CSP 571 Data Preparation and Analysis

Project Report on

Analyzing Divvy Bike Usage Patterns in Chicago

Date: 08/02/2024

Mohammad Hamza Piracha
mpiracha@hawk.iit.edu
A20554741

Poshan Pandey
Ppandey6@hawk.iit.edu
A20519852

Usman Matheen Hameed
uhameed@hawk.iit.edu
A20564338

Abstract

This project report presents an in-depth analysis of Divvy bike usage patterns in Chicago for the year 2023. The primary objective is to extract meaningful insights into ridership trends and behaviors, employing various data preparation, cleaning, transformation, and analysis techniques. The study leverages machine learning algorithms for outlier detection and clustering, along with visualization tools to uncover patterns. Additionally, the project explores feature selection and model training using Random Forest and XGBoost algorithms to predict trip duration accurately. The findings aim to enhance the operational efficiency and user satisfaction of the Divvy bike-sharing system.

1. Introduction

1.1 Objective

The objective of this project is to analyze the usage patterns of Divvy bikes in Chicago for the year 2023. By performing thorough data preparation, cleaning, transformation, and analysis, the study aims to uncover significant insights into ridership behaviors and trends. The project also involves the application of machine learning algorithms for outlier detection and clustering to understand data structures better. Additionally, feature selection and model training using Random Forest and XGBoost algorithms are conducted to accurately predict trip duration. The goal is to enhance the efficiency and user experience of the Divvy bike-sharing system through data-driven recommendations.

1.2 Literature Review

The literature review encompasses a survey of existing studies and articles on the analysis of bike-sharing systems, specifically focusing on Divvy bikes in Chicago. Key references include exploratory data analysis (EDA) of Divvy bike datasets, research on variations in Divvy bike station usage volumes, and studies on historical trip records. These sources provide foundational insights into bike-sharing trends, seasonal variations, and user demographics. The review also covers methodologies for data cleaning, outlier detection, clustering, and predictive modeling, highlighting best practices and innovative approaches from previous research. By synthesizing these findings, the literature review sets the stage for the project's methodological framework and analysis.

2. Methodology

2.1 Data Preparation

Our dataset covers ride data for the year 2023, with individual datasets downloaded monthly. These datasets were combined into a single file (`merged_data.csv`) to facilitate thorough analysis. The combined dataset was then prepared for further cleaning and analysis to extract insights into ridership patterns and trends.

2.2 Data Cleaning and Transformation

2.2.1 Data Cleaning

The dataset for the year 2023 contains comprehensive ride data, encompassing details from each month. To ensure the dataset's quality and relevance, several key actions were undertaken during the cleaning process:

- Unnecessary columns were removed, focusing on essential variables such as ride ID, rideable type, timestamps, station locations, and rider type (member or casual).
- Missing values were systematically addressed through the removal of incomplete data points.
- Timestamps for ride start and end times were standardized into datetime format to facilitate accurate time-based analysis.
- Essential features like trip duration, day of the week, and hour of the day were derived to enhance the dataset's analytical depth.
- Outliers, including negative durations and trips exceeding 24 hours, were filtered out to maintain dataset integrity.
- Checks for data consistency ensured that each ride's end time logically followed its start time.

These actions collectively ensure that the cleaned dataset is robust and ready for in-depth analysis of bike-sharing behaviors and trends in 2023.

2.2.2 Outlier Detection

Apart from general analysis and missing value reduction, Outlier detection and removal is also performed using these following approaches:

2.2.2.1 Z-Score Method

The Z-score method standardizes the dataset and identifies outliers based on a threshold, usually 3 standard deviations from the mean.

2.2.2.2 Isolation Forest

Isolation Forest is a machine learning algorithm that isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

2.2.3 Data Transformation

Following approaches were employed for data transformation:

2.2.3.1 Normalization and Standardization

Normalization: Rescales the data to a range of [0, 1]. **Standardization**: Centers the data to have a mean of 0 and standard deviation of 1.

2.2.3.2 Log Transformation

Log transformation can help stabilize variance and make the data more normally distributed.

2.2.3.3 Date and Time Features

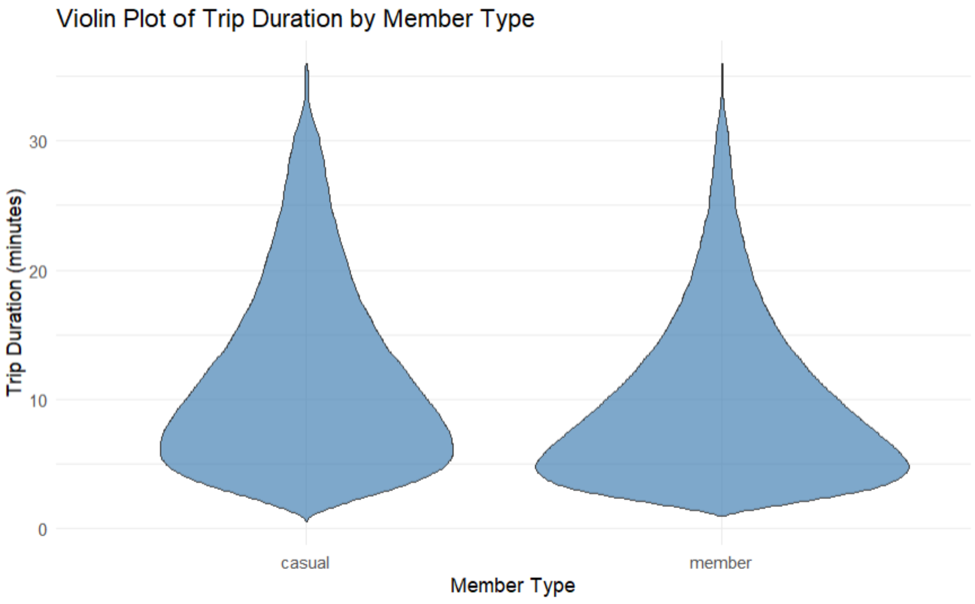
Extracting useful features from datetime columns, such as the day of the week, month, hour, etc.

2.2.4 Distribution Analysis

Distribution analysis refers to the process of examining the statistical distribution of data within a dataset. Extensive usage of histograms and graphs for distribution analysis.

Analysis Used in Distribution Analysis:

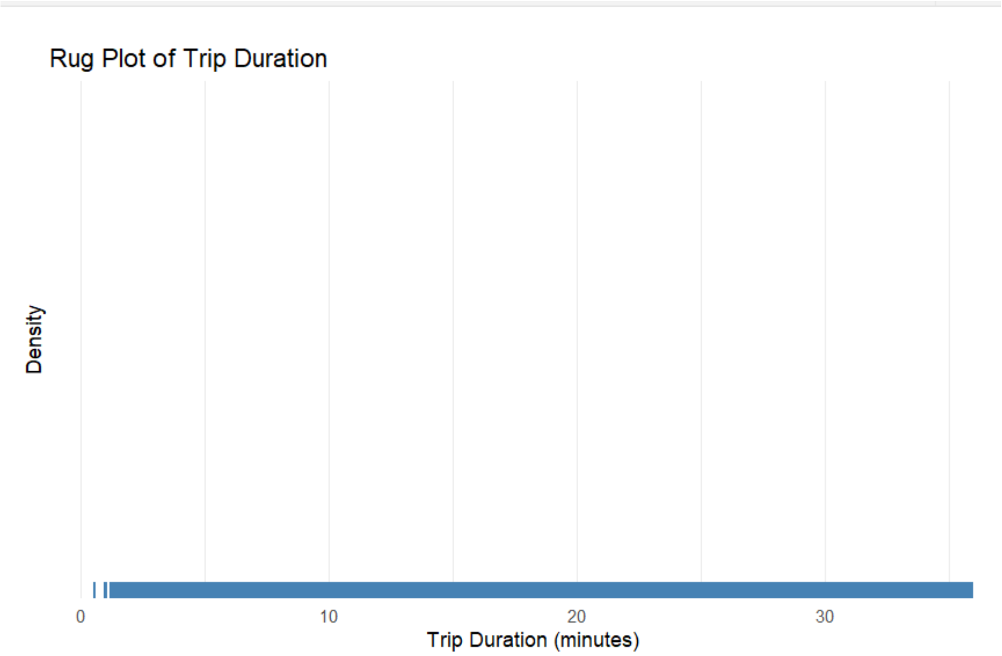
a) Violin Plot for Member Type vs Trip Duration
The violin plot shows the distribution of trip durations for members and casual riders. It reveals that casual riders tend to have longer trip durations compared to members.



This insight suggests that casual users, possibly tourists or infrequent users, tend to explore more or take leisurely rides compared to members who may use the service for commuting or shorter, routine trips.

b) Rug Plot of Trip Duration

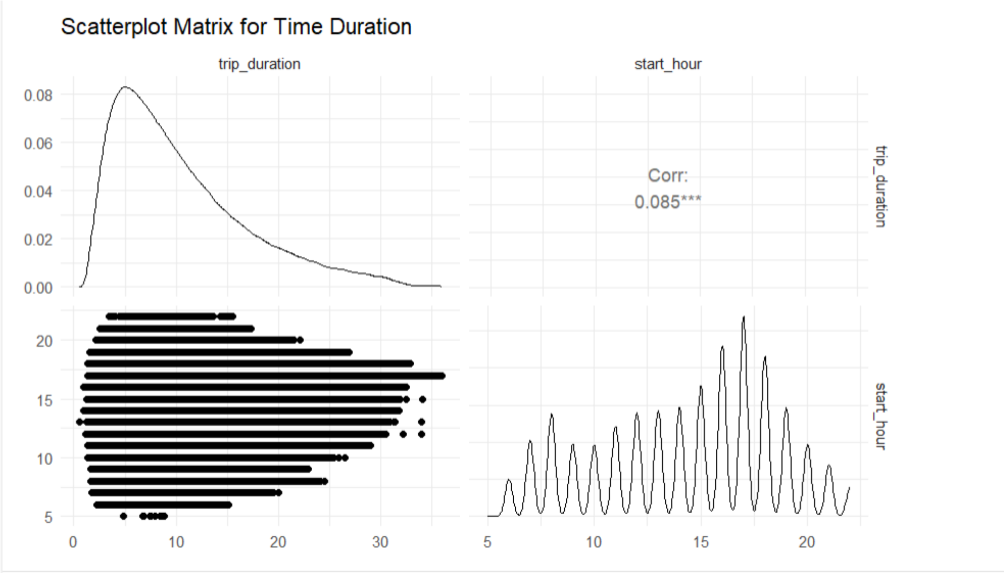
The rug plot displays individual trip durations as small ticks along the x-axis, providing a clear view of the density and spread of trip durations.



showing a dense concentration of trips with shorter durations. This finding indicates that most trips are relatively short, which aligns with typical urban bike-sharing usage patterns where bikes are used for quick, short-distance travel.

c) Scatterplot Matrix for time duration

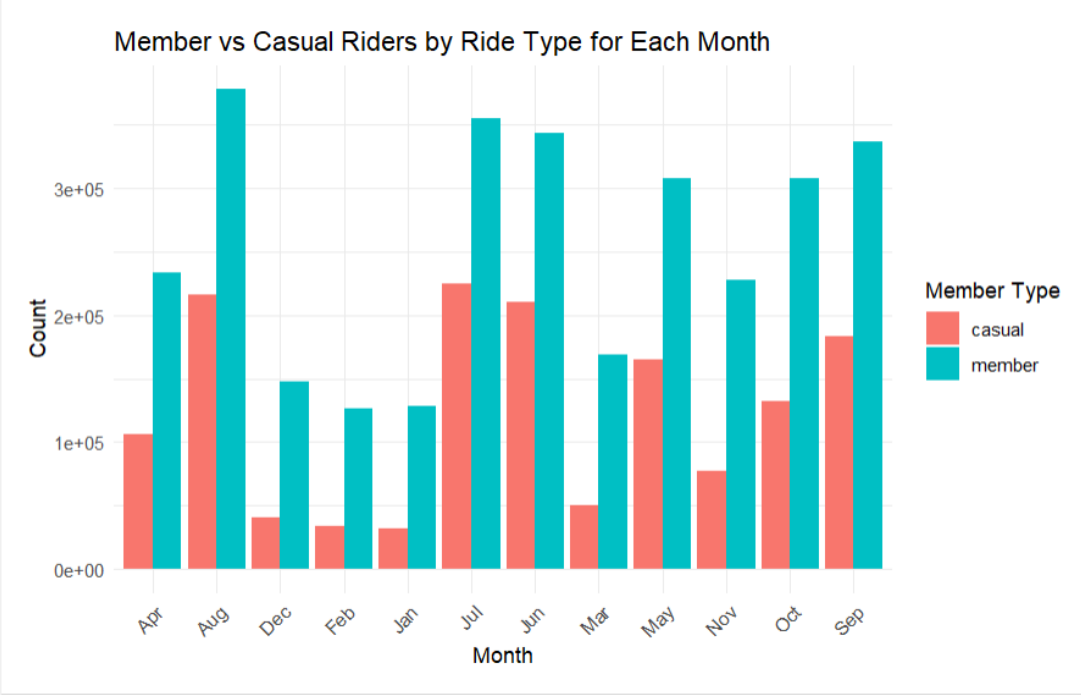
The scatterplot matrix helps in identifying correlations and patterns between trip duration and start hour. It shows how these variables interact and highlights any notable relationships.



It suggests that there are certain hours during the day when trips tend to be longer, particularly during late morning and early afternoon hours. This pattern could be due to leisure rides or flexible travel times outside of peak commuting hours.

d) Member vs Casual Riders by Ride Type for Each Month

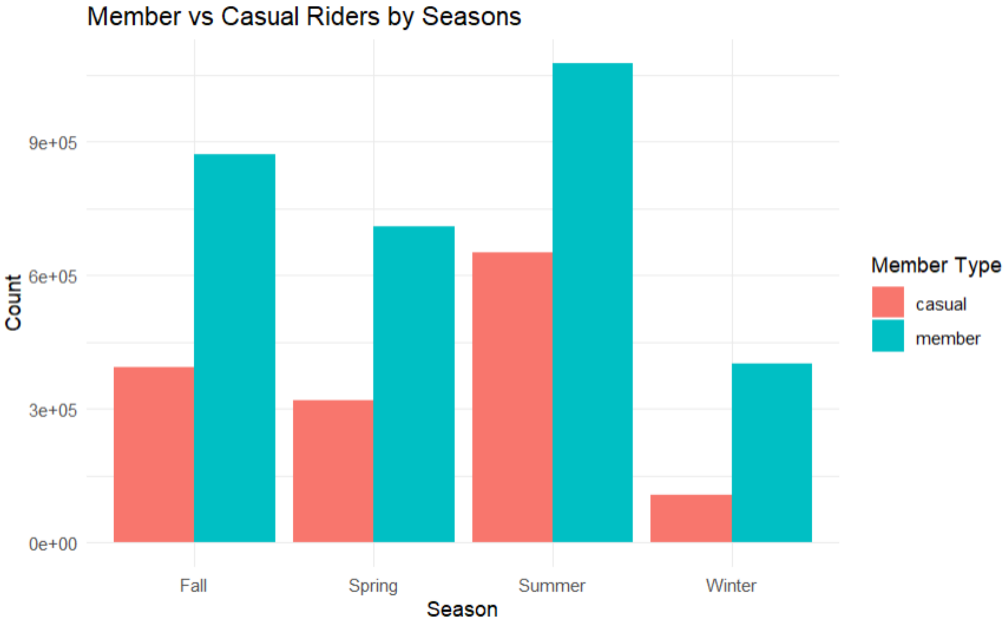
This bar plot illustrates the variation in ride types used by members and casual riders across different months. It highlights any seasonal preferences for ride types among the two groups.



Members consistently prefer classic bikes, while casual riders have a more varied preference, especially in summer months. This seasonal trend indicates that casual riders are more likely to try different ride types during peak tourist seasons.

e) Member vs Casual Riders by Seasons

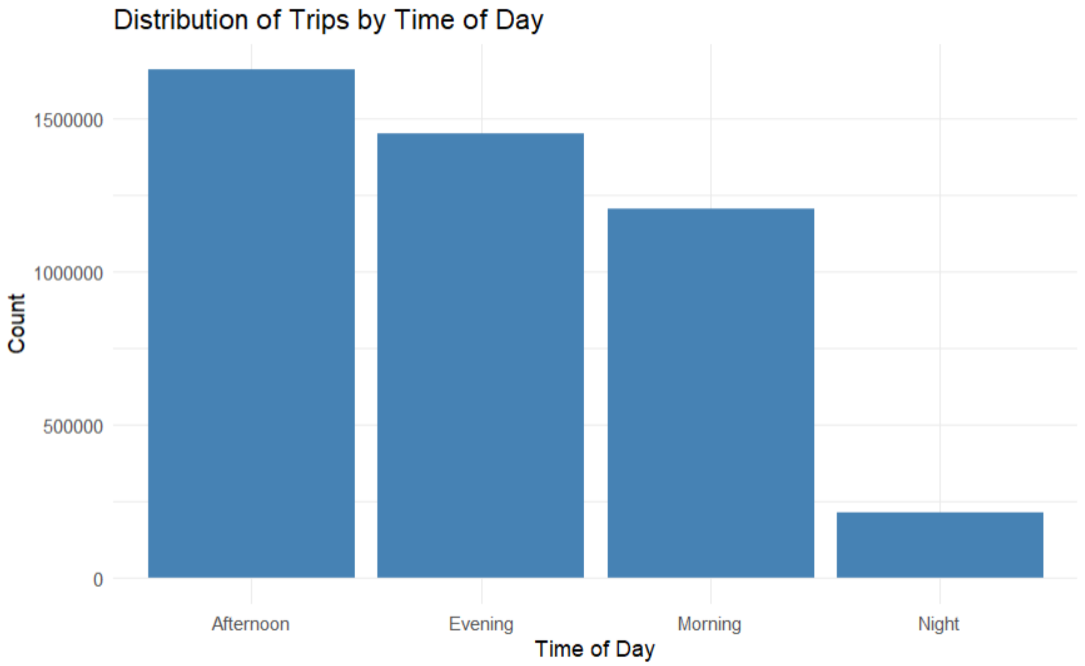
This bar plot shows the distribution of ride types used by members and casual riders across different seasons. It provides insights into how seasonal changes affect the preferences for ride types among these groups.



This analysis shows a clear seasonal trend where casual riders increase significantly in the summer, while members show a more consistent usage throughout the year. Casual riders predominantly prefer the summer months, suggesting a correlation with warmer weather and tourism activities.

f) Time of Day Distribution

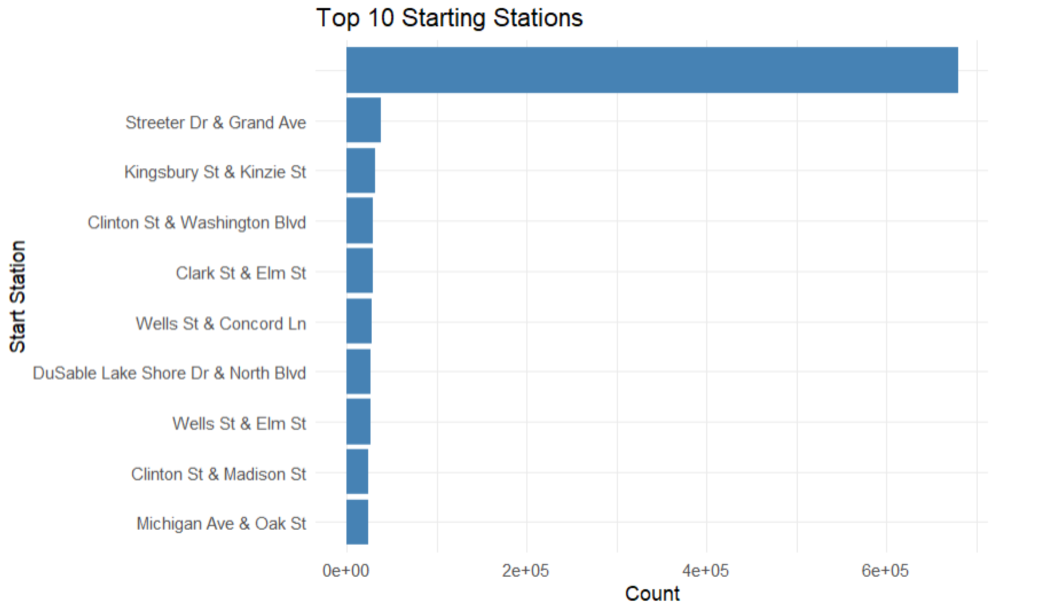
The histogram illustrates the distribution of trips starting at different hours of the day. It shows peak usage times, indicating when the bike-sharing service is most frequently used.



The histogram reveals peak usage hours, with the highest number of trips occurring during afternoon and evening rush hours. This pattern is typical of a commuting usage pattern, indicating that many users rely on the bike-sharing service for their daily commute.

g) Starting Station Distribution (Top 10)

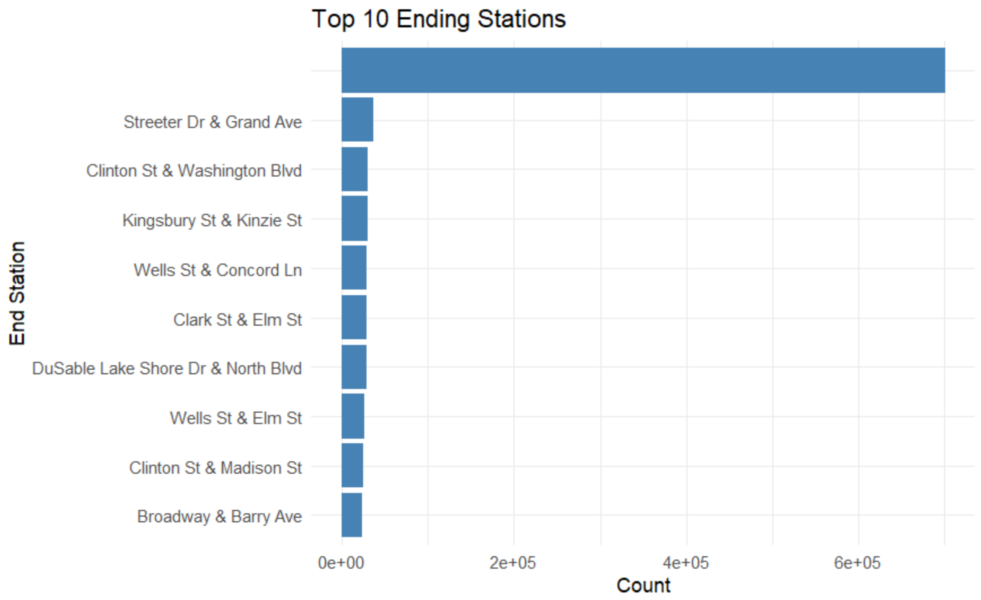
The bar plot shows the top 10 starting stations by trip count, providing insights into the most popular starting points for trips.



The top 10 starting stations are primarily located in busy, central areas, indicating high demand in these locations. This information is crucial for station placement and ensuring availability in high-traffic areas.

h) End Station Distribution (Top 10)

The bar plot shows the top 10 end stations by trip count, highlighting the most common destinations for trips.



Similarly, the end station distribution shows the most popular destinations. The alignment with starting stations suggests common travel routes and high-traffic areas. This data helps in understanding user flow and optimizing bike redistribution strategies to maintain service balance.

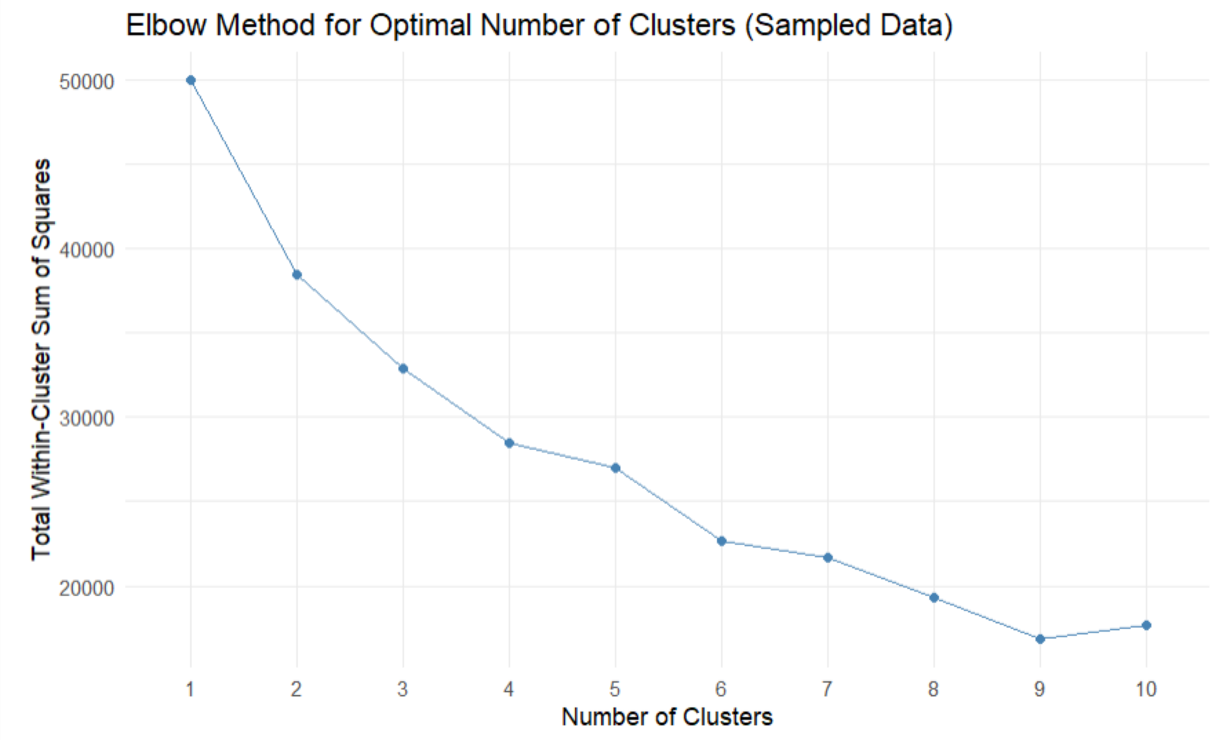
2.2.5 Clustering

Clustering is a crucial unsupervised learning technique used to group similar data points into clusters. In this project, we applied clustering analysis to understand the patterns and structures in our transformed dataset. This section details the steps and methodologies used to perform clustering, evaluate the optimal number of clusters, and analyze the resulting clusters.

To identify the optimal number of clusters, we employed three methods: the Elbow Method, the Silhouette Method, and the Gap Statistic Method. Each method provides a different perspective on selecting the optimal number of clusters.

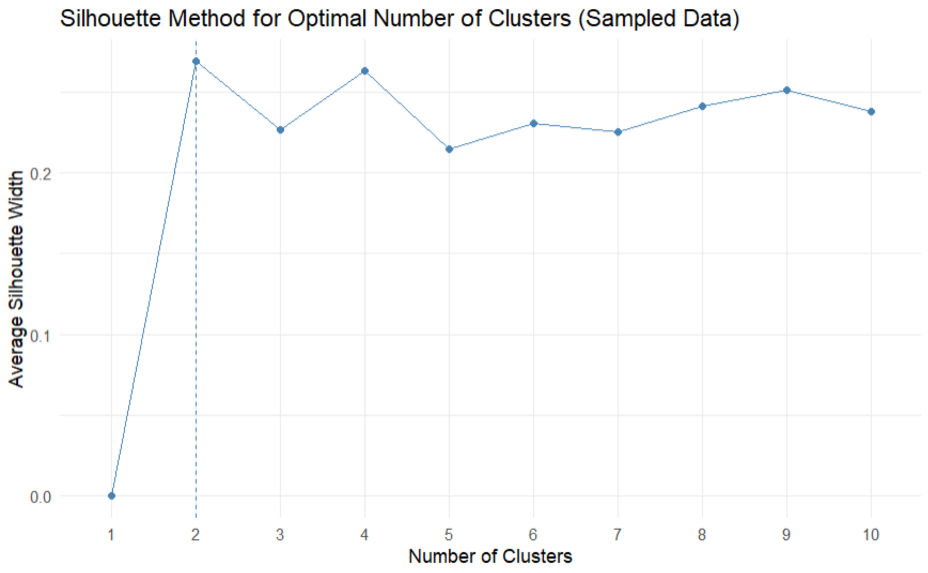
2.2.5.1 Elbow Method

The Elbow Method involves plotting the total within-cluster sum of squares (WSS) against the number of clusters. The point where the WSS starts to diminish significantly (the "elbow") indicates the optimal number of clusters. From the plot, we observed an elbow point at 3 clusters.



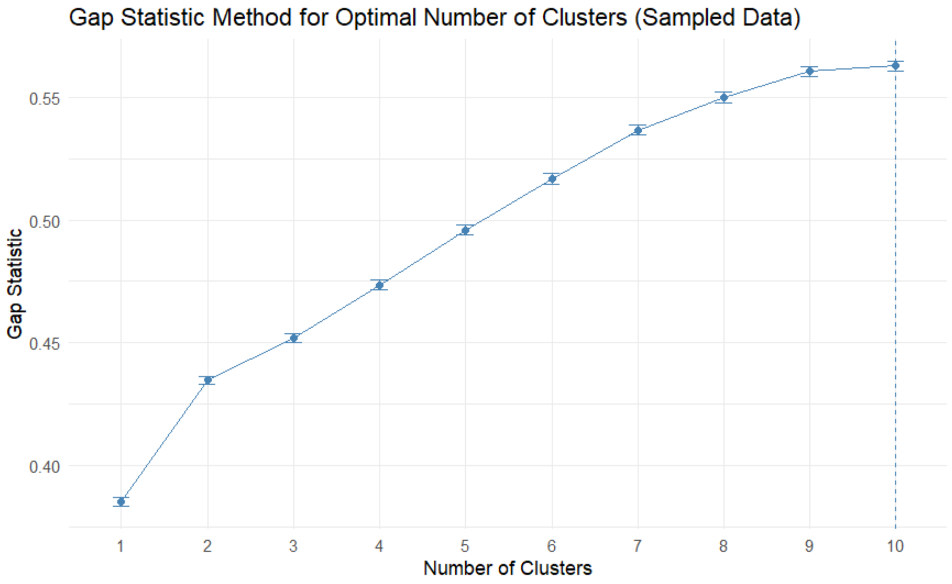
2.2.5.1.1 Silhouette Method

The Silhouette Method measures how similar each data point is to its own cluster compared to other clusters. The optimal number of clusters maximizes the average silhouette width. Our analysis indicated the highest silhouette width for 3 clusters.



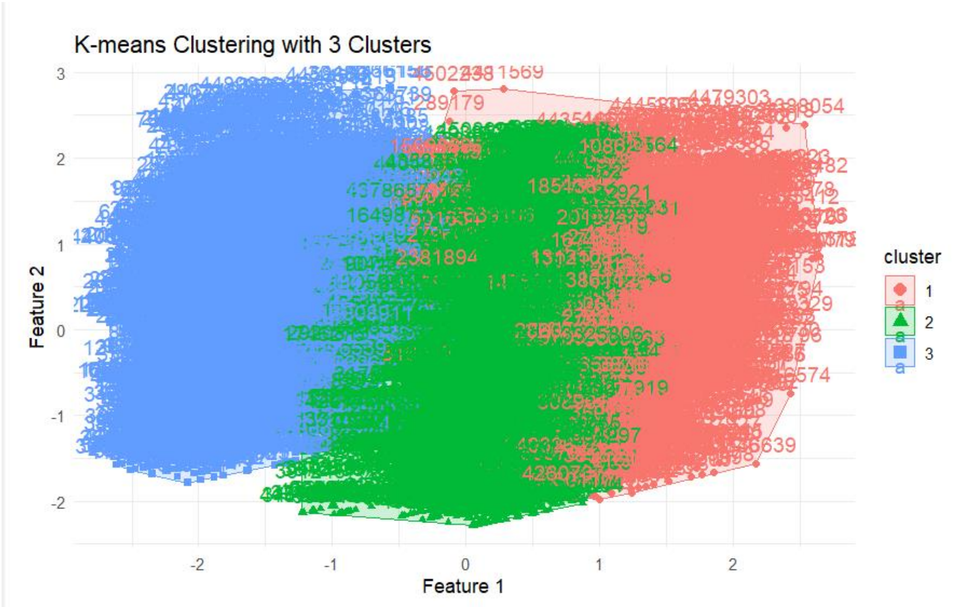
2.2.5.2 Gap Statistic Method

The Gap Statistic Method compares the total within intra-cluster variation for different numbers of clusters with their expected values under null reference distribution. The optimal number of clusters is chosen as the value where the gap statistic is maximized. This method also suggested 3 clusters.



2.2.5.3 K-means Clustering

Based on the results from the above methods, we determined the optimal number of clusters. For this analysis, we chose `optimal_clusters` (replace with the determined number of clusters). We then performed K-means clustering using this optimal number of clusters.



2.2.5.4 Clustering Evaluation

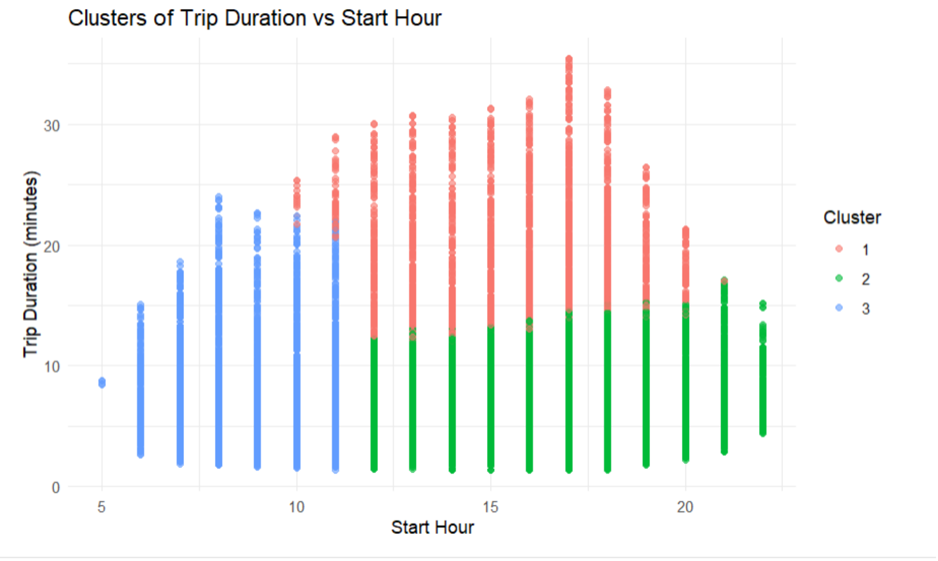
We evaluated the clustering results using the Total Within-Cluster Sum of Squares (WSS) and the average silhouette width.

2.2.5.5 Visualization of Clusters

To visualize the clustering results, we used ggplot2 to plot the clusters along with the data points. The visualizations include scatter plots showing the distribution of clusters based on `start_hour`, `trip_duration`, `day_of_week`, `season`, and `time_of_day`.

2.2.5.6 Clusters of Trip Duration vs Start Hour

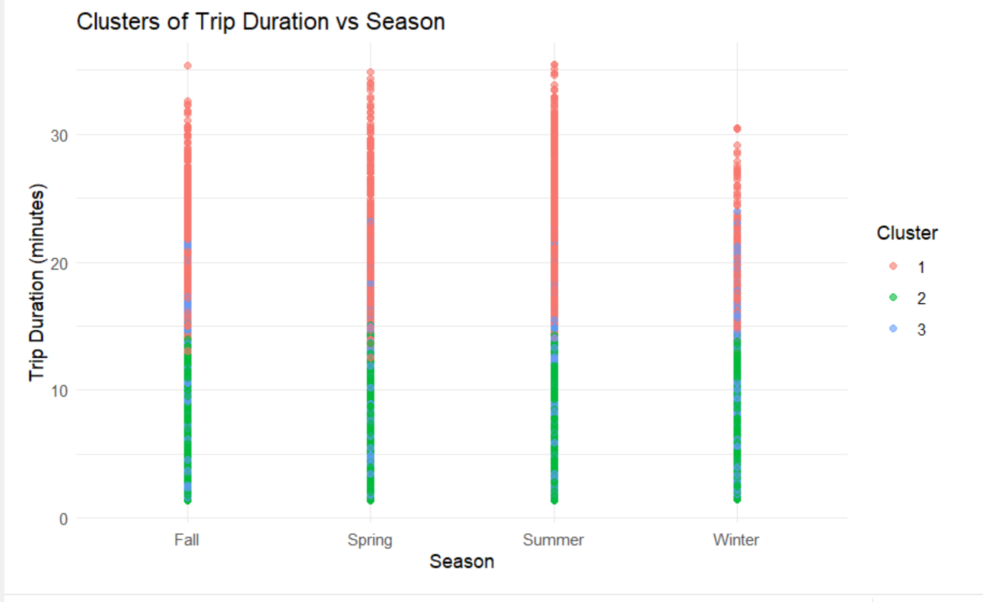
This plot visualizes the relationship between trip duration and start hour, color-coded by cluster. It helps identify patterns in trip duration based on the time of day.



- The clusters indicate distinct groups with varying trip durations and start hours.
- One cluster (Cluster 1) tends to have shorter trip durations and is more spread out across different start hours.
- Another cluster (Cluster 2) has trips that start predominantly during peak hours (morning and evening) and have moderate durations.
- A third cluster (Cluster 3) represents trips with longer durations that occur at various times throughout the day.

2.2.5.7 Clusters of Trip Duration vs Season

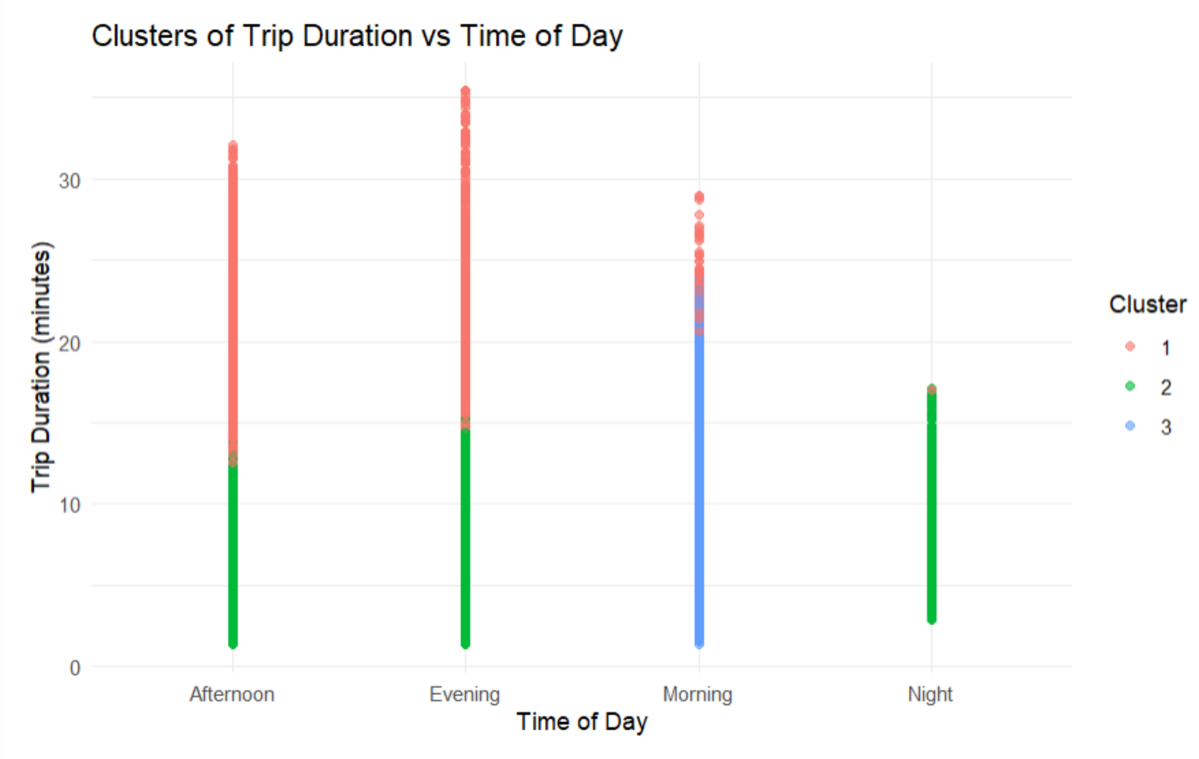
This plot visualizes the relationship between trip duration and season, color-coded by cluster. It helps identify seasonal patterns in trip durations.



- Clusters are distinctly separated by both trip duration and seasonal variation.
- One cluster contains shorter trips, prevalent in all seasons, but with a higher density in the summer.
- Another cluster has longer trips, more common during spring and summer seasons.
- The third cluster, consisting of trips of moderate durations, is distributed across all seasons but is slightly more frequent in summer and fall.

2.2.5.8 Clusters of Trip Duration vs Time of Day

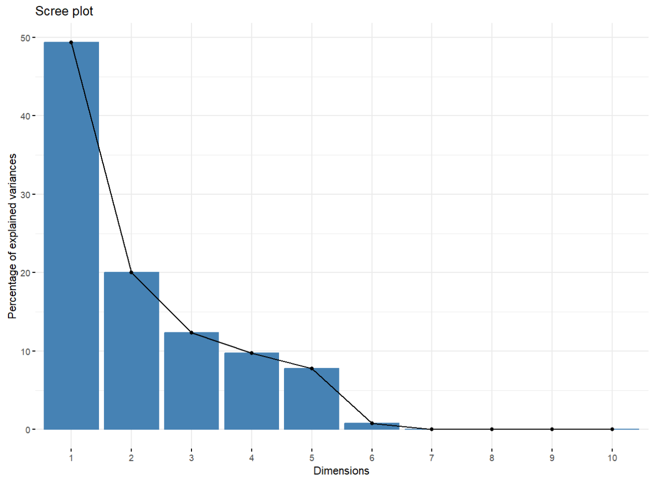
This plot visualizes the relationship between trip duration and time of day, color-coded by cluster. It helps identify how trip durations vary at different times of the day.



- One cluster includes shorter trips that are evenly spread across different times of the day.
- Another cluster features trips with moderate durations, more concentrated during morning and evening hours.
- A third cluster comprises longer trips, occurring at various times but with a slight preference for mid-day and afternoon periods.

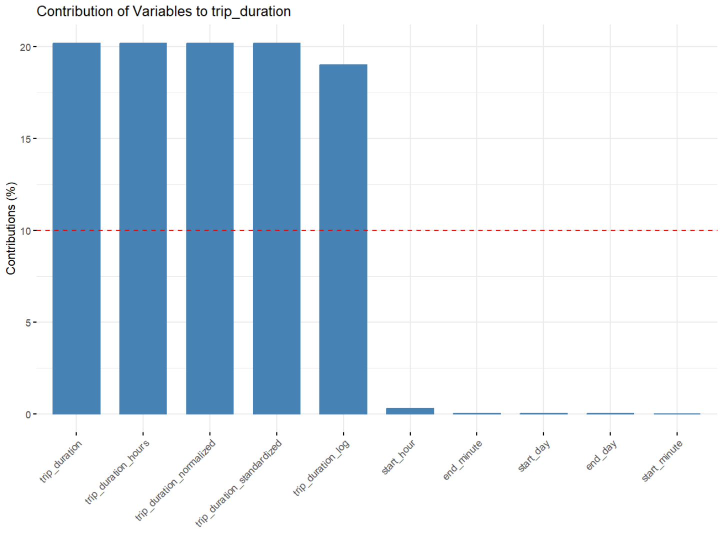
2.2.6 Dimensionality Reduction

Scree Plot:

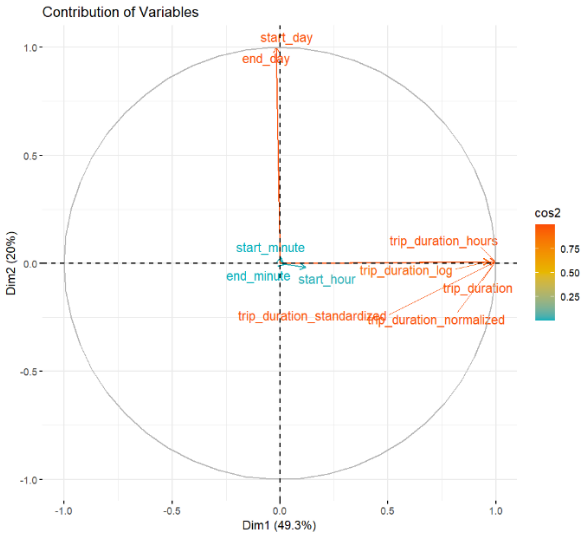


The scree plot is used to determine the number of principal components to retain where the x-axis shows dimensions representing the principal components. In contrast, the y-axis represents the percentage of total variance in the data for each component. Assume that the first component is trip_duration then it has the highest variance in the data of nearly 50%.

Contribution of variables towards trip_duration:

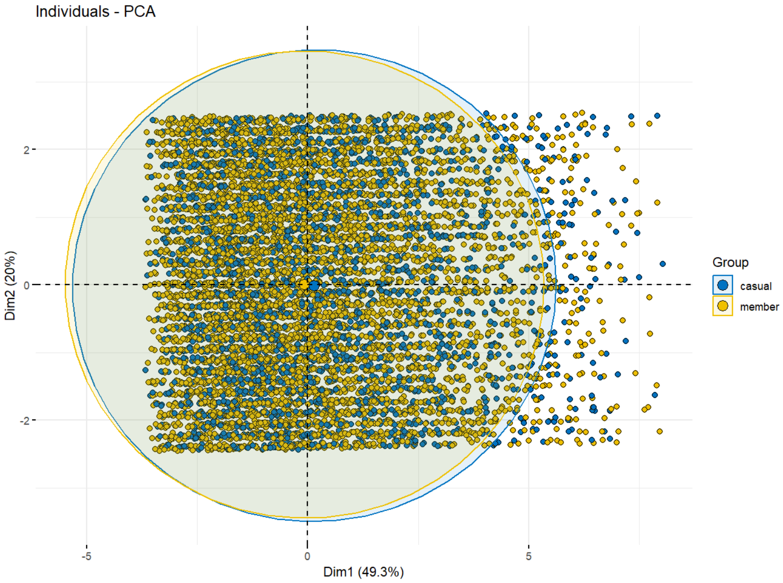


This bar chart shows how different variables contribute to the duration of a trip where the x-axis represents the variables that affect trip duration whereas the y-axis represents the percentage contribution of each of the variables. The variables trip_duration, trip_duration_hours, trip_duration_ normalized, trip_duration_standardized, trip_duration_log has the highest percentage of contributing towards trip duration.



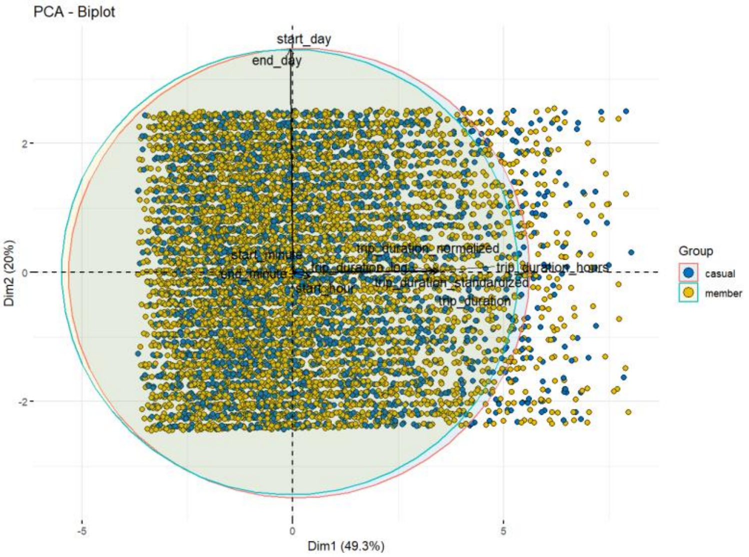
This biplot visualizes the contributions of various variables to two principal components. The axis represents the first two principal components, which together explain a significant portion of the variance in the data (49.3% for Dim1 and 20% for Dim2). Variables like trip_duration_hours and trip_duration_log have strong contributions with Dim1, indicating they explain a lot of the variance captured by this component whereas variables like start_day and end_day contribute more to Dim2, suggesting they explain variance captured by this component.

Individual PCA Biplot:



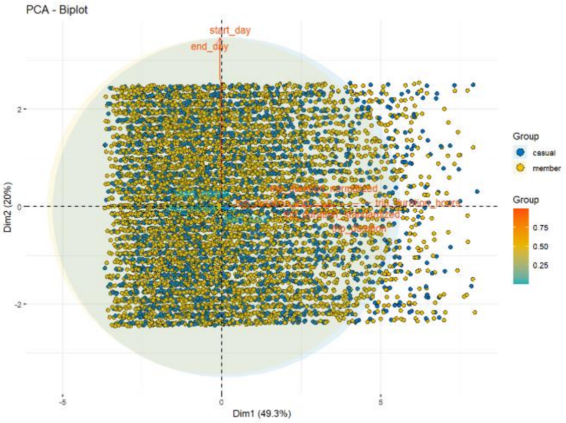
This plot shows two components on each axis with 49.3% of the variance on Dim1 compared to 26.9% on Dim2 of the variance in the data that captures a significant portion of the variability in the dataset. Each point represents an individual observation, and they are based on two groups: casual and member. Points that are close to each other on the plot are similar in terms of the principal components whereas the points that are far apart are dissimilar and the separation between the yellow and blue points indicates differences between the casual and member groups.

Combined PCA Biplot:



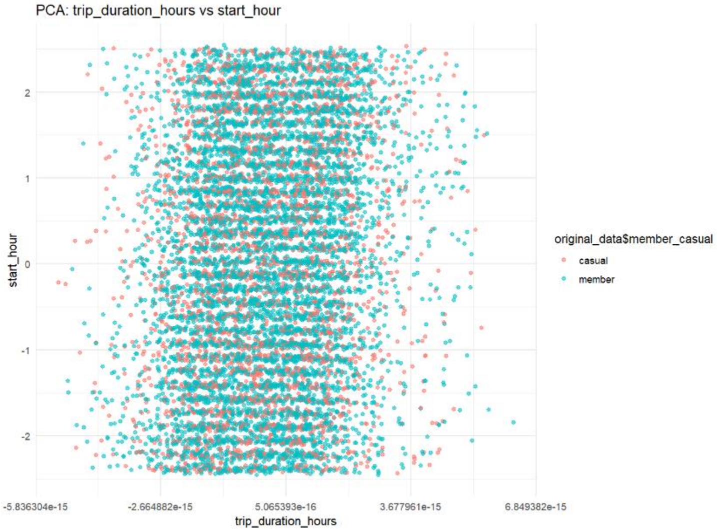
This plot shows two components on each axis with 49.3% of the variance on Dim1 compared to 26.9% on Dim2 of the variance in the data that captures a significant portion of the variability in the dataset. Each point represents an individual observation, and they are based on two groups: casual and member. The labeled vectors indicate the direction and magnitude of each variable's contribution to the principal components such as variables like start_day and end_day have arrows pointing in specific directions, showing how they contribute to Dim1 and Dim2. The green ellipse encompasses most of the data points and shows a cluster for both groups which indicates that the majority of the data points share similar characteristics. Points that are close to each other on the plot are similar in terms of the principal components whereas the points that are far apart are dissimilar and the separation between the yellow and blue points indicates differences between the casual and member groups.

Combined PCA Biplot with cos2:

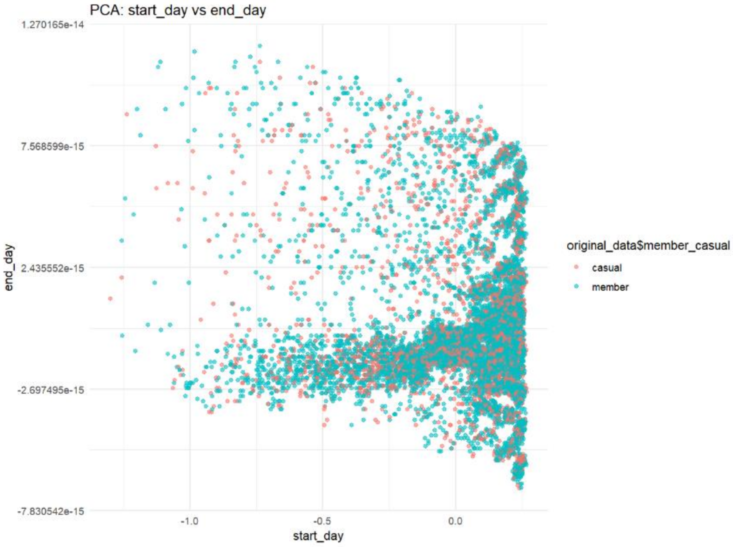


This plot shows two components on each axis with 49.3% of the variance on Dim1 compared to 26.9% on Dim2 of the variance in the data that captures a significant portion of the variability in the dataset. Each point represents an individual observation, they are based on two groups: casual and member where the shading of the points indicates the value of another variable ranging from 0.25 to 0.75. The labeled vectors indicate the direction and magnitude of each variable's contribution to the principal components such as variables like start_day and end_day have arrows pointing in specific directions, showing how they contribute to Dim1 and Dim2. The green ellipse encompasses most of the data points and shows a cluster for both groups which indicates that the majority of the data points share similar characteristics. Points that are close to each other on the plot are similar in terms of the principal components whereas the points that are far apart are dissimilar and the separation between the yellow and blue points indicates differences between the casual and member groups.

Additional PCA Plots:

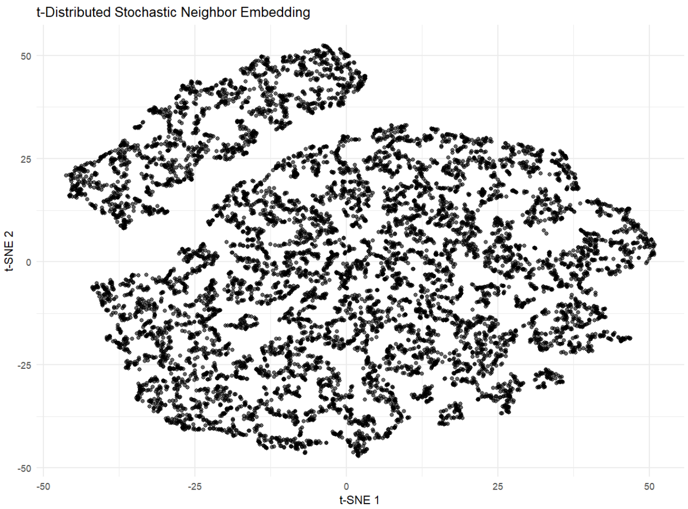


This scatter plot visualizes the relationship between trip duration (in hours) and the hour of the day when trips start where The x-axis represents trip_duration_hours which is the duration of trips in hours and the y-axis represents start_hour which is the hour of the day when trips start. Each point represents an individual trip and the points are color-coded into two groups: casual users and members. The points are densely plotted, creating a cloud-like distribution across the graph but there are no immediately discernible patterns or clusters, indicating that trip duration and start hour are spread out across the dataset.



This scatter plot visualizes the relationship between the start day and end day of trips where the x-axis represents start_day which is the day when trips start whereas the y-axis represents end_day which is the day when trips end. Each point represents an individual trip and the points are color-coded into two groups: casual users and members. Points that are close to each other on the plot are similar in terms of the principal components whereas points that are far apart are dissimilar and the separation between the red and blue points indicates differences between the two groups.

t-sne plot:



This plot visualizes high-dimensional data in a two-dimensional space where each point represents an individual observation and the points are distributed across the plot forming two distinct clusters. The presence of two distinct clusters suggests that the t-SNE algorithm has identified two groups within the data based on their similarity where one cluster is larger than the other which indicates a difference in the number of observations within each group. Points that are close to each other on the plot are similar in terms of the high-dimensional data whereas points that are far apart are dissimilar.

2.2.7 Feature Selection

Feature selection is a critical process in data analysis and modeling, aimed at identifying the most relevant features that contribute to the predictive power of a model. Initially, a correlation analysis is conducted to identify and remove highly correlated features, reducing redundancy in the dataset. Subsequently, recursive feature elimination (RFE) is applied to further refine the feature set. RFE works by recursively fitting a model and removing the least important features, as determined by the model's performance. This iterative process continues until the optimal subset of features is identified, balancing model accuracy and complexity. To handle large datasets efficiently and manage memory usage, the data.table package is used. data.table provides optimized and fast data manipulation capabilities, which are particularly useful when dealing with large volumes of data. By selecting a subset of the most significant features and using data.table for efficient data processing, we enhance the model's interpretability and efficiency, ultimately improving its predictive performance.

2.2.8 Model Selection and Training

In our analysis, we focused on two robust machine learning algorithms: Random Forest and XGBoost. These models were chosen for their effectiveness in handling complex datasets and their proven track record in delivering high predictive accuracy.

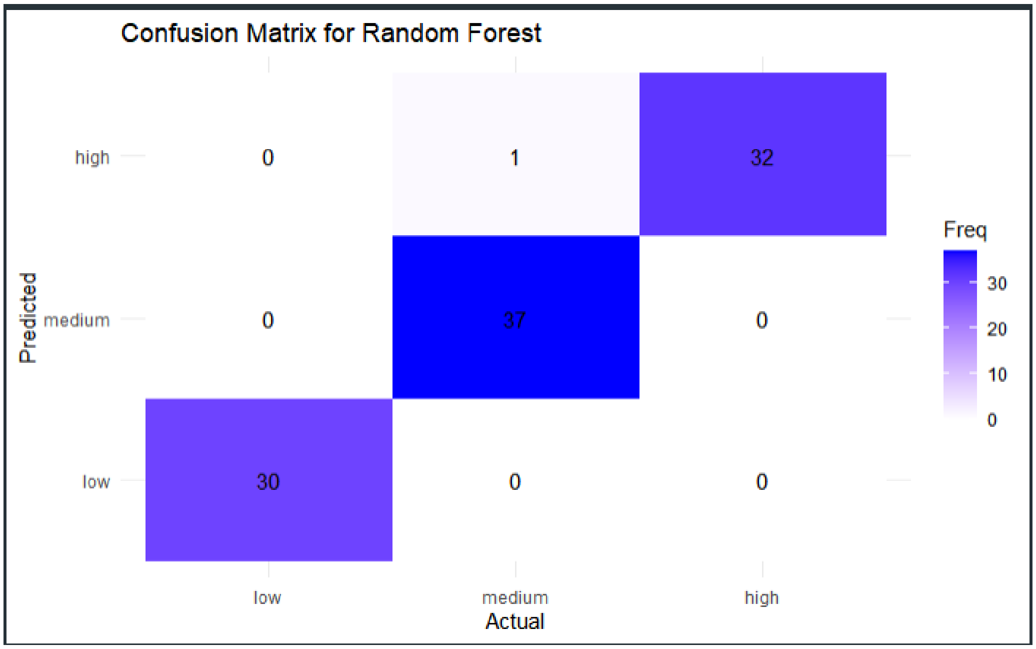
Random Forest:

In this analysis, we employed the Random Forest algorithm to predict bike trip durations based on a range of features such as ride type, member status, station names, and time-related factors. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. This approach helps to improve predictive accuracy and control overfitting compared to a single decision tree.

We first preprocessed the data by converting date columns to numeric format and ensuring all categorical variables were consistent between training and test datasets. To adapt the continuous trip duration into a classification problem, the data was binned into three categories: low, medium, and high trip durations. The model was trained using cross-validation to ensure robustness and generalization across different data subsets.

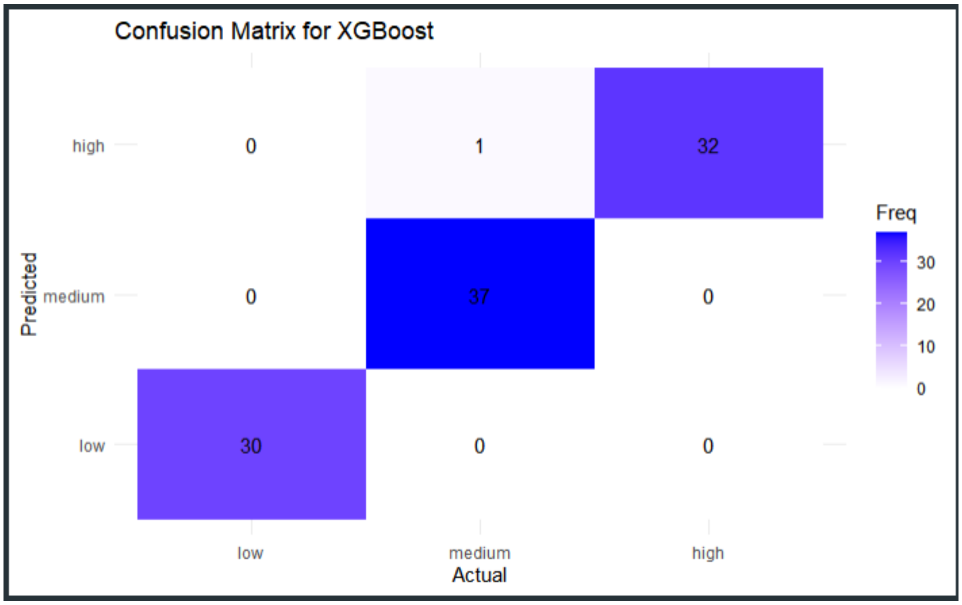
The results demonstrated a high accuracy rate of 99%, indicating that the Random Forest model was highly effective in categorizing trip durations into their respective bins. The confusion matrix (see below) showed that the model correctly predicted most instances with minimal misclassification, particularly excelling in predicting 'low' and 'high' categories with perfect sensitivity and specificity. The kappa statistic of 0.9849 further supports the model's strong agreement beyond chance. These results suggest that the Random Forest model effectively captured the relationships between the predictor variables and trip durations, providing reliable predictions for the dataset analyzed.

Confusion Matrix:



XGBoost:

The XGBoost model demonstrated excellent performance in predicting bike trip durations, achieving an accuracy of 99% across the test dataset. The model's high accuracy indicates that it was highly effective in categorizing trip durations into the predefined bins of low, medium, and high. This strong performance is further evidenced by the 95% confidence interval for accuracy, which ranges from 94.55% to 99.97%, suggesting consistent reliability across different samples.



The confusion matrix reveals that the XGBoost model correctly classified most of the instances, with only a single misclassification observed in the high-duration category. The sensitivity, which measures the true positive rate, was perfect for the low and high categories, with values of 1.0, and nearly perfect for the medium category at 0.9737. Specificity, which measures the true negative rate, was also 1.0 for the low and medium categories, indicating no false positives for these categories, and slightly lower for the high category at 0.9853.

The positive predictive value (PPV) and negative predictive value (NPV) were also high across all categories, demonstrating the model's accuracy in making positive predictions and correctly identifying negative cases. The kappa statistic of 0.9849 highlights strong agreement between the predicted and actual classifications, beyond what would be expected by chance alone.

Overall, these results suggest that the XGBoost model effectively captured the underlying patterns in the data, providing accurate and robust predictions. The high sensitivity and specificity across categories indicate that the model is reliable in distinguishing between different trip duration bins. The single misclassification in the high category indicates that there may be room for further fine-tuning, but overall, the XGBoost model proves to be a powerful tool for this classification task.

From all these analysis it can be seen that random forest r-squared performs better with and the accuracy is almost 100 with that.

2.3 Software Packages and Tools

- **R Language:** The primary programming language for analysis.
- **R Studio:** Integrated development environment for R.

Libraries:

- dplyr: Data manipulation and transformation.
- ggplot2: Data visualization.
- tidyr: Data tidying and reshaping.
- dbscan: Density-based spatial clustering of applications with noise for outlier detection.
- isotree: Isolation forest for outlier detection.
- lubridate: Date and time manipulation.
- caret: Machine learning model training and evaluation.
- randomForest: Implementation of Random Forest algorithm.
- xgboost: Implementation of XGBoost algorithm.
- data.table: High-performance data manipulation.
- base: Core functions for R operations.
- lattice: Improved base R graphics.
- stringr: String manipulation and regular expressions.
- stats: Statistical functions and algorithms.
- e1071: Support vector machines and statistical learning.
- Matrix: Sparse and dense matrix classes and methods.
- factoextra: Extract and visualize the results of multivariate data analyses.
- cluster: Cluster analysis for datasets.
- gridExtra: Arrange multiple grid-based figures on a page.
- pROC: Display and analyze receiver operating characteristic (ROC) curves.
- mlbench: Machine learning benchmark problems.
- plotly: Interactive web-based graphics.

- dendextend: Extending 'dendrogram' functionality in R.
- factoextra: Extract and visualize multivariate analysis results.
- This list of libraries provides the necessary tools and functionalities to perform data manipulation, visualization, clustering, modeling, and evaluation for our analysis of Divvy bike usage patterns in Chicago. Each library plays a crucial role in ensuring the accuracy and efficiency of our data analysis and model building processes.

3. Challenges

One of the significant challenges encountered in this project was the large size of the original combined dataset, which was 1.2 GB. After performing outlier detection and feature selection, the dataset was reduced to 830 MB. Despite this reduction, it was still challenging to run models on the entire dataset due to memory and processing constraints. Consequently, we had to use a subset of the data for modeling to ensure efficient and feasible analysis. This limitation required careful consideration in data sampling to maintain the integrity and representativeness of the dataset.

4. Conclusion

In this project, we successfully analyzed Divvy bike usage patterns using machine learning models, specifically focusing on Random Forest and XGBoost, to predict trip durations. The dataset was meticulously cleaned and enriched with features such as time of day, season, and trip metrics to enhance the models' predictive capabilities. Both Random Forest and XGBoost demonstrated high accuracy in classifying trip durations into low, medium, and high categories, as evidenced by their performance metrics and confusion matrices. The project highlighted the models' ability to uncover key drivers of trip duration variability and their potential to inform operational strategies for bike-sharing services. By predicting average trip durations for January 2024, we provided actionable insights that can aid in optimizing resource allocation and improving service efficiency. Future work could explore integrating additional data sources, such as weather conditions, to further refine predictions and support strategic planning for urban mobility solutions.

5. Future Scope

- **Integration with Public Transit:** Analyze the relationship between Divvy bike docks and CTA bus stops to optimize the multimodal transportation network in Chicago.
- **Population Analysis:** Correlate Divvy usage data with population density data to identify areas with unmet demand for bike docks and cycles.

- **Expansion to Other Cities:** Apply the analysis methodology to other cities with bike-sharing programs to compare and improve overall urban mobility.
- **Real-Time Data Integration:** Incorporate real-time data feeds to provide dynamic recommendations for bike redistribution and dock availability.
- **User Experience Enhancement:** Analyze user feedback and usage patterns to improve overall user experience.
- **Sustainability Impact:** Evaluate the environmental benefits of the bike-sharing program and suggest improvements for increasing its positive impact on urban sustainability.

6. References:

L. Czarlnski, "Exploratory Data Analysis (EDA) of the Chicago Divvy Bikes Dataset," Medium. [Online]. Available: https://medium.com/@leonczarlnski/exploratory-data-analysis-eda-of-the-chicago-divvy-bikes-dataset.

"Exploring variations in Divvy bike station usage volume: from historical trip records to Google Street view images," MACS 37000 (Spring 2021) Thinking with Deep Learning for Complex Social & Cultural Data Analysis, uchicago.edu. [Online]. Available: https://uchicago.edu/macs37000/divvy-bike-station-usage.

"Divvy Trips," City of Chicago, Data Portal. [Online]. Available: https://data.cityofchicago.org/Transportation/Divvy-Trips.

Shivaniwac, "Quarterly Success: Divvy Bike's 2024 Growth Analysis," Medium, May 2024. [Online]. Available: https://medium.com/@shivaniwac/quarterly-success-divvy-bikes-2024-growth-analysis-e49927841eaf.

"Index of bucket 'divvy-tripdata'," [Online]. Available: https://divvy-tripdata.s3.amazonaws.com/index.html.