DSC 365 Milestone 3

Group 2: Transit Trackers
Sachit Patel
Rohan Dhoyda
Kevin Thompson
Abbie Van Nuland
Logan Williams

Data Source:
https://www.meteoblue.com/en/weather/historyclimate/weatherarchive/chicago_united-states_4887398

Current Plan for Final Visualizations:

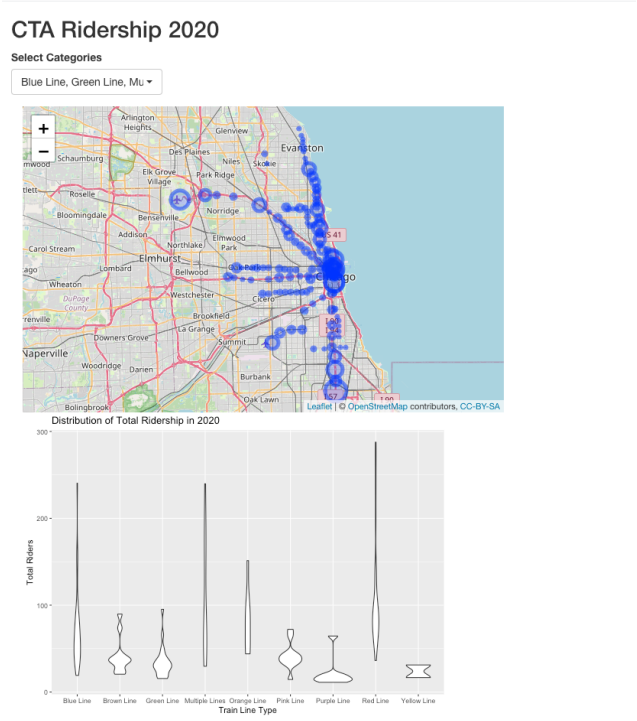
For our 2 lines of research, we're building off our dataset by relating ridership to certain conditions. The conditions we'd be looking at are COVID's impact on ridership, and weather-related factors that may dissuade ridership numbers (temperature, rainfall, and snowfall). There are some limitations with the datasets, weather-related data can not provide accuracy towards every corner of the city and is moreso just an average, and the data only spans 2022 and 2023. The further back we go, the more we have to rely on monthly averages instead of daily data, which gives less overall data to work with.

The weather dataset is sourced from meteoblue, which hosts a large archive for weather-related datasets pertaining to specific locations. Unfortunately, the dataset available only spans 18 months, as data from 2021 and before is locked behind a paywall. While there is data for 2023, it ends off on June 30th. The interface on the website allows for selecting different variables to include as well as different timeframes for the variables (daily or hourly for temperature for example). The variables taken in a daily context for our research were the temperature (given in a minimum, maximum, and mean per day) as well as total precipitation and snowfall per day. This can be combined with the dataset within this timeframe on a left join to show all overlap, and add context to the dataset.

Draft Explanatory Visualizations:

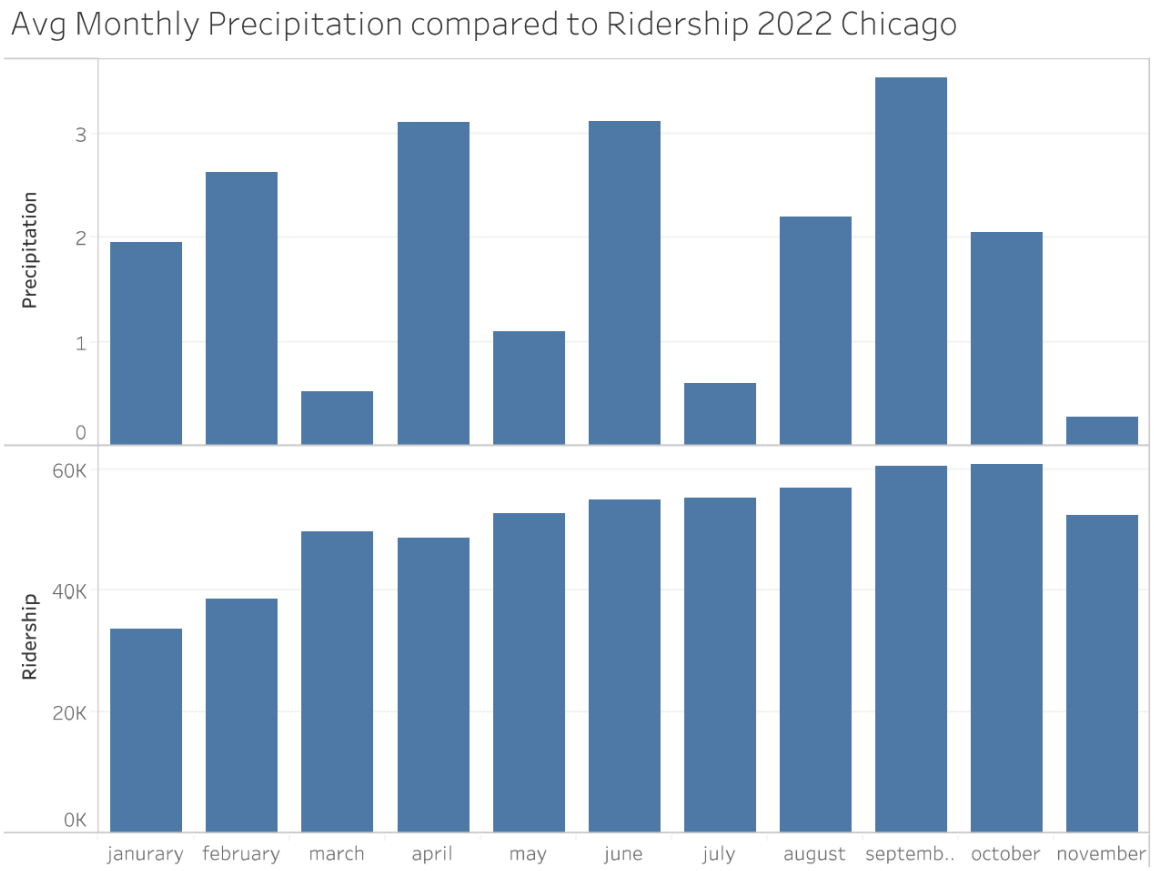
This is a sample output of an interactive shiny app with the dataset regarding 2020 (peak of COVID). It allows for selecting different lines to be displayed on the map, and multiple can be selected at once. This helps see numbers and tie them to the geographical context of how the train lines are laid out. Below the map is an automatically-generated violin plot for the selected train lines. This is just a draft, but is something we'd like to flesh out more. As we continue to explore the data it will allow us to create visualizations tailor made towards a specific message that we are trying to convey.

The idea is to have these visualizations included in the Shiny App so that users can interact with the data, seeing what happens when we control for a certain line type or time period. The goal is to have some more detailed and complex visualizations included that highlight the contrast between pre and post COVID ridership. As you can see below, the app is currently only for 2020 (the peak of COVID).



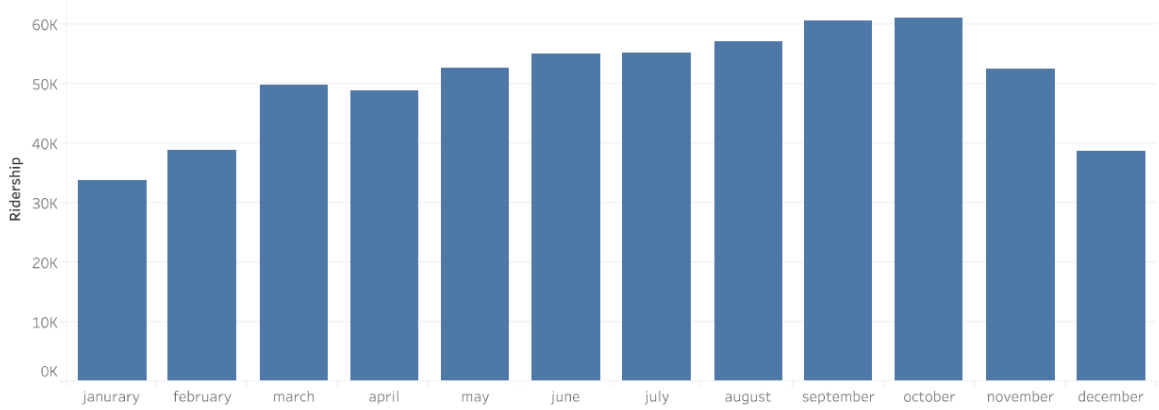
Ridership and Precipitation Averages:

Below are various plot outputs regarding total rainfall, total ridership, and average monthly precipitation compared to ridership. We don't see a monthly trend but I think we would only really see a significant change on especially rainy days. I doubt weather affects a month very much. However, what's a bit interesting is that ridership really goes up during the winter months. I suspect this is due to holidays, and actually bad weather. People are probably more willing to take a subway rather than drive if the weather is bad. Whilst some may not want to make the walk or commute to the train (let alone use it), it's safer than driving in these conditions.

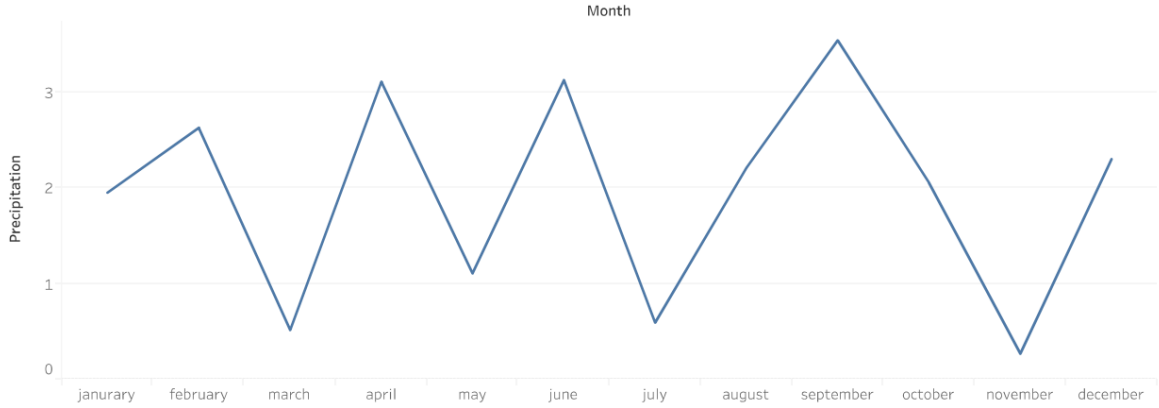


Precipitation and Ridership for each Month.

Total Ridership Chicago 2022

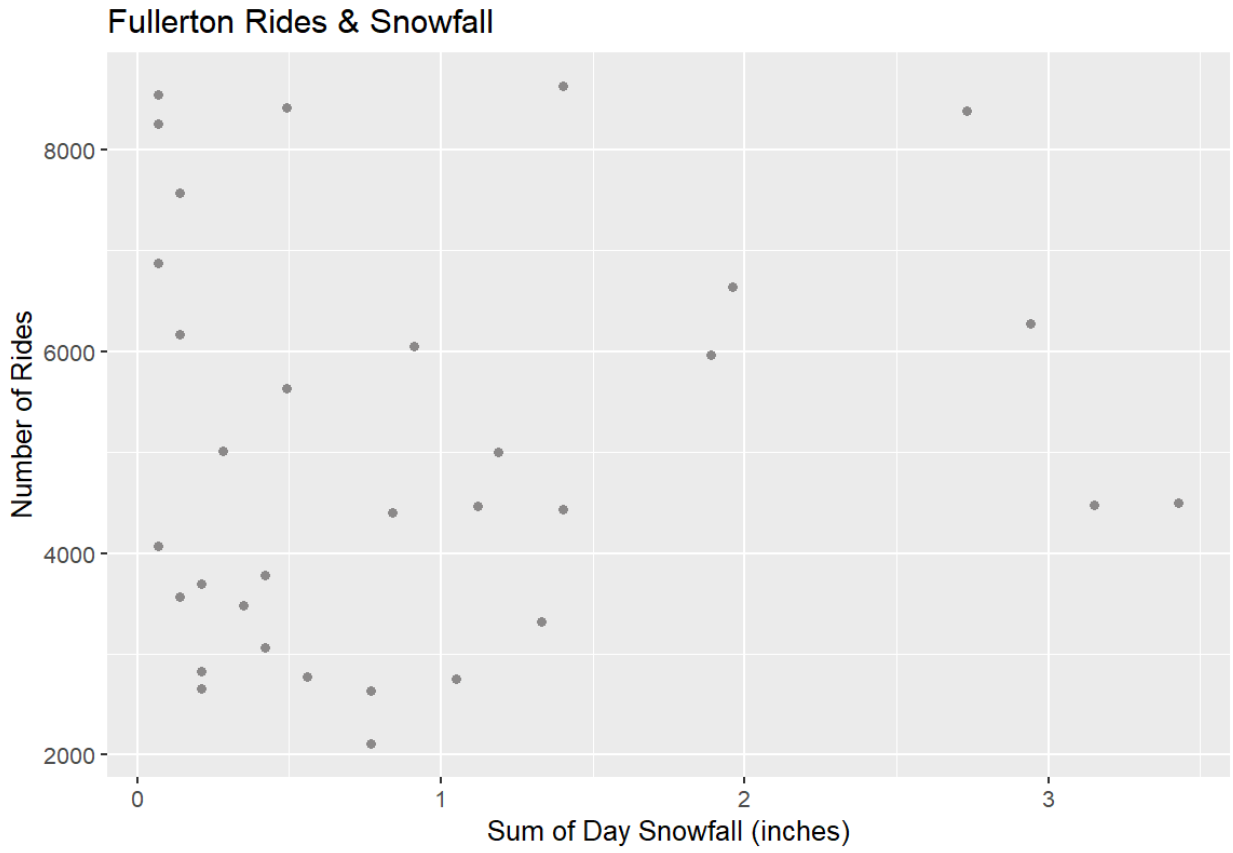


Avg. Monthly Rainfall



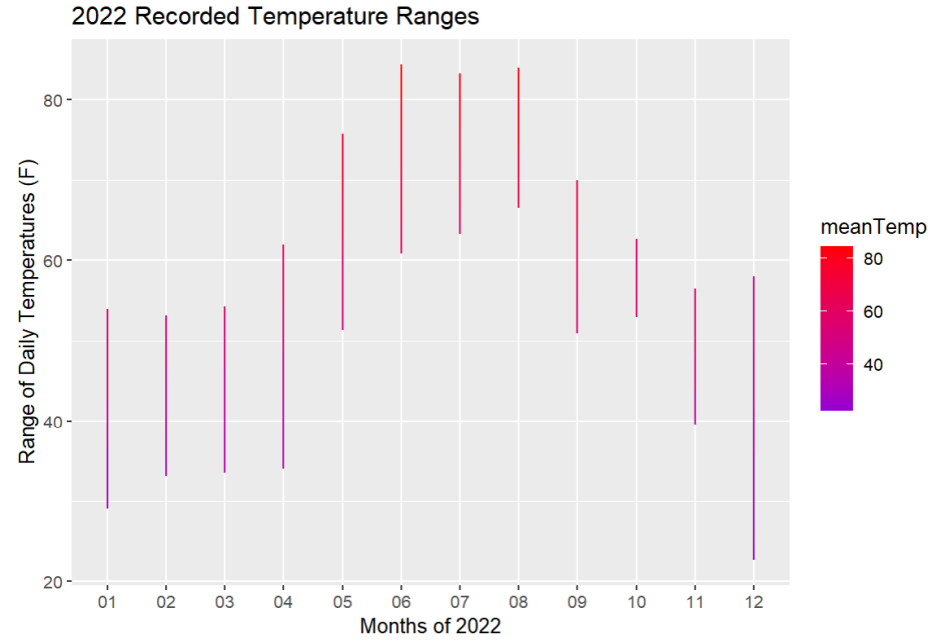
Specific Querying Potential with Left Join (combined datasets):

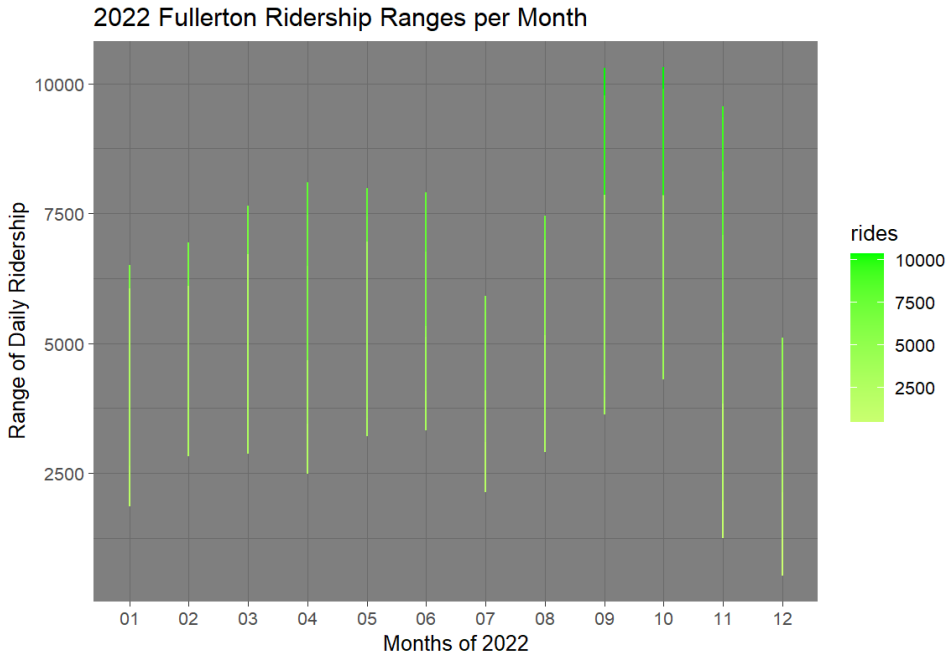
Below is a sample output after cleaning and combining datasets with a left join. In this output, the days that had snowfall are plotted against the ridership that day for a specific station (Fullerton in this case, a station with Red, Brown, and Purple Line stops). There isn't correlation here, this explanatory visualization is meant to show the depth in which information can be requested thanks to the inner join and utilizing station_id's, dates, and weather/precipitation as factors.



Ranges Over the Year Plots (combined datasets):

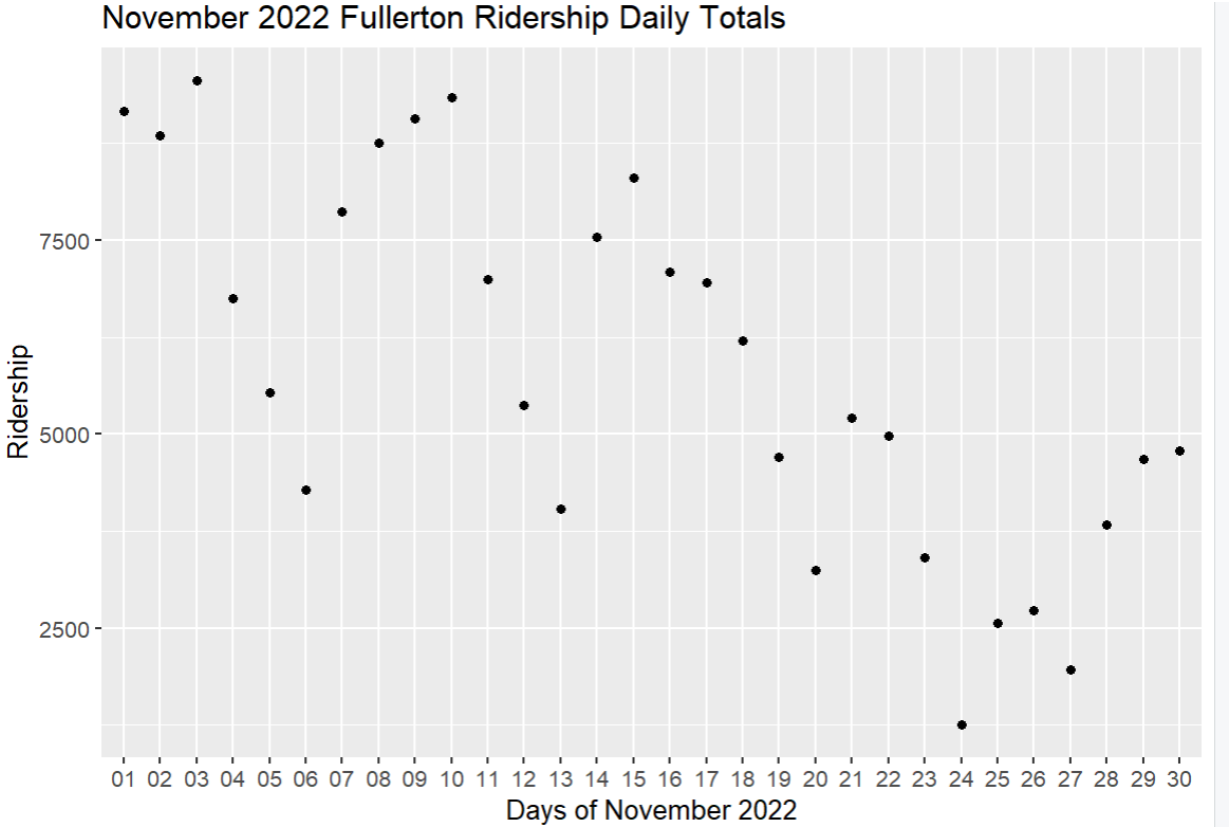
Below are outputs of range plots for different specifications. One is the range of mean daily temperatures for each month. This allows us to see the ranges we're working with, not much insight to draw from it. The second is the range of daily ridership for each month, **which lets us see the highest and lowest points of the year for a specific station.** We could combine multiple stations into this search, but for the explanatory output, I chose to use Fullerton again.





Extended Visualizations:

The range in ridership for Fullerton had an extremely wide range, a mix of the consistent pattern and low-points with the combination of weekends and holidays. It can be seen that in early November, ridership ranged from a minimum of 4000 to a maximum of 8000. However, as winter reached its peak, daily ridership dropped to less than 2500. Towards the end of November, there was a noticeable attempt to recover and reach a peak of around 5000, possibly influenced by Thanksgiving or improved weather conditions.



Multi-plot visualizations:

Here are some extra samples of ridership ranges for red line stations. It's worth noting Addison's higher range and scale. **This is because the station is the closest to the Cubs stadium, where many fans go to watch games.** If we were to use the shiny app, we could see what days reflect the cubs games, and perhaps get a rough idea of **where the fans are commuting from, since other stations would also have increased ridership for that day.**