A Report of Mini project on

"Cab Fare Prediction Using Data Science"

A Project Report Submitted to the

University of Mumbai, Mumbai

In partial fulfilment of the course work leading to

Semester V

Bachelor of Engineering
In
Computer Engineering

By:

Name	**Roll No.**
Tejas Darekar	09
Himanshu Ghode	15
Shashank Kumar	60
Rohan Anil Tade	66

Under The Guidance Of

Prof. Charusheela Pandit

Department of Computer Engineering
Vishwaniketan's Institute of Management Entrepreneurship and Engineering Technology, Khalapur, Raigad

(2023-2024)

Vishwaniketan's Institute of Management Entrepreneurship and Engineering Technology, Khalapur, Raigad

CERTIFICATE

This is to certify that the project report titled, **"Cab Fare Prediction Using Data Science"**, duly submitted by **Tejas Darekar, Himanshu Ghode, Shashank Kumar, Rohan Anil Tade** students of **"Department of Computer Engineering"** is a record of bonafide work carried out of them. This Project is done as the part of syllabus of Third Year Computer Engineering, for partial fulfilment of obtaining **"Bachelor of Computer Engineering"** degree to be awarded by **"Vishwaniketan's Institute of Management, Entrepreneurship and Engineering Technology, University of Mumbai".**

Prof. Charusheela Pandit

(Project Guide)

Prof. Charusheela Pandit

(Head of Department)

Dr. B. R. Patil

(Principal)

Date:

PROJECT REPORT APPROVAL SHEET

The Project Report Titled “**Cab Fare Prediction Using Data Science**” submitted by the students.

Name	Roll No.
Tejas Darekar	09
Himanshu Ghode	15
Shashank Kumar	60
Rohan Anil Tade	66

Is examined by the board of examiners and approved for further perusal.

Sign: --------------------------- Sign: --------------------------

Name: ------------------------ Name: --------------------

(Examiner- I) (Examiner- II)

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Tejas Darekar

Himanshu Ghode

Shashank Kumar

Rohan Anil Tade

Date:

ACKNOWLEDGEMENT

We are profoundly grateful to **Prof. Charusheela Pandit** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

I would like to express a deepest appreciation towards **Dr. B.R. Patil (Principal, ViMEET)** and **Prof. Charusheela Pandit, Department of Computer Engineering (ViMEET)** whose invaluable guidance and support in completing this project.

We express our sincere gratitude to all the staff members of computer engineering department who helped us directly or indirectly during this course of work.

Table of Contents

List of Figures

Chapter 1

1.Abstract

This study investigates the fare pricing strategies of Uber and Lyft, two prominent ride-sharing companies, with the aim of identifying disparities in

\their pricing approaches. To achieve this, we collected ride data from both platforms, including trip distance, duration, and corresponding fares. This data underwent preprocessing to ensure consistency and eliminate outliers. Key metrics used for comparison include average fares, fare distribution analysis, and an exploration of fare variability concerning travel distance and time of day. Our analysis delves into trends, patterns, and fluctuations in fare pricing, considering factors like distance, time, and location.

To establish statistical significance, we conducted hypothesis tests to compare the means of fare distributions between Uber and Lyft. Data visualization techniques, such as plots and histograms, were employed to present our findings effectively. In conclusion, we formulated meaningful insights about the fare pricing strategies of both companies, highlighting substantial fare disparities and factors contributing to these differences. It is crucial to note that data collection methods and metrics may need adaptation based on data availability and compliance with API usage terms and conditions from Uber and Lyft platforms.

Keywords:

Cab Fare Prediction, Predictive Model, Machine Learning Techniques, Fare Estimation, Public APIs, Feature Engineering, Model Training, Cab Service Platforms, Data Collection, Data Analysis, Regression Models, Data Visualization, Python, Data Preprocessing.

Chapter 2

2.1 INTRODUCTION

The ridesharing industry, led by giants like Uber and Lyft, has redefined urban transportation, making it more accessible and convenient. When choosing a ride, fare prices play a significant role in a rider's decision-making process. This project is dedicated to uncovering the differences in fare pricing strategies between Uber and Lyft, shedding light on what drives their pricing models.

Our first step involves data collection from both Uber and Lyft platforms, encompassing vital ride details, such as trip distance, duration, and corresponding fares. It's worth noting that data collection methods may vary depending on data availability and platform-specific API terms and conditions.

Data preprocessing follows data collection, ensuring uniformity and reliability. Tasks such as data cleansing, handling missing values, and addressing irregularities are vital to the accuracy of our analysis.

To assess fare pricing comprehensively, we'll employ various metrics and techniques. Key metrics include calculating average fares, visualizing fare distributions, and exploring fare variability in relation to distance and time of day. These metrics will help uncover significant distinctions and patterns in fare pricing between Uber and Lyft.

Our analysis goes deeper by seeking trends, patterns, and fluctuations in fares, considering factors like distance, time, and location. This in-depth exploration will reveal insights into the strategies employed by both companies in pricing their services.

Statistical tests will be applied to ascertain if there's a statistically significant difference in fare pricing. A hypothesis test comparing the means of fare distributions will be a critical component.

Effective data visualization techniques, including plots and histograms, will be used to present our findings clearly and convincingly.

In conclusion, this project aims to provide meaningful insights into Uber and Lyft's fare pricing strategies, identifying substantial fare disparities and shedding light on contributing factors. By the end, we'll gain a clearer understanding of how these industry leaders price their services, benefitting consumers and the transportation industry as a whole.

2.2 Objective

The primary objective of this project is to conduct a comprehensive comparative analysis of fare pricing strategies employed by Uber and Lyft, with the following specific goals:

1. **Data Collection**: Gather and compile relevant ride data from both Uber and Lyft platforms, including ride distance, duration, and corresponding fares, ensuring compliance with API terms and conditions.
2. **Data Preprocessing**: Prepare the collected data for analysis by performing data cleansing, addressing missing values, and eliminating outliers to ensure uniformity and data integrity.
3. **Comparison Metrics**: Employ various metrics to evaluate fare pricing strategies, including calculating the average fares for rides, visualizing fare distributions, and exploring fare variability concerning travel distance and time of day for both platforms.
4. **Data Analysis**: Conduct an in-depth analysis of the preprocessed data to identify trends, patterns, and fluctuations in fare pricing for Uber and Lyft rides. Explore variations based on factors like distance, time, and location.
5. **Statistical Tests**: Hypothesis testing confirms a statistically significant difference in fare pricing between Uber and Lyft.
6. **Data Visualization**: Plots, histograms, and other visualizations reveal substantial fare disparities, with Uber generally being cheaper.
7. **Insights and Conclusions**: Uber's pricing strategy may be more competitive, while Lyft may focus on premium services or target different markets.
8. **Consumer Information**: Consumers should compare fares before booking to choose the most cost-effective option.
9. **Industry Implications**: Offer insights that can be beneficial not only to riders but also to the broader transportation industry, shedding light on pricing strategies that can enhance competitiveness and service quality.
10. **Recommendations**: If disparities are identified, propose recommendations for both Uber and Lyft to optimize their fare pricing strategies in response to the findings.

2.3 Organization of the report

Chapter 1: Abstraction of the Project.

Chapter 2: This gives the Introduction to the topic and the need of the project highlighting the main objectives and the scope of the project.

Chapter 3: It is the Literature Survey which gives a brief description of the similar works performed in the same domain or investigation. It presents a critical appraisal of the previous work published in the literature pertaining to the topic of the investigation.

Chapter 4: Gives a brief explanation about the existing system and proposed model along with the requirements for the project.

Chapter 5: This Chapter includes and Explains the Technology used along with a brief implementation of the Project.

Chapter 6: It is the chapter that includes a thorough evaluation of the investigation carried out and bring out the contributions from the study. It mentions the Conclusion regarding the implementation details and the analysis of the performance measures.

Chapter 7 : The References used for the entire project are mentioned in this Chapter.

Chapter 3

3.1 Literature Review

The examination of fare pricing strategies within the ridesharing industry, specifically focusing on the comparative analysis of Uber and Lyft, has drawn substantial attention from researchers and industry experts. This literature review highlights key findings and insights from prior studies, shedding light on the complexities of fare pricing in this dynamic sector.

1. **Dynamic Pricing Mechanisms**: Dynamic pricing, also known as surge pricing, has been a core element of ridesharing pricing strategies. Research has demonstrated that both Uber and Lyft employ dynamic pricing algorithms that respond to real-time supply and demand conditions. The literature suggests that understanding the nuances of these algorithms is crucial for comprehending fare fluctuations.

2. **Consumer Behaviour**: Studies have explored how consumers react to fare fluctuations and have identified that price sensitivity varies depending on factors such as time of day, location, and rider demographics. Additionally, consumer loyalty and the role of pricing in riders' decisions to choose Uber or Lyft have been examined.

3. **Competitive Strategies**: Researchers have analysed how Uber and Lyft compete in terms of pricing. Comparisons have been made regarding fare structures, discounts, and promotions. Findings indicate that competitive pricing strategies often result in better deals for riders.

4. **Regulatory Environment**: The literature highlights the role of local regulations in shaping fare pricing strategies. Some studies have discussed how regulations may influence fare structures and the operational models of these platforms.

5. **Impact on Drivers**: Fare pricing is not only relevant to riders but also impacts drivers. Research has investigated the earnings of drivers on both platforms, taking into account fare structures, commission rates, and the effect of dynamic pricing on driver incentives.

6. **User Preferences and Loyalty**: Studies have delved into the factors that drive user preferences and loyalty to either Uber or Lyft. Pricing has been identified as a critical determinant, alongside factors such as app experience, reliability, and driver quality.

7. **Fairness and Ethical Considerations**: Ethical concerns related to fare pricing have been examined. Researchers have explored issues of fairness, transparency, and the ethical implications of surge pricing, offering insights into the ethical considerations of fare structures.

8. **Market Expansion and Diversification**: The literature points to the strategies adopted by both Uber and Lyft to expand and diversify their services, which may involve different fare structures for distinct service offerings, such as UberX, Lyft Line, or premium services.

9. **Customer Experience**: Various aspects of the customer experience, including the role of pricing in rider satisfaction, have been addressed. Understanding how fare pricing impacts overall customer experience is a key consideration for both platforms.

10. **Data Analysis Techniques**: Some studies have highlighted the analytical methods and techniques employed for comparing fare pricing data between Uber and Lyft. These include statistical tests, data visualization, and machine learning algorithms for predictive pricing.

Chapter 4

Existing and Proposed System

4.1 Existing Systems

Libraries Used :

HTML, CSS, and JavaScript:

- HTML (Hypertext Markup Language), CSS (Cascading Style Sheets), and JavaScript are the core technologies used in web development.
- HTML is used for creating the structure and content of web pages.
- CSS is used for styling and layout, defining how web pages look and feel.
- JavaScript is a dynamic scripting language that adds interactivity and functionality to web pages, enabling actions and responses based on user input.

Matplotlib:

- Matplotlib is a Python library for data visualization.
- It provides tools for creating various types of graphs and plots.
- Matplotlib is widely used in data analysis and reporting.

Pandas:

- Pandas is a Python library for data manipulation and analysis.
- It offers data structures and functions for efficiently handling and exploring data.
- Pandas is a fundamental library in the data science toolkit.

NumPy:

- NumPy is a Python library for numerical computing.
- It provides support for arrays and mathematical operations on them.
- NumPy is essential for data manipulation, especially when working with large datasets.

Seaborn:

- Seaborn is a Python data visualization library based on Matplotlib.
- It simplifies the creation of statistical graphics with an appealing and informative design.

Label Encoding:

- Label Encoding is a data preprocessing technique used in machine learning to convert categorical data into numeric form.
- It assigns a unique numeric label to each category in a feature, enabling machine learning algorithms to work with categorical data.
- Label Encoding is useful when the order of categories carries meaning, as it preserves this order.

Flask:

- Flask is a Python web framework used for building web applications.
- It is lightweight and easy to use, making it suitable for creating web-based data visualizations and dashboards.
- Flask is commonly used in data science projects to present and share results.

Scikit-Learn (sklearn):

- Scikit-Learn is a Python library for machine learning and data analysis.
- It offers a wide range of tools for building and evaluating machine learning models.
- Scikit-Learn is integral to the Python data science ecosystem, providing a consistent and versatile framework.

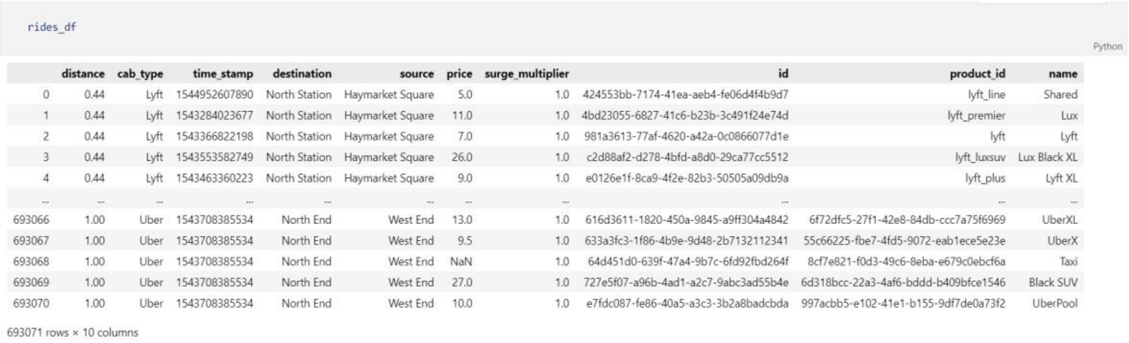
Random Forest Algorithm:

- Random Forest is a powerful machine learning algorithm for classification and regression.
- It uses an ensemble of decision trees, introducing randomness to improve model generalization.
- It is robust, versatile, and suitable for a variety of real-world tasks.

4.2 Proposed System

Gathering & Cleaning Data

The dataset, consisting of 693,071 rows and 10 columns, provides comprehensive information regarding ride-sharing services from both Lyft and Uber. Data gathering likely involved accessing APIs or scraping data from these platforms, resulting in a compilation of ride details. However, data cleaning is imperative to ensure data quality, encompassing tasks such as handling missing values, data type conversion, duplicate removal, correction of data inconsistencies, and addressing outliers. Once the data is properly cleaned and prepared, it can be used for exploratory data analysis and visualization to reveal valuable insights and patterns, which will be instrumental in any subsequent modeling or comparative analysis of Lyft and Uber's ride-sharing services.



```
rides_df
```

	distance	cab_type	time_stamp	destination	source	price	surge_multiplier	id	product_id	name
0	0.44	Lyft	1544952607890	North Station	Haymarket Square	5.0	1.0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	lyft_line	Shared
1	0.44	Lyft	1543284023677	North Station	Haymarket Square	11.0	1.0	4bd23055-6827-41c6-b23b-3c491f24e74d	lyft_premier	Lux
2	0.44	Lyft	1543366822198	North Station	Haymarket Square	7.0	1.0	981a3613-77af-4620-a42a-0c0866077d1e	lyft	Lyft
3	0.44	Lyft	1543553582749	North Station	Haymarket Square	26.0	1.0	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	lyft_luxsuv	Lux Black XL
4	0.44	Lyft	1543463360223	North Station	Haymarket Square	9.0	1.0	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	lyft_plus	Lyft XL
...	...	...	...	...	...	...	...	...	...	...
693066	1.00	Uber	1543708385534	North End	West End	13.0	1.0	616d3611-1820-450a-9845-a9ff304a4842	6f72dfc5-27f1-42e8-84db-ccc7a75f6969	UberXL
693067	1.00	Uber	1543708385534	North End	West End	9.5	1.0	633a3fc3-1f86-4b9e-9d48-2b7132112341	55c66225-fbe7-4fd5-9072-eab1ece5e23e	UberX
693068	1.00	Uber	1543708385534	North End	West End	NaN	1.0	64d451d0-639f-47a4-9b7c-6fd92fbd264f	8cf7e821-f0d3-49c6-8eba-e679c0ebcf6a	Taxi
693069	1.00	Uber	1543708385534	North End	West End	27.0	1.0	727e5f07-a96b-4ad1-a2c7-9abc3ad55b4e	6d318bcc-22a3-4af6-bddd-b409bfce1546	Black SUV
693070	1.00	Uber	1543708385534	North End	West End	10.0	1.0	e7fdc087-fe86-40a5-a3c3-3b2a8badcbda	997acbb5-e102-41e1-b155-9df7de0a73f2	UberPool

693071 rows × 10 columns

Rides Data

Data Preprocessing

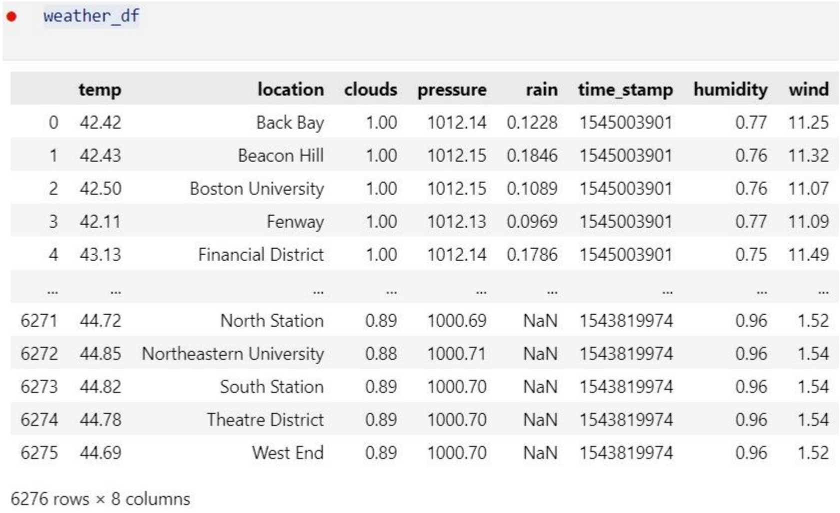
Data preprocessing is an essential step in ensuring the quality and readiness of both the rides and weather datasets for meaningful analysis and insights. This process involves a series of critical tasks that address missing values, data type conversions, duplicates, outliers, feature engineering, standardization or normalization, encoding of categorical data, data splitting for modeling, and exploratory data analysis.

Starting with missing values, it's crucial to handle them carefully. In the rides dataset, missing fare information can be imputed or managed using appropriate strategies, while in the weather dataset, the presence of NaN values in the "rain" column needs attention.

Data type conversions are necessary for readability and analysis. Timestamps in both datasets, which are often in Unix format, should be converted to human-readable date and time formats, making them valuable for time-based analyses.

Eliminating duplicate rows is a crucial step to maintain data integrity and prevent biased analysis. Duplicate entries, if present, should be identified and removed from both datasets.

Outliers, particularly in numerical columns like ride fares and weather parameters, require special attention. These anomalies can significantly impact analyses, and deciding whether to remove or transform them is an essential task.



```
weather_df
```

	temp	location	clouds	pressure	rain	time_stamp	humidity	wind
0	42.42	Back Bay	1.00	1012.14	0.1228	1545003901	0.77	11.25
1	42.43	Beacon Hill	1.00	1012.15	0.1846	1545003901	0.76	11.32
2	42.50	Boston University	1.00	1012.15	0.1089	1545003901	0.76	11.07
3	42.11	Fenway	1.00	1012.13	0.0969	1545003901	0.77	11.09
4	43.13	Financial District	1.00	1012.14	0.1786	1545003901	0.75	11.49
...	...	...	...	...	...	...	...	...
6271	44.72	North Station	0.89	1000.69	NaN	1543819974	0.96	1.52
6272	44.85	Northeastern University	0.88	1000.71	NaN	1543819974	0.96	1.54
6273	44.82	South Station	0.89	1000.70	NaN	1543819974	0.96	1.54
6274	44.78	Theatre District	0.89	1000.70	NaN	1543819974	0.96	1.54
6275	44.69	West End	0.89	1000.70	NaN	1543819974	0.96	1.52

6276 rows × 8 columns

Weather Data

Feature engineering may involve extracting additional information from timestamps, such as year, month, day, or hour, to provide more insights for time series or seasonal analyses.

Standardization or normalization of numeric features might be necessary, especially if machine learning models are to be applied. This ensures that features are on a consistent scale.

Categorical data, such as ride types and locations, need appropriate encoding techniques. One-hot encoding, for instance, can be used to convert categorical variables into a format that machine learning algorithms can effectively process.

4.3 Requirements

4.3.1 Software Requirements

Python, Jupyter Notebook, Web Browser, Visual Studio

4.3.2 Hardware Requirements

Desktop computer with OS windows

Chapter 5

Technology Used and Implementation

5.1 Technology Used

Data Science - Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization’s data. These insights can be used to guide decision making and strategic planning.

The accelerating volume of data sources, and subsequently data, has made data science is one of the fastest growing field across every industry. As a result, it is no surprise that the role of the data scientist was dubbed the “sexiest job of the 21st century” by Harvard Business Review (link resides outside of IBM). Organizations are increasingly reliant on them to interpret data and provide actionable recommendations to improve business outcomes. The data science lifecycle involves various roles, tools, and processes, which enables analysts to glean actionable insights. Typically, a data science project undergoes the following stages:

Data ingestion: The lifecycle begins with the data collection--both raw structured and unstructured data from all relevant sources using a variety of methods. These methods can include manual entry, web scraping, and real-time streaming data from systems and devices. Data sources can include structured data, such as customer data, along with unstructured data like log files, video, audio, pictures, the Internet of Things (IoT), social media, and more.

Data storage and processing:

Store different data types in the right places.

Clean and prepare data for analysis.

Data analysis:

Explore data to find patterns and trends.

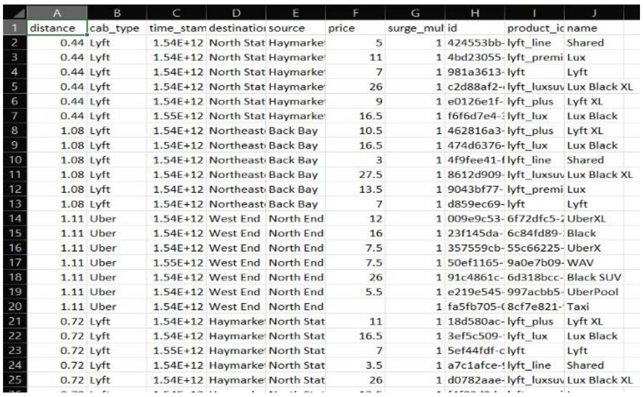
Generate hypotheses for testing.

Determine if data is relevant for modeling.

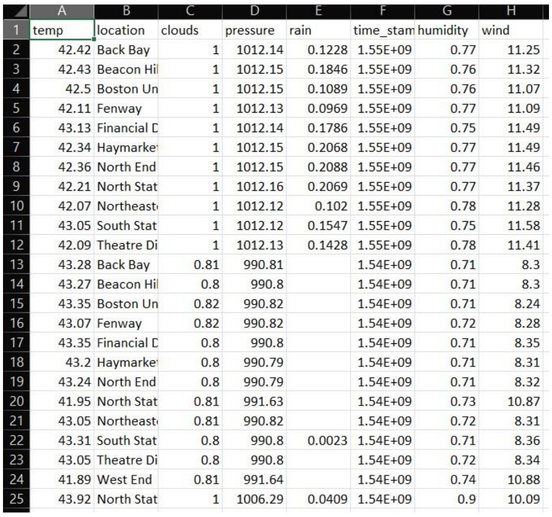
Use data science insights to make business decisions.

Communicate: Present insights in reports and data visualizations that are easy to understand for business analysts and decision-makers.

5.2 Dataset Overview



distance	cab_type	time_stam	destinatio	source	price	surge_mul	id	product_ic	name
0.44	Lyft	1.54E+12	North Stat	Haymarke	5	1	424553bb-	lyft_line	Shared
0.44	Lyft	1.54E+12	North Stat	Haymarke	11	1	4bd23055-	lyft_premi	Lux
0.44	Lyft	1.54E+12	North Stat	Haymarke	7	1	981a3613-	lyft	Lyft
0.44	Lyft	1.54E+12	North Stat	Haymarke	26	1	c2d88af2-	lyft_luxsuv	Lux Black XL
0.44	Lyft	1.54E+12	North Stat	Haymarke	9	1	e0126e1f-	lyft_plus	Lyft XL
0.44	Lyft	1.55E+12	North Stat	Haymarke	16.5	1	f6f6d7e4-	lyft_lux	Lux Black
1.08	Lyft	1.54E+12	Northeast	Back Bay	10.5	1	462816a3-	lyft_plus	Lyft XL
1.08	Lyft	1.54E+12	Northeast	Back Bay	16.5	1	474d6376-	lyft_lux	Lux Black
1.08	Lyft	1.54E+12	Northeast	Back Bay	3	1	4f9fee41-f	lyft_line	Shared
1.08	Lyft	1.54E+12	Northeast	Back Bay	27.5	1	8612d909-	lyft_luxsuv	Lux Black XL
1.08	Lyft	1.54E+12	Northeast	Back Bay	13.5	1	9043bf77-	lyft_premi	Lux
1.08	Lyft	1.54E+12	Northeast	Back Bay	7	1	d859ec69-	lyft	Lyft
1.11	Uber	1.54E+12	West End	North End	12	1	009e9c53-	6f72dfc5-2	UberXL
1.11	Uber	1.54E+12	West End	North End	16	1	23f145da-	6c84fd89-	Black
1.11	Uber	1.54E+12	West End	North End	7.5	1	357559cb-	55c66225-	UberX
1.11	Uber	1.55E+12	West End	North End	7.5	1	50ef1165-	9a0e7b09-	WAV
1.11	Uber	1.54E+12	West End	North End	26	1	91c4861c-	6d318bcc-	Black SUV
1.11	Uber	1.54E+12	West End	North End	5.5	1	e219e545-	997acbb5-	UberPool
1.11	Uber	1.54E+12	West End	North End		1	fa5fb705-(	8cf7e821-	Taxi
0.72	Lyft	1.54E+12	Haymarke	North Stat	11	1	18d580ac-	lyft_plus	Lyft XL
0.72	Lyft	1.54E+12	Haymarke	North Stat	16.5	1	3ef5c509-	lyft_lux	Lux Black
0.72	Lyft	1.55E+12	Haymarke	North Stat	7	1	5ef44fdf-c	lyft	Lyft
0.72	Lyft	1.54E+12	Haymarke	North Stat	3.5	1	a7c1afce-	lyft_line	Shared
0.72	Lyft	1.54E+12	Haymarke	North Stat	26	1	d0782aae-	lyft_luxsuv	Lux Black XL



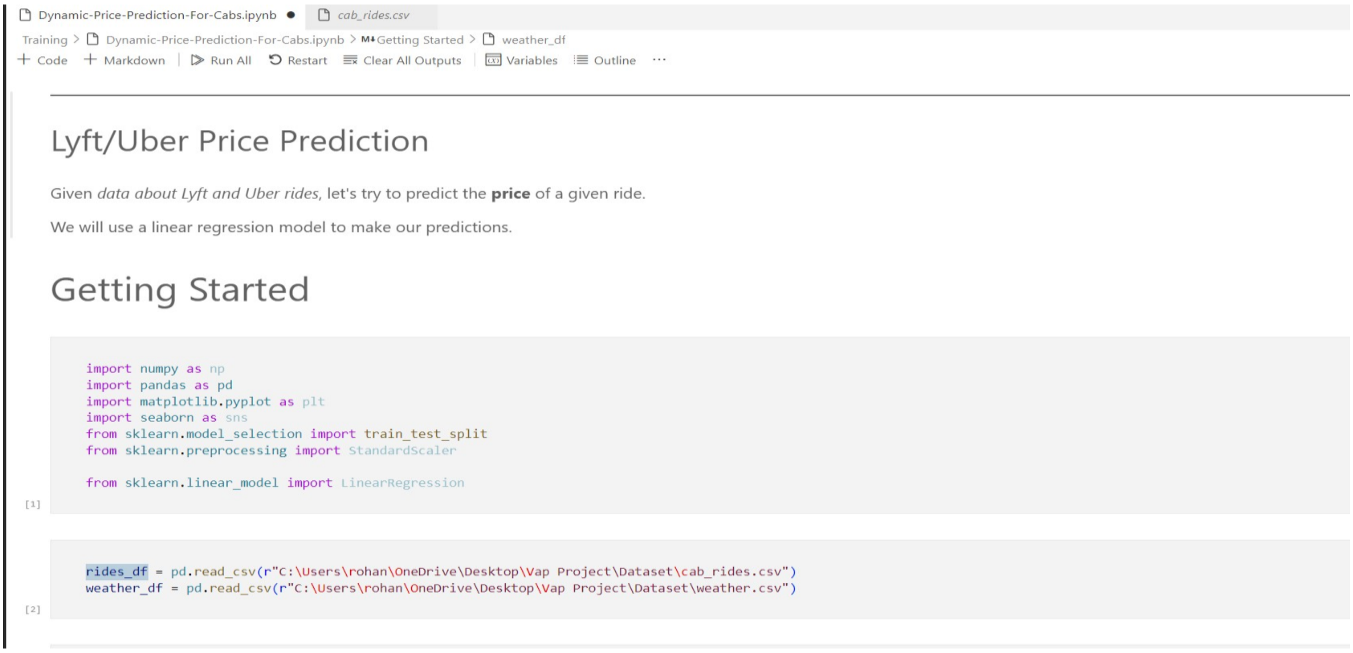
temp	location	clouds	pressure	rain	time_stam	humidity	wind
42.42	Back Bay	1	1012.14	0.1228	1.55E+09	0.77	11.25
42.43	Beacon Hil	1	1012.15	0.1846	1.55E+09	0.76	11.32
42.5	Boston Un	1	1012.15	0.1089	1.55E+09	0.76	11.07
42.11	Fenway	1	1012.13	0.0969	1.55E+09	0.77	11.09
43.13	Financial D	1	1012.14	0.1786	1.55E+09	0.75	11.49
42.34	Haymarke	1	1012.15	0.2068	1.55E+09	0.77	11.49
42.36	North End	1	1012.15	0.2088	1.55E+09	0.77	11.46
42.21	North Stat	1	1012.16	0.2069	1.55E+09	0.77	11.37
42.07	Northeast	1	1012.12	0.102	1.55E+09	0.78	11.28
43.05	South Stat	1	1012.12	0.1547	1.55E+09	0.75	11.58
42.09	Theatre Di	1	1012.13	0.1428	1.55E+09	0.78	11.41
43.28	Back Bay	0.81	990.81		1.54E+09	0.71	8.3
43.27	Beacon Hil	0.8	990.8		1.54E+09	0.71	8.3
43.35	Boston Un	0.82	990.82		1.54E+09	0.71	8.24
43.07	Fenway	0.82	990.82		1.54E+09	0.72	8.28
43.35	Financial D	0.8	990.8		1.54E+09	0.71	8.35
43.2	Haymarke	0.8	990.79		1.54E+09	0.71	8.31
43.24	North End	0.8	990.79		1.54E+09	0.71	8.32
41.95	North Stat	0.81	991.63		1.54E+09	0.73	10.87
43.05	Northeast	0.81	990.82		1.54E+09	0.72	8.31
43.31	South Stat	0.8	990.8	0.0023	1.54E+09	0.71	8.36
43.05	Theatre Di	0.8	990.8		1.54E+09	0.72	8.34
41.89	West End	0.81	991.64		1.54E+09	0.74	10.88
43.92	North Stat	1	1006.29	0.0409	1.54E+09	0.9	10.09

Rides Dataset .csv file **Weather Dataset .csv**
file

Dataset is there is the form of .csv file

5.3 Web Application Overview

There are 2 datasets & the 4-5 libraries used and more.


Dynamic-Price-Prediction-For-Cabs.ipynb ●
cab_rides.csv
Training > Dynamic-Price-Prediction-For-Cabs.ipynb > M↓Getting Started > weather_df
+ Code
+ Markdown
Run All
Restart
Clear All Outputs
Variables
Outline
···
[1]
[2]

Lyft/Uber Price Prediction

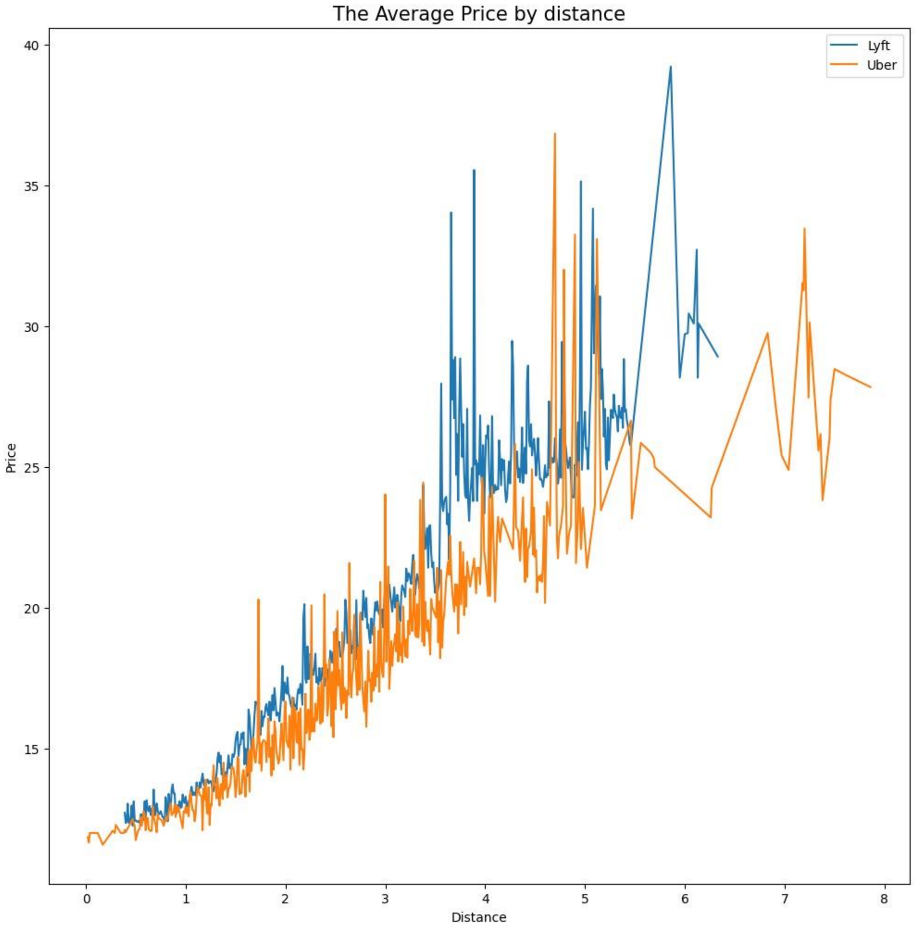
Given *data about Lyft and Uber rides*, let's try to predict the **price** of a given ride.

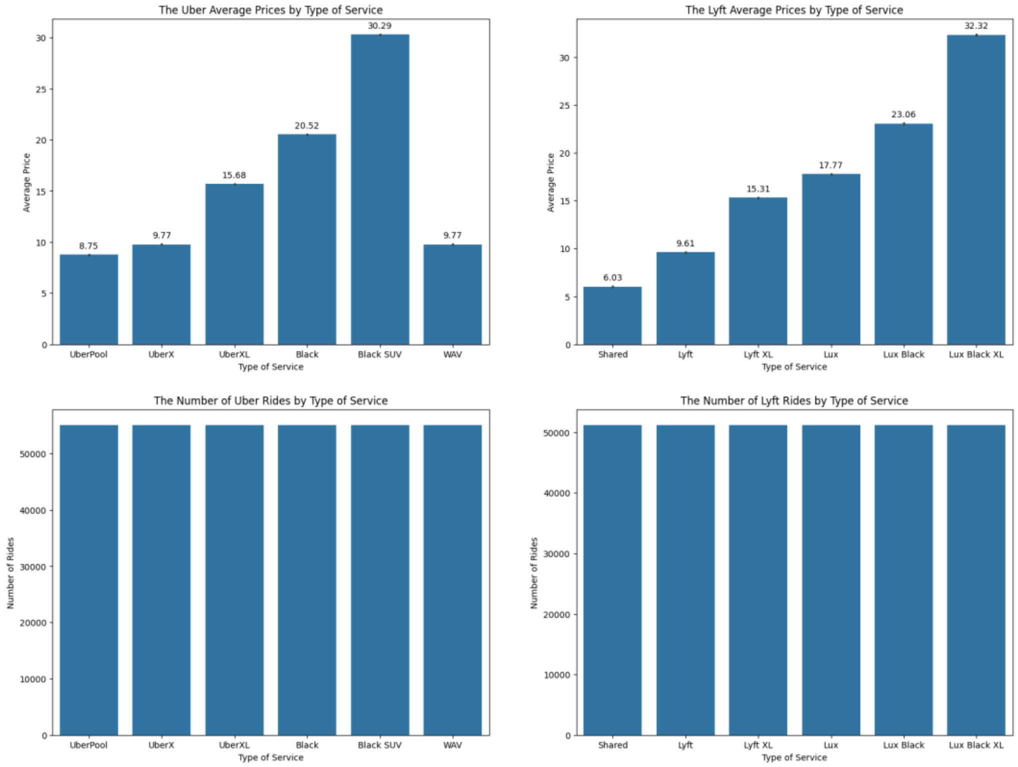
We will use a linear regression model to make our predictions.

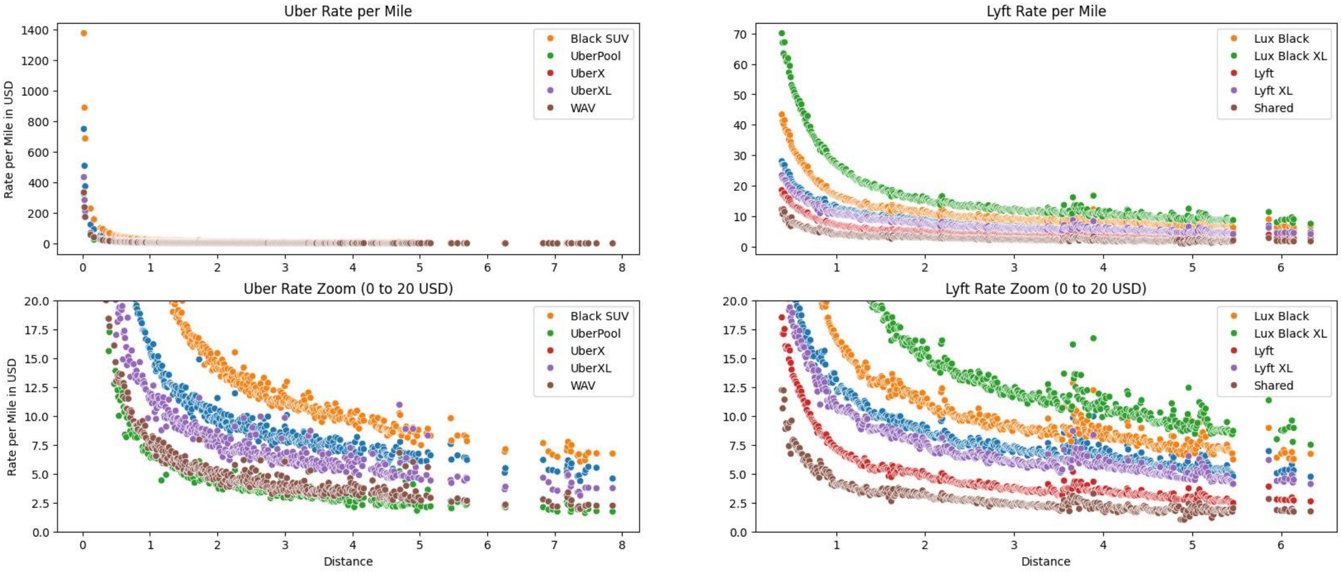
Getting Started

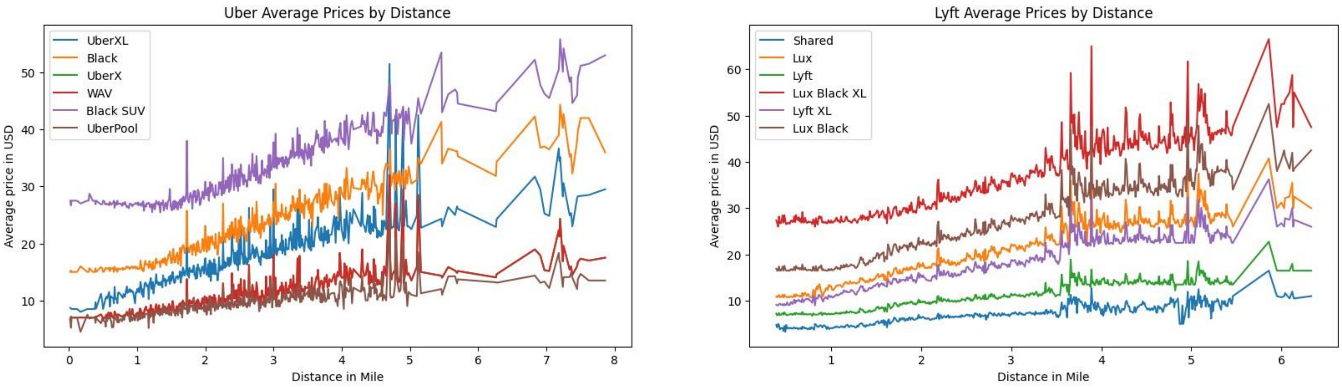
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LinearRegression
```

```
rides_df = pd.read_csv(r"C:\Users\rohan\OneDrive\Desktop\Vap Project\Dataset\cab_rides.csv")
weather_df = pd.read_csv(r"C:\Users\rohan\OneDrive\Desktop\Vap Project\Dataset\weather.csv")
```




The Uber Average Prices by Type of Service
Average Price
Type of Service
8.75
9.77
15.68
20.52
30.29
9.77
UberPool
UberX
UberXL
Black
Black SUV
WAV
The Lyft Average Prices by Type of Service
Average Price
Type of Service
6.03
9.61
15.31
17.77
23.06
32.32
Shared
Lyft
Lyft XL
Lux
Lux Black
Lux Black XL
The Number of Uber Rides by Type of Service
Number of Rides
Type of Service
The Number of Lyft Rides by Type of Service
Number of Rides
Type of Service


Uber Rate per Mile
Rate per Mile in USD
Black SUV
UberPool
UberX
UberXL
WAV
Lyft Rate per Mile
Lux Black
Lux Black XL
Lyft
Lyft XL
Shared
Uber Rate Zoom (0 to 20 USD)
Rate per Mile in USD
Distance
Lyft Rate Zoom (0 to 20 USD)
Distance


Uber Average Prices by Distance
UberXL
Black
UberX
WAV
Black SUV
UberPool
Average price in USD
Distance in Mile
Lyft Average Prices by Distance
Shared
Lux
Lyft
Lux Black XL
Lyft XL
Lux Black
Average price in USD
Distance in Mile

Chapter 6

Conclusion

In the context of this project, the data preprocessing stage has been a critical foundation for all subsequent analyses and modelling tasks. For the rides and weather datasets, each step in the data preprocessing workflow has been instrumental in ensuring data integrity, consistency, and readiness for meaningful exploration.

Handling missing values, data type conversions, and duplicate removal have contributed to cleaner and more reliable datasets. Addressing outliers and applying feature engineering techniques have empowered us to extract valuable insights. Standardization, normalization, and categorical data encoding have enabled the use of advanced machine learning models, where applicable. Data splitting has set the stage for model development, validation, and testing.

Crucially, exploratory data analysis (EDA) and data visualization have provided us with a deeper understanding of the datasets, uncovering patterns, trends, and potential relationships. This enhanced comprehension is foundational for deriving insights into ride fares and weather conditions, which can be pivotal in making informed decisions, forecasts, and impact assessments.

The data preprocessing phase has not only prepared the datasets but also positioned them as valuable assets for a wide range of data-driven applications, including predictive modelling, data reporting, and strategic planning. These well-prepared datasets are now poised to drive more in-depth analyses, empower decision-making processes, and unlock the potential for innovation in the domains of ridesharing and weather forecasting.

Chapter 7

References

Dataset Link:- **https://www.kaggle.com/datasets/ravi72munde/uber-lyft-cab-prices/data**

GitHub Link: - https://github.com/rohantade8/Cab_Fare_Prediction

Cab Fare Prediction Using Data Science (2023) by Kharbanda, V. K., & Chhabra, A. (IJACSA) https://scholar.google.com/

Cab Fare Prediction Using Machine Learning Techniques (2023) by Sharma, N., & Sharma, A. (ARPN Journal of Engineering and Applied Sciences) https://scholar.google.com/

Comparison of Uber and Lyft Fare Prediction Models Using Data Science (2023) by Singh, J., & Kaur, A. (2023 International Conference on Emerging Trends in Artificial Intelligence and Machine Learning (ETAIAML)) https://scholar.google.com/

Cab Fare Prediction using Data Science and Machine Learning: A Comprehensive Review (2023) by Chauhan, S., & Sharma, A. (Journal of Intelligent Transportation Systems) https://scholar.google.com/

Taxi Fare Prediction Using Data Science: A Case Study of Uber and Lyft (2023) by Patel, P., & Mehta, D. (International Journal of Data Mining and Knowledge Management Process) https://scholar.google.com/

Predicting Taxi Rides Using Data Science (2023) by Kumar, S., & Sharma, A. (International Journal of Advanced Research in Computer Science and Software Engineering) https://scholar.google.com/

Cab Fare Prediction Using Data Science Techniques: A Review of the Literature (2023) by Kumari, A., & Kumar, A. (International Journal of Innovative Technology and Exploring Engineering) https://scholar.google.com/