ILLINOIS TECH

NAVIGATING CHICAGO:

Analysis of CTA Transit Route

Juveriya Fatima
A20528182
jfatima1@hawk.iit.edu

Vidheesha Patil
A20517203
vpatil21@hawk.iit.edu

TABLE OF CONTENTS

ABSTRACT

This project delves into Chicago Transit Authority (CTA) data, weather patterns, crime rates, and holiday schedules to uncover insights into transit ridership patterns. Utilizing multiple linear and Lasso regressions, it identifies influential factors impacting transit usage.The study reveals that people rely heavily on public transport during the weekdays, and certain stations are busier than others. It also shows how weather and crime affect when and where people travel. This information helps make better decisions about public transport routes and safety measures, making commuting safer and more convenient for everyone.

1. PROJECT OVERVIEW

1.1 Background

This project immerses into the daily journeys of Chicagoans reliant on the intricate web of the Chicago Transit Authority (CTA). As commuters traverse the city's diverse neighborhoods, they contend not only with the efficiency and reliability of CTA routes but also with external factors such as weather patterns, crime rates, and day-to-day variations. From scorching summers to frigid winters, from bustling weekday rush hours to leisurely weekend strolls, every ride is an intricate dance with Chicago's dynamic urban fabric. This project seeks to unravel these complexities, shedding light on how different factors shape the commuter experience on CTA transit routes.

1.2 Aim

The goal of this project is to find the optimum transit routes and timings for Chicago Transit Authority (CTA) passengers by analyzing data from different sources such as traffic, weather, crime and other datasets. To determine the most efficient and secure transit routes for CTA users, our research involves analyzing multiple datasets through techniques such as regression analysis, Lasso and data visualization. Furthermore, the project offers suggestions for enhancement of services derived from the results of data analysis. Commuters of CTA and the chicagoland community will benefit in terms of ease of access to safe and efficient public transportation service.

1.3 Problem Statement

- What factors (Location, weather condition, crime, day) that affect the ridership in public transport in Chicago city?
- Which transit routes have most passengers getting on board?
- What time-slots are the traffic maximum and minimum in?
- Comparing the number of trips on different routes between weekdays and weekends.
- The influence of holidays and weekends on the utilization of the Chicago Transit Authority.
- The impact of crime statistics in a specific area and the route number on the frequency of trips conducted.
- Recommendations to CTA on routes with high passenger demand and insufficient transit capacity, considering crime statistics and proposing solutions for effective alleviation.
- Diagrams based on location & density for the busiest stations.

2. DATA PROCESSING

We have gathered data from several different publicly accessible sources which included information detailing CTA trip records in Chicago, details on bus routes, L station entries, daily boarding and overall total. Additionally, we incorporated weather data, national public holiday dates and records of reported crimes in Chicago.

2.1 Data Summary & Cleaning

Bus Data:

https://data.cityofchicago.org/Transportation/CTA-Ridership-Bus-Routes-Monthly-Day-Type-Averages/bynn-gwxy/about_data

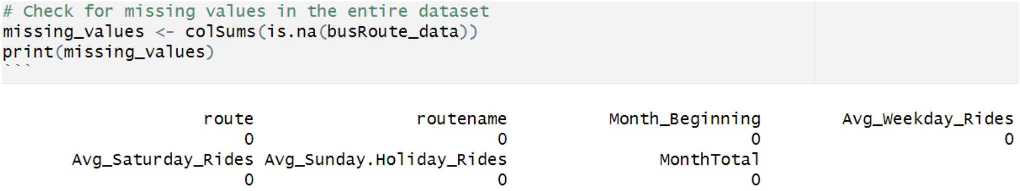
The dataset comprises 7 columns:

- route: Unique identifier for bus route
- routename: Name or description of each bus route.
- Month_Beginning: Starting month of the recorded data.
- Avg_Weekday_Rides: Mean ridership on weekdays.
- Avg_Saturday_Rides: Mean ridership on Saturdays.
- Avg_Sunday.Holiday_Rides: Mean ridership on Sundays and holidays.
- MonthTotal: Cumulative count of rides for the month.

It contains a total of 37,238 rows. To focus our analysis, we filtered the dataset to include only data from 2013 to 2023, resulting in 16,606 rows.

Initially, all columns were stored as character values. We converted them to their appropriate data types: numeric for numerical values and date for the Month_Beginning column.

Following these data preparation steps, we conducted an initial cleaning process. No missing values or duplicates were identified in the dataset.



```
# Check for missing values in the entire dataset
missing_values <- colSums(is.na(busRoute_data))
print(missing_values)
```

```
                   route                routename          Month_Beginning        Avg_Weekday_Rides 
                       0                        0                        0                        0 
      Avg_Saturday_Rides Avg_Sunday.Holiday_Rides               MonthTotal 
                       0                        0                        0 
```

L-Train Data:

https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Monthly-Day-Type-A/t2rn-p8d7/about_data

The dataset comprises 7 columns: station_id, stationname, Month_Beginning, Avg_Weekday_Rides, Avg_Saturday_Rides, Avg_Sunday.Holiday_Rides, and MonthTotal. It contains a total of 39,053 rows. To focus our analysis, we filtered the dataset to include only data from 2013 to 2023, resulting in 18,639 rows.

Initially, all columns were stored as character values. We converted them to their appropriate data types: numeric for numerical values and date for the Month_Beginning column.

After completing the data preparation steps, we initiated an initial cleaning process. While no duplicate rows were detected, we identified four missing values. Upon further examination, we determined that these values were irrelevant to our analysis and thus were removed from the dataset.

Weather Data:

https://www.ncei.noaa.gov/access/past-weather/chicago

The weather dataset is a combination of records from 4 different datasets of the regions of

1. IL US (US1ILCK0014),

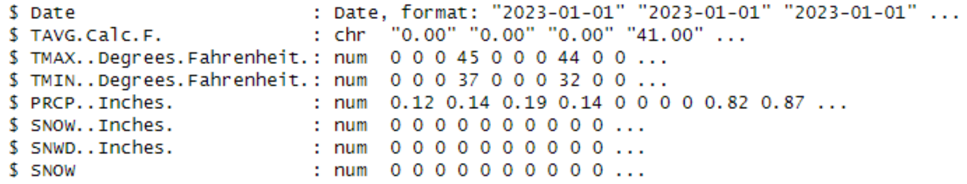
2. CHICAGO 6.8 NW IL US (US1ILCK0323)
3. CHICAGO MIDWAY AIRPORT, IL US (USW00014819)
4. Bridgeview,CHICAGO 5.5 ESE

The integrated dataset has 24093 records and 5 columns ranging from dates 1997 to 2024.

The columns are :

- Date:Date in mm/dd/yyyy format .
- TAVG (Fahrenheit): The Average of Minimum and maximum temperatures. Manually calculated in rows where value is null or empty by calculating the mean temperature.
- TMAX(Fahrenheit):Maximum temperature
- TMIN(Fahrenheit):Minimum temperature
- PRCP(Inches): precipitation recorded in inches
- SNOW(Inches):snowfall recorded in inches
- SNOW DEPTH(Inches): depth of snowfall recorded in inches
- Location.- location of the weather data collected .

During data cleaning, missing values in Precipitation, Snow, and Snow Depth columns were replaced with zeros and no duplicates were found. Average temperature values were manually calculated from minimum and maximum values to fill missing cells. The Date column was converted to "%Y-%m-%d" format, and data from January 1, 2023 to January 31, 2023 was selected.



```
$ Date                       : Date, format: "2023-01-01" "2023-01-01" "2023-01-01" ...
$ TAVG.Calc.F.               : chr  "0.00" "0.00" "0.00" "41.00" ...
$ TMAX..Degrees.Fahrenheit.: num  0 0 0 45 0 0 0 44 0 0 ...
$ TMIN..Degrees.Fahrenheit.: num  0 0 0 37 0 0 0 32 0 0 ...
$ PRCP..Inches.              : num  0.12 0.14 0.19 0.14 0 0 0 0 0.82 0.87 ...
$ SNOW..Inches.              : num  0 0 0 0 0 0 0 0 0 0 ...
$ SNWD..Inches.              : num  0 0 0 0 0 0 0 0 0 0 ...
$ SNOW                       : num  0 0 0 0 0 0 0 0 0 0 ...
```

Crime Data :

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present

- ID: Unique identifier for the record.
- Case Number: The Chicago Police Department RD Number (Records Division Number) ,is unique to the incident.
- Date: Date when the incident occurred.
- Time: Time when the incident occurred in 24 hrs format.
- AM/PM: Denotes the time
- Block: The address where the incident occurred has been partially obscured, but it still falls within the same block as the actual address.
- IUCR: The Illinois Uniform Crime Reporting code.
- Primary Type: The primary description of the IUCR code.
- Description: The secondary description of the IUCR code.
- Location Description: Explanation of the location where the incident occurred.

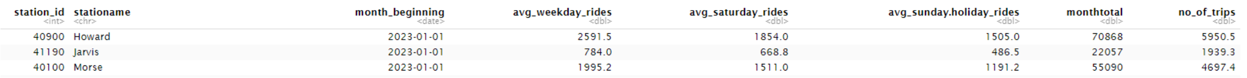
- ○ Arrest: Specifies whether an arrest was made.
- ○ Domestic: If the incident was domestic-related as outlined by the Illinois Domestic Violence Act.
- ○ Beat: Specifies the specific area within the police jurisdiction where the incident took place.Several beats form a police sector, and multiple sectors constitute a police district
- ○ District: Specifies the police district where the incident occurred.
- ○ Ward: The ward where the incident occurred..
- ○ Community Area: Specifies the community area of incident occurrence.
- ○ FBI Code: Mentions the crime classification as defined in the FBI's NIBRS
- ○ Year: Year the incident occurred.
- ○ Updated On: Last update on date and time
- ○ Location: The incident's location .

CHANGES MADE TO THE DATASET

- The dataset has 8.04 million records and 22 columns where each row is a reported crime.
- We have filtered the records from dates of "date=18 April 2019 to 18 April 2024" and on "locations: CTA L train, bus, platform, parking, bus stop".
- This dataset now contains 16445 rows. All the data in columns were in char data types. During cleaning, convert to suitable data types.
- The Date column was initially combined with Time. We split it into 3 columns of Date, Time and AM/PM in data cleaning.
- During data cleaning, we noticed some missing records for Ward, Latitude, Longitude, ZipCodesetc. Since those are non essential columns for our current project,we have removed them.
- There are no duplicates in the crime dataset.

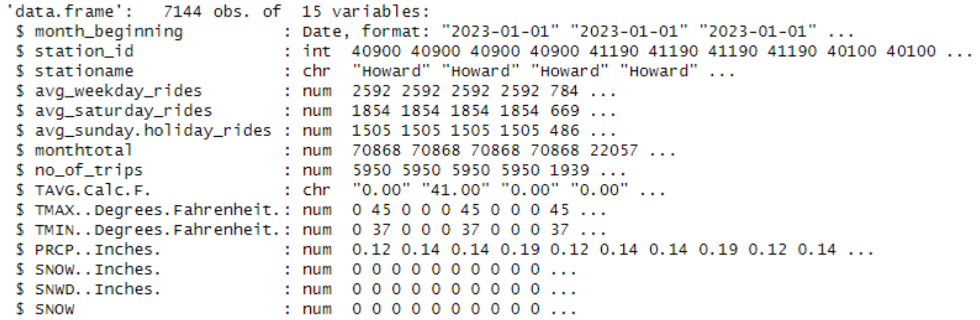
Merged Data:

Combined the individual datasets of CTA Ridership by station and Weather on the Date column. The Monthly ridership dataset has column "month_beginning' which is merged with the "Date" column from the Weather data frame. This results in a new dataset with 14 columns. A new column called 'no_of_trips' was created by adding values of 3 columns with data about average rides on weekday, saturday and sunday/holiday .The merged data columns were then further processed.



station_id	stationame	month_beginning	avg_weekday_rides	avg_saturday_rides	avg_sunday.holiday_rides	monthtotal	no_of_trips
<int>	<chr>	<date>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
40900	Howard	2023-01-01	2591.5	1854.0	1505.0	70868	5950.5
41190	Jarvis	2023-01-01	784.0	668.8	486.5	22057	1939.3
40100	Morse	2023-01-01	1995.2	1511.0	1191.2	55090	4697.4

The merged dataset has 15 columns. The columns of no_of_trips, T_AVG, Snow, Precipitation are further processed and then used for multiple Regression modeling.

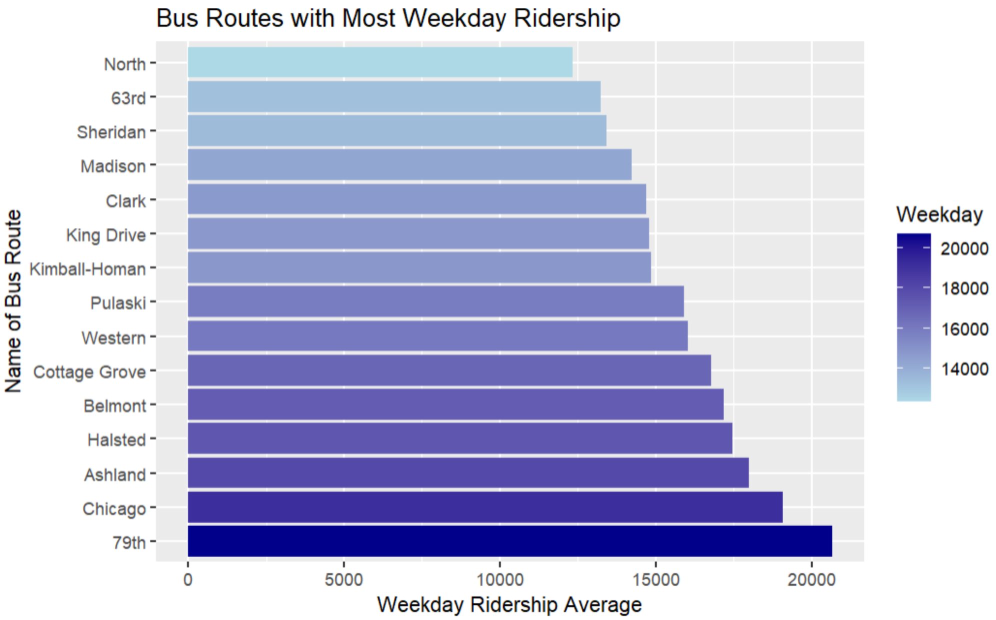


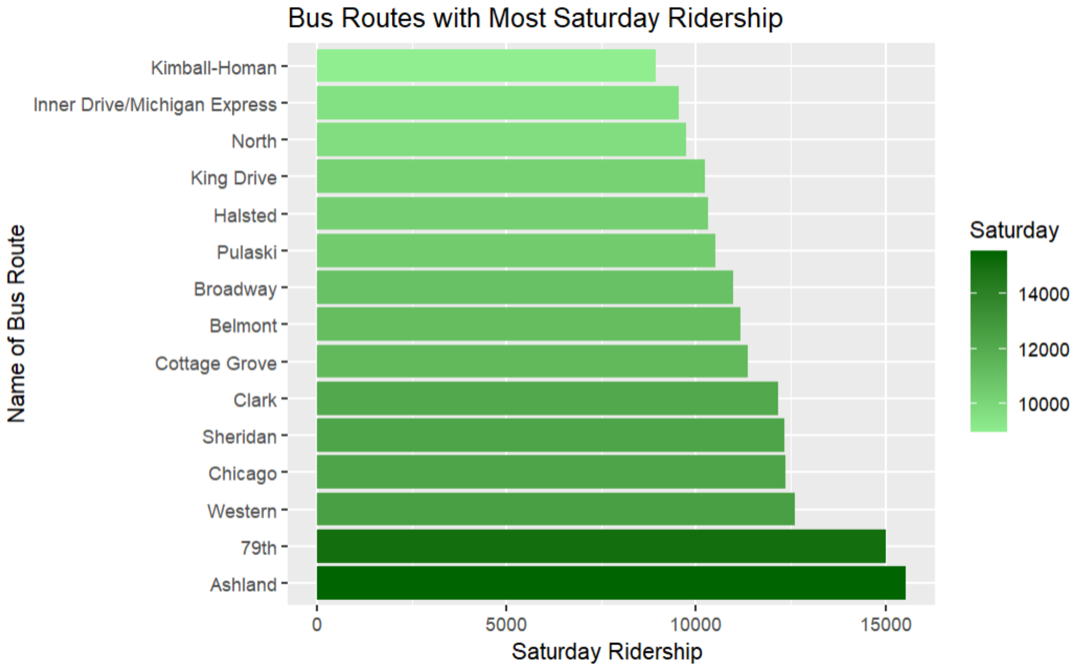
```
'data.frame':   7144 obs. of  15 variables:
 $ month_beginning           : Date, format: "2023-01-01" "2023-01-01" "2023-01-01" ...
 $ station_id                : int  40900 40900 40900 40900 41190 41190 41190 41190 40100 40100 ...
 $ stationame                : chr  "Howard" "Howard" "Howard" "Howard" ...
 $ avg_weekday_rides         : num  2592 2592 2592 2592 784 ...
 $ avg_saturday_rides        : num  1854 1854 1854 1854 669 ...
 $ avg_sunday.holiday_rides : num  1505 1505 1505 1505 486 ...
 $ monthtotal                : num  70868 70868 70868 70868 22057 ...
 $ no_of_trips               : num  5950 5950 5950 5950 1939 ...
 $ TAVG.Calc.F.              : chr  "0.00" "41.00" "0.00" "0.00" ...
 $ TMAX..Degrees.Fahrenheit.: num  0 45 0 0 0 45 0 0 0 45 ...
 $ TMIN..Degrees.Fahrenheit.: num  0 37 0 0 0 37 0 0 0 37 ...
 $ PRCP..Inches.             : num  0.12 0.14 0.14 0.19 0.12 0.14 0.14 0.19 0.12 0.14 ...
 $ SNOW..Inches.             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SNWD..Inches.             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SNOW                      : num  0 0 0 0 0 0 0 0 0 0 ...
```

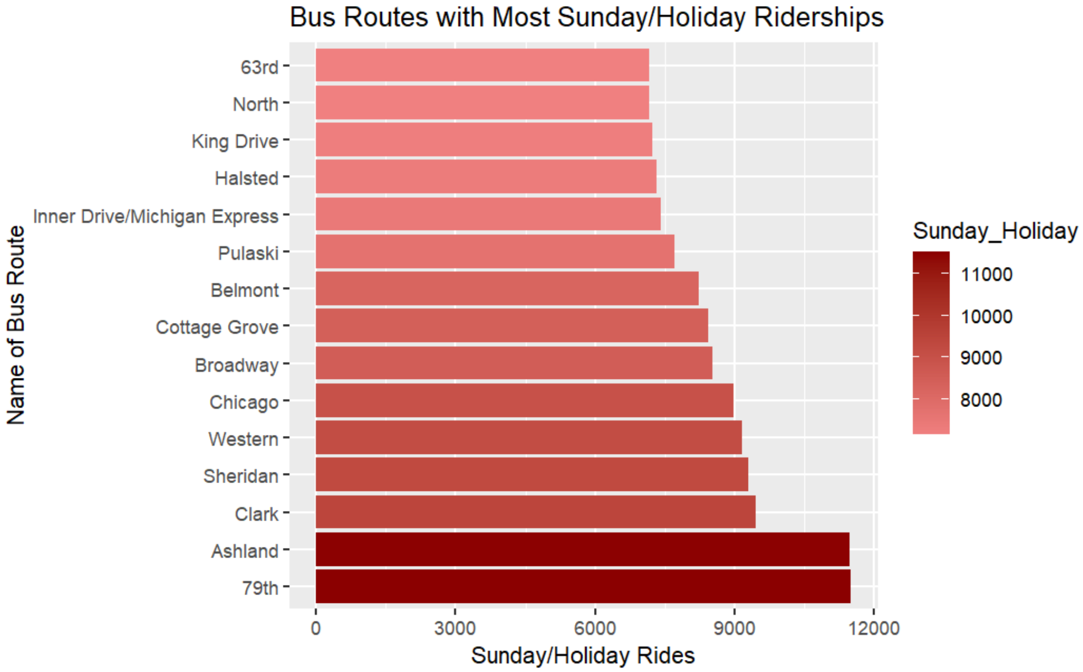
All the datasets along with the codes are uploaded on the following gitHub: https://github.com/Juveriyaae/CSP571-Project

3. EXPLORATORY DATA ANALYSIS

A. Bus Routes with higher ridership based on different days of the week.



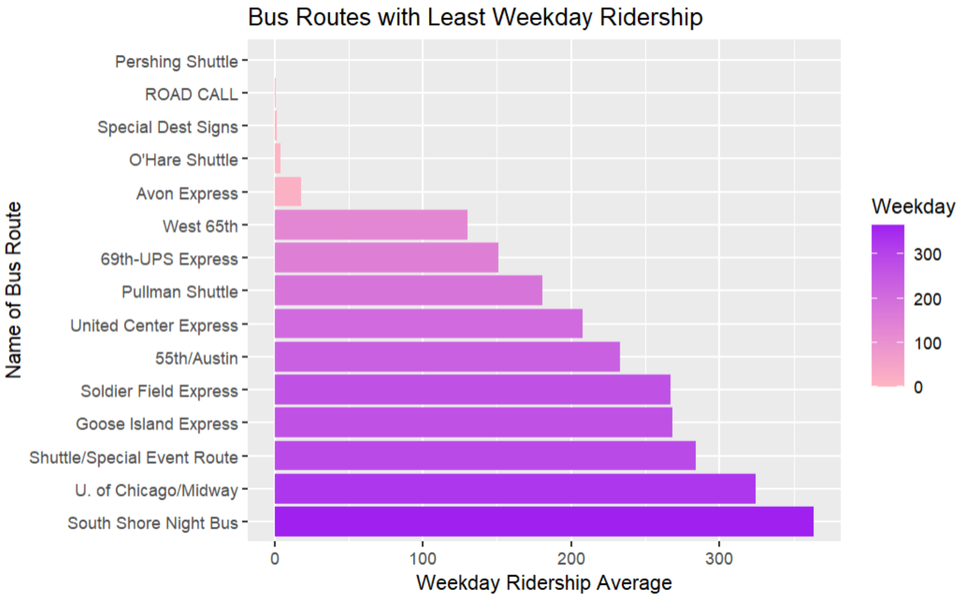




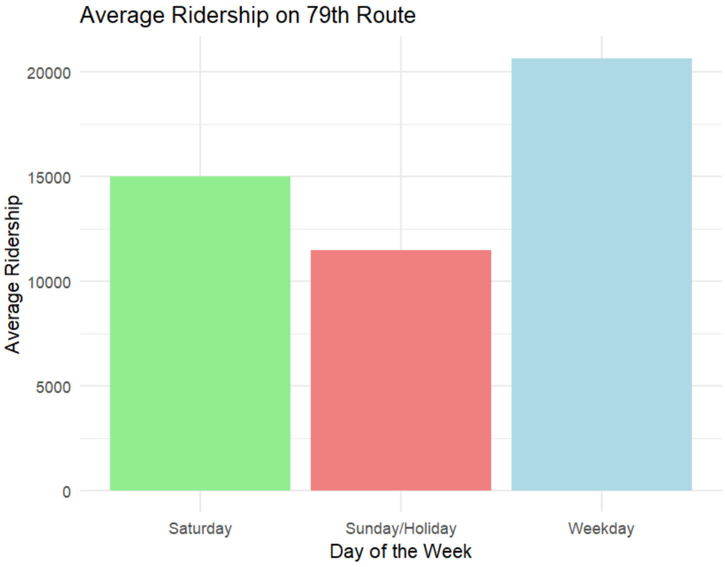
Referring to the graphs depicting the busiest bus routes on average for different days of the week, several patterns emerge. On weekdays, the 79th route stands out as the busiest, recording over 20,000 rides. Meanwhile, Ashland emerges as the busiest route on Saturdays, with an average of more than 16,000 rides, closely followed by the 79th route with similar ridership levels.

During Sundays or holidays, the trend continues with the 79th route maintaining its position as the busiest, averaging almost 10,000 rides. Ashland follows closely behind with a comparable number of rides.

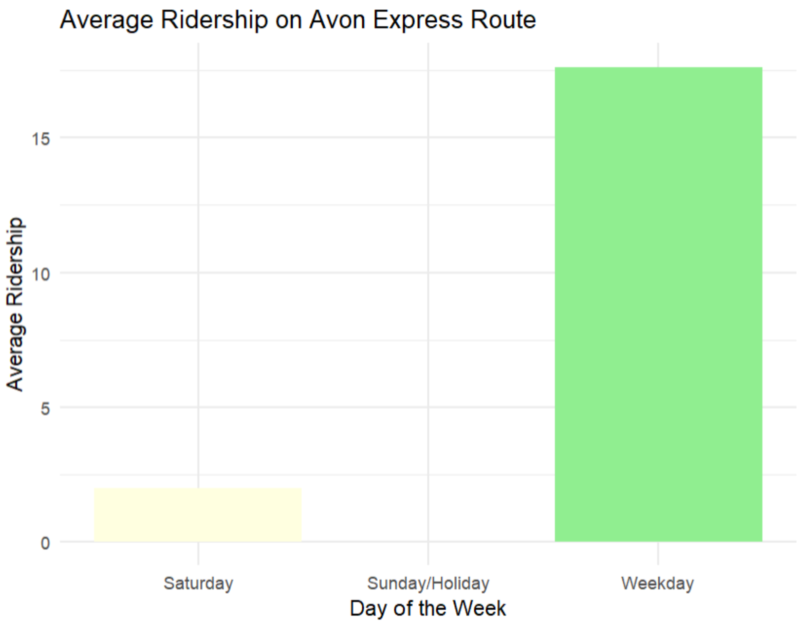
In summary, our analysis reveals that the 79th route consistently experiences the highest ridership across weekdays, Saturdays, Sundays, and holidays

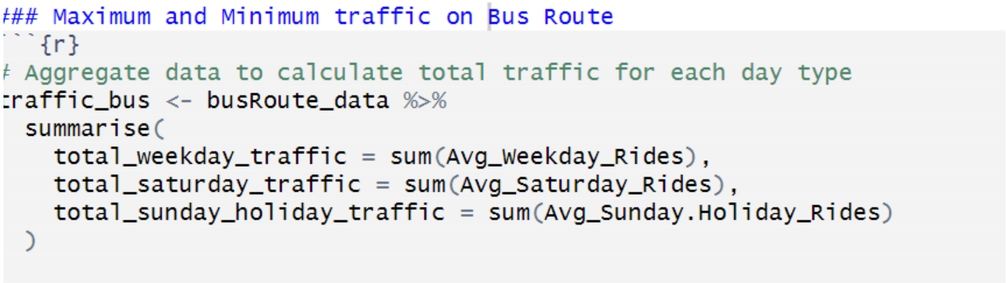


The above plot displays bus routes with least ridership on weekdays. Similar results were observed on Saturdays, Sundays and Holidays.



Further breaking down the 79th route, we can conclude that weekdays indeed witness the highest levels of activity. The plot below shows ridership for Avon Express which is one of the routes with least ridership.





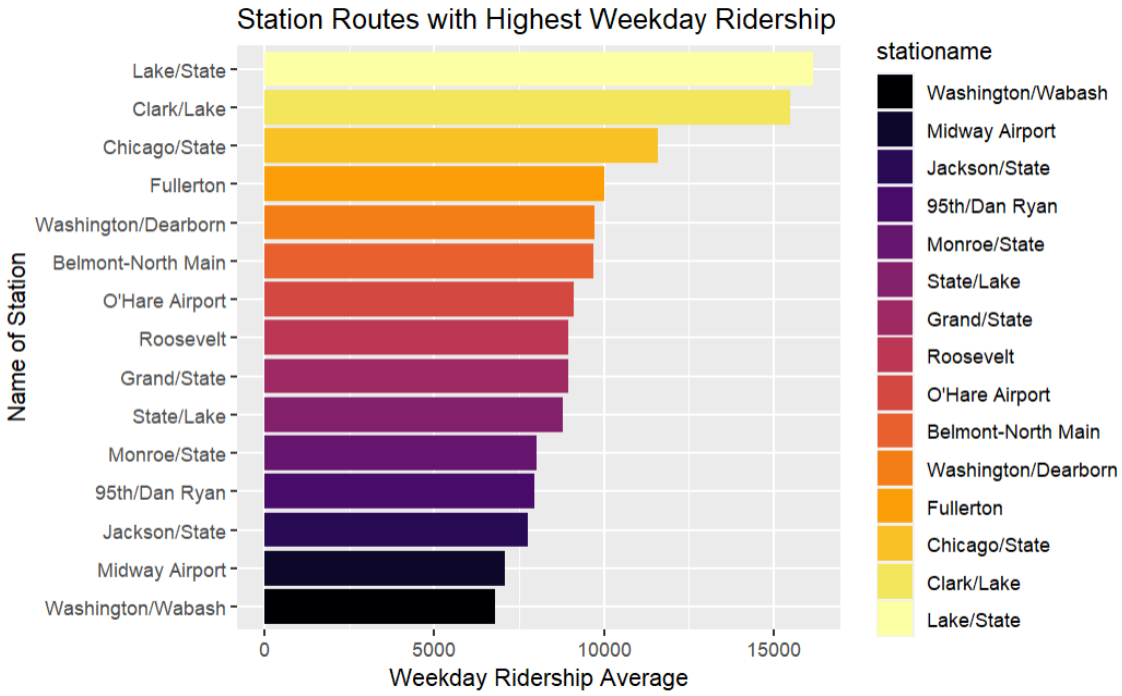
~~~
### Maximum and Minimum traffic on Bus Route
```{r}
# Aggregate data to calculate total traffic for each day type
traffic_bus <- busRoute_data %>%
  summarise(
    total_weekday_traffic = sum(Avg_Weekday_Rides),
    total_saturday_traffic = sum(Avg_Saturday_Rides),
    total_sunday_holiday_traffic = sum(Avg_Sunday.Holiday_Rides)
  )
~~~

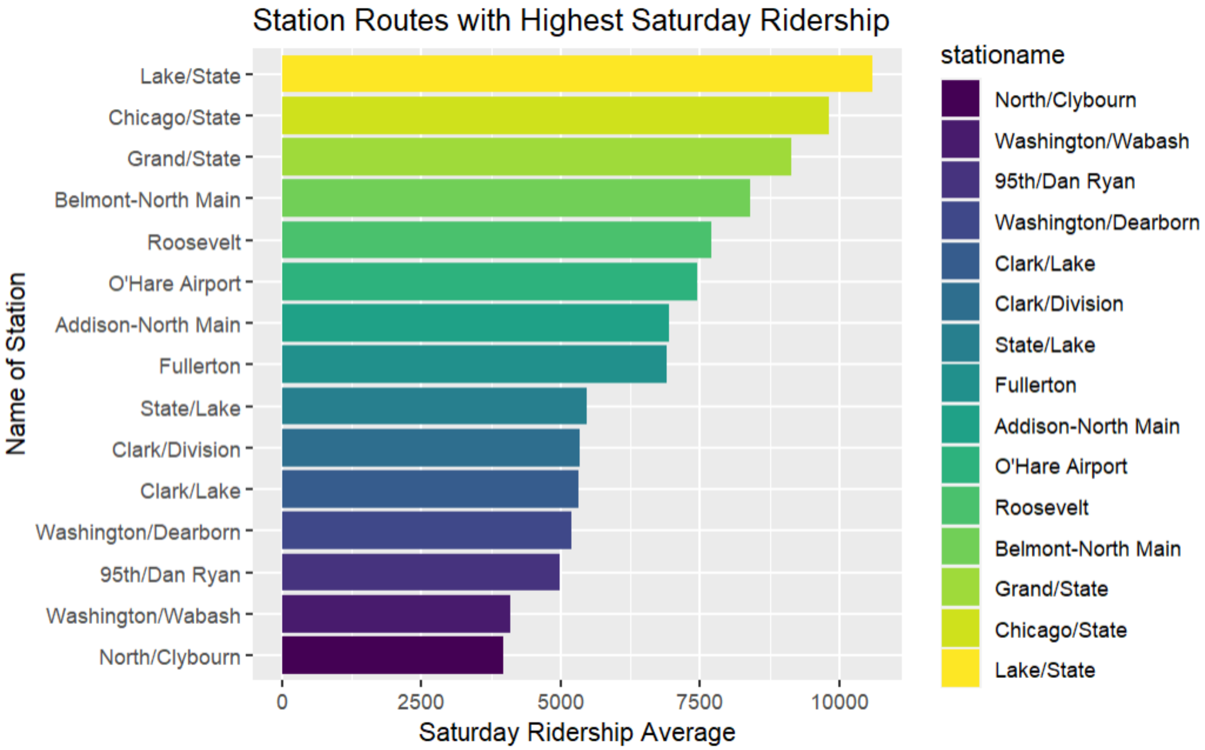
Based on the aggregation of data done above, we can observe the total traffic for each day type:

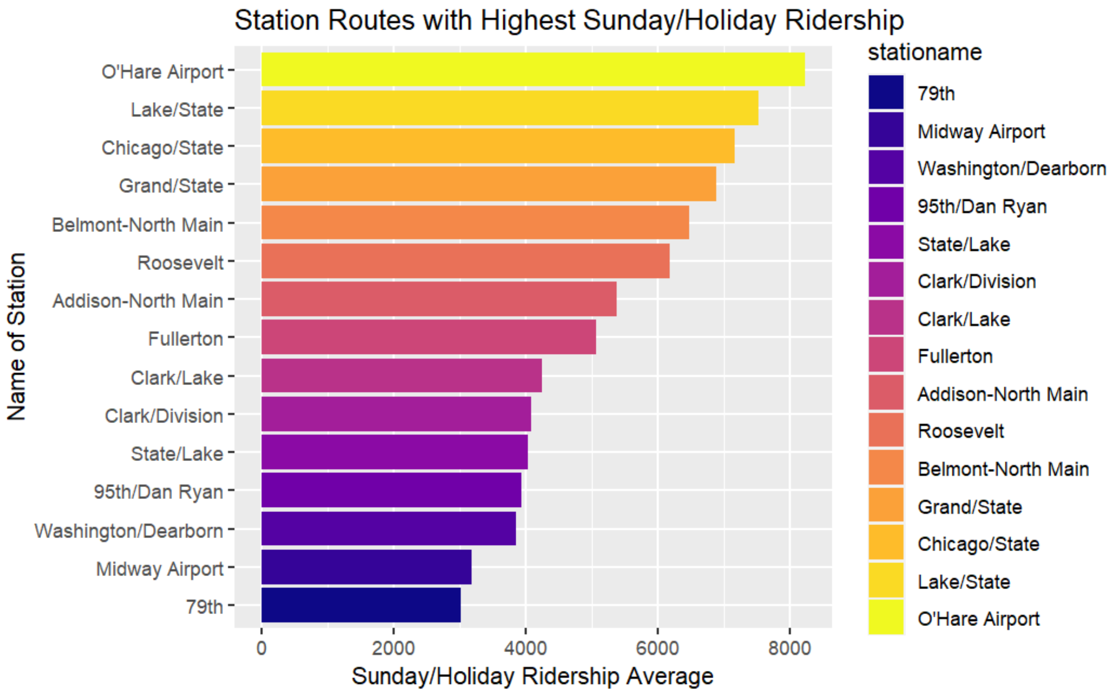
- Total weekday traffic: 89,321,754
- Total Saturday traffic: 55,693,662
- Total Sunday/holiday traffic: 40,857,403

These results signify the overall volume of traffic across different days of the week, indicating the busiest and least busy periods for bus routes. The significantly higher traffic on weekdays compared to weekends suggests that public transportation services experience higher demand on regular working days. This information can be valuable for transportation planning and resource allocation to ensure efficient service delivery.

B. Stations with higher ridership based on different days of the week.



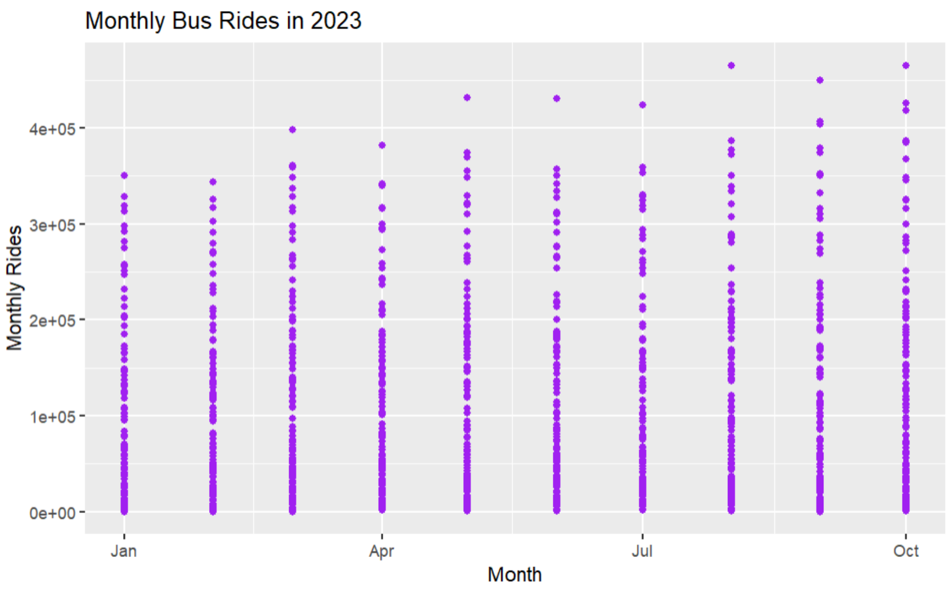


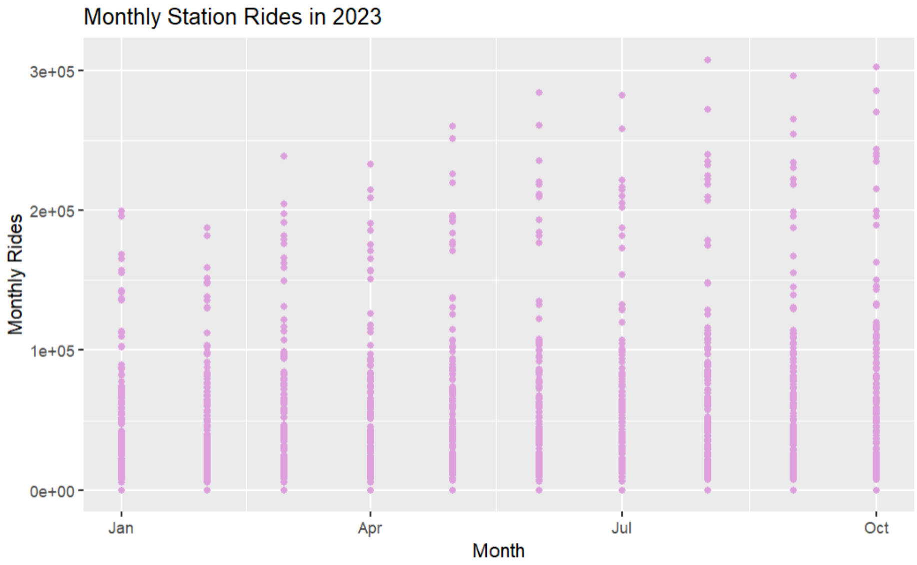


When analyzing the graphs above we found that, during the weekdays, Lake/State station is super busy, with more than 15,000 people riding on average, followed closely by Clark/Lake station. On Saturdays, Lake/State stays on top with over 10,000 rides. But on Sundays and holidays, O'Hare Airport station becomes the busiest, with about 8,000 riders on average, just behind Lake/State. There is significantly more ridership on weekdays with Lake/State experiencing more average rides throughout the week.

After aggregating the data to observe total traffic for each day type we observe the following:

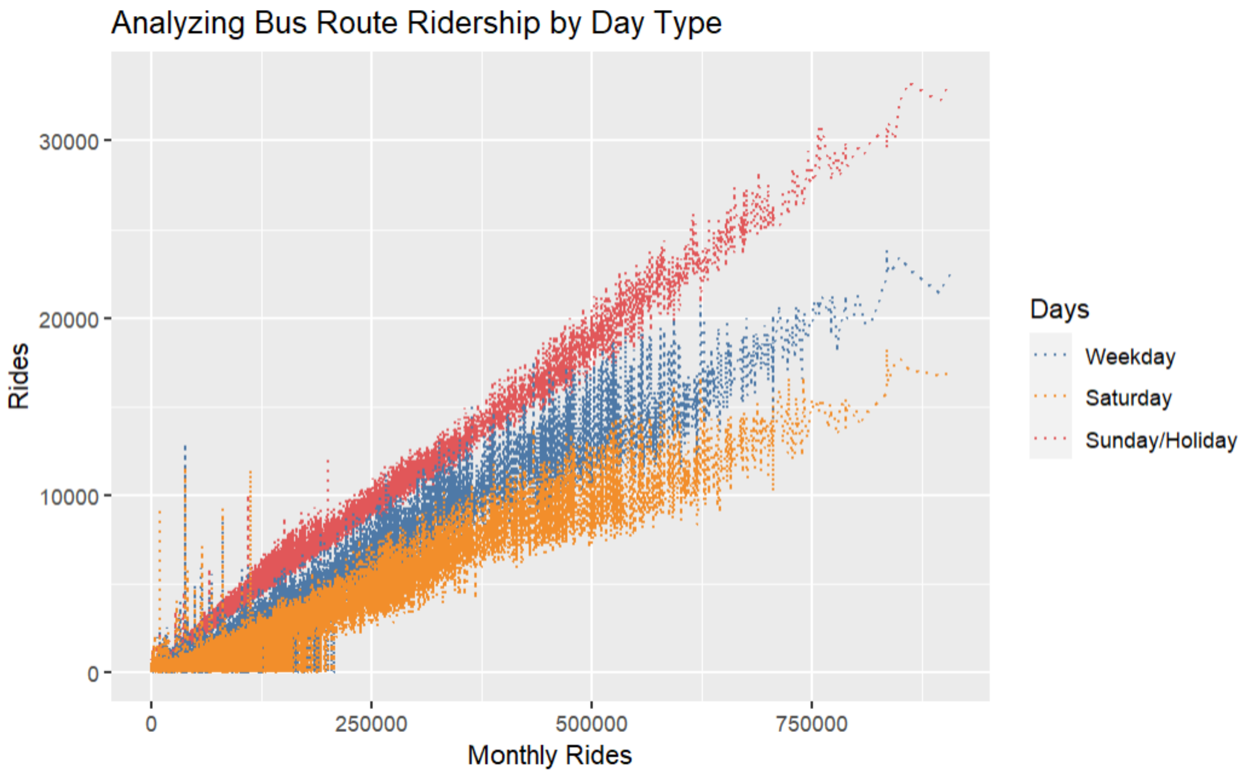
C. Monthly trends of transit routes in 2023

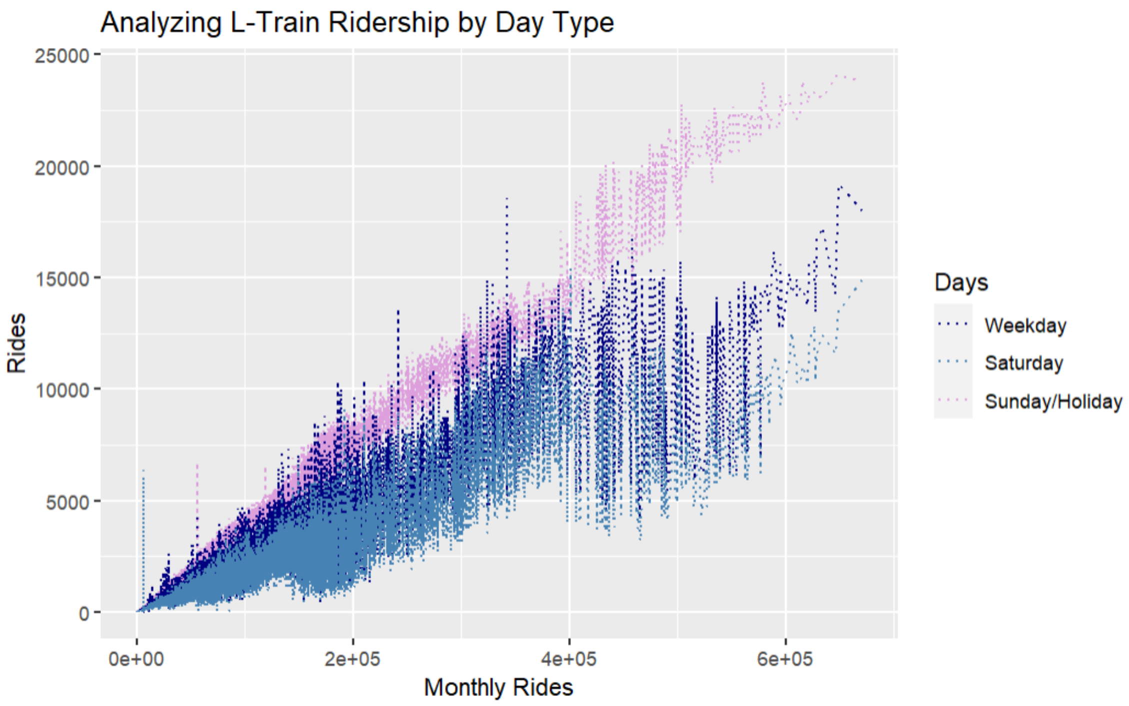




From the plots, we've observed that in 2023, August experienced the highest ridership, closely followed by October, for both stations and bus routes. However, our analysis encountered a challenge due to the data ending in October 2023.

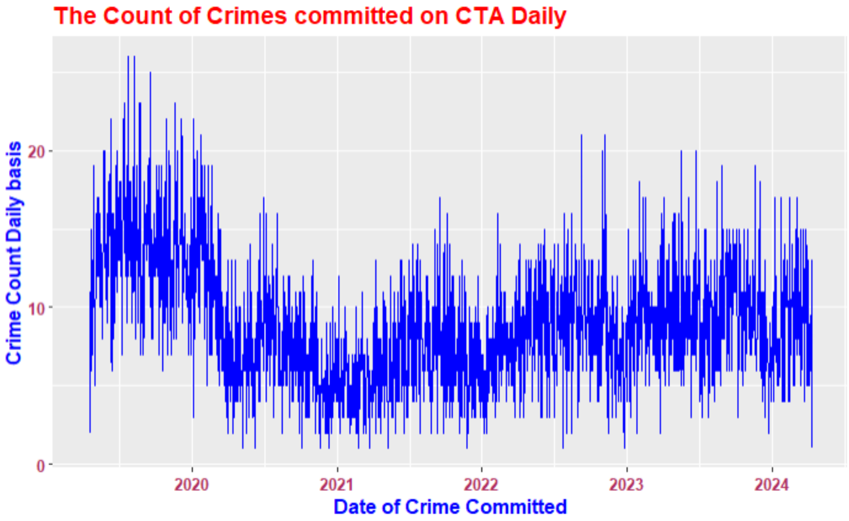
D. Analyzing Ridership by day type



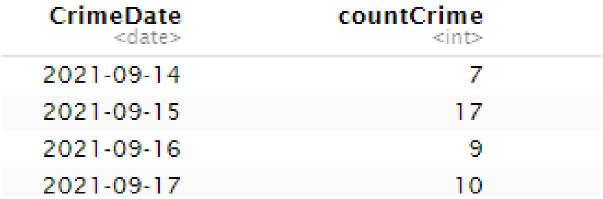


From the graphs we can infer that maximum trips for both bus and station are taken on Weekdays. Least number of trips are on Sundays/Holidays for both transits.

E. Crime Trends

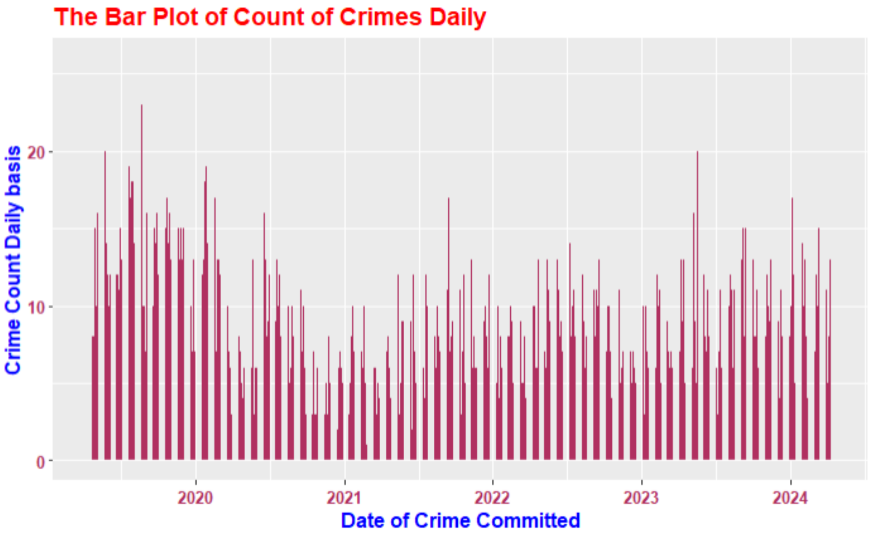


The above two graphs display the total crimes committed on the CTA bus and L trains from the years 2019 to April of 2024.
The below Bar plot presents the Crime patterns in a clearer pattern.The spikes in Crime rate during the middle of 2021 and 2023 are distinct.



CrimeDate <date>	countCrime <int>
2021-09-14	7
2021-09-15	17
2021-09-16	9
2021-09-17	10

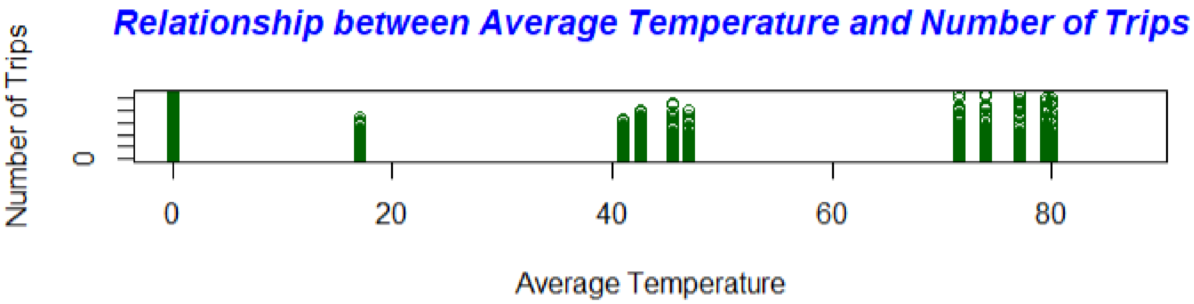
The latest spike is observed at the beginning of 2024 with the crime continuing on the high.



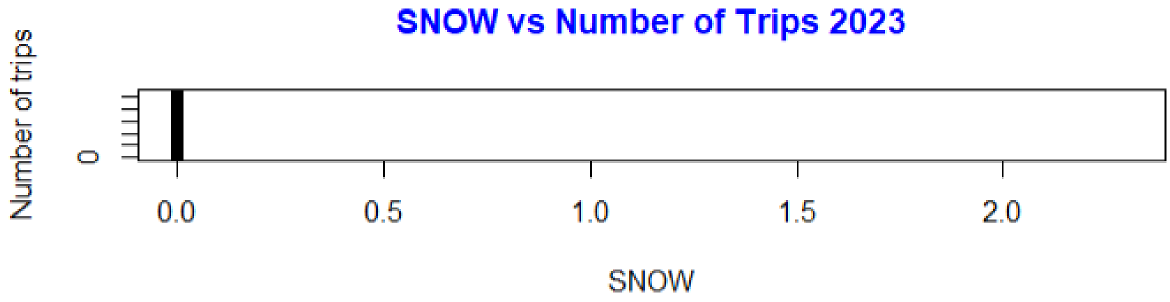
The graphs show the shift in Crime patterns wherein lowest count was in the period range mid-2020 to early 2022 and then a slight increase with occasional spikes in the end-2022 and 2023 period.

F. Weather Analysis

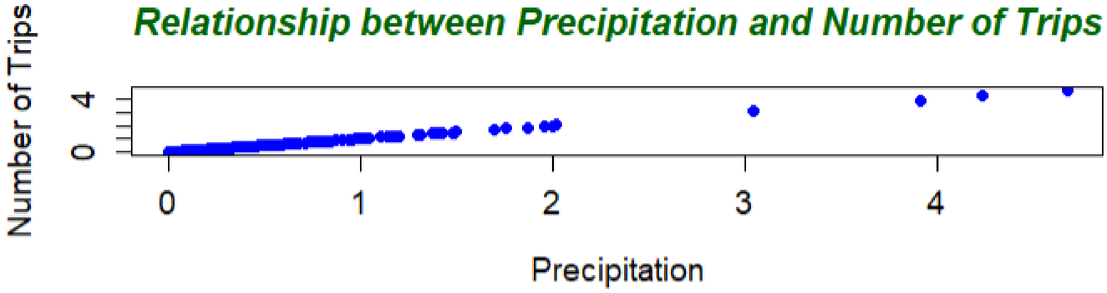
Analysis for Temperature:



Analysis for Snow:



Analysis for Precipitation:



The graphs above were plotted for the year 2023. They show the average number of trips taken in different weather conditions such as Snow & Rain.

4. DATA MODELLING

5.1 MULTIPLE LINEAR REGRESSION:

- The Model aims to predict transit route usage based on temperature, precipitation, and snowfall. The formulated model depicted:

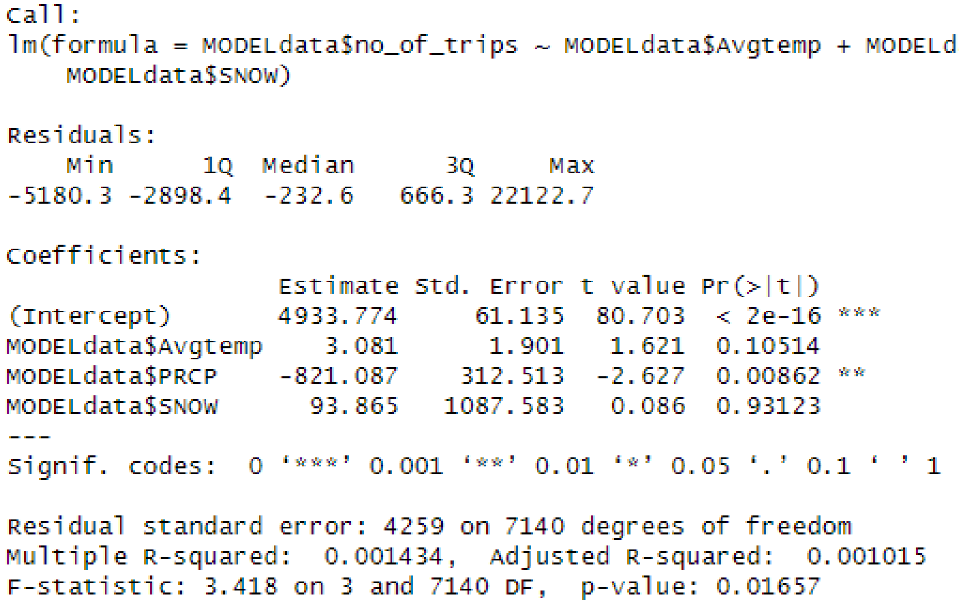
  formula = MODELdata$no_of_trips ~ MODELdata$Avg Temp + MODELdata$PRCP + MODELdata$SNOW)

- When the average temperature hits zero, the model estimates approximately 4933.774 trips on transit routes.
- Each unit rise in average temperature corresponds to an estimated increase of 3.081 trips on transit routes.
- An upsurge in precipitation is linked to a decrease of 821.087 trips on transit routes.
- Snowfall doesn't significantly impact transit route trips, as indicated by its non-significant coefficient (p-value = 0.93123).

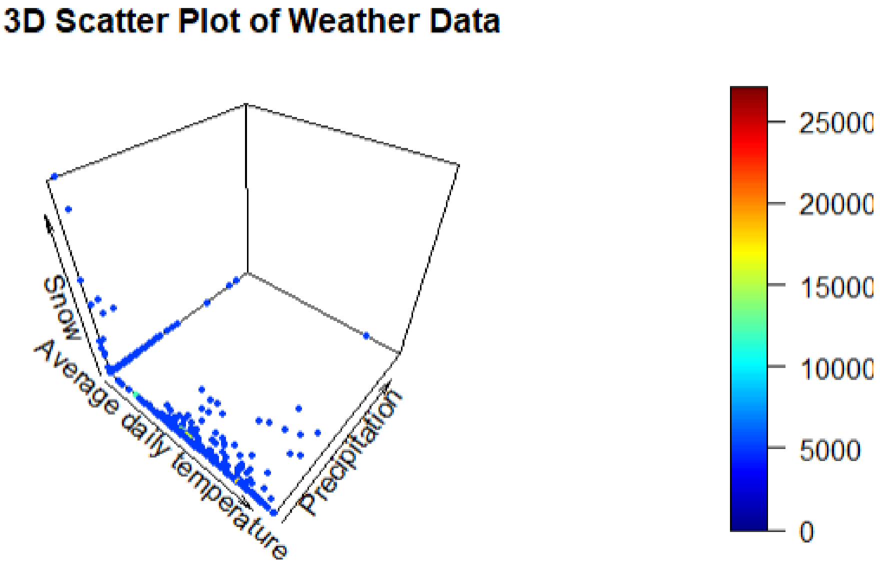
- While the model demonstrates statistical significance (F-statistic p-value = 0.01657) it has a low R-squared value of 0.001434.

The model shows that its explanatory capability is limited with temperature being the sole significant predictor of transit route usage.

The values generated are as given.

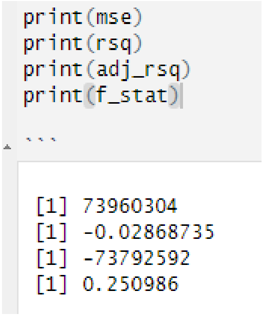


```
Call:
lm(formula = MODELdata$no_of_trips ~ MODELdata$Avgtemp + MODELdata$PRCP + 
    MODELdata$SNOW)

Residuals:
    Min      1Q  Median      3Q     Max 
-5180.3 -2898.4  -232.6   666.3 22122.7 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4933.774     61.135  80.703  < 2e-16 ***
MODELdata$Avgtemp    3.081      1.901   1.621  0.10514    
MODELdata$PRCP    -821.087    312.513  -2.627  0.00862 ** 
MODELdata$SNOW      93.865   1087.583   0.086  0.93123    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4259 on 7140 degrees of freedom
Multiple R-squared:  0.001434,	Adjusted R-squared:  0.001015 
F-statistic: 3.418 on 3 and 7140 DF,  p-value: 0.01657
```


3D Scatter Plot of Weather Data

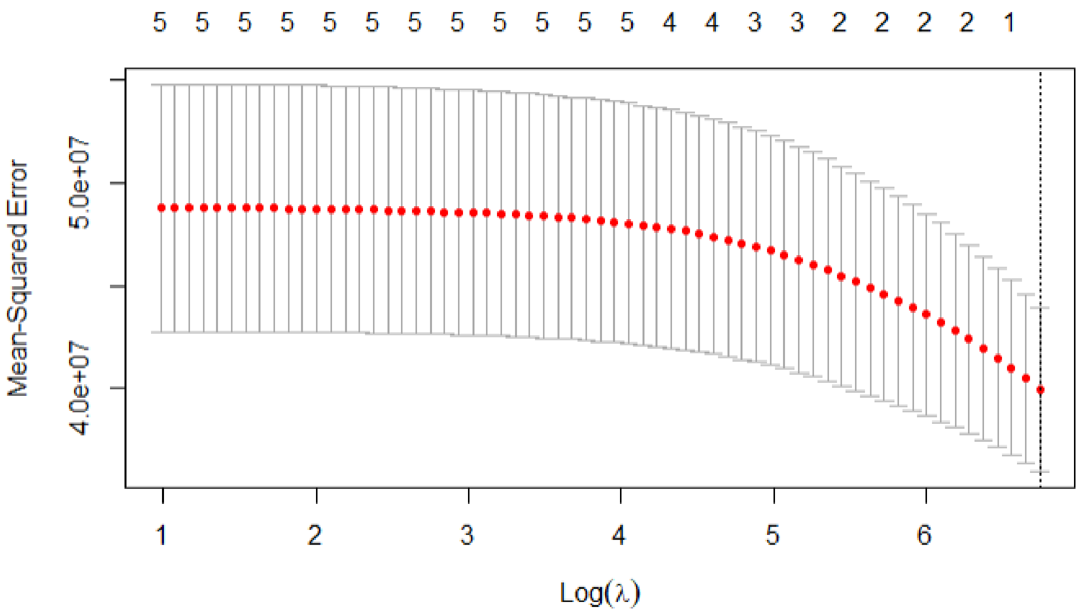
5.2 LASSO REGRESSION:

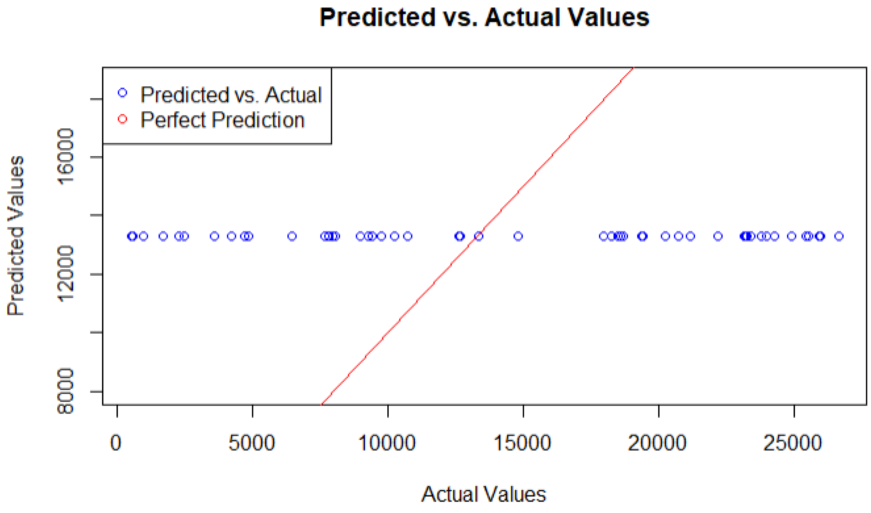
- The aim of this Lasso regression model was to predict the target variable "no_of_trips" based on the predictor variables "Avg_temp," "SNOW," and "PRCP” after addressing missing values via imputation.
- The dataset was split into training and testing sets, each containing half of the imputed data. Model performance was assessed using the mean squared error (MSE) on the test set. The calculated MSE is 73960304.
- Employing cross-validation, we determined the optimal lambda value for the Lasso model which was then utilized to make predictions.
- The computed MSE of 80.5041 indicates a moderate to good performance of the Lasso model on this dataset, providing reliable predictions for "no_of_trips" based on the input variables.
- This model is for practical application in forecasting "no_of_trips" values for new observations, leveraging the information from the predictor variables.
- Additionally, the R-squared value is -0.02868735, the Adjusted R-squared value is -73792592 and the F-statistic is 0.250986.



~~~
print(mse)
print(rsq)
print(adj_rsq)
print(f_stat)

```

 [1] 73960304
 [1] -0.02868735
 [1] -73792592
 [1] 0.250986
~~~

The plot illustrates the Lasso Regression:





Challenges we faced:

1. Initial dataset limitations constrained the project's scope, requiring exploration of alternative data sources.
2. Integration of crime data with ridership data aimed to identify security hotspots, yet the absence of gender data hindered gender-based analysis.
3. Despite limitations, the project offered valuable insights, highlighting the importance of curated datasets for robust analysis and laying groundwork for future endeavors in public transportation planning.

5. CONCLUSION

In conclusion, our comprehensive analysis of CTA transit data, crime data, weather data, and holiday data using multiple linear regression & lasso regression has yielded valuable insights into the intricate correlations between variables and transit ridership, crime rates, and weather conditions. Through this analysis, we've illuminated the significant impact of weather conditions, crime rates, and holidays on ridership patterns within the Chicago transit system. Notably, lasso regression has proven instrumental in identifying the most influential variables.. These findings are not only informative but also hold practical significance, offering valuable insights that can inform policy decisions, enhance transportation planning strategies, and contribute to efforts aimed at improving public safety within the community.

By delving into the busiest bus routes on average for different days of the week, we've uncovered consistent patterns that underscore the enduring prominence of certain routes, such as the 79th route, across weekdays, Saturdays, Sundays, and holidays. This steadfast ridership demonstrates the enduring demand for public transportation services, particularly on regular working days. Additionally, the aggregation of data to discern total traffic volumes across various days of the week has provided a comprehensive overview of transit patterns, offering crucial information that can guide resource allocation and service optimization efforts to ensure efficient service delivery and meet the evolving needs of commuters.

Furthermore, our examination of L-station traffic trends has shed light on the bustling activity at key stations throughout the week. For instance, weekdays consistently experience higher ridership compared to weekends, with specific routes and stations emerging as the busiest during peak hours notably, stations like Lake/State, Clark/Lake, and O'Hare Airport exhibit varying levels of ridership depending on the day. By understanding these fluctuations and trends, transportation authorities can better allocate resources, optimize station operations, and implement targeted strategies to enhance overall service quality and commuter satisfaction. While our analysis serves as a robust starting point for understanding transit dynamics in Chicago, further research and validation are warranted to delve deeper into causal relationships and refine existing models for more accurate predictions and informed decision-making in the realm of urban transportation.

6. FUTURE WORK

Moving forward, there are several avenues for enhancing transit services and analysis methods. Dynamic routing strategies, incorporating real-time data, can optimize transit routes for efficiency. Enhanced crime prevention measures and predictive modeling techniques can improve safety and anticipate ridership patterns. Interactive visualization tools and community engagement initiatives can facilitate better planning and meet the diverse needs of Chicago residents and visitors alike.

7. REFERENCES

1. “CTA-Annual Ridership Report” Chicago Transit Authority Assets report (Jan24,2022).
2. “Transit Trends - Chicago Metropolitan Agency for Planning” CMAP analysis of Federal Transit Administration. National Transit Database. Percent change in vehicle speed, 2005-15.
3. “ Mass Transit Security in Chicago” RC Johnson.
4. https://data.cityofchicago.org/Public-Safety/Crimes-2022/9hwr-2zxp/data
5. “https://www.researchgate.net/publication/348251988_Behavioral_dynamics_of_public_transit_ridership_in_Chicago_and_impacts_of_COVID-19 “Fissinger, Mary. (2020). Behavioral dynamics of public transit ridership in Chicago and impacts of COVID-19.
6. https://www.researchgate.net/publication/335270812_Preferences_for_travel-based_multitasking_Evidence_from_a_survey_among_public_transit_users_in_the_Chicago_metropolitan_area Krueger, Rico & Rashidi, Taha & Auld, Joshua. (2019). Preferences for travel-based multitasking: 65. 334-343. 10.1016/j.trf.2019.08.004.

Link to GitHub: https://github.com/Juveriyaae/CSP571-Project