

|  |
| --- |
| **FINAL CAPSTONE PROJECT**  **Rideshare Analysis & Price Prediction – Uber & Lyft** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Developed by**  **Members:**   |  |  |  | | --- | --- | --- | | **No.** | **Student Name** | **Student ID** | | 1 | Nguyen Huynh Ngoc Han (Jack) | Student1503324 | | 2 | Tran Minh Kiet | Student1521253 | | 3 | Vu Manh Trung Hai | Student1522966 |  * **Class No.: DT2307L-NK** * **Start Date: 10th January 2025.** * **End Date: 24th February 2025.** * **Name of the Coordinator: Teacher Mr. Ho Nhat Minh** * **Date of Submission: 26th February 2025.** |

|  |
| --- |
| Rideshare Analysis & Price Prediction – Uber & Lyft |

***This final project includes 7 sections as below:***

I - Project Overview & objectives

II - Project Workflow

III - Data Cleaning & Exploration

IV - Data Analyzing & Visualizing

V - Data Preprocessing

VI - Regression Model With PySpark

VII - Communicate The Insights & Results

*Please refer to this link for the github link of this project: [Link added here]*

**I. PROJECT OVERVIEW & OBJECTIVES**

* + **Project Introduction:** The rapid development of ride-hailing services has transformed urban transportation, making it more accessible, convenient, and cost-effective.

However, pricing in ride-hailing services is highly dynamic, influenced by various factors such as demand, traffic conditions, weather, and service availability, etc.

In this project, we aim to analyze and figure out key factors that affect ride costs and build a predictive model to estimate ride prices based on these influencing variables.

* **Questions Hypotheses:**

Question 1: Ride fares increase during peak hours due to higher demand and surge pricing?

Question 2: Weather conditions, such as rain or cloudy, lead to higher fares as supply decreases and demand rises?

Question 3: Fluctuations in temperature affect price fluctuations, due to changing demand?

Question 4: Longer travel distances contribute to increased ride prices?

Question 5: Pickup and drop-off locations influence ride pricing?

Question 6: How to predict estimated ride prices based on these influencing variables?

* **Key Objectives:**.

- Identify key factors influencing ride-hailing prices.

- Build and evaluate a predictive model to predict ride fares.

- Provide data-driven insights for both ride-hailing companies and consumers.

This project will contribute to a better understanding of pricing dynamics in the ride-hailing industry, supporting more informed decision-making for both service providers and users.

* **Dataset Overview & Key Notes:** Uber and Lyft Dataset (Boston, MA) - a very popular and beginner-friendly dataset about ride-hailing data. This dataset includes various attributes such as: ride price, ride distance, car type, surge pricing multipliers, time of day, temperature information and weather conditions.

The source of the dataset is referenced in the following link:

<https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma/data>

***>>> Dataset Strengths:***

✅ Large and diverse dataset covering variety factors related ride-hailing.

✅ Includes both Uber and Lyft, enabling comparative analysis.

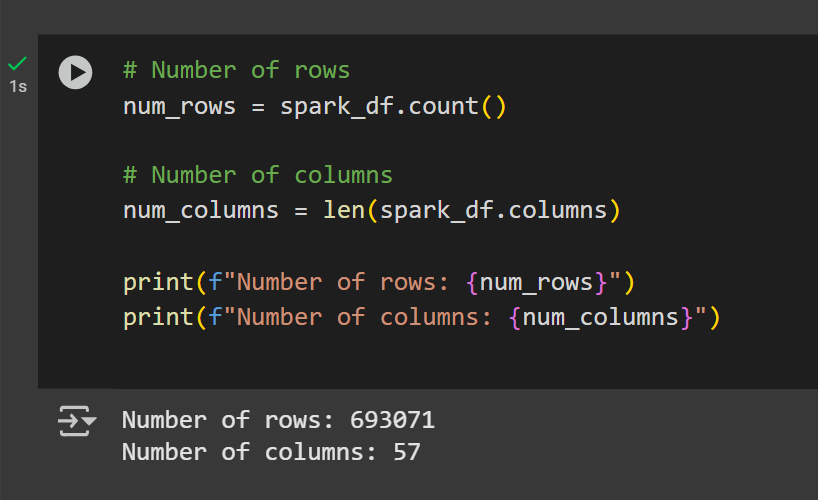
✅ Contains weather & temperature data, allowing us to analyze its impact on ride fares.

***>>> Dataset Limitations:***

❌ Limited to a single geographical area (Boston, MA), and the raw data collected in only 2 months, which may affect generalizability.

❌ No real-time traffic data, which could be a crucial factor in price variations.

❌ Does not include rider demographics or behavioral patterns, limiting user-segmentation analysis.





***>>> Definition of dataset columns as below*** *- Total 57 columns:*

- id: unique ride ID.

- timestamp: timestamp of the ride request.

- hour: hour of the ride request.

- day: day of the ride request.

- month: month of the ride request.

- datetime: date and time of the ride request.

- timezone: timezone of the ride request.

- source: starting location of the ride.

- destination: ending location of the ride.

- cab\_type: ride-hailing company (Uber/Lyft).

- product\_id: unique ride product identifier.

- name: type of ride (UberX, Lyft XL, etc.)

- price: ride fare in USD.

- distance: Trip distance in miles.

- surge\_multiplier: surge pricing multiplier.

- latitude: coordinates of source/destination.

- longitude: coordinates of source/destination.

- temperature: temperature in Fahrenheit.

- apparentTemperature: feels-like temperature.

- short\_summary: weather condition (e.g., Clear, Rainy).

- long\_summary: detailed weather description.

- precipIntensity: rainfall intensity.

- precipProbability: probability of precipitation.

- humidity: air humidity percentage.

- windSpeed: wind speed in mph.

- windGust: wind gust speed.

- windGustTime: wind gust timestamp.

- visibility: Visibility in miles.

- temperatureHigh

- temperatureHighTime

- temperatureLow

- temperatureLowTime

- apparentTemperatureHigh

- apparentTemperatureHighTime

- apparentTemperatureLow

- apparentTemperatureLowTime

- icon: types of weather

- dewPoint: điểm sương

- pressure: áp suất

- windBearing: hướng gió

- cloudCover: Cloud coverage percentage

- uvIndex: UV radiation index

- visibility.1

- ozone

- sunriseTime

- sunsetTime

- moonPhase: Phase of the moon (0 = New Moon, 1 = Full Moon)

- precipIntensityMax

- uvIndexTime

- temperatureMin

- temperatureMinTime

- temperatureMax

- temperatureMaxTime

- apparentTemperatureMin

- apparentTemperatureMinTime

- apparentTemperatureMax

- apparentTemperatureMaxTime

**II. PROJECT WORKFLOW**



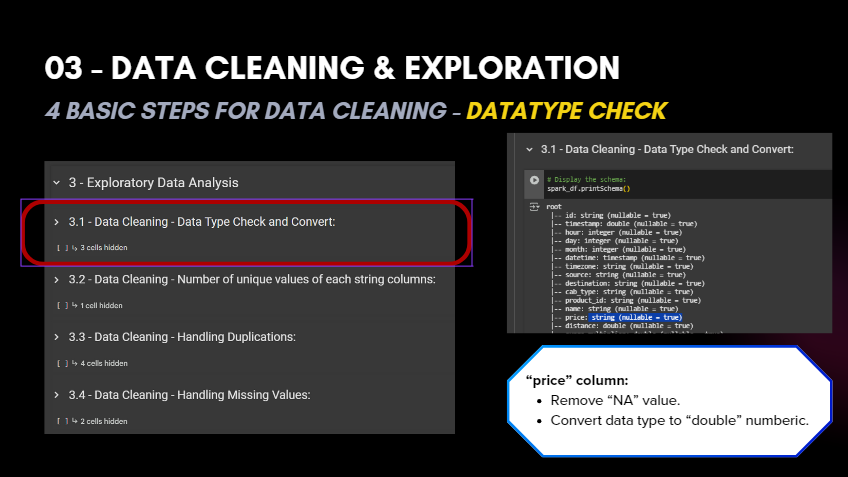
**III. DATA CLEANING & EXPLORATION**

**4 basic steps for data cleaning:**

- **Step 1:** Check data type across columns and convert those not correct.

At this step, after displaying the data types of all columns, it was observed that the "price" column had an unusual data type “**string”** instead of **double (numeric)**.

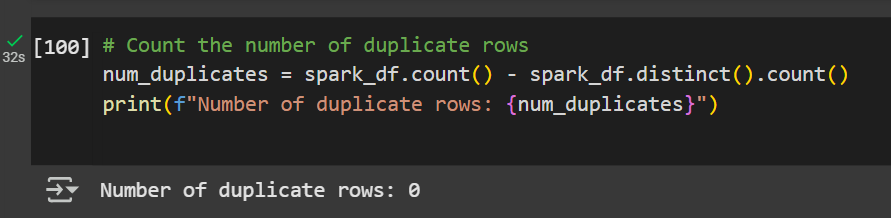
Upon investigating the cause, it was found that the "price" column contained the value **"NA"**.  
**Resolution:** Remove the **"NA"** values and convert the **"price"** column to **double (numeric)**.

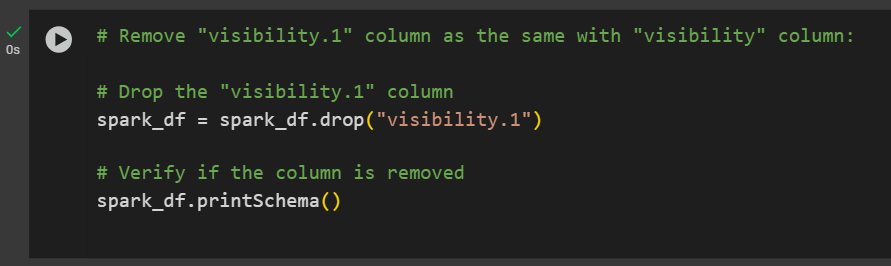


- **Step 2:** Check unique values of all string columns to understand more about the represented meaning of each column and detect “not normal values”.

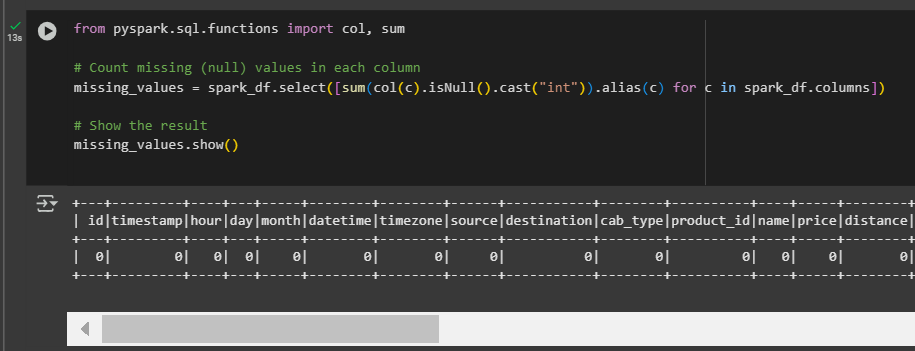


- **Step 3:** Handle duplications: Not found duplicated values - Only remove "visibility.1" column as the same with "visibility".





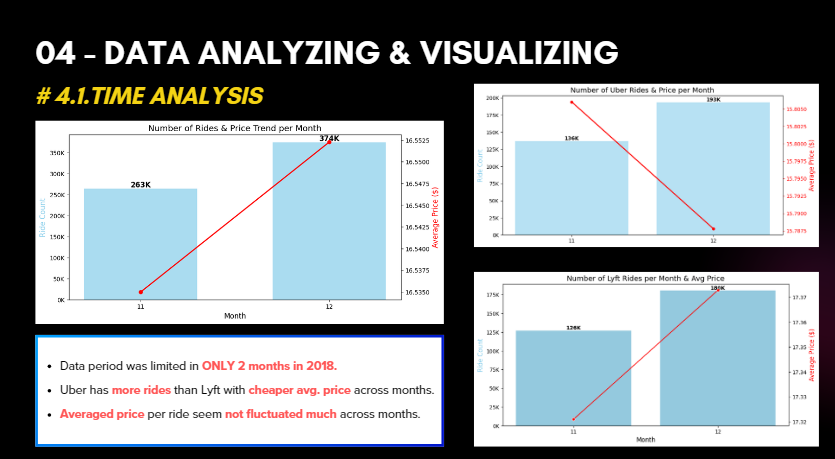
- **Step 4:** Handle missing values - - Not found any missing values as well.



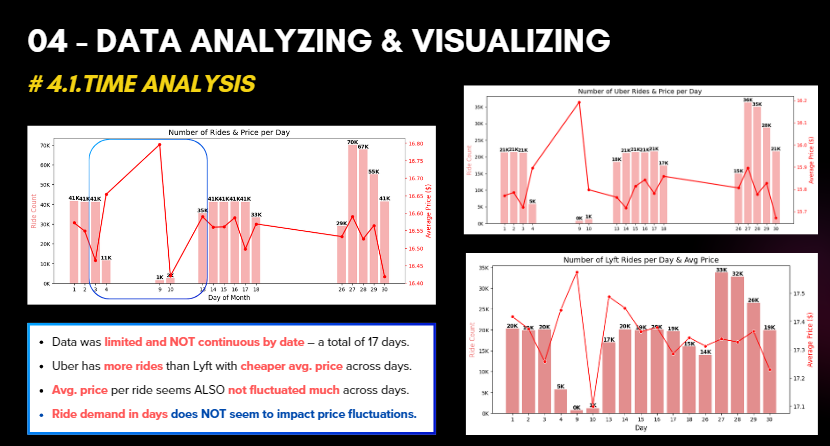
**IV. DATA ANALYZING & VISUALIZING**

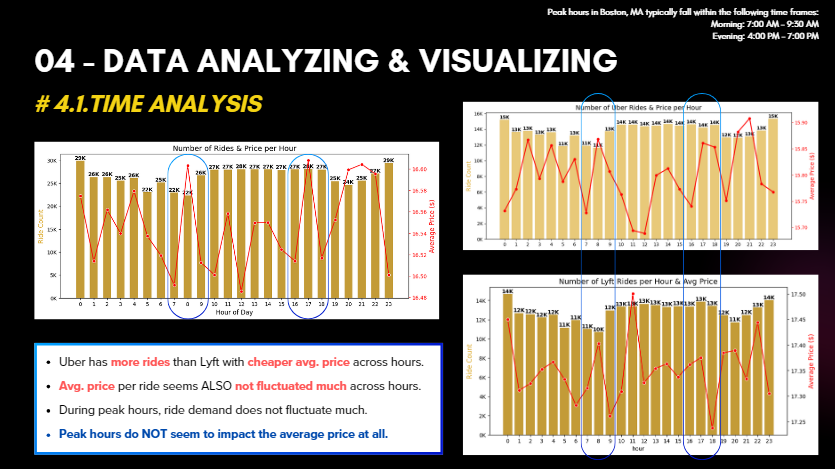
The objective is to develop a comprehensive understanding of the dataset's structure and characteristics, ensuring that users of the project outcomes have a clear and detailed perspective on the data utilized for model training. This section is structured into 5 key components.

***#1 – TIME ANALYSIS***

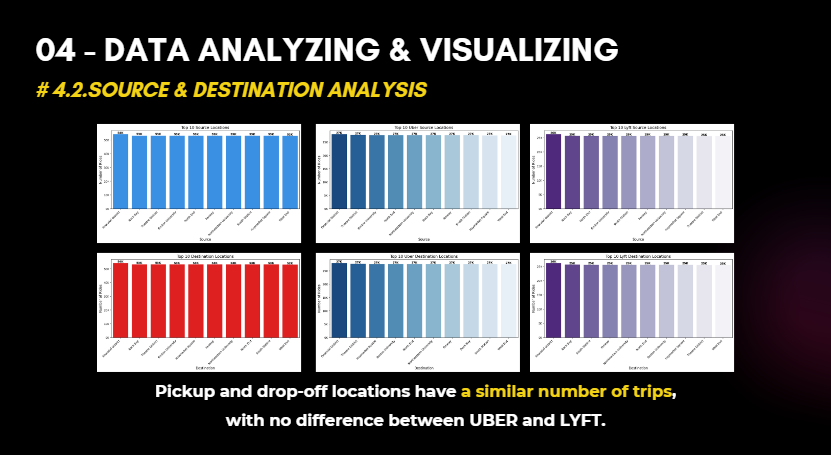
****

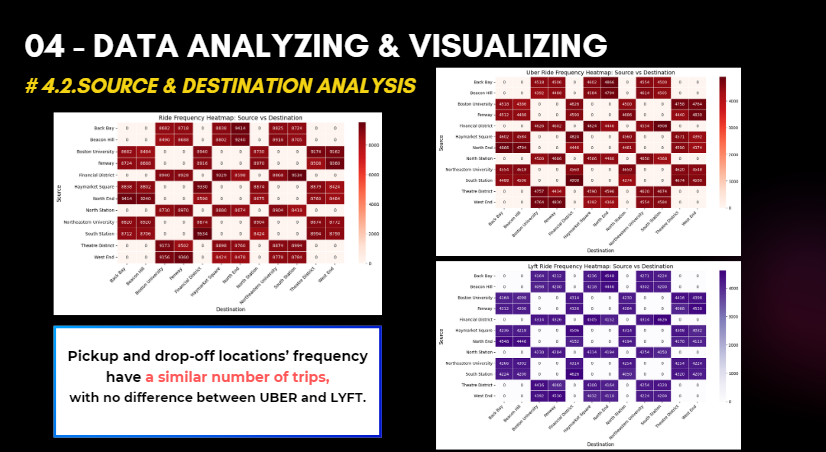
The most significant limitation of this dataset is that it only covers the last 2 months of 2018. As a result, it is not possible to conduct a deeper analysis of trends and changes over time.

****

****

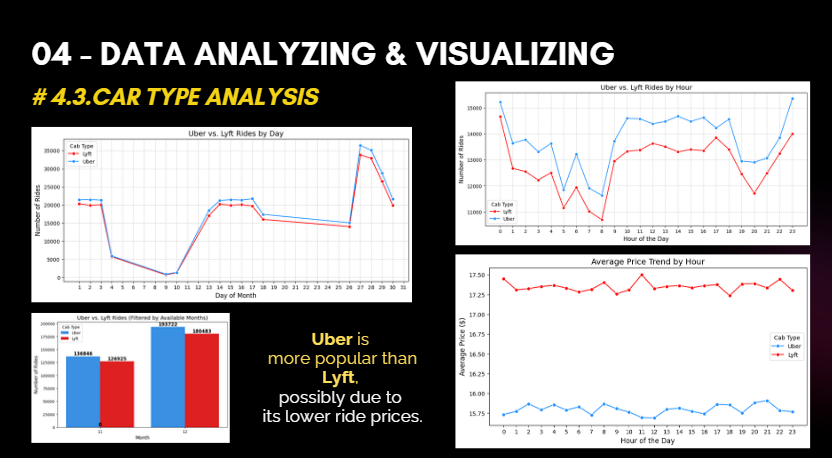
***#2 – SOURCE & DESTINATION ANALYSIS***

******

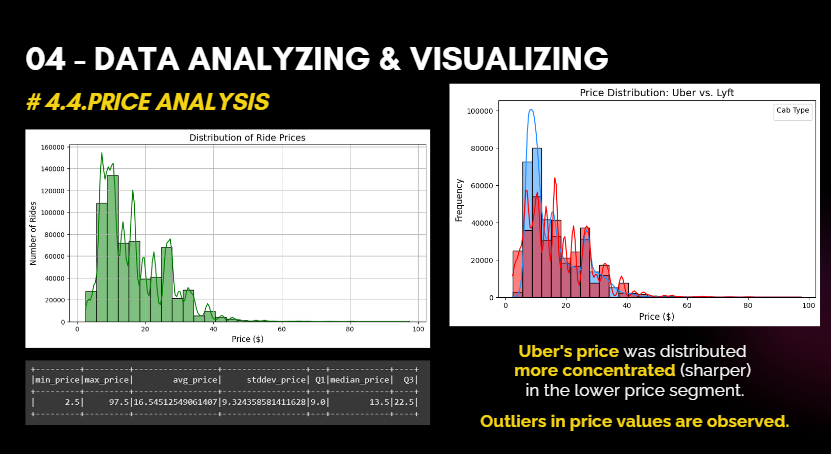
******

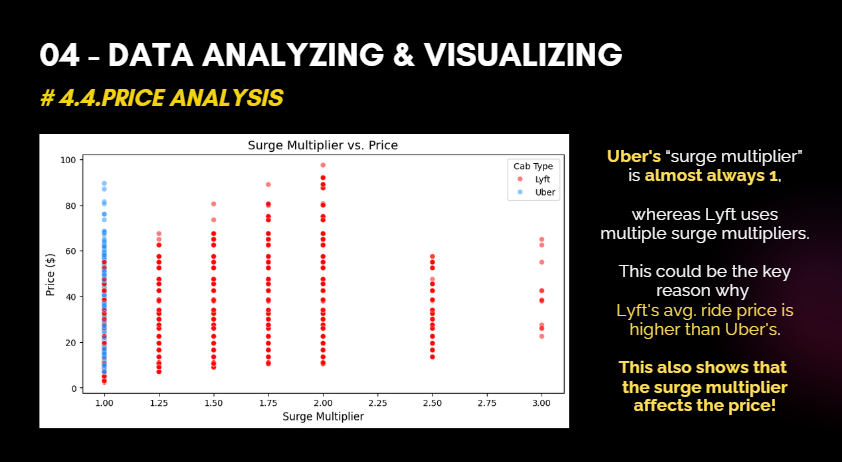
***#3 – CAR TYPE ANALYSIS***

******

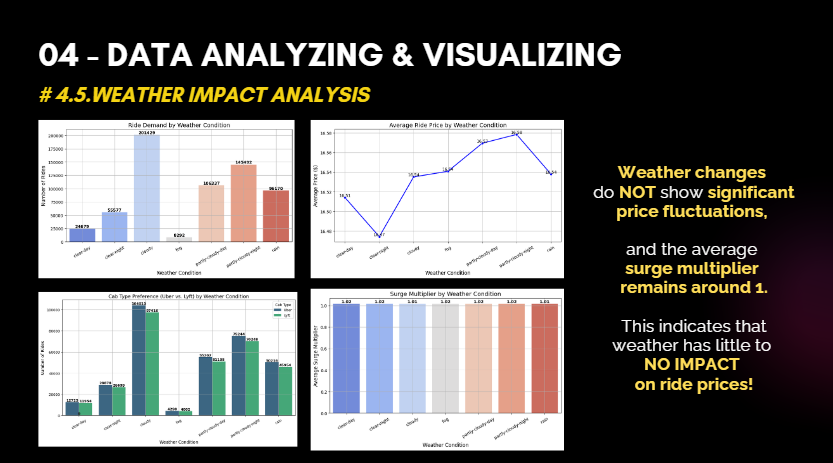
******

***#4 – PRICE ANALYSIS***

******

******

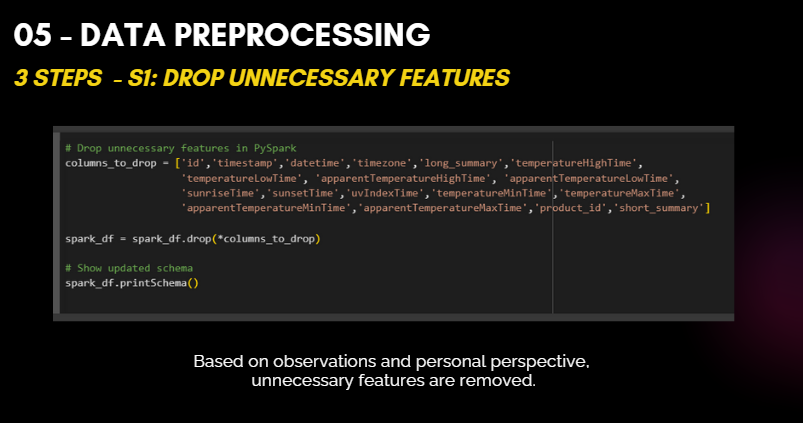
***#5 – WEATHER IMPACT ANALYSIS***



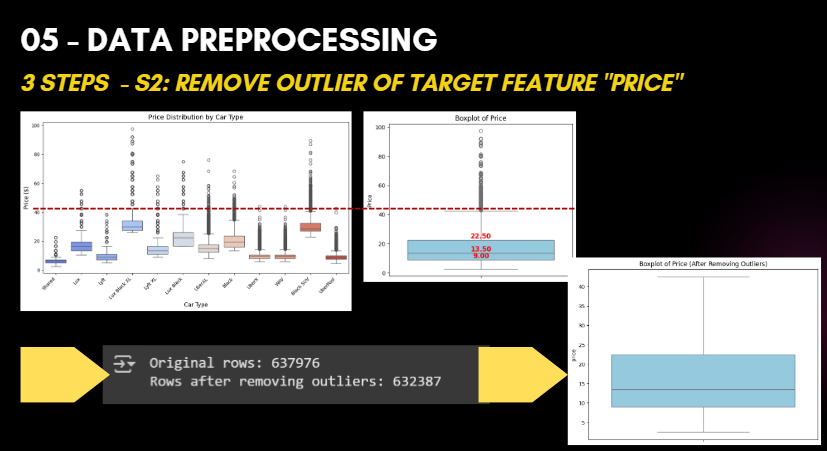
**V. DATA PREPROCESSING**

The objective of this section is to refine the dataset by eliminating unnecessary data fields, preprocessing the target variable "price" before model training, and conducting a correlation analysis to identify variables that have a relationship with "price".

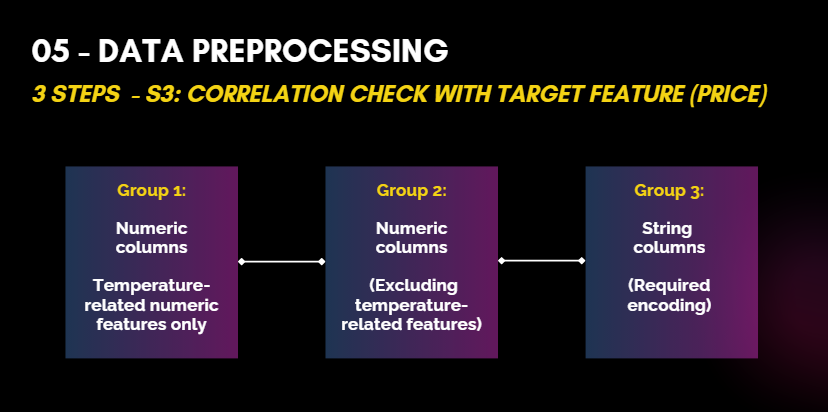
***#1 – Removing Unnecessary Features***



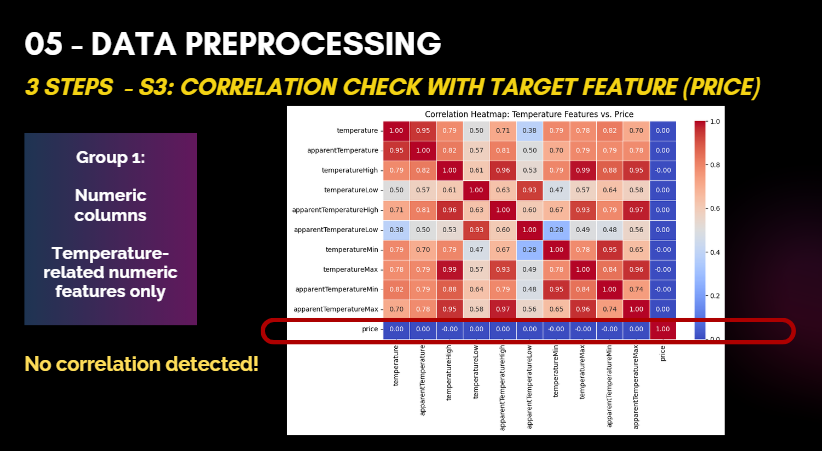
***#2 – Checking target feature "price" and remove outlier***

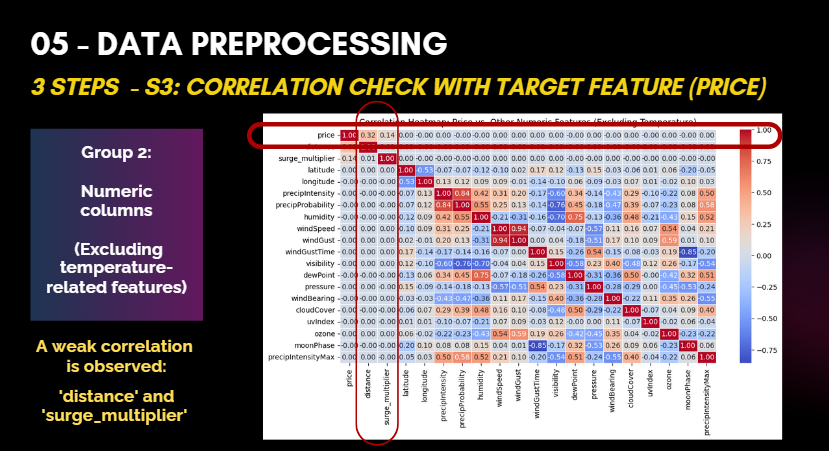


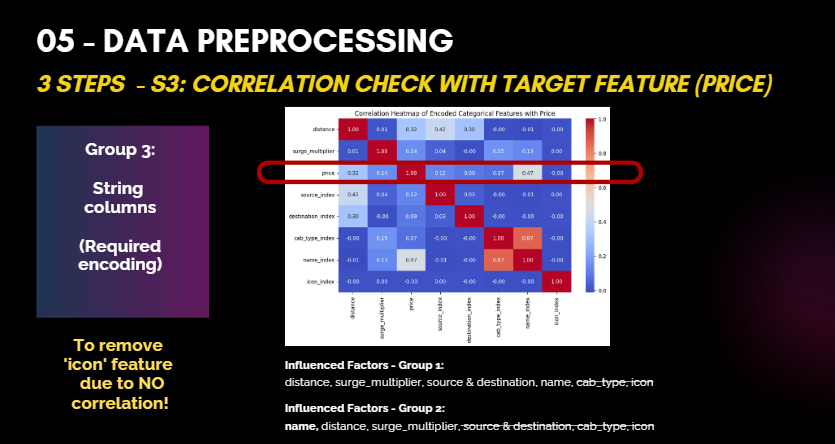
***#3 – Checking correlation between features and target feature (price)***



Due to the large dataset with numerous data fields, it is necessary to group variables when performing correlation analysis with the target variable **"price."**

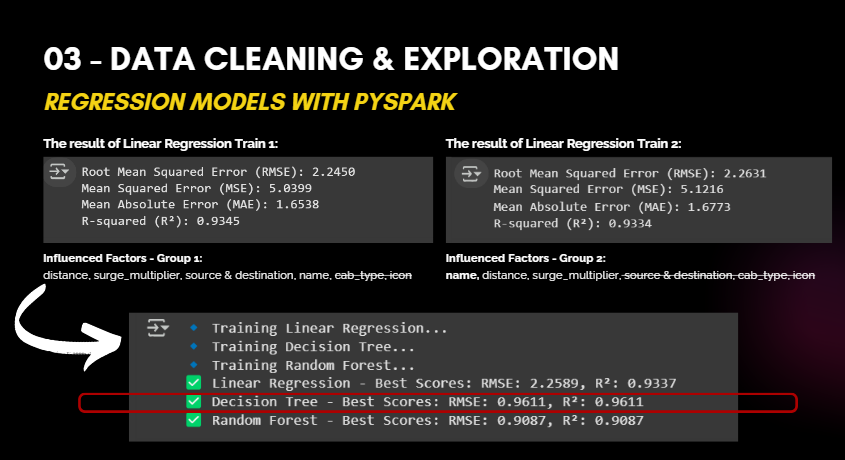






Consider selecting two dependent variable groups to include in the model training, aiming to determine which grouping yields better performance metrics.

**VI. REGRESSION MODEL WITH PYSPARK**



**Linear Regression Model Evaluation – Group 1:**

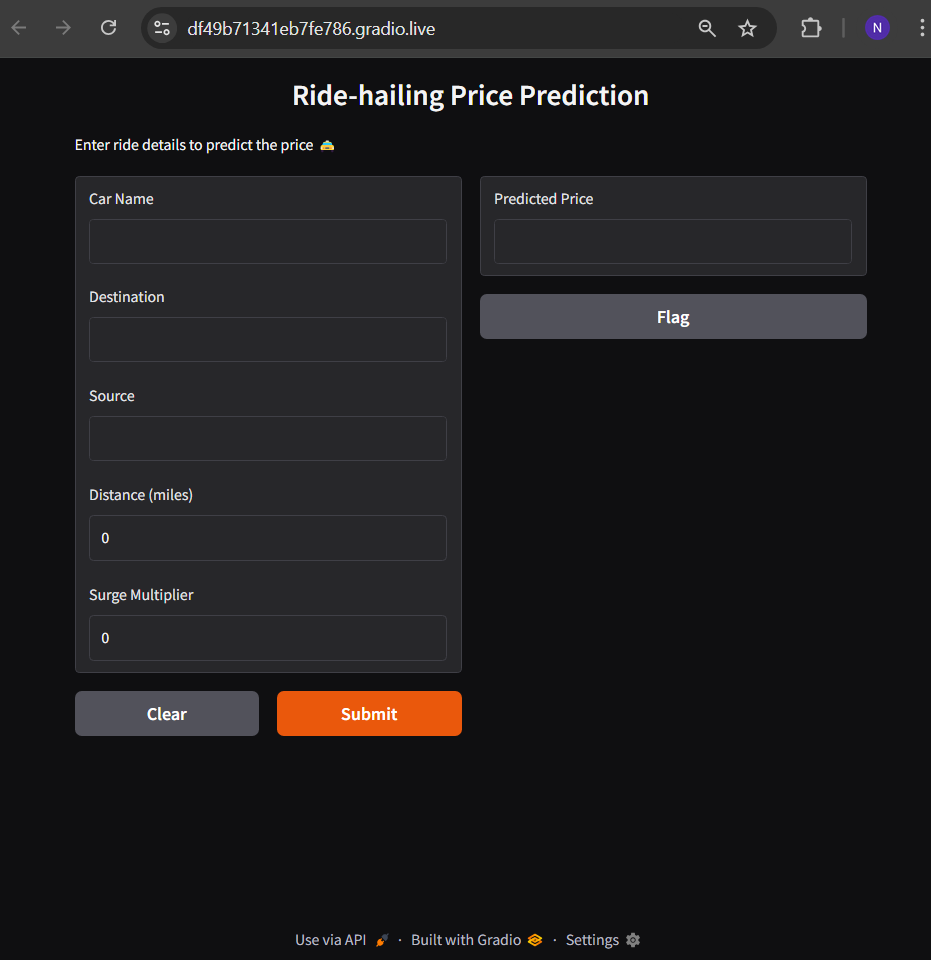
* **RMSE (2.2450):** Average prediction error is **2.25 units**, indicating a relatively low error.
* **MSE (5.0399):** Squared error measure; slightly higher due to sensitivity to outliers.
* **MAE (1.6538):** Average absolute error of **1.65 units**, showing consistent performance.
* **R² (0.9345):** The model explains **93.45%** of variance, indicating a strong fit.

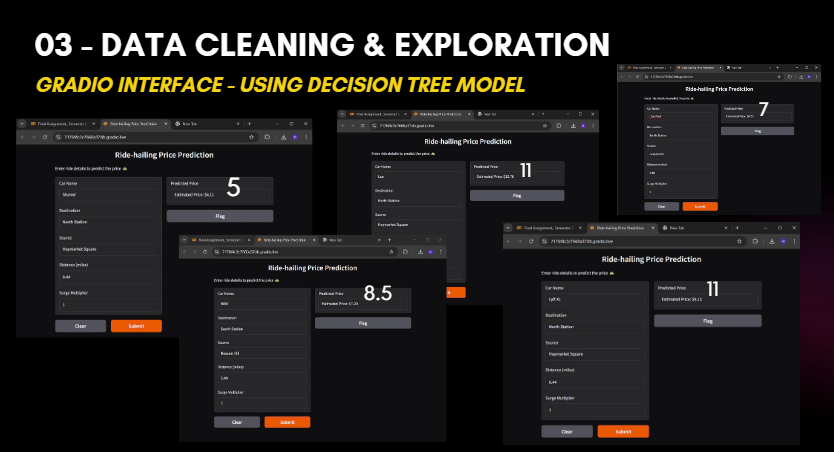
**Conclusion:** The model performs well, but potential outliers and overfitting should be considered. Further improvements could include feature engineering, regularization, or alternative models.

Based on the Linear Regression training results, Group 1 of dependent variables demonstrated better performance metrics.

Therefore, Group 1 was used to train multiple regression models simultaneously to identify the model with the best predictive performance. The results indicated that the **Decision Tree model achieved the highest accuracy with R2 = 96.11%.**

Finally, the Decision Tree model was implemented in a Gradio interface, allowing users to interact with the model and predict ride prices based on influential features.





**VII. INSIGHTS & RESULTS**

* **Peak hours** do **NOT** exhibit significant fluctuations in **ride demand** and have NO notable impact on **ride prices.**
* **Weather changes and temperature variations** do NOT lead to substantial increases or decreases in ride prices.
* **Greater travel distances** are associated with higher total costs.
* 5 factors: pickup and drop-off locations, travel distance, car models’ name, and surge multiplier have demonstrated their influence on ride pricing.
* To **estimate ride prices** based on these influencing factors, the Gradio interface can be utilized (using the best model - Decision Tree Model).

**~~~ END ~~~**