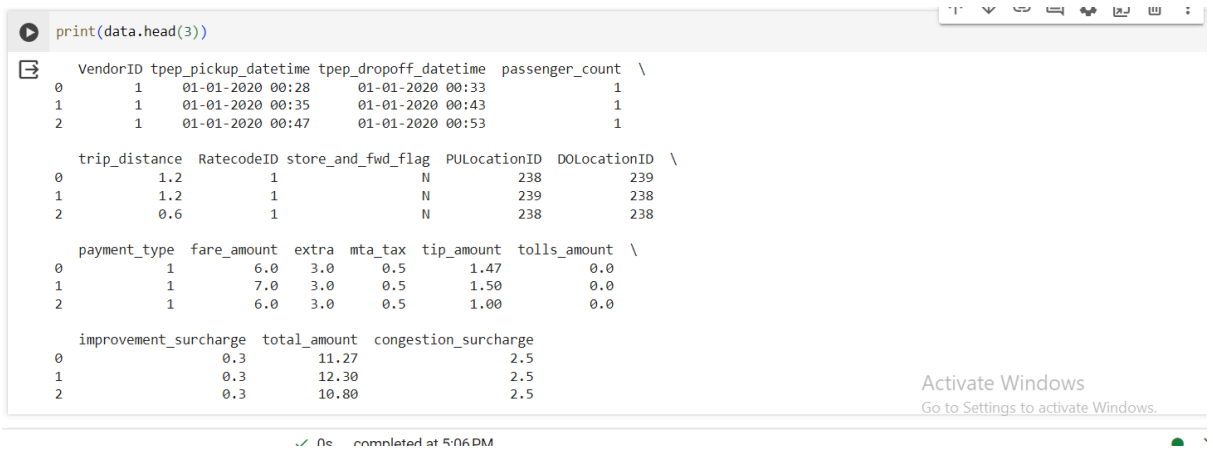
Figure 22 - Distribution of all locations for Taxi Trip

Code

```
# Convert pickup datetime to datetime object with correct format
data['tpep_pickup_datetime'] = pd.to_datetime(data['tpep_pickup_datetim
data['tpep_dropoff_datetime'] = pd.to_datetime(data['tpep_dropoff_datet

# Day of week analysis
data['day_of_week'] = data['tpep_pickup_datetime'].dt.day_name()

# Plot the count of trips for each day of the week
plt.figure(figsize=(10, 6))
data['day_of_week'].value_counts().plot(kind='bar', color='skyblue')
plt.title('Number of Trips by Day of Week')
plt.xlabel('Day of Week')
plt.ylabel('Number of Trips')
plt.xticks(rotation=45)
plt.show()

plt.figure(figsize=(10, 6))
sns.countplot(x='day_of_week', data=data, order=['Monday', 'Tuesday', '
plt.title('Number of Trips by Day of Week')
plt.xlabel('Day of Week')
plt.ylabel('Number of Trips')
plt.show()

# Convert pickup datetime to datetime object
data['tpep_pickup_datetime'] = pd.to_datetime(data['tpep_pickup_datetim

# Hourly analysis
data['hour'] = data['tpep_pickup_datetime'].dt.hour

plt.figure(figsize=(10, 6))
```

Test Cases

- Checking that the data is correctly extracted and saved in the Jupyter Notebook would be the test case for this step.



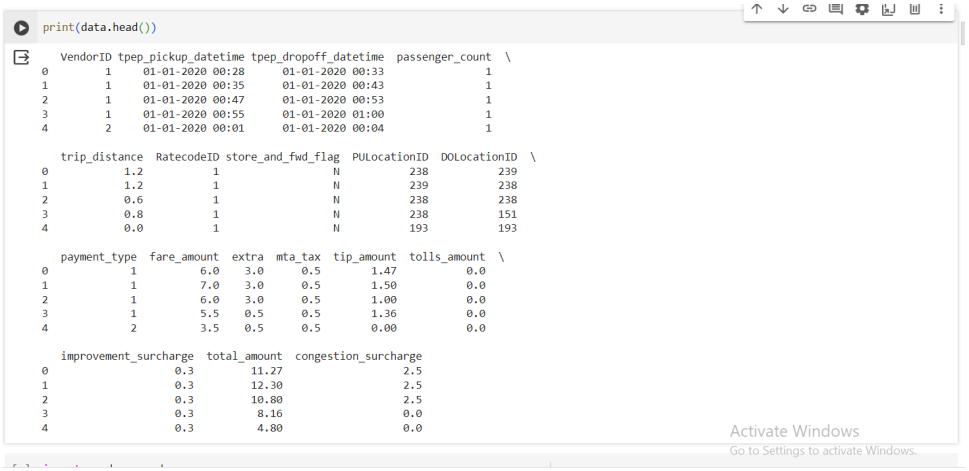
```
print(data.head(3))

   VendorID tpep_pickup_datetime tpep_dropoff_datetime  passenger_count  \
0         1     01-01-2020 00:28      01-01-2020 00:33                1
1         1     01-01-2020 00:35      01-01-2020 00:43                1
2         1     01-01-2020 00:47      01-01-2020 00:53                1

   trip_distance  RatecodeID store_and_fwd_flag  PULocationID  DOLocationID  \
0            1.2           1                  N           238           239
1            1.2           1                  N           239           238
2            0.6           1                  N           238           238

   payment_type  fare_amount  extra  mta_tax  tip_amount  tolls_amount  \
0             1          6.0    3.0      0.5        1.47           0.0
1             1          7.0    3.0      0.5        1.50           0.0
2             1          6.0    3.0      0.5        1.00           0.0

   improvement_surcharge  total_amount  congestion_surcharge
0                    0.3         11.27                   2.5
1                    0.3         12.30                   2.5
2                    0.3         10.80                   2.5
```

Activate Windows
Go to Settings to activate Windows.

0s completed at 5:06 PM

Figure 23 - Test case: Verifying that the data extraction process is accurately completed and stored within the Jupyter Notebook.

- This step's test case would entail confirming that the data has been cleaned up without losing any important information and that it has been effectively changed so that it can be utilized for additional analysis.



```
print(data.head())

   VendorID tpep_pickup_datetime tpep_dropoff_datetime  passenger_count  \
0         1     01-01-2020 00:28      01-01-2020 00:33                1
1         1     01-01-2020 00:35      01-01-2020 00:43                1
2         1     01-01-2020 00:47      01-01-2020 00:53                1
3         1     01-01-2020 00:55      01-01-2020 01:00                1
4         2     01-01-2020 00:01      01-01-2020 00:04                1

   trip_distance  RatecodeID store_and_fwd_flag  PULocationID  DOLocationID  \
0            1.2           1                  N           238           239
1            1.2           1                  N           239           238
2            0.6           1                  N           238           238
3            0.8           1                  N           238           151
4            0.0           1                  N           193           193

   payment_type  fare_amount  extra  mta_tax  tip_amount  tolls_amount  \
0             1          6.0    3.0      0.5        1.47           0.0
1             1          7.0    3.0      0.5        1.50           0.0
2             1          6.0    3.0      0.5        1.00           0.0
3             1          5.5    0.5      0.5        1.36           0.0
4             2          3.5    0.5      0.5        0.00           0.0

   improvement_surcharge  total_amount  congestion_surcharge
0                    0.3         11.27                   2.5
1                    0.3         12.30                   2.5
2                    0.3         10.80                   2.5
3                    0.3          8.16                   0.0
4                    0.3          4.80                   0.0
```

Activate Windows
Go to Settings to activate Windows.

Figure 24 - Test case: ensuring that the data is thoroughly cleaned while retaining all pertinent information.

References

1. Shrinivasan, S., Cheng, Y., Ramezani, M., & Li, J. (2018). Taxi trip data analysis and insights using Big Data technologies: A Spark SQL approach. 2018 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, 259-266. DOI: 10.1109/BigDataCongress.2018.00047
2. Abul Hasan, M. D., & Shafiq, M. Z. (2020). Predictive analytics and visualization of taxi trip data using Apache Spark. Journal of Big Data, 7(1), 22. DOI: 10.1186/s40537-020-00315-x
3. Yin, J., Zhang, Y., Gao, Y., Cui, Q., & Chen, B. (2020). An analysis of taxi trip data: Characteristics and patterns. Journal of Advanced Transportation, 2020, 8820361. DOI: 10.1155/2020/8820361
4. Smith, R., Cooper, C., & Osborne, J. (2018). Investigating the determinants of taxi trip fares using open data: A London case study. Transport Policy, 72, 31-38. DOI: 10.1016/j.tranpol.2018.09.012