**Introduction**

Our goal in this study was to utilize the publicly available CTA ridership data with detailed demographic information from the US Census Bureau to develop a robust model capable of accurately predicting CTA station usage. Additionally, our goal was to create a predictive tool that could be applied to the newly planned Red Line stations—103rd St, 111th St, 116th St, and 130th St. This model would not only enhance our understanding of current station usage but also provide valuable insights for future transit planning and infrastructure development as these new stations become operational.

**Data Sources**

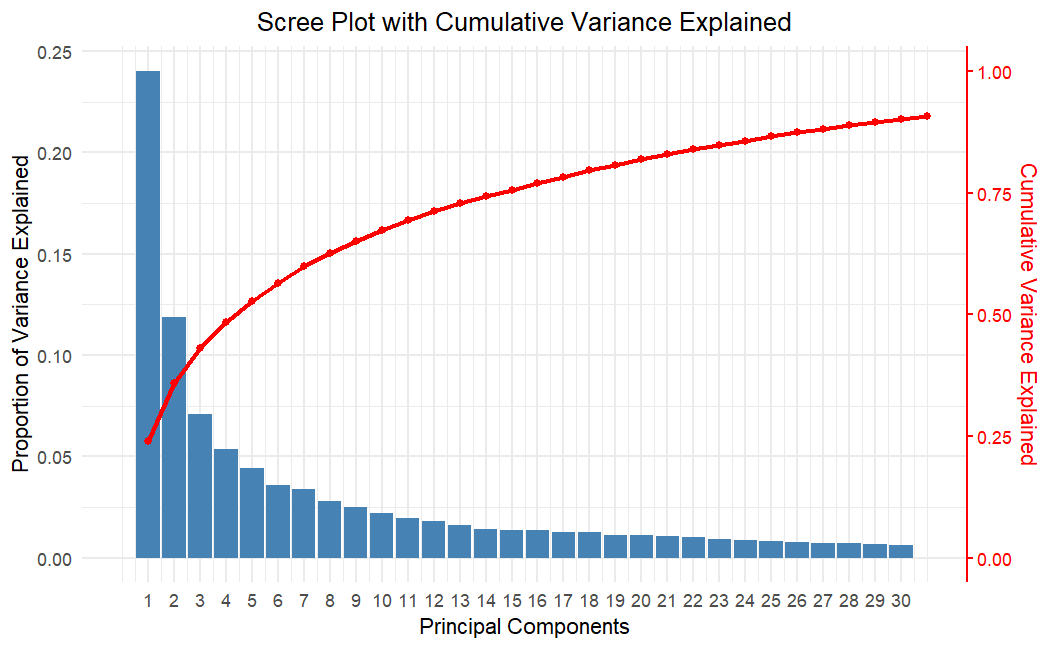
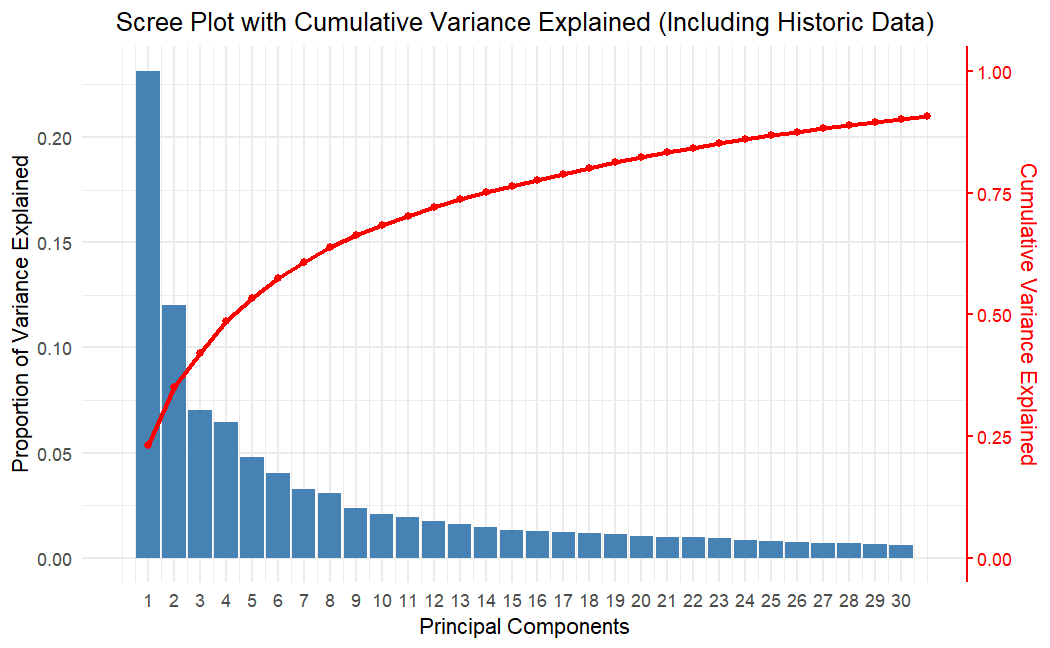
To begin to develop a model that could predict CTA station usage based on demographic information we first had to combine the 4 datasets that were downloaded from the US Census Bureau’s data website. This data was downloaded in four separate datasets and then merged using R. The individual datasets consisted of occupation and commute data, housing data, education and income data, and age and gender data. Each row of these datasets was a census tract and they were joined using the Census tract code which was the index for each dataset. This gave us a data set of 99 attributes.

Additionally, four data sets were downloaded from the CTA database, Annual Ridership, Daily Boarding Totals, and L Station Entries and L station locations which contains the coordinates of the L stations. The L Station entries dataset will be the only one used for developing our models as it was the most detailed and contained the most information. The coordinates were used to match the stations to the correct census tract using the Tigris library in R.

**Processing the Datasets**

The first challenge encountered was merging the census data sets with the CTA data and aggregating the station entries. The station entries dataset had 1,189,058 rows with columns for station ID number, station name, the date, the type of day, and the number of rides from that station. There are three types of days which are Weekdays, Saturdays, and Sundays/Holidays. These are abbreviated W, A, and U in the data set (W=Weekday, A=Saturday, U=Sunday/Holiday). The data was sorted and aggregated so that the average daily riders for each station based on the year was calculated and added to the dataset. This created 12 new columns: A\_2020, U\_2020, W\_2020, A\_2021, U\_2021, W\_2021, A\_2022, U\_2022, W\_2022, A\_2023, U\_2023, and W\_2023. The 2020 – 2022 columns are used as historical data to build the model and our target variables will be the three types of days in 2023. The cut-off of 2020 was chosen due to the large change in ridership in early 2020 due to the COVID-19 pandemic as shown in the Trend of Daily Total Rides Over Time graph in the exploratory analysis section. Additionally, four new stations were added to the dataset. These are the 103rd St, 111th St, 116th St, and 130th St. stations which are part of the CTA’s new Red Line Extension Project. To get the census data for this, the station locations were mapped in Google Earth and the coordinates were used to match them to the correct census tracts using Tigris in R. The past ridership data was estimated for these stations by taking the closest stations on the red line to them and averaging the ridership for them for each year and day type.

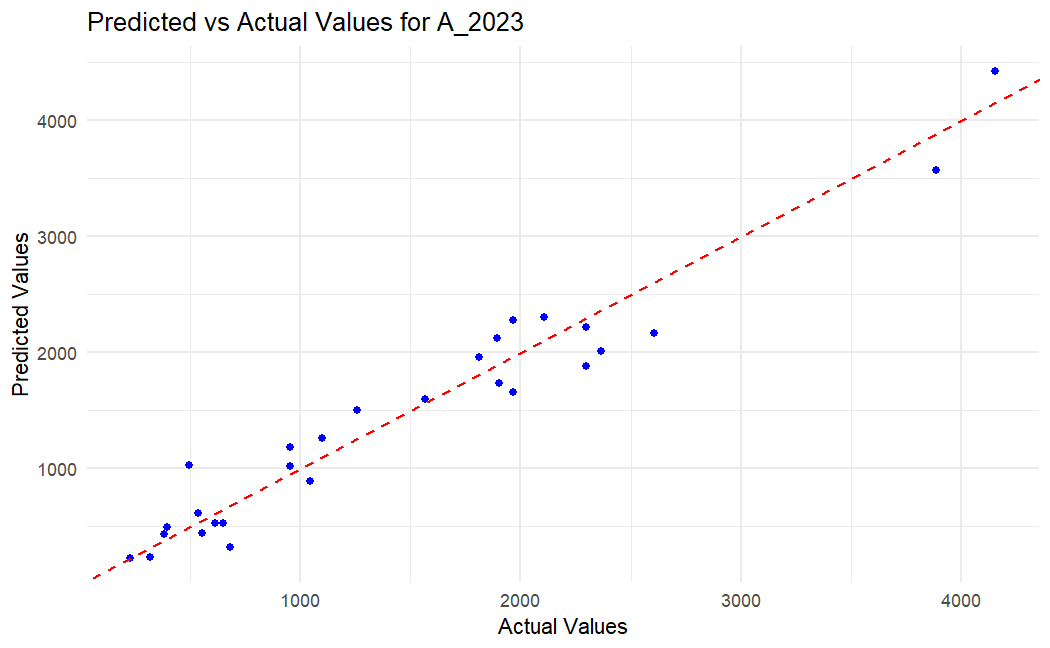
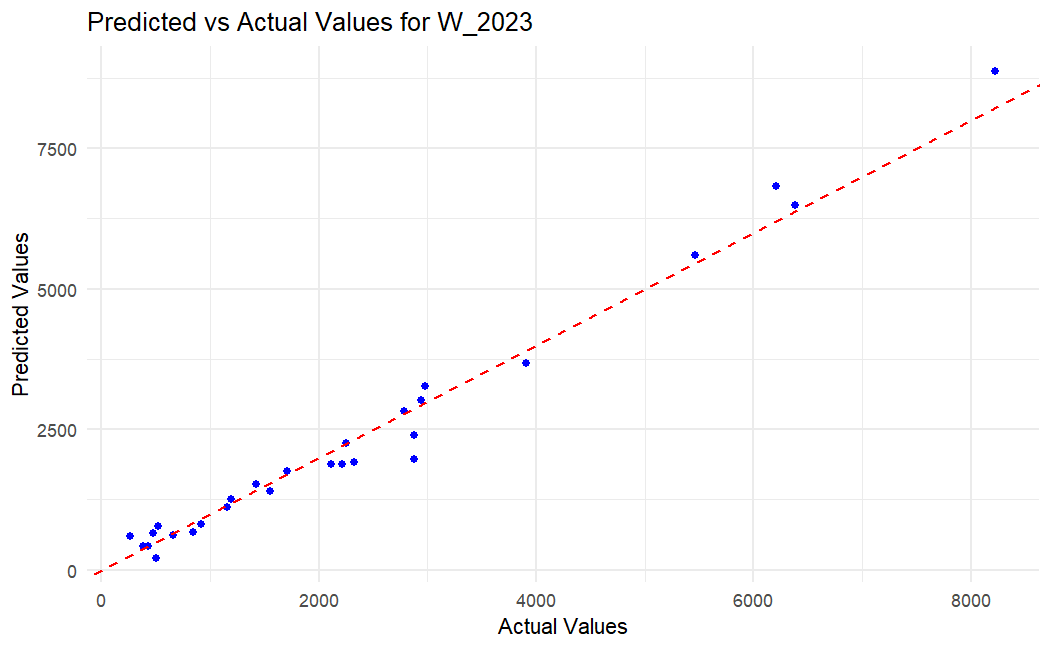
These new columns were then joined to the merged CTA datasets to complete the dataset used to train the models. The following variable columns were removed as they contained mostly 0s across all census tracts in our dataset: Workers in military occupations, Agriculture forestry fishing and hunting and mining, Armed forces, Unpaid family workers, American Indian and Alaska Native, and Native Hawaiian and Other Pacific Islander. The final data set and all variables are shown in Appendix A.

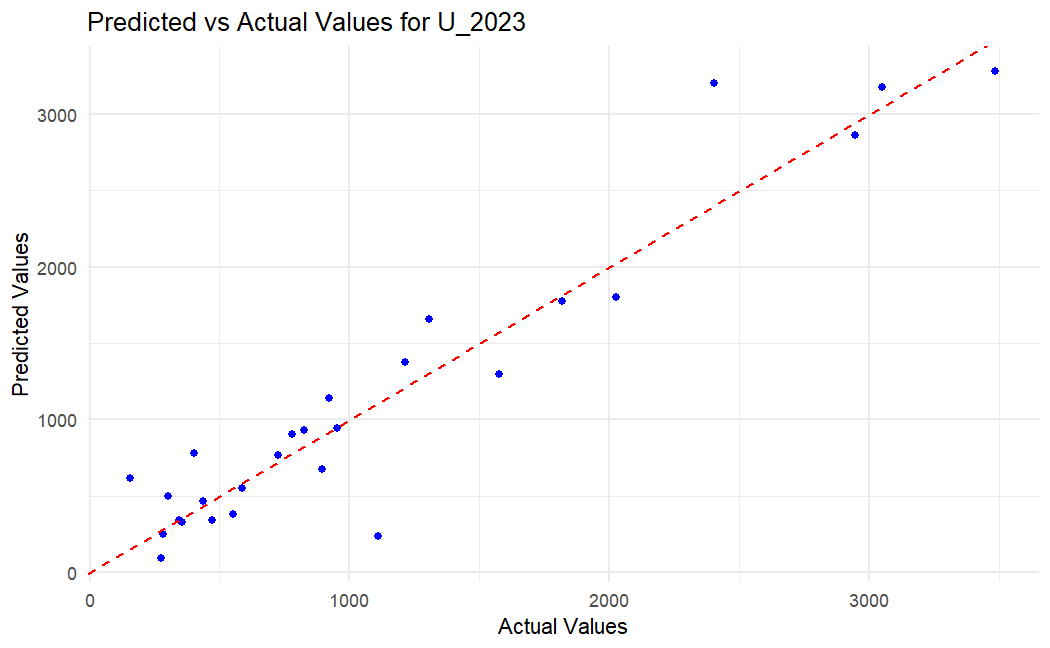
**Model Development** Several different models were tested using RMSE as the metric to compare the results. The first model was done using only the census data (all historical ridership data was removed) with the target variable of A\_2023 (Saturdays in 2023). This was done to see how accurately the census data alone could predict the ridership. The data was scaled and then a linear regression model was applied. This showed that this model was not an effective way of predicting ridership with an RMSE of 754.64, a Multiple R-squared of 0.7654, and an Adjusted R-squared of 0.3772. This difference between the R-squared and Adjusted R-squared implies overfitting or multicollinearity issues.

Given the large number of variables, PCA was applied to reduce the dimensions and to handle the multicollinearity issues that were shown in the first model. Two different versions of PCA were conducted. The first excluded the historical ridership and the second included it. This was again done to see how accurately the census data alone could predict the ridership. The first PCA created 93 PCs and the first Scree Plot shows the first 30 which accounted for approximately 90% of the cumulative variance. The historical ridership data created a total of 102 PCs. The Scree plot below shows the first 30 PCs which accounted for approximately 90% of the cumulative variance. Models created using the PCs without historical data included performed poorly compared to the models with historic data, so the final models would include the PCs created with the historic ridership data.

The PCs were used as the new data set and a training and testing split was created with .8 used for training and .2 used for testing. 10-fold cross-validation was also used for the creation of the model. The initial model was a linear regression model that used stepwise feature selection to select the best PCs to use. This method selected 91 PCs and returned an RMSE of 1616 when predicting the Saturday ridership in 2023 (Variable A\_2023). Additionally, the Multiple R-squared and Adjusted R-squared were both .99 implying overfitting of the model. To reduce the number of PCs used and reduce overfitting, a new model was trained that only used the most significant PCs. This created a new model using 67 PCs with a Multiple R-squared of 0.925, Adjusted R-squared of 0.8753, and RMSE of 646.64 showing an improvement. Additionally, three SVM models were trained and tested using grid search to find the best parameters. These were a Linear SVM (RSME: 573), Polynomial SVM (RSME: 799), and RFB SVM (RSME 872). These SVM’s RMSE scores did not show a major improvement over the previous regression models.

**Elastic Net Models**

 The Elastic Net Regression method was then applied to see if this could improve over the simplified model above. To create this model a new testing and training split was created and an Elastic Net model was trained using grid search to find the optimum Lambda and Alpha parameters. The grid search was done for Alpha between 0 (Ridge Regression) and 1 (Lasso Regression) and for a Lambda between 0 – 20. The search was done in sequence of .01 to ensure the best parameters were found. The A\_2023 (Saturday ridership) variable was again used as the target variable. The Alpha parameter chosen was .66 showing that it is favoring Lasso Regression but is still using some Ridge Regression. The Lambda chosen was 7.83 showing a high level of regularization to avoid overfitting. The RSME of 242.9 was the best RSME of all models so far.

 Two additional models were created using an elastic net model and grid search for the weekday (W\_2023) and Sunday/holiday (U\_2023) variables. These also returned RMSEs that were similar to the first elastic net model with U\_2023 having an RSME of 294 and an Alpha of .98 and Lambda of 1.04. W\_2023 had an Alpha of .59 and a Lambda of 3.94. Creating a separate model for each day type ensured that the differences in ridership between weekdays and weekends/holidays did not lead to underfitting of the model.

**Conclusions**

The Elastic Net models proved to be the most effective at predicting the ridership for CTA stations. They outperformed the other models significantly with only the linear SVM having a RMSE of under 600. This comparison shows the RMSE results of all models tested:

|  |  |
| --- | --- |
| Model | Test RMSE |
| Linear Reg. - Scaled Data | 754.6361 |
| Linear Reg. - PCA w/o Historical data - All PCs | 7294.794 |
| Linear Reg. - PCA w/o Historical data - Stepwise PC Selection | 6746.441 |
| Elastic Net - PCA w/o Historical data | 1500.494 |
| Linear Reg. - PCA Historical data - Stepwise PC Selection | 1616.786 |
| Linear Reg. - PCA Historical data - PCs with significances <= 0.001 | 636.6444 |
| SVM - Linear | 573.8628 |
| SVM - Poly | 798.9153 |
| SVM - RBF | 871.5119 |
| Elastic Net with grid search for A\_2023 with historical data | 242.8974 |
| Elastic Net with grid search for W\_2023 with historical data | 313.7893 |
| Elastic Net with grid search for U\_2023 with historical data | 294.2801 |

The Elastic New with grid search and the historical data clearly developed the best model with all RMSE being the lowest of any model.

While the newly planned 103rd St, 111th St, 116th St, and 130th St stations will not be open in 2024, these models could be used to estimate their usage once they become open. These models could be updated with the ridership levels for 2024 and the following years as the project is being worked on as well as updating the Census data if new data becomes available. This would allow the CTA to better predict the usage of these stations as they get closer to opening. Additionally, the trend of overall ridership will likely begin to return to higher levels as ridership returns to normal after the post-pandemic drop. If the stations were to have opened already, these would be the predicted daily average uses for 2023:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Average Daily Riders | |  |
| New Station | Weekdays | Saturdays | Sunday/Holiday |
| 103rd St | 2734 | 2037 | 1647 |
| 111th St | 2755 | 1949 | 1588 |
| 116th St | 2820 | 2278 | 1647 |
| 130th St | 2765 | 1659 | 1605 |

**Future Work**

There are several ways that this project could continue and how the models could be improved upon. During our research, we noticed that many other transit systems collected additional ridership information such as a passenger’s exit station. This allowed the transit authority to see passengers' entire trips. While the CTA does not require an exit swipe on a CTA card or ticket which could be used to track the data, recording a count of people exiting the stations would at least provide information on the most common destinations. If the CTA wanted to develop more accurate models that could account for the entire trip of a passenger, we would recommend that they create new systems to gather additional data.

Another way that these models could be improved upon would be to create individual models based on the time of year (winter vs. summer etc.) since there are likely major differences in usage during these times. However, due to the impact of the Covid 19 pandemic on the most recent data, it would likely be beneficial to have data for more post-pandemic years to get a better idea of what transit usage will be going forward. Additionally, models could be developed similarly to the ones used in the Atalanta (Santanam, 2024), where they developed models to predict usage during sudden increases in ridership. Since we noticed that the most used stations and days all were related to Cubs events this could be beneficial to the CTA. This would again need more data to be gathered to have a better understanding of how passengers are using the CTA during sporting events.

Appendix A

Combined CTA and Census Bureau Dataset

|  |  |
| --- | --- |
| Variable | Data Type |
| GEOID | Census Code |
| Management, Sci, Arts | % of Pop in Work Type |
| Service | % of Pop in Work Type |
| Sales and Office | % of Pop in Work Type |
| Natural Resources, Construction | % of Pop in Work Type |
| Production, Transportation | % of Pop in Work Type |
| Construction | % of Pop in Work Type |
| Manufacturing | % of Pop in Work Type |
| Wholesale trade | % of Pop in Work Type |
| Retail trade | % of Pop in Work Type |
| Transportation and warehousing and utilities | % of Pop in Work Type |
| IT Finance Real Estate | % of Pop in Work Type |
| Professional or Scientific Mgmt. | % of Pop in Work Type |
| Social Services | % of Pop in Work Type |
| Arts Entertainment or Food Services | % of Pop in Work Type |
| Other Services | % of Pop in Work Type |
| Public Admin | % of Pop in Work Type |
| Private Wage and Salary Workers | % of Pop in Private Sector |
| Government Workers | % of Pop in Public Sector |
| Self-Employed Workers | % of Pop in Self Employed |
| Workers who did not work from home | # of non-work from home workers |
| 12:00 AM to 4:59 AM | % of Pop commuting during these hours |
| 5:00 AM to 5:29 AM | % of Pop commuting during these hours |
| 5:30 AM to 5:59 AM | % of Pop commuting during these hours |
| 6:00 AM to 6:29 AM | % of Pop commuting during these hours |
| 6:30 AM to 6:59 AM | % of Pop commuting during these hours |
| 7:00 AM to 7:29 AM | % of Pop commuting during these hours |
| 7:30 AM to 7:59 AM | % of Pop commuting during these hours |
| 8:00 AM to 8:29 AM | % of Pop commuting during these hours |
| 8:30 AM to 8:59 AM | % of Pop commuting during these hours |
| 9:00 AM to 11:59 PM | % of Pop commuting during these hours |
| Less than 10 minutes | % of Pop with this commute time |
| 10 to 14 minutes | % of Pop with this commute time |
| 15 to 19 minutes | % of Pop with this commute time |
| 20 to 24 minutes | % of Pop with this commute time |
| 25 to 29 minutes | % of Pop with this commute time |
| 30 to 34 minutes | % of Pop with this commute time |
| 35 to 44 minutes | % of Pop with this commute time |
| 45 to 59 minutes | % of Pop with this commute time |
| 60 or more minutes | % of Pop with this commute time |
| No Vehicle Available | % of Pop with No Vehicle |
| 1 Vehicle Available | % of Pop with 1 Vehicle |
| 2 Vehicles Available | % of Pop with 2 Vehicles |
| 3 or More Vehicles Available | % of Pop with 3 or 3+ Vehicles |
| Drove alone | # of workers who drove along |
| Carpooled | # of workers who carpooled |
| Public transportation | # of workers who used public transit |
| Owned House/Apt | # of people who owned residence |
| Renting House/Apt | # of people renting residence |
| White | % of Pop of this Race/Ethnicity |
| Black or African American | % of Pop of this Race/Ethnicity |
| Asian | % of Pop of this Race/Ethnicity |
| Some Other Race | % of Pop of this Race/Ethnicity |
| Hispanic or Latino | % of Pop of this Race/Ethnicity |
| <High School | % of Pop with this level of education |
| High School | % of Pop with this level of education |
| Some college or Assoc. | % of Pop with this level of education |
| Bachelors | % of Pop with this level of education |
| Graduate | % of Pop with this level of education |
| Income $1 to $9,999 or Loss | % of Pop in income bracket |
| Income $10,000 to $14,999 | % of Pop in income bracket |
| Income $15,000 to $24,999 | % of Pop in income bracket |
| Income $25,000 to $34,999 | % of Pop in income bracket |
| Income $35,000 to $49,999 | % of Pop in income bracket |
| Income $50,000 to $64,999 | % of Pop in income bracket |
| Income $65,000 to $74,999 | % of Pop in income bracket |
| Income $75,000 or More | % of Pop in income bracket |
| Median income dollars | Median income in $USD for census tract |
| Est. Pop Below Poverty Line | % Pop below 100% of poverty line sometime in last 12 months |
| Total population | Population of Census Tract |
| Under 5 years | # of people in age range in census tract |
| 5 to 9 years | # of people in age range in census tract |
| 10 to 14 years | # of people in age range in census tract |
| 15 to 19 years | # of people in age range in census tract |
| 20 to 24 years | # of people in age range in census tract |
| 25 to 29 years | # of people in age range in census tract |
| 30 to 34 years | # of people in age range in census tract |
| 35 to 39 years | # of people in age range in census tract |
| 40 to 44 years | # of people in age range in census tract |
| 45 to 49 years | # of people in age range in census tract |
| 50 to 54 years | # of people in age range in census tract |
| 55 to 59 years | # of people in age range in census tract |
| 60 to 64 years | # of people in age range in census tract |
| 65 to 69 years | # of people in age range in census tract |
| 70 to 74 years | # of people in age range in census tract |
| 75 to 79 years | # of people in age range in census tract |
| 80 to 84 years | # of people in age range in census tract |
| 85 years and over | # of people in age range in census tract |
| Median age | Median age of census tract |
| Sex ratio males per 100 females | Male to Female ratio in census tract |
| Male Total population | Male total population |
| Male Median age | Male median age |
| Female Total population | Female total population |
| Female Median age | Female median age |
| STATION DESCRIPTIVE NAME | Station Name with Line |
| Station id | CTA Station ID code |
| Station name | Station Name |
| A\_2020 | Saturday ridership for given year |
| U\_2020 | Sunday/holiday ridership for given year |
| W\_2020 | Weekday ridership for given year |
| A\_2021 | Saturday ridership for given year |
| U\_2021 | Sunday/holiday ridership for given year |
| W\_2021 | Weekday ridership for given year |
| A\_2022 | Saturday ridership for given year |
| U\_2022 | Sunday/holiday ridership for given year |
| W\_2022 | Weekday ridership for given year |
| A\_2023 | Saturday ridership for given year |
| U\_2023 | Sunday/holiday ridership for given year |
| W\_2023 | Weekday ridership for given year |