STAT 480 Final Project

Raghav Goyal (raghavg4@illinois.edu) Brian Lim (brianl8@illinois.edu)
Daniel Alonso Gonzalez (da39@illinois.edu)

2024-12-15

Introduction

Project Definition

In today's urban landscapes, shared mobility solutions like bikeshare programs play a critical role in enhancing transportation efficiency, reducing environmental impact, and fostering healthier lifestyles. The *Chicago Divvy Bikeshare Program*, operated under Lyft and owned by the Chicago Department of Transportation, is a cornerstone of of the city's urban mobility infrastructure, offering an extensive network of stations and bikes to support diverse commuter needs. By analyzing data from this program, we aim to uncover actionable insights that can optimize bikeshare operations and enhance user experiences.

Project Objectives

This project leverages Divvy data spanning a full 12 months (November 2023 to October 2024), which provides comprehensive information on bike rides, station details, and user demographics. The primary objectives are: - *Optimizing Station Placements and Bike Availability - Member Retention and Engagement*

Data Preprocessing, Feature Engineering, and Initial Analysis

We have loaded 12 month historic data ranging from November 2023 to October 2024. Each of these CSV files had over 100,000 observations and had sizes ranging from 25MB to 160MB. After combining, let's get an idea of the overall dataset.

```
## [1] "1.9 Gb"
```

Looks like we are working with just under 2GB worth of data, which is definitely a large dataset in terms of both memory/storage, as well as its potential computational expenses.

The overall dataset has almost 6 million rows across 13 distinct columns. On first glance, we can see that a decent chunk of start and end station names and ids are empty and a few values for ending longitude and ending latitude are missing. To get up and running, we can do some quick cleaning to make the data a bit more manageable. This includes dropping the initial ride id column as it contains no discernible information, changing certain categorical columns into factors, and changing our start and end trip times to a proper datetime format to extract as much information.

```
##
##       2023 2024
##   Jan    0  288
##   Feb    0  318
##   Mar    0  396
##   Apr    0  479
##   May    0  824
##   Jun    0 1141
##   Jul    0 1137
##   Aug    0  978
##   Sep    0  703
##   Oct    0  564
##   Nov  350    0
##   Dec  239    0
```

It seems the 7417 missing values in the ending coordinates are pretty spread out across each month, since this is a fairly negligible amount compared to the near 6 million observations, we can simply drop them. Especially, considering that these coordinates are important pieces of information with no real potential way of imputing.

Let's take a closer look at the empty rows in the start and end station columns. These empty occurrences are spread across the time frame, so it is not the case that 1 or a few months are missing this data.

```
##
##                     casual member
##   classic_bike         240    150
##   electric_bike     571220 984626
##   electric_scooter   59475  37035
```

We can see that most empty station data are for electric bikes and scooters and this makes sense as only classic bikes need to be operated at Divvy specific docking stations and e-bikes and scooters are eligible to be parked at any public racks for free and other legal public locations for a small cost. Since the values for the classic bike are actually "missing" we will drop these. Also, there are no instances where station id contains information and station name does not or vice-versa and there is no discernible information for station id so we will drop the two corresponding columns for this.

For the rest of the station NAs, we can fill them up by using a two-element tuple containing the longitude and latitude of of their start and end trip. Dropping these values would not be advisable as they constitute the majority of total observations containing e-bike and scooter specific data.

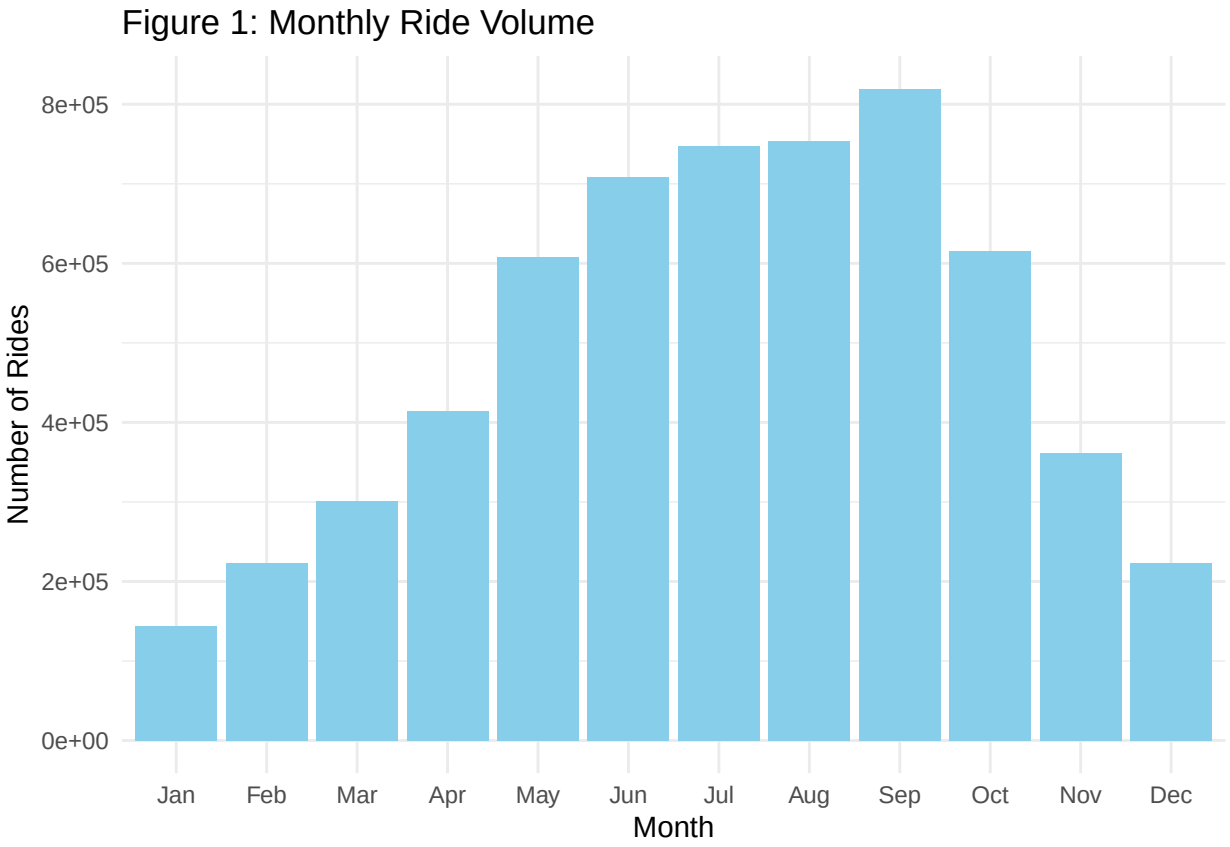
The data we currently have has the trip start and end time down to the second. By finding the differences between these two columns, we can create a new feature to represent the trip duration in minutes. To ensure data quality we will also filter out "negative" trip durations or overly long times like over 500 minutes long.

Let's also create a few additional time-based features. These include, extracting the specific months, day, and hours of the rides. We can also create a entirely new feature called part of the day, where we segment mornings, afternoons, evenings, and nights. Below is a quick preview of these new features.

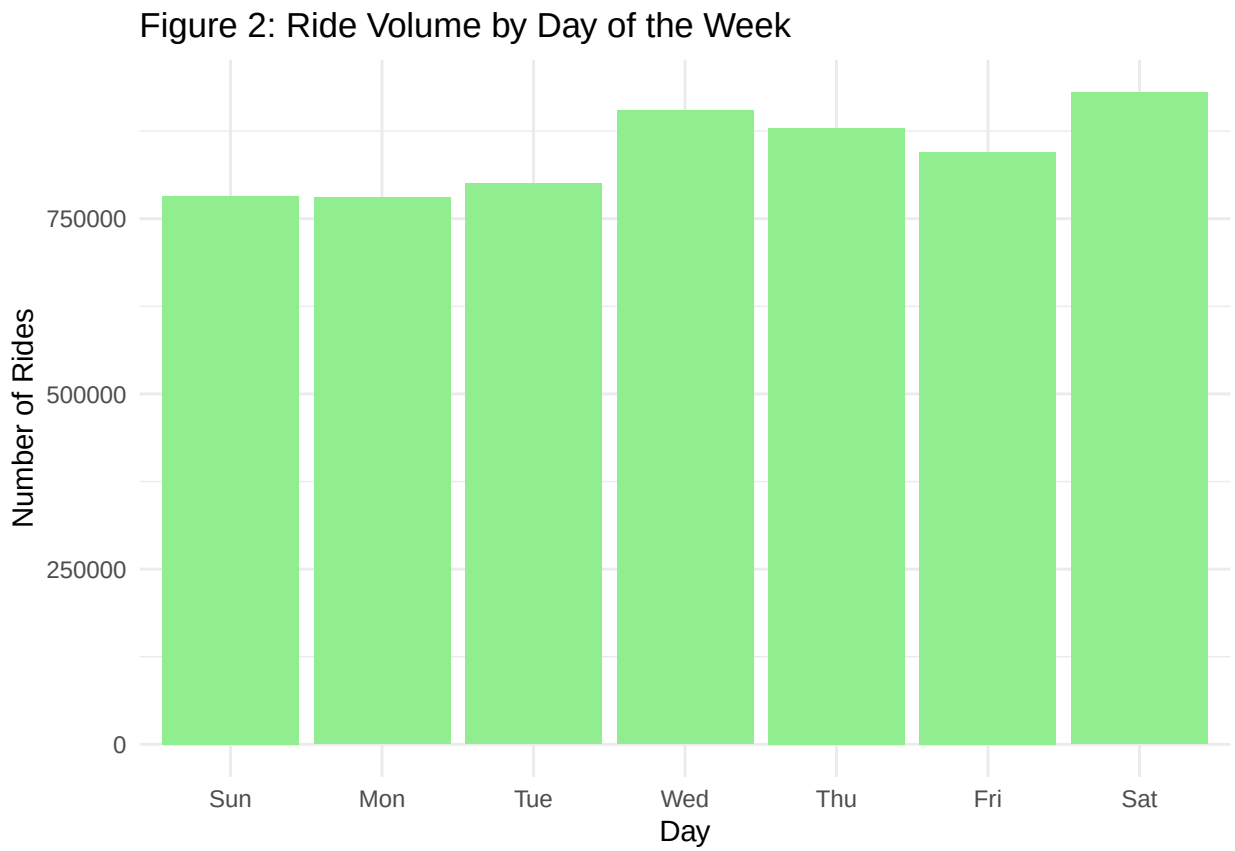
```
##            started_at month day hour part_of_day
## 1 2023-11-30 21:50:05   Nov Thu   21       Night
## 2 2023-11-03 09:44:02   Nov Fri    9     Morning
## 3 2023-11-30 11:39:44   Nov Thu   11     Morning
## 4 2023-11-08 10:01:45   Nov Wed   10     Morning
## 5 2023-11-03 16:20:25   Nov Fri   16   Afternoon
## 6 2023-11-30 16:15:53   Nov Thu   16   Afternoon
```

Further Exploratory Analysis and Visualization

We have completed a fair amount of data cleaning and feature engineering to this point. Let's now get into the crux of the data and explore further and generate some useful visualizations and potentiall actionable insights.

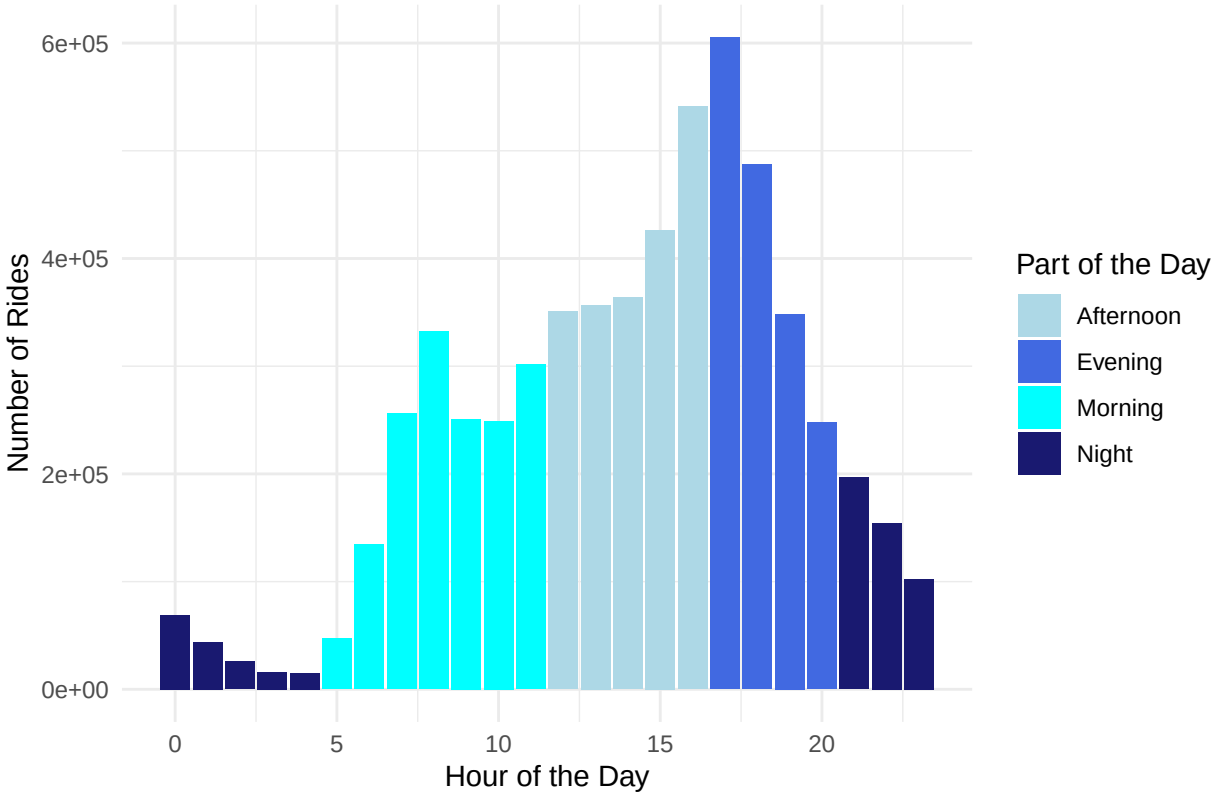


In the entire 12-month period we have 5.9 million trips, and we can see that most of these trips occur in the summer months, with some decent amount in the fall and spring as well. It is clear to see that winter trips are not as popular which makes sense due to the harsh conditions we see in Chicago during that time.



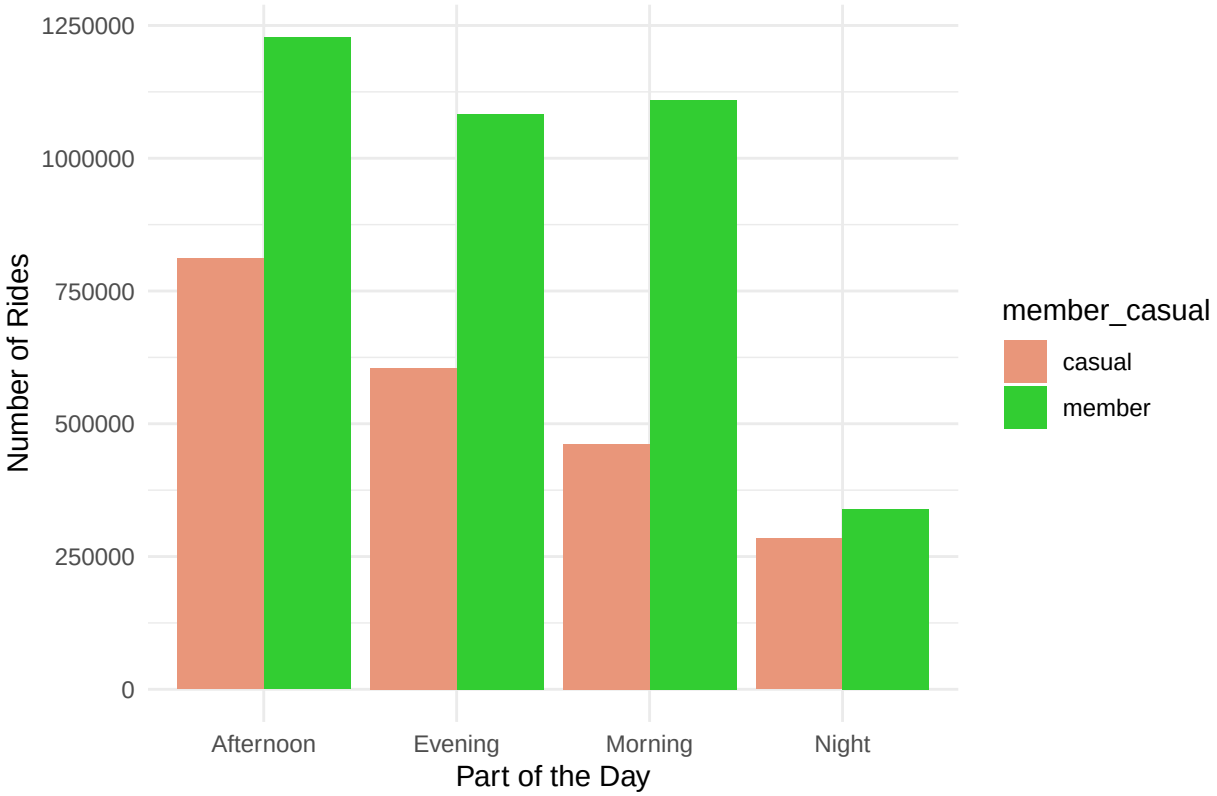
Rides are extremely consistent throughout the week. It may be the case that during the weekdays they are used for peak commute (school, work, etc), while on the weekends they can be used for leisure.

Figure 3: Ride Volume by Hour of the Day and Part of the Day



Most rides occur during the late afternoon and early evening portions of the day, from around 3PM to 6PM. This makes sense as this time is also associated with peak commute and is typically considered "pleasant" on off-days around the city. There is still a decent amount of rides throughout the day, although this undertandibly drops significantly during the dead of night.

Figure 4: Ride Volume by Part of the Day and User Type

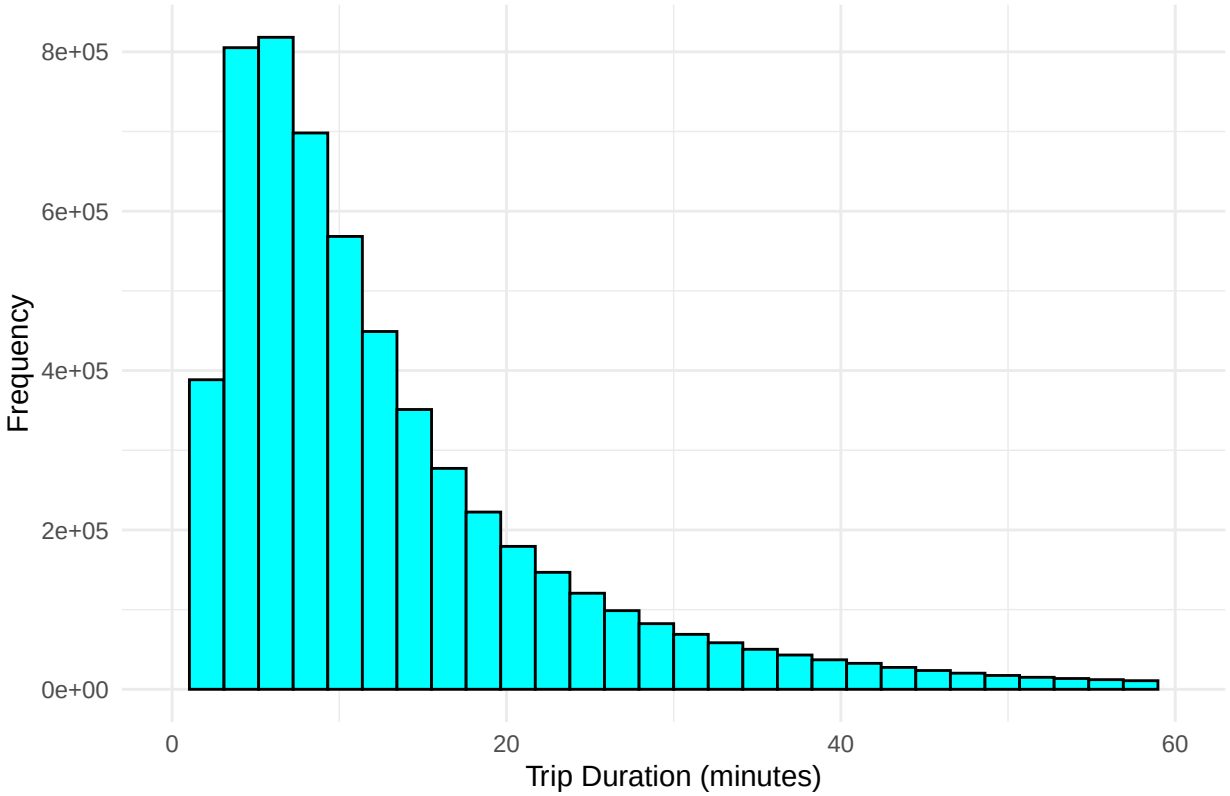


For overall trips, we have around 63% of them being from members and the remaining 37% of them being casual or single-use riders. Pretty much across the board time-wise, we see a consistent trend of member riders being more frequent than casual, however, this difference is less for night rides. Let's check how long these rides typically are.

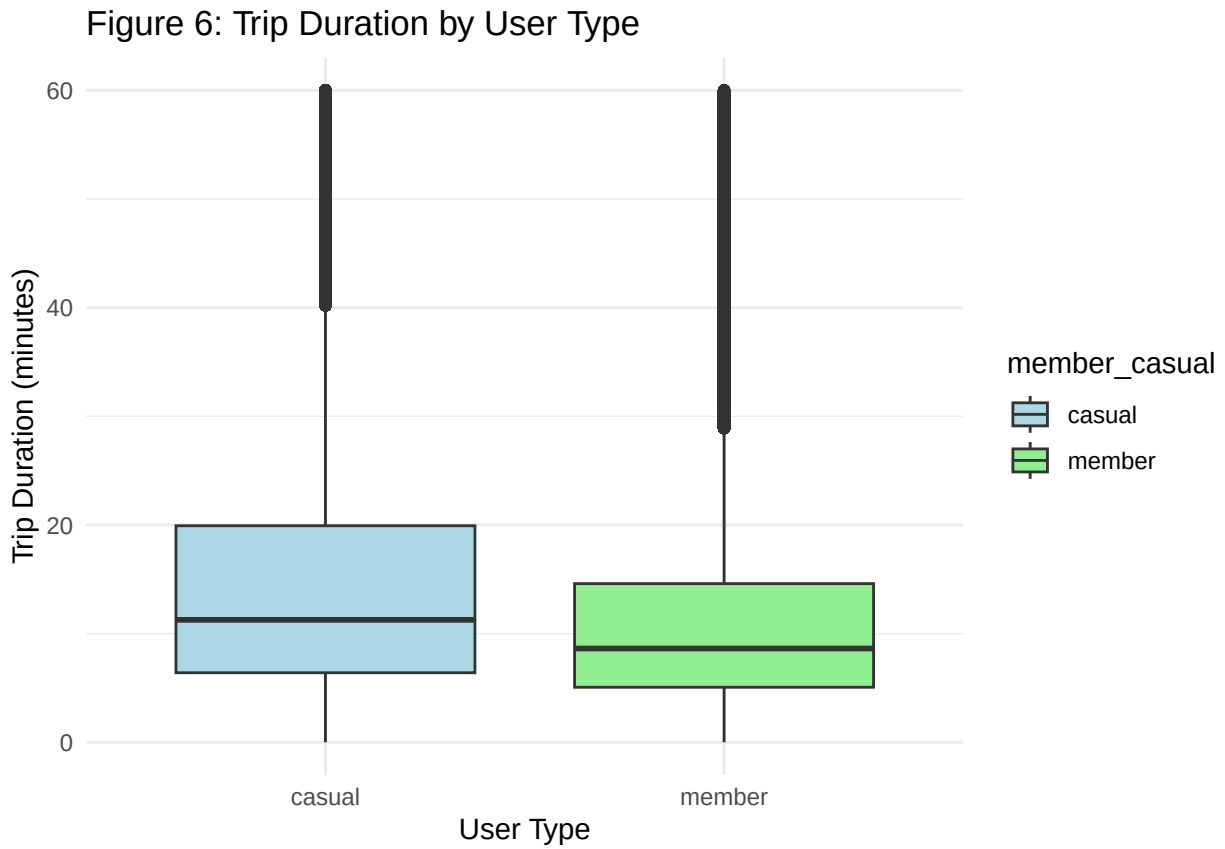
```
## [1] 143097
```

There are 143 thousand trips that are longer than 1 hour. This means that the vast majority of trips are within the 1-hour time frame, so let's check the distribution of trip length from 0-60 minutes.

Figure 5: Distribution of Trip Durations



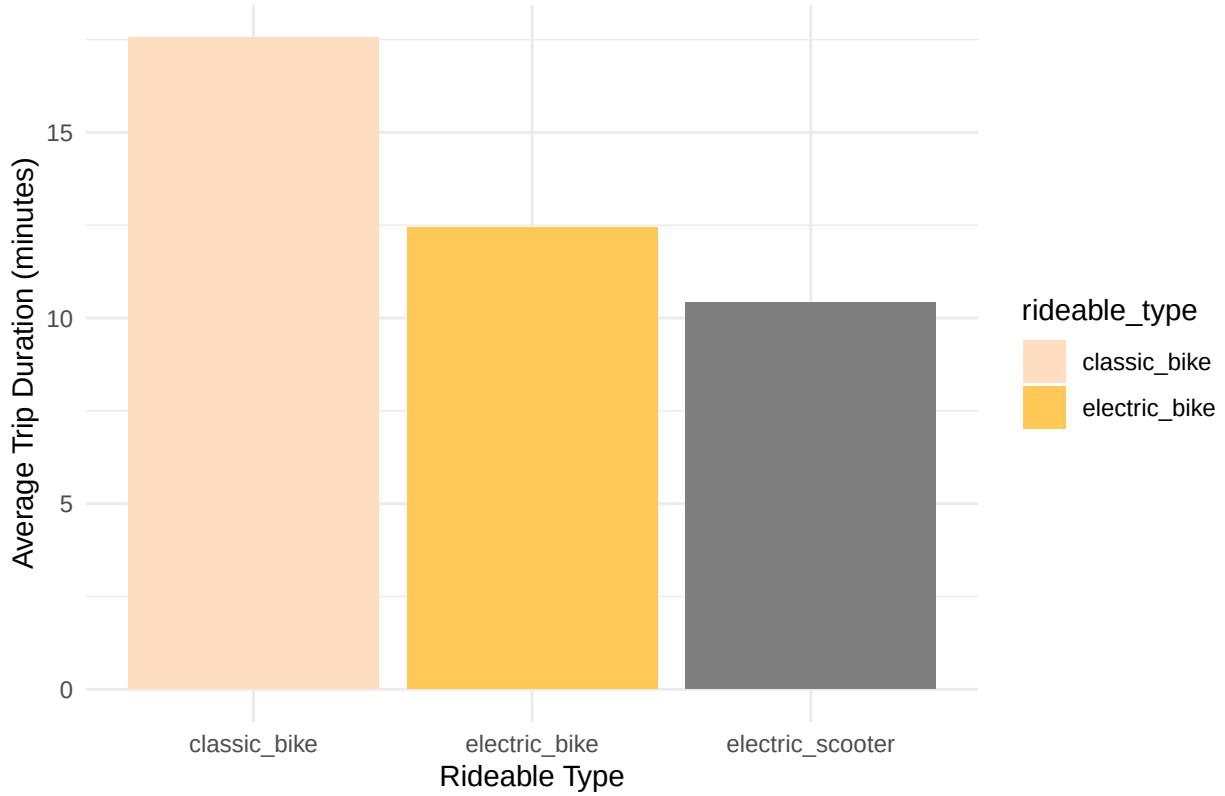
In general, most trips are fairly short and run from 2-15 minutes. This is expected as people likely use these vehicles for short, convenient commutes across the city.



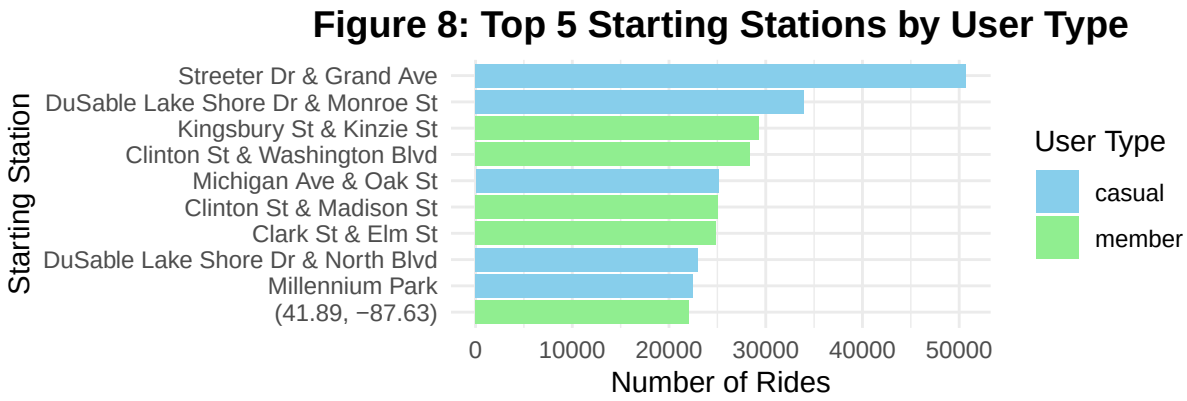
Casual riders are more likely to have longer trips. This may be the case due to pricing. Members do not have to pay an unlocking fee and variable rates are lower, so members are not subject to greater costs upfront to deter them from shorter rides. Single-use riders however, may feel the need to get the most out of their unlocking fee and opt for longer rides.

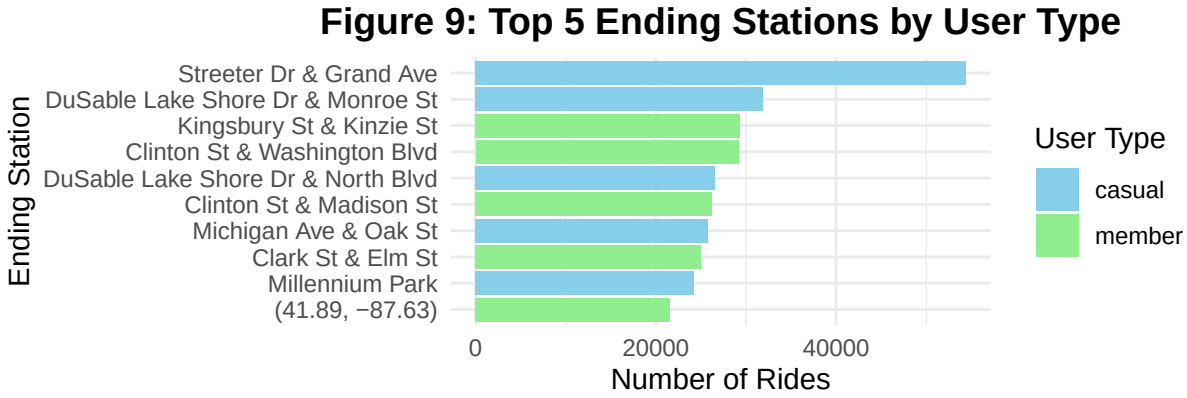
Classic bikes and electric bikes take up more than 95% of total ride volume, but let's see if there are any differences in trip duration for each ride type.

Figure: 7 Average Trip Duration by Rideable Type



It seems trips with classic bikes have longer trip times compared to the e-vehicles. This could be so due to the e-vehicles having higher operational speeds, leading to a smaller duration. Let's now check out what some of the most popular starting and ending docking stations are.





From figures 8 and 9 above, we can see that a lot of the starting and ending stations are same for both casual riders and members. Most of these are official docking stations like "Streeter Dr & Grand Ave", however there is one very popular general location of (41.89, -87.63) which happens to be in the River North Area where the Courthouse Place is. This is nationally recognised as a historical landmark which may point to its popularity.

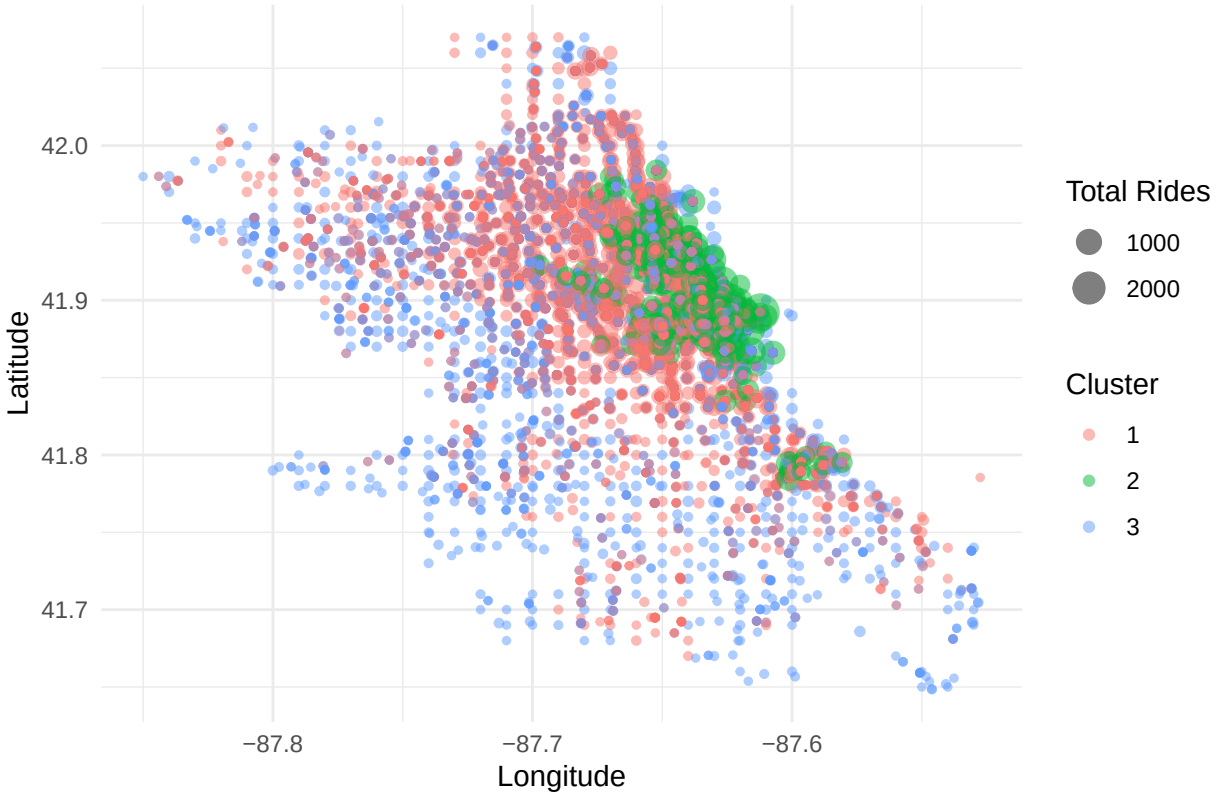
Modeling

We have been able to gather a lot of useful information and learned more about the data we are dealing with. However, it is time to move into some more advanced analytics and modeling. After all, we aim to optimize station placements and bike availability, as well as understand membership vs casual riders.

Cluster Analysis

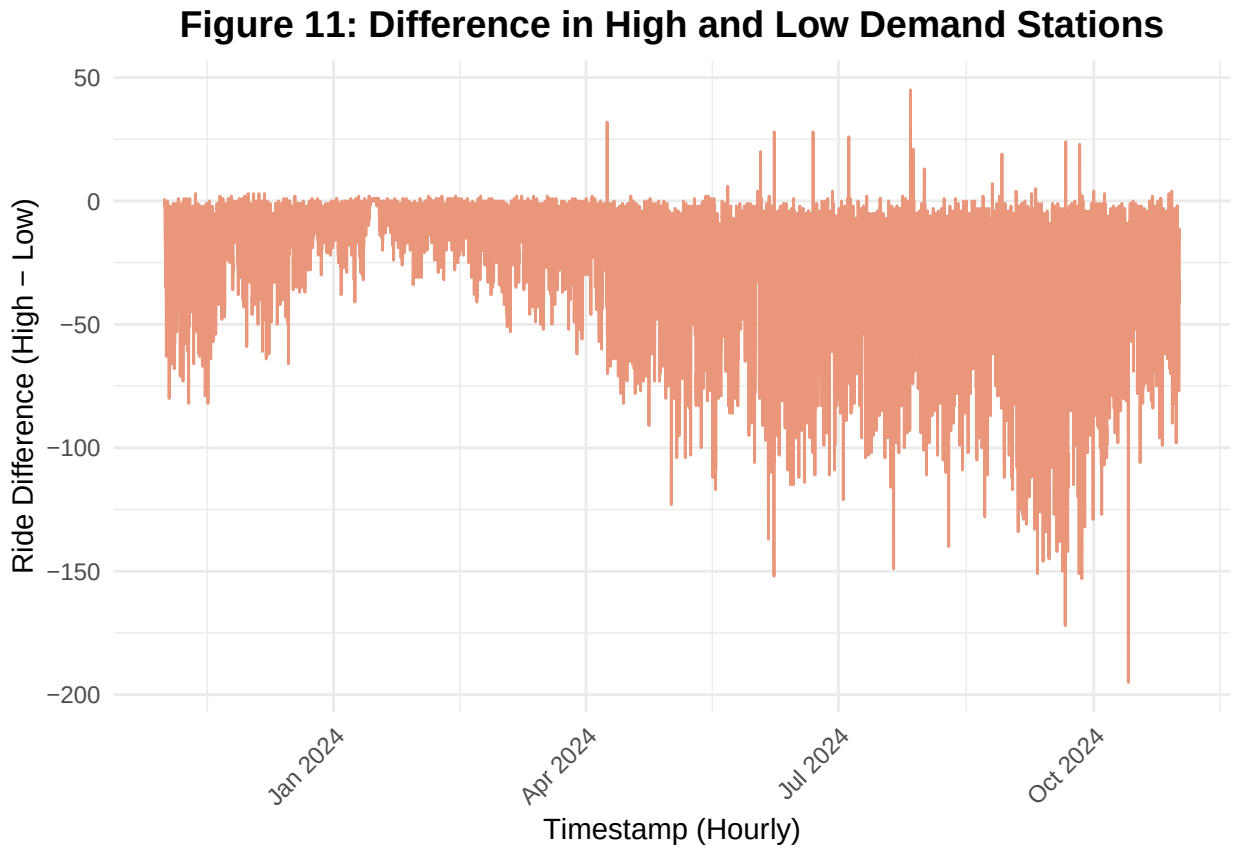
We will first cluster rides based on station location and usage. Given the dataset size, clustering on the entire dataset will be very computationally expensive, so instead, we can create a stratified sample of 5% to ensure stations are still well-represented.

Figure 10: Clustered Stations Based on Usage Patterns

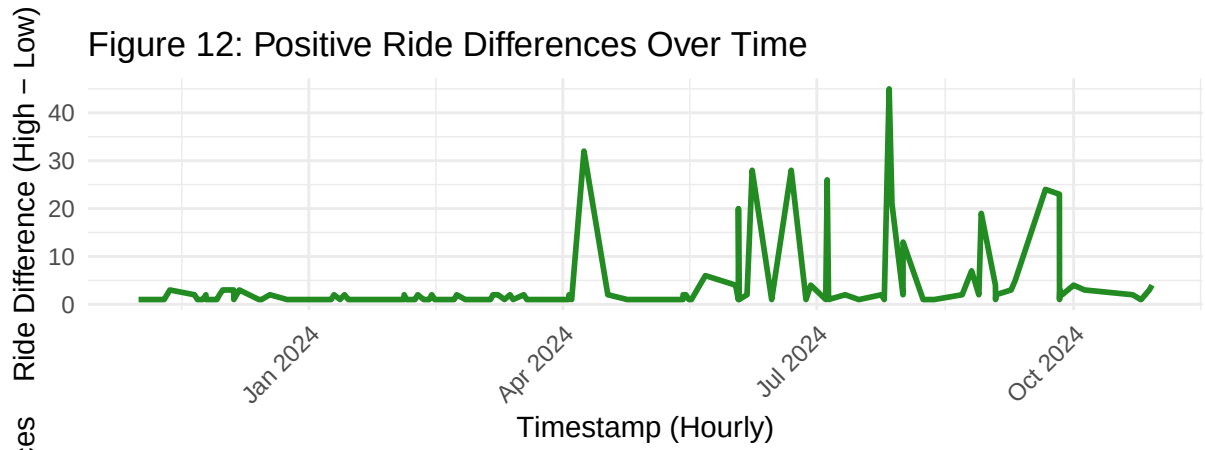


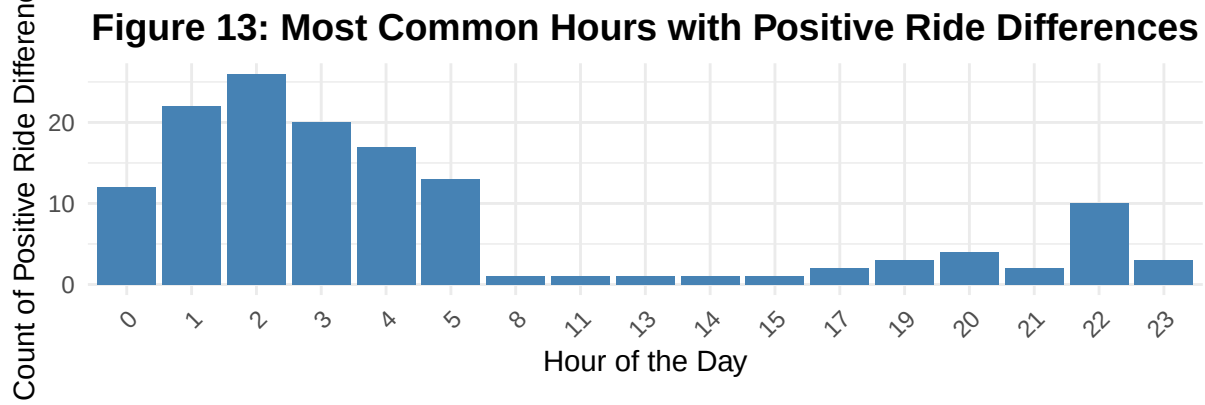
In Figure 10 above, we have clustered stations based on usage patterns like total rides, average trip duration, and member ratio. There is one cluster that covers stations more widespread geographically, covering areas both in the periphery and less dense areas. Another cluster dominates the plot with a heavy concentration in mid-density areas. The last cluster is highly concentrated in downtown Chicago and represents high-demand stations.

The high-demand stations in concentrated cluster are critical for operational efficiency and user satisfaction. It would make sense to increase bike availability during peak hours and to meet demand and other potential redistribution depending on temporal patterns, considering these are starting stations. We can focus on the demand differences between Cluster 3 and 1, and keep Cluster 2 as is since it acts like an intermediary between the two.



From Figure 11 above we can see that there are typically more rides in low-demand stations as compared to high-demand stations. This might seem counter-intuitive at first, however, it is important to note that there are many more low-demand-stations than high-demand. There are still around 300 or so observations that fall in the positive range and there are few large spikes as well, which are essential to optimize vehicle redistribution.



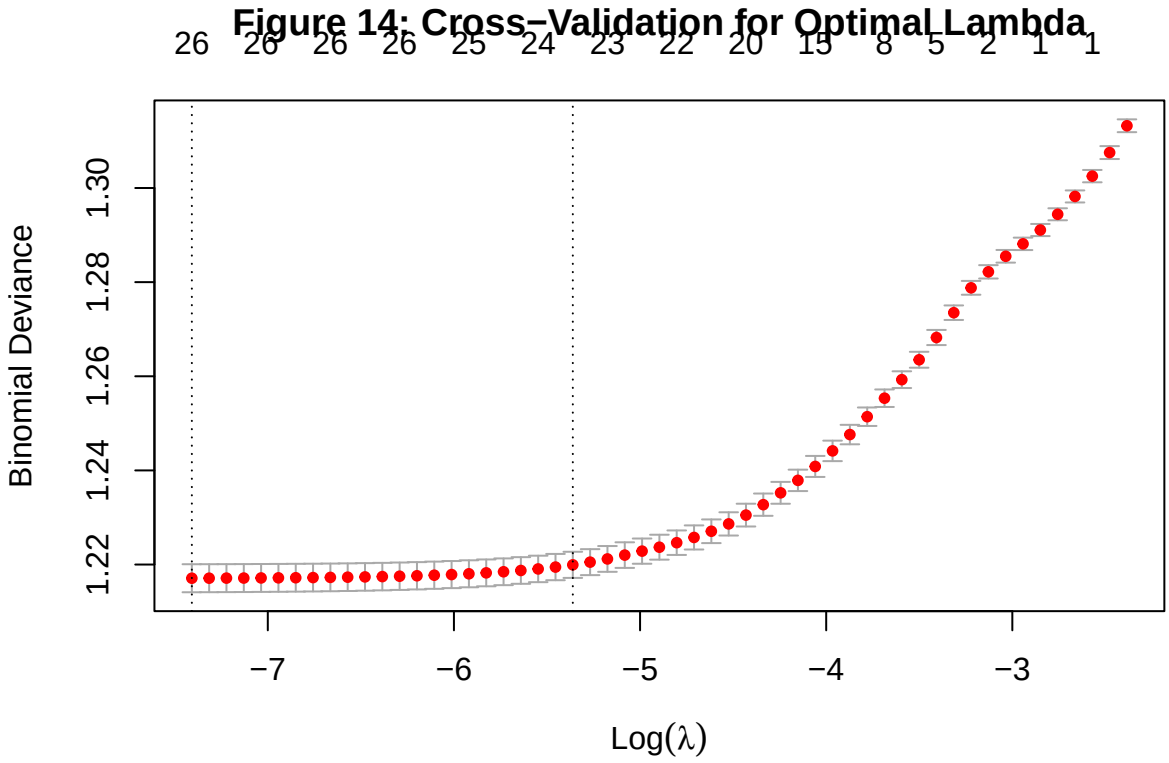


From Figure 12 and 13 it is clear to see that there are imbalances in high-demand areas during the warmers months in the night time. We can infer that low-demand stations are popular for people to commute for school and work from largely locations in the city outskirts. The flow of rides seems to be balanced in that aspect, however, during warmers months in early mornings and late nights, demand for rides in the downtown areas pick-up despite the relative low "overall rides" during these times as we saw back in Figure 3. It makes sense as the lower overall activity in these times would not encourage a a proper redistribution of vehicles.

As a side note, a time series forecasting model was trialed to fit both the values for high-demand stations and the differences between high-and-low using an ARIMA. We received some decent-enough results, however, the the plot of residuals in the ACF and PACF still indicated a lot of seasonality and cyclical elements that would have required significant tuning to generate more accurate forecasts - which at this point of time is out of the scope of our project.

Classification

Let's switch gears and understand what goes into the user preference of being a member rider against being a casual rider. To do this we will use a LASSO path logistic regression to predict whether a ride was a ridden by a member. We will use the vehicle type, our clusters of low, medium, and high demand stations, trip_duration, and month, day, and part of the day as our initial features. Again, due to computational limits, we will not use the entire set of 6 million observations, but use 80,000 for training and 20,000 for testing. The LASSO path will also help alleviate burden on the predictor aspect by performing feature selection.



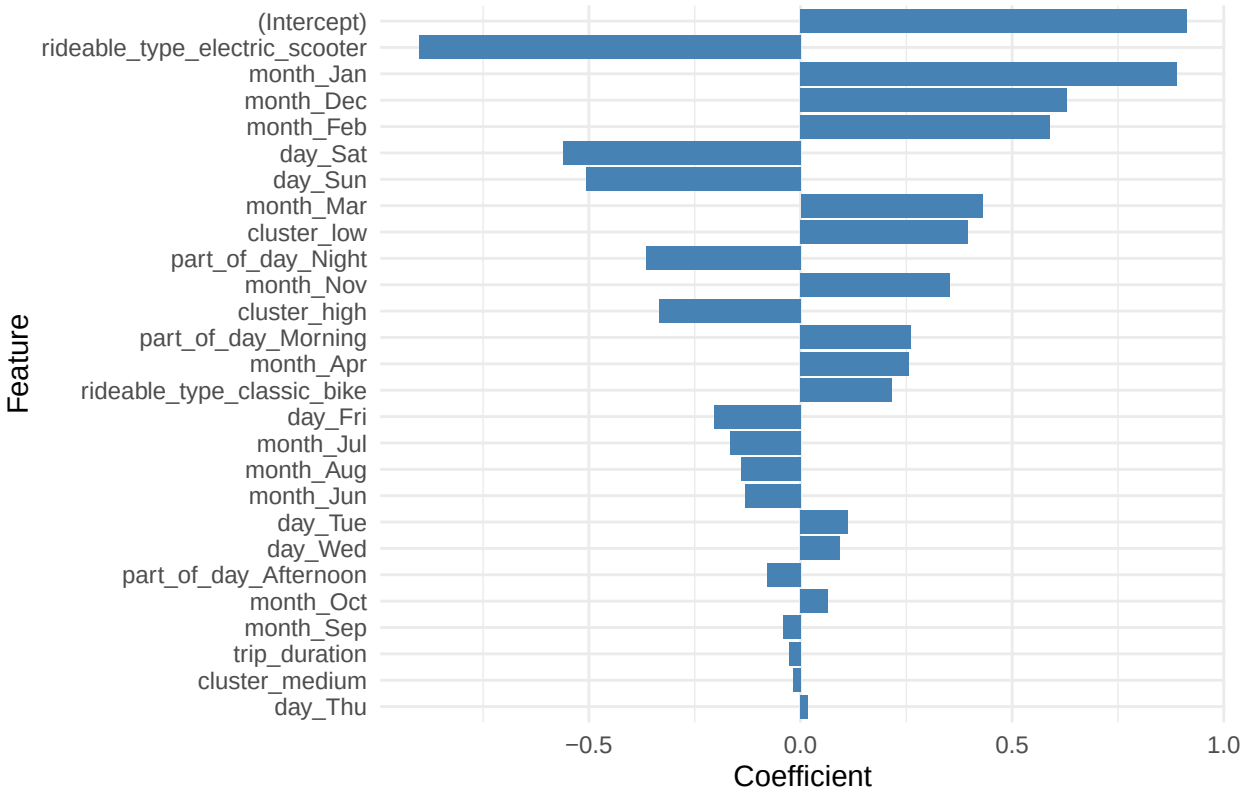
Using cross-validation, we have extracted an optimal lambda that minimizes deviance. This value is around 0.00007. Let's use this best model to generate the predictions.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0  1997  1200
##          1  5294 11509
##
##                Accuracy : 0.6753
##                  95% CI : (0.6688, 0.6818)
##     No Information Rate : 0.6354
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.2039
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.27390
##             Specificity : 0.90558
##          Pos Pred Value : 0.62465
##          Neg Pred Value : 0.68494
##              Prevalence : 0.36455
##          Detection Rate : 0.09985
##    Detection Prevalence : 0.15985
```

```
##        Balanced Accuracy : 0.58974
##
##         'Positive' Class : 0
##
```

We get an overall accuracy of around 67% which is not bad. However, we get a recall for class 0 as 25.9% which is quite low, suggesting the model struggles to predict the "Casual" class correctly. This is not the end of the world however, as we get a recall for class 1 as 90.7%. This is very high meaning the model does well at predicting members which is our primary objective. The poor performance for the negative class could be partly due to the imbalance, with it consisting of 36% of the data. If we wanted the best model for accuracy it might have made sense to utilize something like a random forest or gradient boosting model or to tackle to imbalance with re-sampling methods. Our main objective however, is to understand what features go into being a member, so let's check some of these out.

Figure 15: Feature Importance (LASSO Logistic Regress



We can see that some of the largest negative coefficients are rides with an electric scooter, or rides on the weekend. These areas are certainly areas of improvement to convert casual riders to members. It also makes sense that the largest positive coefficients relate to rides being taken in the winter months as members will be more dedicated. We can definitely takes these insights and make recommendations to Divvy on how to maximize member conversion.

Conclusion and Evaluation

Our analysis of the Divvy Bikeshare Program data uncovered significant patterns and actionable insights to optimize station placements, enhance bike availability, and understand the differences between member and casual riders.

Optimizing Station Placements and Bike Availability:

Clustering Analysis: We identified three demand clusters: -Cluster 1 (Low-Demand): Widespread, peripheral stations with lower activity. -Cluster 2 (Medium-Demand): Intermediary stations with moderate activity. -Cluster 3 (High-Demand): Downtown stations with concentrated and peak activity.

Recommendations: -Improve bike availability at high-demand stations during early morning and late evening hours. -Introduce dynamic bike redistribution based on hourly demand forecasts. -Monitor and adjust for seasonal shifts in ride patterns, particularly during summer months.

Member Retention and Engagement:

LASSO Logistic Regression: The model, trained on 100,000 samples with features like rideable_type, trip duration, time-related factors (month, day, part_of_day), and station clusters, achieved: -Accuracy: ~67% -High Recall for Members: ~90.7%, indicating strong performance in identifying members.

Recommendations: -Target casual riders using electric scooters and weekend trips with discounts or promotions for membership conversions. -Leverage the winter months as an opportunity to strengthen membership retention through targeted campaigns. -Promote high-demand stations with exclusive membership benefits like priority access or discounted fares.

Strengths: -Robust Methodology: Our combination of clustering, time-series analysis, and logistic regression provided a comprehensive view of station-level demand and user behavior. -Feature Selection: LASSO logistic regression effectively selected meaningful predictors, reducing noise and ensuring interpretability. -Actionable Insights: Clear recommendations for bike availability optimization and membership engagement were derived from our results.

Limitations:

Model Simplifications: -Time-series forecasting was limited due to seasonal and cyclical complexities. The ARIMA model required significant tuning to improve residual behavior. The logistic regression achieved moderate accuracy but struggled with the imbalanced dataset, particularly for predicting casual riders.

Data Constraints: -Our sampling (100,000 observations) was necessitated by computational limits. While representative, the full 6 million observations could provide additional insights. Unobserved factors, like weather conditions, user preferences, and special events, were not accounted for but could significantly influence ride patterns.

Station-Level Assumptions: -Clustering was based solely on aggregated ride metrics. Real-world conditions like road access, infrastructure, or user density were not incorporated.