Data Wrangling Project
Public Transportation Dispatch Optimization

1. Introduction

As a key provider of public transportation in Chicago, the CTA plays an important role in the lives of both residents and visitors. The CTA trip database, provided by the Chicago government, offers valuable insights that can help city planners design an effective transportation system and optimize bus and train schedules for riders. By analyzing this data, we can gain a better understanding of the CTA's audience and identify the best times to use the CTA.

In this project, we will utilize the CTA travel data from the Chicago Government Database Portal to study the impact of weather and holidays on ridership. This analysis will allow us to provide recommendations for the CTA to improve costs, service quality, and transit capacity. Overall, our goal is to help the Chicago government enhance the CTA's services and provide an even better experience for riders.

2. Data

This project has three primary sources of data, Chicago Data on CTA Ridership, Chicago Data on Chicago Beach Front Weather, and Kaggle on US Holiday from 2004 to 2021.

2.1 CTA Ridership

The "CTA - Ridership - Total Daily Trips" data provided by the Chicago government database is the main source of information for this project. This dataset records daily trip data for various modes of transportation in Chicago, covering a period from 2001 to 2022. The dataset consists of 7944 observations and includes information on boarding numbers for buses and trains. This comprehensive data allows for a thorough analysis of transportation trends in Chicago over the past two decades. The Chicago government places CTA data in one entity. It is available to users as a CSV and an API. In this project we will read this data as a CSV in our R script.

2.2 Chicago Beach Weather

The beach weather dataset provided by the Chicago government database includes 147194 observations recorded from various observation stations. The data covers a period from May 2015 to October 2022, and includes a range of weather data such as wind speed, temperature, and total rainfall. In this analysis, we will focus on using the wind speed, temperature, and total rainfall data in combination with the ridership dataset to gain insights into the impact of weather on public transportation usage in Chicago. By merging these datasets, we can better understand how factors such as wind speed, temperature, and rainfall affect ridership trends

and make informed recommendations for the CTA. The Chicago government places weather data in one entity. It is available to users as a CSV and an API. In this project we will read this data as a CSV in our R script.

In our project, we aim to calculate the average data from various weather stations on a daily basis. This will reduce the number of observations in our dataset to 2625. Additionally, we plan to create a new feature based on the existing rain and wind speed data. We will categorize heavy rain, drizzle, calm, and breeze into more easily readable categories. This will provide a more intuitive and user-friendly representation of the data.

2.3 US Holiday Data

Kaggle maintains a database of US holiday data ranging from 2004 to 2021 in the form of a downloadable CSV file. However, this dataset is limited in scope and only covers a specific time period. In order to extend the dataset and include holidays from 2022, we conducted research on the US federal government website and manually created a new CSV file. We then utilized the R programming language to merge the two datasets and create a single, comprehensive dataset of US holiday data from 2004 to 2022.

2.4 Merging and cleaning

In order to merge the three distinct datasets that we possess, we will utilize the date function in the R programming language. We will first ensure that the dates in each dataset are formatted in the same manner. After this step, we will proceed to combine the three tables into a single dataset. However, there may be instances where certain dates do not have any corresponding data. In such cases, we will utilize the average value of the previous five days in order to impute the missing data. This will allow us to create a comprehensive and cohesive dataset.

2.5 Links

The lead links for the download and API are:
https://data.cityofchicago.org/Transportation/CTA-Ridership-Daily-Boarding-Totals/6iiy-9s97
https://data.cityofchicago.org/Parks-Recreation/Beach-Weather-Stations-Automated-Sensors/k7hf-8y75
https://www.kaggle.com/datasets/donnetew/us-holiday-dates-2004-2021

Column Name	Description	Type
date	Format in YYYY-MM-DD	date
day	W = weekday, A = Saturday, U = Sunday/Holiday	character
bus	Total bus ride in the day	integer
rail_boardings	Total railway boardings	integer
total_rides	Total rides in CTA from the day	integer
holiday	Federal Holiday, NA for normal days	character

temperature	Average temperature of the day in celsius	numeric
rain	Total rain of the day in millimeter	numeric
wind_speed	Average wind speed of the day in meter	numeric
rain_type	Calculated based on total rain of the day. Less than 10 rain is clear, between 10 and 40 is drizzle, between 40 and 80 is rain, more than 80 is heavy rain.	character
wind_type	Calculated based on average wind speed. Less than 10 is calm, between 10 and 20 is light breeze, between 20 and 30 is breeze, between 30 and 40 is strong breeze, between 40 and 50 is gale, more than 50 is storm	character

3. Analysis

The main objective of this project is to optimize the CTA experience through improved scheduling and a deeper understanding of CTA riders. By analyzing the CTA trip data, we aim to identify the impact of weather and holidays on ridership, as well as the potential impact of the COVID-19 pandemic. Additionally, we will examine the best times for travel using the CTA and explore how the weather may influence the mode of transportation chosen by riders. Through this analysis, we hope to provide recommendations for the CTA to enhance its services and improve the experience for riders.

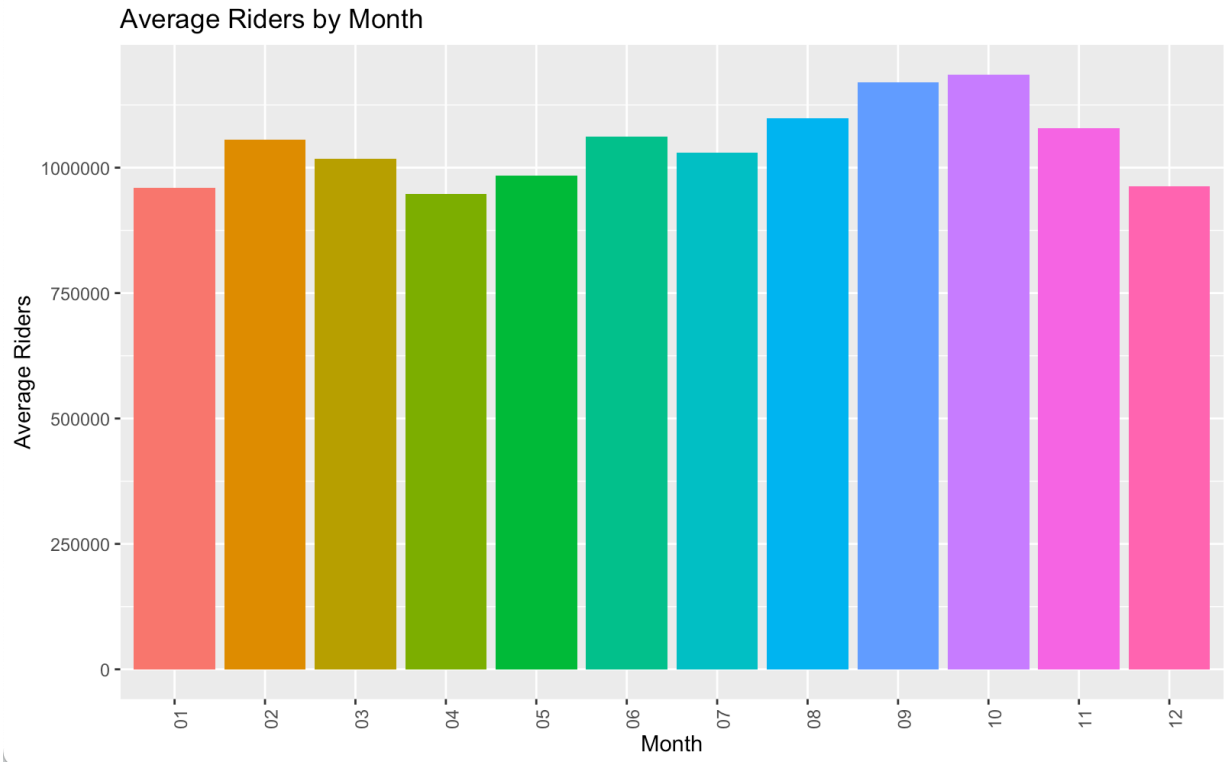
The main objectives of this project are to

1. Determine the impact of the COVID-19 pandemic on CTA ridership.
2. Explore the effects of weather and holiday data on CTA ridership.
3. Identify the best times to travel using the CTA.
4. Determine the types of days on which the CTA is most heavily used.

Additionally, we will also seek to

1. Understand the demographics of CTA riders and how they may impact ridership trends.
2. Analyze the impact of different modes of transportation (e.g. buses vs. trains) on ridership.
3. Identify potential recommendations for the CTA to improve its services and provide a better experience for riders.
4. Investigate the relationship between CTA ridership and other factors, such as traffic congestion and economic conditions.

3.1 Average Ridership by Month



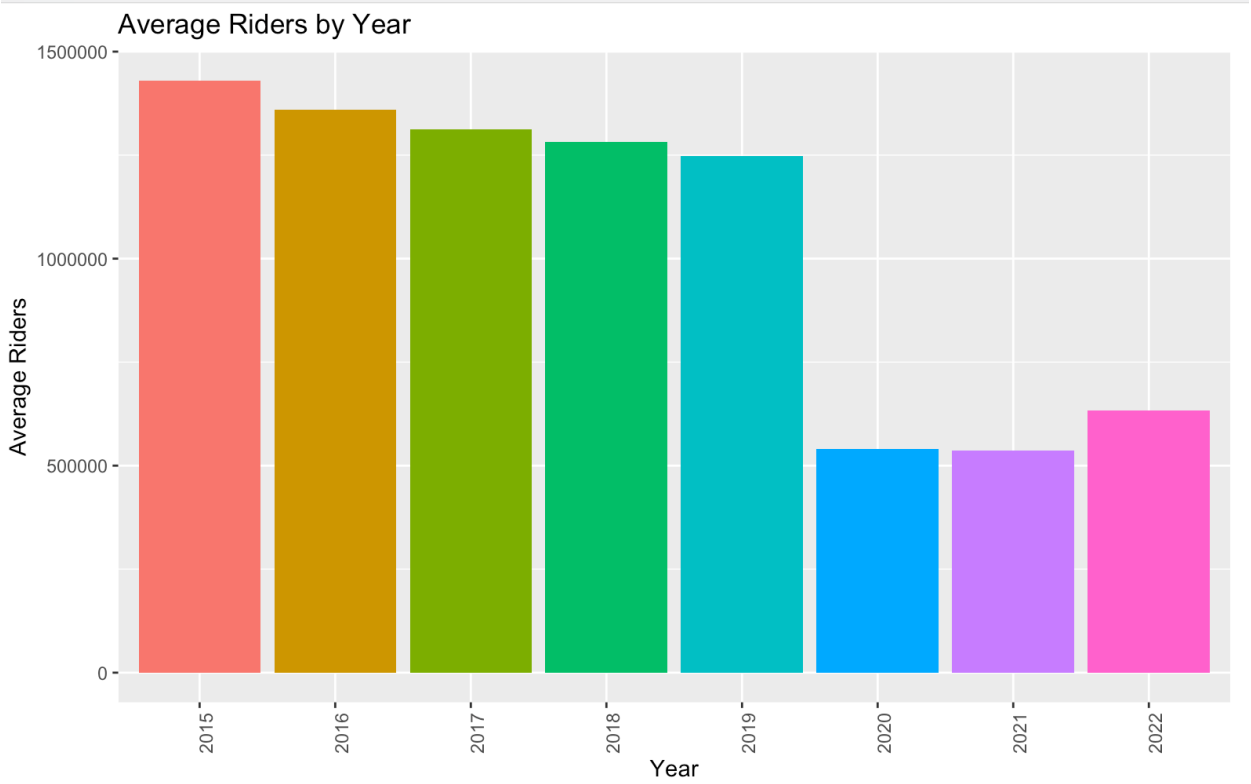
The provided data indicates that the average monthly ridership of the Chicago Transit Authority (CTA) varies over the course of the year. The busiest months are typically in the summer, with the highest ridership occurring in June, July, and August. The least busy months are typically in the winter, with the lowest ridership occurring in December, January, and February.

One potential explanation for this pattern is that the weather influences CTA ridership. During the summer months, when the weather is warmer, people may be more likely to use the CTA for outdoor activities and leisure. In the winter months, when the weather is colder and more unpredictable, people may be more inclined to use other modes of transportation or to stay indoors.

Another factor that may affect CTA ridership is the impact of holidays. The summer months often coincide with popular vacation times, when people may be more likely to use the CTA to travel within the city or to access tourist attractions. In the winter, holidays such as Christmas and New Year's Day may result in lower ridership due to changes in people's travel patterns.

In summary, the CTA appears to experience fluctuations in ridership over the course of the year, with the busiest months occurring during the summer and the least busy months occurring during the winter.

3.2 Average Ridership by Year



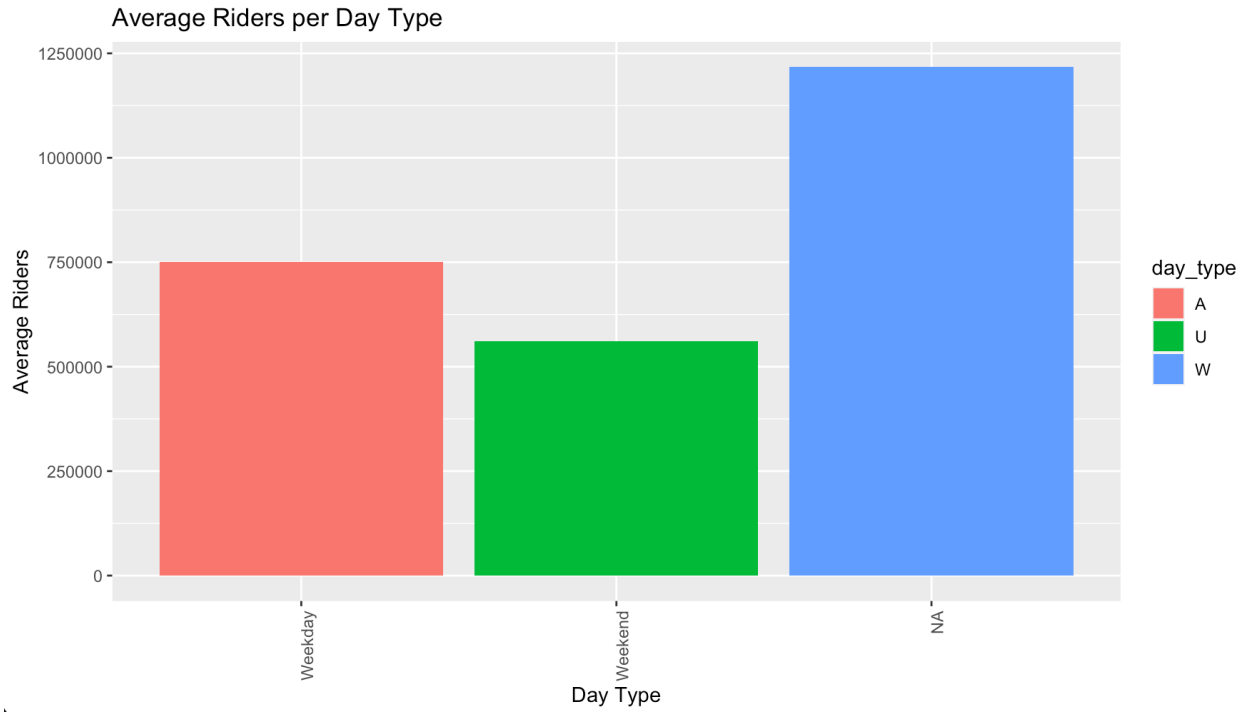
The provided data shows that the average annual ridership of the Chicago Transit Authority (CTA) has declined significantly since the onset of the COVID-19 pandemic. Prior to the pandemic, the ridership was relatively stable, but in 2020, there was a sharp decrease in ridership compared to previous years.

One potential explanation for this decline in ridership is that the pandemic has caused many people to alter their travel behavior. With widespread social distancing measures and concerns about the spread of the virus, people may be more reluctant to use public transportation and may prefer alternative modes of transportation or to stay at home.

Another factor that may have contributed to the decline in ridership is the economic impact of the pandemic. Many people have lost their jobs or have had their hours reduced, which may have reduced the need for them to use the CTA for commuting. Additionally, the decrease in tourism and business travel may have reduced the demand for the CTA's services.

In conclusion, the COVID-19 pandemic has had a significant negative effect on CTA ridership, leading to a sharp decrease in the average annual ridership compared to previous years.

3.2 Ridership Impact by Day Type



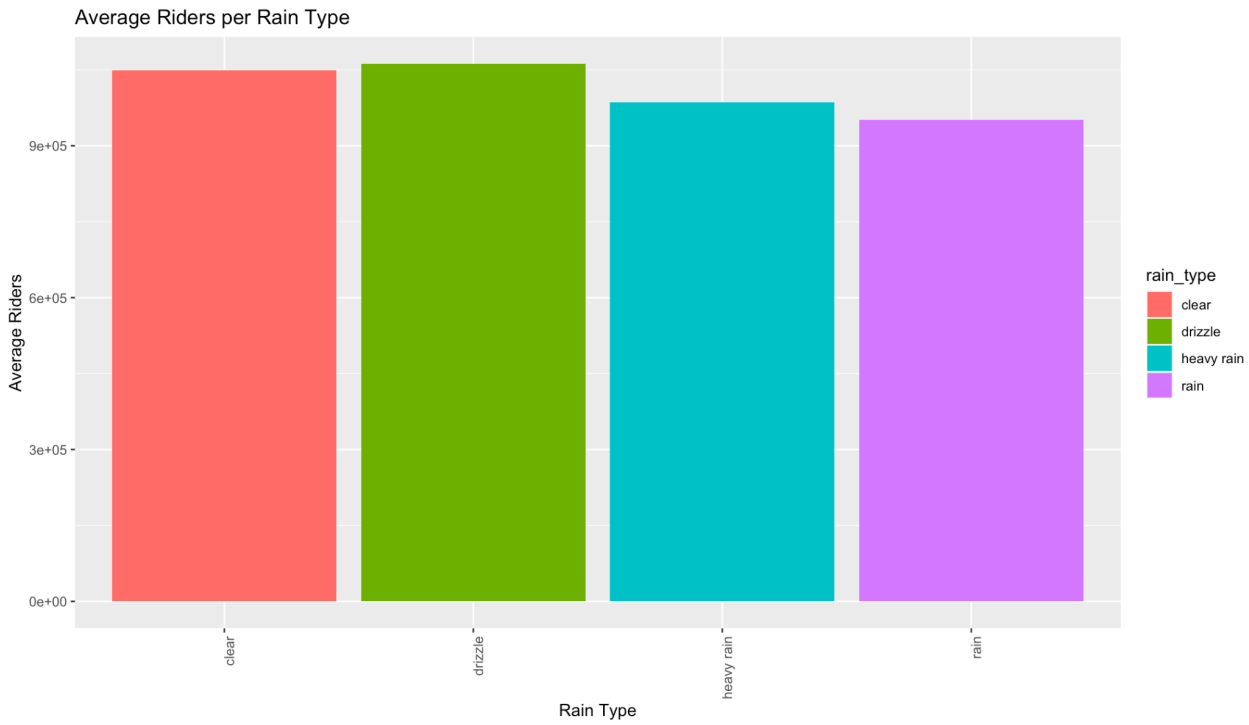
Based on this graph which shows which day people use the transit, the Chicago Transit Authority (CTA) experiences the highest ridership on weekdays (Monday through Friday). On weekends, the ridership is lower, with the least busy days being Saturday and Sunday.

One potential explanation for this pattern is that the CTA is primarily used for commuting to and from work. During the week, when people are working, the CTA is more heavily utilized. On the weekends, when people are not commuting to work, the CTA is used less.

Another factor that may contribute to the ridership pattern is that weekday evenings are likely to be busier than weekday mornings. This could be due to people using the CTA for social activities in the evenings, in addition to using it for commuting.

In conclusion, the CTA appears to be an important and heavily used mode of transportation for many people in Chicago, particularly during the week.

3.3 Ridership Impact by Rain Type



Based on the data presented in the graph, it appears that weather has a limited impact on ridership for public transportation. However, there is a clear trend indicating that heavy rain or rain results in a decrease in ridership, likely due to individuals opting to drive instead of taking public transportation.

On the other hand, drizzle seems to result in higher average ridership. Given the limited data available, it is difficult to determine the exact reasons for these trends. However, it is possible that individuals who typically walk to work are more likely to take buses on days with drizzle. This suggests that the Chicago Transit Authority (CTA) could potentially save costs by reducing the number of scheduled trains or buses on days with heavy rain or rain.

Alternatively, the CTA could implement promotional strategies to encourage individuals to use public transportation on days with bad weather. This could potentially improve traffic conditions in the city on rainy days. Further research and data collection would be necessary to confirm these suggestions and determine the most effective approach.

All code for analysis and visualization is included in the R script “Juanxi_Angel_project_analysis.R.”

4. Conclusion

Through our analysis of the "CTA - Ridership - Total Daily Trips" data, we were able to uncover several key observations that could assist the Chicago Transit Authority (CTA) in its future endeavors. By merging the CTA data with weather and holiday data, we were able to gain insight into the impact of these factors on ridership. Our findings showed that there was a steep drop in ridership after the onset of the COVID-19 pandemic, indicating that the pandemic has had a significant impact on the use of public transportation in Chicago.

Furthermore, our analysis revealed that clear days and weekdays are the most common times for CTA ridership. This information could be useful for the CTA in terms of scheduling and staffing, as they could anticipate higher ridership on these days and plan accordingly. However, we were limited in our ability to calculate the differences between different routes and their impact on ridership, as we did not have sufficient data on this factor.

Based on the available data on the CTA system, we recommend that the city of Chicago invest in upgrading and expanding the CTA's transit network. This could include adding more trains and buses, as well as improving the infrastructure of the existing network by adding new rail lines or bus lanes. Investing in public transportation could have numerous benefits for the city, including economic growth and improved quality of life. It could also reduce traffic congestion and emissions, benefiting both air quality and the environment. In addition, the CTA could consider implementing new technologies and services to enhance the customer experience, such as mobile ticketing and real-time information displays. Overall, these measures could help make public transportation in Chicago even more effective and efficient.