Exploratory Analysis for CTA Data:

The CTA stations data was taken from the City of Chicago’s open data site. It consisted of 3 data sets, L Station Entries, which had the daily entries for every station in the system, Annual ridership which contained the total ridership each year up to 2018 and daily ridership which contained the total amount of riders for all of the transit system for each day. L Station Entries would prove to be the most useful as it allowed us to compare individual stations. During our investigation of the data, we noticed as few interesting points. First, when graphing the trend of daily ridership or time we could see a clear drop off during the covid-19 pandemic and the limited recovery of ridership afterwards as seen in figure 1. Additionally, we could see one large spike in 2016 that we wanted to investigate as it could be an outlier.

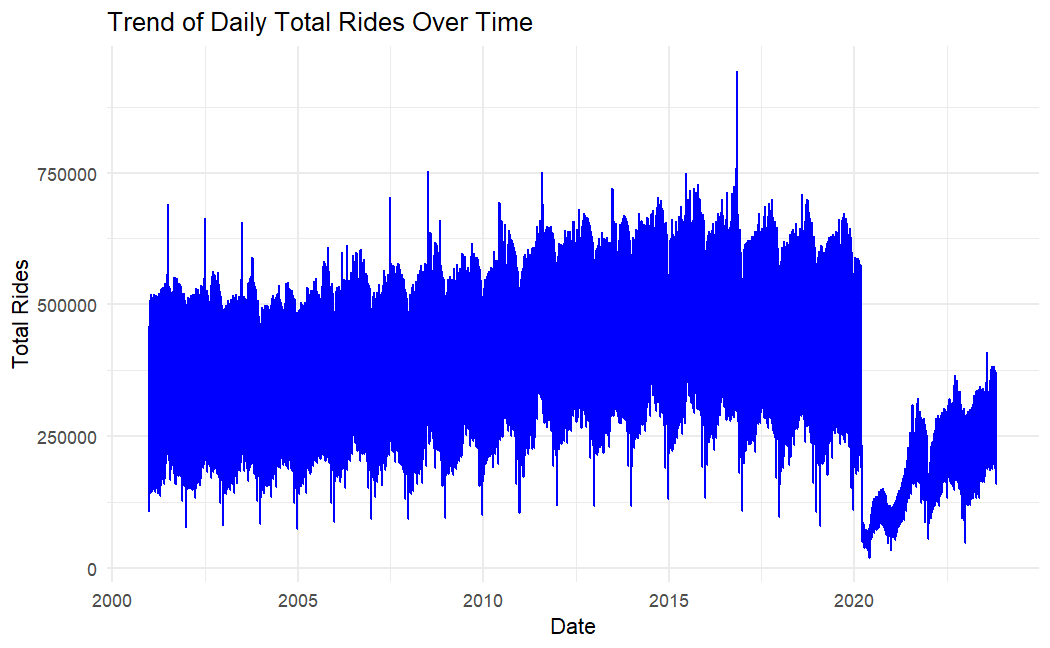
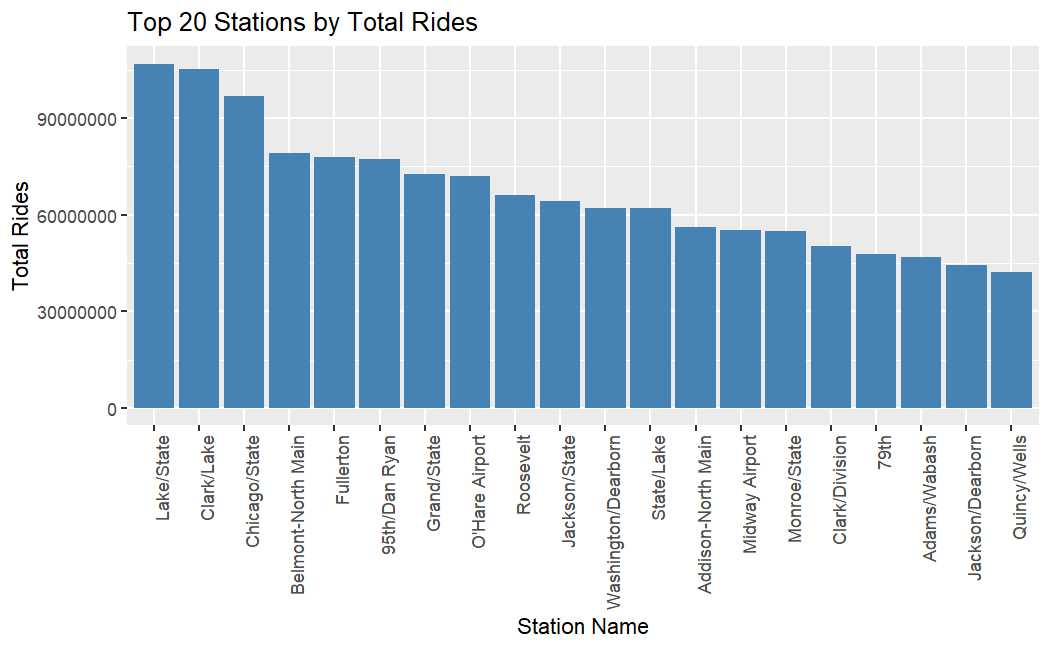
Next, we investigated the most used (figure 2) which showed us that the majority of the top 20 stations were located in the Loop or downtown with a few exceptions. Belmont-North Main and Addison North-Main are the main stations to access Wriggly Field, Fullerton is the main DePaul stop. both Midway and O’Hare stops were also in the top 20. More interestingly, 95th St. and 79th St. stations were also in the most used. Both of these stations are on the far south of the Redline. This likely means that they are pulling in more riders from the surrounding areas that do not have any stations. Given this, the CTA’s new Redline Extension project that will add four new stations to the south of 95th St. makes sense as it will give more access to the CTA and relieve pressure on those stations. We also noticed that the largest rides at individual stations were all related to the Cubs (figure 3). We could see that Belmont North-Main was the main station used and the two others (Lake/State and Clark/Lake) were both on Nov. 4th, 2016, the day of the Cubs’ World Series Victory Parade.

Figure 1: Trend of Ridership

Figure 1: Most used Stations

Figure 2: Daily Ridership Trend

Lastly, we compared the differences between station usage on the weekend vs during the week. There were some differences between the satiations used between the two as seen in figures 4 and 5.

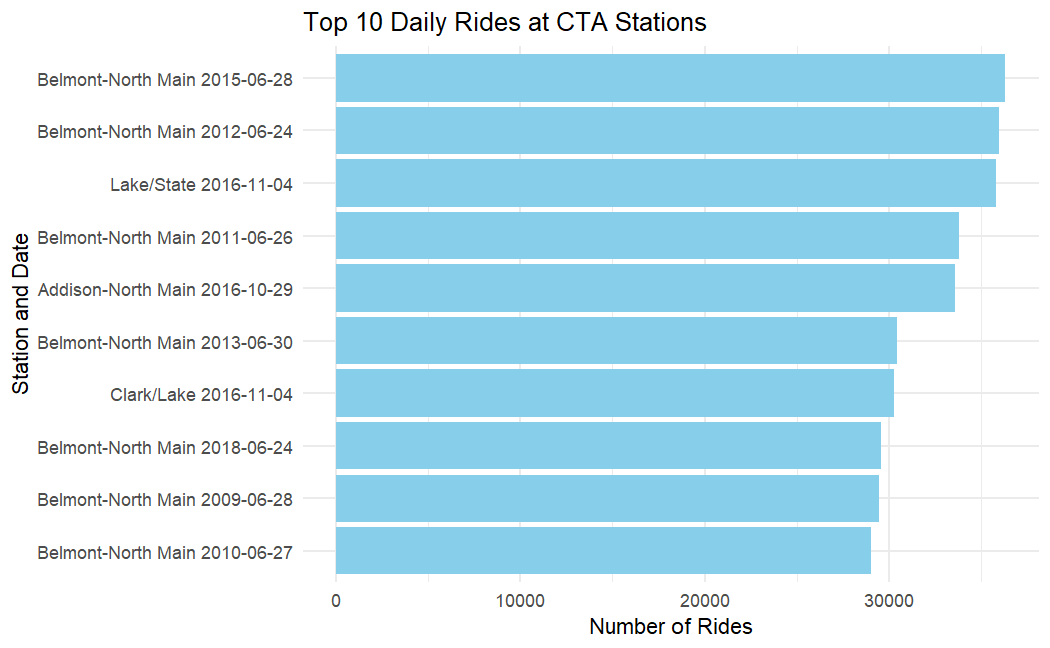
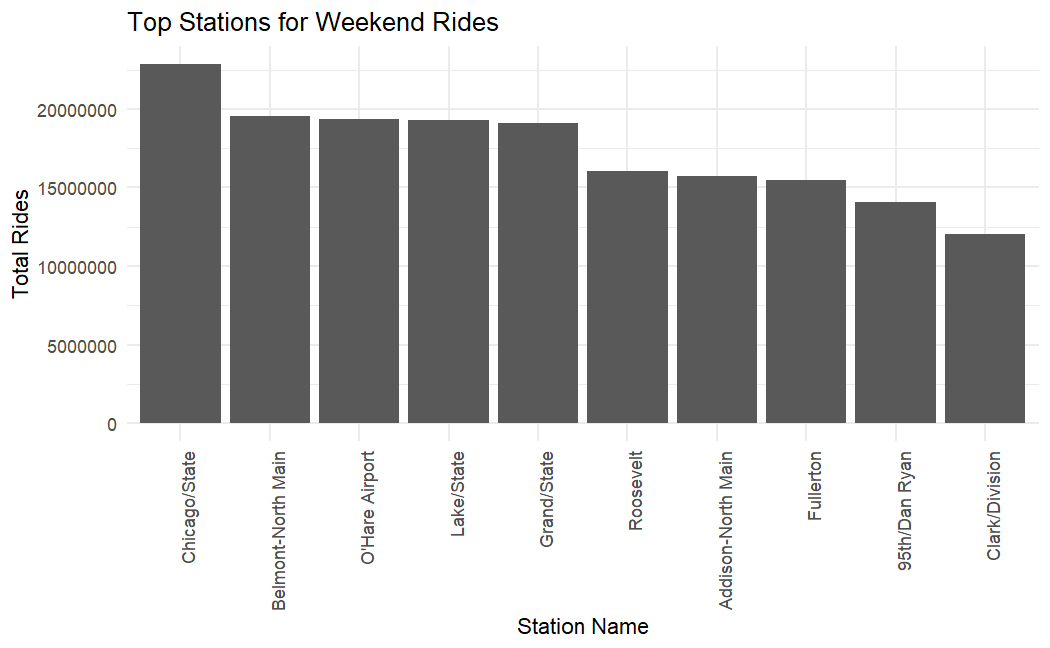
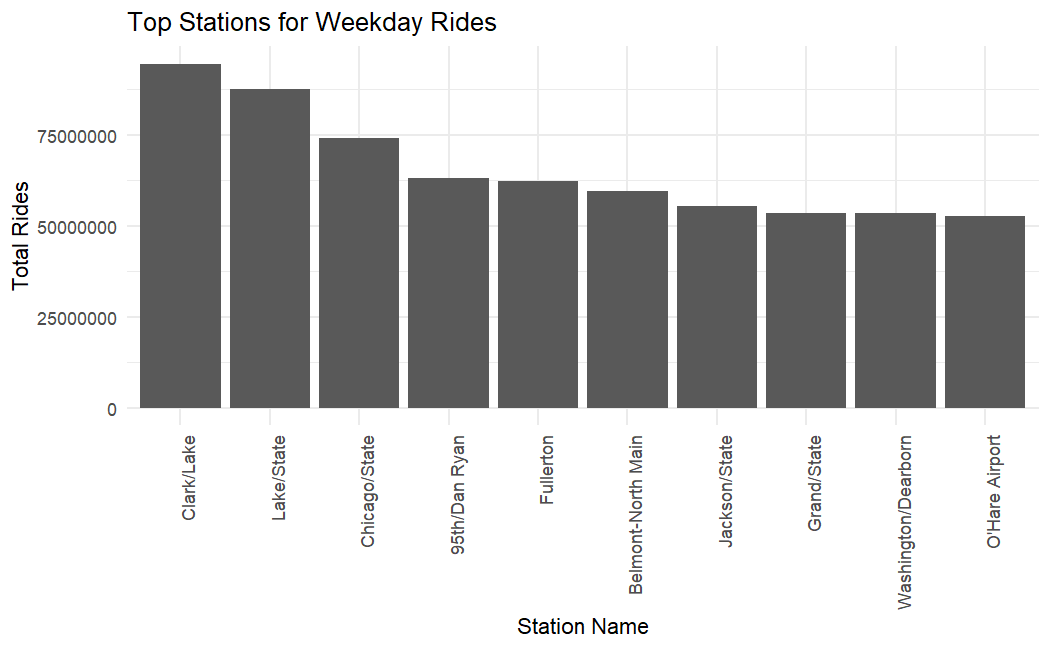


Figure 5: Most used weekday stations.

Figure 4: Most used weekend stations.

Figure 3: Top 10 days with highest ridership by station.

Exploratory Analysis of Census data:

The census data we used was taken from the Census Bureaus publicly available datasets and was downloaded in several different datasets and then combined into one full data set. The data was taken from the 10-year census as well as the American Community Survey which collects additional demographic information. The merged data set contained 100 columns (99 variables plus a column containing the census code to match it to a geographic census tract) and 2072 columns (One for each census tract in IL). This data set included variables that accounted for occupation type, income level, car ownership, public transportation usage, what time people left for work, commute time to work, owning vs renting, race, sex, education and age. Education, Age and Income all showed generally normal distributions with some skew. Median income did have an outlier census tract with a median income of $200,000. We can also see that the majority of the population is white, with Hispanic and Black or African American making up the other two largest parts of the population. Commute times were pretty evenly spread out with the majority of commutes being under an hour and about 14% being 60 minutes or more. This data would be combined with the CTA station data to create our final data set to conduct our analysis with.

