NAVIGATING CHICAGO

Analysis of CTA Transit Routes

- Project by:
- Juveriya Fatima
- Vidheesha Patil

INTRODUCTION

Background

- The Chicago Transit Authority (CTA) serves millions of passengers daily.
- Understanding ridership patterns can enhance transit planning and commuter experience.

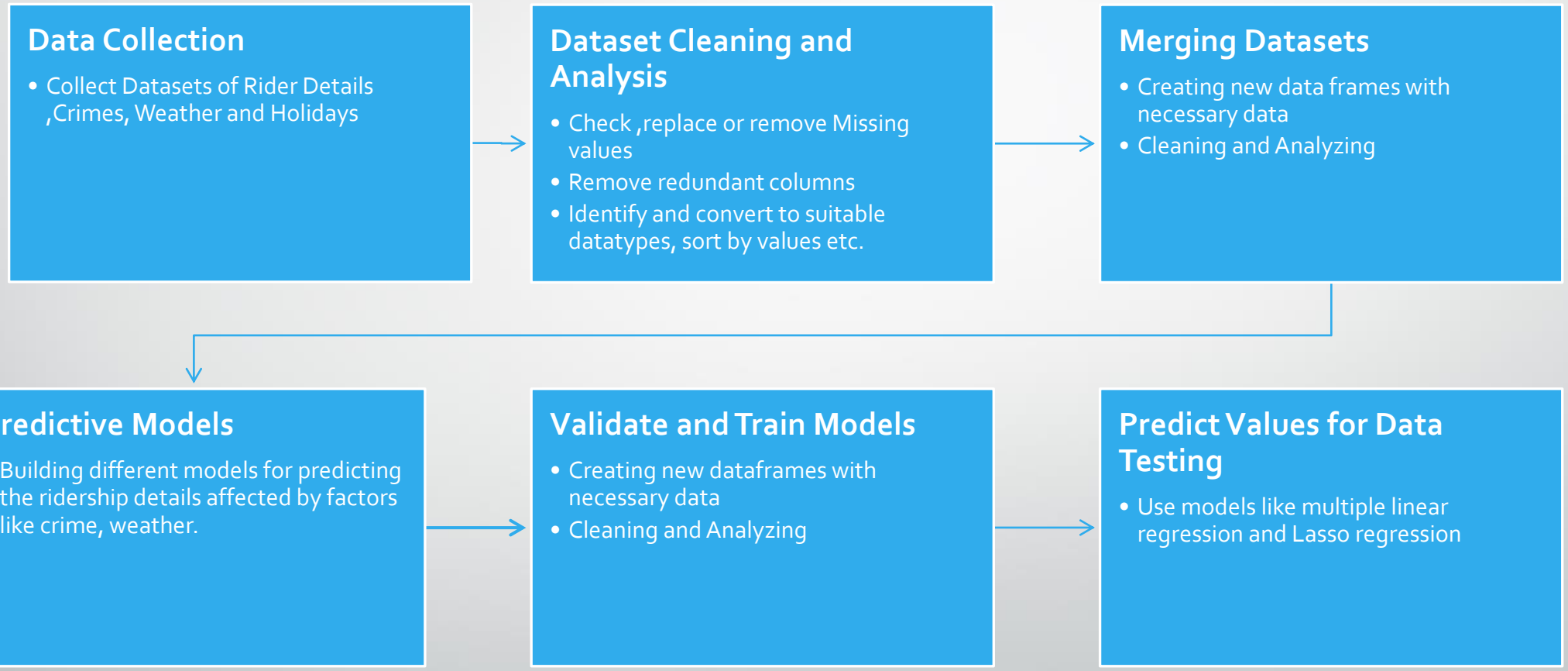
Aim

- Analyze CTA data along with weather and crime statistics collected over a fixed time period.
- Identify factors influencing transit ridership including weather,crime,riders details.
- Provide insights for improving transit services and safety measures.

Objectives

- Analyze ridership trends across different days, routes, and weather conditions.
- Examine the impact of crime rates and holidays on transit usage.
- Develop predictive models to forecast transit demand.

2. DATA PROCESS FLOW



3. DATA COLLECTION

Data Sources

1.CTA Ridership Data

https://www.transitchicago.com/data/

2.Weather Data (NOAA)

https://www.noaa.gov/

3.Crime Data (City of Chicago)

https://qgis.org/en/site/

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data

Data Types

Ridership: Bus and L-Train data

Weather: Temperature, precipitation, snowfall, snowfall depth

Crime: Incidents related to CTA transit areas

Period Covered:

Data spans from 1967 to 2024 for comprehensive analysis.

Actual processing done on data from 2013 to 2024 for detailed analysis.

Github Repository for Code

- https://github.com/Juveriyaae/CSP571-Project

4.Data Cleaning

1.CTA Ridership Data

A. Bus Data

Converted to appropriate data types: Date for the Month_Beginning column.

Check for 0 and missing values eliminated.

B. L-Train Data

Worked on 2 datasets-Daily entries and Monthly entries.

Filtered data from 2013 to 2023.

Four missing values found from a redundant column.Column removed from analysis.

4.Data Cleaning

2.Weather Data (NOAA)

Weather data collected from 4 different regions in Chicago city.

Data from January 1 2023 to January 31 2023 was selected.

Missing values in Precipitation, Snow, and Snow Depth columns were replaced with zeros and no duplicates were found.

Average temperature values were manually calculated.

The Date column was converted to "%Y-%m-%d" format

3.Crime Data (City of Chicago)

Data filtered from 18 April 2019 to 18 April 2024 and locations of CTA L train, bus, platform, parking, bus stop were selected.

Date column split it into 3 columns of Date, Time and AM/PM.

The columns X.Coordinate, Y.Coordinate, Latitude and Location were eliminated

records missing for non-essential columns Ward, Latitude, Longitude, ZipCodes

4.Data Cleaning

4.Merged Data

Combined the individual datasets of CTA Ridership by station and Weather on the Date column.

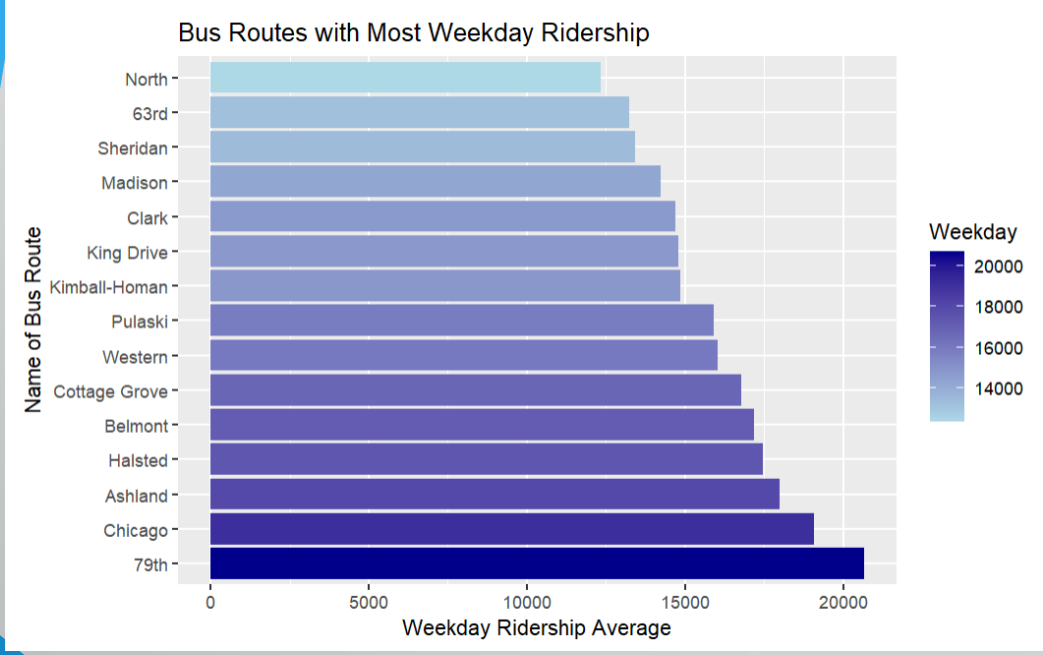
Created new column ‘no_of_trips’ by adding 3 columns of average rides on weekday, saturday and sunday/holiday.

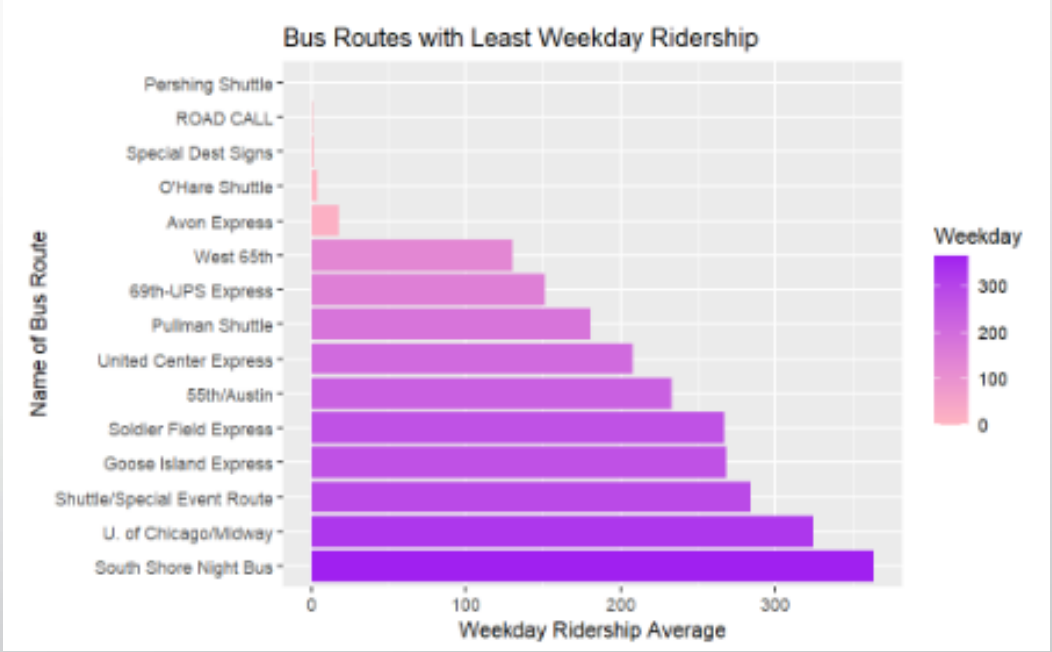
Date column converted to format of mm-dd-yyyy .

The merged dataset has 15 columns.

4. EXPLORATORY DATA ANALYSIS

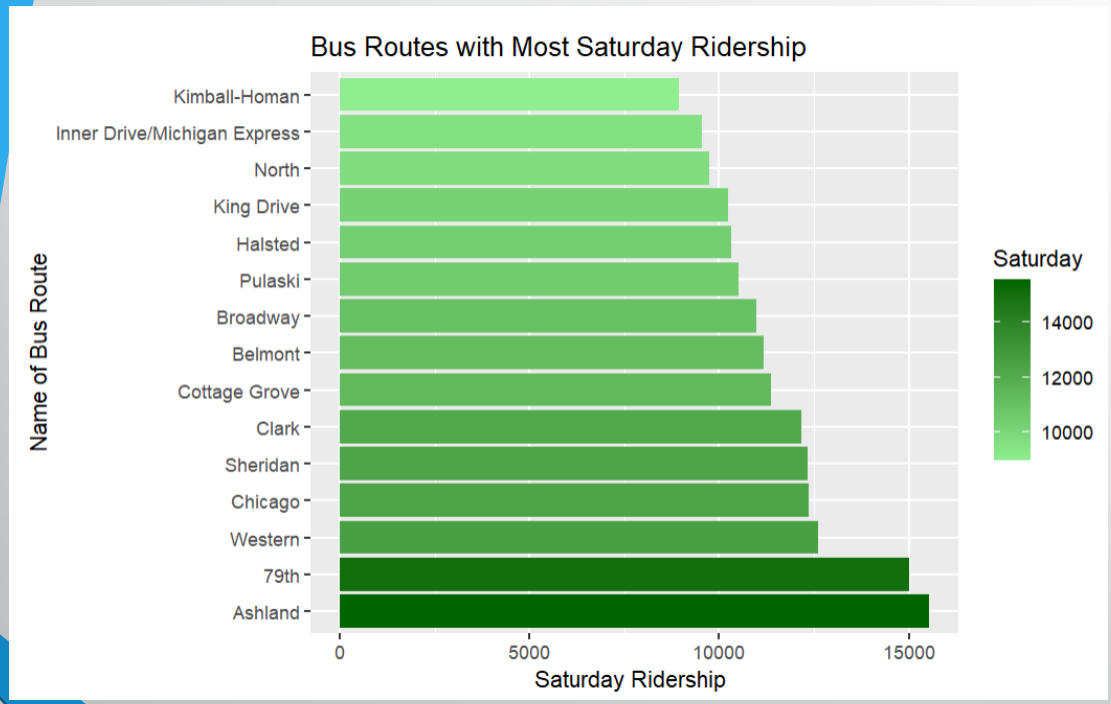
1. **Bus Routes with highest and Lowest ridership on Weekdays**

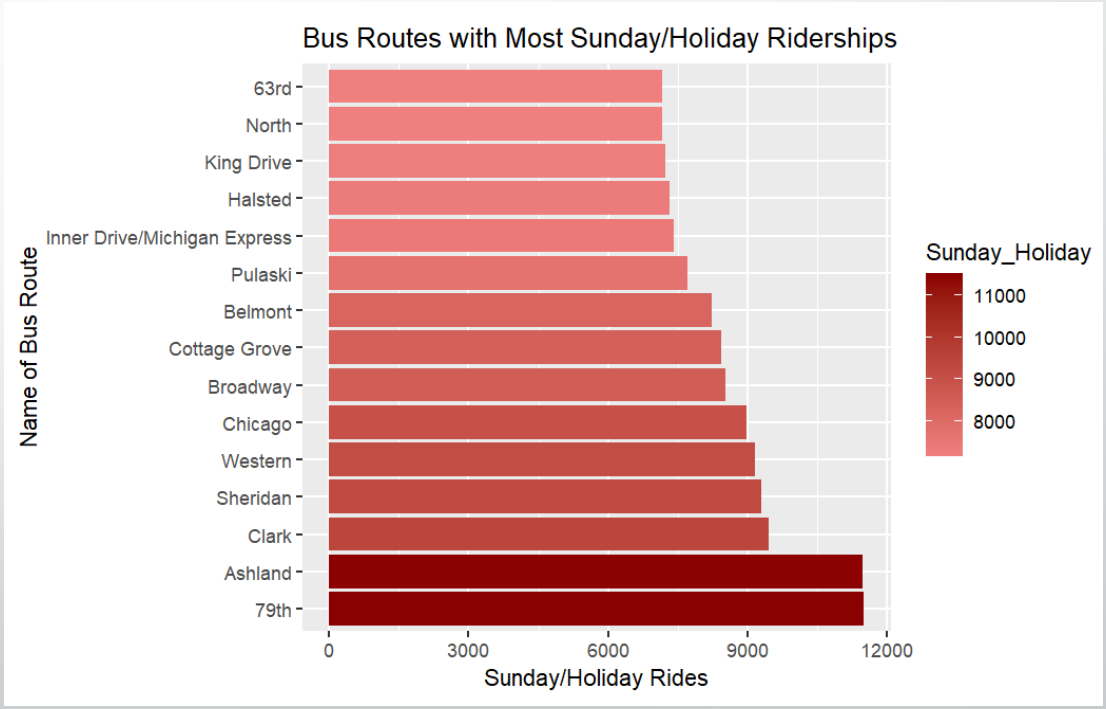




4. EXPLORATORY DATA ANALYSIS

1. Bus Routes with highest Weekend Riders





4. EXPLORATORY DATA ANALYSIS

- **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

```
### 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)
  )

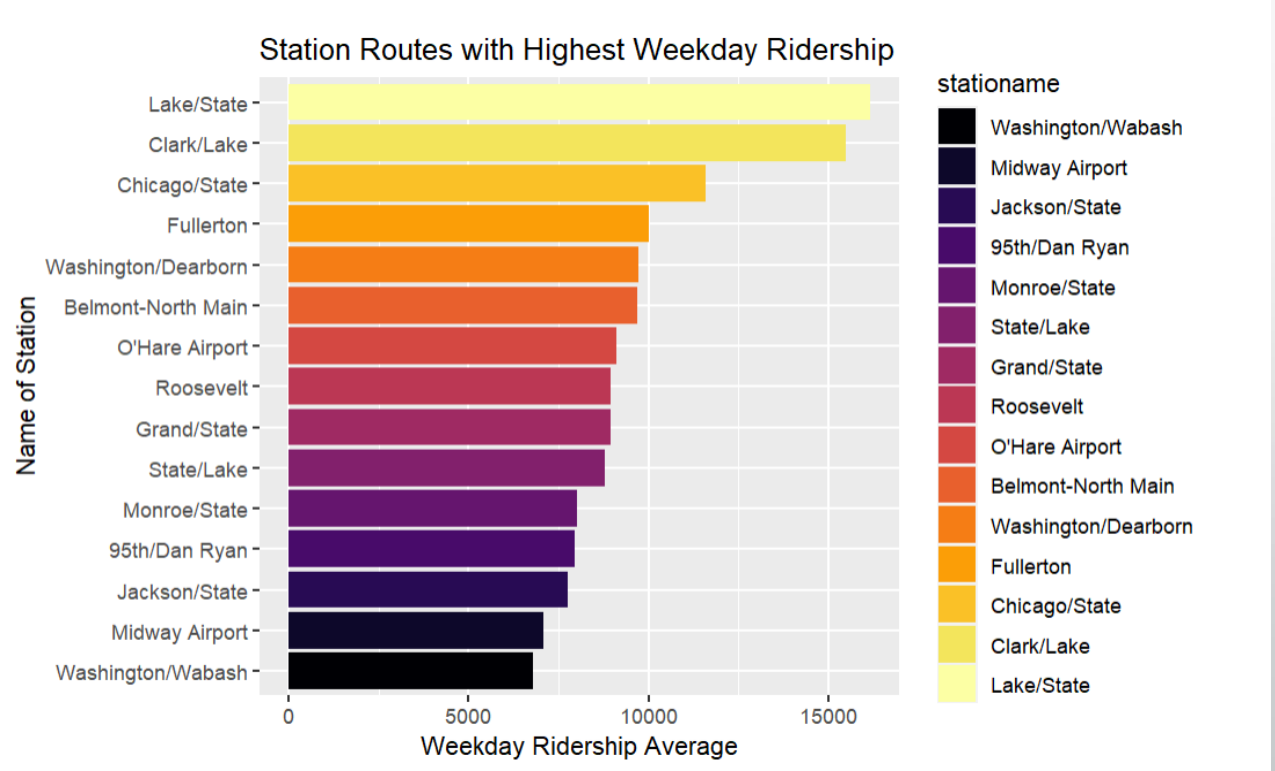
# Print the total traffic for each day type
print(traffic_bus)
```
```

Description: df [1 × 3]

total_weekday_traffic <dbl>	total_saturday_traffic <dbl>	total_sunday_holiday_traffic <dbl>
89321754	55693662	40857403

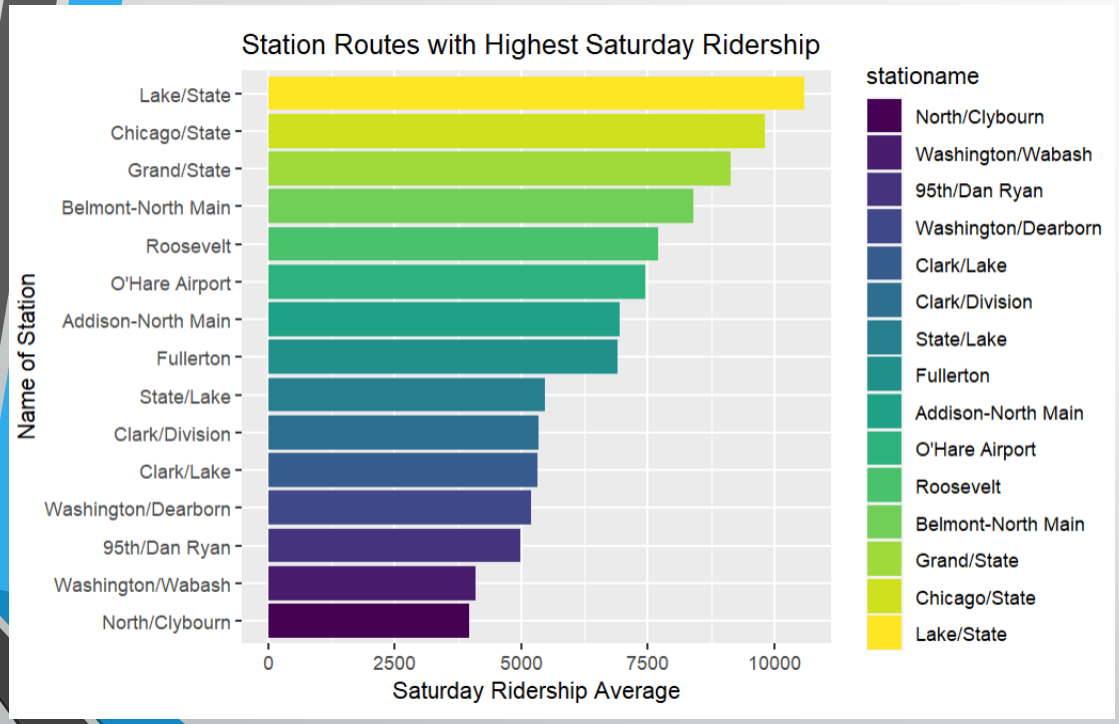
4. EXPLORATORY DATA ANALYSIS

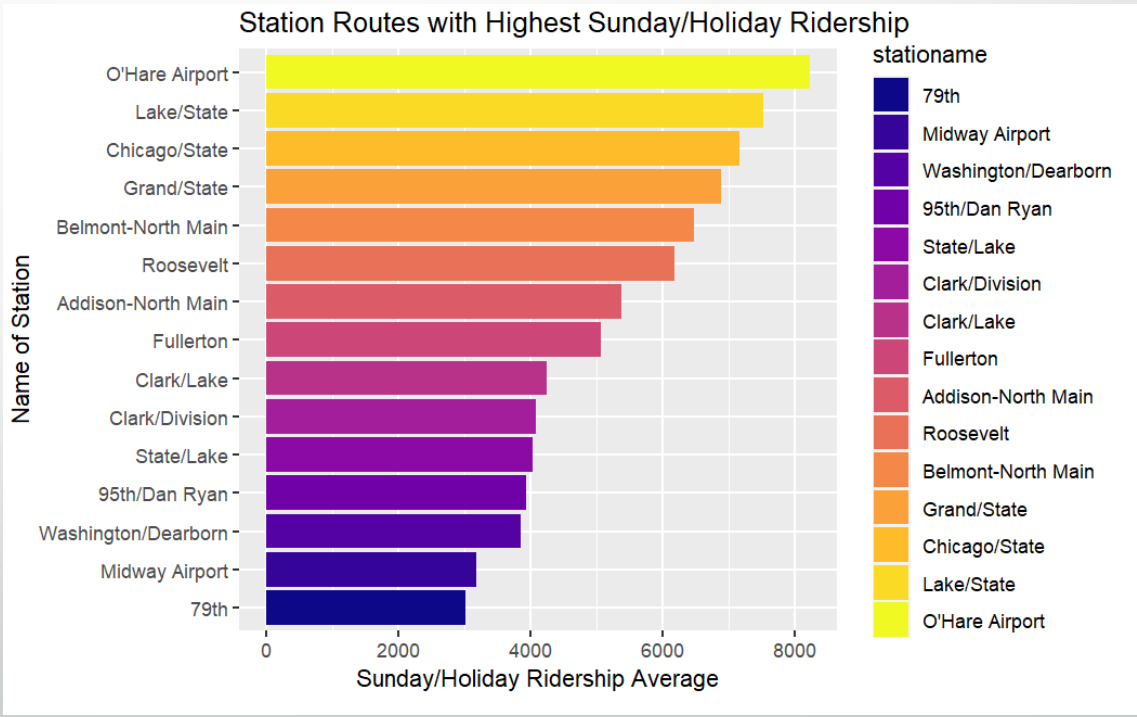
- 2.**Train Routes with highest ridership on Weekdays**



4. EXPLORATORY DATA ANALYSIS

- **2.Train Routes with highest ridership on Weekends**



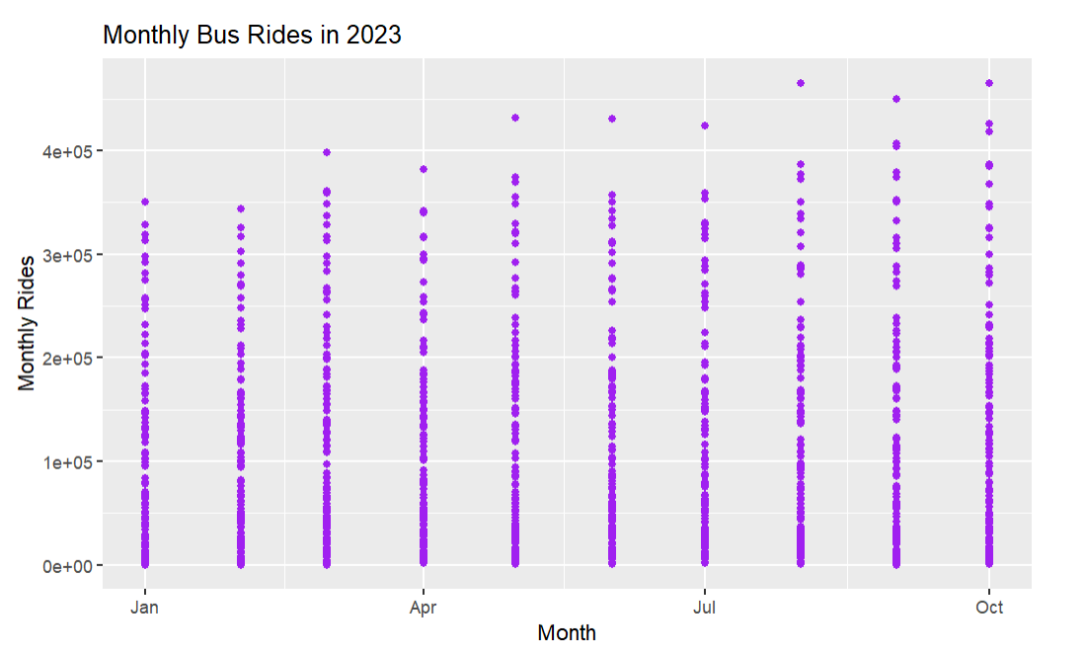


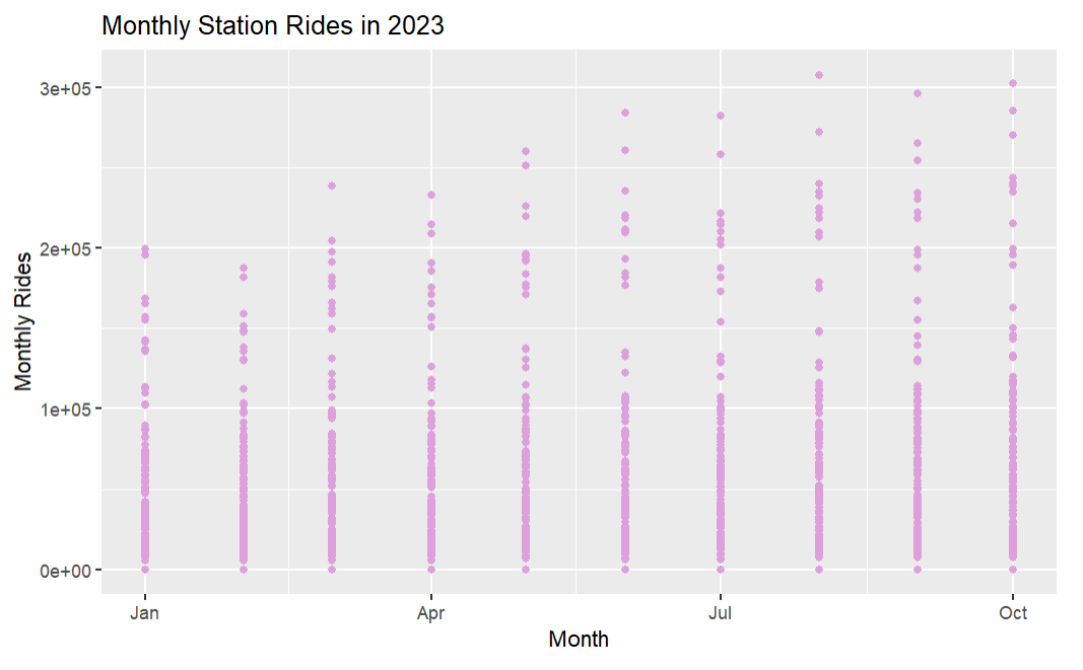
4. EXPLORATORY DATA ANALYSIS

- **2.Train Routes**

  The results –
  The cumulative traffic volume at L-Stations from 2013 to 2023,
  On weekdays, there were approximately 62 million rides,
  On Weekends, Saturdays saw around 38 million rides
  and Sundays or holidays had roughly 29 million rides.

3.Trends in Bus and Train Routes

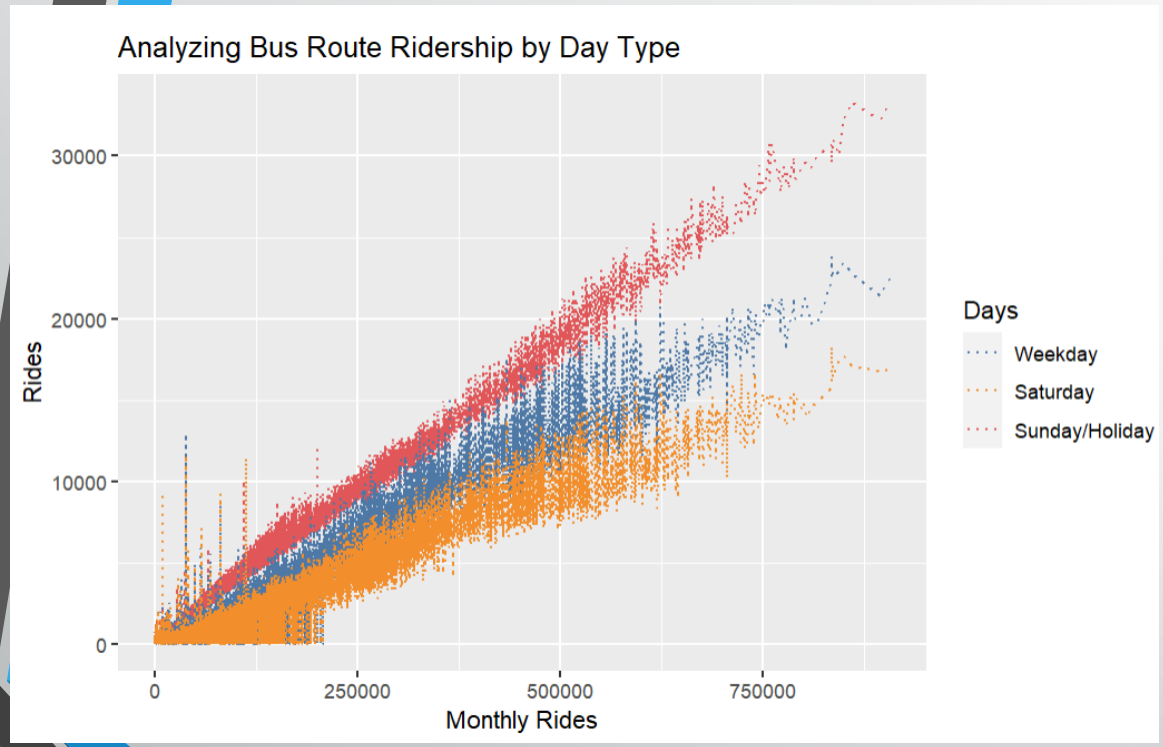


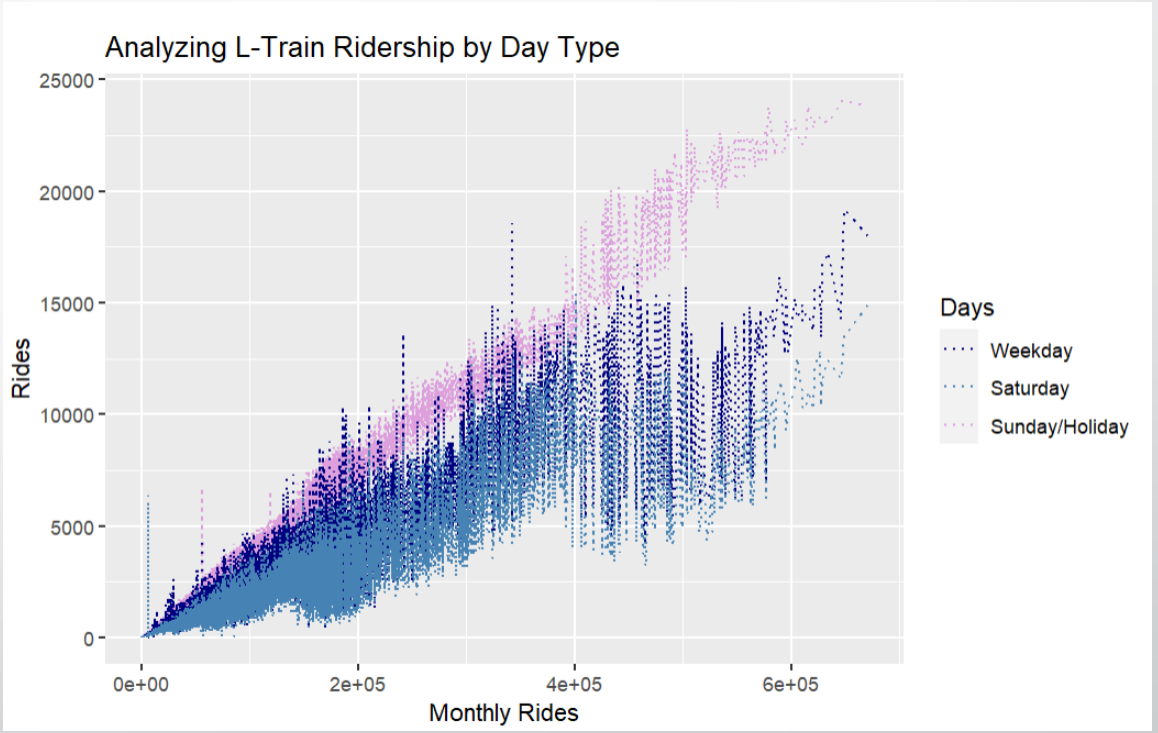


In 2023, August experienced the highest ridership, closely followed by October for both stations and bus routes.
However, our analysis encountered a challenge due to the data ending in October 2023.

EXPLORATORY DATA ANALYSIS

4.Trends in Bus and Train Routes

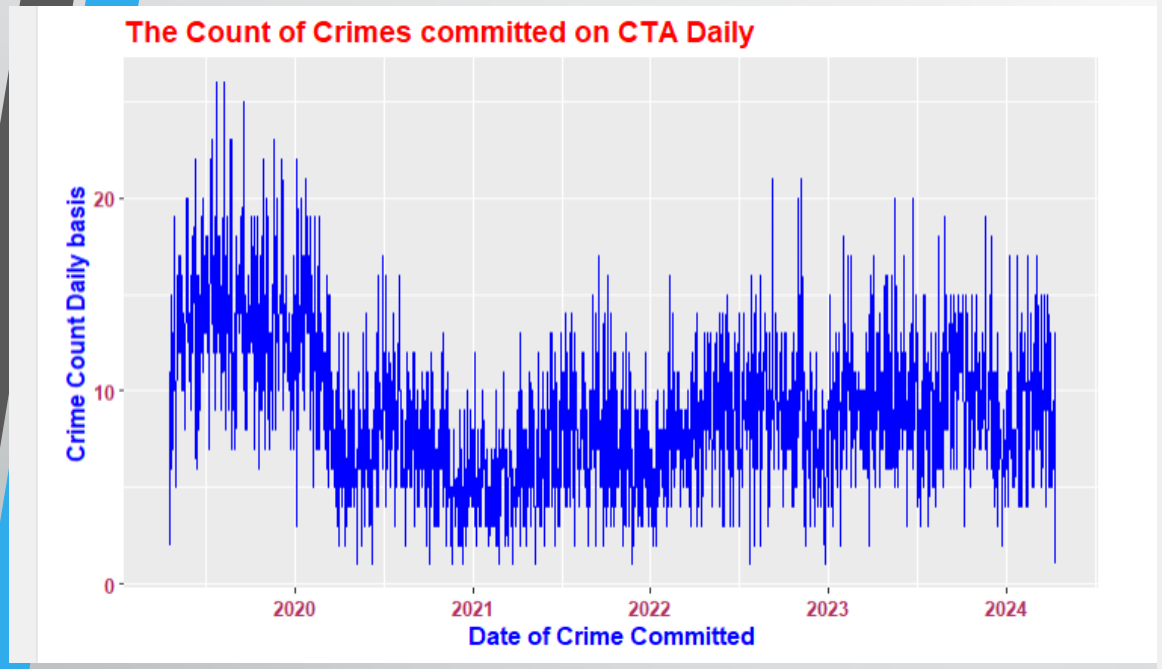


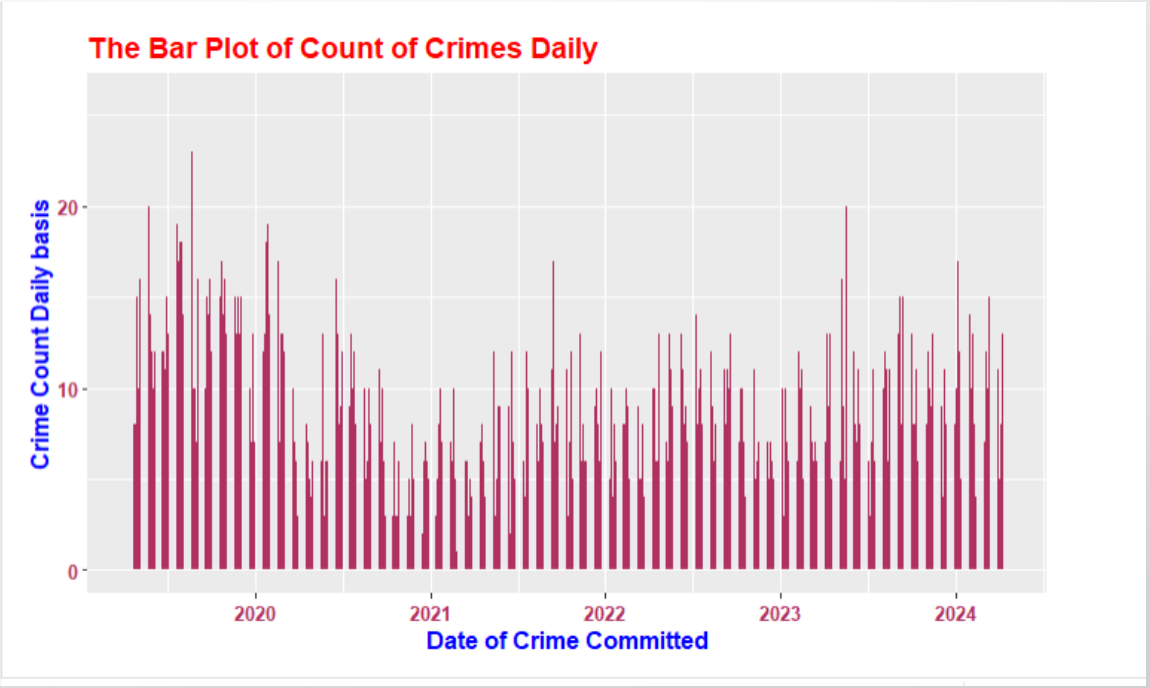


Maximum trips for both bus and station are taken on Weekdays.
Least number of trips are on Sundays/Holidays for both transits.

4. EXPLORATORY DATA ANALYSIS

- **5. Crime Rates**

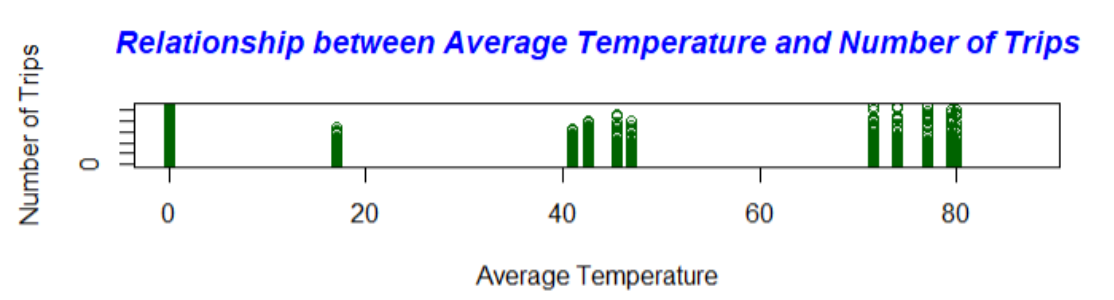


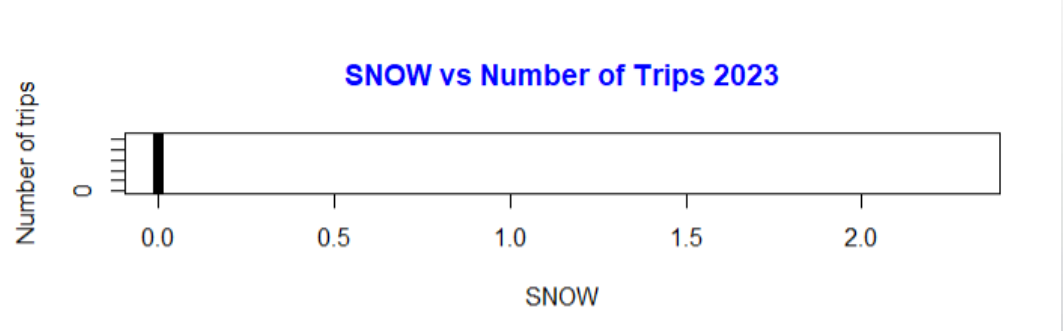


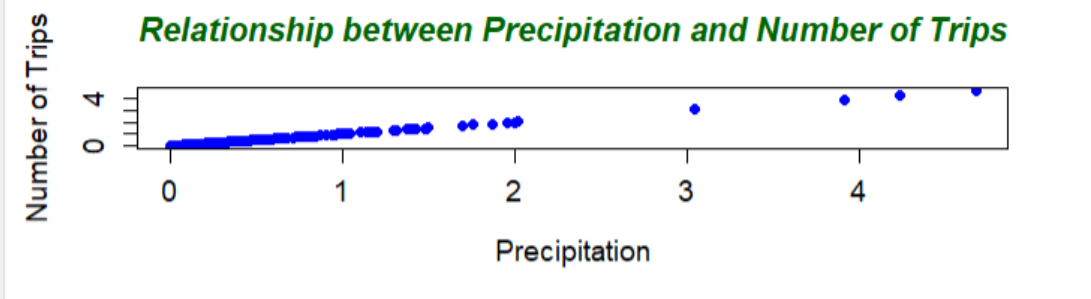
The graphs show the shift in Crime patterns wherein lowest count was in the period range mid-2020 to early 2022 and then a slight increase with occasional spikes in the end-2022 and 2023 period.

4. EXPLORATORY DATA ANALYSIS

- **6. Weather Impact Analysis**







5.DATA MODELLING

MULTIPLE LINEAR REGRESSION:

Results:
When the average temperature was zero,4933.774 trips on transit routes.
A surge in precipitation is led to a decrease of 821.087 trips on transit routes.
Snowfall had no significantly impact transit route trips as indicated by its non-significant coefficient (p-value = 0.93123).
The Model demonstrated statistical significance (F-statistic p-value = 0.01657) , but it has a low R-squared value of 0.001434.

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

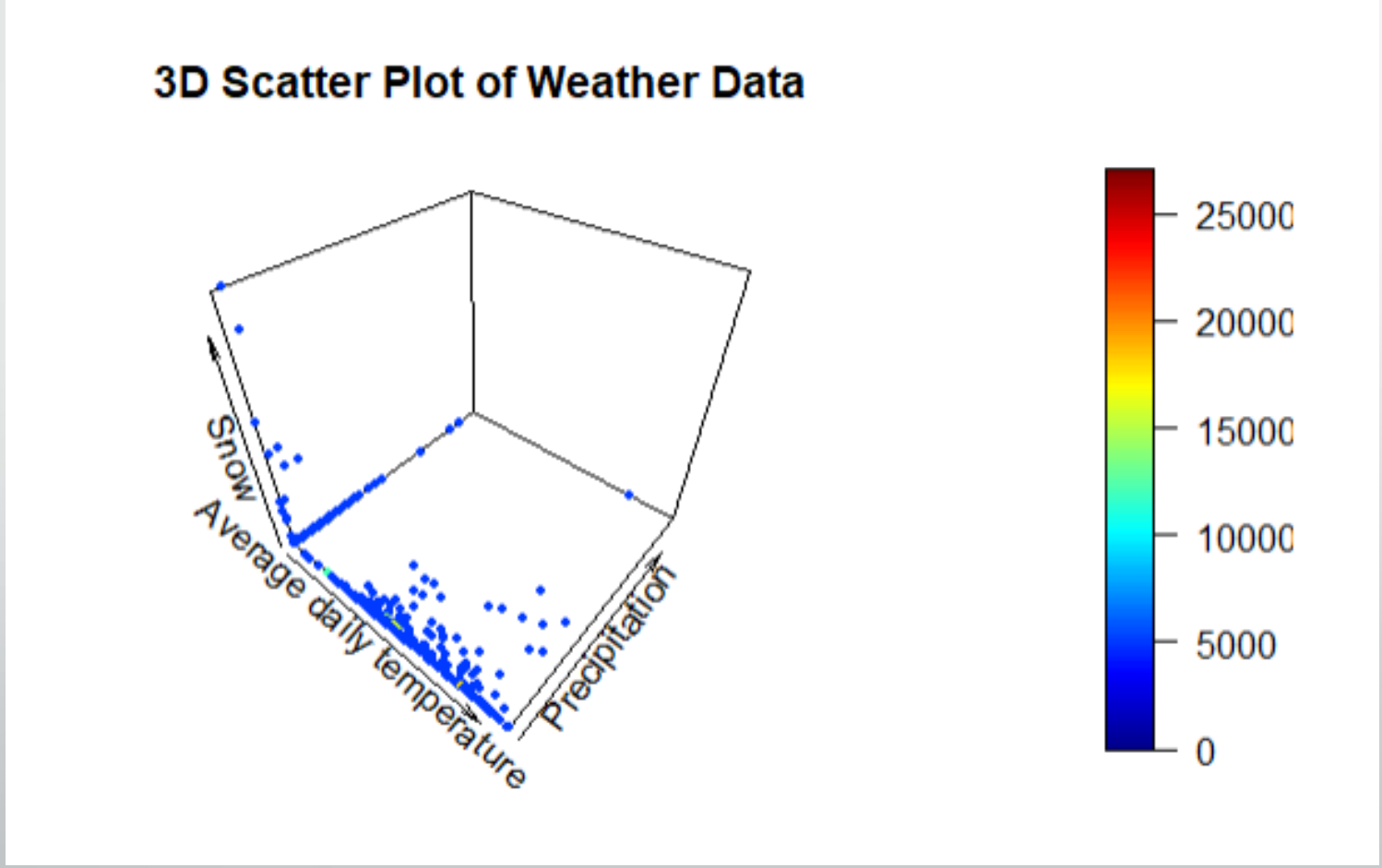
Residuals:
    Min      1Q  Median      3Q     Max
-5180.3 -2898.4  -232.6   666.3 22122.7

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
```

5.DATA MODELLING

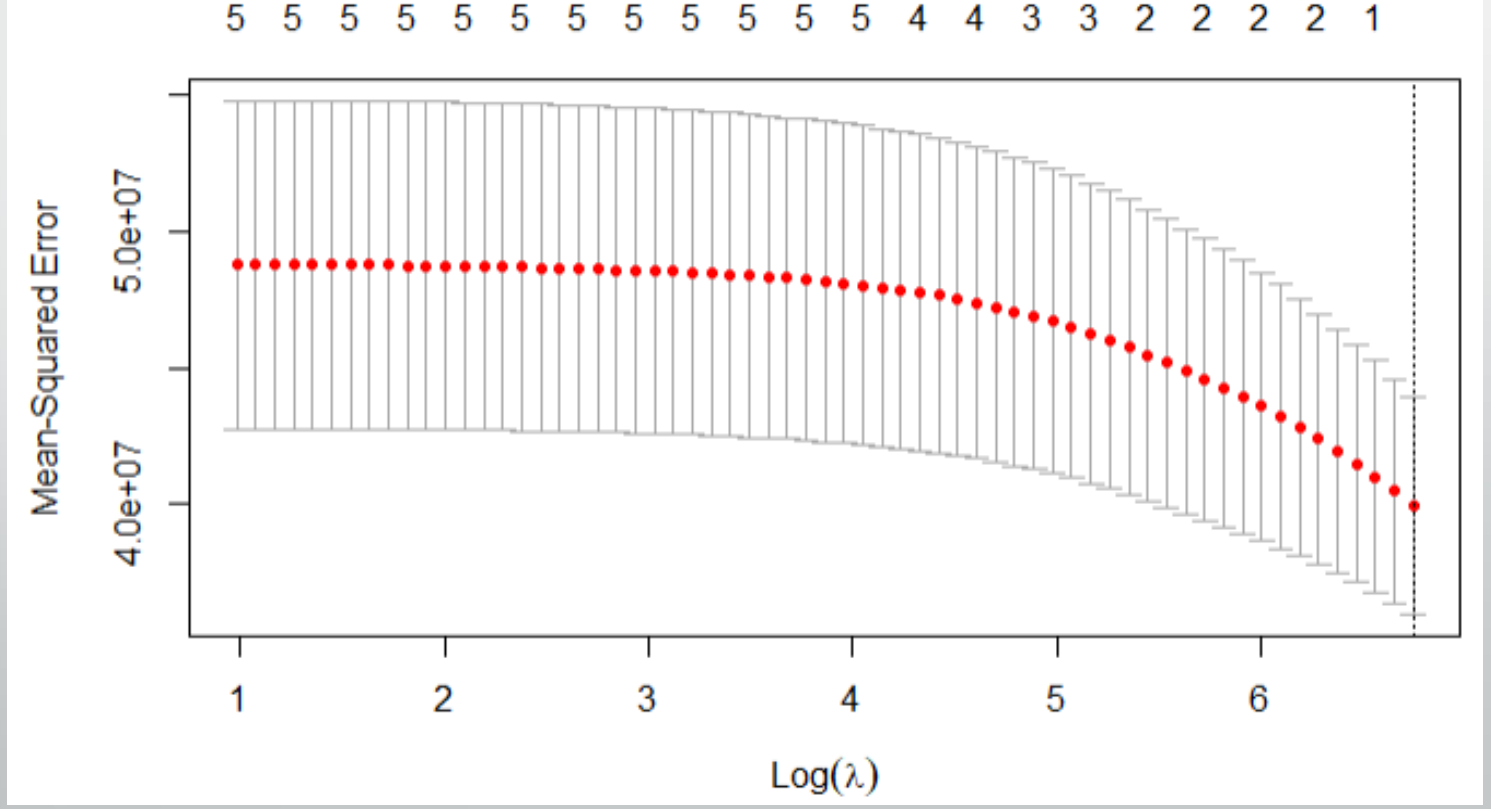
MULTIPLE LINEAR REGRESSION:



5.DATA MODELLING

LASSO REGRESSION:

```
train_idx <- sample(1:nrow(data_imputed), nrow(data_imputed) / 2)
train_data <- data_imputed[train_idx, ]
test_data <- data_imputed[-train_idx, ]
x_train <- model.matrix(no_of_trips ~ ., data = train_data)[, -1]
y_train <- train_data$no_of_trips
lasso_fit <- glmnet(x_train, y_train, alpha = 1)
cv_fit <- cv.glmnet(x_train, y_train, alpha = 1)
```



6.Challenges Faced:

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

8.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.

THANK YOU.