CTA “L” Ridership Summary

We chose our specific project because we are all from Chicago and ride the L weekly, if not daily. If you live in Chicago, it’s potentially a big part of how you get around the city. We initially set out with the question of “Has L ridership recovered from the COVID-19 pandemic?” However, the data that CTA provided did not have enough ridership data for 2021 available yet, so we pivoted to the final question of “How has L ridership changed over the years?” We also wanted to find out if there was a seasonal component to ridership and which L lines see the most ridership. After we finalized those questions, our initial hypothesis was that we would see an increase in ridership across all lines, the summer months would be the most popular for L’s, and that the Red Line would have the most ridership in Chicago.

Our data source was directly from CTA using <https://data.cityofchicago.org/> . From their homepage, we navigated to the Transportation tab. We used the CTA “L” Ridership-Daily Totals spreadsheet (https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Daily-Totals/5neh-572f ) , the CTA - Ridership – “L” Station Entries - Monthly Day-Type Averages & Totals (<https://data.cityofchicago.org/Transportation/CTA-Ridership-L-Station-Entries-Monthly-Day-Type-A/t2rn-p8d7>) , and lastly the CTA System Information – List of “L” stops to find which station correlated with which line (<https://data.cityofchicago.org/Transportation/CTA-System-Information-List-of-L-Stops/8pix-ypme>).

The first step in cleaning our data was the get these CSV files into Pandas so we could merge them. Every station has a unique Station ID, which made merging the DataFrames easy. We merged the L-Stops with the Monthly Totals. We wanted to be able to split the Months away from the Year, so we used pd.to\_datetime on the ‘Month Date’ column. This made it easy for us to a dedicated Month and Year column and then drop the initial ‘Month Date’ column. Then we renamed some columns for easier identification and dropped more unnecessary columns from the dataframe. After this, we noticed that during the merge, some duplicate line items were added. A simple df.drop\_duplicates() fixed that issue. Finally, we had a manageable DF that we could start splitting, grouping by, and manipulating in order to answer our questions.

The first question we posed was How has L ridership changed over the period we pulled data for. We noticed initially that most lines had an increase from 2014 to 2015. 2015 for most of the lines, was the highest year total. However, from years 2015-2019, we see a decrease in ridership across most of the L lines, including Red, Blue, Pink, Green, and Purple. As to be expected, a substantial drop in ridership occurred in 2020 due to the pandemic.

The second question we looked to answer was what L Lines had the highest ridership. Without question from the chart used to answer the first question we posed, Red and Blue have the highest ridership. We analyzed all the stops across these 2 lines to figure out the Top 5 busiest L stops in Chicago (95th/Dan Ryan, Chicago/State, Grand/State, Lake/State, Washington/Dearborn) in order to further analyze trends. We noticed the same amount or a small increase in ridership from 2014-2015 in all the stations. From 2016-2020, we noticed a decrease in ridership across all stations except Washington/Dearborn.

The third question we looked to answer was if there was a seasonal component to ridership. We split the months in to 4 equal groups to signify the seasons. January, February and December were Winter. March, April, May were Spring. June, July, August were Summer. September, October, November were Fall. When we grouped our data across 2014-2020 into these buckets, we noticed that there is only a slight seasonal component to ridership. The winter months category was the only category to not hit 300 million riders. Surprising to us, the Fall category had the highest ridership.

The final question posed was to view how L ridership was impacted in 2020. As to be expected, we noticed a sharp decrease in ridership in April 2020. This coincides with the COVID lockdowns that occurred in March 2020. We notice a small spike in ridership starting in June and ending around October 2020.

The additional questions we would like to address would be to see how ridership has changed into 2021 now that businesses are reopening, and the vaccine is widely available. We would need to wait for CTA to publish the data for 2021 and then run the same analysis on it that we did for this presentation. Another question we would like to pose is “Why are the most popular L lines so popular?” We could use the Google Maps API to pull some results of what’s near each stop ie. Restaurants, Apartments, Businesses, etc. Also, we just generally could get more granular on our process as well. For seasonal, we could try to pinpoint the highest months of each season. We could then plot each day of the month’s ridership. Also, we could do analysis on Weekday vs Weekend ridership and attempt to figure out if one category is significantly higher than the other.