Chicago Transit Authority Data Analysis

Connecting Chicago, One Ride at A Time

Submitted by:

Pranit Kotkar

Vaishnavi Shankar

Siddhi Shukla

Rewa Deshpande

Anushka Chaubal

GOAL

Analysis of Boarding Passengers on Transit Routes:

- Analyze boarding patterns to identify peak hours, popular routes, and areas with high demand.
- Insights from analysis can help CTA to optimize its transit service and improve customer satisfaction.

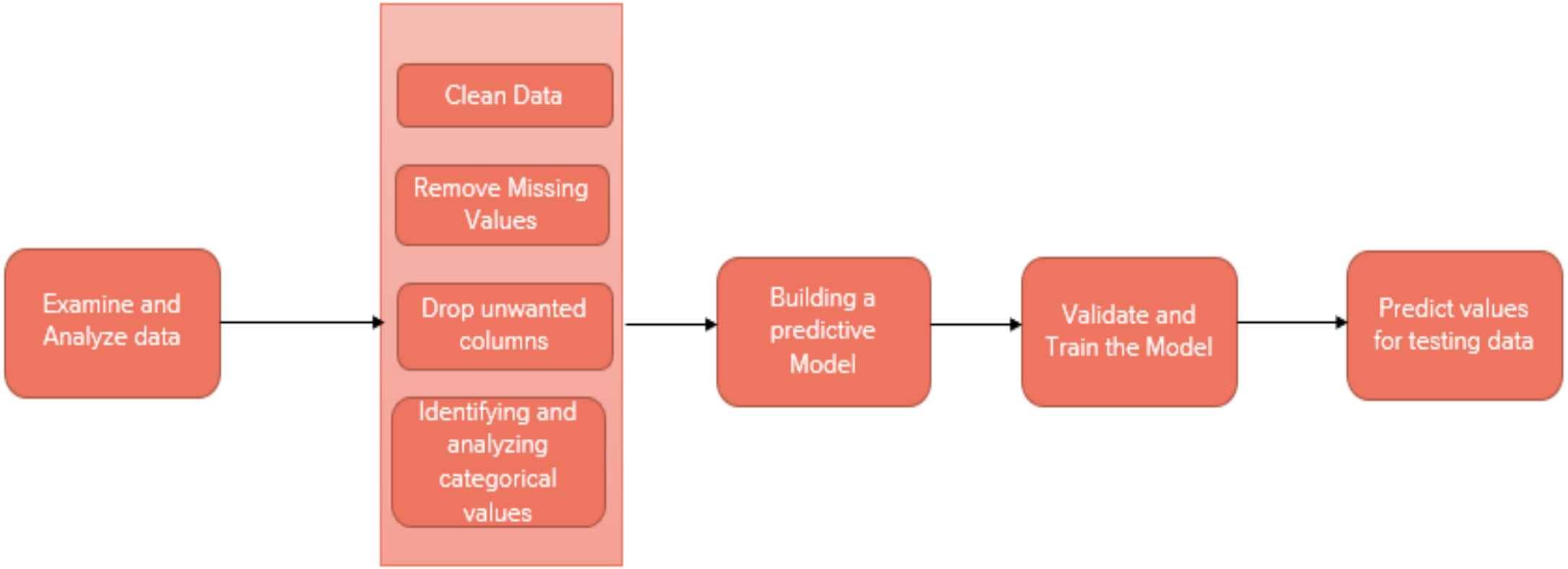
Factors Impacting CTA Usage:

- Identify factors that influence ridership, such as weather, events, and socioeconomic changes.
- This analysis can help CTA to anticipate changes in ridership and adjust service levels to meet customer demand.

Recommendations for CTA:

- Based on analysis, provide recommendations to improve service quality and increase ridership.
- The recommendations should focus on enhancing the overall customer experience, improving service efficiency, and addressing any identified issues or challenges.

Proposed Methodology



Exploring Data - Data Source

- **CTA Bus data:**

CTA-Ridership-Bus-Routes-Monthly-Day-Type-Averages

  - The Month_Beginning column was changed to the date format.
  - NA values were eliminated.

- **CTA L-Train data:**

CTA-Ridership-L-Station-Entries-Daily-Totals

CTA-Ridership-L-Station-Entries-Monthly-Day-Type

  - Changed the month_beginning column to date format, specifically changed to m/d/yyyy format.

- **CTA Daily Ridership data:**

CTA-Ridership-Daily-Boarding-Totals

Exploring Data - Data Source

- **Weather data collected from National Centers for Environmental Information:**

past-weather-Chicago

  - Some TAVG values were missing. Calculated the TAVG value by taking the average of TMAX and TMIN.
  - Replaced NA values in PRCP, SNOW and SNOWD with 0.

- **Holidays - US Holiday Dates (2004-2021):**

Holiday

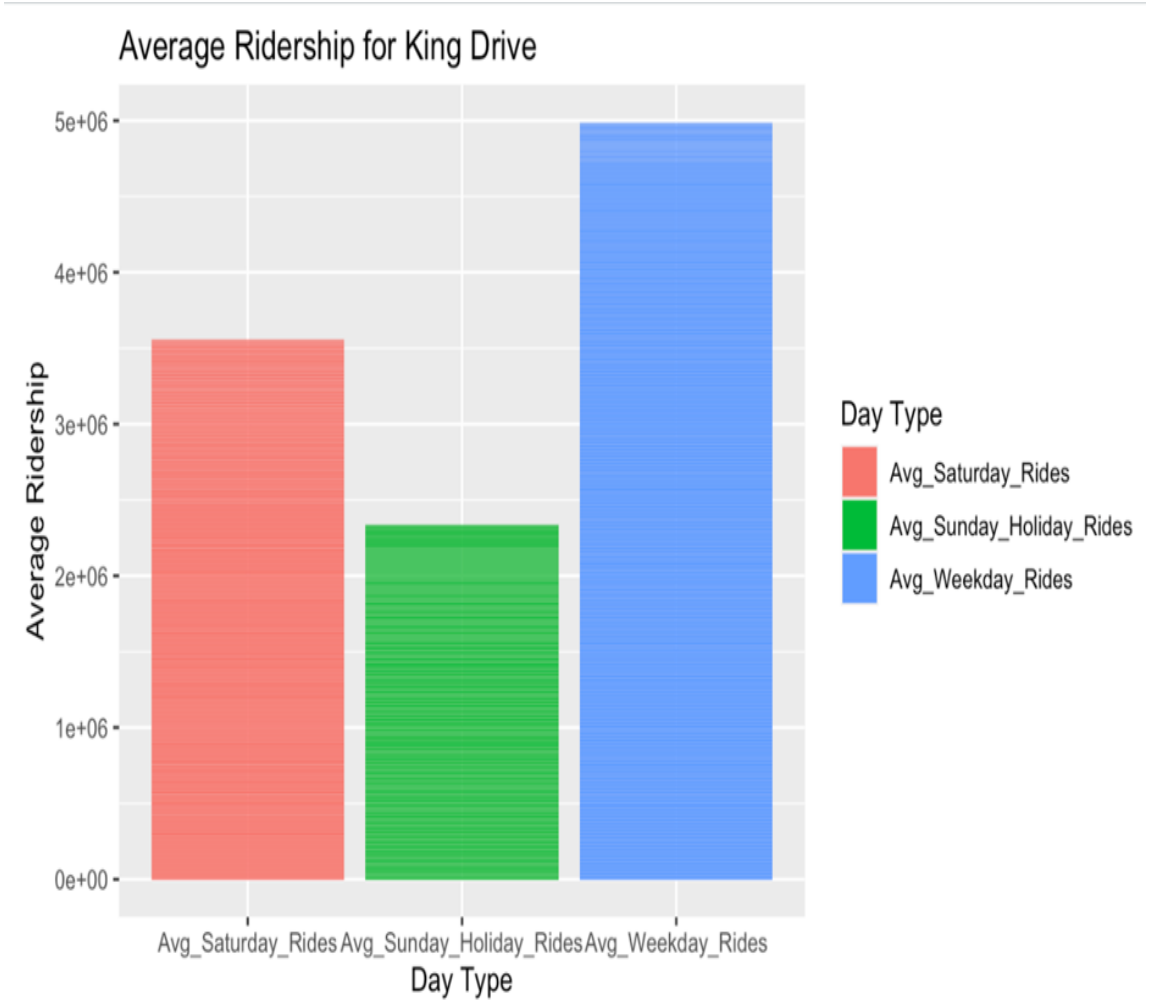
- **Crime data collected from Crimes - 2011:**

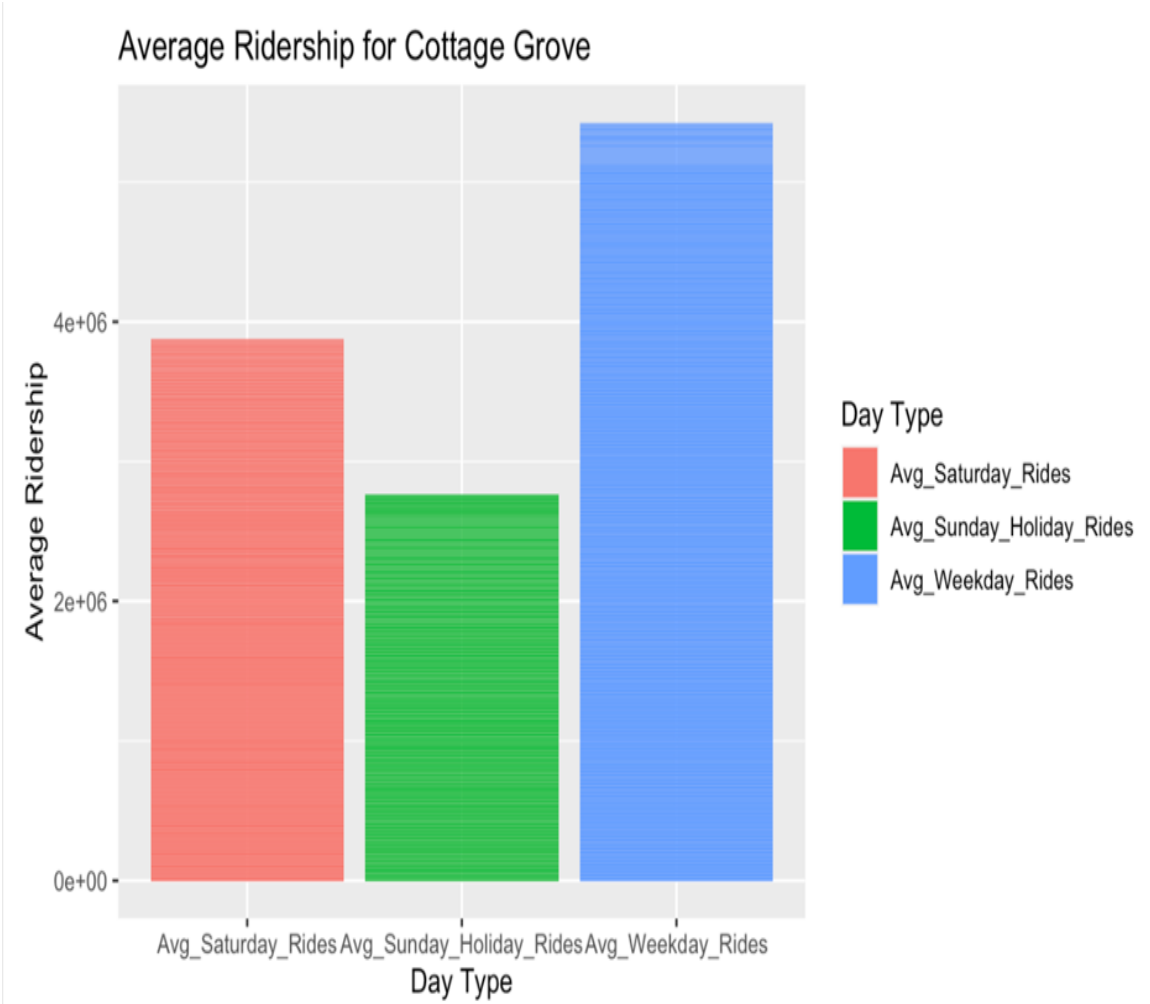
CTA-Crime

  - The columns X.Coordinate, Y.Coordinate, Latitude and Location were eliminated.
  - The Date column consisted of both the date and time of the crime. This is split into two different columns, Date and Time.
  - Splitting caused the dates to appear in the format MM/DD/00YY. This has been changed to MM/DD/YYYY and the column type is changed from character to date.
  - The Time column is changed to 24-hour format using ITime.
  - A new column ActiveOrInactive is added to the dataset which describes whether the crime occurred during the active or inactive hours of the day, based on the Time.

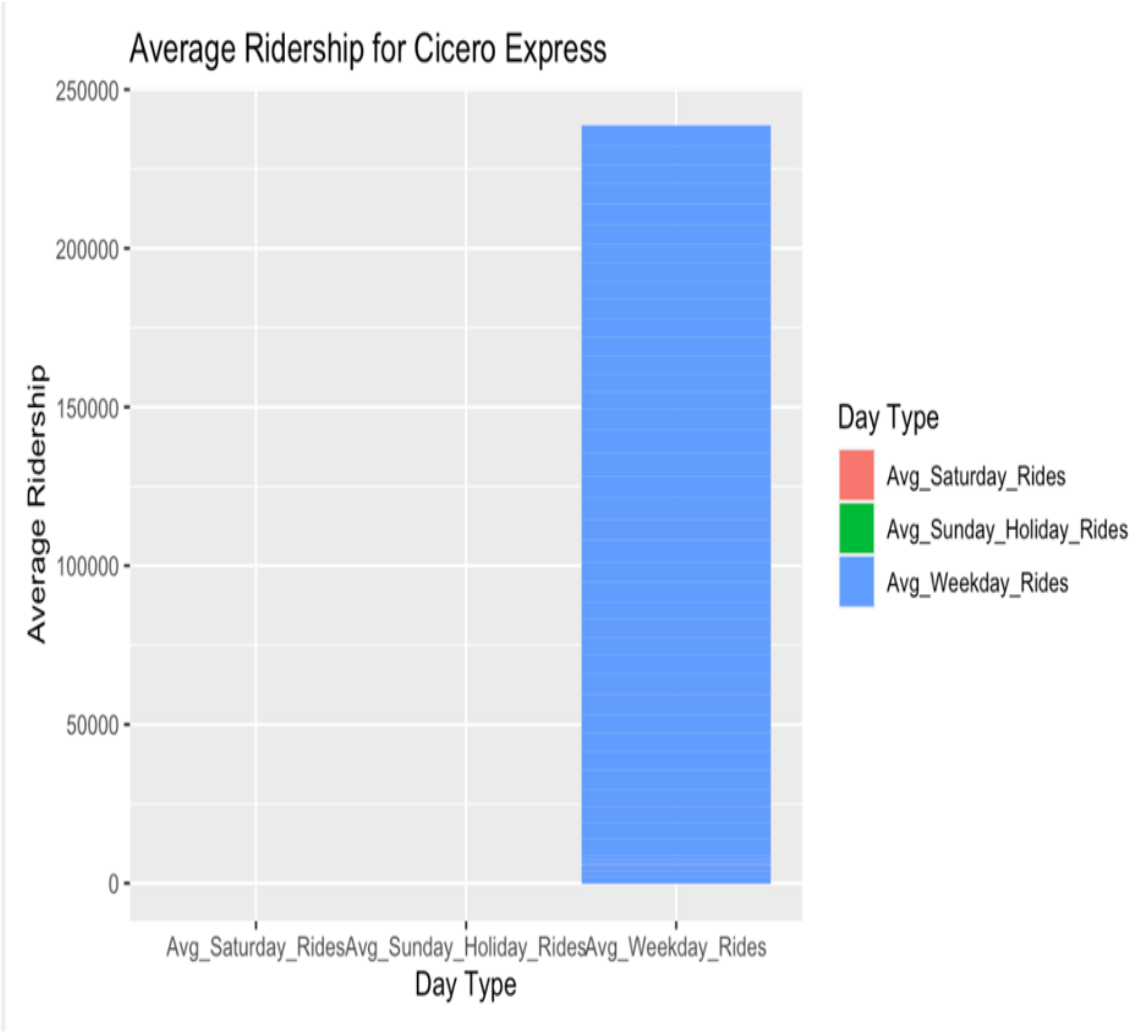
Exploratory Data Analytics Highlights

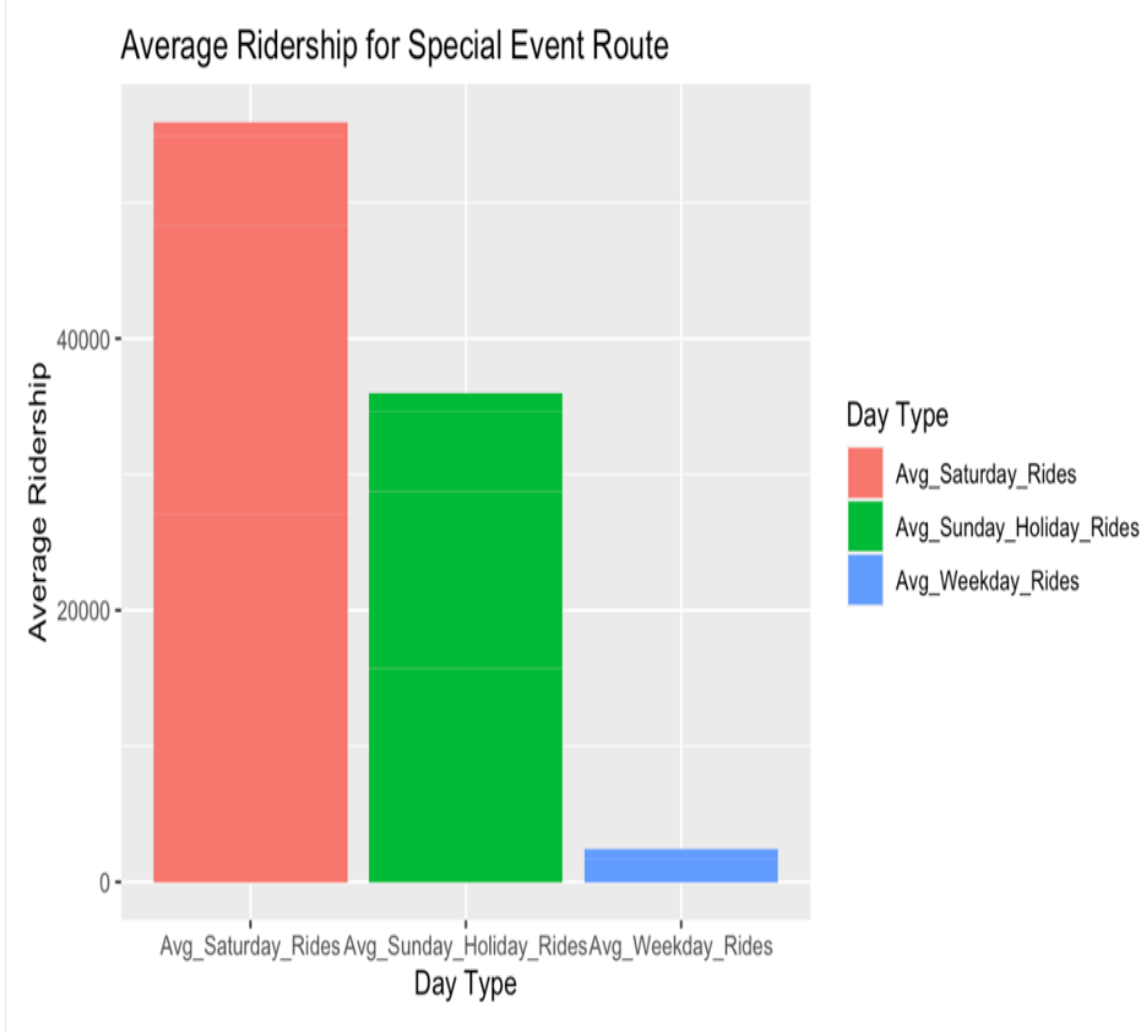
Average number of trips



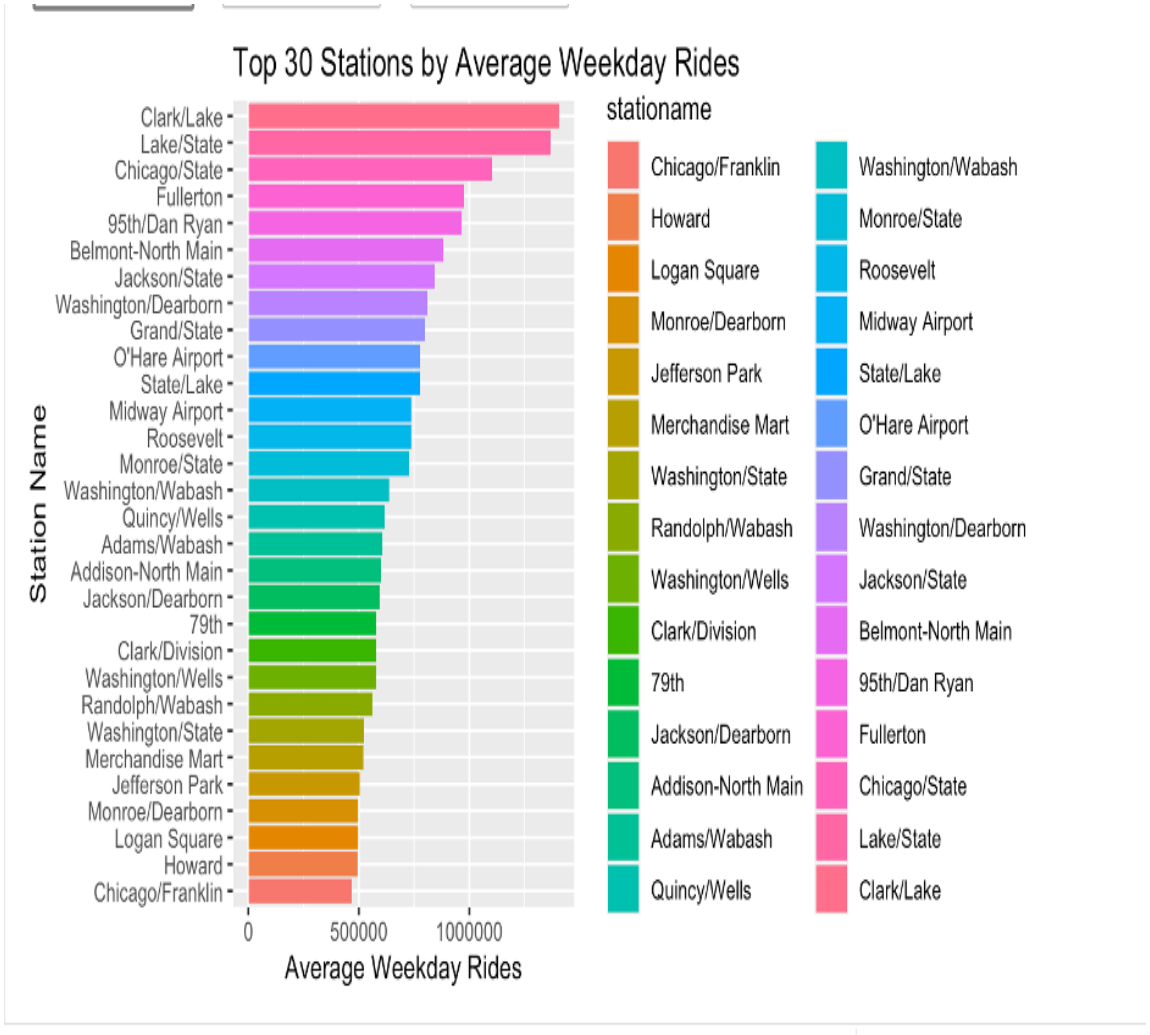


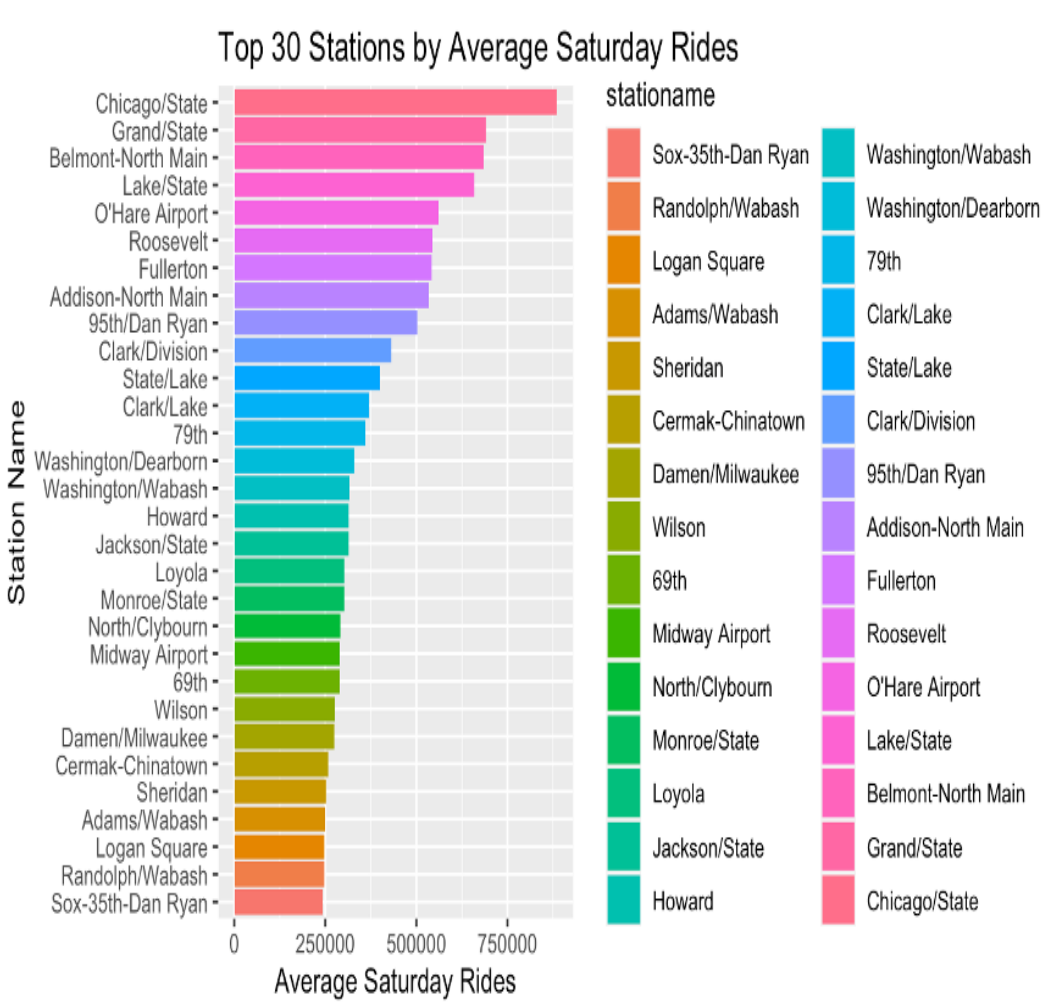
Average number of trips



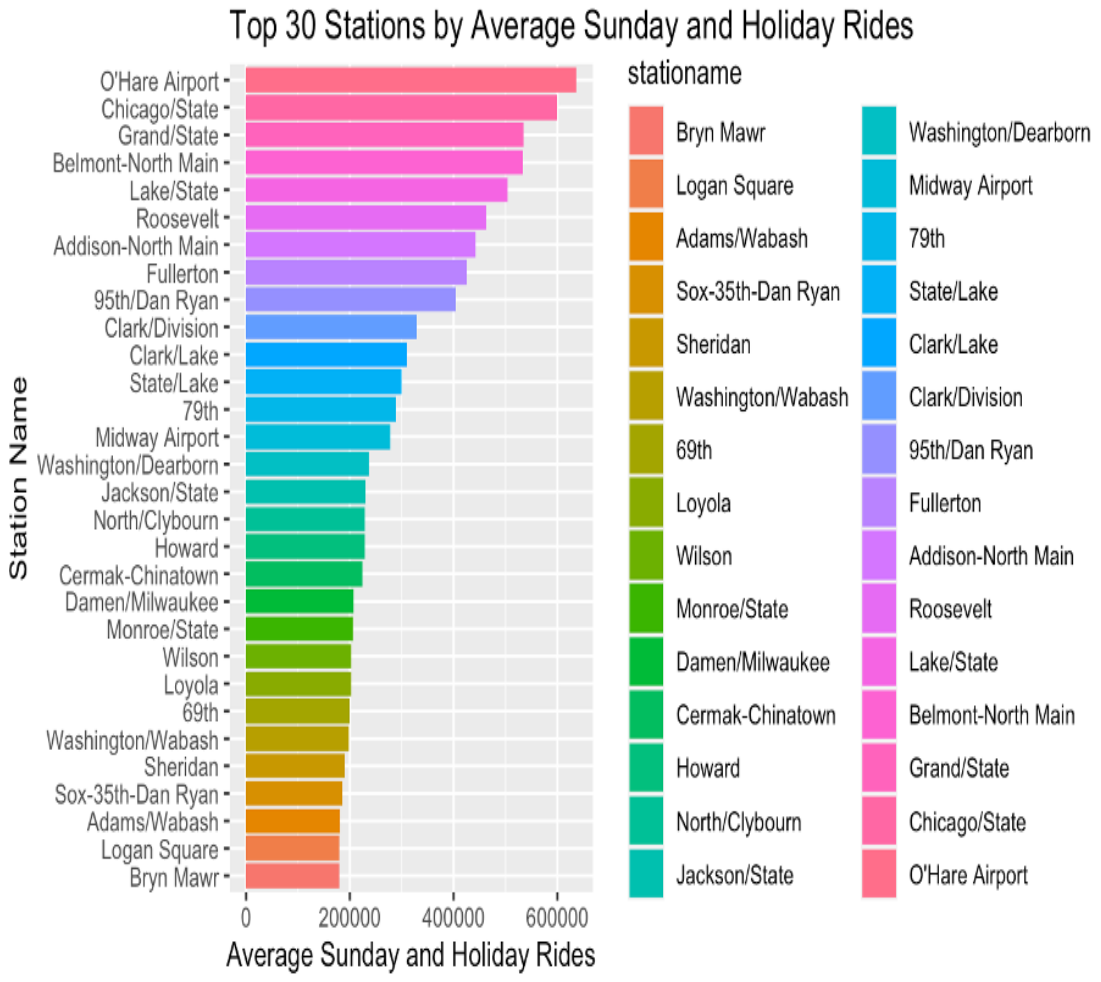


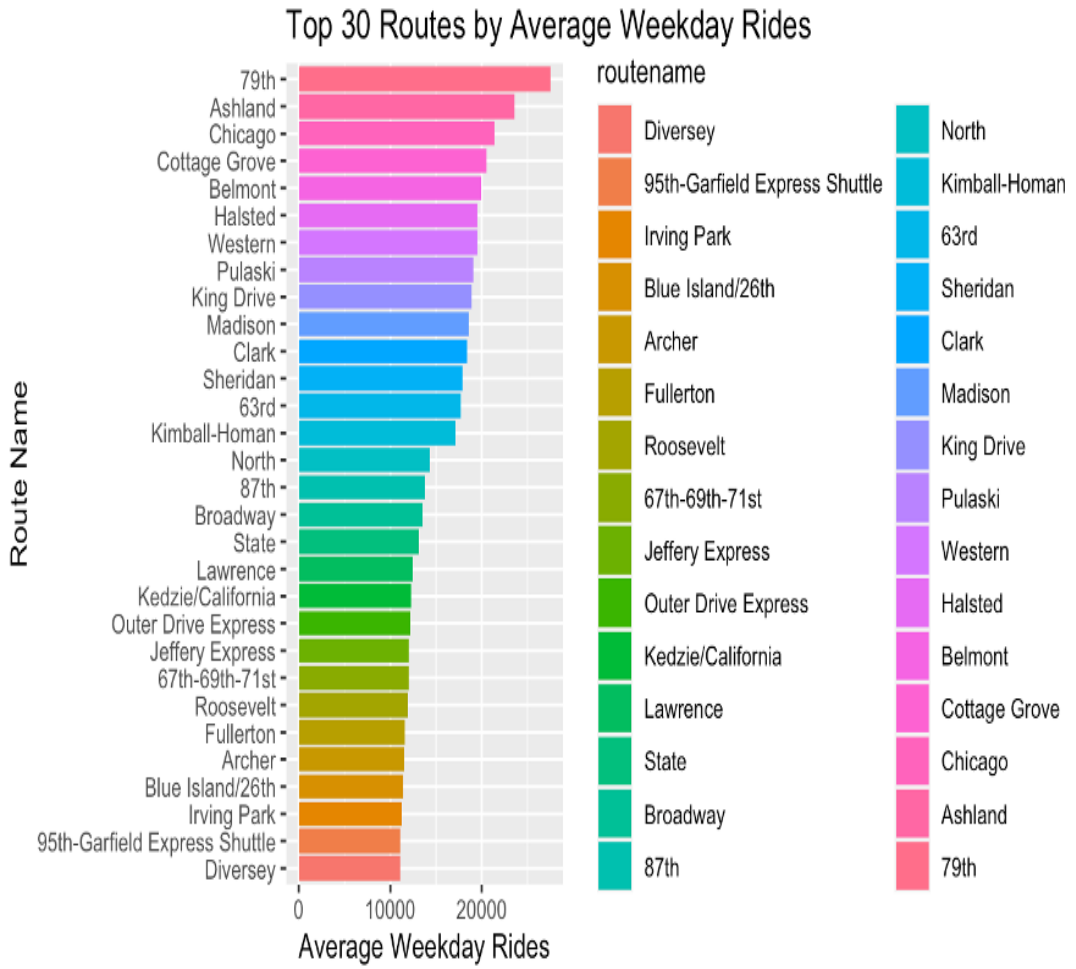
Average number of trips on Weekdays and Weekends



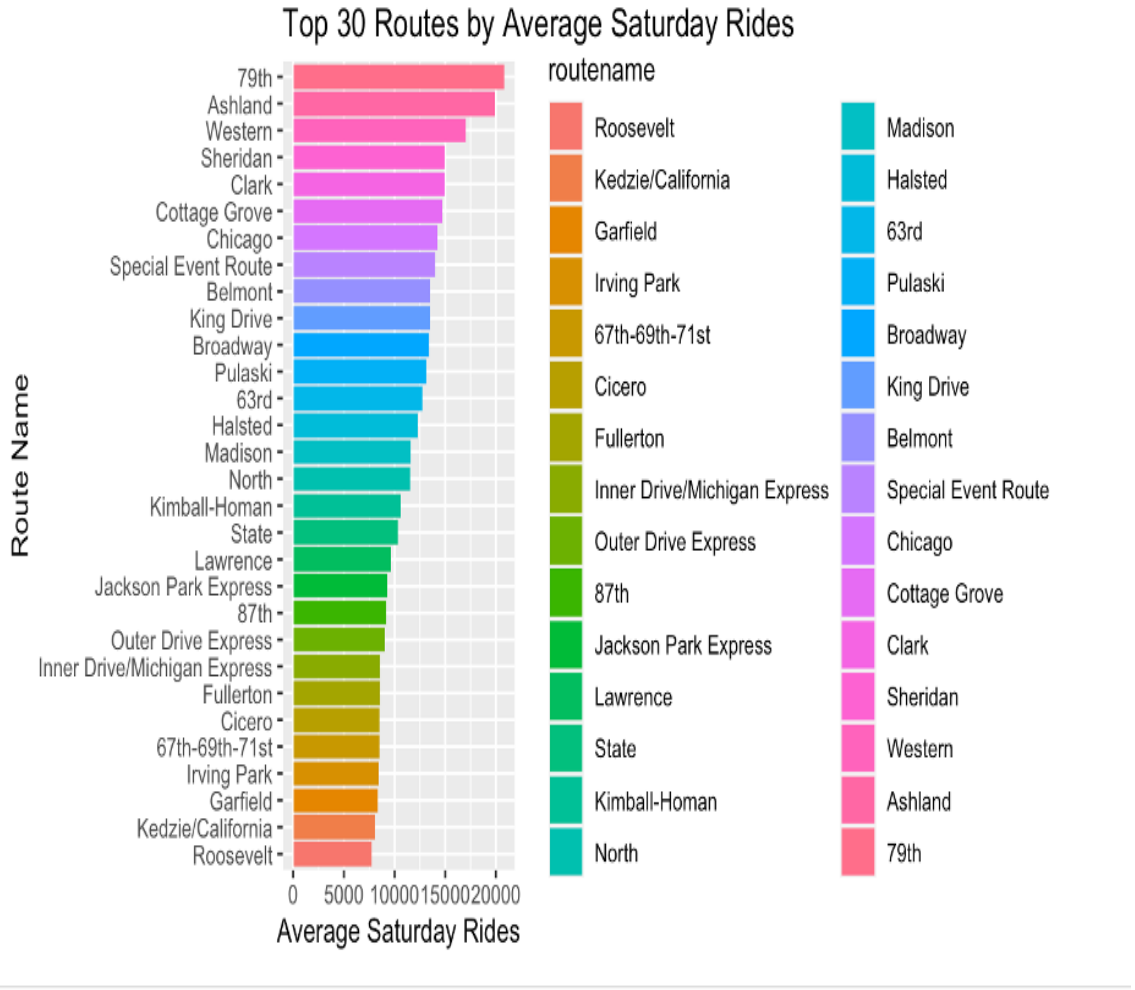


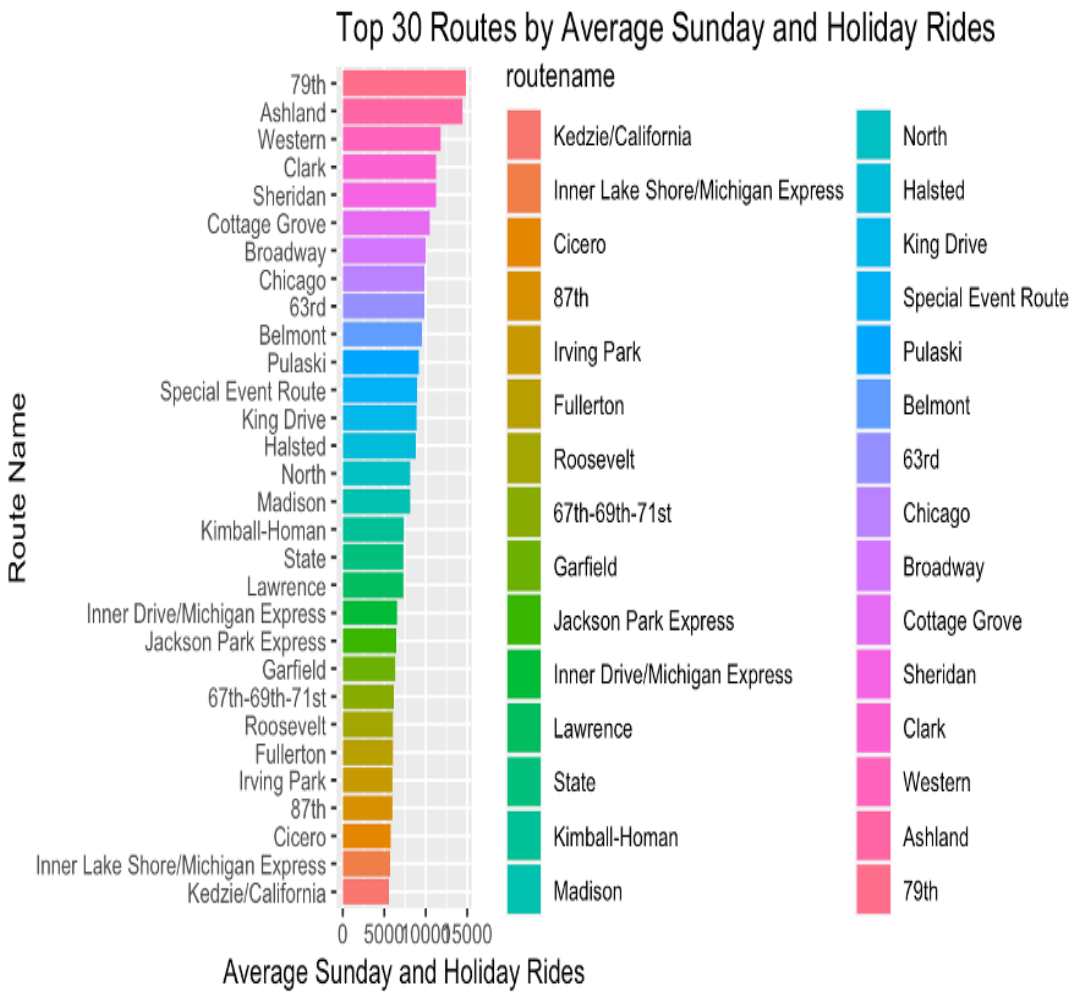
Average number of trips on Holidays and Weekdays



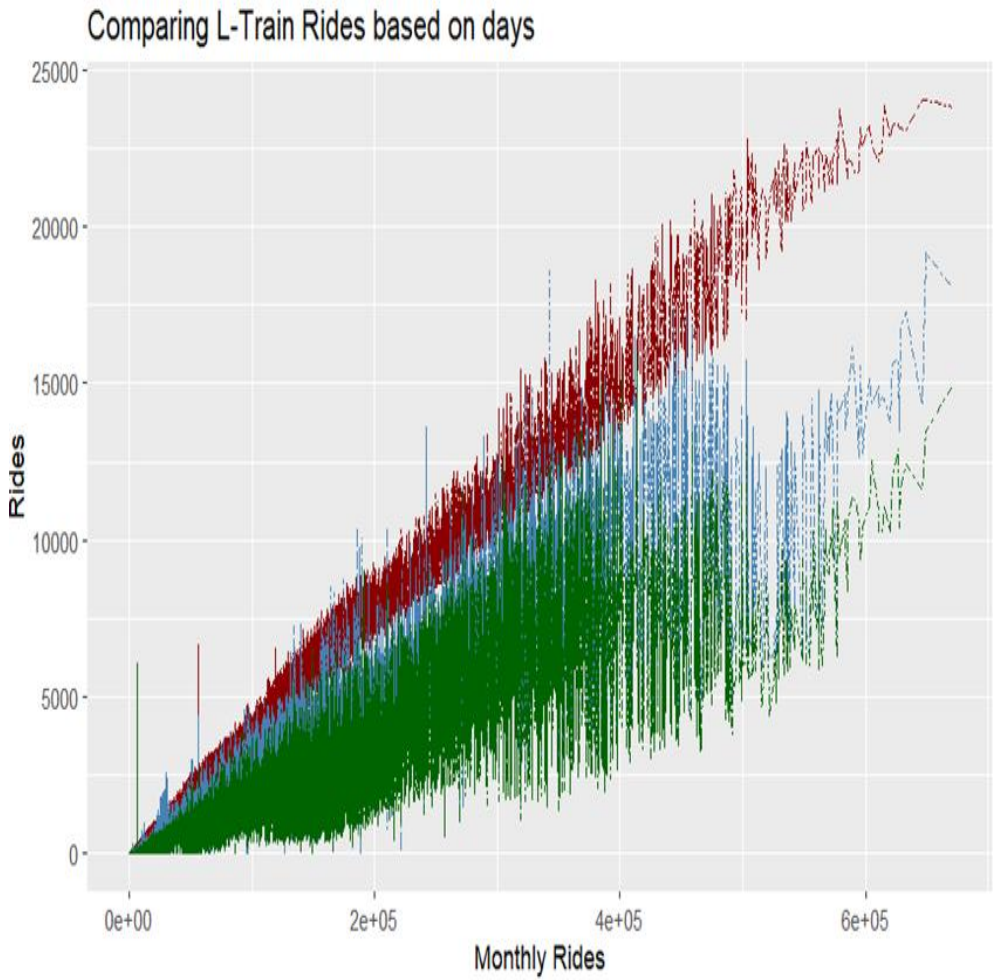


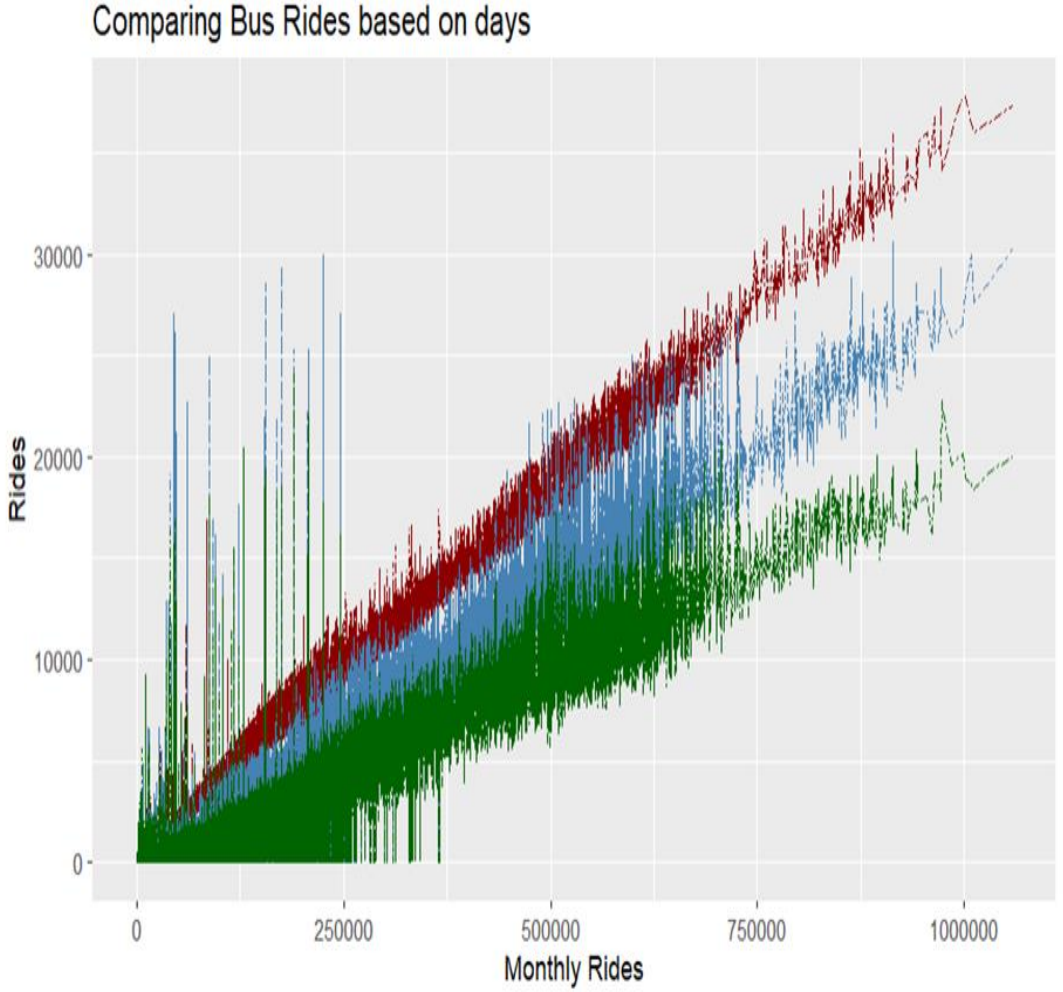
Average number of trips on Saturday and Sunday



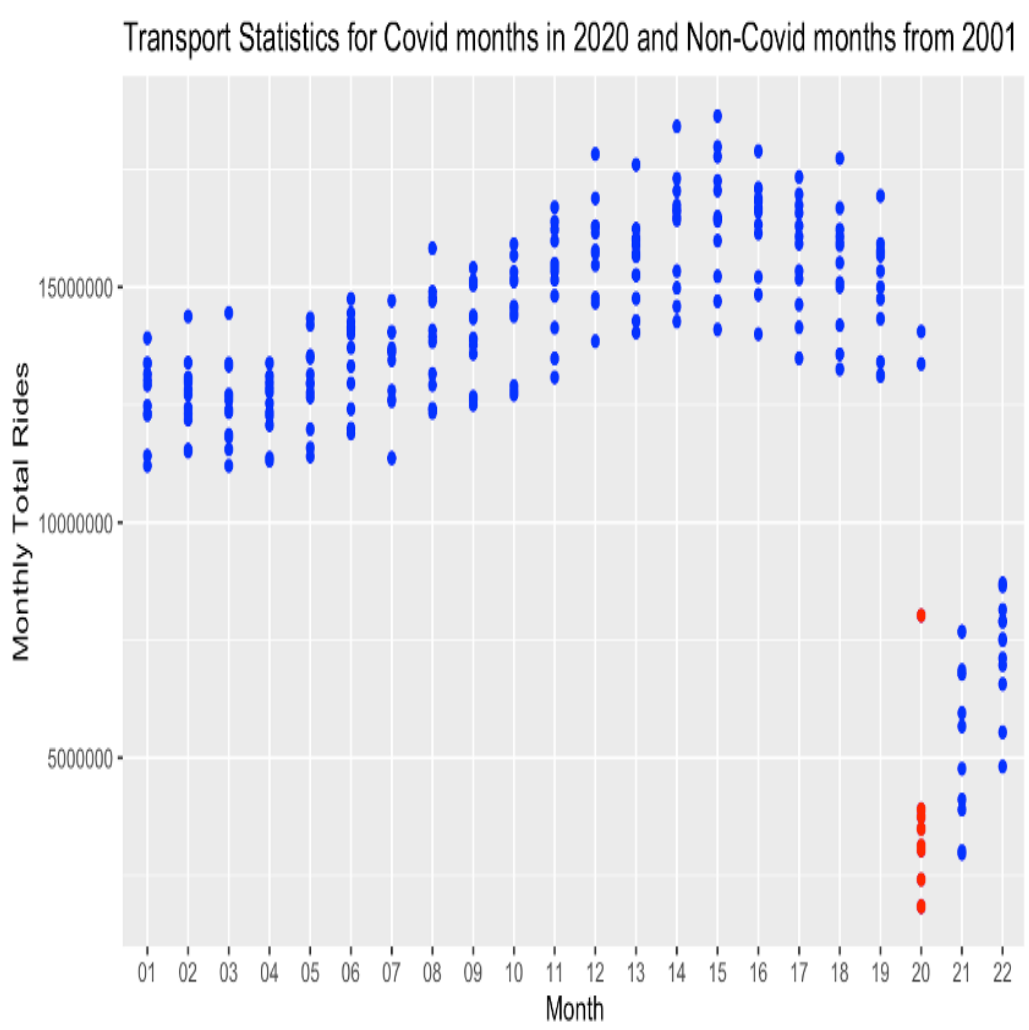


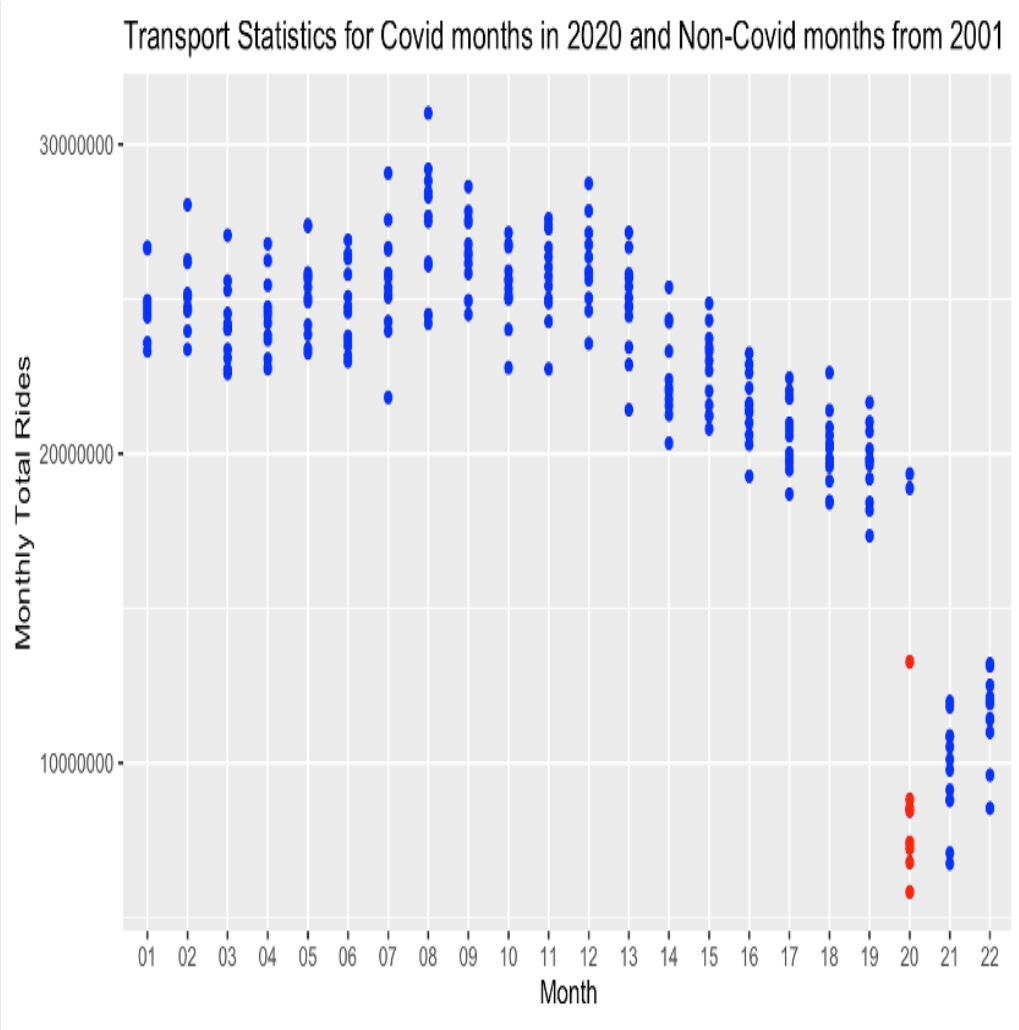
Comparing L-Train and Bus Rides



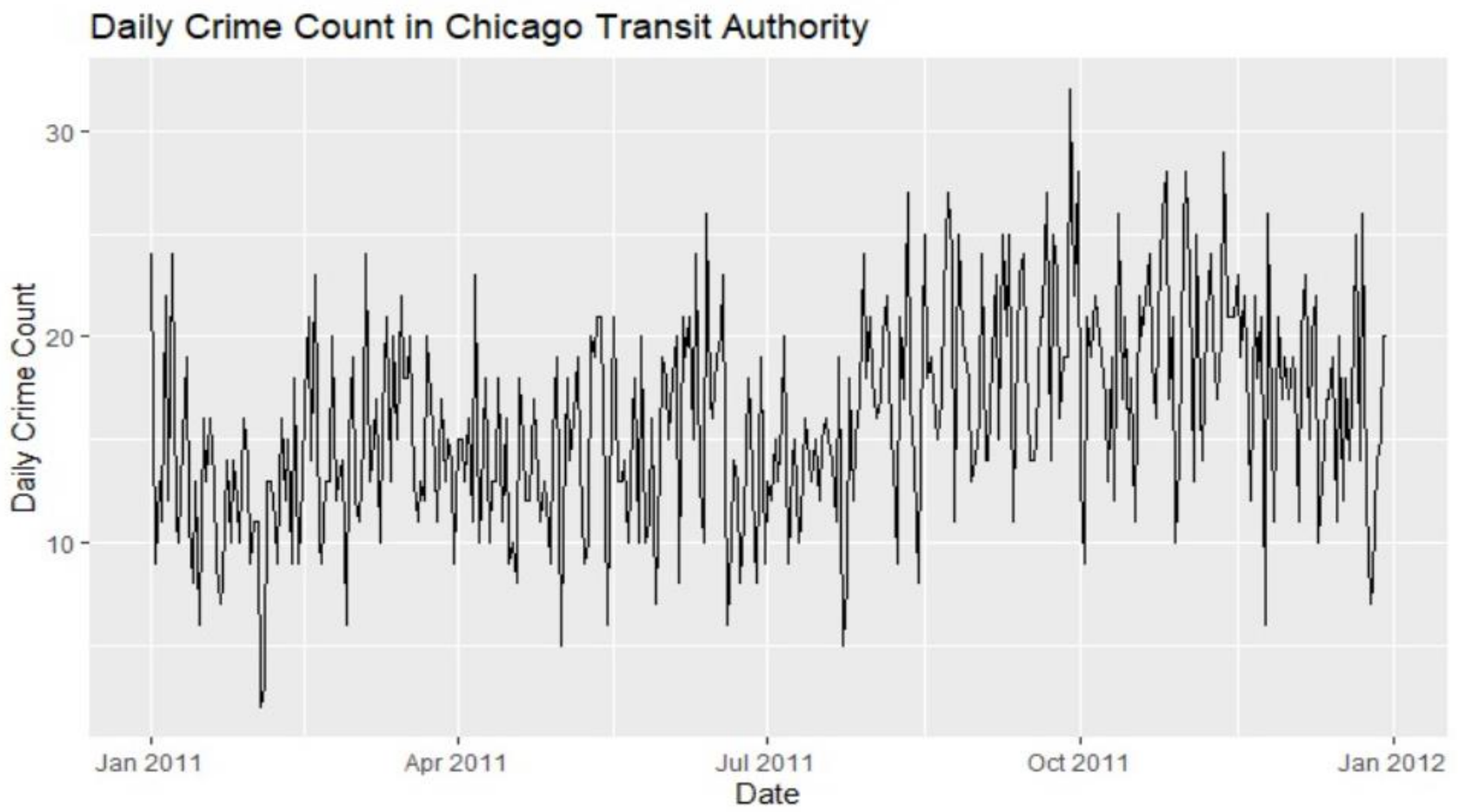


COVID Data Analysis



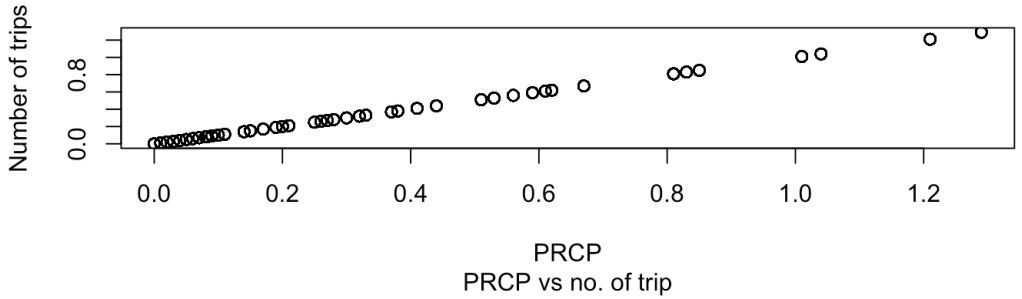


Crime Data Analysis

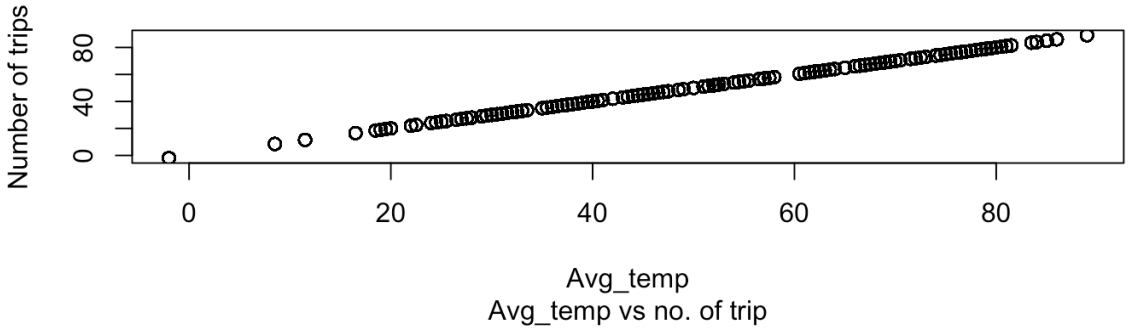


Weather Data Analysis

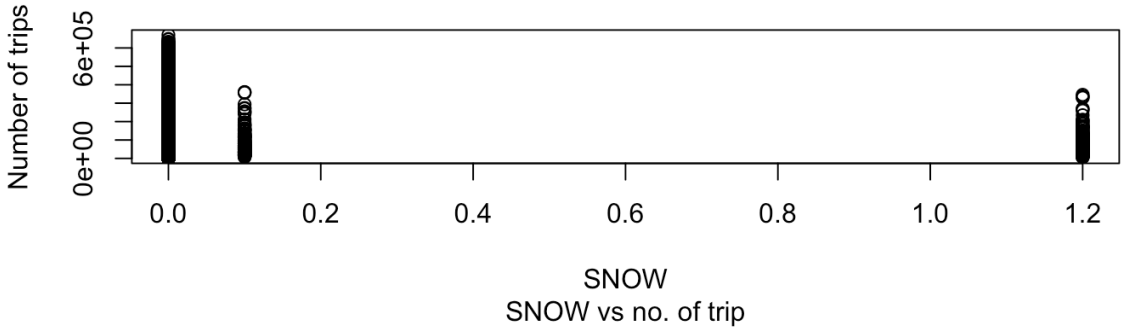
Precipitation versus the number of trips:



Temperature versus the number of trips:



Snow versus the number of trips:



Lasso Regression to eliminate Predictors

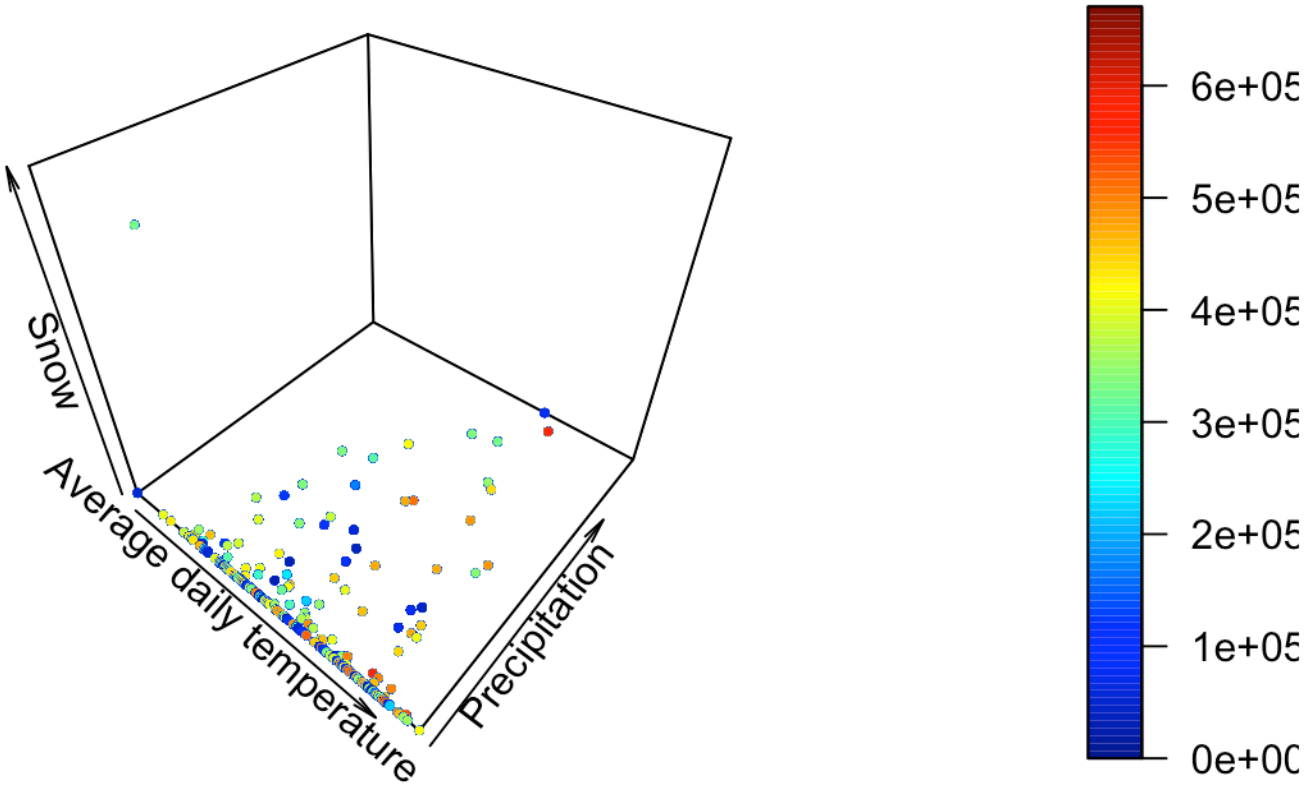
Before modeling, using Lasso coefficients, we removed few predictors. Below are the few parameters which we have taken into consideration

```
# use the selected lambda value to make predictions on the test set
x_test <- model.matrix(y ~ ., data = test_data)[,-1]
y_test <- test_data$y
lasso_pred <- predict(lasso_fit, s = cv_fit$lambda.min, newx = x_test)
```

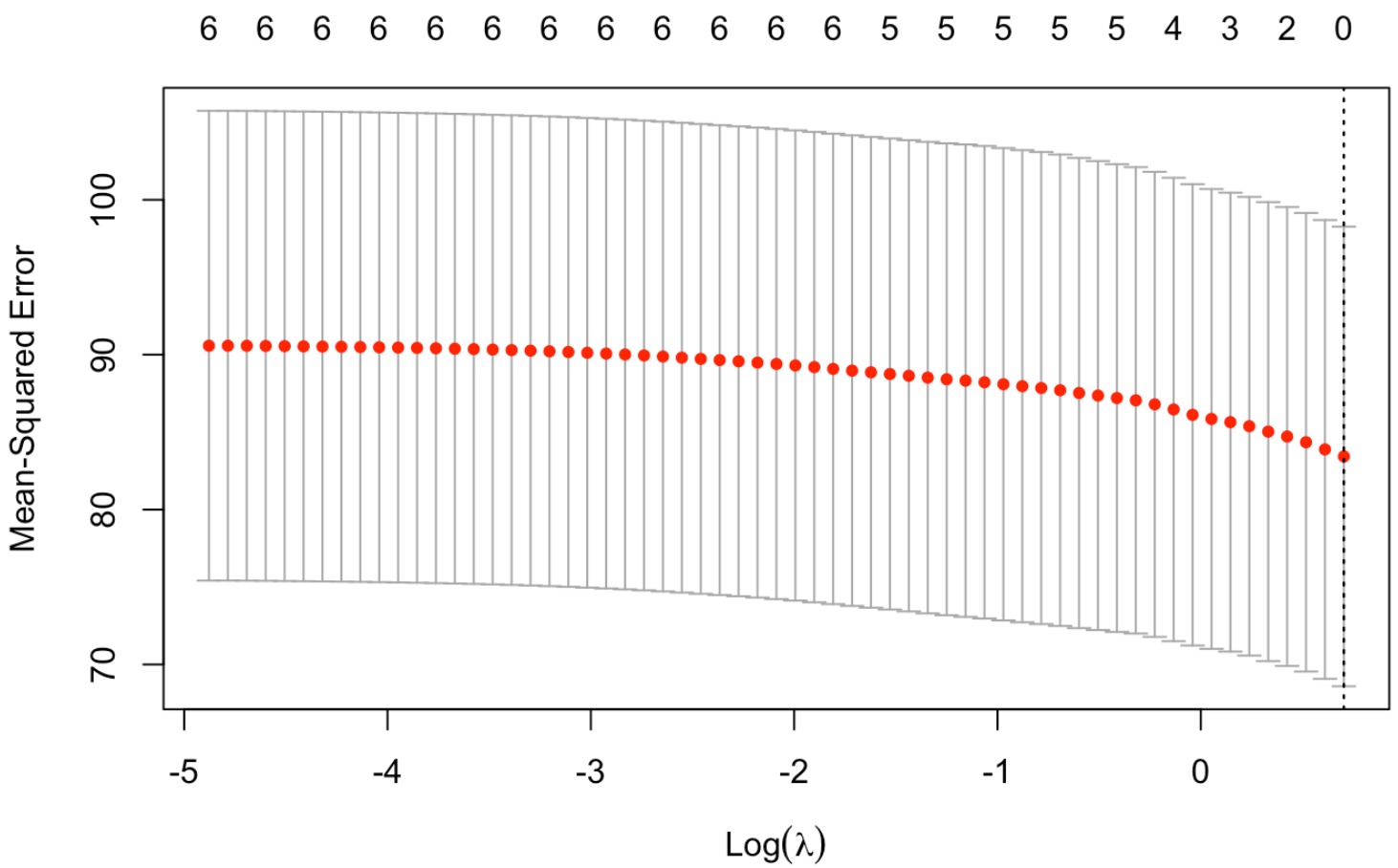
```
# fit a Lasso regression model using glmnet
x_train <- model.matrix(y ~ ., data = train_data)[,-1]
y_train <- train_data$y
lasso_fit <- glmnet(x_train, y_train, alpha = 1)
```

Using above predictors, we have done Linear model.

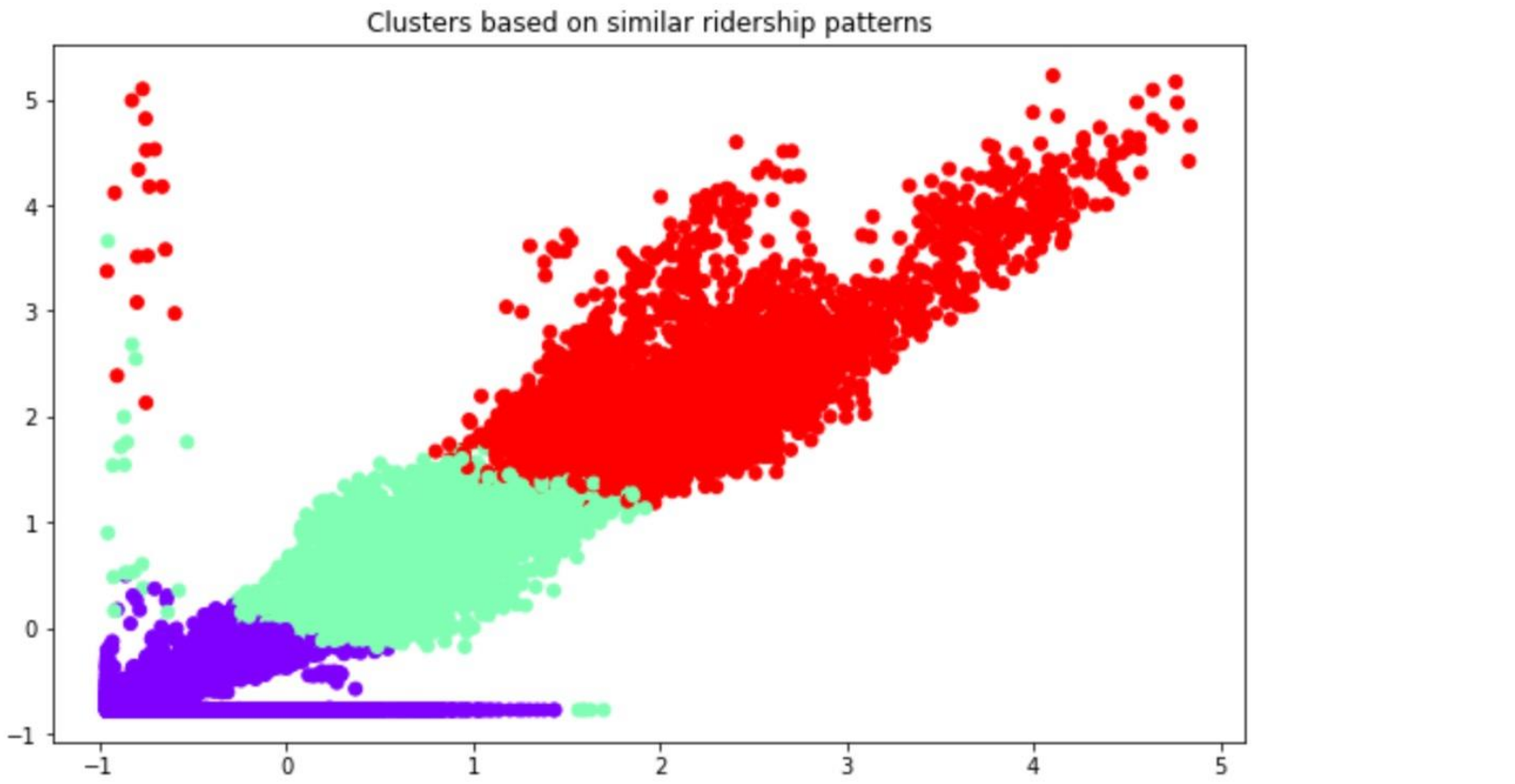
Multiple Linear Regression



Lasso Regression



K-means Clustering



Model Evaluation

For Multiple Linear Regression

Residual standard error: 85990

Multiple R-squared: 0.001946

Adjusted R-squared: 0.001865

F-statistic: 24.26

For Lasso Regression

Residual standard error: 9.564569

Mean squared error: 80.50327

Multiple R-squared: 0.83

Adjusted R-squared: 0.87

F-statistic: 0.0000251788

For K-Means Clustering

Silhouette score:

0.633605652308159

Calinski-Harabasz score:

104757.73820429464

Davies-Bouldin score:

0.566558442099883

CONCLUSION

- Limited dataset availability constrained project scope, hindering comprehensive analysis. However, we identified potential data sources and alternative collection solutions.
- Analyze CTA crime data to provide insights on station/area safety, spotting patterns in criminal activity. Crime/ridership data combo can pinpoint hotspots, aiding resource allocation to enhance security.
- COVID-19 affected CTA routes and schedules. Analyzing CTA and COVID-19 data identifies areas for cost-effective service adjustments. Linking COVID-19 dataset with CTA routes dataset was unfeasible, limiting pandemic impact analysis.
- Despite the limitations, the project was able to provide valuable insights into the available data and highlight the importance of having a well-curated dataset for conducting data analysis.

FUTURE WORK

- Create a new model with ARIMA that can handle time series data and can be used for forecasting with higher accuracy.
- Include the Gender Specific dataset into the CTA dataset. This will enhance the analysis and provide deeper insights into gender-based differences and preferences. Incorporating such data can improve the accuracy and relevance of your findings.
- Analyze CTA dataset by gender to adjust service levels to demand. By studying ridership by gender, identify differences on routes/times. This can optimize service and improve efficiency, but gender data was unavailable in the CTA dataset.

ILLINOIS TECH

THANK YOU