

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

Krishna Chatrathi
(A20520106)

Ranjitha Aswath
(A20526404)

Sanketkumar Patel
(A20523237)

1 Project Proposal

1.1 Objective

The goal of this project is to examine the Chicago Transit Authority (CTA) system in order to determine the busiest routes and stations, track the change in ridership for a bus route daily, identify service delays and disruptions for particular routes and stations, evaluate the performance of various neighborhoods in the city, and compare travel wait times with scheduled service times. The ultimate objective is to offer insightful information about the effectiveness of the CTA system.

1.2 Problem statement

Below are the questions our project seeks to address:

1. Which routes and stations are busiest in the CTA system?
2. What is the change in ridership for a bus route daily in a week?
3. For particular routes and stations are there any service delays and disruptions?
4. Are there areas of cities with better performance than others?
5. What are the travel wait times, and how do these compare to the scheduled service times?

1.3 Methodology

To conduct our data analysis, we follow below steps:

1. **Data Collection:** We collect relevant data from the CTA website and ensure integrity.

2. **Data Preprocessing:** Clean and transform the raw data into a format that can be easy to understand and analyze by the model.
3. **Exploratory Data Analysis (EDA):** Perform EDA to get initial insights into the data to understand patterns and spot anomalies and outliers.
4. **Data Analysis:** Analyze the data by applying different techniques such as regression models, cluster algorithms, etc.
5. **Data Visualization:** The results of the analysis are shown with different visualization methods such as graphs etc to understand the result better.

2 Project Outline

2.1 Related work and 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/>

2.2 All Data Sources and Tools

Assumptions:

- We have assumed the wait time by riders in 3 categories
1-5 minutes
6-10 minutes
More than 10 minutes.
- Above are initial assumptions, we are conducting the below survey to understand the nature of the wait times
[Wait periods](#) .

Data columns:

Below are the different variables used in our dataset.

- Origin: The station where a particular bus starts (Categorical)
- Day of the week: The day of the week (categorical)
- Decimal Time: Time of the day when a bus arrives (numerical)
- Destination: The station where a particular bus arrives. (Categorical)
- Travel time: Time taken to reach the goal. (Numerical)

- Route Number: A number is given to a particular route taken by a set of buses. (categorical)
- Station Name: Name assigned to a particular location where a bus on that route stops. (categorical)
- Station number: A unique identifier is given to every station for all CTA bus/train stations. (Categorical)
- longitude/latitude: Geospatial coordinates of the stations. (Categorical)
- Connecting Routes: Mentioning the routes that are connected to the particular station which a rider can choose from. (Categorical)
- Trip_Id: A unique identification number given to a particular trip, for a particular bus on that route at a particular time of the day. (Categorical)
- Arrival_time: Scheduled arrival time of the bus at a particular station.
- departure_time: Scheduled departure time of the bus at a particular station.
- Stop_id: It is the same as the station number. (Categorical)
- Date: It is the calendar date of a particular day. (Categorical)
- Daytype: To identify the day as either weekday or weekend or a holiday. (Categorical)
- Rides: Number of rides taken for the particular route on that particular day. (Numerical)
- Bus_Boardings: Number of total rides taken for all buses on that particular day. (Numerical)
- Rail_Boardings: Number of total rides taken for all trains on that particular day. (Numerical)
- Total rides: Bus_Boardings+Rail_Boardings(numerical)
- Wait time: The difference between the scheduled arrival time of a bus to a particular station and the actual arrival time. (numerical)

Tools:

- Libraries/Packages: Tibble, DT, knitr, tm, ggplot2, fitdistrplus, plotly, plyr.
- Softwares : RStudio, R project management, Python.

Future Aspirations:

1. How does the weather conditions affect the delay time of buses or service disruptions?
2. How does the weather affect ridership?