Problem Statement

Objective

This assignment aims to give you an idea of applying EDA in real-world scenarios. In addition to using the techniques that you have learnt in the EDA module, you will develop a basic understanding of analytics for real-world situations, how to integrate them into a narrative and draw usable insights.

Business Value

In this case study, you will learn about exploratory data analytics with the help of a data set on yellow taxi rides in New York City. Taxis play a crucial role in New York City's urban transport network. With the city's dynamic environment, taxi companies need to continuously adapt and optimise their operations to meet changing demand patterns, ensure profitability and enhance customer satisfaction.

As an analyst at an upcoming taxi operation in NYC, you are tasked to use the 2023 taxi trip data to uncover insights that could help optimise taxi operations. The goal is to analyse patterns in the data that can inform strategic decisions to improve service efficiency, maximise revenue and enhance passenger experience.

Dataset Overview

Context

The yellow taxi trip records include fields capturing details of the trips taken everyday by yellow taxis in New York City. The data is stored in Parquet format (*.parquet*). The data set is from 2009 to 2024. However, for this assignment, we will only be using the data from 2023. The data for each month is present in a different Parquet file. You will receive 12 files, one for each of the months in 2023. The data was collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers such as vendors and taxi-hailing apps.

Content

- The dataset contains a set of 12 *Parquet*files having records on attributes such as vendor ID, time and locations of pick-up and drop-off, payment type, fare amount and tips.
- Apart from the trip data, you will find a zipped folder *taxi_zones*. This folder has a shapefile containing geometric data about the taxi zones. You will learn more about this while attempting the assignment.

You can find the data dictionary here: Data Dictionary.

Acknowledgements

This dataset is free and is publicly available at the New York City TLC. Do not download the dataset from this page.

Scoring and Penalty

- **Total Marks**: **200** (140 for code notebook and 60 for report)
- **Extension and penalty**: As given in your learner handbooks

Instructions

1. Each learner should attempt this assignment individually.
2. Programming Language: Python
3. You will be provided with the data set and a starter notebook. You have to perform analyses in the starter notebook only.
4. You must not change any headings, subheadings, questions or tasks in your notebook as this can cause problems with grading.
5. For analyses and processing tasks, you should use only the following libraries: NumPy, Pandas, Matplotlib, Seaborn and Plotly.
6. The data will have inconsistencies and outliers. Handle them as per your understanding and mention them in your report.
7. You are encouraged to search the web and consult AI tools for conceptual understanding. However, using plagiarized code is strictly prohibited and using purely AI-generated code is strongly discouraged.
8. Submitting plagiarized and purely AI-generated code or reports will result in significant penalties to your scores.

Submission Guidelines

1. To submit your solution, push your submission to GitHub and submit the GitHub link here.
2. You are required to upload your solution as a **zip file** titled *"EDA_Optimising_NYC_Taxis_<your_name>.zip"* in a public GitHub repository.
3. The repository should be named after the assignment.
4. The zipped folder that you upload should contain two files: an Interactive Python Notebook (.ipynb) that contains your code, and a Report Document (.pdf) that presents your visualizations, analysis, results, insights, and outcomes.
5. Note that these files should only be generated from the starter files provided to you.
6. Both your Jupyter notebook and report should contain your name and the assignment title.
7. Mention all assumptions made in the report.
8. Your answers to all the tasks mentioned in the starter notebook should be present in the report. Any graphs/plots you generate for analysis should also be attached in the report.

Results Expected from Learners

Present the overall approach of the analysis in a report document. Mention the problem statement and the analysis approach briefly.

In the starter notebook, you will find headings, subheadings, and checkpoints stating the tasks you need to perform. The marks associated with each checkpoint will also be mentioned in the notebook. Keep in mind not to edit the cells with marking schemes and questions. You can find a brief description of the tasks below.

1. Data Preparation [5 marks]

There are total twelve files. The data first needs to be combined into one file.

- Sample an appropriate fraction from each file
- Combine the sampled parts into one

2. Data Cleaning [30 marks]

(a) Fixing Columns [10 marks]

(b) Handling Missing Values [10 marks]

(c) Handling Outliers [10 marks]

Hints:

- Note that it is not necessary to replace the missing value in EDA, if you have to replace it, what should be the approach? Mention the approach.
- Identify if there are outliers in the dataset. Also, mention why you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

3. Exploratory Data Analysis [90 marks]

(a) **General EDA: Finding Patterns [40 marks]**

i. Classify variables into categorical and numerical
ii. Analyse the distribution of taxi pickups by hours, days of the week, and months
iii. Filter out the zero/negative values in fares, distance and tips
iv. Analyse the monthly revenue trends
v. Find the proportion of each quarter's revenue in the yearly revenue
vi. Analyse and visualise the relationship between distance and fare amount
vii. Analyse the relationship between fare/tips and trips/passengers
viii. Analyse the distribution of different payment types
ix. Load the taxi zones shapefile and display it
x. Merge the zone data with trips data
xi. Find the number of trips for each zone/location ID
xii. Add the number of trips for each zone to the zones dataframe
xiii. Plot a map of the zones showing number of trips
xiv. Conclude with results

(b) **Detailed EDA: Insights and Strategies [50 marks]**

i. Identify slow routes by comparing average speeds on different routes
ii. Calculate the hourly number of trips and identify the busy hours
iii. Scale up the number of trips from above to find the actual number of trips
iv. Compare hourly traffic on weekdays and weekends
v. Identify the top 10 zones with high hourly pickups and drops
vi. Find the ratio of pickups and dropoffs in each zone
vii. Identify the top zones with high traffic during night hours
viii. Find the revenue share for nighttime and daytime hours
ix. For the different passenger counts, find the average fare per mile per passenger
x. Find the average fare per mile by hours of the day and by days of the week
xi. Analyse the average fare per mile for the different vendors
xii. Compare the fare rates of different vendors in a distance-tiered fashion
xiii. Analyse the tip percentages
xiv. Analyse the trends in passenger count
xv. Analyse the variation of passenger counts across zones
xvi. Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.

4. **Conclusion [15 marks]**

Final insights and recommendations:

(a) Propose recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies

(b) Provide suggestions on strategically positioning cabs across different zones to make best use of insights

(c) Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates

Points to note:

- Conclude the analysis, draw final insights and propose recommendations.
- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms..
- Include visualisations and summarise the most important results in the report. You are free to choose the graphs that explain the numerical/categorical variables. Insights should explain why the variable is important.

Evaluation Guidelines

The following rubrics will be used while judging your solutions to the above tasks.

Table 1: Rubrics

Criteria	Meets expectations	Does not meet expectations
Data Understanding	1. All data quality issues are correctly identified and reported. 2. Wherever required, the meanings of the variables are correctly interpreted and written either in the comments or text.	1. Data quality issues are overlooked or are not identified correctly such as missing values, outliers and other data quality issues. 2. The variables are interpreted incorrectly or the meaning of variables is not mentioned.
Data Cleaning and Manipulation	1. Data quality issues are addressed in the right way (missing value imputation analysis and other kinds of data redundancies, etc.). 2. If applicable, data is converted to a suitable and convenient format to work with using the right methods. 3. Manipulation of strings and dates is done correctly wherever required	1. Data quality issues are not addressed correctly. 2. The variables are not converted to an appropriate format for analysis. 3. String and date manipulation is not done correctly or is done using complex methods

Continued on next page

Table 1: Rubrics (Continued)

Criteria	Meets expectations	Does not meet expectations
Data Analysis (EDA) *These come from the analyses in your code and insights reported in your report. The code and report will be graded for following a reasonable order along with the notebook and implementing the points mentioned here.*	1. The right problem is solved which is coherent with the needs of the business. The analysis has a clear structure and the flow is easy to understand. 2. The analyses successfully identify at least the 5 important driver variables. 3. Business-driven, type-driven and data-driven metrics are created for the important variables and utilised for analysis. The explanation for creating the derived metrics is mentioned and is reasonable. 4. Multivariate analysis is performed correctly and is able to identify the important combinations of driver variables. The combinations of variables are chosen such that they make business or analytical sense. 5. The most useful insights are explained correctly in the comments. 6. Appropriate plots are created to present the results of the analysis. The choice of plots for respective cases is correct. The plots should clearly present the relevant insights and should be easy to read. The axes and important data points are labelled correctly.	1. The analyses do not address the right problem or deviate from the business objectives. The analysis lacks a clear structure and is not easy to follow. 2. The analysis is not performed in sufficient detail and thus some crucial insights are missed out. The analyses are not able to identify enough important driver variables. 3. New metrics are not derived wherever appropriate. The explanation for creating the derived metrics is either not mentioned or the metrics are not reasonable. 4. Derived metrics are not analysed correctly/are insufficiently utilised. 5. Important insights are not mentioned in the report. 6. Relevant plots are not created. The choice of plots is not ideal and the plots are either difficult to interpret or lack clarity or neatness. The plots do not clearly present relevant insights. The axes and important data points are not labelled correctly/neatly.

Continued on next page

Table 1: Rubrics (Continued)

Criteria	Meets expectations	Does not meet expectations
Presentation and Recommendations	1. The report has a clear structure, is not too long, and explains the most important results concisely in simple language. 2. The recommendations to solve the problems or the outcomes and insights, whichever is applicable, are realistic, actionable and coherent with the analysis. 3. If any assumptions are made, they are stated clearly.	1. The report lacks structure, is too long or does not put emphasis on the important observations. The language used is complicated for business people to understand. 2. The recommendations to solve the problems or the outcomes are either unrealistic, non-actionable or incoherent with the analysis. 3. Contains unnecessary details or lacks important ones. 4. Assumptions made, if any, are not stated clearly.
Conciseness and readability of the code	1. The code is concise and syntactically correct. Wherever appropriate, built-in functions and standard libraries are used instead of writing long code (if-else statements, loops, etc.). 2. Custom functions are used to perform repetitive tasks. 3. The code is readable with appropriately named variables and detailed comments are written wherever necessary.	1. Long and complex code is used instead of shorter built-in functions. 2. Custom functions are not used to perform repetitive tasks resulting in the same piece of code being repeated multiple times. 3. Code readability is poor because of vaguely named variables or lack of comments wherever necessary.