Exploring Performance and Efficiency in the CTA System: A Data Driven Analysis of Routes and Stations

Krishna Chatrathi (A20520106) Ranjitha Aswath (A20526404) Sanketkumar Patel (A20523237)

Abstract:

This project aimed to evaluate the effectiveness of the Chicago Transit Authority (CTA) system by comprehensively analyzing various operational aspects. The objectives were to identify the busiest routes and stations, monitor daily changes in L-line ridership, detect service delays and disruptions, evaluate neighborhood performance, and compare travel wait times with scheduled service times. The study uncovered significant findings, including seasonal ridership patterns, the impact of COVID-19 on ridership, and key insights from geographical data visualizations. Additionally, average travel wait times at specific stations were identified, providing valuable information for improving commuter experiences. The project proposes future steps to enhance the analysis, including extending geo-spatial analysis to bus routes and utilizing real-time data through CTA's Application Programming Interface (API). By combining historical trends and advanced analyses, this research aims to provide insightful information, ultimately contributing to the overall efficiency and effectiveness of the CTA system and enhancing public transportation experiences for Chicago residents and visitors.

Overview:

The main objective of this project was to assess the effectiveness of the Chicago Transit Authority (CTA) system by conducting a comprehensive analysis of various operational aspects. We aimed to answer specific questions, including identifying the busiest routes and stations, monitoring daily ridership changes for Red-Line, detecting service delays and disruptions, evaluating neighborhood performance, and comparing travel wait times with scheduled service times.

The study yielded significant findings. We observed a clear ridership pattern, noting increased ridership during summer and decreased ridership during winter. Moreover, the impact of the COVID-19 pandemic on ridership was carefully examined, providing valuable before-and-after comparisons to understand the system's resilience during challenging times. Geographical data visualization, with the aid of packages in R Studio, allowed us to create heat maps showcasing the busiest stations for each CTA line, aiding in route optimization and resource allocation decisions.

Additionally, we identified the average travel wait times at specific station throughout the week. This crucial information shed light on commuter experiences, which could serve as a basis for making improvements to enhance passenger satisfaction.

To enhance the analysis further, the study outlines future steps. We plan to extend the geo-spatial analysis to encompass bus routes, which would offer a comprehensive understanding of the entire public transit network. By incorporating real-time data through CTA's Application Programming Interface (API), the project aims to provide timely insights into service delays and disruptions at specific routes and stations. This approach would enable comparisons of performance across different areas within the city, assisting in identifying areas for targeted improvements.

By combining historical ridership trends, and advanced geo-spatial analyses, this project aims to provide valuable insight. Ultimately, the research contributes to enhancing the efficiency and effectiveness of the CTA system, leading to improved public transportation experiences for both the residents and visitors of Chicago.

Project Motivation and Problem Statement:

The motivation behind this project is to drive positive changes in the CTA system, benefitting commuters, the city, and the environment by providing data-driven insights that aid in improving the overall effectiveness and efficiency of the public transportation network. A few featured problem statements are as follows:

- Which routes and stations are busiest in the CTA system?
- What is the change in ridership for a bus route daily in a week?
- How did covid effect the ridership of CTA system?
- For routes and stations are there any service delays and disruptions?
- Are there areas of cities with better performance than others?
- What is the travel wait times, and how do these compare to the scheduled service times?

Proposed Methodology and Approach:

For this project, the primary objective was to evaluate the effectiveness of the Chicago Transit Authority (CTA) system, focusing on several key aspects of its operations.

Data Gathering:

The data gathering phase involved downloading relevant data from the CTA website, including station information and ridership data spanning from January 2001 to January 2023, for both weekdays and weekends. Additionally, shape files were obtained from the Chicago Data Portal to facilitate the visualization of stations and routes across the entire city. To complement historical data, we requested access to real-time data from the Transit App, which was granted later on as an API, which we wish to implement in our future. To gain an understanding of commuters' expectations, a Google survey was distributed to fellow residents, seeking insights into their wait time preferences during different times of the day.

Data preprocessing and Pipeline details:

This step was a crucial step in preparing the datasets for analysis. We meticulously ensured data consistency by appropriately assigning data types to each data point. To handle any missing values, imputation techniques were employed using statistical measures like mean, median, or mode, where applicable. Invalid entries were addressed, and redundant or unnecessary columns were removed to streamline the datasets. Outlier detection techniques were also applied to refine the data further.

Data Cleaning: Issues and Adjustments:

Data cleaning was performed with a focus on geospatial analysis of L-stations. We removed irrelevant columns from the available L-stations dataset and filtered out a specific rail line (e.g., red line) to concentrate on detailed analysis of that route. Location coordinates (x, y) were separated into distinct columns, X and Y, and given numeric data types to facilitate plotting on maps. Redundant location data was eliminated. The ridership data for weekdays, Saturdays, and Sundays were merged with the L-stations data using common variables like Map_ID and station_id. Further unnecessary columns were removed, and data types were adjusted as necessary for analysis. This process was repeated for other relevant datasets as well, ensuring a consistent and standardized preprocessing approach.

Data Analysis: Visualization:

Figure 1 illustrates an interactive map showcasing all L-stations in the Chicago Transit Authority (CTA) system. The map was generated using the Leaflet package in R, offering an engaging and user-friendly visualization of the station locations across the city.

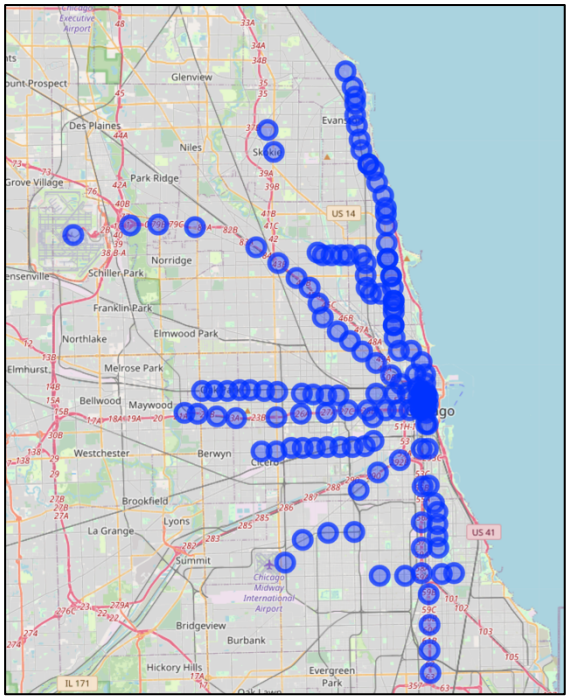


Figure 1 All CTA L-stations plotted on Map.

To enhance the map's insight, the busyness of each station was incorporated as an overlay. The busyness information provides a visual representation of passenger density and activity at each station. The intensity of the busyness overlay varies across the map, with darker regions indicating higher passenger volumes and greater activity.

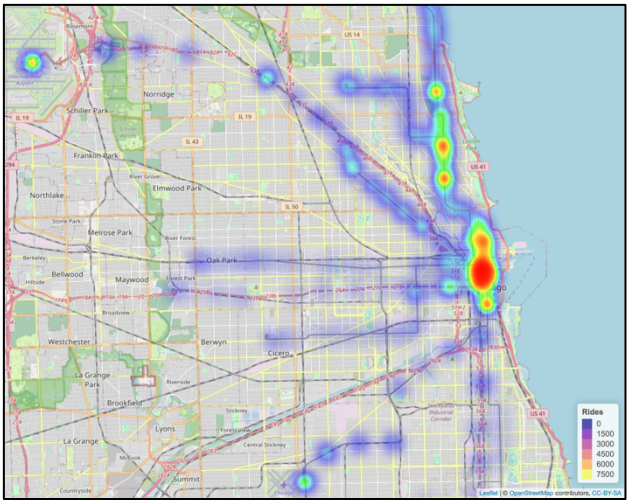
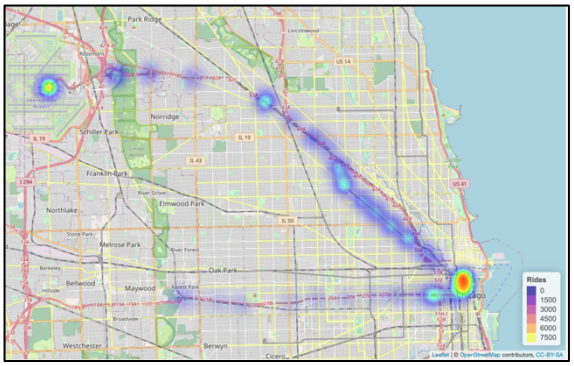
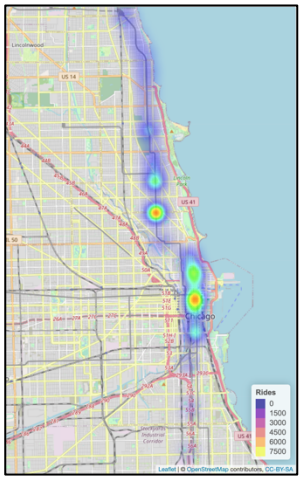


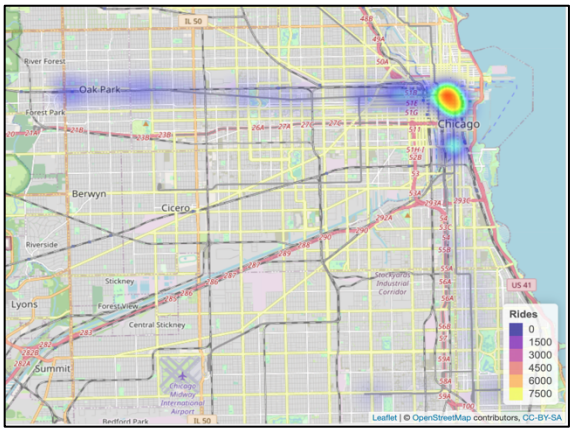
Figure 2 Heatmap showing Busyness of L-station.

Based on the analysis of the below images, the following conclusions can be drawn:

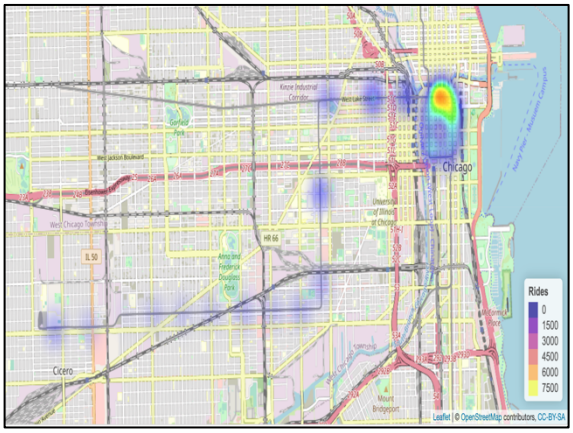
1. Stations situated in the downtown Loop area, as well as O'Hare Airport and Midway Airport, exhibit significantly higher levels of activity compared to other stations. These locations serve as major transit hubs with substantial passenger traffic.
2. By examining the respective images for each rail line, we can easily identify the busiest station along each route. The visualization allows us to discern the stations with the highest passenger volume on each line, providing valuable insights into key transit points.

Blue Line	Red Line
	

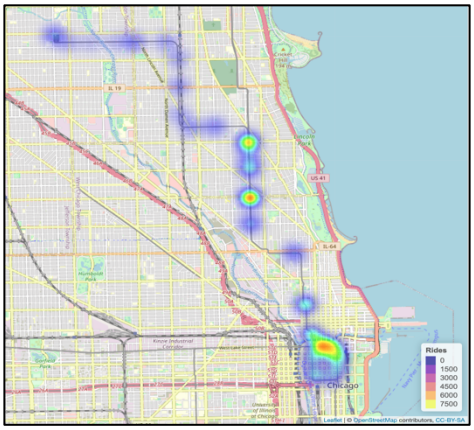
Green Line



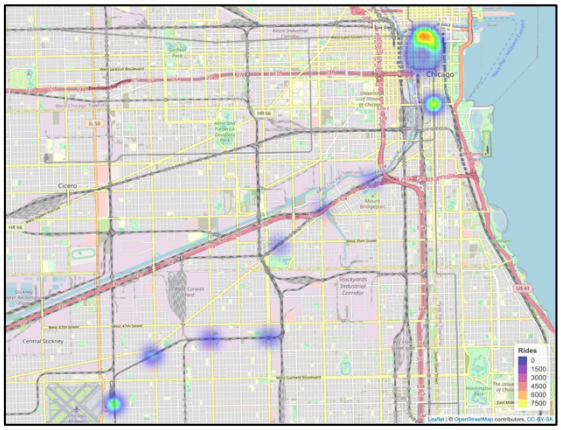
Pink Line



Brown Line



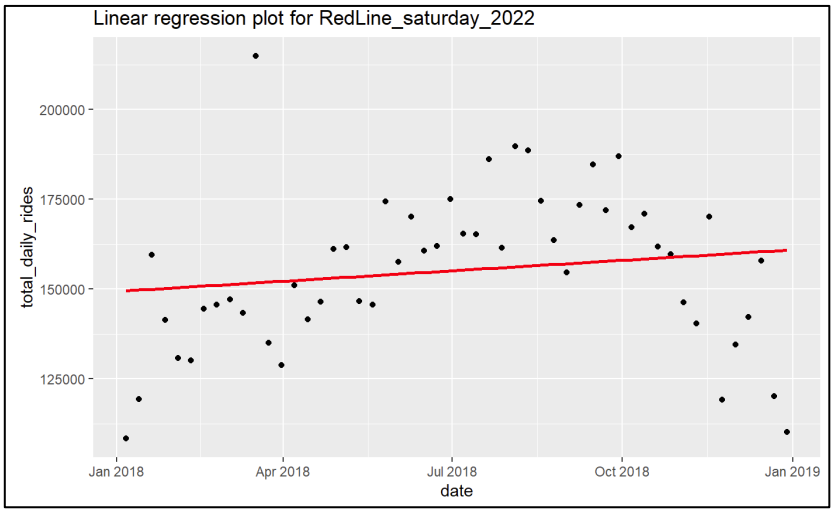
Orange Line



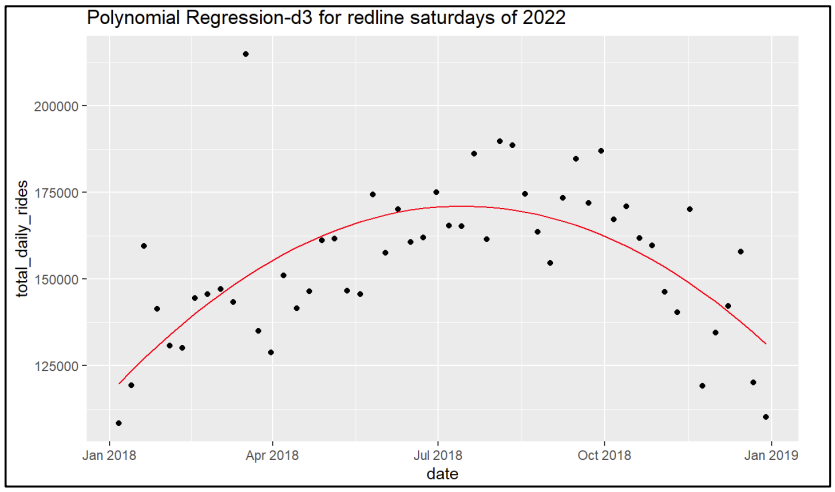
Model Selection and Validation:

Our focus is on a quantitative response variable, with single predictor and single dependent variable. For this kind of data simple linear regression and polynomial regressions methods can be used. These models enable us to evaluate on which days have higher ridership.

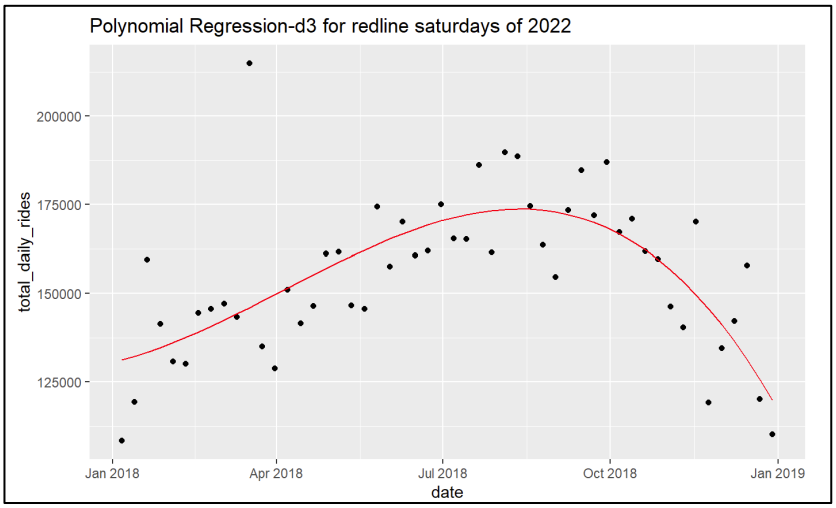
We have selected all Saturdays of 2022 of Red-Line ridership to select the regression models. We have observed that for simple Linear regression and polynomial regression with degree 2 and 3 and we have selected the model Polynomial Regression degree 3 because it had the higher accuracy in all three models and the increase in accuracy from degree 3 to 4 is not significantly more. Since all our datasets have similar trend, we use model polynomial degree 3 for them.



(a)



(b)



(c)

Figure 3 (a),(b) & (c) fitting polynomial regression of Saturdays average ridership of 2022 Red Line

Overall, the quantitative-based techniques we use give us a solid way to examine the data and determine which model performs the best. With the help of this method, we can better understand the connection between the predictors and the response variable and improve the performance of our models.

Table 1 Model's performance on testing set

Model	Accuracy (Rsquared)	RSME
Simple Linear Regression	0.1031	18809
Polynomial D-2	0.729	10333
Polynomial D-3	0.745	10021

Results:

- Based on heatmap, we can visualize the ridership on particular Rail line stations and can make out the mentioned conclusions like the ridership is more in loop and airport terminals.
- Overall, CTA's L is contributing to the city's economy by providing cost-effective and environment friendly alternate mode of transportation for people every year.

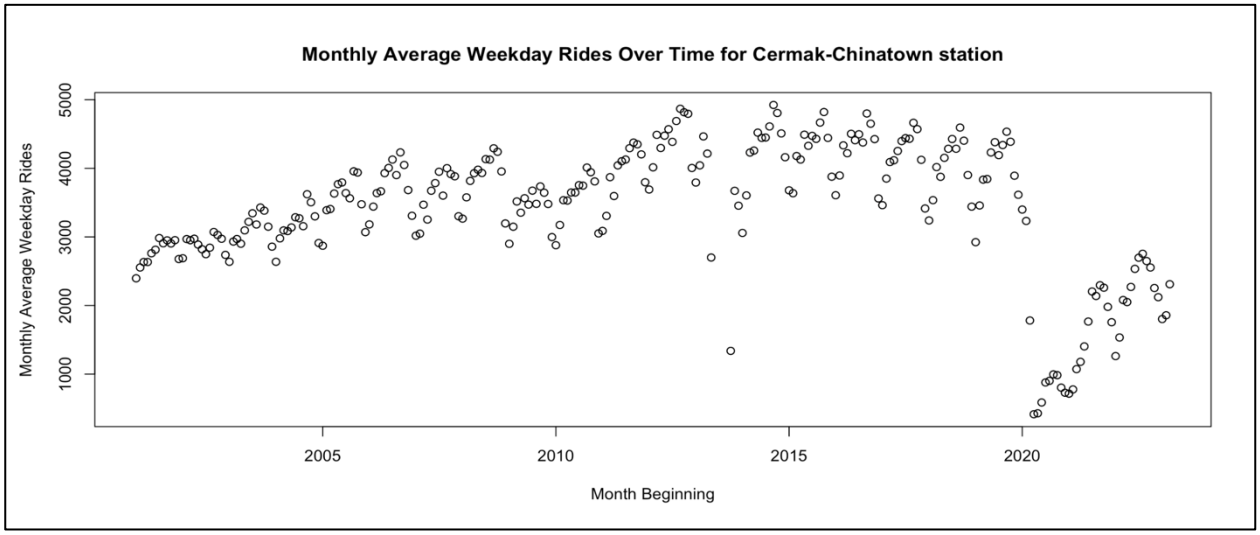


Figure 4 Monthly Average Weekday Rides over time for a particular station

- We can see that from above figure, CTA observed a significant drop in ridership in every station (example, Cermak-Chinatown station) during COVID pandemic which the system has not yet recovered to the pre-pandemic levels.

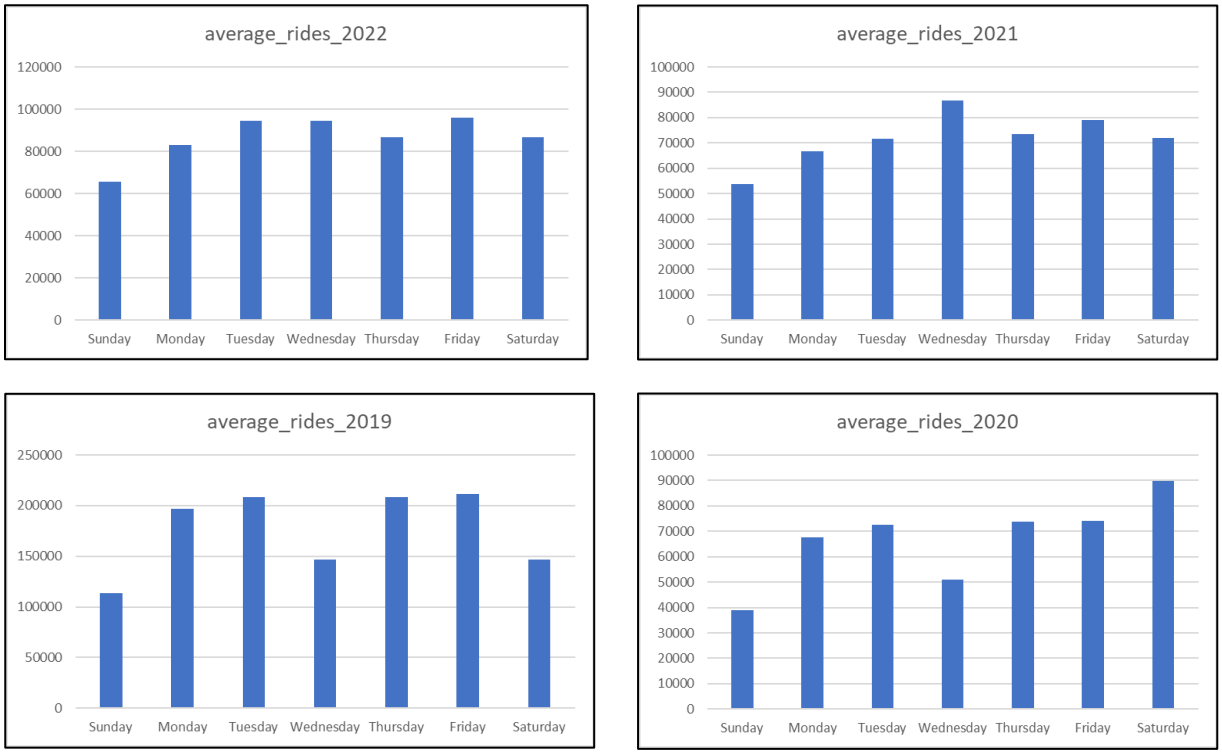


Figure 5 Average Rides vs day type, for a specific year for Red line

- We observe that CTA observed a significant drop in Ridership in 2020 due to COVID.
- Post Covid we can observe that there is a recovery in ridership but not to pre-pandemic levels.

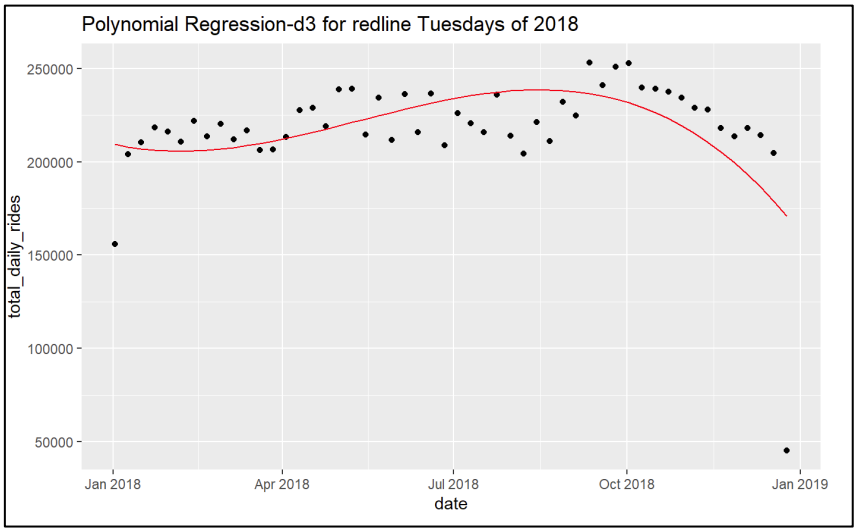


Figure 6 Fitting polynomial regression on ridership

- We see that ridership increases during summer and decreases during winter.

Conclusion:

1. High Ridership in Loop and Airport Terminals: The heatmap visualization indicates that the rail line stations located in the Loop area and near the airport terminals experience higher ridership compared to other stations. This suggests that these stations are crucial transit hubs and play a significant role in the city's transportation system.

2. Contribution to the City's Economy: CTA's L system contributes significantly to the city's economy by providing a cost-effective and environmentally friendly alternative mode of transportation for people. It serves as a vital link in the public transportation network, facilitating the movement of residents and visitors throughout the city.

3. Impact of COVID-19 on Ridership: The analysis shows that CTA experienced a substantial drop in ridership during the COVID-19 pandemic, with many stations, including the Cermak-Chinatown station, witnessing a significant decline in passengers. As of the time of the analysis, the ridership had not fully recovered to pre-pandemic levels, indicating the lingering effects of the pandemic on public transportation.

4. Post-COVID Recovery: There is evidence of a gradual recovery in ridership post-COVID; however, the system has not yet reached its pre-pandemic levels. This indicates that while people are starting to use public transportation more as restrictions ease, it may take some time to fully regain the pre-pandemic ridership levels.

5. Seasonal Variation: The analysis also reveals a seasonal variation in ridership, with increased ridership observed during the summer months and decreased ridership during winter. This pattern is common in many public transportation systems, as weather conditions and seasonal activities can influence people's travel patterns.

Data Source:

Google Drive Link - For Data Source

Source Code:

Google Drive Link - For Source Code

References:

a. https://github.com/sabrinadchan/ctabus
b. *Online documents by category/date*. (n.d.). CTA. Retrieved 11 June 2023, from https://www.transitchicago.com/documents/
c. *Ridership reports—Performance*. (n.d.). CTA. Retrieved 11 June 2023, from https://www.transitchicago.com/ridership/
d. *When things go wrong (How we mitigate delays)*. (n.d.). CTA. Retrieved 11 June 2023, from https://www.transitchicago.com/performance/wtgw/