CSP 571 DATA PREPARATION AND ANALYSIS PROJECT

ANALYZING DIVVY BIKE USAGE PATTERNS IN CHICAGO

DATE: 08/02/2024

AUTHORS

MOHAMMAD HAMZA PIRACHA | POSHAN PANDEY | USMAN MATHEEN HAMEED

INTRODUCTION

- **Objective**: Understand the intricacies of Divvy bike usage in Chicago in 2023.

- **Goals**:
  - Uncover significant ridership trends and behaviors.
  - Utilize machine learning algorithms to analyze data.
  - Accurately predict trip durations and enhance user satisfaction.

LITERATURE REVIEW

•We explored existing studies and articles on bike-sharing systems, focusing on Divvy bikes in Chicago.

•**Key Insights**: Highlighted trends in usage patterns, seasonal variations, and demographic factors.

•Reviewed methodologies for data cleaning, clustering, and predictive modeling to establish a strong foundation for our analysis

DATA PREPARATION AND CLEANING

- **Data Preparation**:
  - Merged monthly ride data into a cohesive dataset, merged_data.csv, for comprehensive analysis.

- **Data Cleaning**:
  - Removed irrelevant columns and standardized timestamps for consistency.
  - Derived features such as trip duration and time of day to enrich our dataset.

OUTLIER DETECTION

- **Methods Applied**:
  - Z-Score: Identified extreme outliers through statistical deviation.
  - Isolation Forest: Leveraged machine learning to isolate anomalous rides.

- **Outcome**: Successfully identified and removed data anomalies, ensuring robust analysis.

DATA TRANSFORMATION

Methods Applied:

- Z-Score Method: Identified extreme outliers by evaluating how many standard deviations away from the mean each data point lies. Observations beyond three standard deviations were considered anomalies.

- Isolation Forest: Employed an ensemble machine learning algorithm that isolates anomalies by randomly selecting features and then randomly partitioning values. This method effectively pinpoints anomalies in less obvious conditions.

Outcome:

Through meticulous application of these methods, we successfully identified and removed data anomalies, ensuring the dataset's integrity and robustness for subsequent analysis.

DISTRIBUTION ANALYSIS

Visual Techniques:

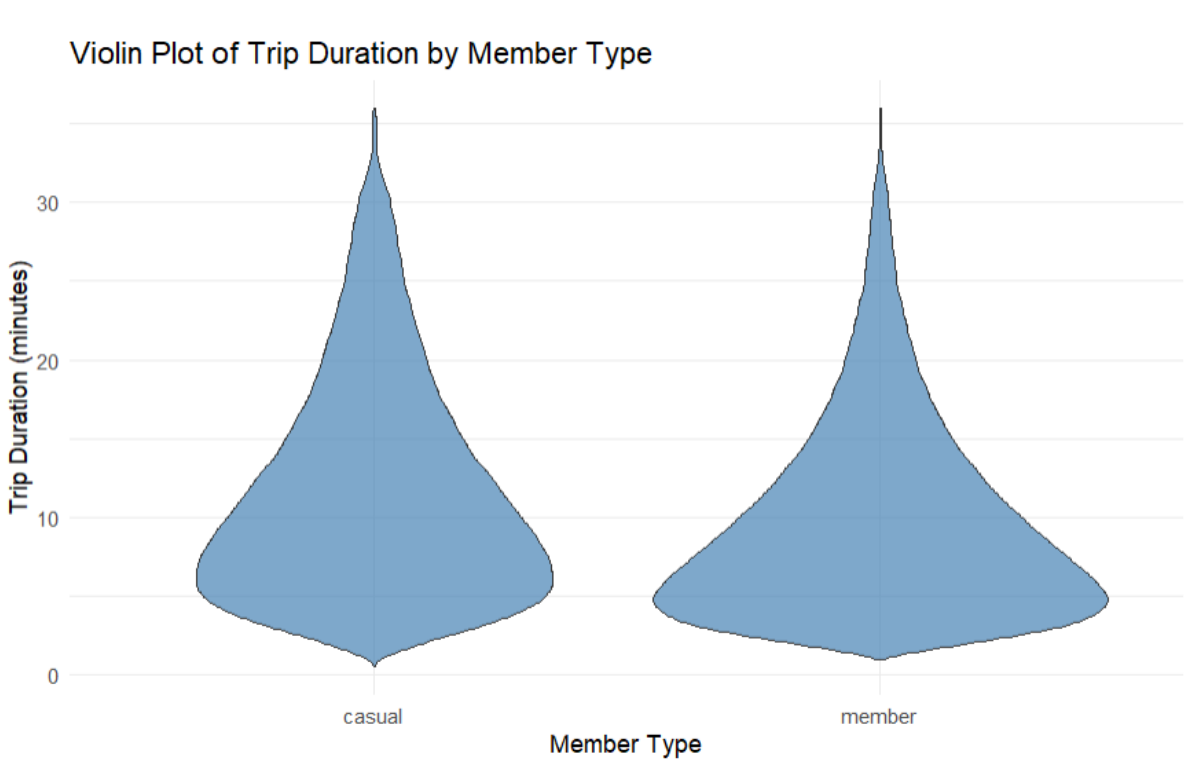
- Violin Plot for Member Type vs Trip Duration
- Rug Plot of Trip Duration
- Scatterplot Matrix for time duration
- Member vs Casual Riders by Ride Type for Each Month
- Member vs Casual Riders by Seasons
- Time of Day Distribution
- Starting and End Station Distribution (Top 10)

DISTRIBUTION ANALYSIS

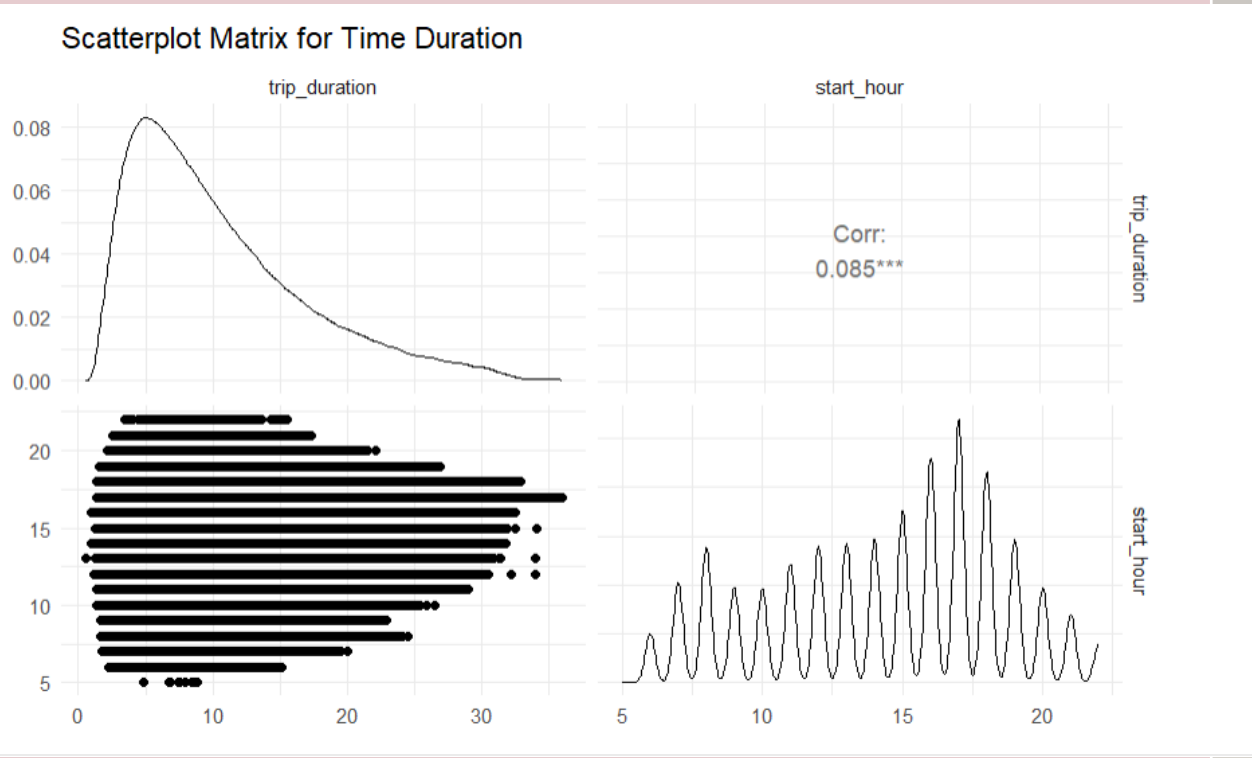
Visual Techniques:

- Violin Plot for Member Type vs Trip Duration
- Rug Plot of Trip Duration
- Scatterplot Matrix for time duration
- Member vs Casual Riders by Ride Type for Each Month
- Member vs Casual Riders by Seasons
- Time of Day Distribution
- Starting and End Station Distribution (Top 10)

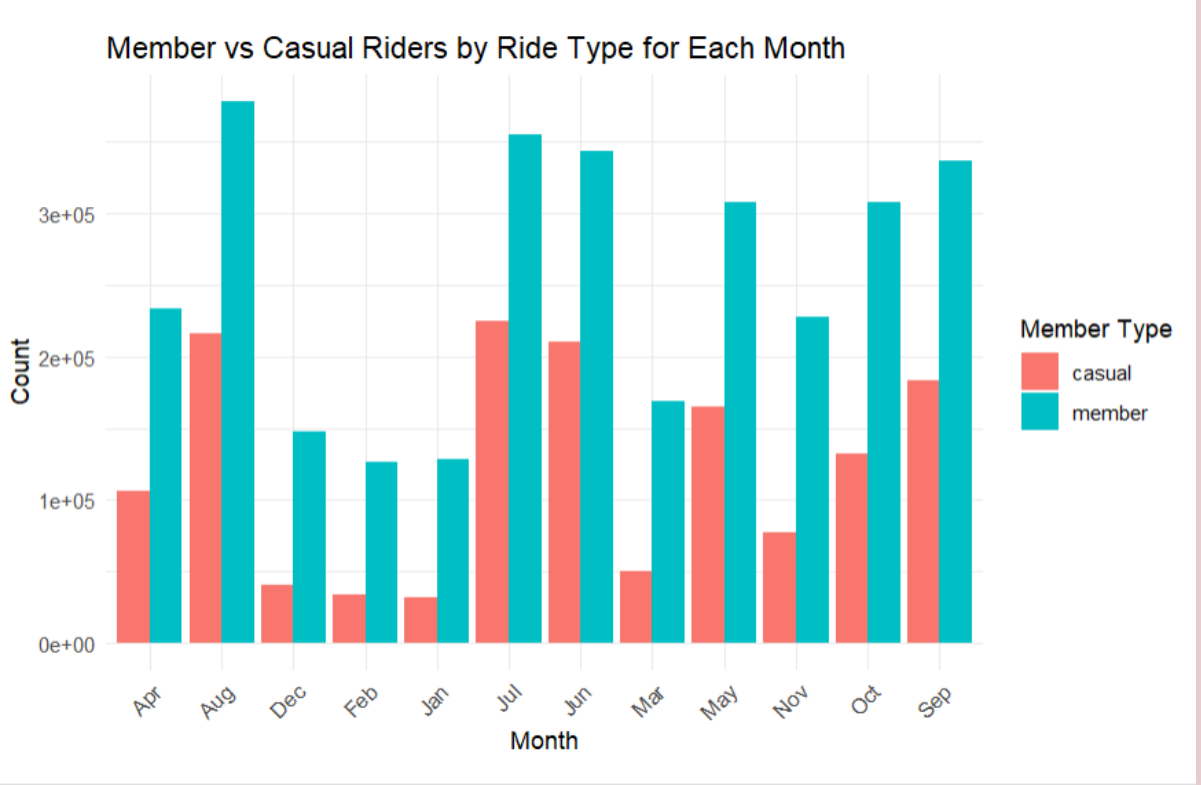
Violin Plot for Member Type vs Trip Duration



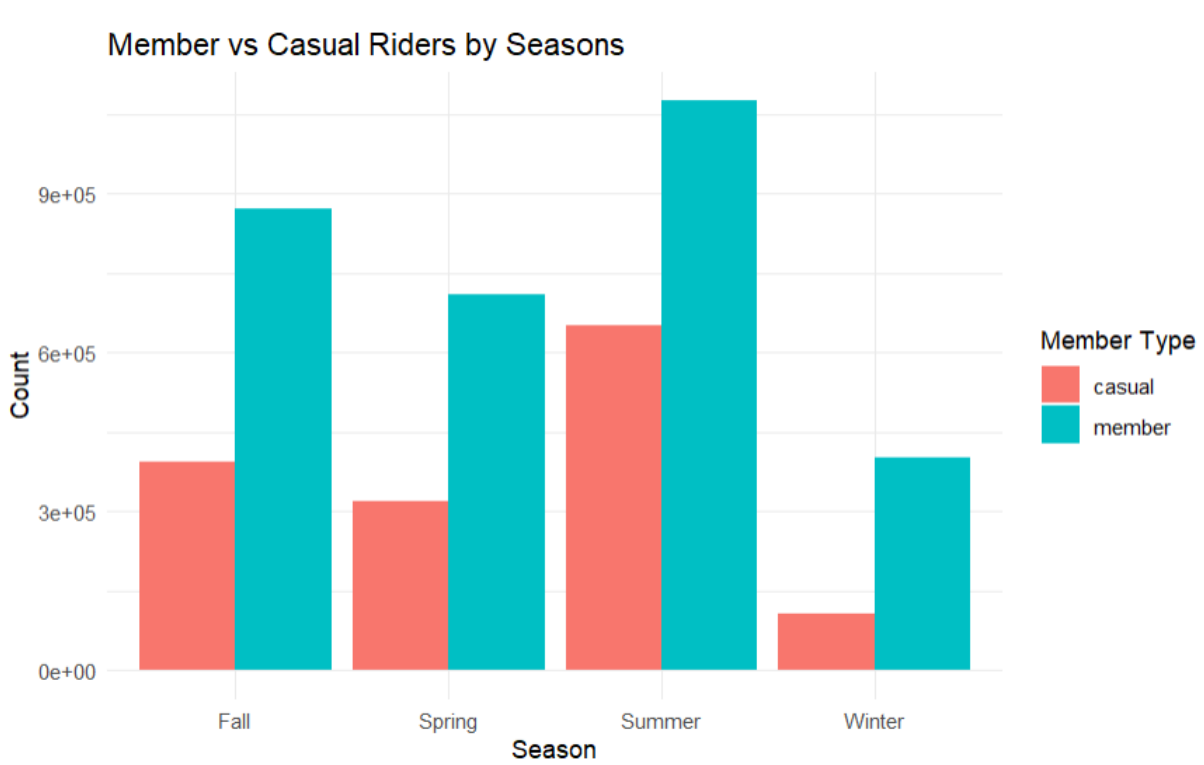
Scatterplot Matrix for time duration



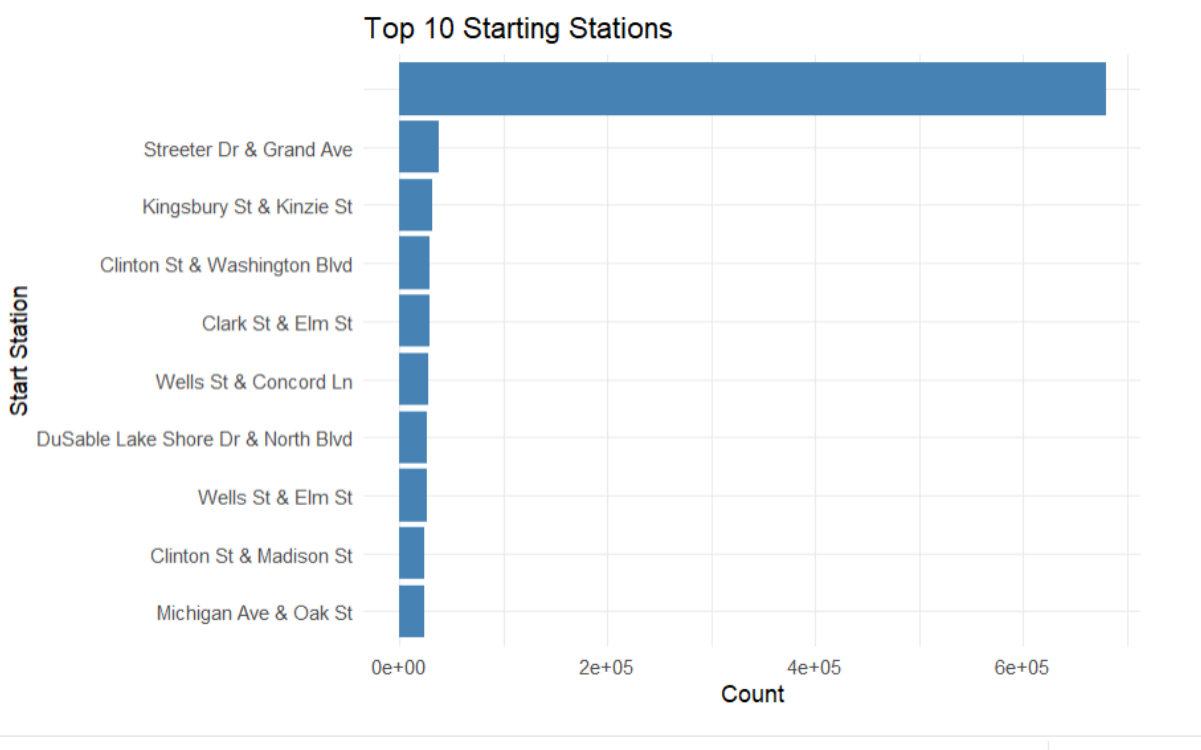
Member vs Casual Riders by Ride Type for Each Month



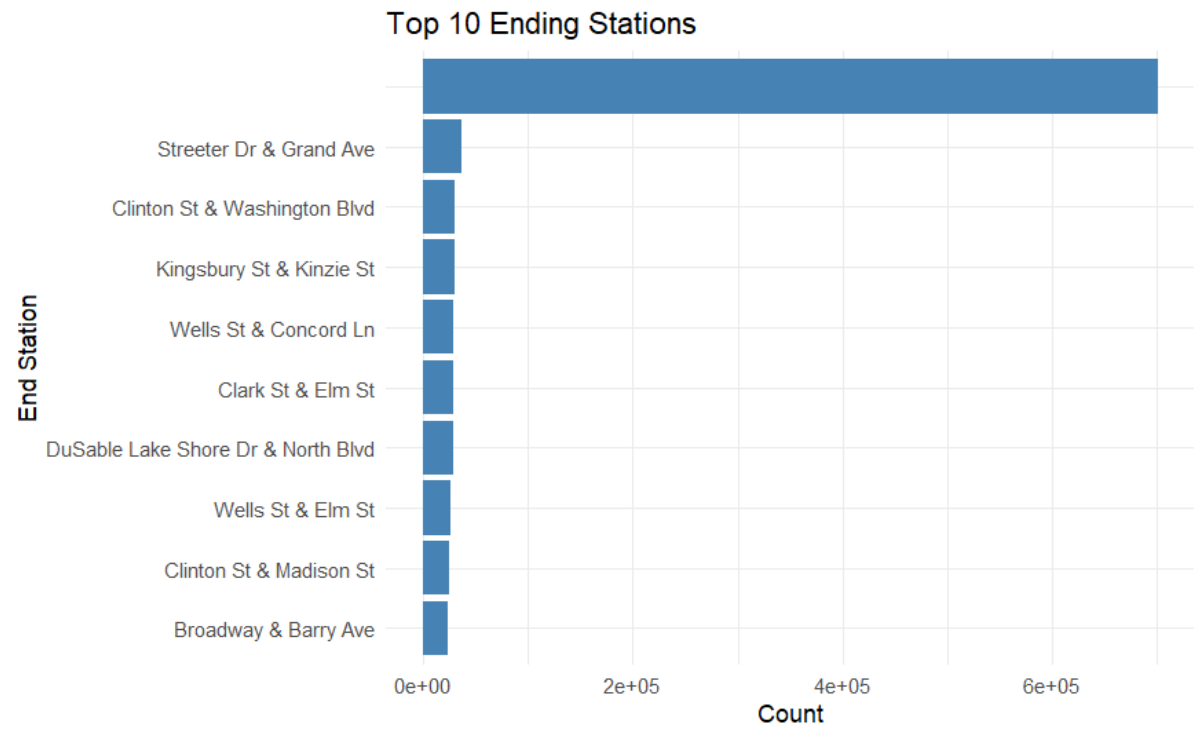
Member vs Casual Riders by Seasons



Top 10 Starting Stations



Top 10 Ending Stations



CLUSTERING

Techniques:

- Utilized Elbow Method, Silhouette, and Gap Statistic to determine optimal cluster numbers.

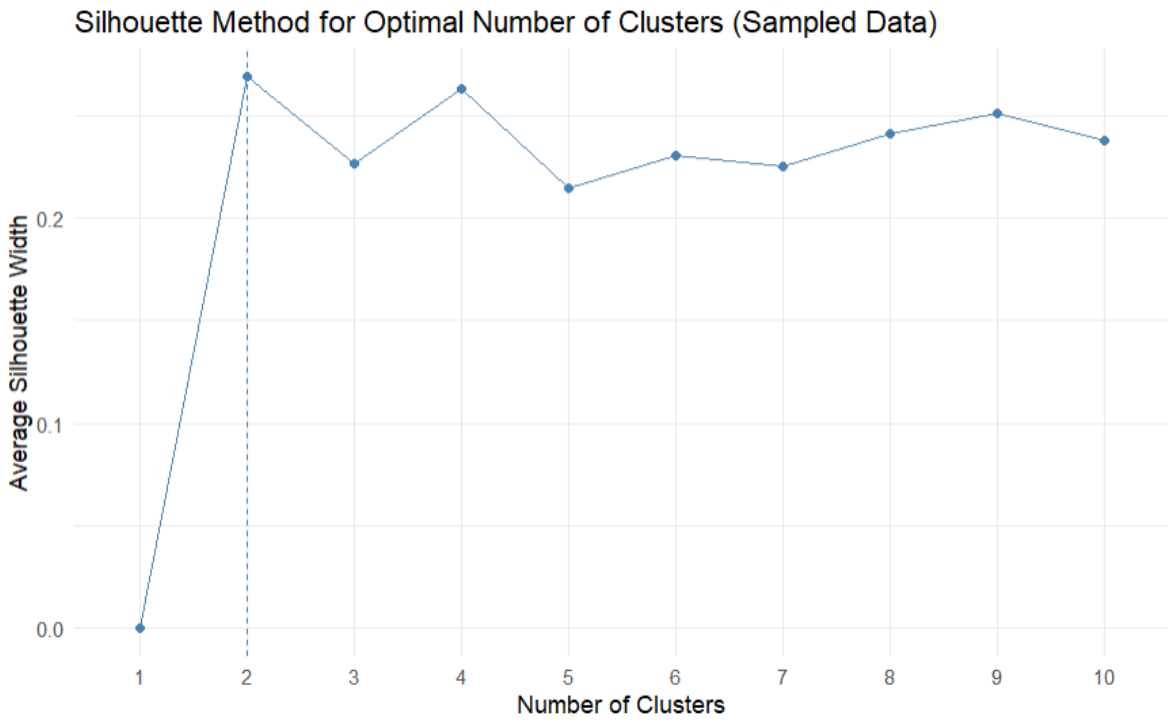
Results:

- Identified 3 key clusters, each representing distinct seasonal and time-based usage patterns.

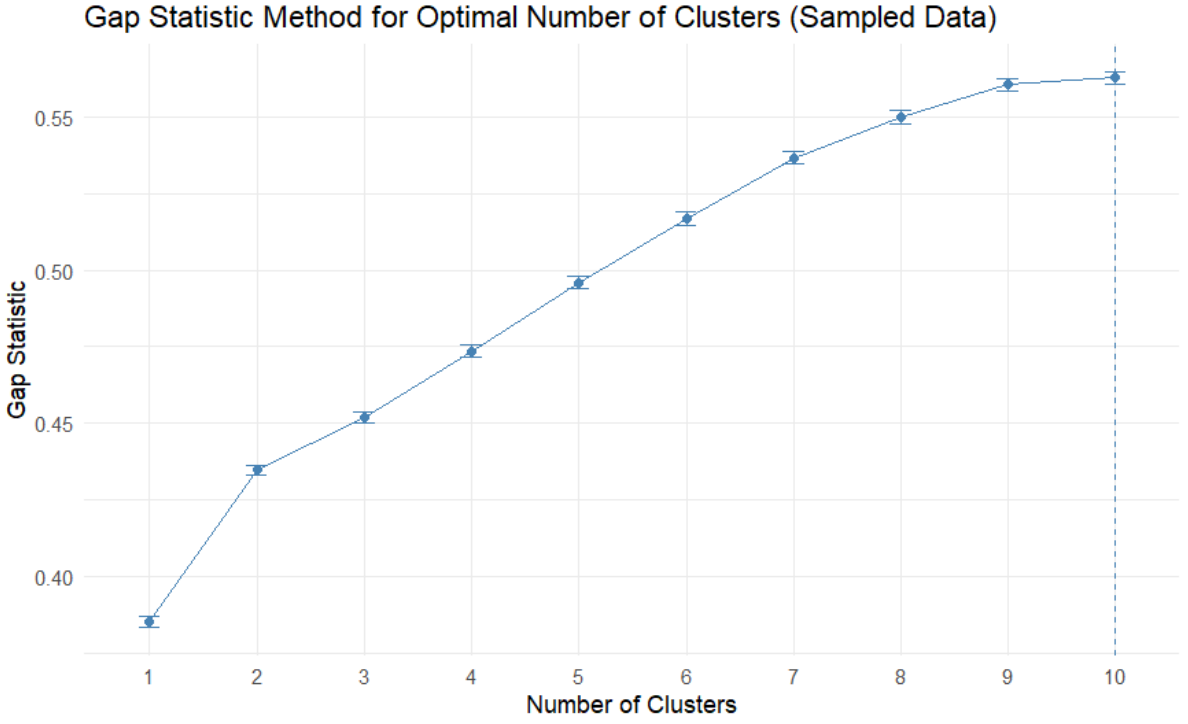
Visualizations:

- Silhouette, k-Means, and Gap statistic method.

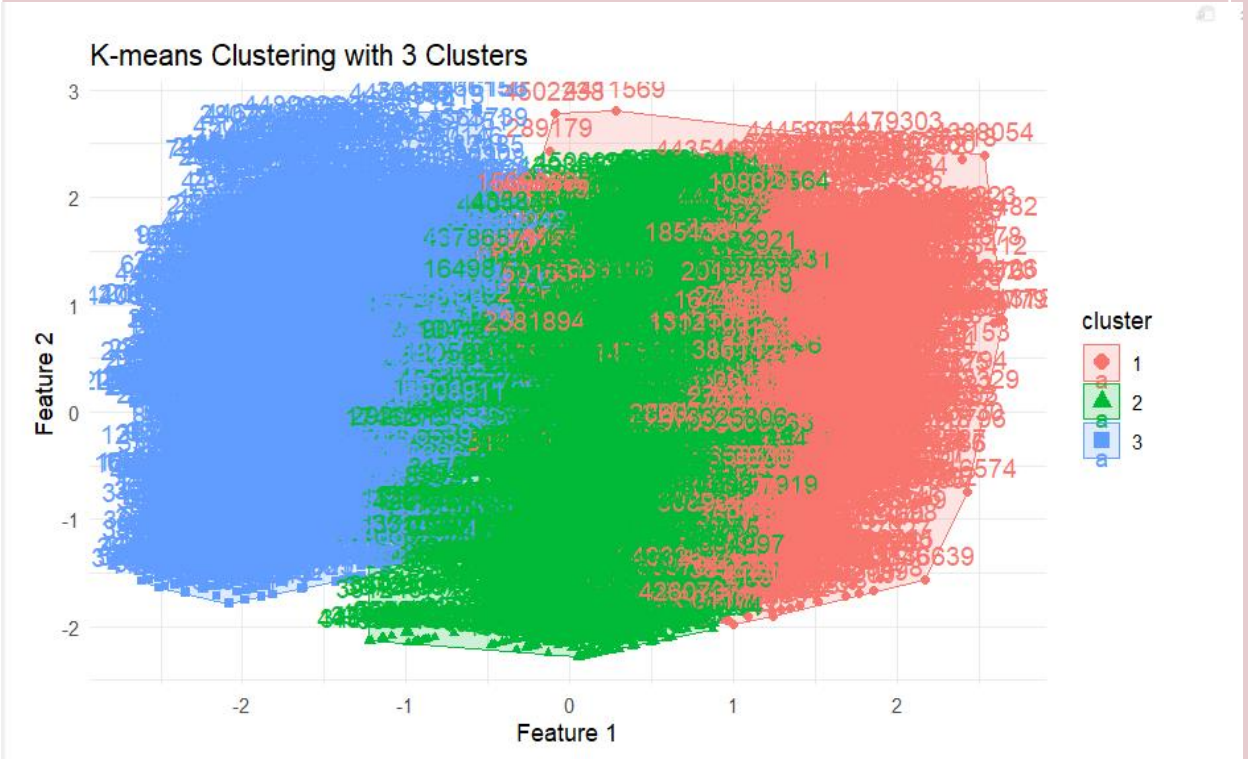
Silhouette Method



Gap Statistics Method



K Means Method



FEATURE SELECTION

Process:

- Conducted correlation analysis to remove redundant features.
- Applied Recursive Feature Elimination to select key predictive variables.

Tools:

- Leveraged data.table for efficient data manipulation and processing.

MODEL TRAINING

Selected Models:

- Random Forest: Excelled in handling complex datasets with high accuracy.
- XGBoost: Proven for classification tasks with robust predictions.

Outcome:

- Both models effectively captured key predictors of trip duration variability.

MODEL OUTPUT ANALYSIS

Random Forest:

- Achieved an impressive accuracy rate of 99%, effectively categorizing trip durations.
- Confusion matrix revealed minimal misclassification, highlighting model precision.

XGBoost:

- Delivered excellent sensitivity and specificity, accurately differentiating trip durations.
- High agreement in predictions, evidenced by kappa statistic.

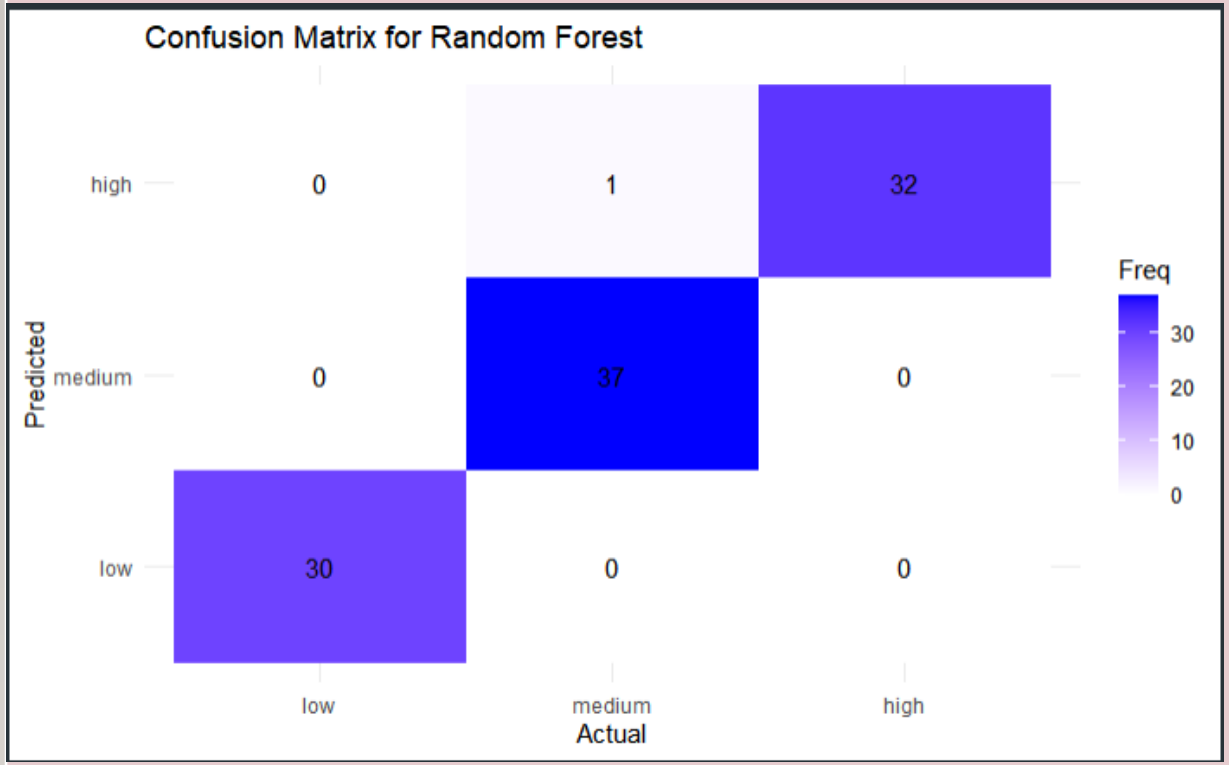
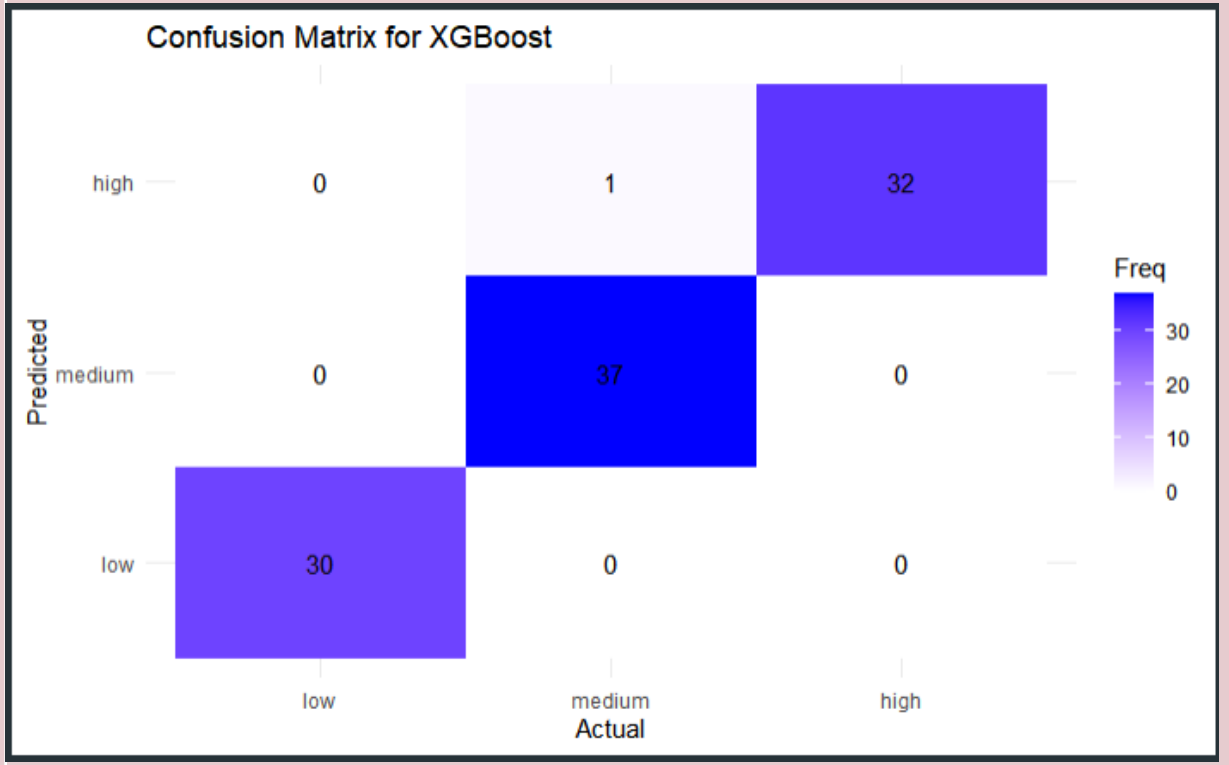
MODEL COMPARISON

Metrics Evaluated:

- Random Forest demonstrated strong R-squared values and high accuracy.
- XGBoost provided consistent predictions with minimal errors.

Visualization:

- Displayed confusion matrices to compare model performance.

Random Forest	XGBoost
	

CONCLUSION

- We successfully analyzed Divvy bike usage patterns and employed machine learning models to predict trip durations.
- Our analysis revealed critical insights into trip duration variability, enhancing strategic planning for bike-sharing operations.
- Future directions include integrating additional data sources to further refine predictive accuracy.

FUTURE SCOPE

- Integration: Explore synergy with public transit data to optimize Chicago's transportation network.
- Demand Analysis: Correlate Divvy usage with population density for targeted service improvements.
- Expansion: Apply methodologies to other cities for comparative analysis.
- Real-Time Data: Incorporate dynamic data feeds for real-time recommendations.
- User Experience: Enhance service based on user feedback and sustainability metrics.

ACKNOWLEDGEMENTS

- Heartfelt appreciation to the professor and TA for continuous support
- Team Contribution: Collaborative efforts led to project success and meaningful insights.
- Future Aspirations: Aim to continue exploring urban mobility solutions through data-driven approaches.

THANK YOU

HAVE A WONDERFUL DAY!