Analyzing Cab Data to Streamline Transportation

Pratima Yadav
A20537717
Computer Science
Illinois Institute Of Technology
pyadav6@hawk.iit.edu

Shreya Padaganur
A20551549
Computer Science
Illinois Institute Of Technology
spadaganur@hawk.iit.edu

Chinmay Chaudhari
A20549646
Computer Science
Illinois Institute Of Technology
cchaudhari1@hawk.iit.edu

Muhammad Rafay Danish
A20494781
Computer Science
Illinois Institute Of Technology
mdanish@hawk.iit.edu

GROUP NO 34

Project Code Github Link:
https://github.com/PratimaYadav22/Data-Preparation-Analysis

CSP 571 - Data Preparation and Analysis
Professor: Oleksandr Narykov
Illinois Institute of Technology

INDEX

➢ABSTRACT

During the past few years, Uber and Lyft have dominated the ride-sharing industry. The secret to Uber and Lyft success is convenience. They provide prompt, affordable transportation services to any area upon request from their clients. The rates for each trip vary depending on the type of service, travel distance, pickup location, etc., making it clear that each firm has a unique pricing strategy. This project focuses on analyzing and comparing the pricing strategies of Uber and Lyft using their ride-sharing data. Our objectives include exploring the dataset to understand correlations between features, their impact on pricing, and any independence assumptions. We will apply various data visualization strategies, including correlation plots and dimensionality reduction techniques like PCA, UMAP, and t-SNE, to uncover patterns and provide actionable insights. Unsupervised learning methods will be used to further explore the data, followed by the identification of an appropriate cross-validation strategy for predictive modeling.

The project will begin with training a simple baseline model, leveraging a validation set for hyperparameter tuning and early stopping. We will evaluate its performance using cross-validation while identifying potential pitfalls. Based on this, we will propose strategies to enhance model performance, such as feature selection, regularization, and model complexity adjustments, and conduct at least two additional experiments to validate these improvements. The goal is to predict ride prices effectively while deepening our understanding of advanced data analysis and visualization techniques.

➢INTRODUCTION

In modern cities, personal transportation has become an essential feature of daily life, providing convenience, speed, and reliability to urban dwellers. Taxicab services have been a dependable mode of transportation for many years, but in recent years, new online ride-sharing services such as Uber, Lyft, and others have emerged and are beginning to gain momentum, threatening to displace traditional taxi services. The market expansion of these ride-sharing services has been quite domineering due to several factors.

One of the primary reasons people are switching to online ride-sharing services is the convenience they offer. Traditional taxis cannot be reserved for a specific time, and you must phone in advance to ensure it arrives on time. In contrast, online taxi services allow you to schedule your pick-up in advance, increasing the driver's dependability as they are closer to the pickup location when you need it. Moreover, online taxi services tend to be less expensive than traditional taxis, with predetermined flat rates that do not increase even if the travel takes longer than anticipated. This makes it easier for passengers to estimate the cost of their ride and budget accordingly. Additionally, the "share my ETA" feature, which allows others who have been granted access by the passenger to view the precise route that is being taken in real-time, enhances safety and peace of mind.

As online ride-sharing services operate primarily online, they generate vast amounts of data daily, and this data can be analyzed to evaluate pricing dynamics and differences between various ride-sharing services. In this project, a basic dataset of Uber and Lyft trips from late 2018 was preprocessed to remove irrelevant data and increase the data's dependability. The preprocessed data was then used to build a model that analyzed the pricing dynamics and differences in special rates between Uber and Lyft, providing insights that could be useful in the transportation industry. Thus, as the ride-sharing services continue to expand and gain momentum, understanding their dynamics and differences can be crucial to their continued growth and success.

➢ PROBLEM STATEMENTS

Some of the problem statements that we are going to answer in this modeling evaluation are as follows-

- Are any features correlated with one another?
- Which features are most relevant?
- Comparison of performance among models.
- The most rides are offered by which company?
- Most common pick-up and drop locations?
- What time of day do most rides happen?
- What are the cheapest and most expensive ride prices?
- How much does each route generally cost?
- How are rides affected by the weather?

➢DATA PROCESSING

- **Data Sources**

  Kaggle Dataset Link

  We will analyze a dataset consisting of Uber and Lyft trips that occurred between November 26, 2018, and December 18, 2018, to gain insight into the factors that influence dynamic pricing and to compare the special prices offered by Uber and Lyft. The dataset used in this project has been sourced from Kaggle and consists of 693,071 rows and 57 columns.

```
# Display the first five rows
print(rideshare_data.head())

# Get a concise summary of the DataFrame
print("info")
print(rideshare_data.info())

# Get descriptive statistics for numerical columns
print("describe")
print(rideshare_data.describe())
```

✓ 0.9s

```
info
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 693071 entries, 0 to 693070
Data columns (total 57 columns):
```

- **Data Cleaning and Data Pre-processing**

  We have undertaken the following data pre-processing steps in order to get reliable results and run the model.

  - The dataset contains numerous missing values and column label inconsistencies across various files.
  - The source and destination names have a high number of missing values, which could affect the accuracy of the analysis.

● Different files have varying date-time formats, making it challenging to merge and analyze the data.

● The same features have different data type values in different files, potentially leading to discrepancies when integrating and analyzing the data.

● Changes made:

  ○ Used the start time and end time of the trip to calculate trip duration and added it wherever it was missing.

  ○ Formatted date time and other feature values in all files to have a single standard format.

  ○ Added additional columns with Day of the week from the date field

  ○ Renamed columns among all files to common labels

  ○ Dropped features that were not present among all files(SunriseTime, SunsetTime, MoonPhase etc)

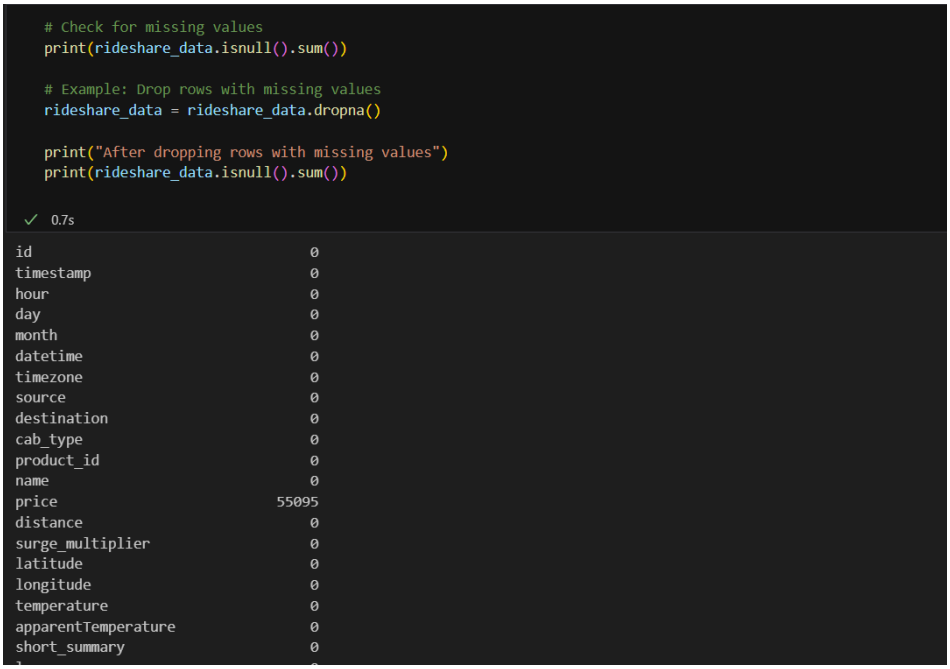
  ○ Used values of latitude longitude to populate missing source and destination values among the files.

● Weather Data:

  ○ The date format is different from the datasets.

● Changes made:

  ○ Date format has been changed to reflect the same format as the dataset.



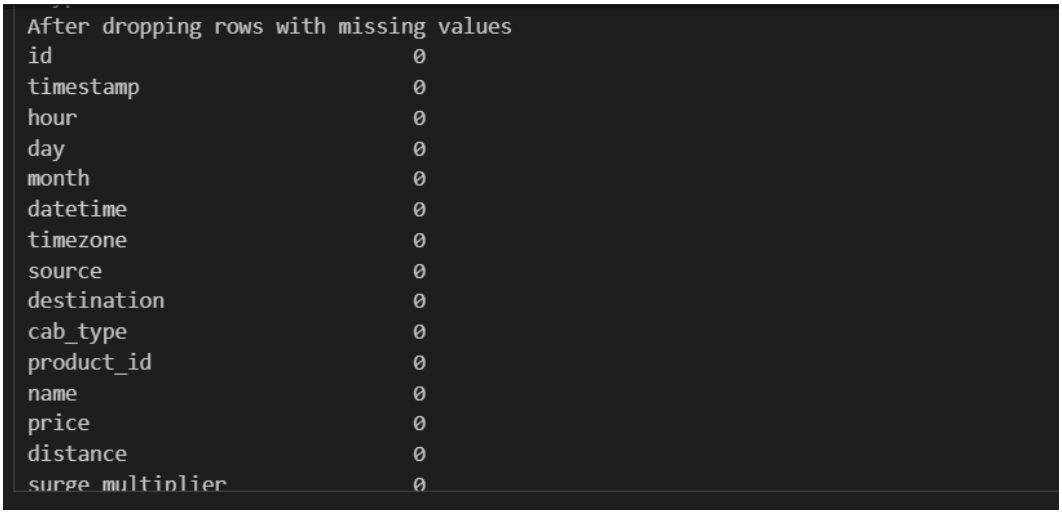
```
# Check for missing values
print(rideshare_data.isnull().sum())

# Example: Drop rows with missing values
rideshare_data = rideshare_data.dropna()

print("After dropping rows with missing values")
print(rideshare_data.isnull().sum())
```

✓ 0.7s

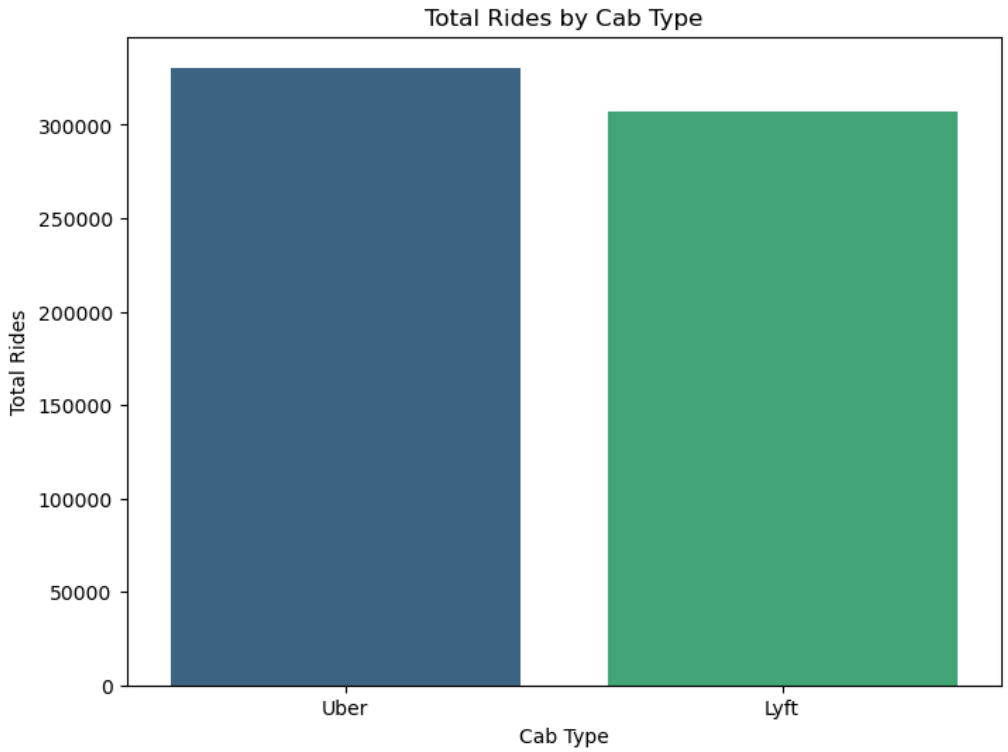
```
id                       0
timestamp                0
hour                     0
day                      0
month                    0
datetime                 0
timezone                 0
source                   0
destination              0
cab_type                 0
product_id               0
name                     0
price                55095
distance                 0
surge_multiplier         0
latitude                 0
longitude                0
temperature              0
apparentTemperature      0
short_summary            0
```



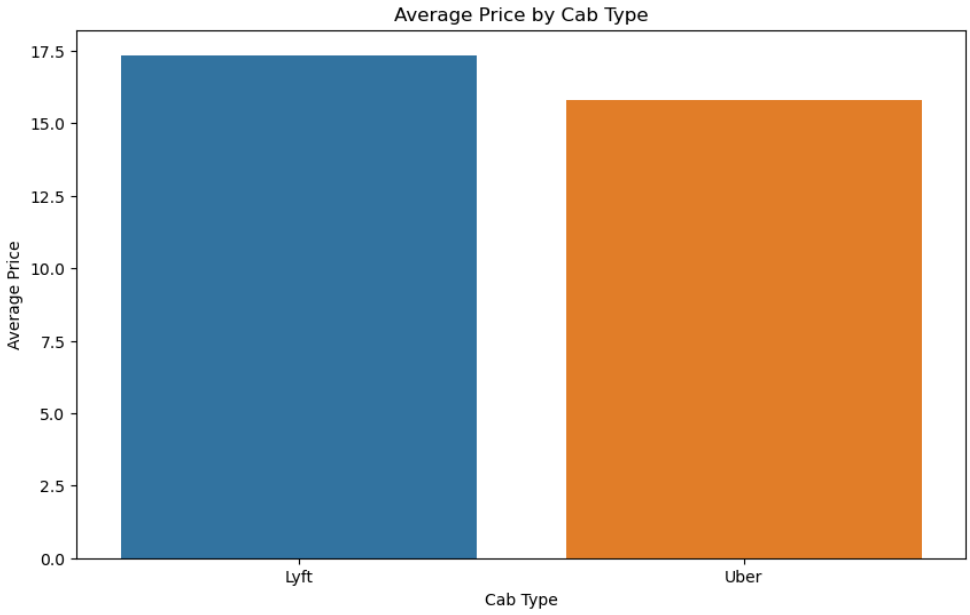
- **Observations:**
  - Collecting and manipulating data is a time-consuming and effort-intensive process, particularly when identifying discrepancies in file formats, column names, and missing data. In this project, a significant amount of time has been dedicated to this task.
  - While working with the current dataset, we did not experience any issues with the laptops we were using. However, as the dataset size grows, it becomes crucial to optimize the code for data manipulation. To achieve this, we rewrote the script using pandas and treated the columns as vectors.

➢EXPLORATORY DATA ANALYTICS

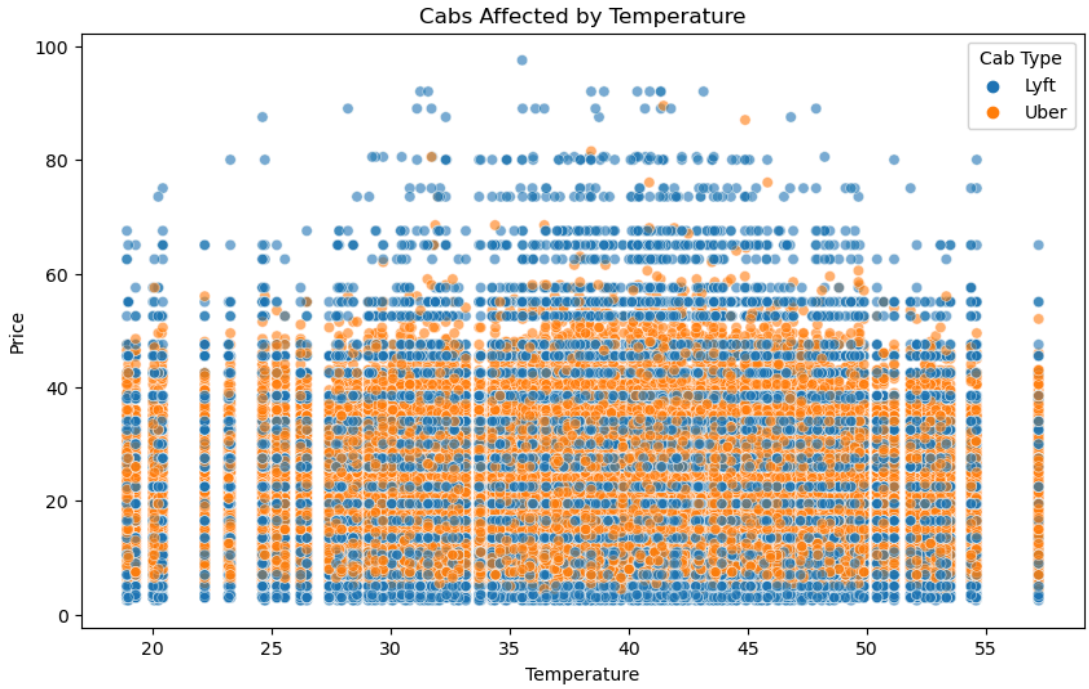
Total rides by cab type: Who offers the most rides, Uber or lyft?



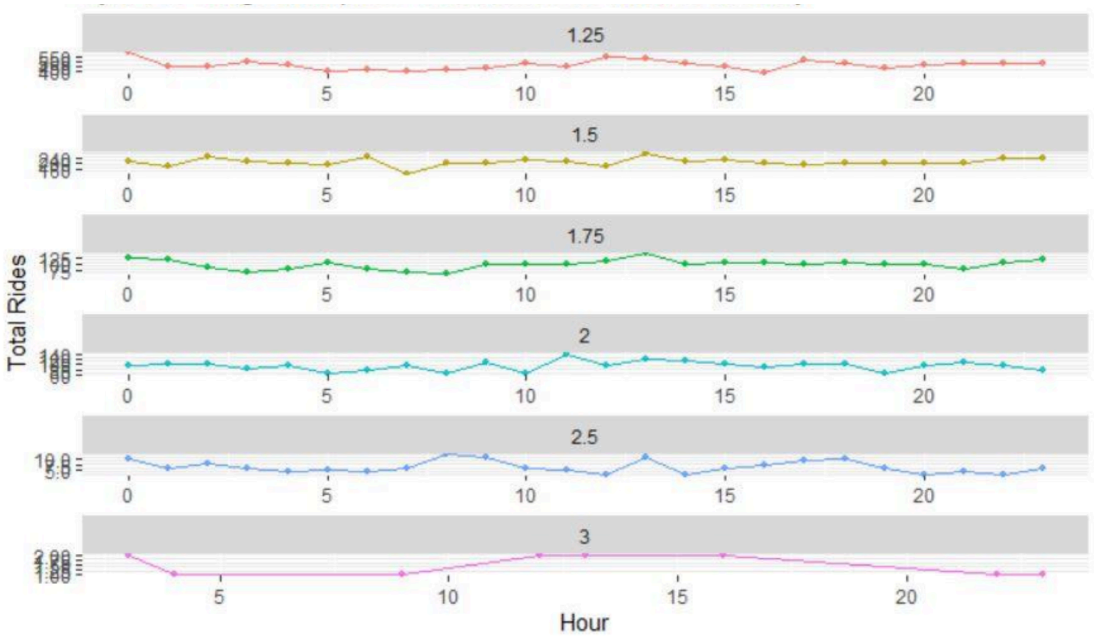
Average price by cab type:



Cabs affected by temperature:



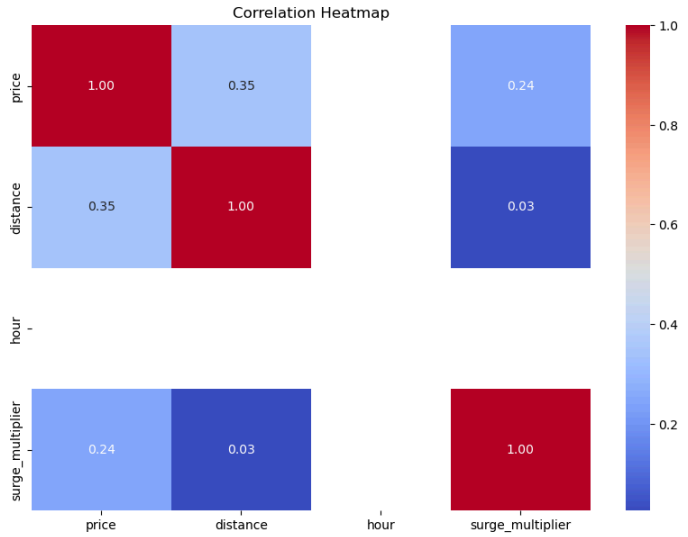
Lyft: total rides and average surge multiplier by hour



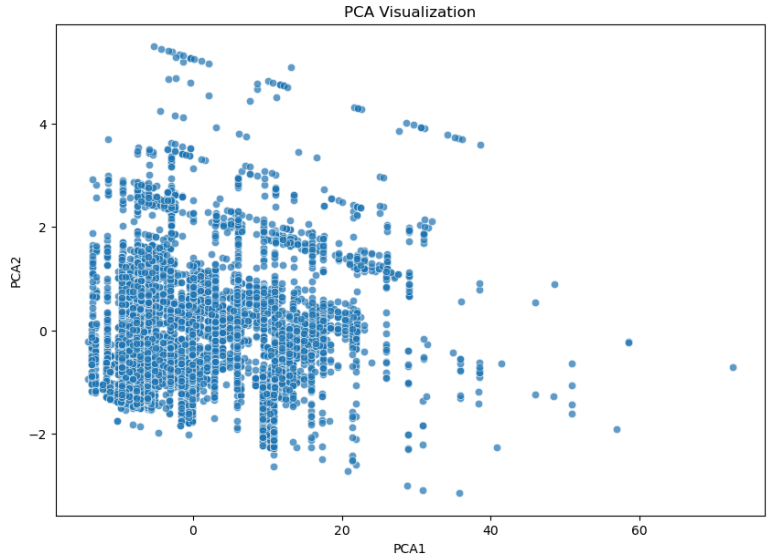
➢DATA VISUALIZATION:

Data visualization is an essential part of the data analysis pipeline, offering ways to uncover hidden structures, relationships, and patterns within the data.Dimensionality reduction techniques like **PCA**, **UMAP**, and **t-SNE** help uncover patterns and clusters in high-dimensional data.

- **Correlation Matrix & Heatmaps:**
  Identifies relationships between features, like how distance and surge multiplier affect price. Heatmaps visualize feature correlations, highlighting dependencies that guide feature selection.

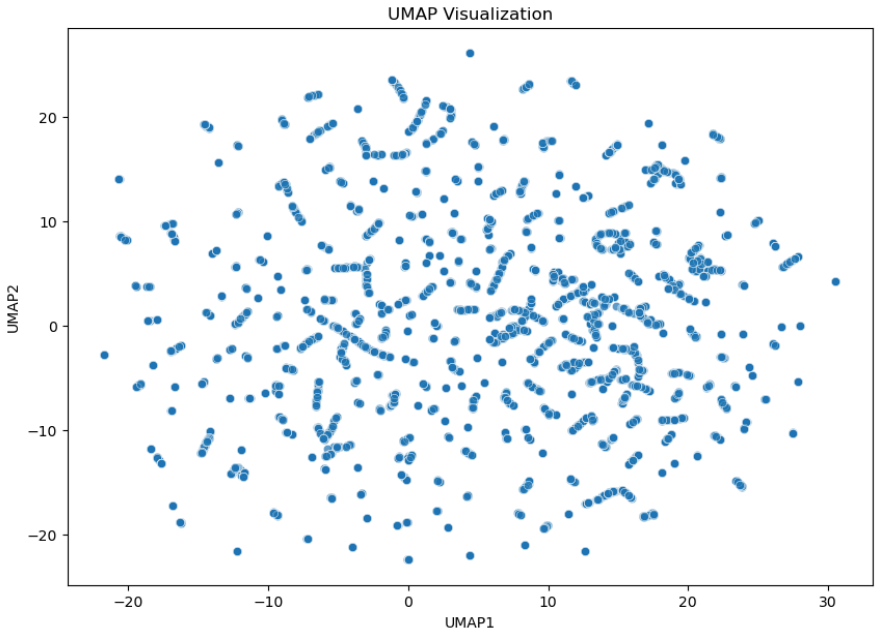


- **PCA Visualisation:**
  Reduces dimensions by finding principal components that capture the most variance in the data. Reveals major trends and relationships in the data, such as the distinction between ride types or price categories.



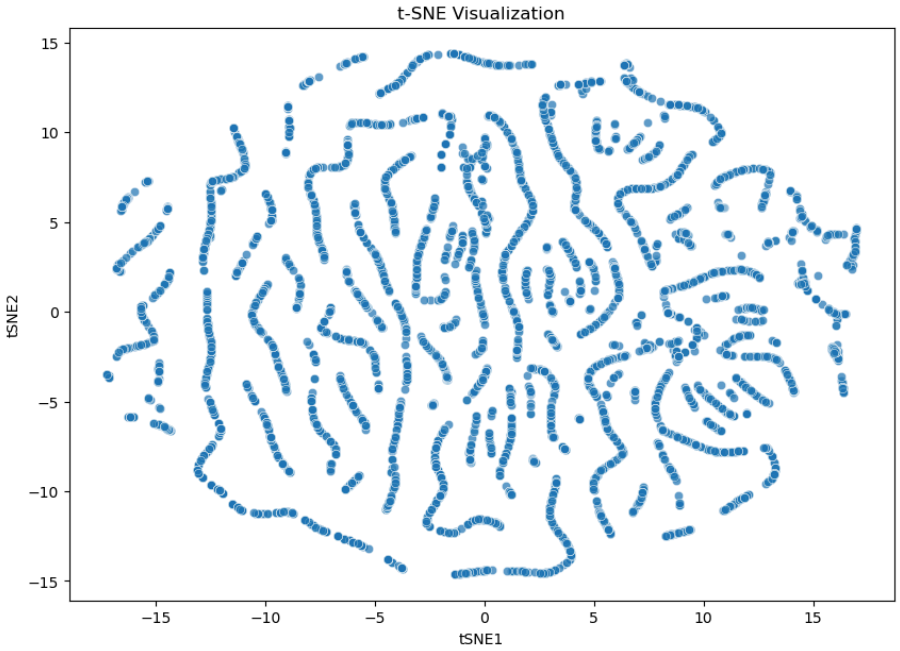
- **UMAP Visualization:**

  A non-linear method preserving both global and local data structures. Effectively identifies clusters based on complex relationships like ride preferences, time, and location.



- **t-SNE Visualization:**

  Focuses on preserving local similarities and helps identify fine-grained clusters. Highlights subtle groupings that may not be captured by linear methods like PCA.



➢UNSUPERVISED LEARNING:

Unsupervised learning techniques, such as clustering and dimensionality reduction, are used to identify patterns and structure in data without predefined labels. In this analysis, we applied the following methods:

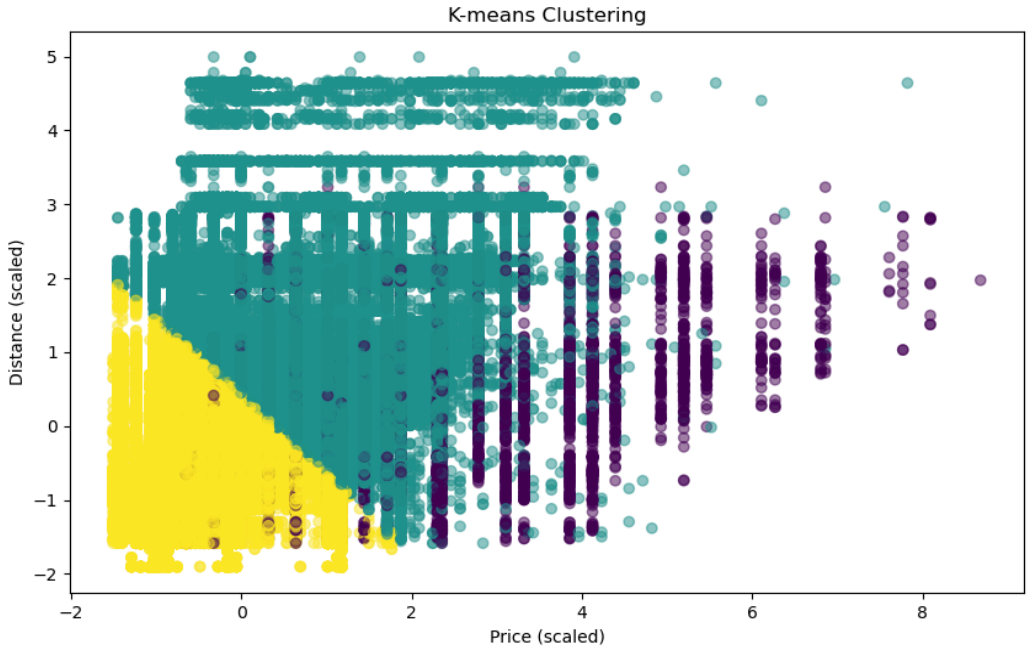
K-Means Clustering:This algorithm partitioned the data into three clusters. The cluster sizes were:Cluster 2: 416,266 samples, Cluster 1: 211,820 samples, Cluster 0: 9,890 samples

Gaussian Mixture Model (GMM): A probabilistic model that also identified three clusters with the following sizes: Cluster 1: 403,688 samples, Cluster 0: 213,307 samples, Cluster 2: 20,981 samples

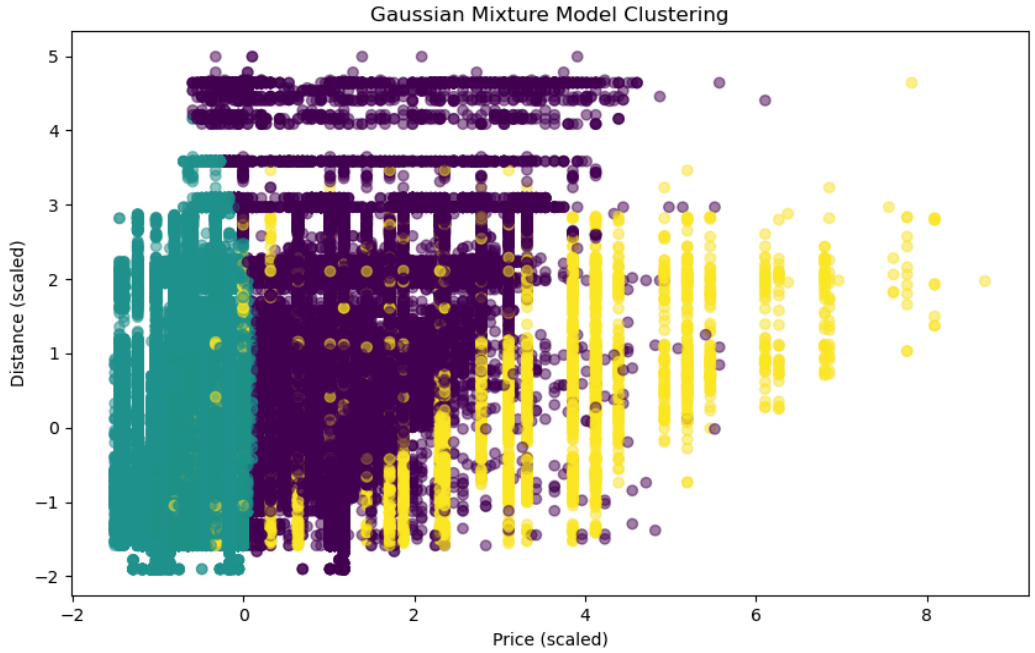
PCA (Principal Component Analysis): PCA was used for dimensionality reduction, revealing that the first two components explained 47.77% and 32.52% of the variance, respectively.

All analysis completed! Check the saved CSV files and plots.
1. rideshare_data_with_all_analysis.csv - Contains all original data with cluster assignments
2. kmeans_cluster_stats.csv - Detailed statistics for K-means clusters
3. gmm_cluster_stats.csv - Detailed statistics for GMM clusters
4. Three visualization plots saved: kmeans_plot.png, gmm_plot.png, pca_plot.png

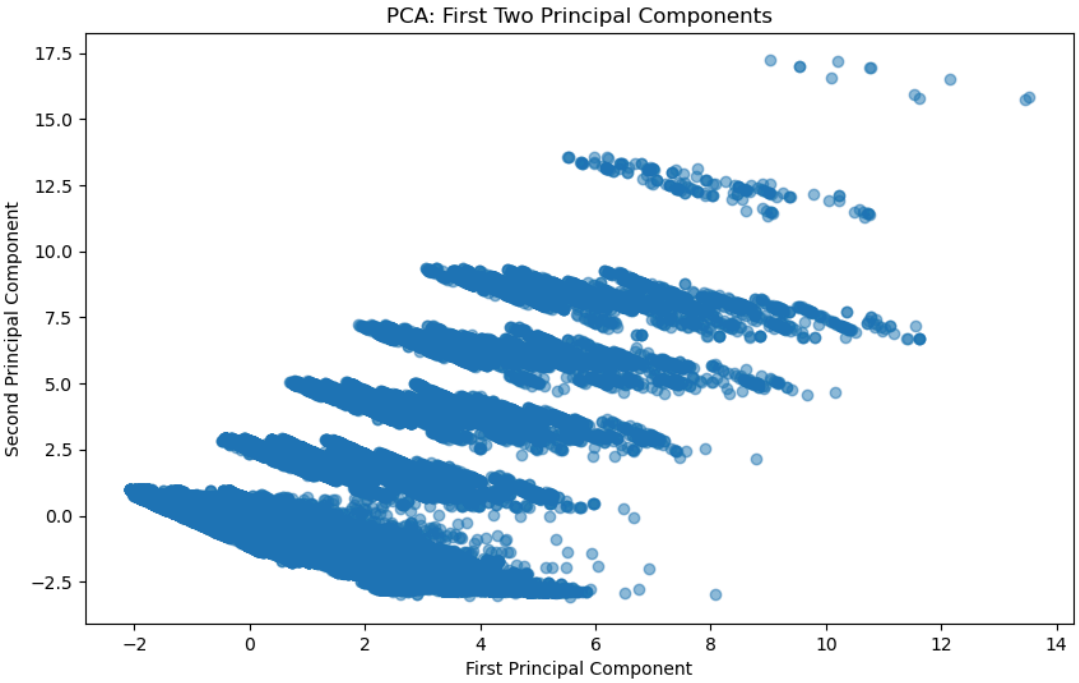
kmeans_plot



gmm_plot:



pca_plot:



➢CROSS-VALIDATION STRATEGY

Cross-validation is a statistical method used to evaluate and improve the performance of machine learning models. It ensures that the model generalizes well to unseen data by splitting the dataset into multiple subsets (folds) and testing the model on these different subsets.

To ensure a balanced evaluation of model performance, **Stratified K-Fold Cross-Validation** was employed. This approach splits the dataset into k folds while maintaining an equal distribution of class labels in each fold, addressing potential issues with class imbalance. By iteratively training and validating the model on these folds, this method ensures robust performance evaluation and prevents overfitting to specific subsets of data.

- **Model Training**

Three models were implemented to compare predictive performance and identify the best approach:

1. **Logistic Regression**:
    - Applied with default parameters, including class weighting, to accommodate imbalanced datasets.
    - A baseline model providing interpretable insights into feature importance.
2. **Decision Tree**:
    - Configured with a maximum depth of 5 to avoid overfitting and maintain model simplicity.
    - Captures non-linear relationships in the data while remaining computationally efficient.
3. **Random Forest**:
    - Constructed with 50 trees and a maximum depth of 5.
    - Utilizes ensemble learning to enhance stability and accuracy, reducing variance from individual trees.

Metrics used to evaluate models include **Accuracy, Precision, Recall, and F1-Score**, ensuring a comprehensive assessment of performance across various aspects of classification quality.

- **Recommendations for Improvement**
    - **Address Class Imbalance**:
        - Implement **SMOTE (Synthetic Minority Oversampling Technique)** to generate synthetic examples for minority classes, balancing the dataset.

    - Adjust class weights in models to better reflect class proportions and improve classification accuracy.
- **Feature Engineering**:
    - Incorporate interaction terms, such as **price × distance**, to capture complex relationships between variables.
    - Remove less informative features (e.g., hour and day), reducing noise and improving model interpretability.
- **Test Additional Models**:
    - Evaluate advanced models such as **Gradient Boosting** (e.g., XGBoost, LightGBM) for improved predictive power.
    - Compare with simpler models, such as **Naive Bayes**, to assess the trade-off between simplicity and performance.
- **Hyperparameter Tuning**:
    - Fine-tune Random Forest parameters, including the minimum number of samples per split and leaf size, to reduce overfitting.
    - Experiment with different configurations to achieve optimal model complexity and generalization.

➢MODEL EVALUATION

To evaluate the performance of different strategies in predicting trip outcomes, we employ various assessment criteria. The following statistics are used to measure their prediction accuracy–

```
[Running] python -u "c:\Users\18477\Desktop\Apps\CS\csp571\project\uvldataset\p4p5p6.py"

=== Model Performance ===

Model: Simple Logistic Regression
Accuracy: 0.74
Classification Report:
              precision    recall  f1-score        support
0              0.985733  0.755602  0.855461  134420.000000
1              0.058179  0.280634  0.096378    4162.000000
2              0.000968  0.454545  0.001931      33.000000
accuracy       0.741269  0.741269  0.741269       0.741269
macro avg      0.348293  0.496927  0.317923  138615.000000
weighted avg   0.957649  0.741269  0.832466  138615.000000

Model: Simple Decision Tree
Accuracy: 0.61
Classification Report:
              precision    recall  f1-score        support
0              0.991825  0.620071  0.763079  134420.000000
1              0.053628  0.364969  0.093514    4162.000000
2              0.000838  0.666667  0.001674      33.000000
accuracy       0.612423  0.612423  0.612423       0.612423
macro avg      0.348764  0.550569  0.286089  138615.000000
weighted avg   0.963419  0.612423  0.742794  138615.000000

Model: Simple Random Forest
Accuracy: 0.66
Classification Report:
              precision    recall  f1-score        support
0              0.989331  0.668435  0.797825  134420.000000
1              0.066560  0.471168  0.116643    4162.000000
2              0.000927  0.515152  0.001851      33.000000
accuracy       0.662475  0.662475  0.662475       0.662475
macro avg      0.352273  0.551585  0.305440  138615.000000
weighted avg   0.961388  0.662475  0.777182  138615.000000

[Done] exited with code=0 in 8.276 seconds
```

The overall accurate model is Simple Logistic Regression with 74% accuracy. Based on different metrics, we can conclude that the Simple Logistic Regression is the best model and thus we could consider this as our final model

➢OBSERVATIONS & CONCLUSION

Observations:

- Exploratory Data Analysis: Key patterns and correlations were identified through EDA, providing a deeper understanding of the dataset.
- Dimensionality Reduction: Techniques like PCA and t-SNE uncovered clusters and simplified the data, highlighting dominant trends.
- Unsupervised Learning: Clustering methods such as K-Means effectively grouped data, showcasing hidden structures.
- Model Training: Basic models trained with cross-validation ensured robust and reliable performance.
- Feature Engineering: Preprocessing methods, including normalization and binning, improved data usability and model accuracy.

Conclusion:

The project effectively demonstrated an end-to-end workflow, combining EDA, advanced analytics, and machine learning to extract actionable insights. Iterative experiments and robust validation enhanced model performance, establishing a strong foundation for real-world applications.

➢PLANNED FUTURE WORK

There are multiple ways to tackle this problem and different perspectives to consider. One approach is to enhance the models to improve the accuracy of the forecasts. For example, we can explore the correlations between variables such as the predictors for distance and types of cabs. We can also analyze the Random Forests prediction errors by optimizing the parameter values. Additionally, we plan to include external data such as traffic conditions and time to our analysis. We aim to investigate the reasons behind the high variability of fares for a given source and destination. With the insights gained from this project, we aim to develop an enhanced model with more useful features and experiment with new data formats such as Avro or Parquet to improve the efficiency of row/column operations.

➢REFERENCES

● Shashank H, "Data Analysis of Uber and Lyft Cab Services," International Journal of Interdisciplinary Innovative Research & Development, 2022,ISSN:2456-236XVol.05 Issue 01 | 2020 http://ijiird.com/wp-content/uploads/050144.pdf

● Mrinalini Sunder, (2021, July 27). “Uber and Lyft Cab Prices Data Analysis and Visualization”Retrievedfrom https://medium.com/mindtrades-consulting/uber-and-lyft-cab-prices-data-analysis-and-visualization-93aca4596f20.

● ACM Digital Library. (n.d.). ACM Digital Library. Retrieved December 4, 2022, from https://dl.acm.org/doi/fullHtml/10.1145/3178876.3186134

● Leo, M. S. (2021, January 24). New York Taxi data set analysis. Medium. RetrievedDecember4,2022,from https://towardsdatascience.com/new-york-taxi-data-set-analysis-7f3a9ad84850

● Liu, L., Qiu, Z., Li, G., Wang, Q., Ouyang, W., & Lin, L. (2019). Contextualized Spatial-T emporal Network for T axi Origin-Destination Demand Prediction.

.