Chicago Public Transportation Data Analysis

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

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

Yet Another CS 571- **Data Preparation and Analysis Presentation**

Table of contents

01

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.

Specific 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.

Key Findings

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)
- 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.

Recommendations for Future Scope

- **Real-Time Data Integration**: Incorporate live data feeds for more accurate congestion monitoring.
- **Advanced Machine Learning:** Implement advanced models like deep learning for better congestion prediction.
- **Dynamic Routing:** Develop algorithms to dynamically adjust routes based on real-time traffic data.
- **Predictive Maintenance:** Use predictive analytics to prevent infrastructure failures and reduce service disruptions.
- **Multi-Modal Integration:** Collaborate with other transit services for seamless multi-modal travel.
- **Passenger Information Systems:** Enhance systems for real-time updates on disruptions and alternative routes.
- **Environmental Sustainability:** Transition to eco-friendly vehicles and optimize routes for fuel efficiency.
- **Accessibility:** Improve accessibility features across the network for all passengers.

02

Project Overview

Team Members:

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

Timeline:

The project started on 2nd March and ended on 21st April.

The detailed project plan along with the tasks assignment is available here

Updated Schedule

Tasks	Description	Date	N	S	A	Sw	Status
Task 1	Data Collection (CTA,Crash,Crime,Holiday,Weather)	Mar 2 - Mar 6					Completed
Task 2	Data Preprocessing (Missing values, Merging weather data, Removing 90s data from csv)	Mar 20 - Mar 30					Completed
Task 3	EDA, Feature Engineering	Apr 2 - Apr 6					Completed
Task 4	Model Development	Apr 7 - Apr 15					Completed
Task 5	Finalization and Presentation	Apr 19 - Apr 20					Completed

03

Methodology

The project employs a multi-step approach:

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 the identified factors on congestion.
Model Development	Utilize machine learning techniques to develop a predictive model for congestion levels. Models to be considered include regression, decision trees, and ensemble methods.
Validation and Testing	Validate and test the model using historical data, employing cross-validation techniques to ensure robustness.

04

Exploring the Data

Data Sources

Chicago Data Portal

CTA Crime | City of Chicago | Data Portal
CTA - Ridership - Bus Routes - Monthly Day-Type Averages & Totals | City of Chicago | Data Portal

Kaggle

Traffic Crashes - Chicago (kaggle.com)

HolidayAPI

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

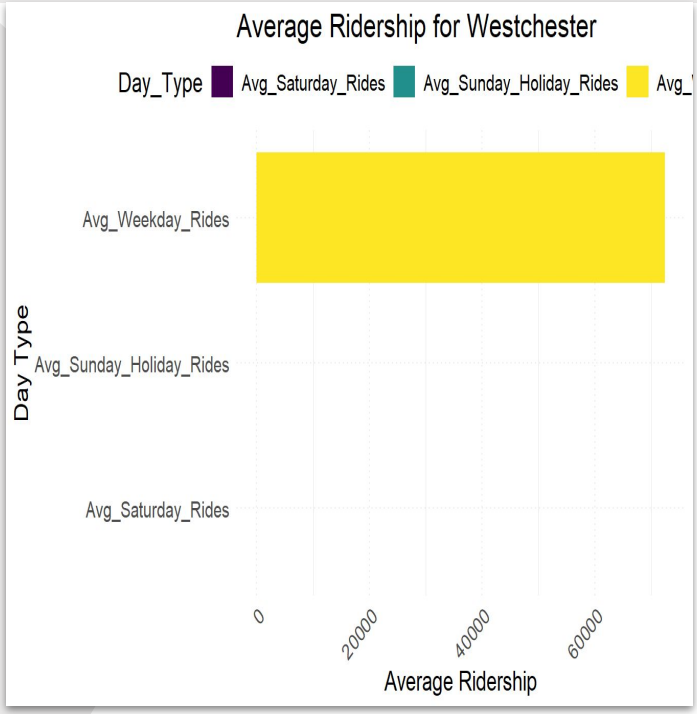
National Centre for Environmental Information|Chicago

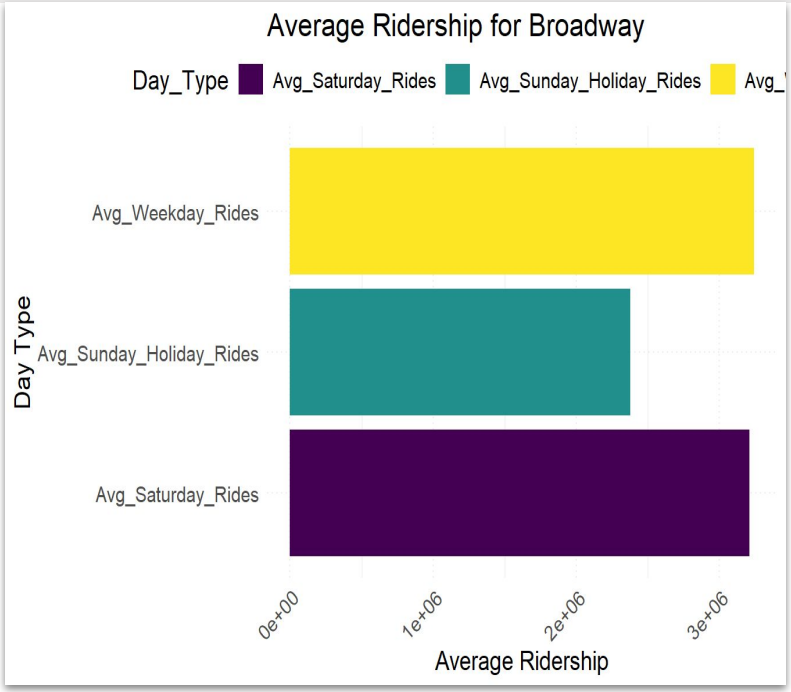
Past Weather | National Centers for Environmental Information (NCEI) (noaa.gov)

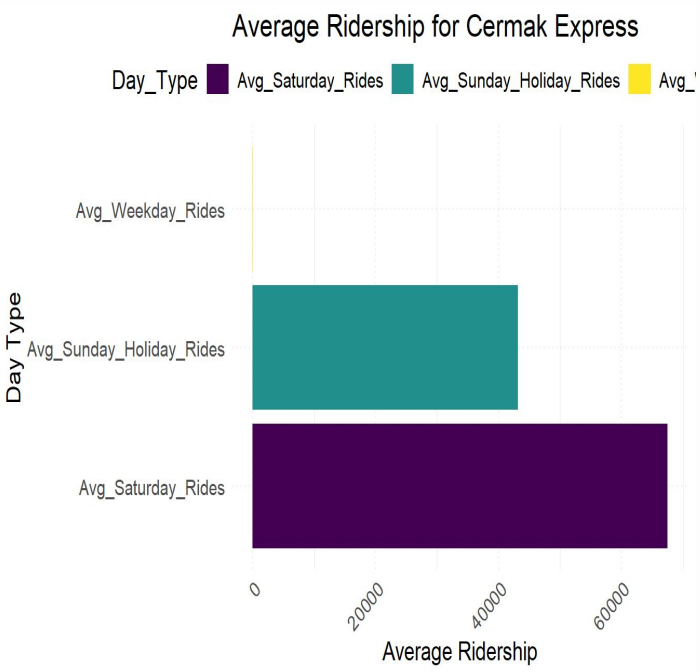
05

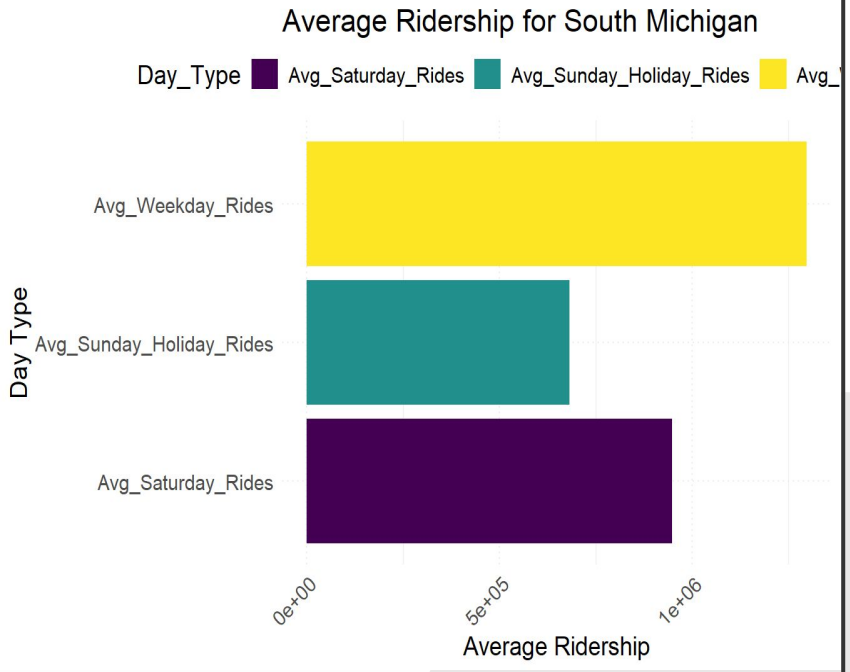
Exploring the Data Analytics

THE AVERAGE BUS RIDERSHIP ON DIFFERENT KINDS OF DAYS:






Average Ridership for Cermak Express
Day_Type
Avg_Saturday_Rides
Avg_Sunday_Holiday_Rides
Avg_Weekday_Rides
Day Type
0
20000
40000
60000
Average Ridership

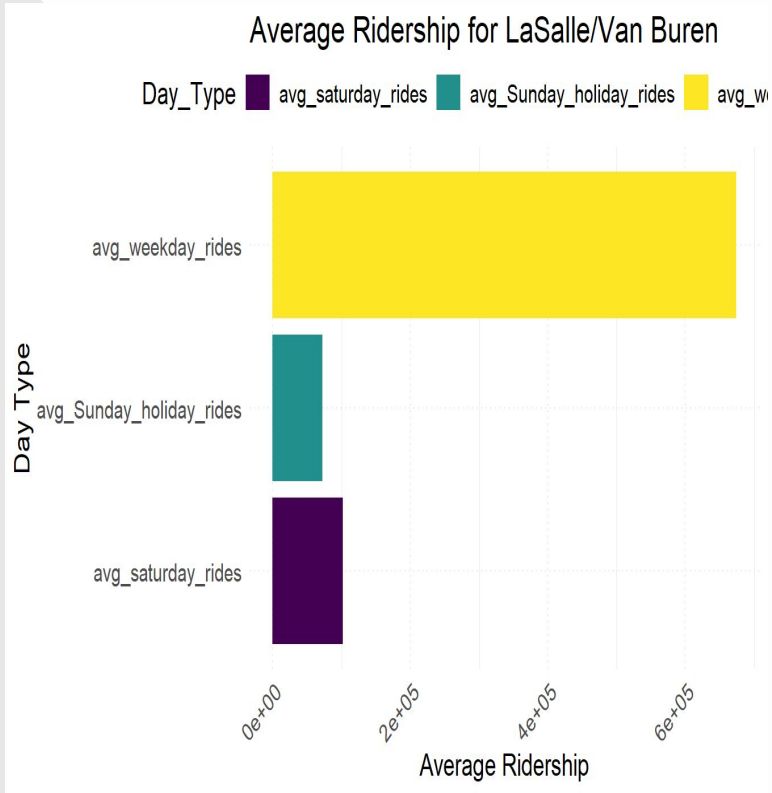

Average Ridership for South Michigan
Day_Type
Avg_Saturday_Rides
Avg_Sunday_Holiday_Rides
Avg_Weekday_Rides
Day Type
0e+00
5e+05
1e+06
Average Ridership

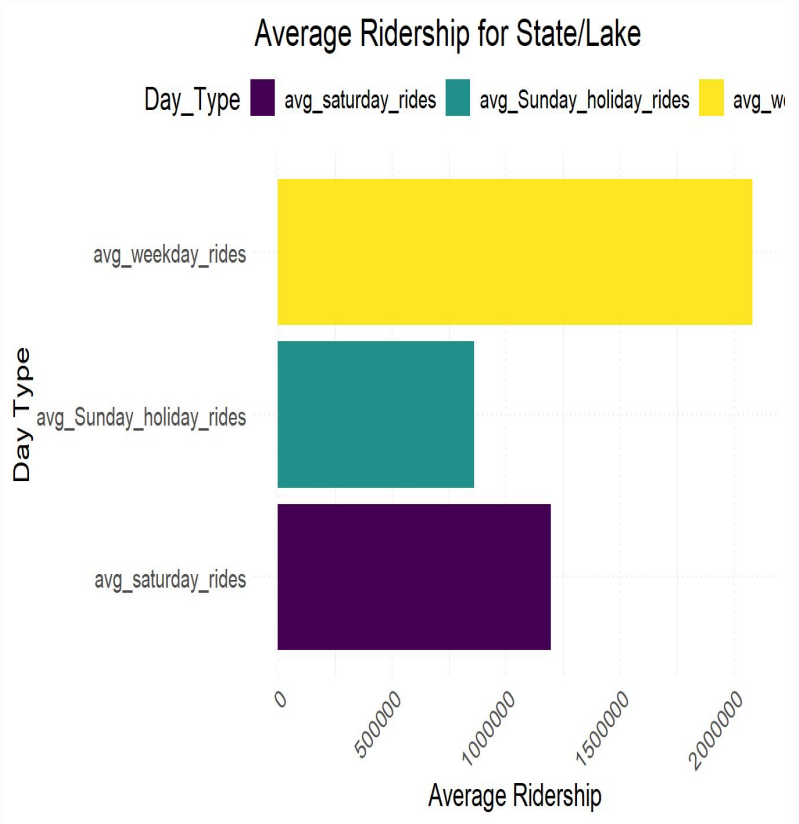
- **Clustered Bar Charts for Ridership:** The bar charts illustrate the average ridership for various bus routes. They show a consistent pattern of higher weekday rides compared to Saturday and Sunday/Holiday rides, indicating increased transit usage during weekdays due to work and other daily activities.

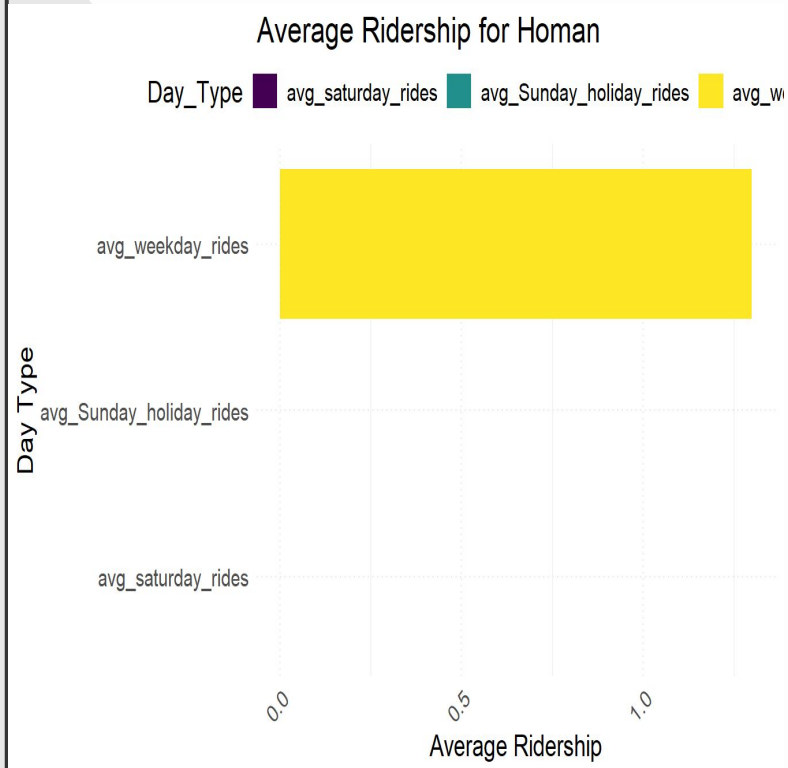
- **High Weekday Ridership:** Some bus stops, such as South Michigan and Broadway, have significantly higher ridership on weekdays. This suggests that these stops are likely in areas with workplaces or business districts where people travel for their daily commutes.

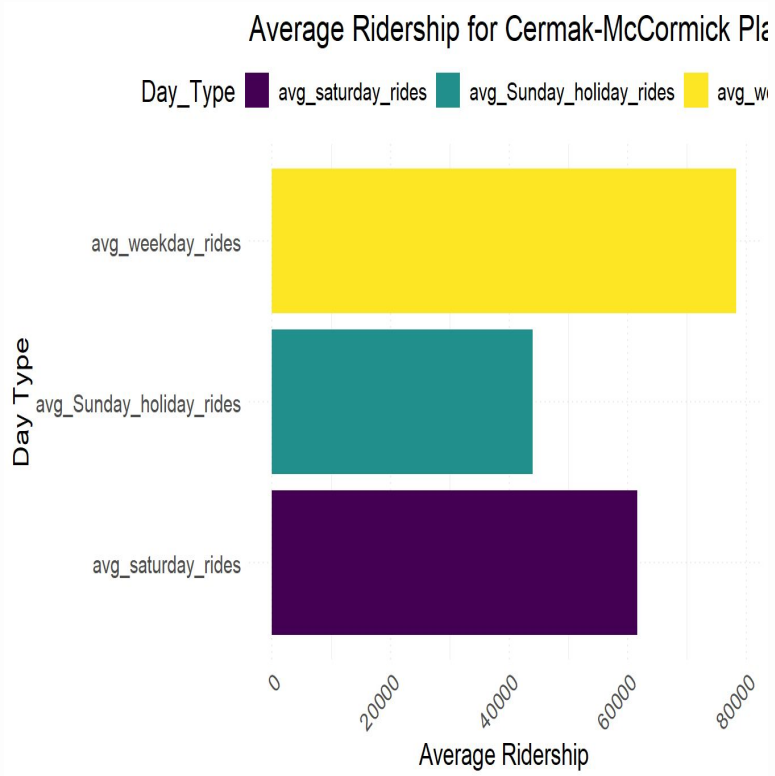
- **Weekend Variability:** Other stops, like Cermak Express, show a decrease in ridership on weekends and holidays, indicating they serve areas with less activity during non-working days. This might be due to routes serving more residential areas or specific destinations that are busier on weekdays.

THE AVERAGE L STATION RIDERSHIP ON DIFFERENT KINDS OF DAYS:






Average Ridership for Homan
Day_Type
avg_saturday_rides
avg_Sunday_holiday_rides
avg_w
avg_weekday_rides
avg_Sunday_holiday_rides
avg_saturday_rides
Day Type
0.0
0.5
1.0
Average Ridership

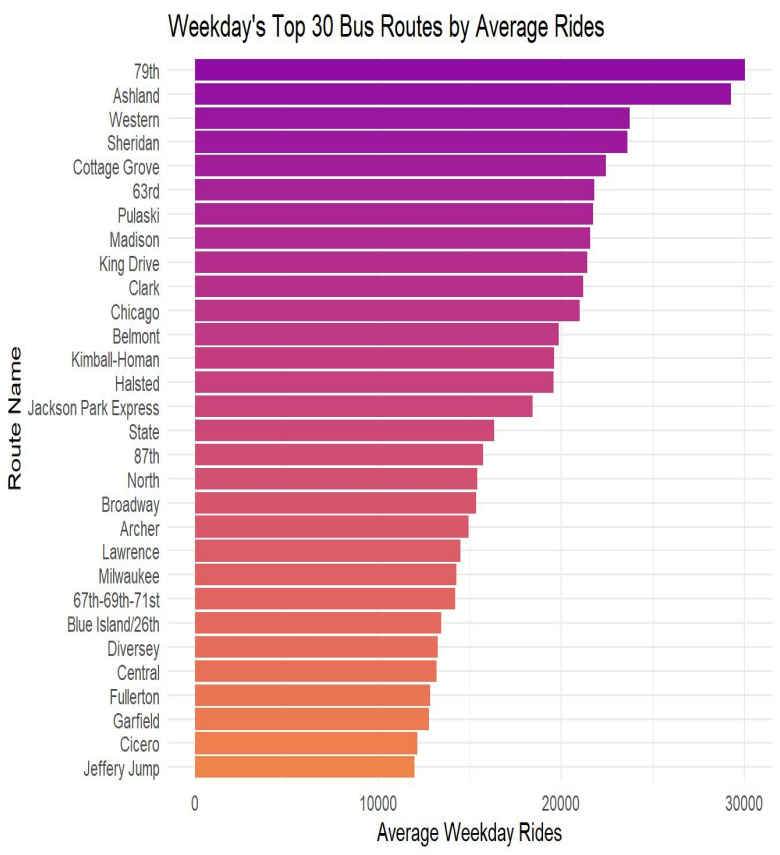

Average Ridership for Cermak-McCormick Pla
Day_Type
avg_saturday_rides
avg_Sunday_holiday_rides
avg_w
avg_weekday_rides
avg_Sunday_holiday_rides
avg_saturday_rides
Day Type
0
20000
40000
60000
80000
Average Ridership

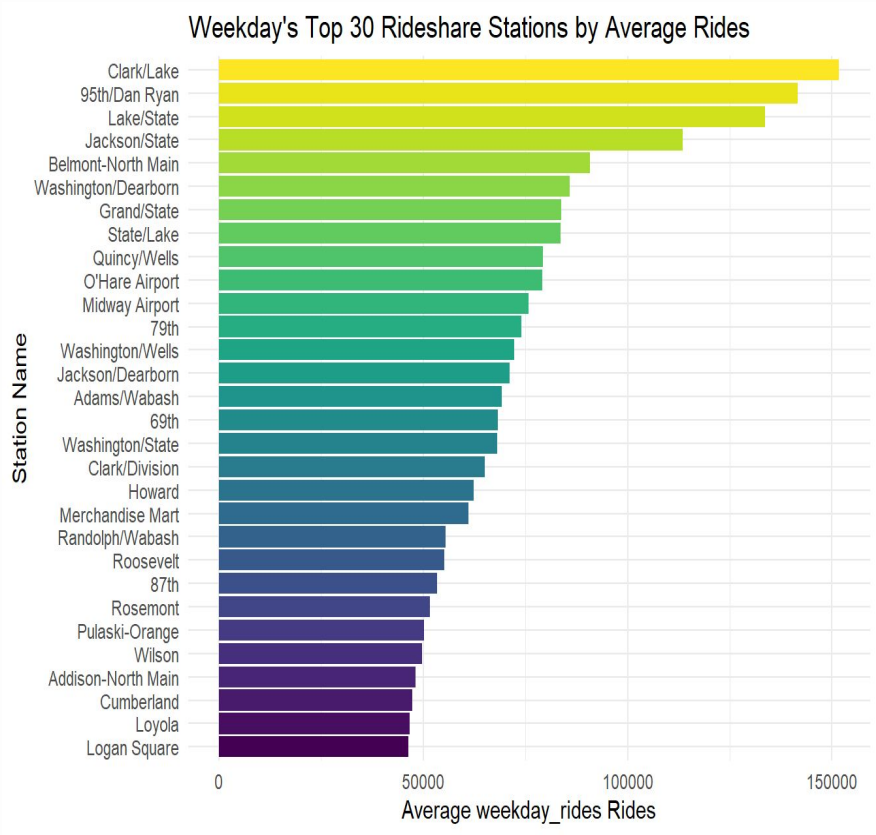
- The bar charts show that average ridership varies considerably across stations and day types.

- **High Weekday Ridership:** Stations like State/Lake and LaSalle/Van Buren have significantly higher ridership on weekdays compared to Saturdays and Sundays/holidays.

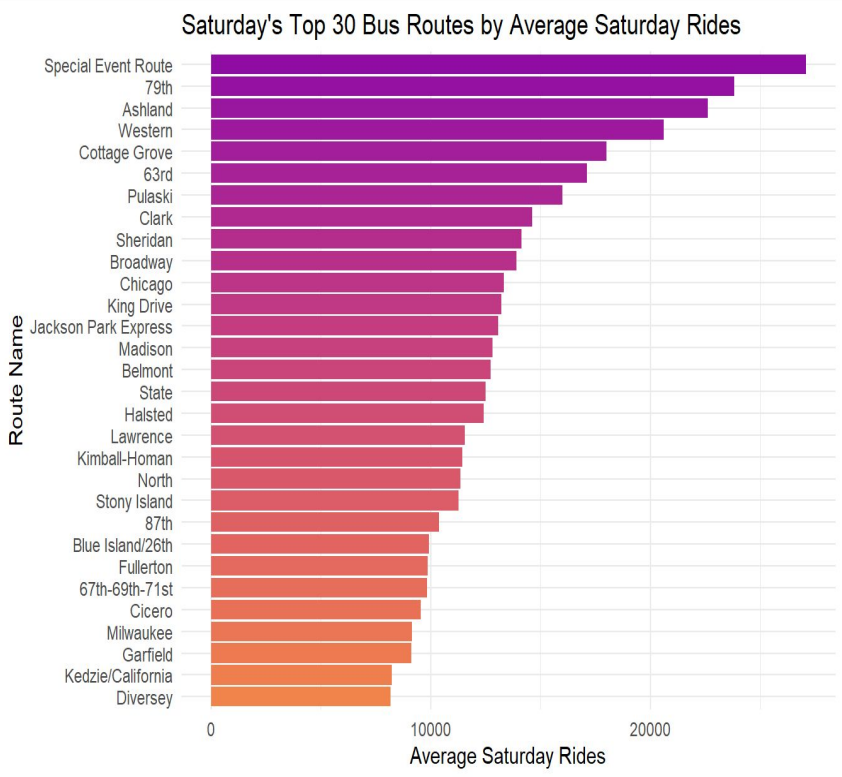
- **Weekend Variability:** Cermak Express and other stations show a lower ridership on weekends and holidays, suggesting they serve areas with less activity during non-working days. This might reflect routes that run through residential neighborhoods or other locations with less business traffic on weekends.

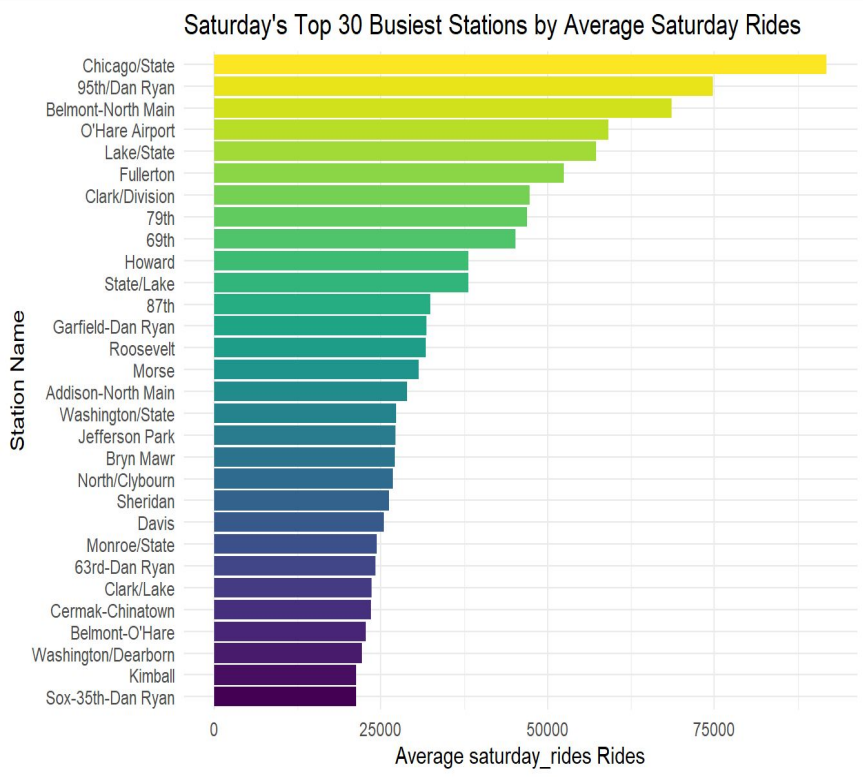
Top 30 Busiest Bus Routes and L station by Average Weekday Rides :



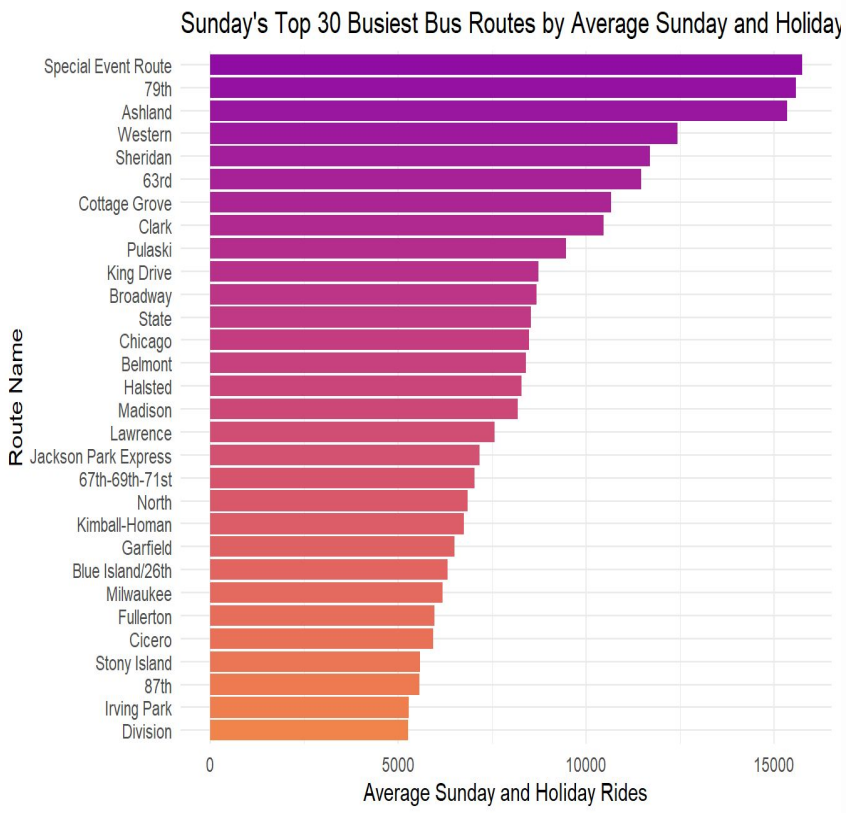


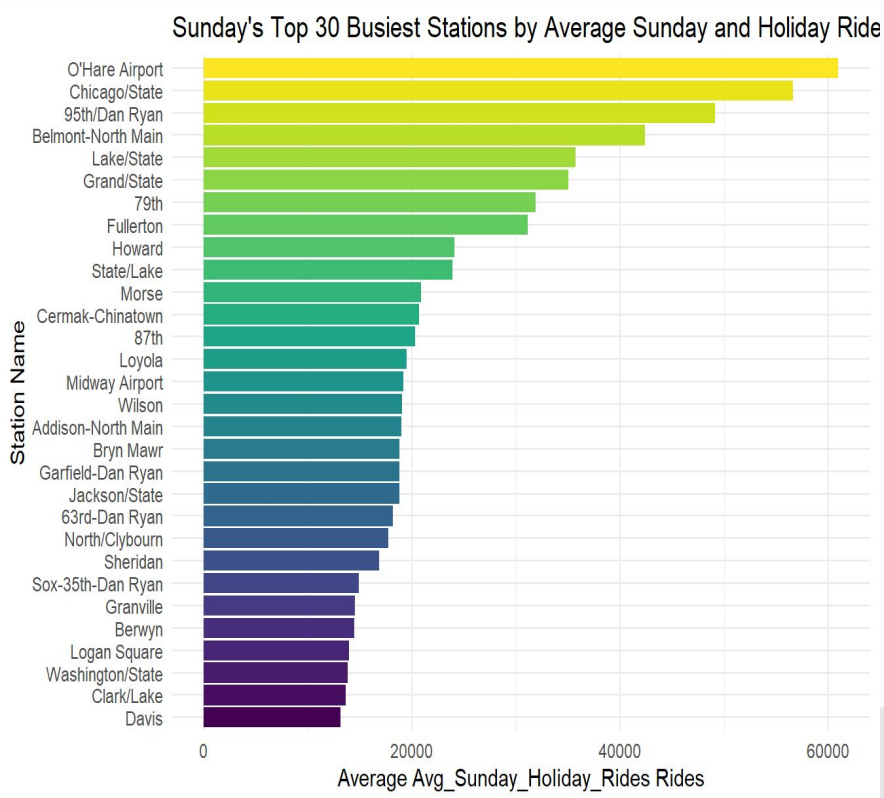
Top 30 Busiest Bus Routes and L station by Average Saturday Rides :



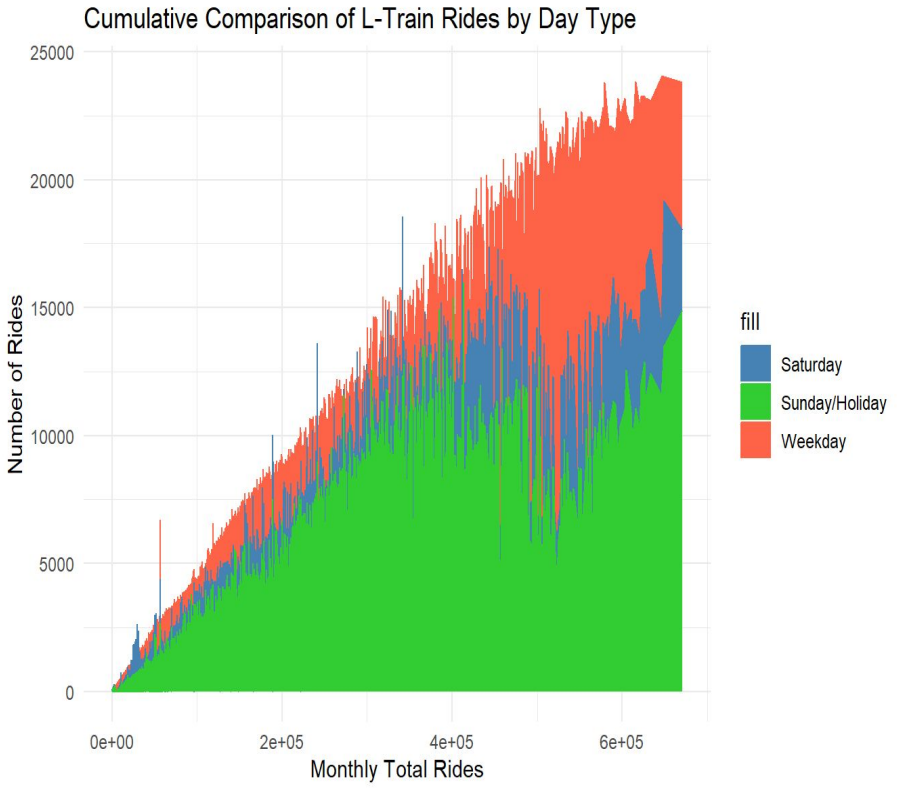


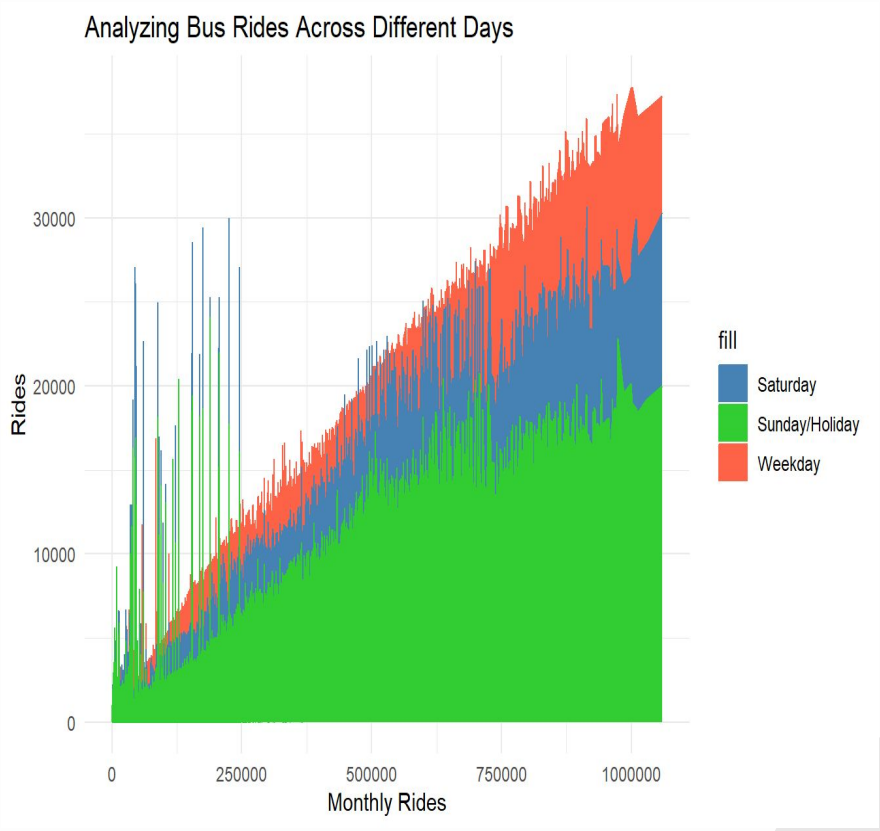
Top 30 Busiest Bus Routes and L station by Average Sunday/Holiday Rides :



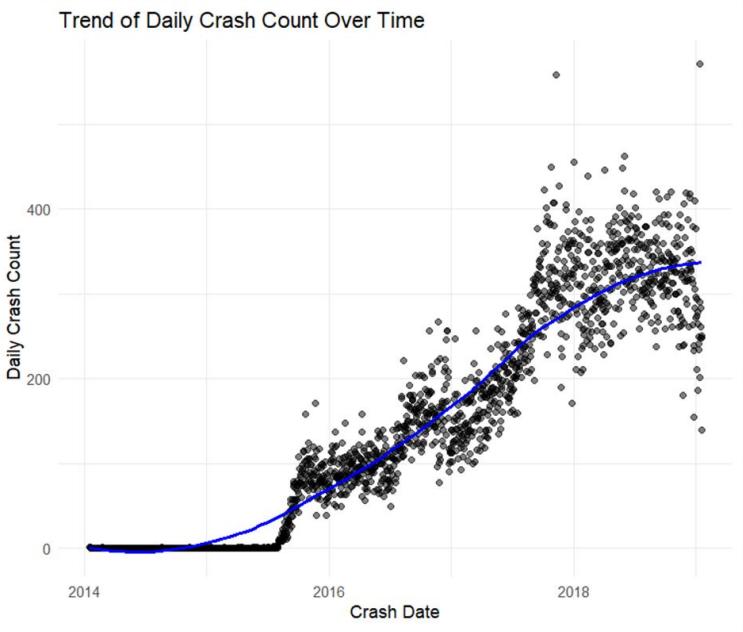


L-Train and Bus ridership according to day of the week:



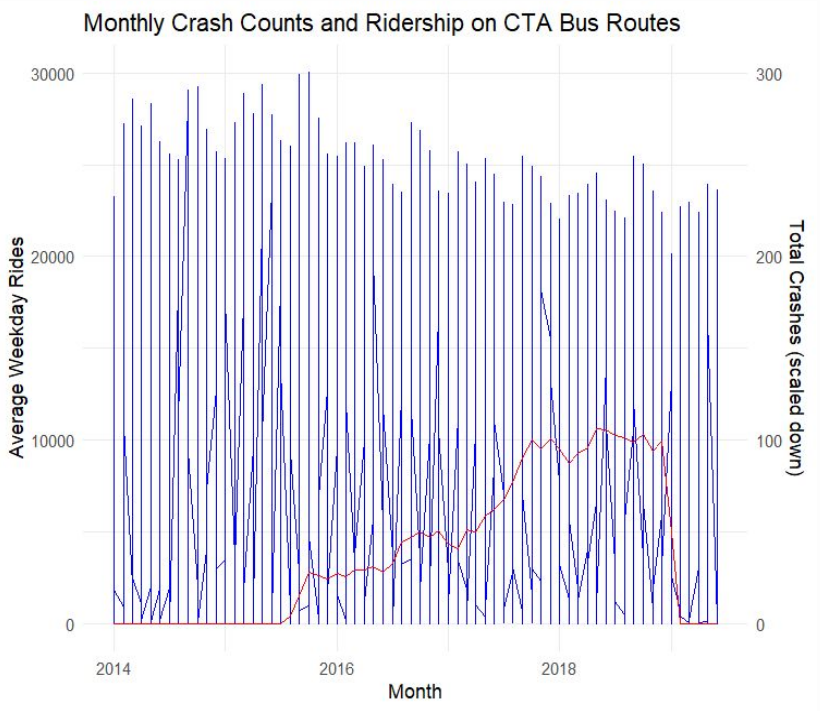


The trend of Daily Crash Count Over Time (Crashes Data)



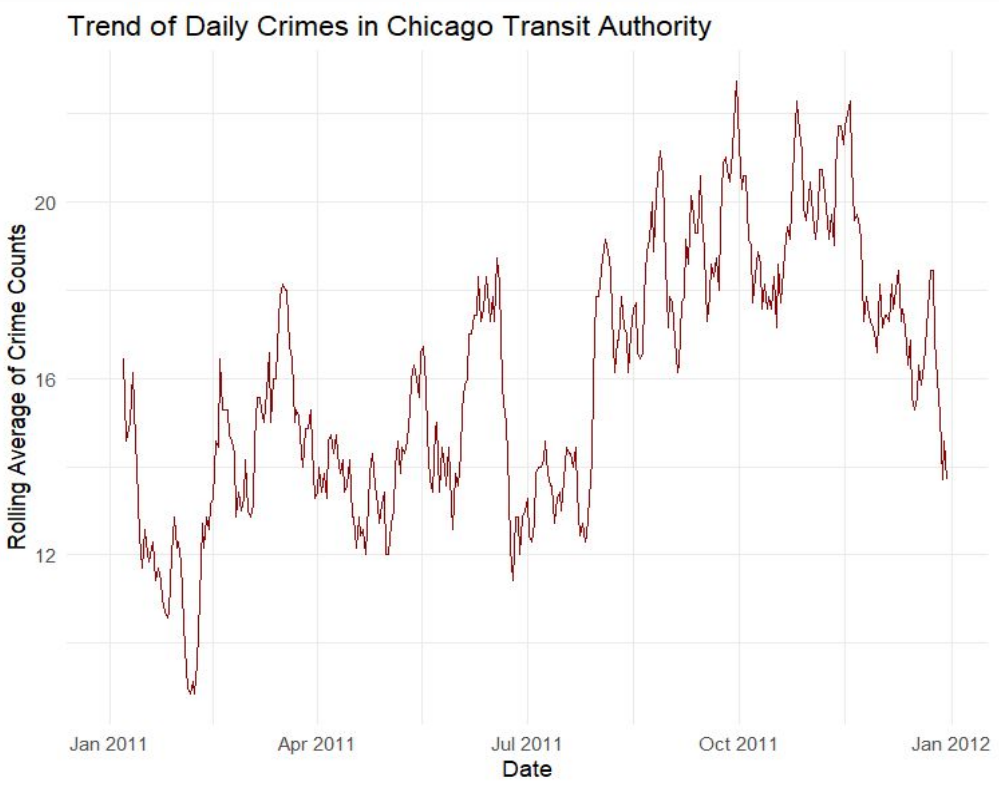
The scatter plot displays daily crash counts from 2014 to 2018, with an upward trend over time. The blue line indicates a significant increase in crashes, peaking in 2018

Monthly Crash Counts and Weekday Ridership Trends on CTA Bus Routes (2014-2018) (Crashes Data)



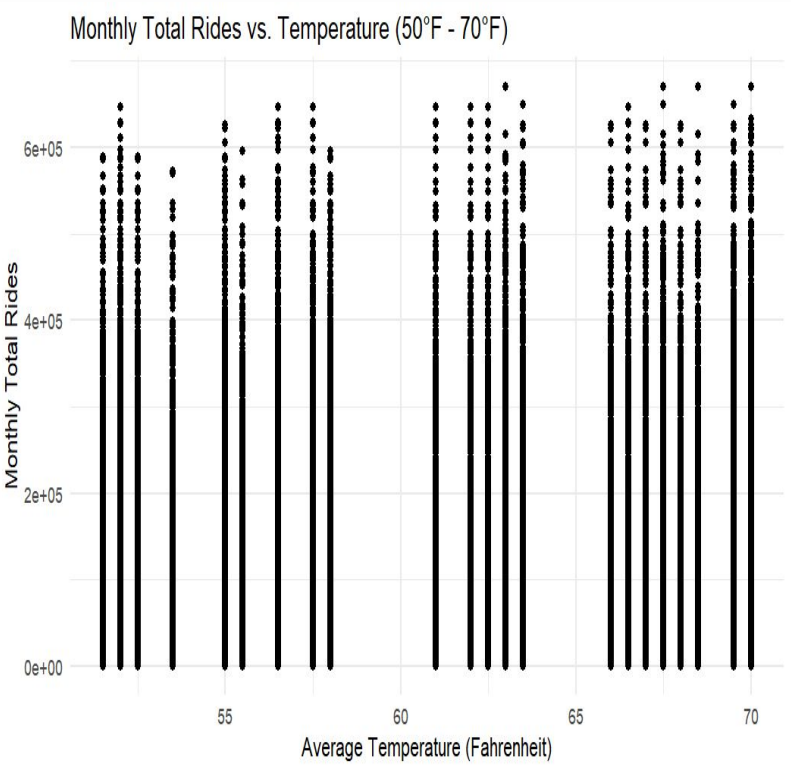
The graph shows that in weekday ridership on CTA bus routes (blue) from 2014 to 2020, indicating steady public transit demand. The red line represents a rising trend in monthly crash counts, highlighting an increase in traffic incidents.

Trends in Daily Crime Rates for Chicago Transit Authority (2011–2012) (Crime Data)



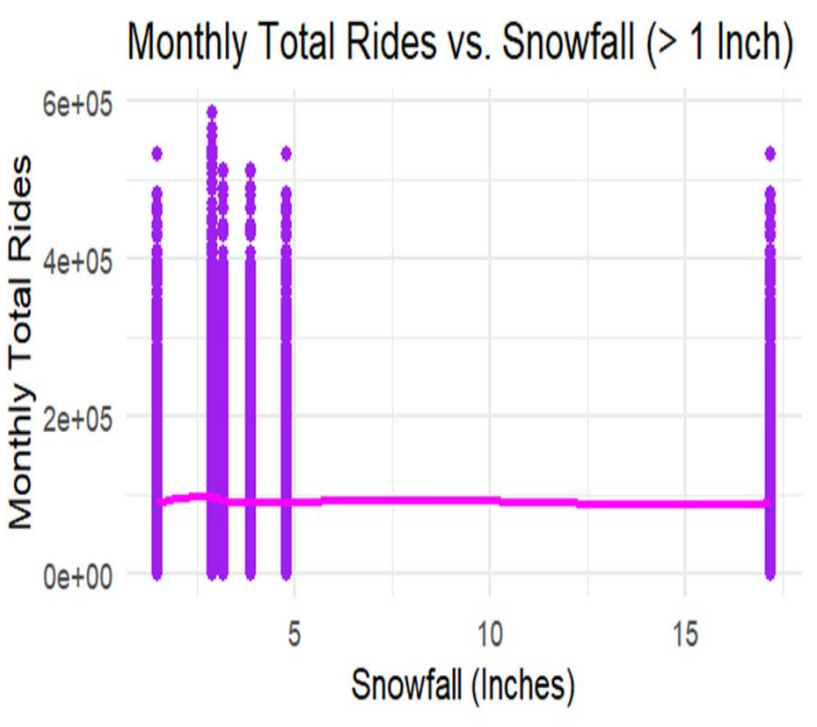
The line graph depicts the rolling average of daily crimes in the Chicago Transit Authority from January 2011 to January 2012. It shows higher crime rates during mid-2011 with peaks in July and November, while the end of the year experiences a decline.

Scatter Plot of Monthly Total Rides vs. Temperature (50°F - 70°F)(Weather Data)



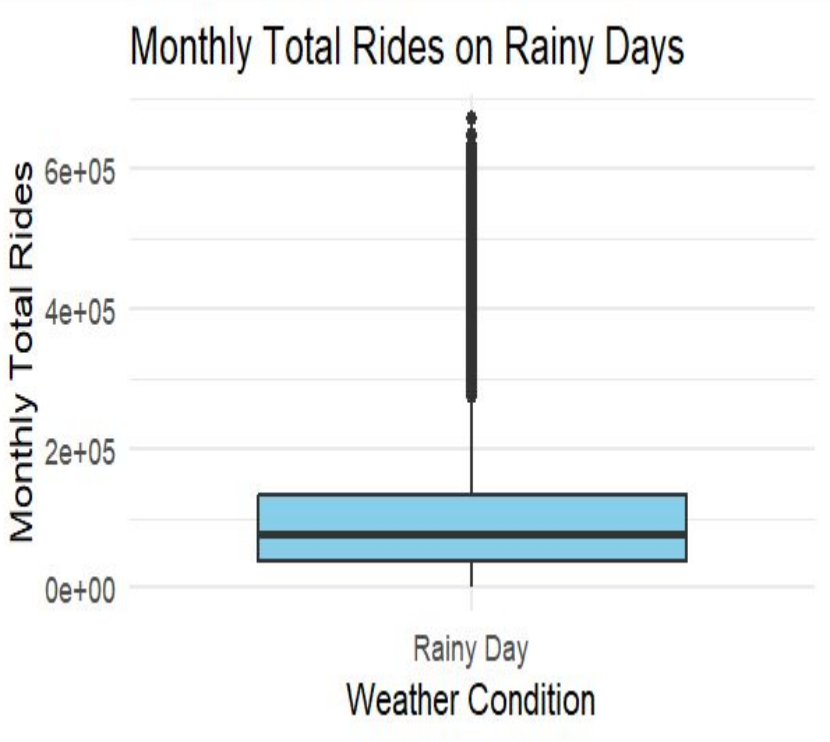
The scatter plot displays the relationship between average temperature (in Fahrenheit) and monthly total rides, with a range of 50°F to 70°F. The positive correlation indicates that as temperature increases, so does the number of monthly rides, suggesting temperature fluctuations impact ridership patterns.

Relationship Between Snowfall and Monthly Total Rides on Public Transit (Weather Data)



The graph depicts the correlation between monthly total rides and snowfall greater than 1 inch, showing a trend of decreasing ridership as snowfall increases.

Monthly Total Rides on Rainy Days (Weather Data)

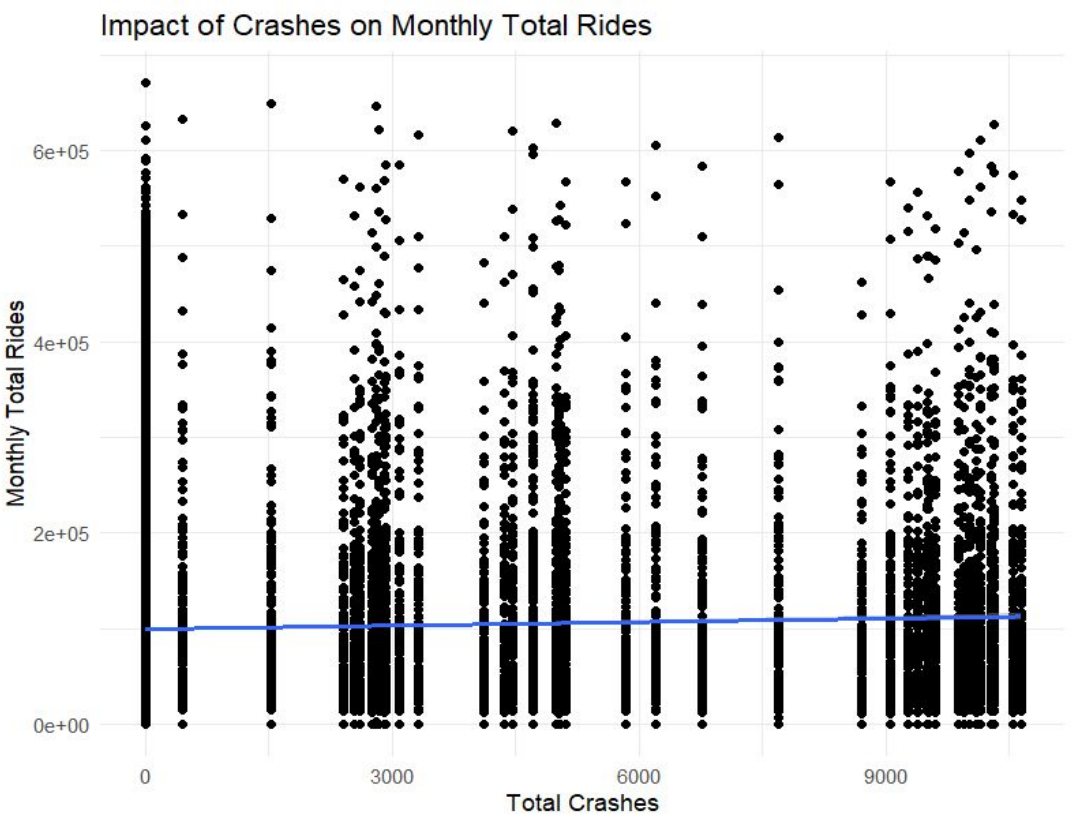


The box plot displays the distribution of monthly total rides specifically on rainy days, with the box's color indicating the ride totals' range. It suggests a variation in ridership during rainy weather, with the box's position and spread reflecting the overall pattern of rides on days with measurable rainfall.

06

DATA MODELLING

Linear Regression (Impact of Crashes on Monthly Rides)



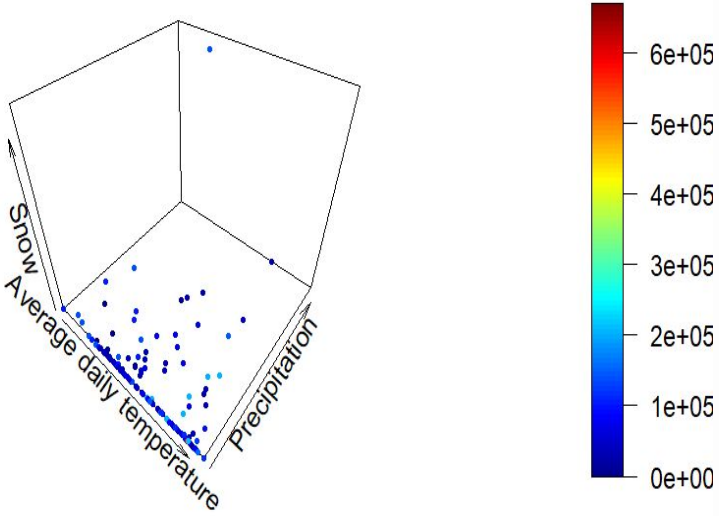
Statistical Significance: The p-value for total_crashes is extremely low (2.69e-08), indicating that it is statistically significant.

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

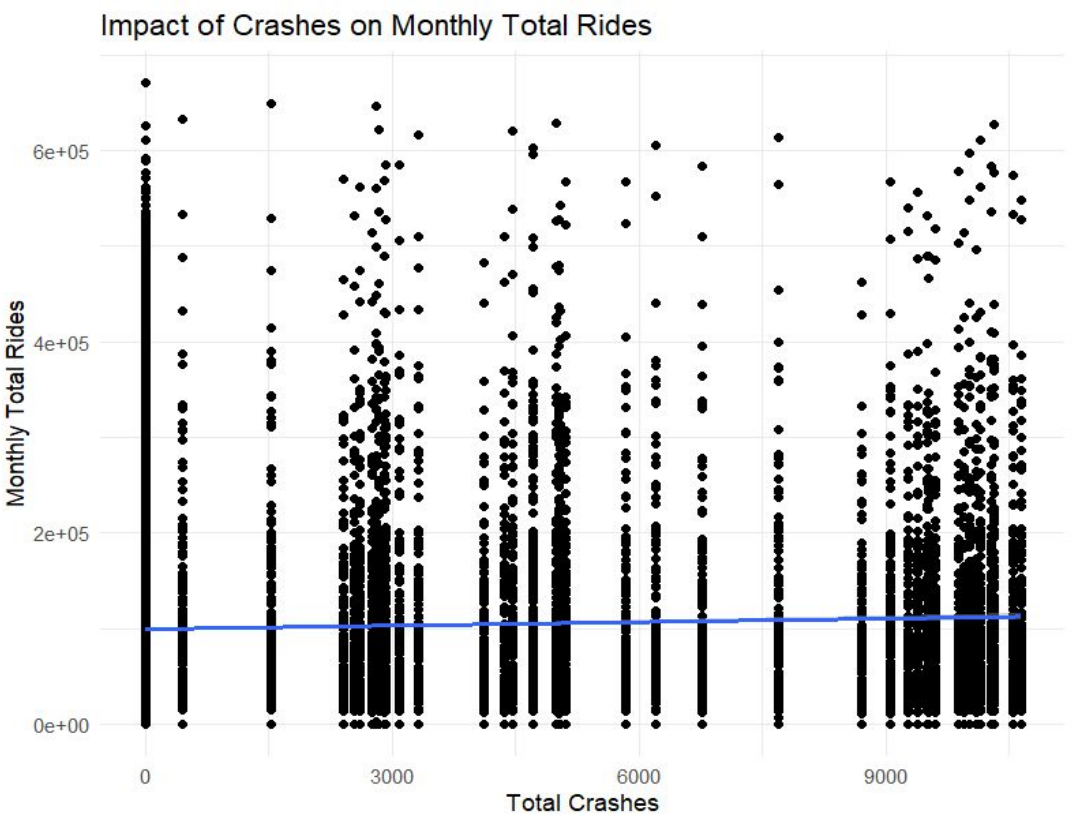
R-Squared: The Multiple R-squared is 0.9971, suggesting that approximately 99.71% of the variance in monthlytotalrides can be explained by the model.

Overall, these results suggest a significant correlation between total crashes and monthly total rides. As monthly total rides also increase, the number of crashes rises, likely due to transit disruptions or changing ridership patterns. Weekday and Saturday rides are also significant predictors of monthly total rides.

Linear Regression (Impact of Weather on Monthly Rides)



Linear Regression (Impact of Crashes on Monthly Rides)



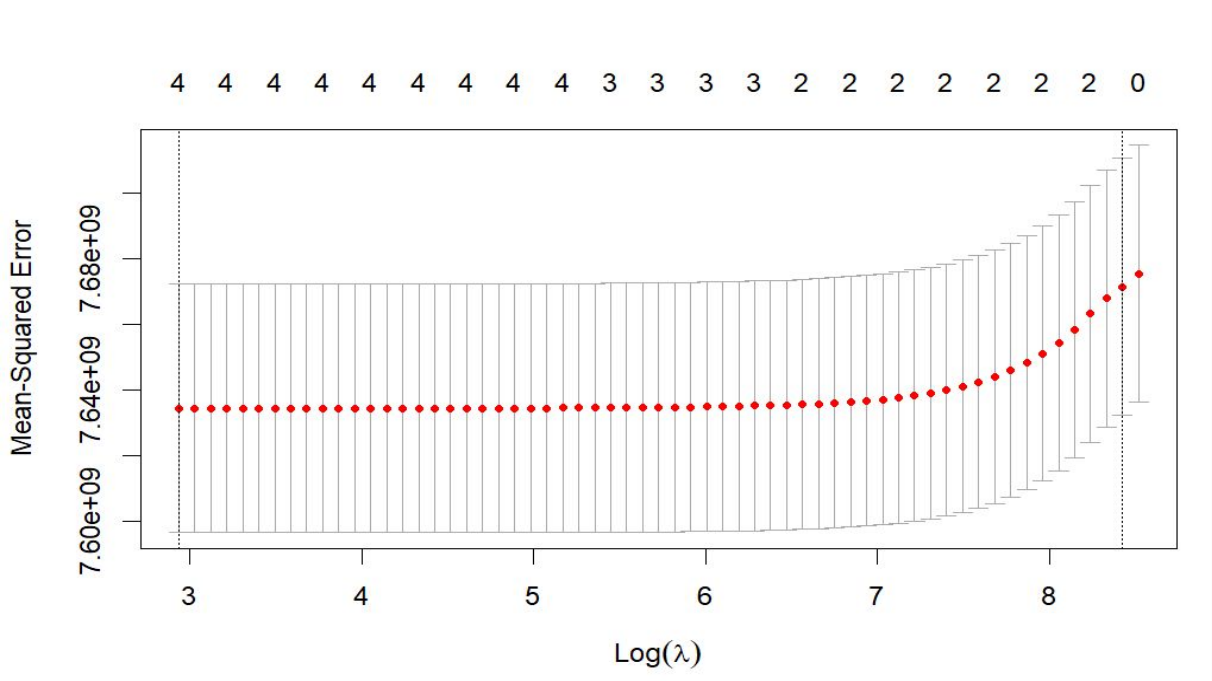
Statistical Significance: The p-value for total_crashes is extremely low (2.69e-08), indicating that it is statistically significant.

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

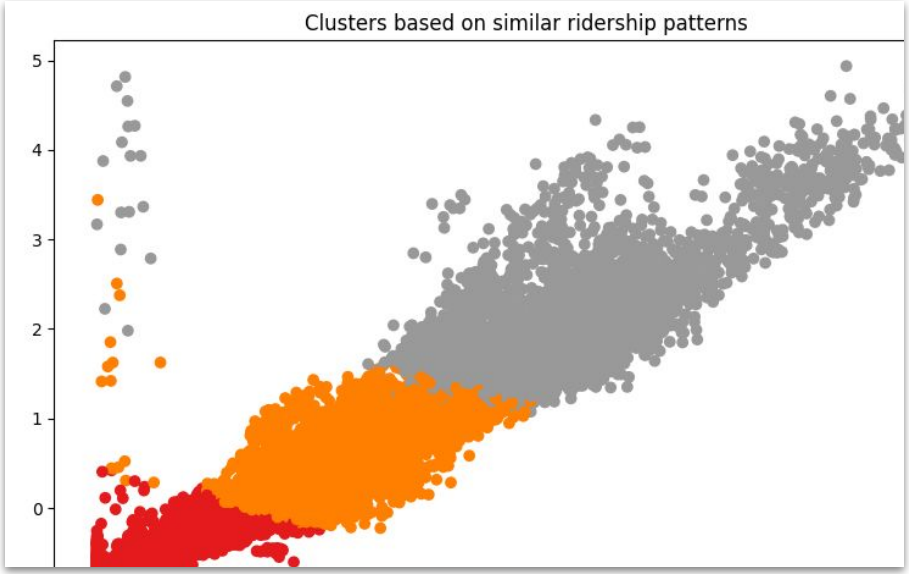
R-Squared: The Multiple R-squared is 0.9971, suggesting that approximately 99.71% of the variance in monthlytotalrides can be explained by the model.

Overall, these results suggest a significant correlation between total crashes and monthly total rides. As monthly total rides also increase, the number of crashes rises, likely due to transit disruptions or changing ridership patterns. Weekday and Saturday rides are also significant predictors of monthly total rides.

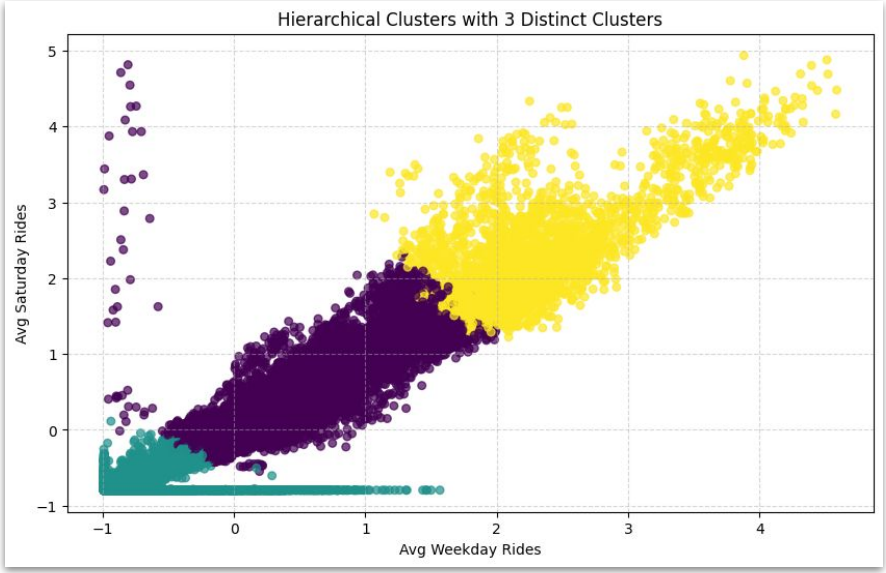
Lasso regression to predict monthly total rides in a transit system using predictor variables like temperature, snowfall, and precipitation.



CTA Bus Routes based on their ridership patterns using clustering methods



K-means clustering



Hierarchical Clustering

07

DISCUSSIONS

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.