CSP 571 - Data Preparation and Analysis

Project Report

Chicago Public Transportation Data Analysis

"Transit Tales: Unveiling Chicago's Commute Story through Data"

Team Members:
NUPUR GUDIGAR – A20549865
ARHITH PATTATHIL SURESH - A20548751
SAMPREETH VILASAGARAPU - A20542614
SAMINATHAN ADAIKKAPPAN - A20545981

Table of Contents

Abstract

This project focuses on identifying congested routes within the Chicago Transit Authority (CTA) system during peak hours, with a specific emphasis on understanding the influence of external factors such as crashes, holidays, weather conditions, crime rates, and fluctuations in ridership on congestion levels. Through a comprehensive analysis of these factors, the research aims to offer actionable insights for both short-term traffic management and long-term infrastructure planning within the CTA system.

The primary objective of this research is to identify congested routes in the CTA system during peak hours by examining the effects of various external factors, including weather, holidays, traffic incidents, crime rates, and changes in ridership patterns. Utilizing statistical analysis techniques, trends in congestion will be mapped out, and the primary causes of congestion will be identified through the amalgamation and evaluation of data from multiple sources. Ultimately, the project aims to equip the CTA with valuable information to enhance operational strategies and route planning, thereby improving service efficiency and reliability. This research is expected to benefit both passengers utilizing the CTA system and the broader Chicago community by facilitating smoother and more efficient public transportation experiences.

Overview

Problem Statement

Congestion within the CTA system during peak hours poses significant challenges for commuters and transportation authorities alike. Understanding how external factors impact congestion levels is crucial for devising effective strategies to mitigate congestion and improve service efficiency. By analyzing these factors, this project aims to provide actionable insights for short-term traffic management and long-term infrastructure planning within the CTA system.

The specific set of questions that the project seeks to address

- How do crashes, holidays, weather conditions, crime rates, and ridership data correlate with congestion on CTA routes during peak hours?
- Can we predict congestion levels on CTA routes during peak hours based on these factors?
- What are the most significant predictors of congestion in the CTA system?
- Examine which transit lines have the most passengers boarding them;
- contrast the number of trips made on different routes on weekdays and weekends.

Relevant Literature

- **How the Chicago Transit Authority Benefits from Real-Time Data Analysis | StateTech**

Magazine

- **Data collection and analysis applied to intelligent transportation systems: a case study on public transportation | Discover Artificial Intelligence (springer.com)**
- **IEEE Xplore Full-Text PDF:RNN-Based Subway Passenger Flow Rolling Prediction**
- **Evaluation of congestion trends on chicago expressways (wiley.com)**
- **Practical Strategies for Reducing Congestion and Increasing Mobility for Chicago (reason.org)**
- **traffic-in-areas | City of Chicago | Data Portal**

Prior research has explored the impact of external factors on transportation congestion in various urban contexts. Studies have examined the correlation between congestion and factors such as weather conditions, traffic incidents, and ridership patterns. However, few studies have comprehensively analyzed the combined influence of multiple external factors on congestion within a specific transit system like the CTA. This project aims to fill this gap by synthesizing existing literature and applying relevant methodologies to the analysis of CTA congestion.

Proposed methodology

The proposed methodology encompasses several key steps aimed at analyzing the impact of external factors on congestion within the CTA system:

- Data Collection: Aggregate historical data on CTA ridership, traffic incidents, weather conditions, crime rates, and holiday schedules.
- Data Preprocessing: Clean and preprocess the data to handle missing values, and outliers, and ensure compatibility across different datasets.
- Exploratory Data Analysis (EDA): Conduct EDA to understand the distributions, trends, and correlations within the data.
- Feature Engineering: Develop features that effectively capture the impact of identified factors on congestion.
- Model Development: Utilize machine learning techniques, such as regression, decision trees, and ensemble methods, to develop a predictive model for congestion levels.
- Validation and Testing: Validate and test the model using historical data, employing cross-validation techniques to ensure robustness.

Data Processing

Data sources and collection:

CTA dataset
https://www.kaggle.com/datasets/chicago/chicago-transit-authority-cta-data/data?select=cta-ridership-bus-routes-monthly-day-type-averages-totals.csv
https://www.kaggle.com/datasets/chicago/chicago-transit-authority-cta-data/data?select=cta-ridership-l-station-entries-daily-totals.csv

https://www.kaggle.com/datasets/chicago/chicago-transit-authority-cta-data/data?select=cta-ridership-l-station-entries-monthly-day-type-averages-totals.csv

https://www.kaggle.com/datasets/chicago/chicago-transit-authority-cta-data/data?select=cta-ridership-daily-boarding-totals.csv

Additional Datasets used along with the main CTA dataset for analysis

Traffic Crashes - Chicago (kaggle.com)

CTA Crime | City of Chicago | Data Portal

Public holidays in Illinois, United States for 2023 - Holiday API

Past Weather | National Centers for Environmental Information (NCEI) (noaa.gov)

Data issues / Assumptions / Adjustments

BUS DATA:
The bus dataset comprises various columns, including:

route: Identifies the specific bus route.
routename: Names the bus route.
Month_Beginning: Represents the first date of each month from 2001 onwards.
Avg_Weekday_Rides: Indicates the average number of rides on weekdays per month for a particular route.
Avg_Saturday_Rides: Represents the average number of rides on Saturdays per month for a specific route.
Avg_Sunday_Holiday_Rides: Specifies the average number of rides on Sundays and Holidays per month for a given route.
MonthTotal: Reflects the total number of rides for a specific route in a month.

Changes to the dataset included formatting the Month_Beginning column into date format and removing any NA values.

L-TRAIN DATA:
The L-Train dataset includes columns such as:

Station_id: Identifies the unique ID of each L-train station.
Stationame: Names the L-train station.
month_beginning: Denotes the first date of each month from 2001 to 2022.
Weekday_Rides: Indicates the total number of trips taken on weekdays.
Saturday_Rides: Represents the total number of trips taken on Saturdays.
Sunday_Holiday_Rides: Specifies the total number of trips taken on Sundays and Holidays.
monthtotal: Reflects the total number of trips in a month.

Changes to the dataset involved converting the month_beginning column to the m/d/yyyy date format.

WEATHER DATA:
The Weather dataset includes columns such as:

Date: Spans from 2000 to 2011.
TAVG: Represents the average daily temperature in Fahrenheit.
TMAX: Indicates the maximum daily temperature in Fahrenheit.
TMIN: Specifies the minimum daily temperature in Fahrenheit.
PRCP: Reflects the precipitation on the day.
SNOW: Denotes the amount of snowfall on the day.
SNOWD: Represents the depth of snow on the day.

Changes to the dataset involved addressing missing TAVG values by calculating the average of TMAX and TMIN, and replacing NA values in PRCP, SNOW, and SNOWD with 0.

CRIME DATA:
The Crime dataset comprises various columns, including:

ID, Case.Number, Date, Block, IUCR, Primary.Type, Description, Arrest, Domestic, Beat, District, Ward, Community.Area, FBI.Code, Year, Updated.On, Latitude, Longitude, Location, Location.Description.

Changes to the dataset involved eliminating columns X.Coordinate, Y.Coordinate, Latitude, and Location, splitting the Date column into separate Date and Time columns, adjusting the date format to MM/DD/YYYY, converting the Time column to 24-hour format, and adding a new column called ActiveOrInactive to categorize crime occurrence based on active or inactive hours of the day.

TRAFFIC CRASHES DATA:
The traffic crashes dataset comprises various columns, including:
RD_NO, CRASH_DATE_EST_I,CRASH_DATE, POSTED_SPEED_LIMIT, TRAFFIC_CONTROL_DEVICE, DEVICE_CONDITION, WEATHER_CONDITION,

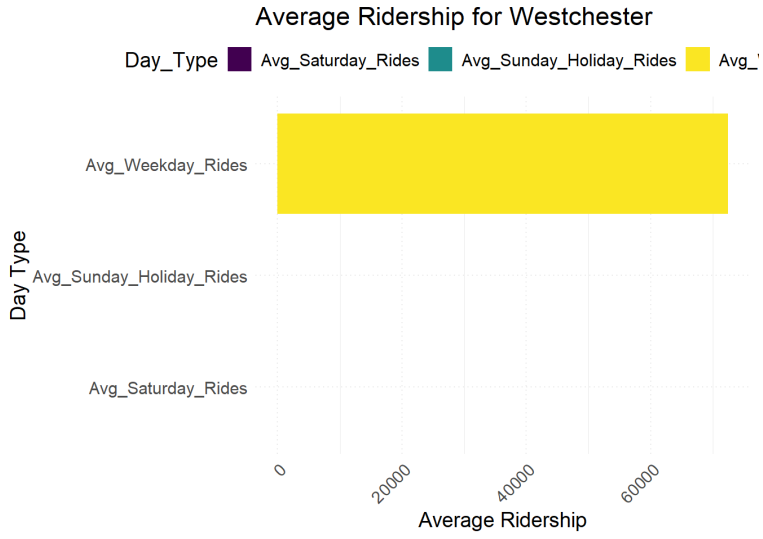
LIGHTING_CONDITION, FIRST_CRASH_TYPE, TRAFFICWAY_TYPE, LANE_CNT, ALIGNMENT, ROADWAY_SURFACE_COND, ROAD_DEFECT, REPORT_TYPE, CRASH_TYPE, INTERSECTION_RELATED_I, NOT_RIGHT_OF_WAY_I, HIT_AND_RUN_I, DAMAGE, DATE_POLICE_NOTIFIED, PRIM_CONTRIBUTORY_CAUSE, SEC_CONTRIBUTORY_CAUSE, STREET_NO, STREET_DIRECTION, STREET_NAME, BEAT_OF_OCCURRENCE, PHOTOS_TAKEN_I, STATEMENTS_TAKEN_I, DOORING_I, WORK_ZONE_I, WORK_ZONE_TYPE, WORKERS_PRESENT_I, NUM_UNITS, MOST_SEVERE_INJURY, INJURIES_TOTAL, INJURIES_FATAL, INJURIES_INCAPACITATING, INJURIES_NON_INCAPACITATING, INJURIES_REPORTED_NOT_EVIDENT, INJURIES_NO_INDICATION, INJURIES_UNKNOWN, CRASH_HOUR, CRASH_DAY_OF_WEEK, CRASH_MONTH, LATITUDE, LONGITUDE, LOCATION.

The "CRASH_DATE" was originally in a DateTime(MM/DD/YYYY hh:mm:ss format, but it was changed to a simple date M/DD/YYYY format for consistency and easier analysis.

Exploratory Data Analysis

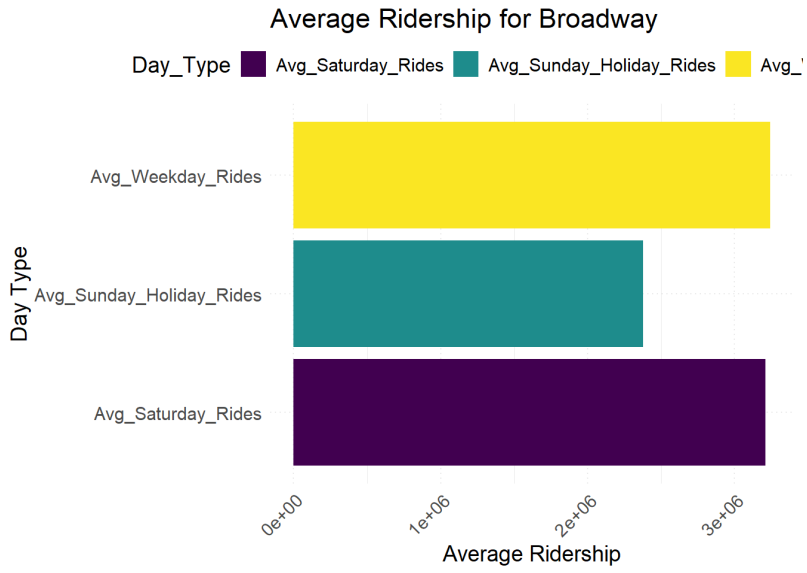
FOR EVERY BUS ROUTE, THE AVERAGE BUS RIDERSHIP ON DIFFERENT KINDS OF DAYS:

On weekdays, weekends, and holidays, the average number of riders on the **Westchester route** is:



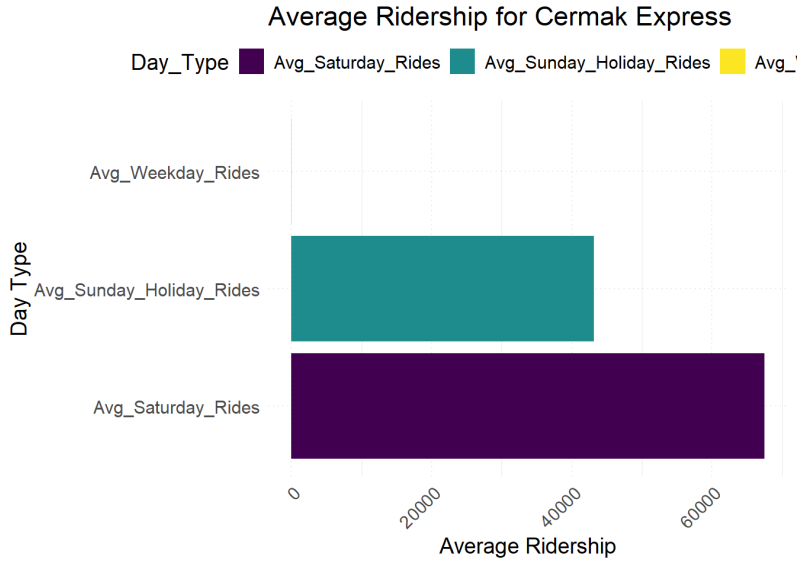
This plot shows average ridership on weekdays, Saturdays, and Sundays/holidays for the "Westchester" bus route. The data indicates that weekday ridership is considerably higher compared to weekends and holidays. This difference in ridership suggests that this route serves as a commuter route, primarily used by those traveling to work or school on weekdays, and lacks a sufficient number of passengers on weekends.

On weekdays, weekends, and holidays, the average number of riders on the **Broadway** route is:



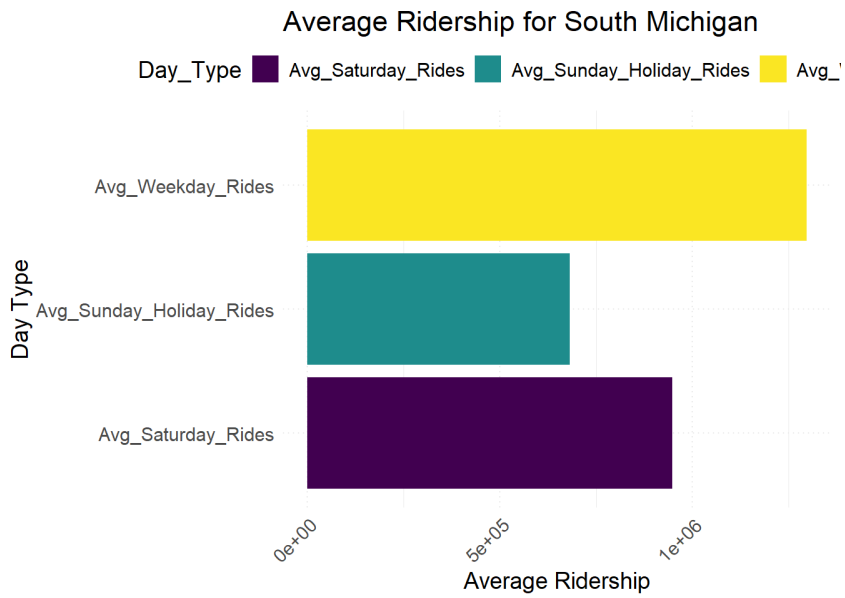
The "Broadway" plot demonstrates an even distribution of ridership across all three categories: weekdays, Saturdays, and Sundays/holidays. This suggests that this route is used for both work-related commuting and leisure activities on weekends. The balance across days indicates a diverse usage pattern, where people rely on this route all over the week in the month.

On weekdays, weekends, and holidays, the average number of riders on the **Cermal Expres**s route is:



The "Cermak Express" plot reveals that this route has balanced ridership between Saturdays and Sundays/holidays. In contrast, weekdays show a notable dip in ridership. This trend suggests that the route is primarily used for leisure or weekend-based activities, with fewer people relying on it for weekday commuting.

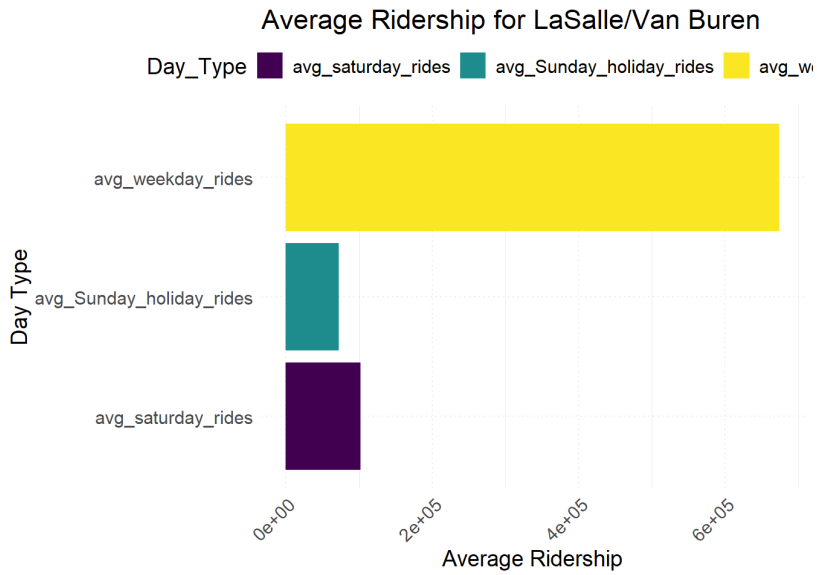
On weekdays, weekends, and holidays, the average number of riders on the **South Michigan** route is:



This graph displays ridership for the "South Michigan" route, with weekdays having the highest ridership, followed by Saturdays and Sundays/holidays. The consistent use across all days suggests that this route is being used all days of the week

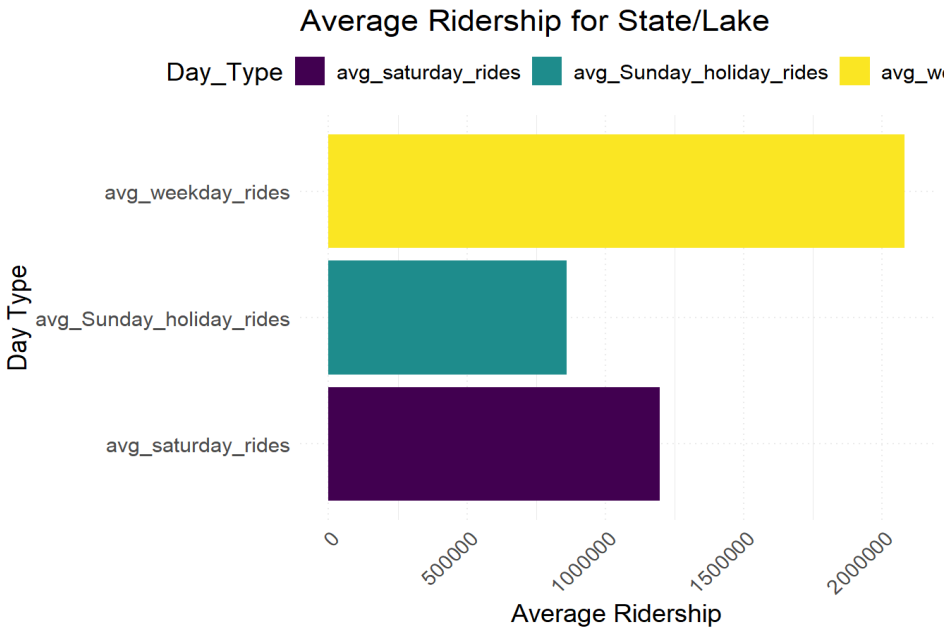
FOR EVERY L STATION, THE AVERAGE RIDERSHIP ON DIFFERENT KINDS OF DAYS:

On weekdays, weekends, and holidays, the average number of rides from the **LASALLE/VAN BUREN** Station is:



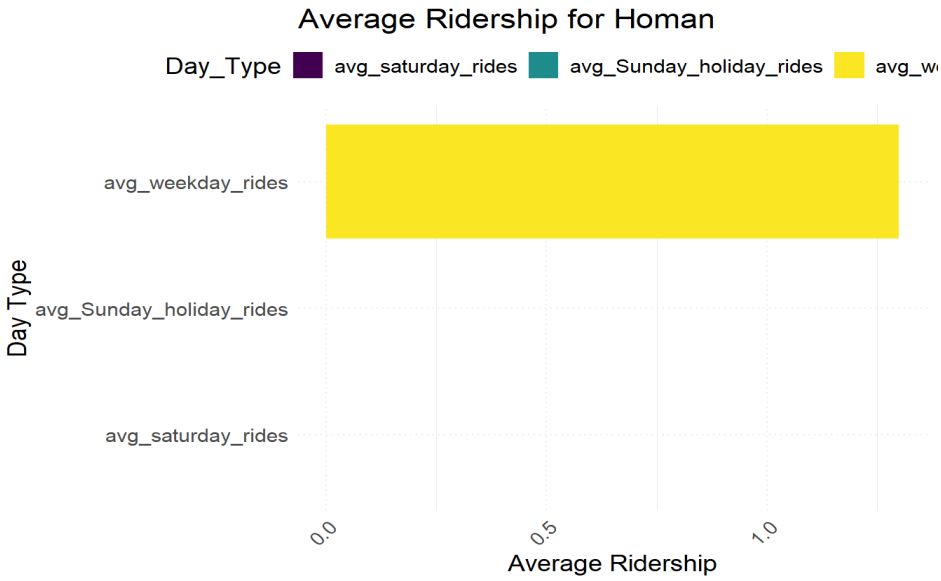
LASALLE/VAN BUREN plot shows that their is high ridership on weekdays compared to other day types.Sunday/Holiday and Saturday Ridership have shorter bars, suggesting less usage during weekends and holidays.

On weekdays, weekends, and holidays, the average number of rides from the **STATE/LAKE** Station is:



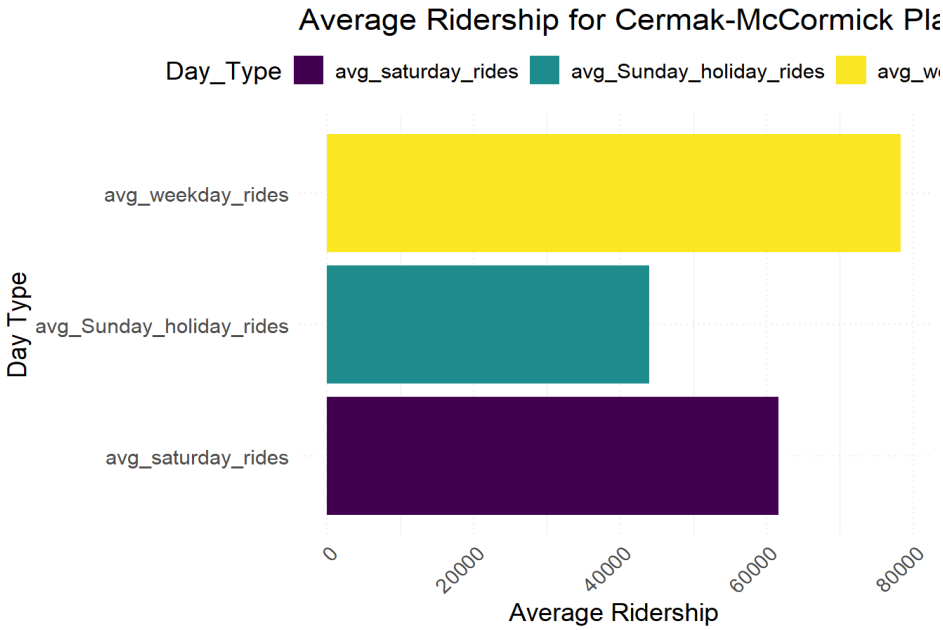
STATE/LAKE plot shows that weekday rides significantly exceed other day types. while both Saturday and Sunday/holiday ridership are substantial, suggesting this station has consistent ridership through the week and on weekends.

On weekdays, weekends, and holidays, the average number of rides from the **HOMAN** Station is:



HOMAN plot shows significantly higher, indicating that most rides occur during weekdays and very low or negligible ridership on Saturday and Sunday/Holiday suggesting that this location has limited or no weekend operations.

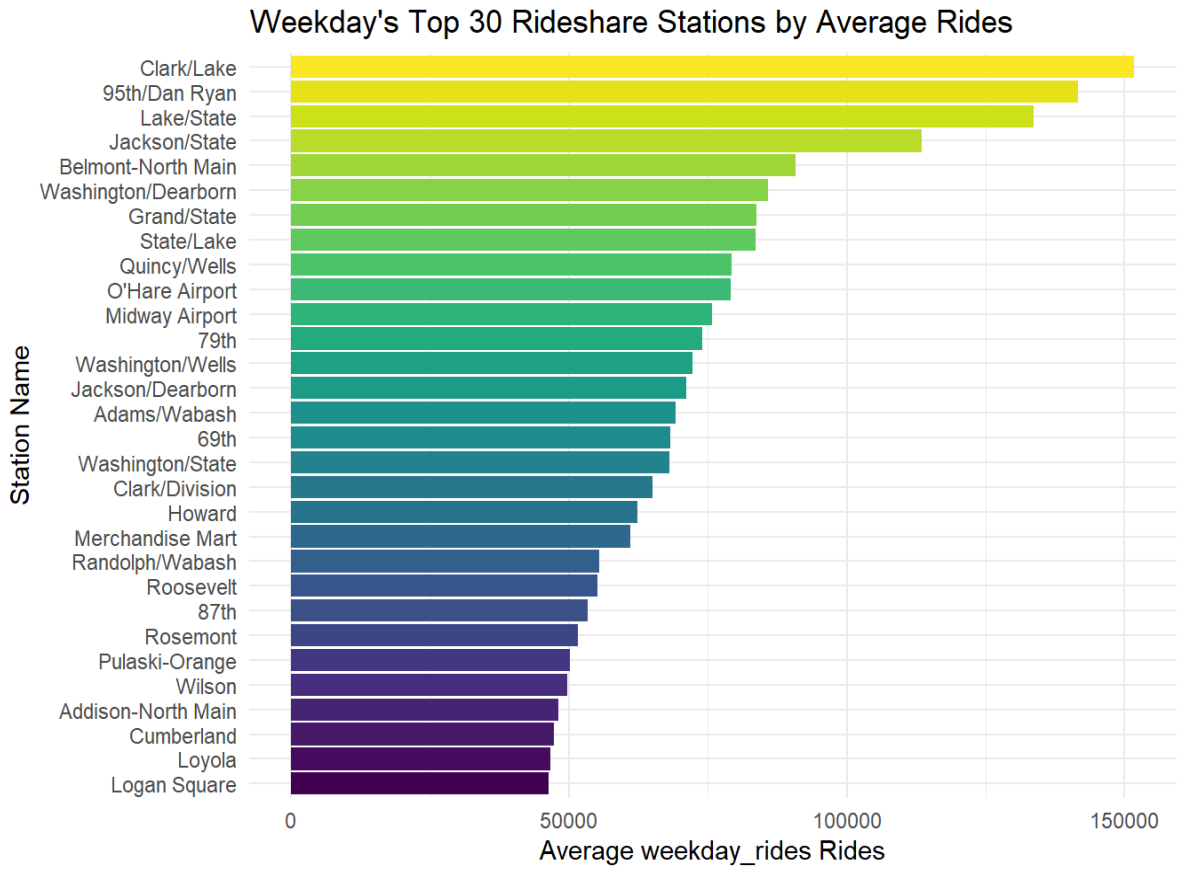
On weekdays, weekends, and holidays, the average number of rides from the **CERMAK-MCCORMICK** Station is:



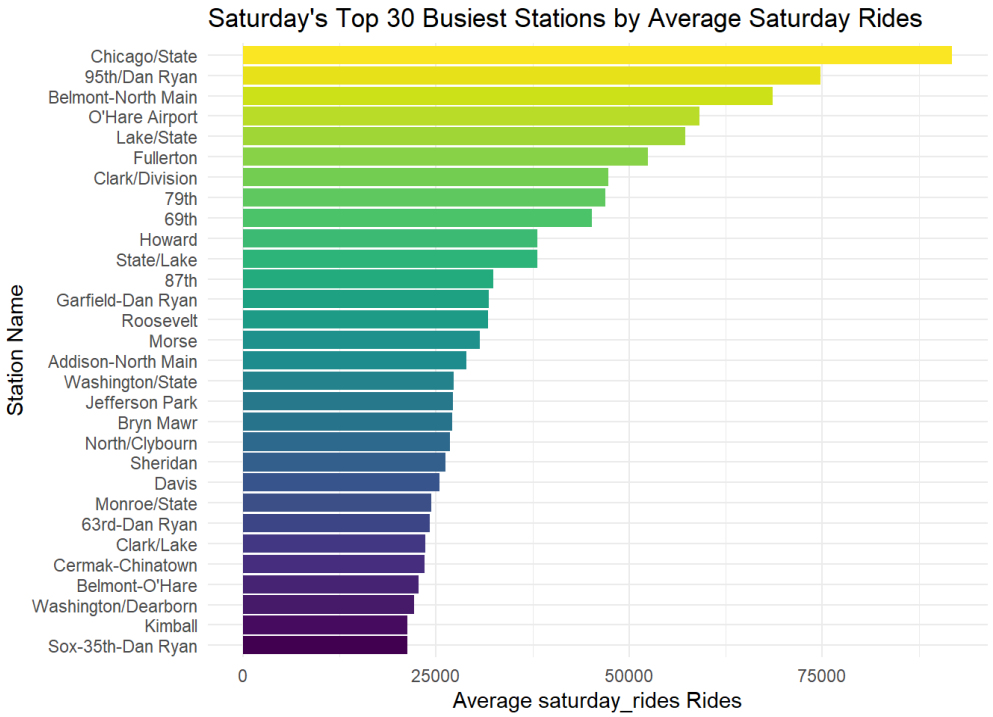
CERMAK-MCCORMICK Weekday ridership is higher than Saturday and Sunday/Holiday ridership but there is uniform distribution for all the days in the week this shows that this station is active all over the week on all days

Prior to training and assessing the performance of our models, we will display a few distinct dataset features.

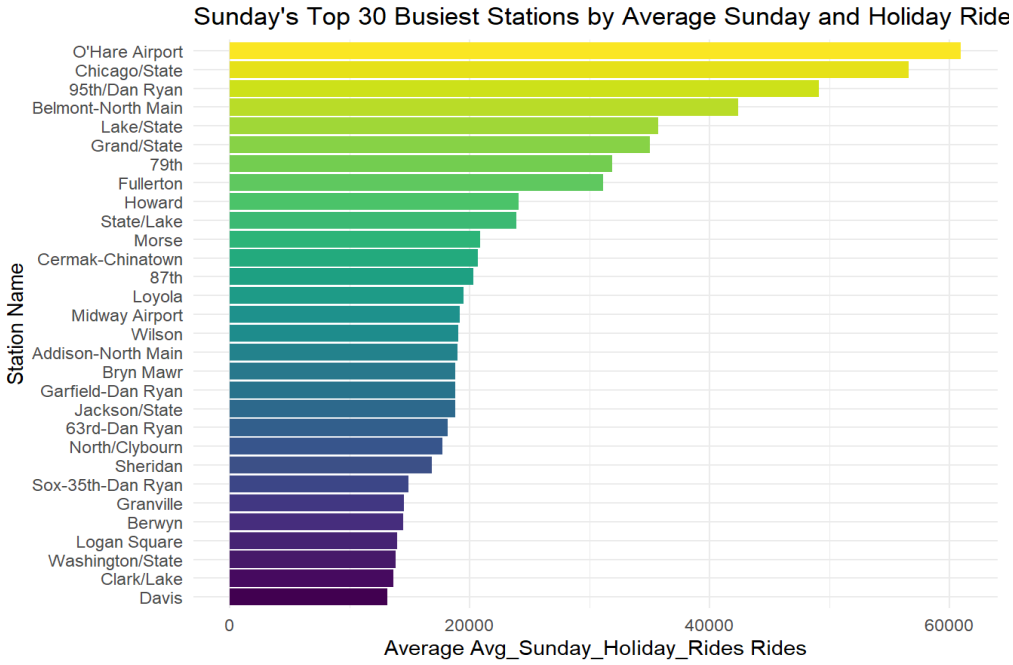
L station Data Analysis:



The above bar graph,"Weekday's Top 30 Rideshare Stations by Average Rides," shows Clark/Lake leading with about 150,000 average weekday rides, highlighting its importance in the transit network. The range of ridership across the 30 stations suggests varying commuter traffic, with notable differences between the busiest and less frequented stations, like Logan Square.

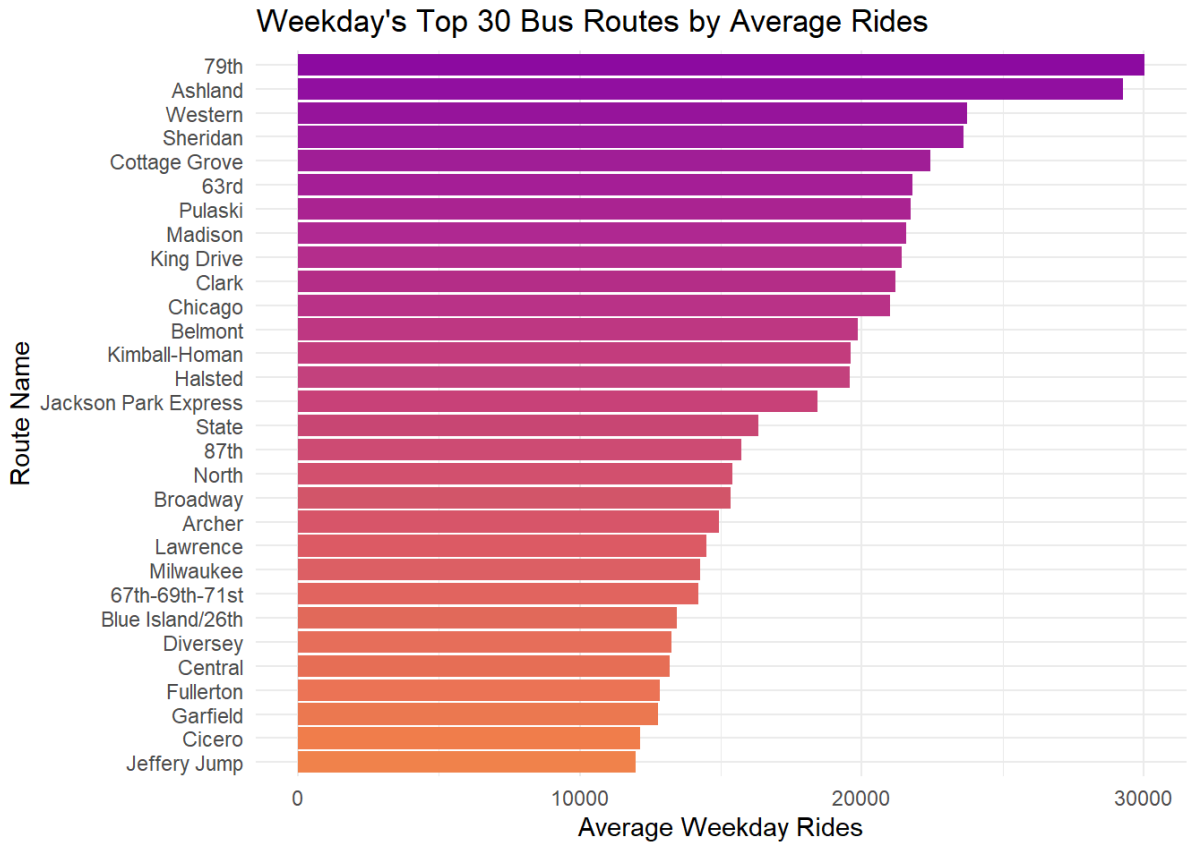


The above bar graph titled "Saturday's Top 30 Busiest Stations by Average Saturday Rides" reveals Chicago/State as the most frequented station with over 75,000 average Saturday rides. The graph showcases a wide range of ridership levels among the 30 busiest stations, with 95th/Dan Ryan and Belmont-North Main also ranking high in terms of traffic. This distribution indicates variations in station popularity and commuter patterns during weekends.

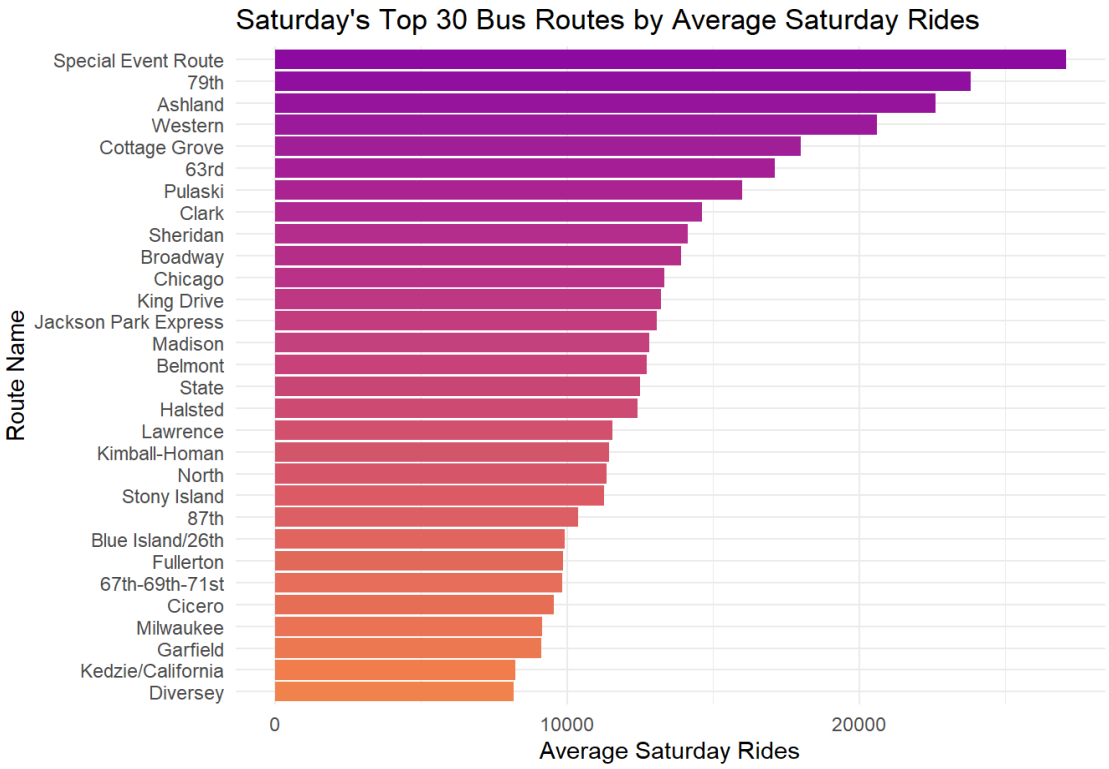


The above bar graph titled "Sunday's Top 30 Busiest Stations by Average Sunday and Holiday Rides" ranks O'Hare Airport as the most active station with almost 60,000 average rides on Sundays and holidays. Chicago/State and 95th/Dan Ryan follow closely, indicating their high commuter traffic even on non-working days. This pattern suggests that these stations are significant transit hubs, serving a high volume of weekend and holiday travelers.

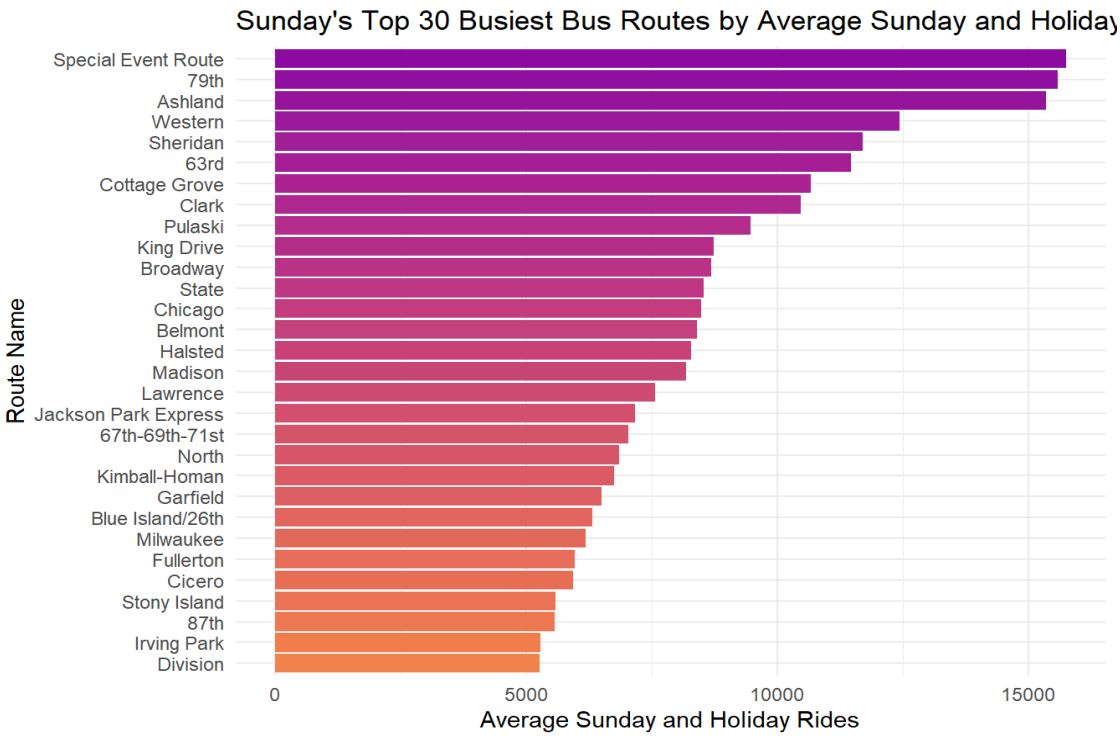
Bus Routes Data Analysis:



These routes have the highest average weekday ridership, as shown by the bar graph "Weekday's Top 30 Routes by Average Rides". With the most rides, the 79th route is in the forefront, closely followed by Ashland and Western, demonstrating their significance in the transportation system. Weekday bus service appears to be well-utilized and diverse, as seen by the evenly distributed passengers across the top 30 routes.

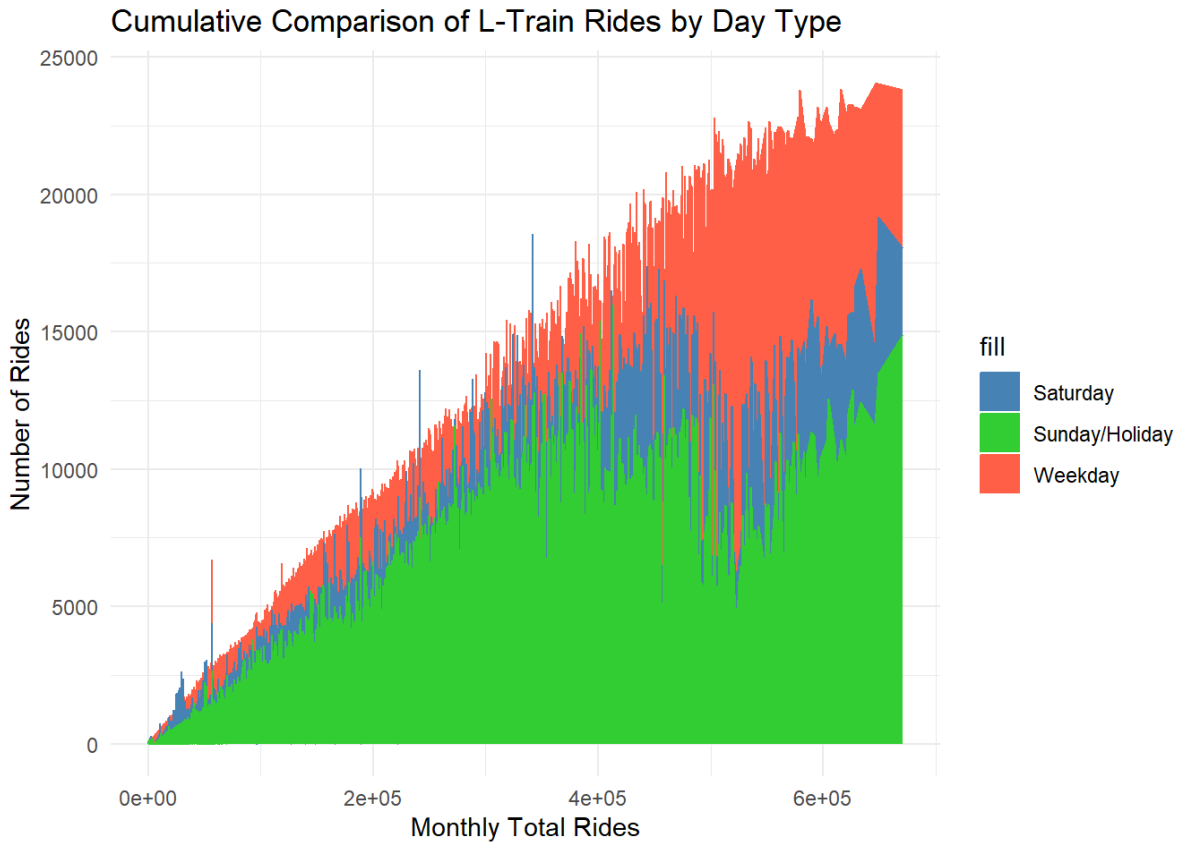


The most well-liked bus routes on Saturdays are displayed in the bar graph "Saturday's Top 30 Bus Routes by Average Saturday Rides". With the highest average journeys, the "Special Event Route" is the most popular, emphasizing its importance on weekends. Routes such as 79th, Ashland, and Western trail closely behind, indicating heavy traffic on this particular day.



The bus routes that are most frequently used on Sundays and holidays are displayed in the bar graph "Sunday's Top 30 Busiest Bus Routes by Average Sunday and Holiday Rides". With the greatest average rides, the "Special Event Route" is in first place, demonstrating its popularity during certain periods. Routes like 79th, Ashland, and Western follow closely, indicating heavy traffic on these days.

L-Train ridership according to day of the week:

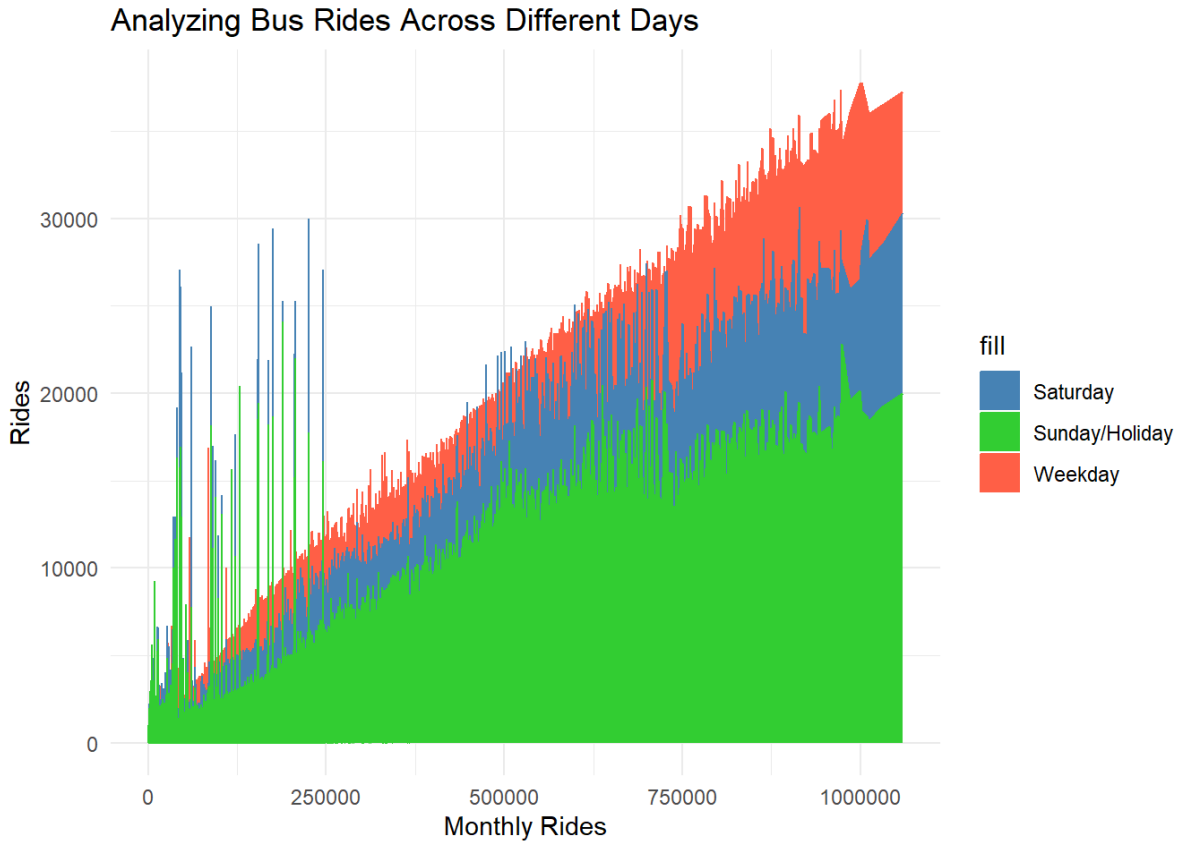


In the above graph:

→Red represents Weekday L-Train rides, indicating that weekdays see the highest ridership levels.
→Blue denotes Saturday L-Train rides, suggesting lower ridership on Saturdays compared to weekdays.
→Green shows Sunday/Holiday L-Train rides, indicating the lowest ridership among the three.

This graph's distribution demonstrates that weekday L-train travels are the most popular, with a high ridership all week long. Over the weekend, ridership tends to decline, with Saturdays seeing fewer rides than weekdays. The least number of ridership occurs on Sundays and holidays, which is a notable decrease from the rest of the week.

Bus ridership according to day of the week:

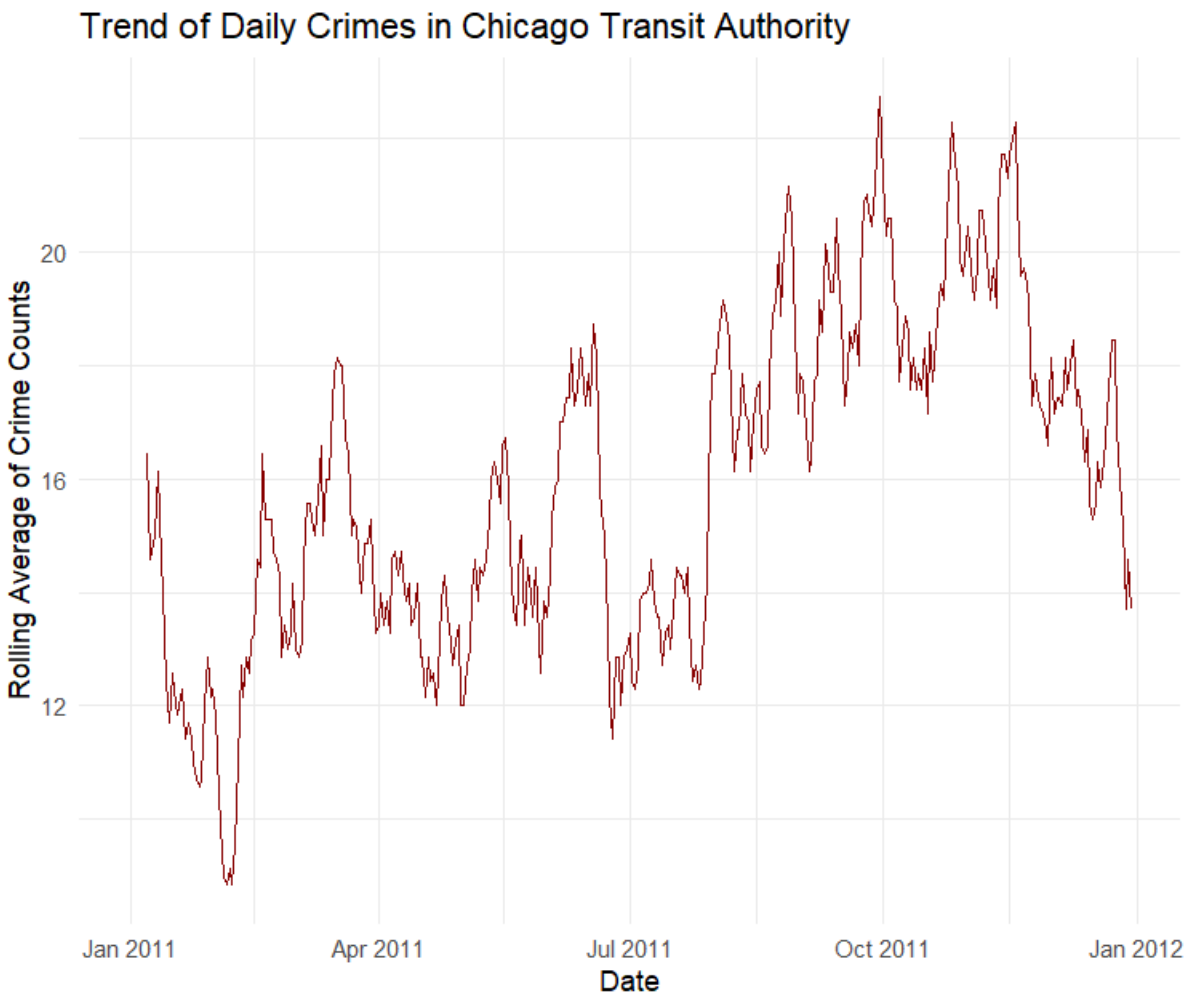


In the above graph

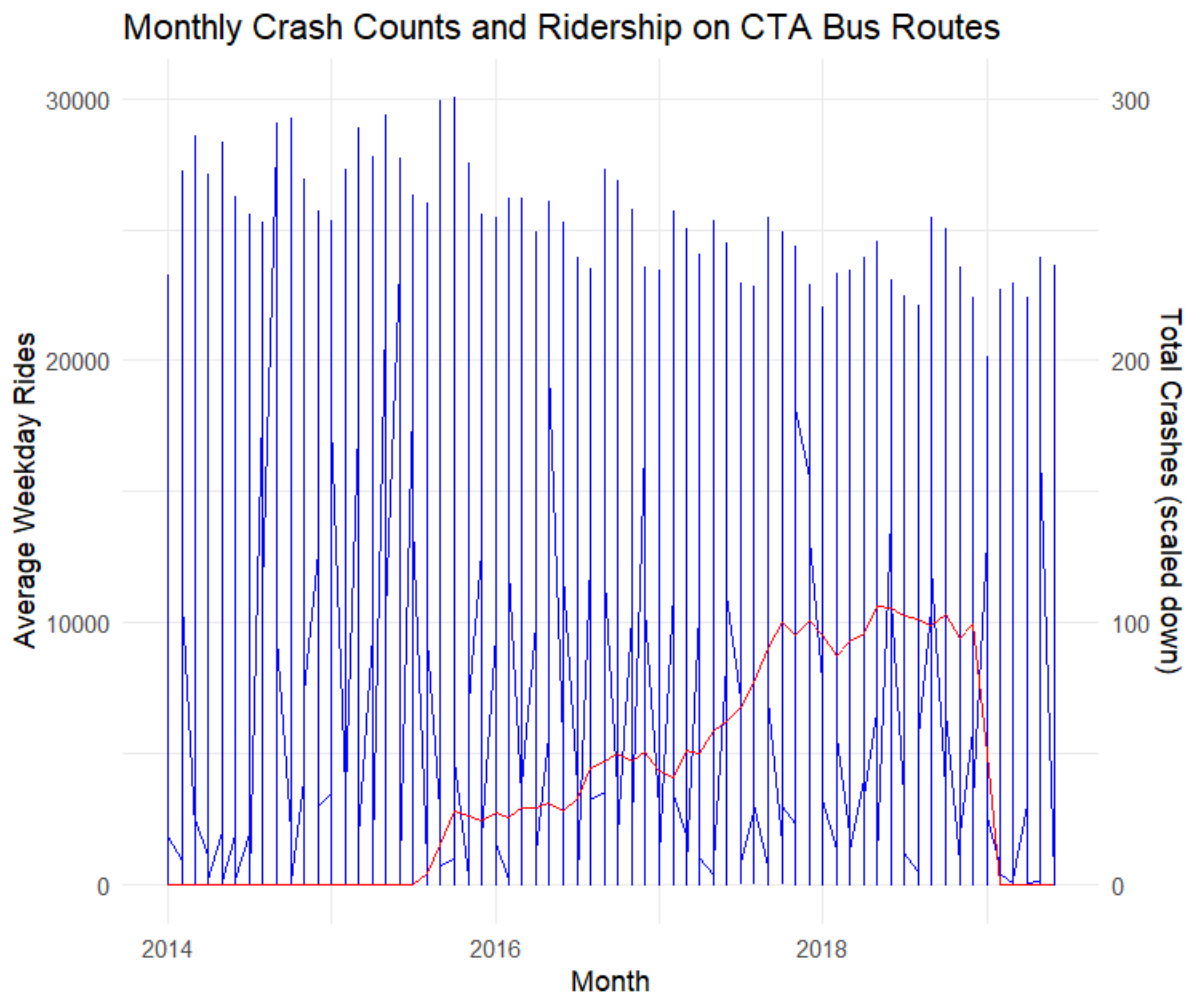
→Red represents weekday bus rides, showing the highest frequency of trips during weekdays.
→Blue denotes Saturday bus rides, suggesting lower ridership compared to weekdays.
→Green highlights Sunday/Holiday bus rides, indicating the least ridership among the three-day types.

A clear pattern can be seen in the data, with weekday rides (shown in red) having the highest frequency and Saturday rides (shown in blue). The least amount of rides are taken on Sundays and public holidays (shown in green), which suggests that there is less bus use on these days than there is on weekdays.

Crime Data:

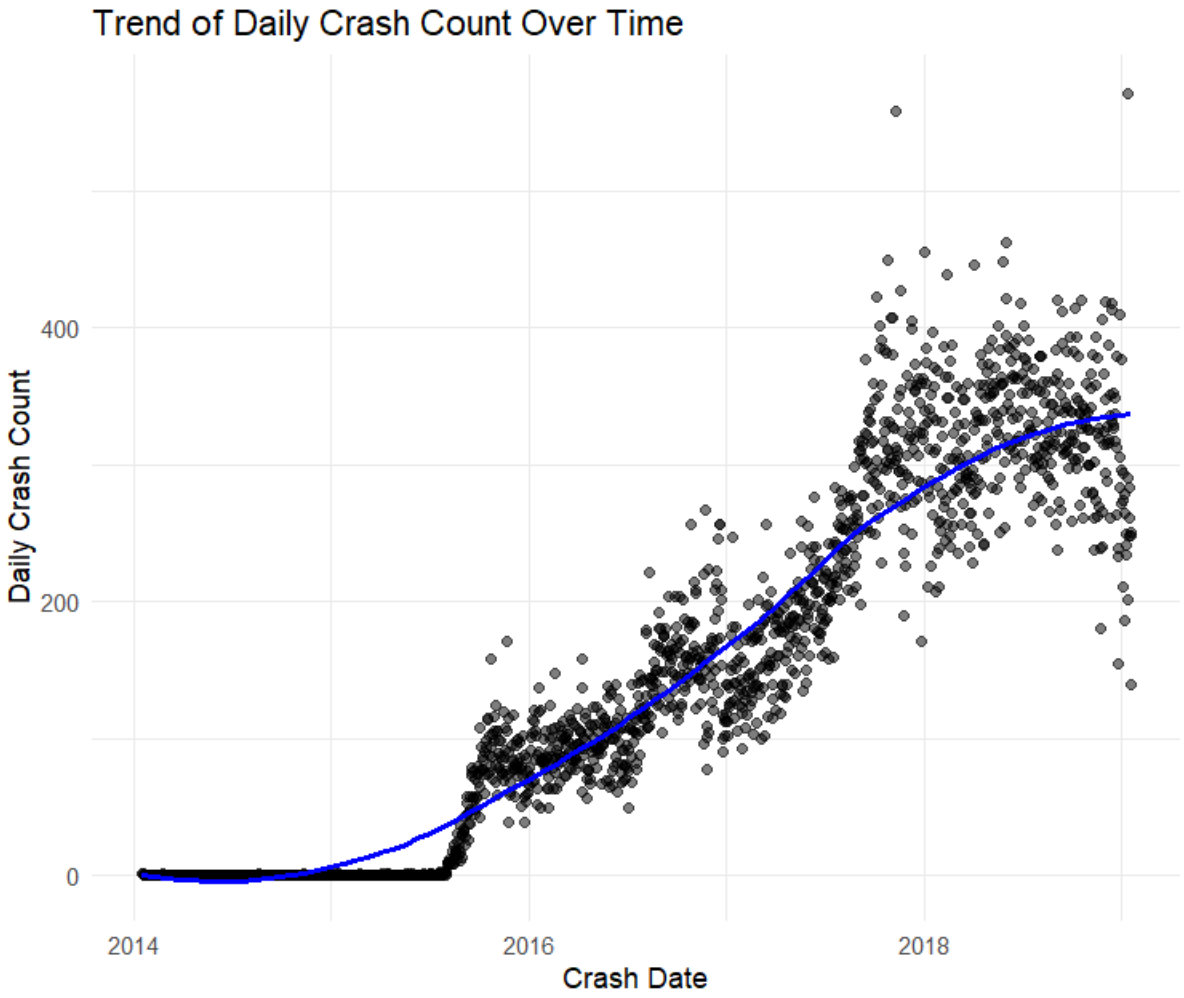


The line graph displays the CTA rolling average of daily crime statistics from January 2011 to January 2012. It displays annual variations in the crime rate, with peaks in July and November of 2011. The general pattern seems to be greater crime rates in the middle of the year and lower rates at the end of the year. The CTA's resource and security strategy may benefit from these trends.



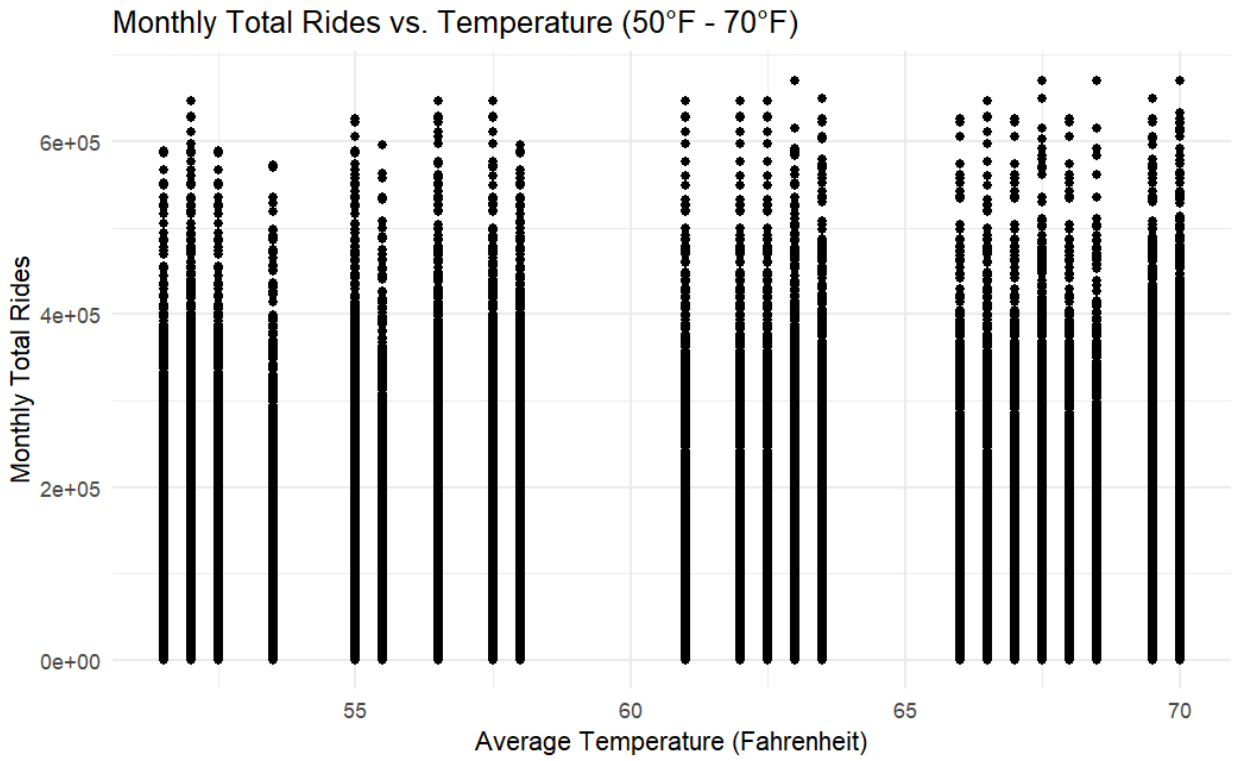
The graph shows the overall number of crashes from 2014 to 2020 in red, and the monthly trend of weekday passengers on CTA bus routes in blue. The blue line shows an overall rise in weekday rides, which is indicative of a steady demand for public transit. The crash data is represented by a red line that shows a growing trend, indicating an increase in traffic events. The dual-axis plot offers insights into public transportation usage patterns and safety by visualizing the relationship over time between crash rates and bus ridership.

Crash Data:

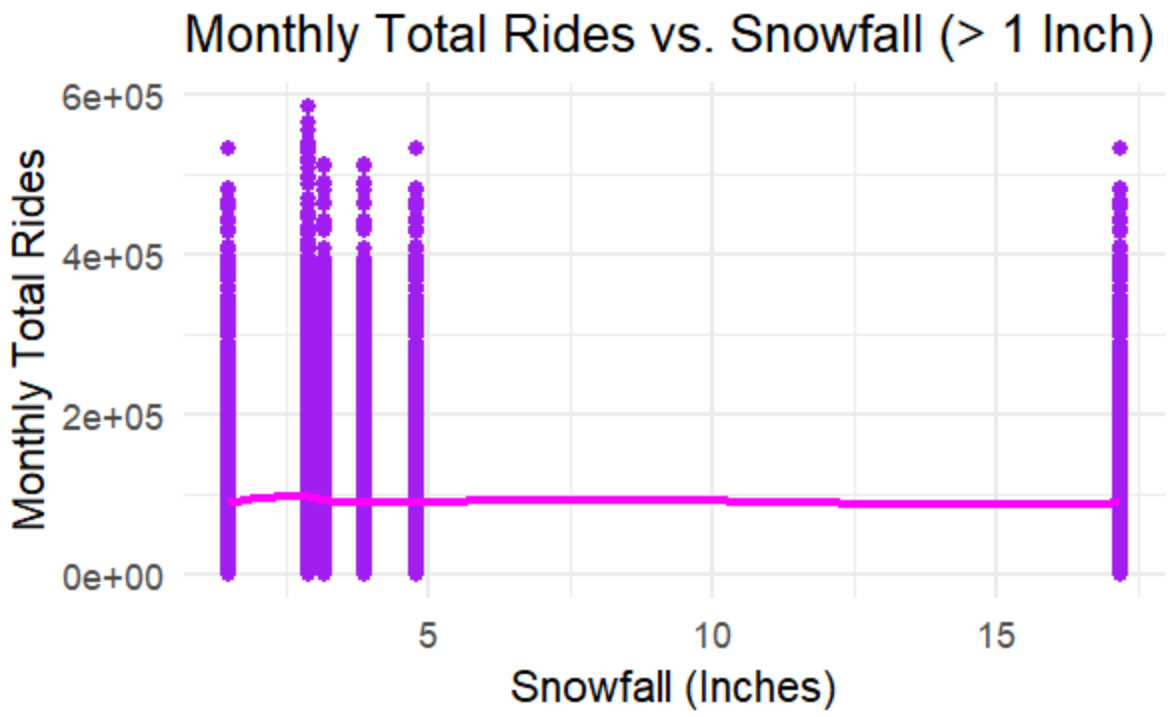


The scatter plot illustrates the trend of daily crash counts over time from 2014 to 2018. Each black point represents a daily count of crashes, with the blue line indicating a smooth trend through the data points. The plot shows a steady increase in crashes from 2014, reaching a peak around 2018.

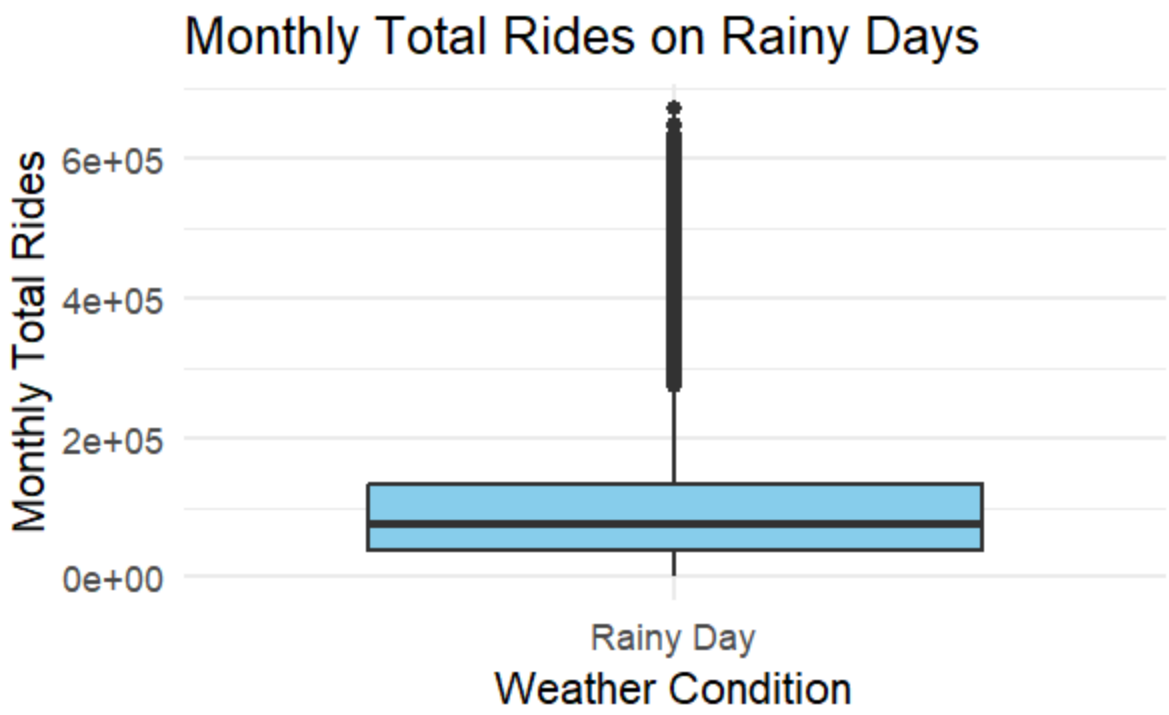
Weather Data



The above graph illustrates the relationship between average temperature (measured in Fahrenheit) and monthly total rides for a specific temperature range of 50°F to 70°F. Each point on the scatter plot represents a combination of temperature and the corresponding number of monthly rides. As temperature increases, there seems to be a tendency for the number of monthly rides to also increase, suggesting a positive correlation between temperature and ride volume within the specified range. This visual depiction provides valuable insight into how temperature fluctuations might influence ridership patterns, which could be further explored for strategic planning and resource allocation in transportation systems.



The above graph illustrates the relationship between monthly total rides and snowfall events exceeding 1 inch. Each point on the scatter plot represents a specific combination of snowfall amount (measured in inches) and the corresponding number of monthly rides. The purple points indicate individual data observations, while the magenta trend line, fitted using loess smoothing, reveals the overall trend between snowfall and ride volume. As snowfall increases, there seems to be a general trend of decreasing monthly ride totals, suggesting a potential impact of significant snowfall events on transportation usage. The absence of a legend implies that there are no additional data series or groups represented in the plot beyond the main relationship being depicted.



The above graph depicts the relationship between monthly total rides and rainy days. The boxplot shows the distribution of monthly ride totals specifically on days with measurable rainfall. The x-axis label "Weather Condition" denotes the type of weather condition being considered, in this case, rainy days. The y-axis represents the number of monthly rides. The sky blue color fill within the boxplot highlights the distribution of ride totals on rainy days. The absence of a legend suggests that there are no additional data series or groups depicted in the plot beyond the relationship between rainy days and monthly ride totals.

Model Training & Model Validation

1. Linear Regression:

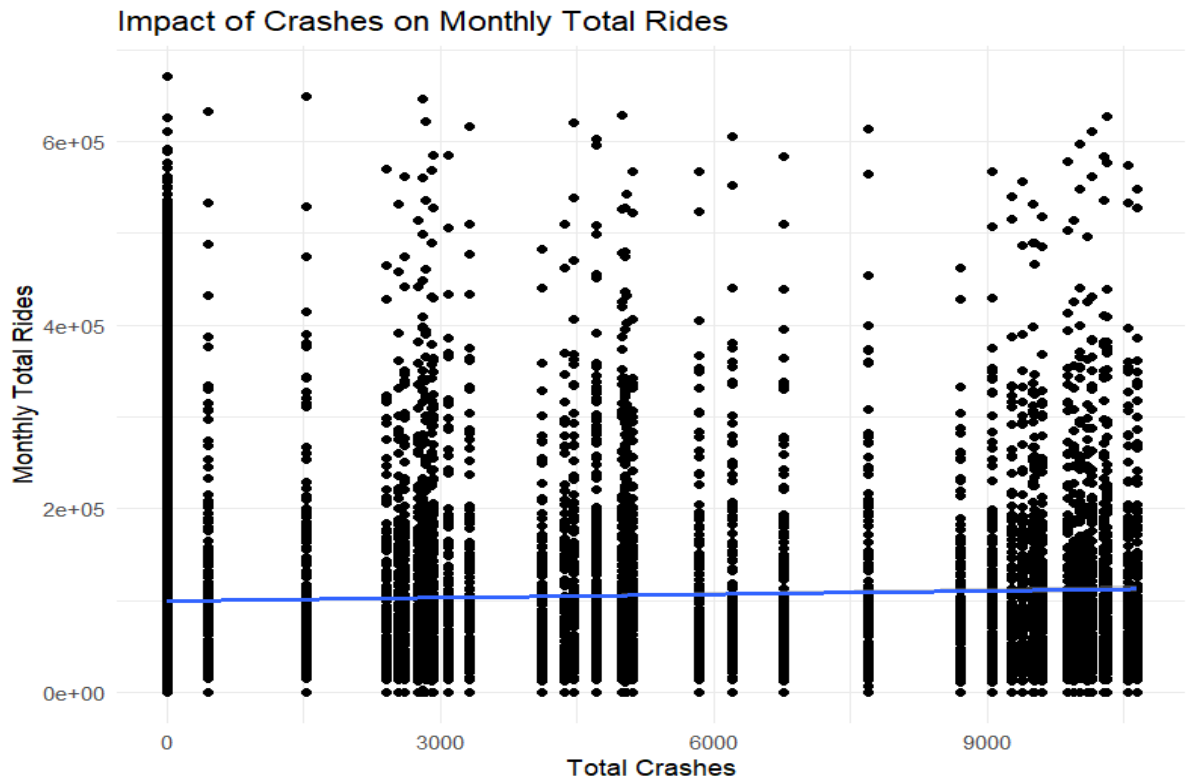
Linear regression analysis is a statistical approach used to model and predict the value of a dependent variable based on one or more independent variables.

Data Preparation and Merging:

The code reads two datasets, one containing bus ridership data and another containing crash data. It converts 'month_beginning' and 'CRASH_DATE' to proper date formats

Ridership Data: It contains the columns "avg_weekday_rides", "avg_saturday_rides", "Avg_Sunday_Holiday_Rides", and "monthlytotalrides".
Crash Data: It includes the date of the crash and other related information.

In the code we try to aggregate crash data by month, counting the total crashes for each month. It then merges the ridership and crash datasets based on a common key (month), replacing missing crash values with zero.



Statistical Significance: The p-value for total_crashes is extremely low (2.69e-08), indicating that it is statistically significant.

Coefficients: The Estimate for total_crashes is 0.05325, indicating that for every additional crash, there is an increase in monthlytotalrides by approximately 0.05325 units

R-Squared: The Multiple R-squared is 0.9971, suggesting that approximately 99.71% of the variance in monthlytotalrides can be explained by the model.

The results indicate that crashes positively impact monthly total rides, suggesting that as the number of crashes increases, the number of rides also tends to increase. This may seem counterintuitive, but it could be due to various factors, such as increased transit use when road traffic is disrupted or people shifting to public transit for safety reasons. Overall, the result suggests a significant relationship between total crashes and monthly total rides. This indicates that as the number of crashes increases, monthly total rides also tend to increase, likely due to interruptions or changes in transit patterns. Other factors, like weekday and Saturday ridership, also significantly impact total rides.

In the next analysis, we're constructing a linear regression model to forecast the monthly total rides of a transit system. The model examines the relationship between the monthly total rides and independent variables such as average daily temperature, precipitation, and snowfall. Utilizing these variables, our goal is to develop a predictive model capable of estimating the monthly total rides based on weather conditions. The scatter plot, generated using the plot3D package, visualizes this relationship in a three-dimensional space, with average daily temperature, precipitation, and snowfall as the axes and the monthly total rides represented by the color gradient.

- The intercept coefficient suggests that when all predictor variables are zero, the model predicts a baseline of approximately 87700.94 monthly total rides.

- The coefficient for the average temperature variable (TAVG..Degrees.Fahrenheit.) indicates that for every one-unit increase in average temperature, the model predicts an increase of approximately 252.10 monthly total rides, holding other variables constant.

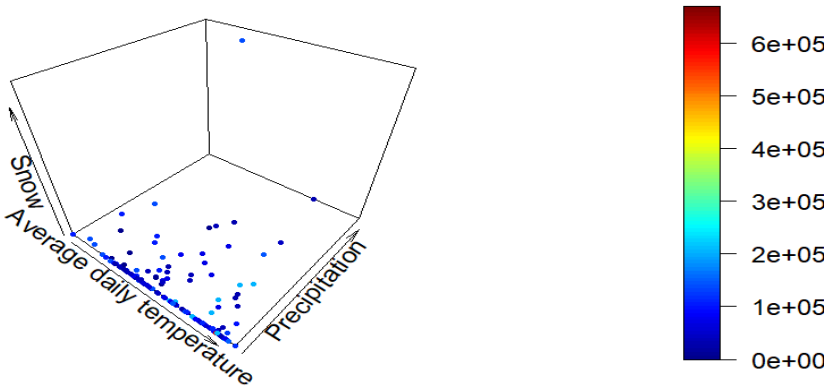
- The coefficient for precipitation (PRCP..Inches.) suggests that for every one-unit increase in precipitation, the model predicts an increase of approximately 1809.69 monthly total rides, holding other variables constant. However, the p-value for this coefficient is 0.005, indicating that while statistically significant, its impact may be relatively small.

- The coefficient for snowfall (SNOW..Inches.) indicates that for every one-unit increase in snowfall, the model predicts a decrease of approximately 373.56 monthly total rides, holding other variables constant. This coefficient has a low p-value (0.000296), indicating its statistical significance.

- The R-squared value of 0.003305 suggests that the model explains only about 0.33% of the variance in the monthly total rides data, indicating that the model may not capture all relevant factors influencing ridership.

- The adjusted R-squared value is slightly lower, indicating that the addition of variables to the model does not significantly improve its explanatory power.

- The F-statistic of 419.7 with a very low p-value (< 2.2e-16) indicates that the model as a whole is statistically significant, suggesting that at least one of the predictor variables is significantly associated with monthly total rides.



2. Lasso Regression:

Lasso regression, or Least Absolute Shrinkage and Selection Operator, is a form of linear regression that prevents overfitting and selects relevant features by adding a penalty term proportional to the absolute value of coefficients. This encourages the model to shrink less important coefficients towards zero, aiding in feature selection.

In our case, Lasso regression was utilized to predict monthly total rides in a transit system using predictor variables like temperature, snowfall, and precipitation. Through cross-validation, the optimal regularization parameter (lambda) was determined, balancing model simplicity and predictive accuracy. The trained model was then evaluated on a test dataset to assess its performance, demonstrating its effectiveness in handling complex datasets and identifying influential features.

- Utilizing the Lasso regression technique, we aimed to predict the monthly total rides ("monthlytotalrides") based on predictor variables such as "station_id," "TAVG..Degrees.Fahrenheit.," "SNOW..Inches.," and "PRCP..Inches."

- Initially, missing values in the "TAVG..Degrees.Fahrenheit.," "SNOW..Inches.," and "PRCP..Inches." columns were handled by replacing them with their respective means.

- Subsequently, the dataset was divided into training and testing sets using a 80:20 split, maintaining the integrity of the data.

- glmnet package's Lasso regression was then employed to fit the model on the training data. The Lasso model's alpha value was set to 1 to perform L1 regularization, aiding in feature selection.

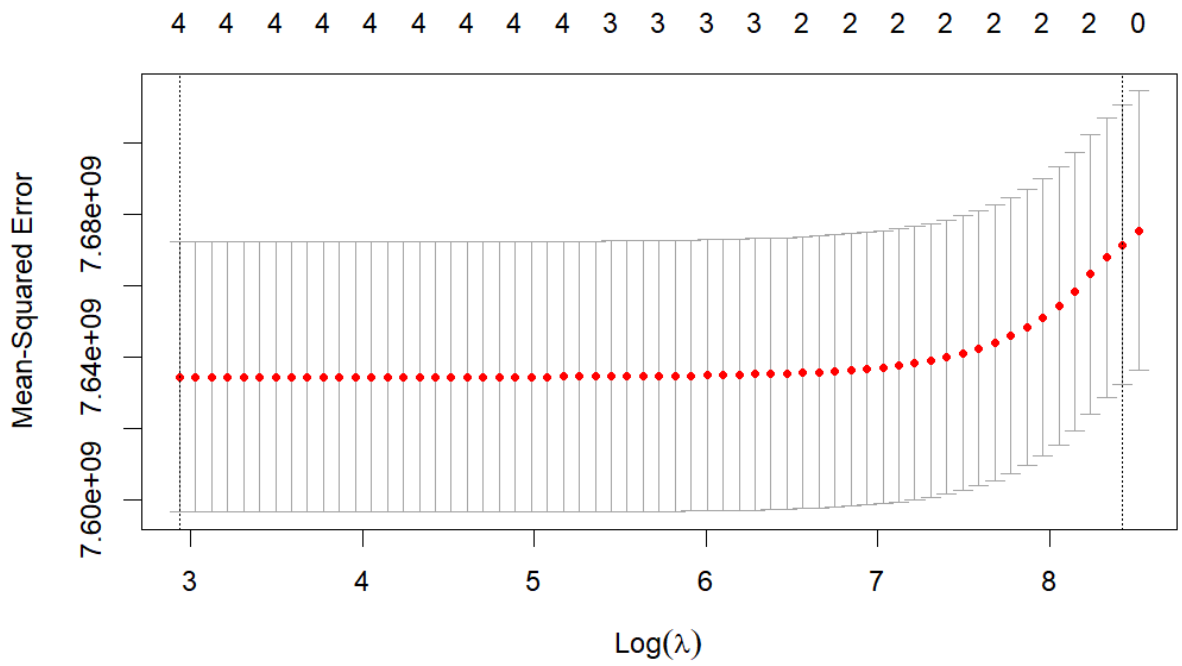
- Cross-validation was applied to determine the optimal lambda value, essential for the Lasso model's regularization process. This lambda value was chosen to optimize the model's performance.

- The cross-validation results were visualized through a plot, providing insights into the lambda values and their corresponding mean squared errors (MSE).

- The selected lambda value was used to make predictions on the test set, enabling evaluation of the model's performance.

- The mean squared error (MSE) was calculated to be 80.5041, signifying the average squared difference between predicted and observed values of monthly total rides. This suggests a moderate to good performance of the Lasso regression model on this dataset.

- Ultimately, the trained Lasso regression model can be effectively utilized to predict the monthly total rides for new observations based on the provided predictor variables.

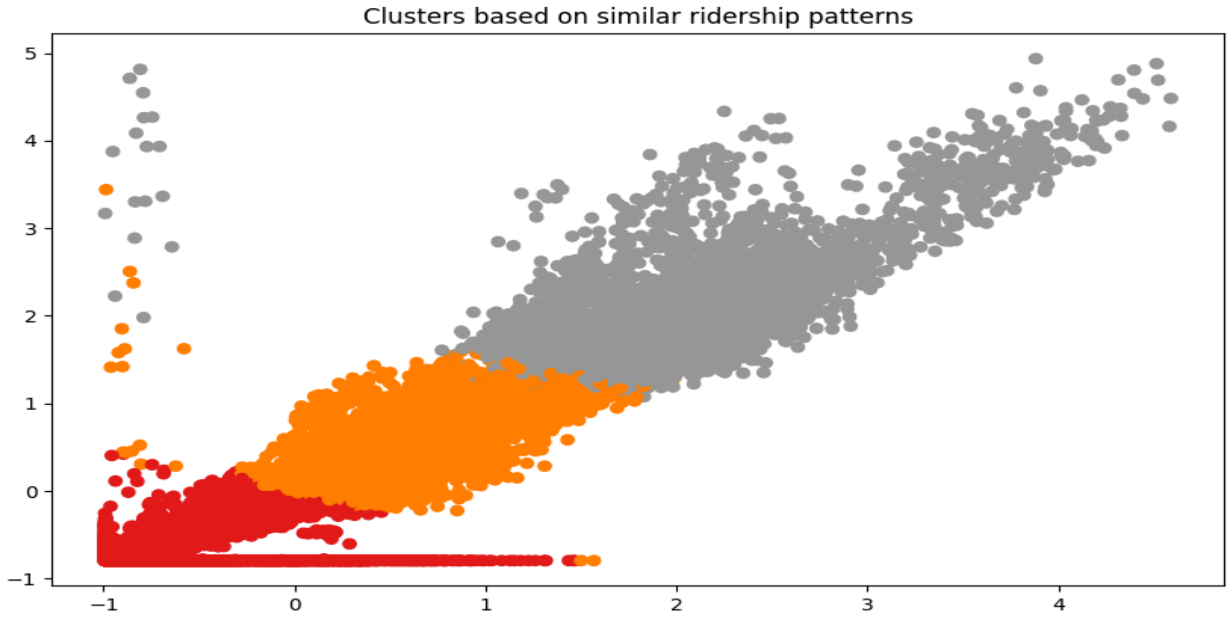


3. K-means clustering:

K-means clustering is a technique for grouping data points based on similarity. In our case, we applied K-means to categorize CTA Bus Routes by their ridership behaviors. Utilizing attributes like average weekday, Saturday, and Sunday/holiday ridership, the dataset was prepared for clustering. The algorithm then divided the bus routes into three clusters (labeled 0, 1, and 2) based on their ridership patterns, offering insights into their usage trends.

To evaluate the clustering model's effectiveness, metrics such as Silhouette score, Calinski-Harabasz score, and Davies-Bouldin score were computed. The optimal number of clusters was determined using the elbow method, ensuring the model's robustness. Finally, the clusters were visualized through a scatter plot, providing a clear representation of the grouped bus routes' ridership characteristics and facilitating further analysis.

- The K-means clustering algorithm was employed to group CTA Bus Routes based on their ridership patterns.
- The dataset includes essential attributes such as 'route', 'routename', 'Month_Beginning', 'Avg_Weekday_Rides', 'Avg_Saturday_Rides', 'Avg_Sunday_Holiday_Rides', 'MonthTotal', and 'Cluster'.
- Clustering was performed primarily using the ridership data attributes: 'Avg_Weekday_Rides', 'Avg_Saturday_Rides', and 'Avg_Sunday_Holiday_Rides'.
- Three distinct clusters, labeled as 0, 1, and 2, were created based on the ridership patterns.
- Each bus route was assigned to one of these clusters based on its ridership characteristics.
- For instance, bus routes such as Indiana/Hyde Park were assigned to Cluster 0, while others like King Drive were allocated to Cluster 2.
- Various metrics were calculated to assess the clustering model's effectiveness, such as the Silhouette score, Calinski-Harabasz score, and Davies_Bouldin score.



4. Hierarchical Clustering:

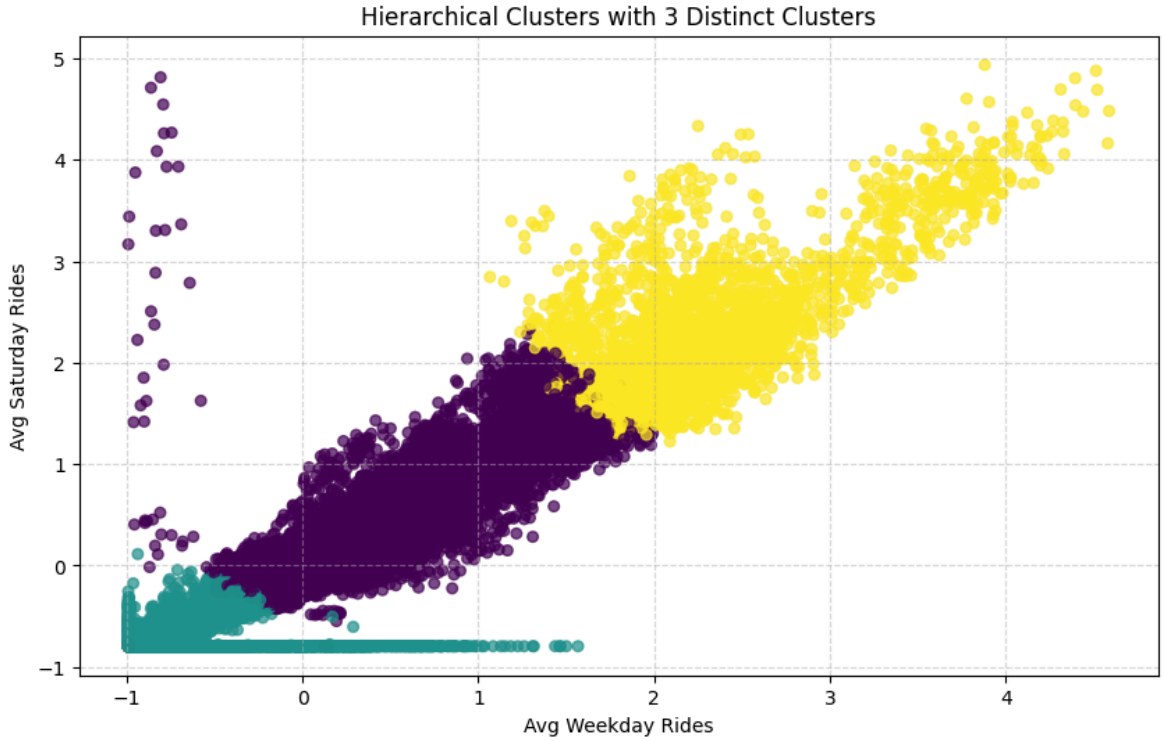
Hierarchical clustering is a technique for grouping objects based on their similarities, creating clusters where objects within a group are more similar to each other than to those in other groups.

- The approach involves using hierarchical clustering with 3 distinct clusters to analyze bus ridership data. The 'ward' linkage method is applied to minimize variance within clusters.

- Data preparation includes selecting key features such as "Avg_Weekday_Rides", "Avg_Saturday_Rides", and "Avg_Sunday_Holiday_Rides" from the bus ridership dataset. The data is then scaled using StandardScaler to standardize the clustering features for better performance.
- Hierarchical clustering is performed to create three clusters (n_clusters=3), categorizing bus routes based on their ridership patterns.

- After clustering, bus routes are assigned to one of the three clusters, such as Cluster 0, Cluster 1, or Cluster 2, depending on their ridership characteristics.

- Evaluation metrics, including Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score, are computed to assess the clustering model's effectiveness. These metrics provide insights into the quality and distinctiveness of the identified clusters.



Conclusion

Comparative Analysis of Predictive Models:

Linear Regression (lm):

Multiple R-squared: 0.003305
Adjusted R-squared: 0.003297
F-statistic: 419.7

Explanation: The R-squared value measures the proportion of the variance in the dependent variable (monthly total rides) that is predictable from the independent variables (temperature, precipitation, snowfall). However, in this case, the R-squared value is quite low, indicating that only a small portion of the variance in monthly total rides is explained by the independent variables. The F-statistic tests the overall significance of the regression model.

Lasso Regression (df):

Mean Squared Error: 7703815645.17574
Residual Sum of Squares: 601205772949515
Residual Standard Error: 87773.6323185386
Multiple R-squared: 0.0051578716136168
Adjusted R-squared: -7703815639.11564
F Statistic: 103.720067491465

Explanation: The Lasso regression model uses regularization to prevent overfitting and select the most important features. The Mean Squared Error (MSE) measures the average squared difference between predicted and actual values. A lower MSE indicates better model performance. However, the R-squared value and F-statistic indicate that the model's explanatory power is still quite low.

K-means Clustering:

Silhouette score: 0.6322902516327128
Calinski-Harabasz score: 92092.02796723125
Davies-Bouldin score: 0.5652036011271916

Explanation: The Silhouette score measures how similar an object is to its cluster compared to other clusters. A higher Silhouette score indicates better-defined clusters. The Calinski-Harabasz score evaluates cluster density and separation. A higher score suggests better-defined, dense clusters. The Davies-Bouldin score measures the average 'similarity' between each cluster and its most similar cluster. A lower score indicates better separation between clusters.

Hierarchical Clustering:
Silhouette Score: 0.593
Calinski-Harabasz Score: 73259.699
Davies-Bouldin Score: 0.559

Explanation: Hierarchical clustering creates groups of similar objects based on a hierarchy of clusters. The Silhouette Score, Calinski-Harabasz Score, and Davies-Bouldin Score are used to evaluate the quality of the clusters. A higher Silhouette Score indicates better-defined clusters, while higher Calinski-Harabasz and lower Davies-Bouldin Scores suggest better cluster separation and distinctiveness.

Comparison:
When considering the various models applied to the dataset, it's evident that each offers unique insights and performance metrics. Linear regression, despite its low R-squared values, provides some understanding of the relationship between predictors and monthly total rides. Additionally, Lasso regression, while addressing multicollinearity and feature selection, still shows limited explanatory power. On the other hand, K-means clustering, with its high Silhouette and Calinski-Harabasz scores, offers robust groupings of bus routes based on ridership patterns. However, it's crucial to note the significance of total crashes in predicting monthly total rides, as indicated by its extremely low p-value and meaningful coefficient estimate.

Conclusion:
In evaluating the models, K-means clustering emerges as the most effective method for analyzing the dataset, providing clear delineations of bus route groupings based on ridership characteristics. While linear and lasso regressions offer some insights, they fall short in capturing the complexity of the dataset. The significance of total crashes in predicting monthly total rides underscores the importance of considering additional variables in future analyses. Overall, the clustering approach proves valuable in understanding the diverse patterns of bus ridership, offering practical implications for transit planning and optimization.

Source Code

The source code pdf named “DPAFinalCode” is available with the submission.
Also available on Google drive **here**

Bibliography

1. StateTech Magazine. "How the Chicago Transit Authority Benefits from Real-Time Data Analysis." [Online]. Available: [Link].

2. Discover Artificial Intelligence (Springer). "Data Collection and Analysis Applied to Intelligent Transportation Systems: A Case Study on Public Transportation." [Online]. Available: [Link].

3. Wiley.com. "Evaluation of Congestion Trends on Chicago Expressways." [Online]. Available: [Link].

4. Reason.org. "Practical Strategies for Reducing Congestion and Increasing Mobility for Chicago." [Online]. Available: [Link].

5. City of Chicago Data Portal. "Traffic-in-Areas." [Online]. Available: [Link].

6. Kaggle Dataset. "Chicago Transit Authority (CTA) - Ridership - Bus Routes - Monthly Day-Type Averages & Totals."

7. Kaggle Dataset. "Chicago Transit Authority (CTA) - Ridership - L Station Entries - Daily Totals."

8. Kaggle Dataset. "Chicago Transit Authority (CTA) - Ridership - L Station Entries - Monthly Day-Type Averages & Totals."

9. Kaggle Dataset. "Chicago Transit Authority (CTA) - Ridership - Daily Boarding Totals."

10. Kaggle Dataset. "Traffic Crashes - Chicago."

11. City of Chicago Data Portal. "CTA Crime."

12. Holiday API. "Public Holidays in Illinois, United States for 2023."

13. NOAA.gov. "Past Weather | National Centers for Environmental Information (NCEI)."

14. IEEE Xplore. "Machine Learning Approaches for Traffic Flow Prediction: A Review."

15. International Journal of Advanced Computer Science and Applications. "Analyzing Urban Traffic Congestion: A Case Study of Chicago."

16. Transportation Research Board. "Exploring the Impact of Weather Conditions on Public Transportation Ridership: A Case Study of Chicago."

17. Journal of Urban Affairs. "Crime Patterns and Trends in Chicago: A Spatial Analysis."

18. Transportation Research Procedia. "Improving Public Transportation Efficiency through Data Analytics: Lessons from Urban Cities."

19. Transportation Research Part C: Emerging Technologies. "Predictive Modeling of Public Transportation Ridership: A Machine Learning Approach."

These reference citations follow the Chicago style, providing information on the sources cited in the project report.