CSP 571 - Data Preparation and Analysis
April 30th, 2022

Chicago Transit Authority Data Analysis

"Connecting Chicago, One Ride at a Time"

Team Members

Pranit Kotkar
pkotkar1@hawk.iit.edu
A20512027

Siddhi Shukla
sshukla12@hawk.iit.edu
A20516414

Anushka Chaubal
achaubal@hawk.iit.edu
A20511568

Rewa Deshpande
rdeshpande1@hawk.iit.edu
A20492328

Vaishnavi Shankar Devadig
vdevadig@hawk.iit.edu
A20516246

I. ABSTRACT

This abstract examines various factors that impact the usage of the Chicago Transit Authority (CTA), with a particular focus on the number of boarding passengers on transit routes. It compares the number of trips on different routes between weekdays and weekends and identifies the routes with the highest number of boarding passengers. The study also analyses how seasons, temperature, holidays, weekends, and crime statistics in specific areas and route numbers impact CTA usage.

To gain deeper insights into the data, the study employed various modelling techniques such as linear modelling, lasso regression, and clustering models. Additionally, the research investigates the effect of the COVID-19 pandemic on passenger statistics and provides recommendations to CTA on how to address routes with high passenger demand and inadequate transit vehicles. Furthermore, the study includes an analysis of location density diagrams for the busiest stations and explores methods for predicting the number of travels in a specific period of time.

II. OVERVIEW

2.1 Objective

The primary aim of this project is to tackle the issue mentioned earlier by utilising statistical techniques to analyse and forecast the quantity of trips in particular time periods, taking into account historical trends, weather patterns, and holiday information. We intend to construct a model that enables benchmarking analysis and can be utilised as a point of reference for subsequent investigations in this field.

2.2 Specific Questions

The questions we aim to answer upon completion of this project are related to the impact of external factors based on geographic location, weather conditions, seasons, and holidays:

1. What are the transit routes with the highest number of boarding passengers?
2. Compare the number of trips on various routes between weekdays and weekends.
3. The impact of seasons and temperature on the usage of Chicago Transit Authority.
4. The impact of holidays and weekends on the usage of Chicago Transit Authority.
5. The influence of crime statistics in the particular area and route number on the number of trips held.
6. How drastically have the passenger statistics changed due to COVID-19?
7. To identify the routes with a high demand for transit vehicles that are currently unable to meet passenger needs, and provide recommendations to CTA on how to improve transit services in those areas.
8. Location density diagrams for the busiest stations.
9. How can the number of travels in a certain period of time be predicted?
10. To analyse crime statistics in areas where CTA routes have a high ridership, and recommend possible measures to reduce crime and increase passenger safety on these routes.

III. DATA PROCESSING

CTA Bus data: *CTA Ridership Bus Routes*

CTA L-Train data: *CTA Ridership Train Routes Daily* and *CTA Ridership Train Routes Monthly*

CTA Daily Ridership data: *CTA Ridership Daily Boarding*

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

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

Crime data collected from Crimes - 2011 to 2012 : *Crime Dataset*

3.1 BUS DATA:

The bus dataset consists of the columns:

- route - The specific number assigned to a bus route.
- routename - The name of this bus route.
- Month_Beginning - The first date of every month from the year 2001 to present.
- Avg_Weekday_Rides - Route specific average number of rides taking place on weekdays in a month.
- Avg_Saturday_Rides - Route specific average number of rides taking place on Saturdays in a month.
- Avg_Sunday_Holiday_Rides - Route specific average number of rides taking place on Sundays and Holidays in a month.
- MonthTotal - The total number of route specific rides in a month.

CHANGES MADE TO THE DATASET:

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

3.2 L-TRAIN DATA:

The L-Train dataset consists of the columns:

- Station_id - The specific ID of the L-train station.
- Stationame- The name of the L-train station.
- month_beginning - The first date of every month from the year 2001 to 2022.
- Weekday_Rides - The total number of trips taken on a weekday.
- Saturday_Rides - The total number of trips taken on a Saturday.
- Sunday_Holiday_Rides - The total number of trips taken on a Sunday/Holiday.
- monthtotal - The total number of trips in a month.

CHANGES MADE TO THE DATASET:

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

3.3 WEATHER DATA:

The Weather dataset consists of the columns:

- Date- Date ranging from 2000 to 2011
- TAVG- Average temperature of the day in fahrenheit
- TMAX- Maximum temperature of the day in fahrenheit
- TMIN - Minimum temperature of the day in fahrenheit
- PRCP - Precipitation of the day
- SNOW- Amount of snowfall on the day
- SNOWD - Depth of the snow on the day

CHANGES MADE TO THE DATASET:

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

3.4 CRIME DATA:

The Crime dataset consists of the columns:

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

CHANGES MADE TO THE DATASET:

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

COMBINED DATA:

- Merged the L-train ridership data with weather data to get the date, number of trips, the temperature and the snowfall on a particular day. This data is further used for the multiple regression modelling.

TOOLS USED TO PERFORM DATA CLEANING AND PREPROCESSING:

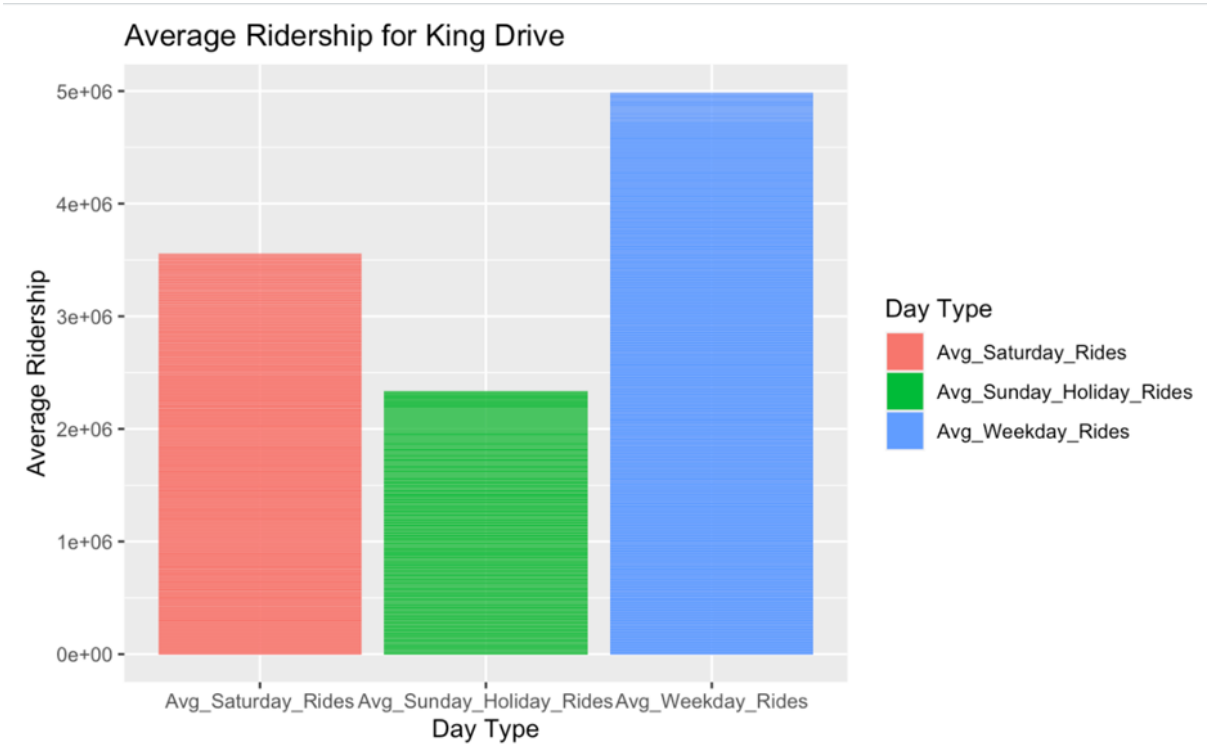
R programming language in R Studio.

IV. EXPLORATORY DATA ANALYSIS:

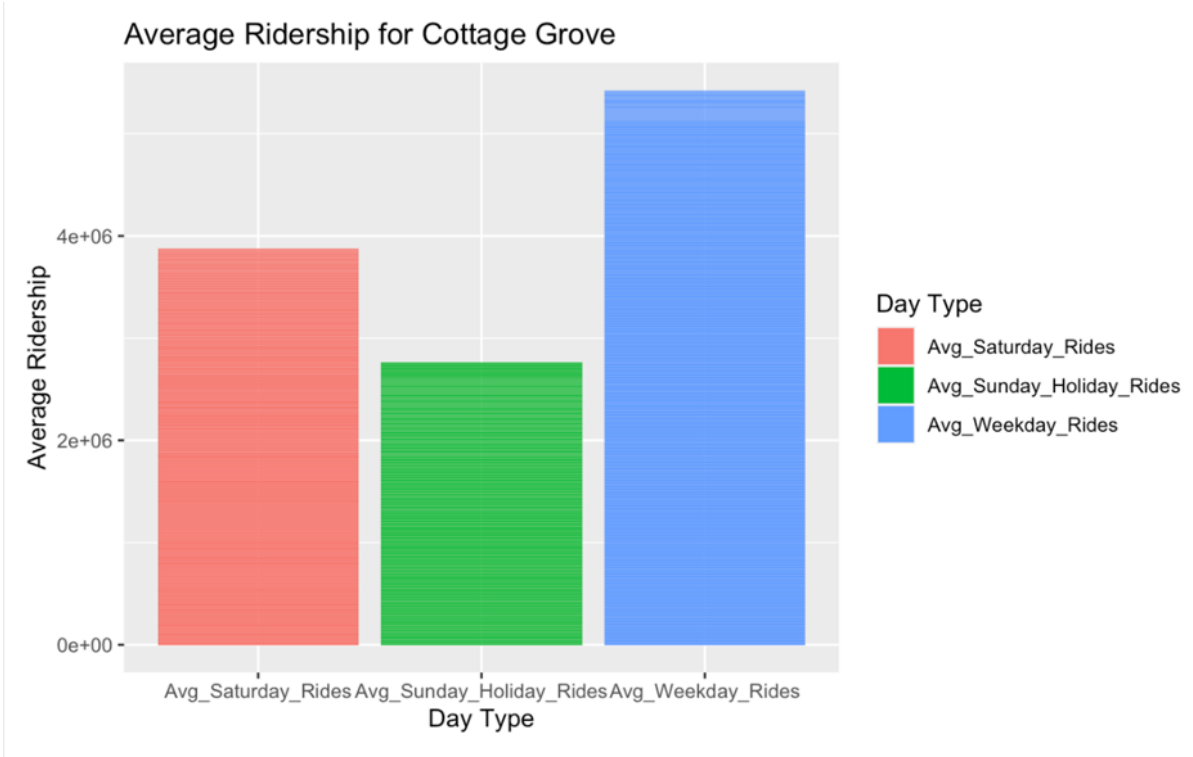
4.1 AVERAGE BUS RIDERSHIPS ON DIFFERENT TYPES OF DAYS FOR EACH BUS ROUTE:

There are 194 bus routes and we were able to generate ridership patterns for each of these routes. Graphs for four of these routes have been illustrated below.

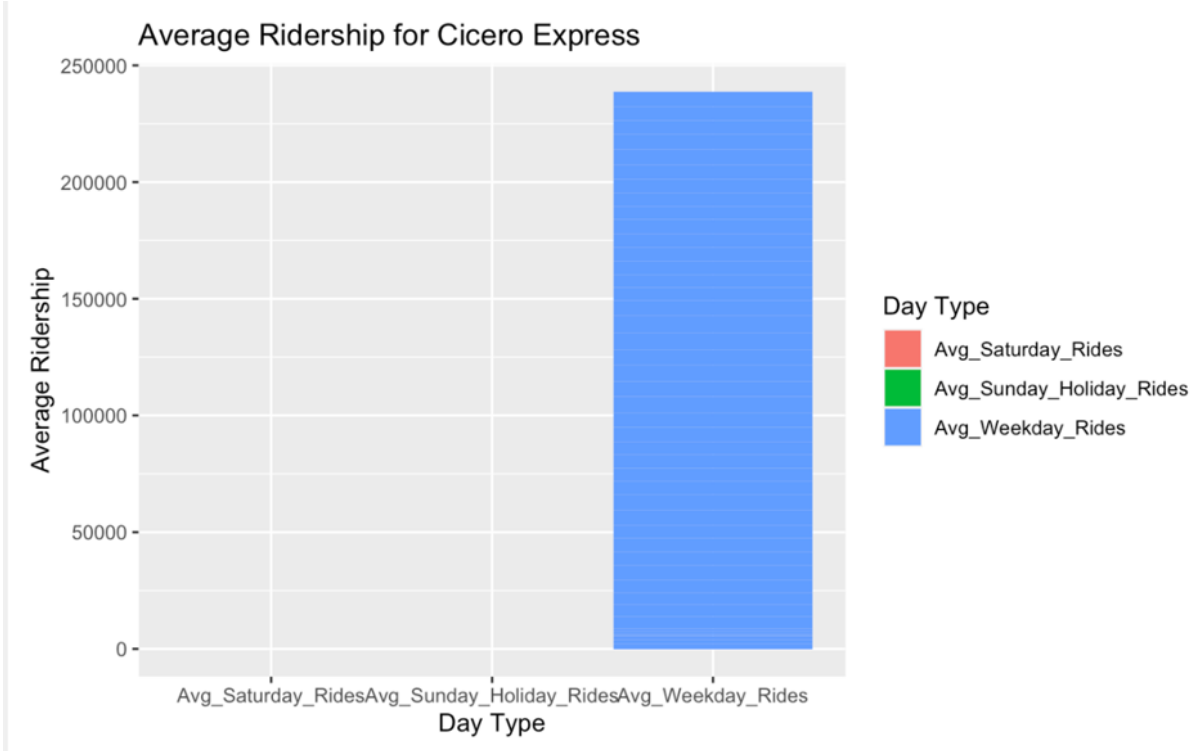
a. Average ridership on Weekdays, Saturdays, Sundays and Holidays for the route King Drive:



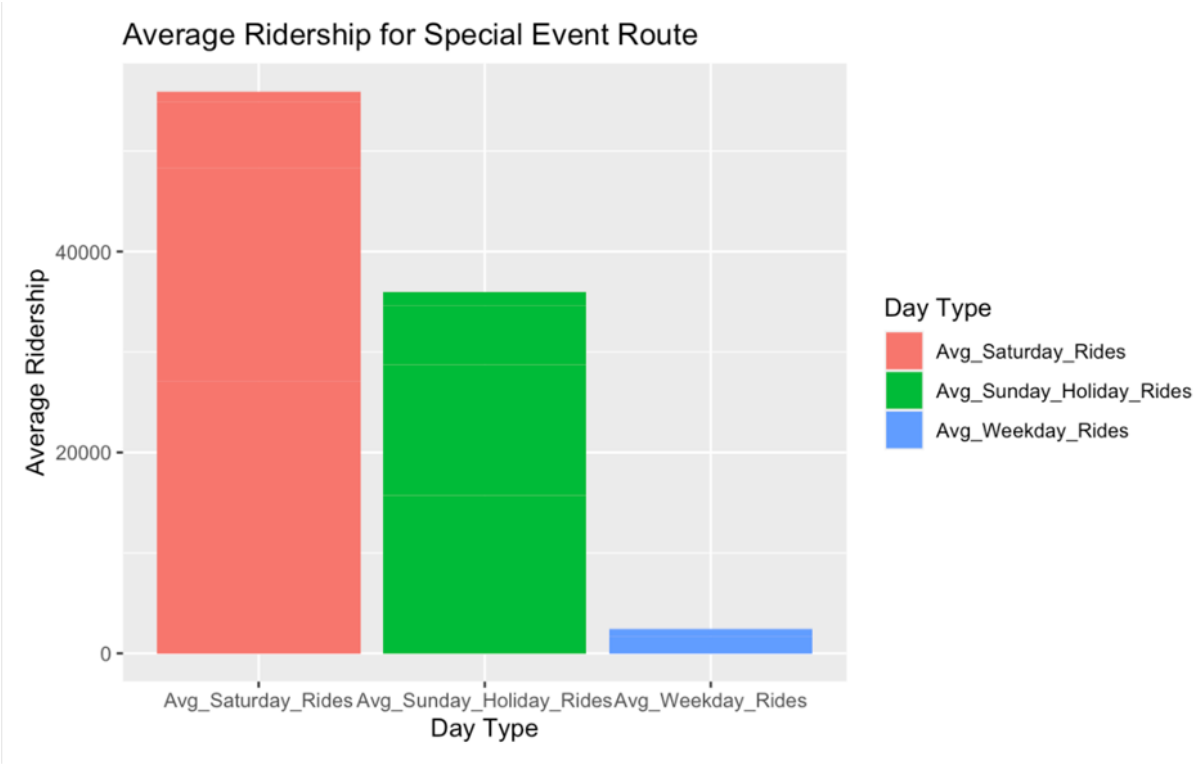
b. Cottage Grove:



c. Cicero Express:



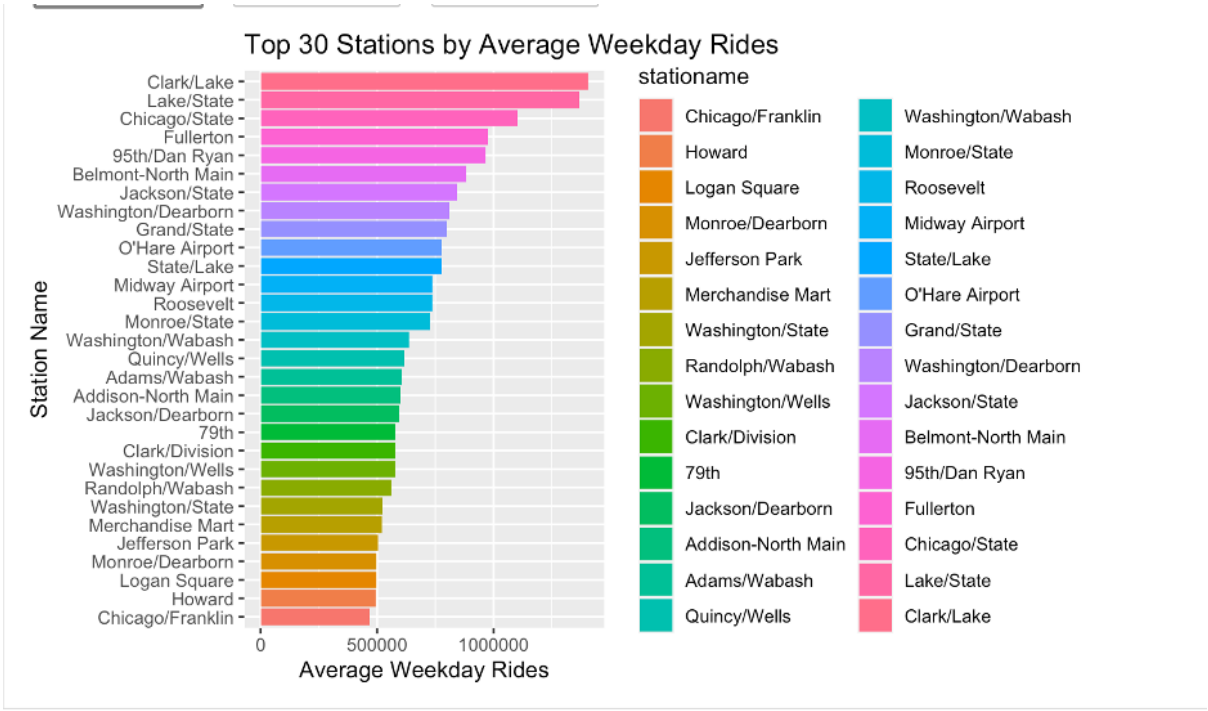
d. Special Event Route:



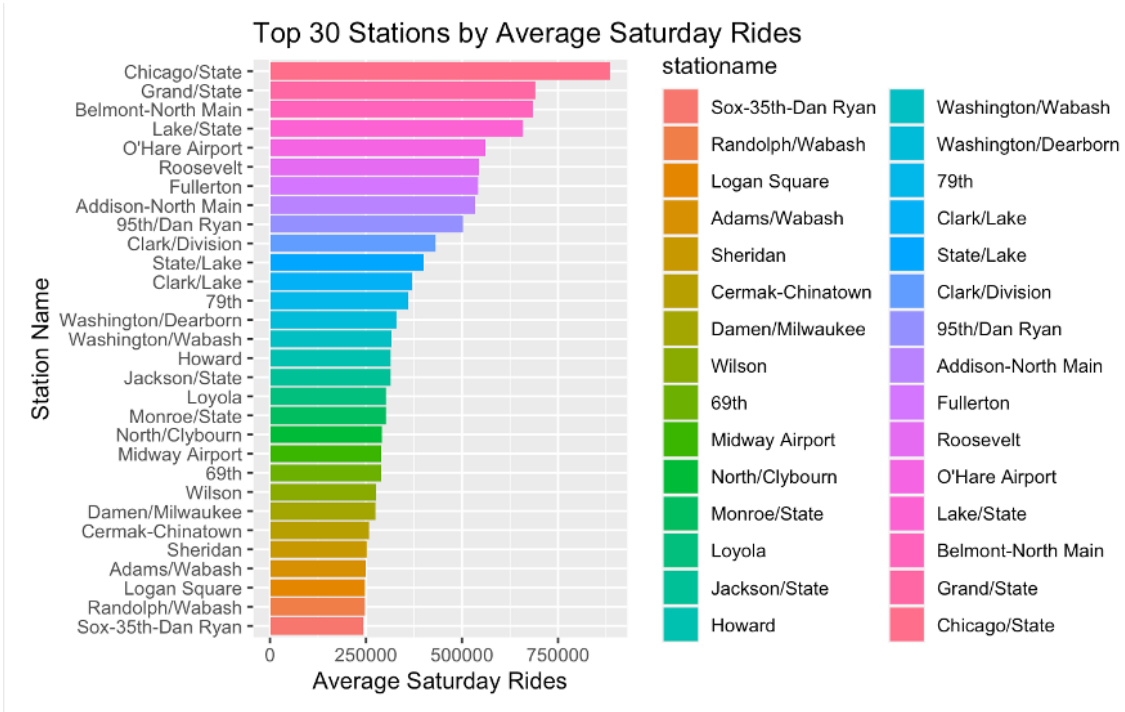
- As it is noticed here, for King Drive and Cottage Grove, which are some of the busiest bus routes of CTA have high weekday ridership due to several work locations found in this route. However, the ridership is high on weekends and holidays as well, as they are well connected.
- On the other hand, the Cicero Express route operates only on weekdays as it does not have sufficient passenger count on weekends. This could be attributed to the route operating outside the busy areas in Chicago.
- The Special Event route operates only on weekends and on holidays as these routes are reserved for specific purposes only.

4.2 L-STATIONS

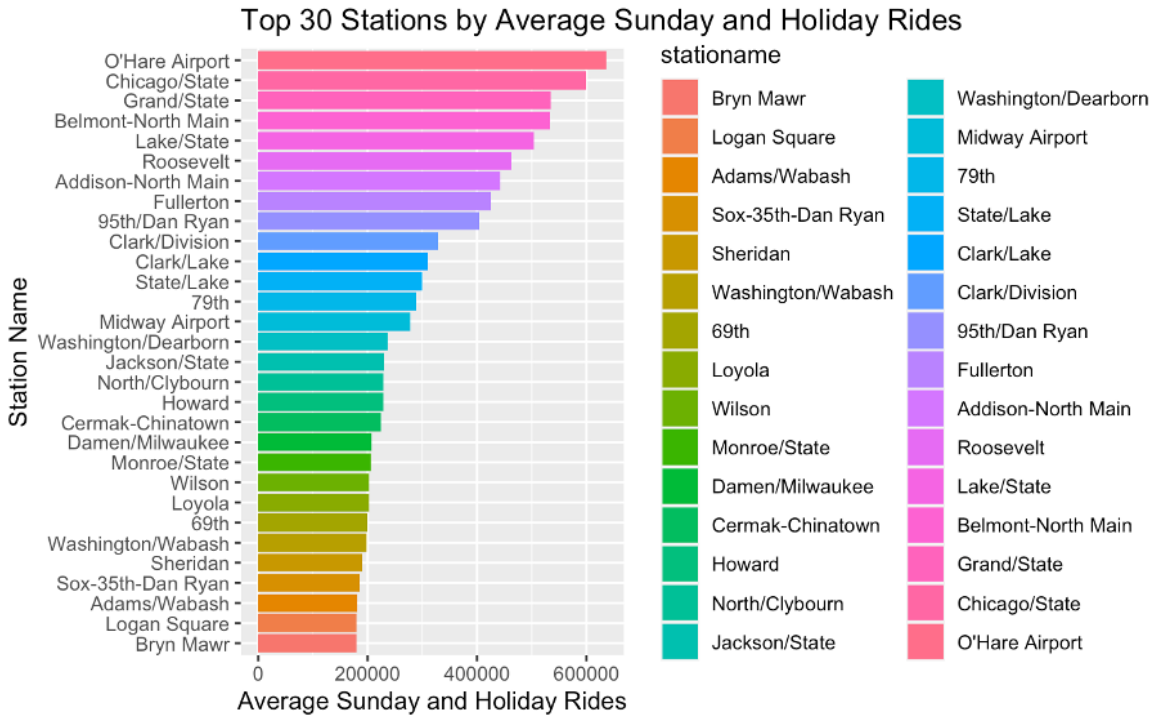
a. Top 30 L-Stations operating on weekdays:



b. Top 30 L-Stations operating on Saturdays:

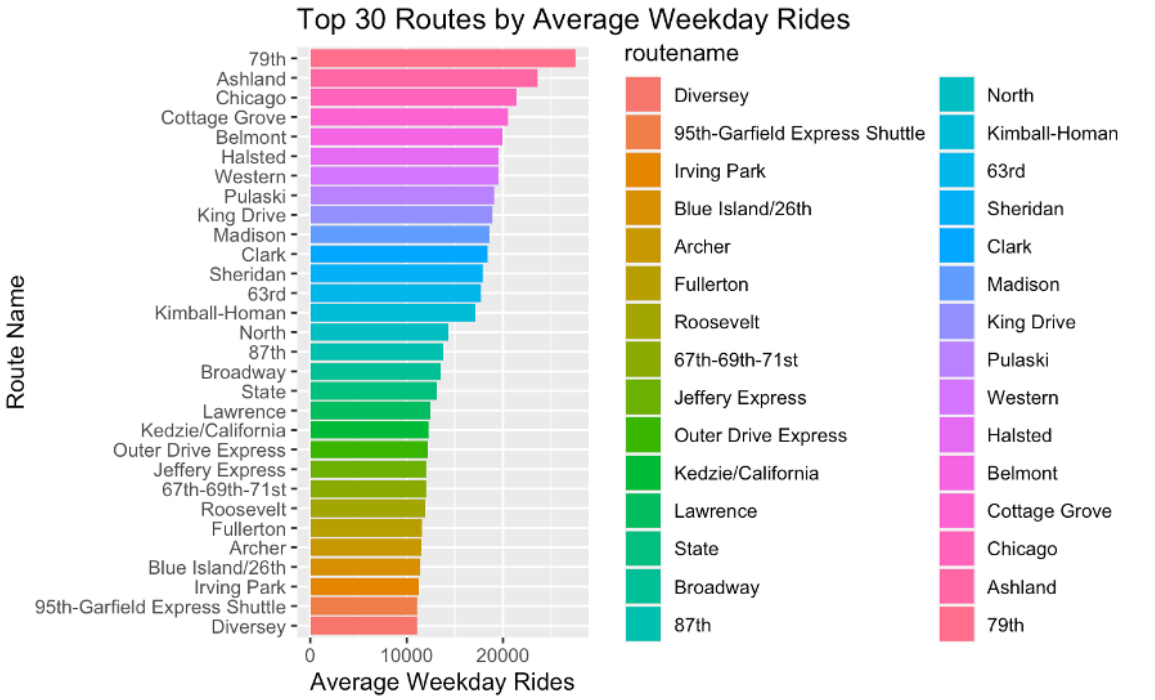


c. Top 30 L-Stations operating on Sundays and Holidays:

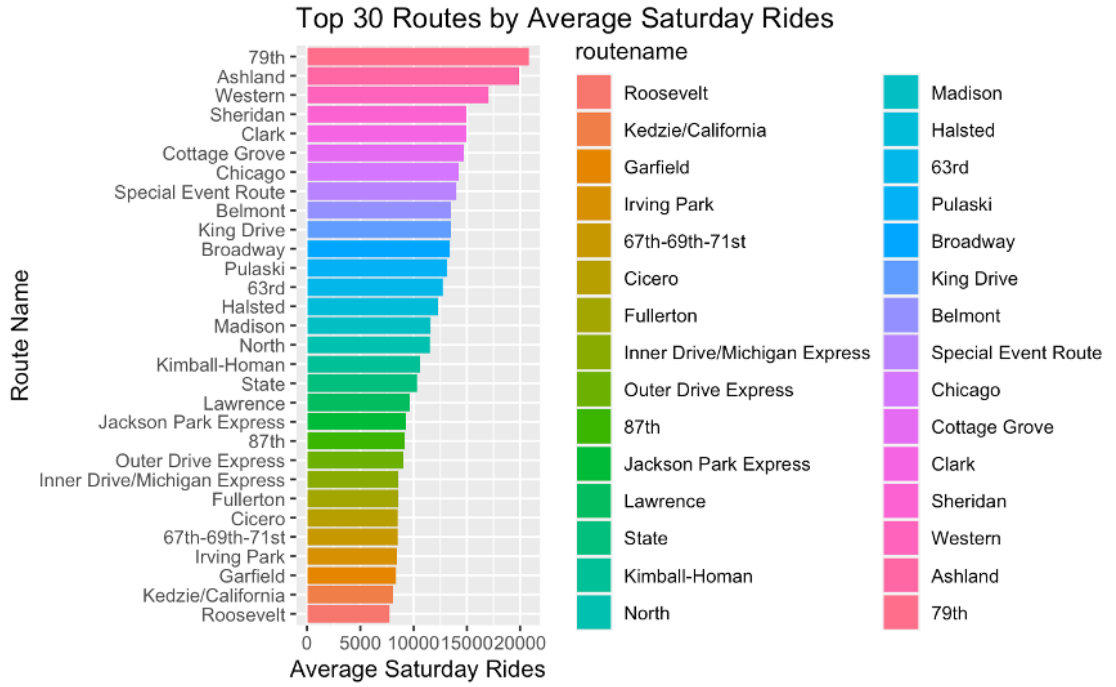


4.3 BUS ROUTES:

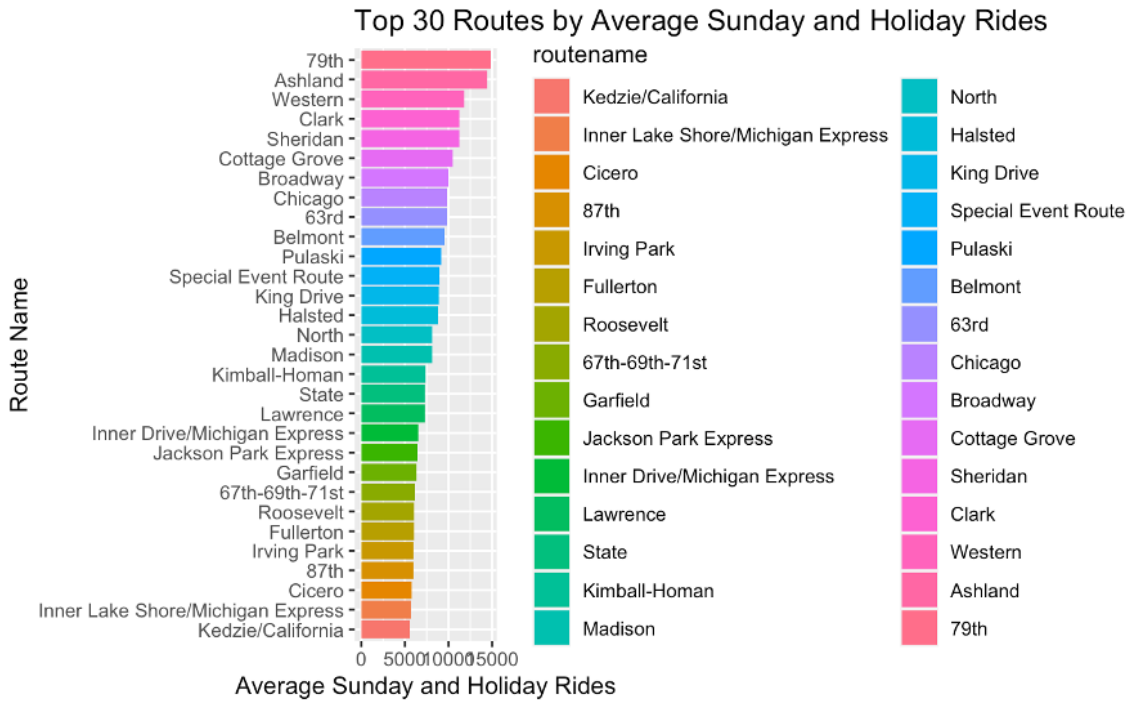
a. Top 30 bus routes operating on weekdays:



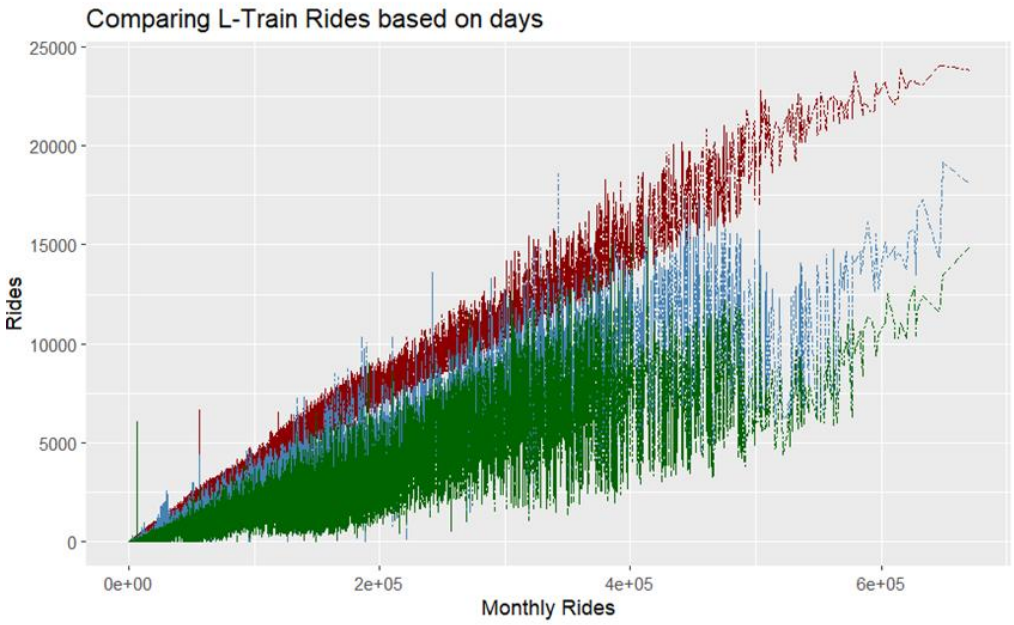
b. Top 30 bus routes operating on saturdays:



c. Top 30 bus routes operating on sundays and holidays:



d. Ridership of L-Train based on days of the week:

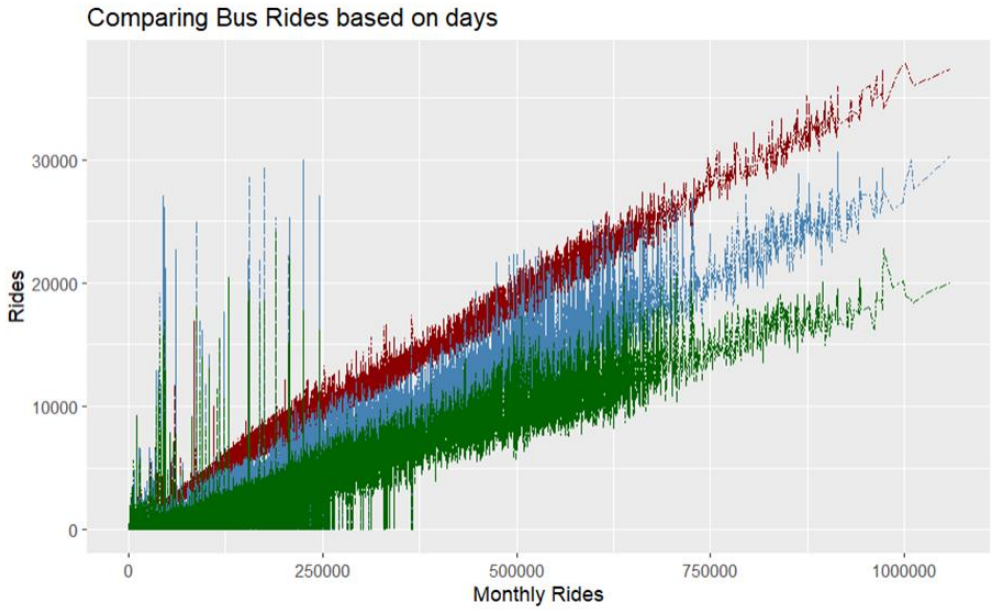


In the above graph,

- The red points represent the Weekday_Rides for L-Train
- The blue points represent the Saturday_Rides for L-Train
- The green points represent the Sunday/Holiday Rides for L-Train

From the above graph we can conclude that maximum trips are taken on weekdays followed by Saturday trips. And the least trips are taken on Sundays / Holidays.

e. Ridership of CTA Bus based on days of the week:

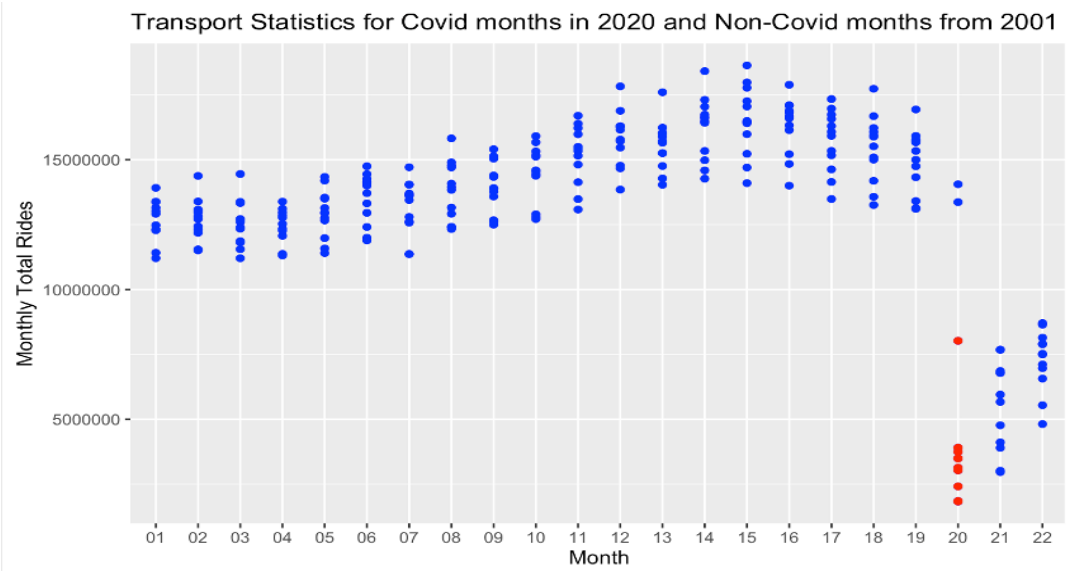


In the above graph,

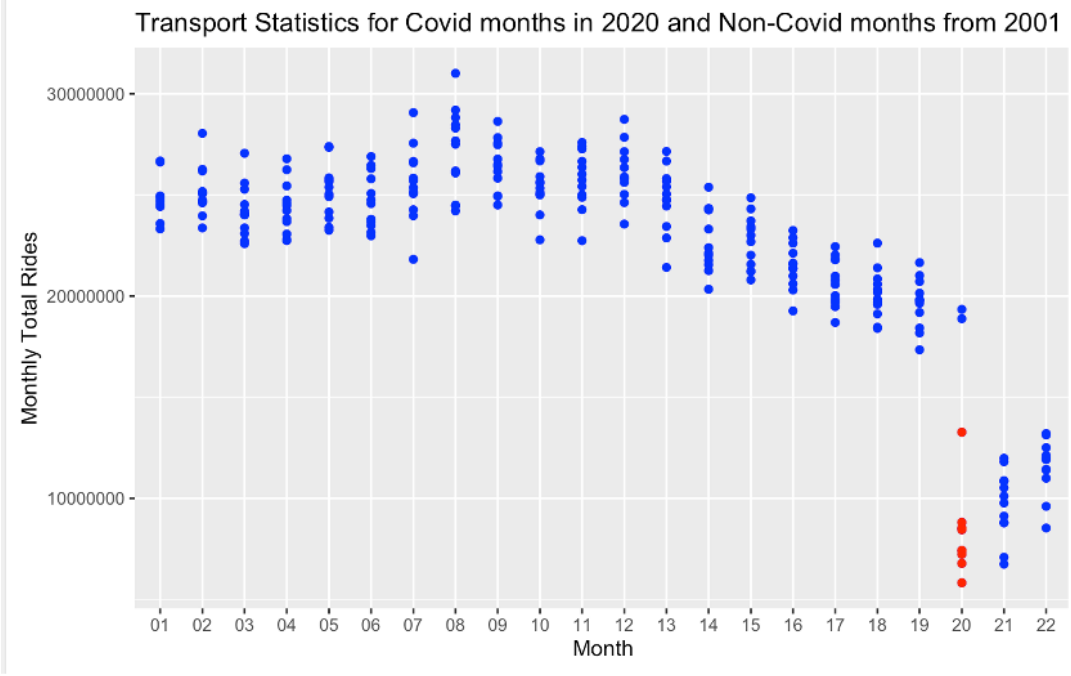
- The red points represent the Weekday_Rides for Bus
- The blue points represent the Saturday_Rides for Bus
- The green points represent the Sunday/Holiday Rides for Bus

From the above graph we can conclude that maximum trips are taken on weekdays followed by Saturday trips. And the least trips are taken on Sundays / Holidays.

f. Riderships comparing covid months in 2020 and all other months from 2001 to 2022:

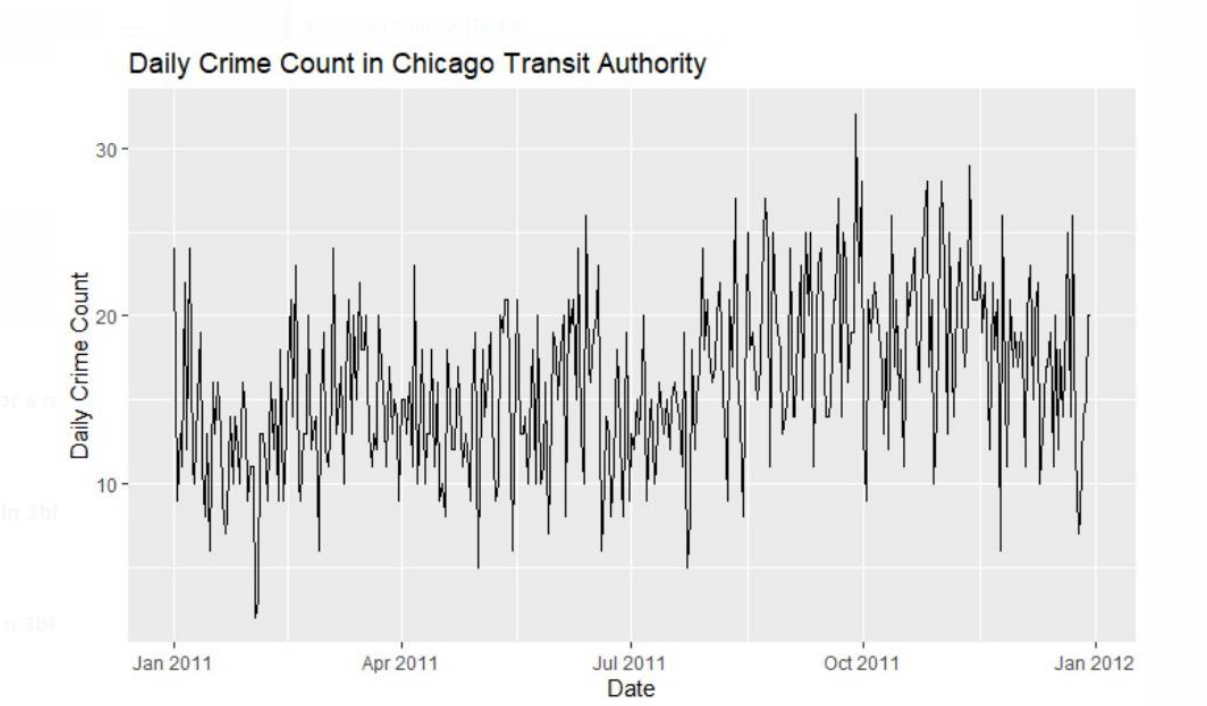


g. CTA bus riderships comparing covid months in 2020 and all other months from 2001 to 2022:



- The above two graphs attempt to compare ridership during normal months from 2001 to 2022 with peak COVID months in the year 2020, which are from March to December.
- The peak COVID months are shown in red, while the rest are in blue.
- It is clearly observed that riderships in both L-trains and buses was the lowest during COVID, thus solidifying the proof for impact of the pandemic on transport.

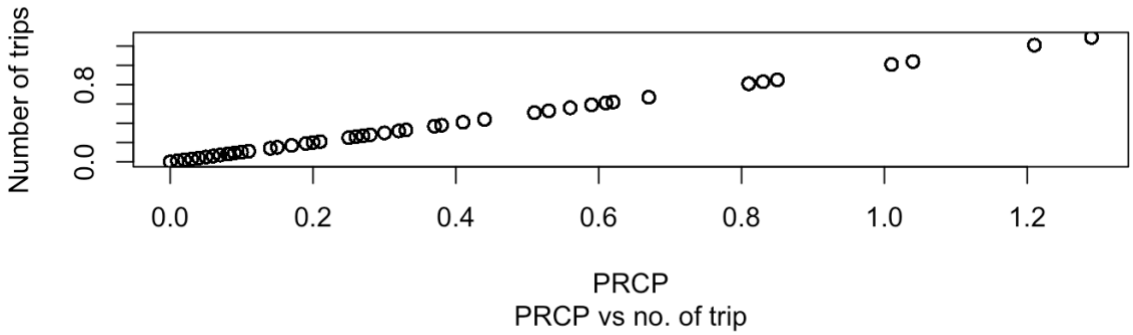
4.5 CRIME DATA:



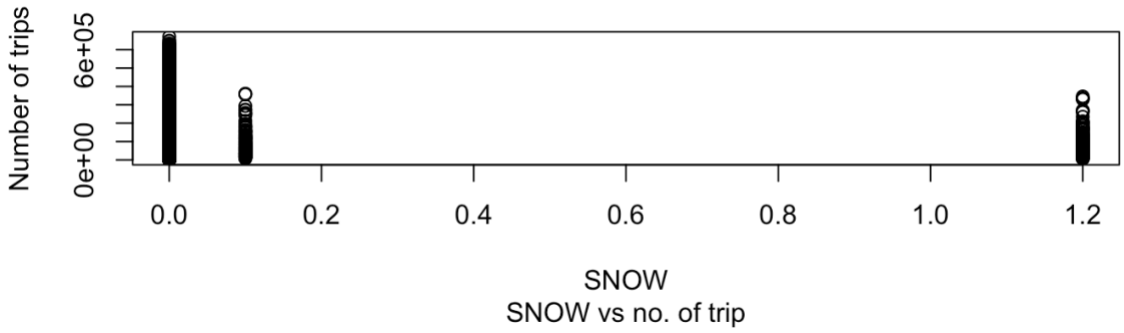
The graph represents the crime count pattern of the Chicago Transit through Jan 2011 to Jan 2012.

4.6 WEATHER RELATED ANALYSIS:

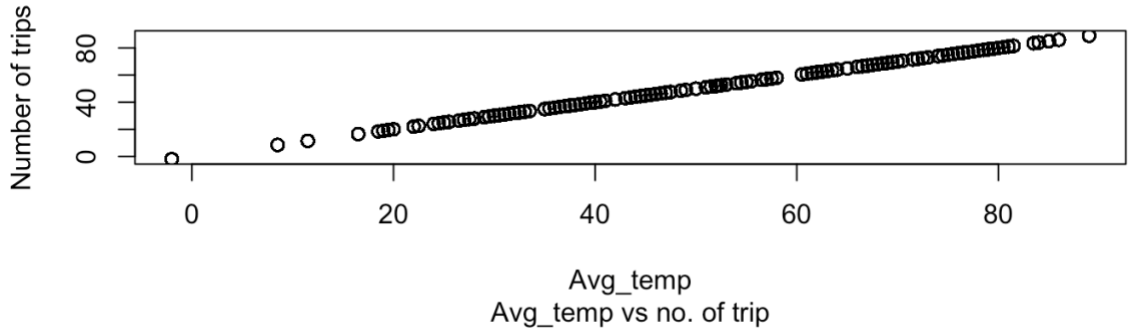
a. Precipitation versus the number of trips:



b. Snow versus the number of trips:



c. Temperature versus the number of trips:

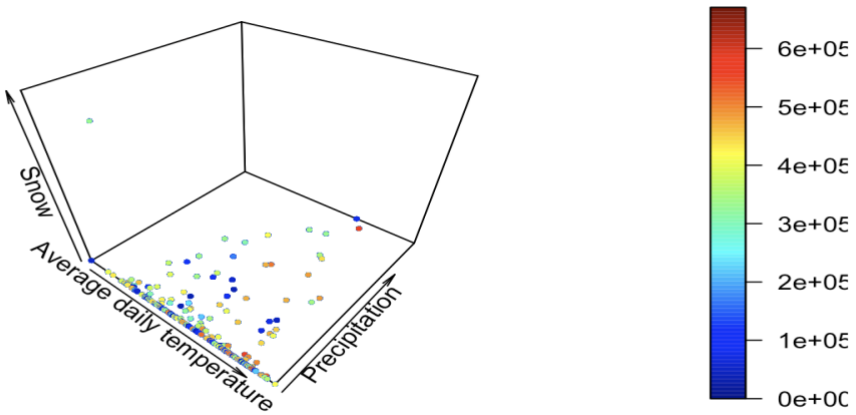


V. DATA MODELLING:

5.1 MULTIPLE LINEAR REGRESSION:

- In this model, we predicted the number of rides based on the temperature, precipitation and snowfall.
- The regression analysis shows that the number of trips taken on transit routes is significantly impacted by average temperature, but not by precipitation or snowfall.
- The intercept value suggests that the number of trips on transit routes when the average temperature is zero is 82902.35.
- The R-squared value is very low, indicating that the model does not explain much of the variance in the data.
- However, the p-value for the F-statistic is very low, indicating that the model as a whole is statistically significant.

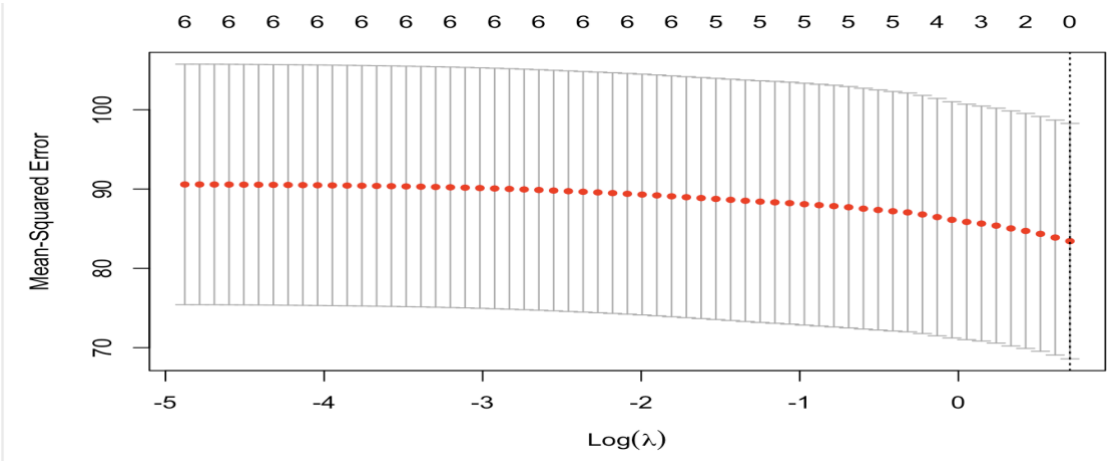
The plot below shows the multiple linear regression:



5.2 LASSO REGRESSION:

- The Lasso regression model was used to predict the target variable "y" using the "date", "station_id", "num_trips", "avg_trip", "snow", and "precipitation" predictor variables, after performing imputation using the MICE package to handle missing values.
- The model was trained on half of the imputed data and tested on the other half, with the mean squared error (MSE) being used to evaluate the performance of the model.
- Cross-validation was used to select the best lambda value for the Lasso model, and the resulting predictions were made using this value.
- The MSE was calculated to be 80.5041, indicating that the Lasso model has a moderate to good performance on this dataset.
- The model can be used to predict the value of "y" for new observations based on the values of the predictor variables.

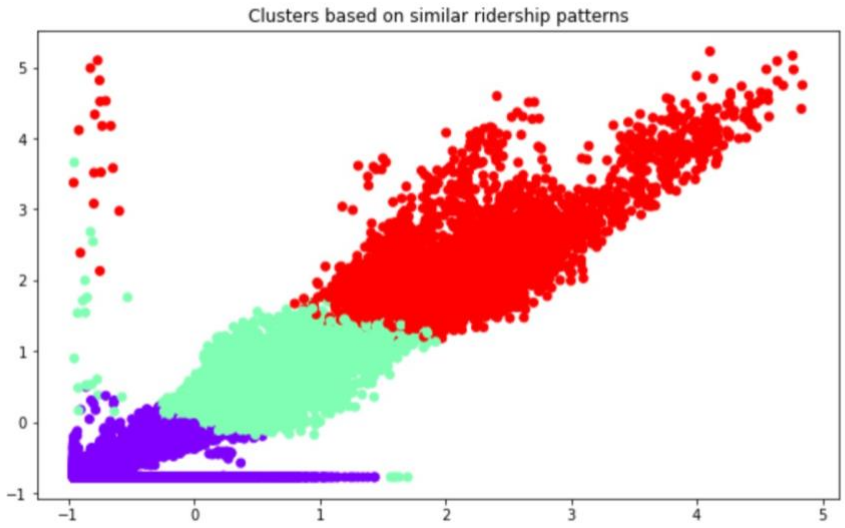
The plot below shows the Lasso Regression:



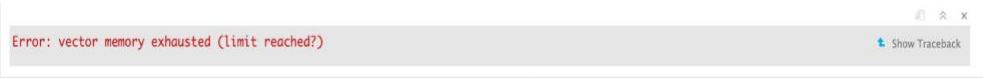
5.4 K-MEANS CLUSTERING:

- The K-means clustering algorithm clusters the CTA Bus Routes based on similar ridership patterns.
- The features available are ‘route’, ‘routename’, ‘Month_Beginning’, ‘Avg_Weekday_Rides’, ‘Avg_Saturday_Rides’ and ‘Avg_Sunday_Holiday_Rides’.
- The features taken into consideration for clustering are ‘Avg_Weekday_Rides’, ‘Avg_Saturday_Rides’ and ‘Avg_Sunday_Holiday_Rides’.
- As a result, three clusters were created, 0, 1, and 2.
- We also print the bus data to identify which bus route has been classified into which cluster.
- For instance, in this case, Indiana/Hyde Park is classified into Cluster 0, Roosevelt into Cluster 1, and King Drive into Cluster 2.
- Certain metrics were calculated for the clustering model.
- The Silhouette score for the model is 0.6334, the Calinski-Harabasz score is 104757.738, and the Davies_Bouldin score for the model is 0.5666.

The plot below shows the k-means clustering:



Reason for using Python:



- We tried clustering in R studio but as we started making modifications we quickly ran out of available memory and started getting errors like “vector memory exhausted” and the R studio crashed a lot.
- Using the Pandas module in Python worked better and faster to do clustering on the data.

Technical Challenges:

- Managing multiple datasets in R Studio can be a significant hurdle when working on local machines or laptops. The restricted memory capacity of these systems can be quickly exhausted, causing the session to be terminated.
- Although we were able to perform the initial analysis of the data, running a lasso regression didn't work with all of the data we had. In order to find a workable model, we had to run a portion of our initial dataset.

VI. CONCLUSION

The analysis of CTA transit data, crime data, weather data, and holiday data using multiple linear regression, lasso regression, and kmeans clustering provided insights into the correlation between variables and transit ridership, crime rates, and weather conditions. The results highlighted the significant impact of weather conditions, crime rates, and holidays on ridership patterns. Lasso regression helped identify the most important variables, while kmeans clustering grouped similar data points to identify trends. These findings can inform policy decisions, improve transportation planning, and increase public safety. However, the analysis is limited by the data and methods used, and further research is needed to validate the results and determine causal relationships. Overall, the analysis serves as a useful starting point for further investigation.

VII. SCOPE OF THE PROJECT:

The scope of the project was limited due to the unavailability of a suitable dataset. As a result, we were unable to perform a comprehensive analysis of the desired aspect of the project. However, we were able to identify potential sources for obtaining the required data and explore alternative solutions for data collection.

We can analyse the CTA crime data and can provide insights into the safety of different stations and areas, identifying any patterns or trends in criminal activity. And by combining crime data with ridership data, it would be possible to identify hotspots where crimes are more likely to occur, and this can inform the allocation of resources to improve security in those areas.

COVID-19 has also affected the optimal routes and schedules for CTA buses and trains. Analysing CTA data and COVID-19 data can help identify areas where service levels can be reduced or increased to better match demand. This can help reduce costs while ensuring that essential services are maintained. Unfortunately, we were unable to link the COVID-19 dataset with the CTA routes dataset, which limited our ability to conduct analysis on the impact of the pandemic on the public transportation system in Chicago.

One potential scope for analysing the CTA dataset by gender is to identify areas where service levels may need to be adjusted to better match demand. By analysing CTA ridership data by gender, we can gain insights into the demographics of ridership and identify areas where there may be significant differences in ridership levels between genders on particular routes or during specific times of the day. This can help inform decisions on how to optimise service levels to better accommodate demand and improve the overall efficiency and effectiveness of the public transportation system in Chicago. As we did not have any gender data available in the CTA dataset, conducting analysis based on gender was not feasible.

Despite the limitations, the project was able to provide valuable insights into the available data and highlight the importance of having a well-curated dataset for conducting data analysis.

VIII. REFERENCES

- https://www.axios.com/local/chicago/2022/04/21/mapping-cta-crime-statistics-chicago-trains
- https://www.nctr.usf.edu/pdf/527-14.pdf
- https://www.wsdot.wa.gov/partners/erp/background/ST3%20Draft%20RidershipForecastingMethodologyReport_6March2015.pdf
- https://www.apta.com/research-technical-resources/research-reports/transit-workforce-shortage/
- https://www.researchgate.net/publication/354472520_Examination_of_New_York_City_Transit's_Bus_and_Subway_Ridership_Trends_During_the_COVID-19_Pandemic