**Predicting CTA Station Usage Using Ridership, Demographic, and Socioeconomic Data**

Lucy Chavez, DePaul University, School of Computing, [lchave28@depaul.edu](mailto:lchave28@depaul.edu)

Denvir Gama, DePaul University, School of Computing, [dlnu1@depaul.edu](mailto:dlnu1@depaul.edu)

Zach Hollis, DePaul University, School of Computing, [zhollis@depaul.edu](mailto:zhollis@depaul.edu)

Danielle Martin, DePaul University, School of Computing, [dmart164@depaul.edu](mailto:dmart164@depaul.edu)

**ABSTRACT**

The Chicago Transit Authority (CTA) operates one of the nation’s largest public transportation systems. Due to the city’s demographic and socioeconomic changes over time, the efficiency and accessibility of public transportation may not be as reliable or optimal as it once was and has the potential to be. With the aim of identifying opportunities for public transportation optimization, we used several machine learning algorithms and geospatial visualizations to analyze multiple datasets.

KEYWORDS: Chicago Public Transportation, CTA, Elastic Net Regression, Neural Network, and GIS

INTRODUCTION

Our goal in this study was to utilize the publicly available CTA ridership data with detailed demographic and socioeconomic information from the U.S. 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.

LITERATURE REVIEW

We conducted a literature review focusing on the use of data science and machine learning to research public transportation trends around the world. These research papers largely focused on finding ways to improve transit systems in major cities by optimizing schedules or finding the optimum place for new routes or stations. The methods used varied considerably between different projects.

**Literature Review of Challenges and Strategic Responses in Public Transportation**

In the United States, some researchers focused on optimizing specific times when the transit system could be overwhelmed by sudden increases in riders (Santanam et al., 2024), while others focused on the availability of transportation for low-income urban areas (Griffin et al., 2016). Lee and Miller’s (2018) time geographic accessibility analysis of residents' accessibility to jobs and healthcare in an underserved neighborhood in Ohio using a high-resolution space-time accessibility measure can be used to measure the accessibility impacts of new or proposed public transit services. Also in Ohio, Wang et al. (2015) developed transport-based spatial autoregressive models to account for the spatial autocorrelation of job accessibility for three modes of transportation (walking, transit, and car) with the goal of improving job accessibility to address equity issues. Sun et al. (2019) presented research on a DDDAS-enabled (dynamic data-driven applications system) smart public transportation decision support system. The results can be used to optimize bus schedules and improve rider satisfaction.

Internationally, researchers took many different approaches. In Taiwan, researchers created a model to find the optimum site of a new metro station based on urban use data (Wey, 2015). In China, K-Means clustering was used to develop models that could predict short term subway usage to better optimize passenger flows (Dong et al., 2023), while in Germany, a study was conducted to see if the current public transportation grid would be able to easily replace commuter cars to reduce emissions (Mocanu et al., 2021). In Australia, Aston et al. (2020) addressed transit mode location biases in research on the built environment. The authors stressed the need for more nuanced geographic and contextual analyses to better understand and improve transit systems and urban development. Other researchers in Australia applied visual analytics related to the study of movement and transportation systems to better inform transportation customers and managers and make infrastructure safer and more efficient (Lock et al., 2020). In South Korea, Caliwag et al. (2022) developed a fault occurrence prediction method based on a machine learning model to predict the remaining useful life of a train subsystem, which can be used to either clear a fault before it occurs or set an alarm for an inevitable fault. The method will increase the reliability and safety of the train system. Kuo et al. (2023) discussed the challenges of designing and operating public transport systems with five performance goals: service, mobility, accessibility, responsiveness, and technology. Their paper summarized recent research on integrated planning, timetable synchronization, feeder services, and disruption management. Branda et al. (2020) developed a methodology, DA4PT (Data Analytics for Public Transport), to predict whether users will buy a ticket or not and to define different dynamic pricing strategies to increase ticket sales and revenue using data from an Italian bus company. Analyzing user-generated event logs revealed that factors like occupancy rate and ticket fare have significant influence on users’ ticket purchasing behavior.

Ahangari et al. (2020) highlighted the significant decline in public transit ridership during the COVID-19 pandemic. To address this issue, the authors proposed several measures for recovery, including enhanced cleaning protocols, financial support, and flexible operations to adapt to the changing circumstances brought about by the pandemic.

**Literature Review of Innovations and Sustainable Planning in Urban Development**

Liu et al. (2023) investigated the spatiotemporal patterns and driving mechanisms of urban expansion in China's Min Delta. Through multi-scenario simulations, the study emphasized the importance of sustainable planning, influenced by economic and policy factors, to manage urban growth effectively. Alexandre, Bernardini, and Pantoja (2023) categorized various applications of machine learning in bus transportation. Their focus was on improving service reliability and passenger information through innovations such as travel time prediction and passenger flow analysis, which are crucial for enhancing the overall efficiency of bus services. Li et al. (2019) proposed an urban transit system that uses a hybrid hub-and-spoke transit model with shared shuttles. They evaluated the performance with trip demand data and were able to reduce the average travel time and aggregate more trips. Ray et al. (2022) studied the importance of spatial data and state-level coordination to identify sites, develop policies, and prioritize transit-oriented development, which is important is building resilient and sustainable cities.

THEORETICAL DEVELOPMENT/MODEL/METHODS

**Data Sources**

We utilized several data sources to gather information on various aspects of Chicago's neighborhoods. The data sources used were:

* Census Data: We obtained census data from the American Community Survey (ACS) 5-year estimates for the years 2013-2018. The data was downloaded using the census data library in Python. The variables included in the dataset were population, median age, median household income, total housing units, median number of rooms, median year built, median gross rent, and mean household size.
* Crime Data: Crime data was obtained from the City of Chicago's Data Portal. The data included crime incidents from 2013 to 2018, with variables such as case number, year, primary type, and location.
* CTA Data: CTA (Chicago Transit Authority) data was obtained from the City of Chicago's Data Portal. The data included information on CTA stations, including their location and distance from each neighborhood.
* Sales Data: Sales data was obtained from the Cook County Assessor's Office. The data included information on property sales, including sale price, age, and location.
* Hardship Data: Data on hardship index scores was obtained from UIC Great Cities Institute/the U.S. Census Bureau. The data included information on the conditions of economic and social hardship within Chicago community areas from 2016 to 2020.

**Data Cleaning and Preprocessing**

Before analyzing the data, we performed several preprocessing steps to ensure the data was clean and consistent. The preprocessing steps included:

* Merging datasets: We merged the census, crime, CTA, and sales datasets based on the GEO\_ID variable, which represents the unique identifier for each neighborhood. We also merged the CTA dataset and the hardship index dataset based on the community area.
* Handling missing values: We handled missing values in the datasets by dropping rows with missing values or imputing them with median values.
* Data transformation: We transformed some variables, such as the sale price and age, to ensure they were in a suitable format for analysis.
* Data reshaping: We reshaped the data from long format to wide format to facilitate analysis.
* Data cleaning: We cleaned the data by removing duplicates and ensuring that the data was consistent across all datasets.

After preprocessing the data, we got a combined dataset with 400,000 rows and 32 columns. To make the analysis easier, we divided this into train and predicting datasets, and now both are ready for analysis and visualization.

To begin to develop a model that could predict CTA station usage based on demographic information we first had to combine the datasets that were downloaded from the U.S. 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, L Station Entries and L Station Locations, which contains the coordinates of the L stations. The L Station Entries dataset was the only one used for developing our models as it is the most detailed and contains the most information. The coordinates were used to match the stations to the correct census tract using the Tigris library in R.

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 because of the large change in ridership in early 2020 due to the COVID-19 pandemic, as shown in Figure 1. 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 zeros 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.

**CTA Ridership**

Exploratory Analysis

The CTA stations data was taken from the City of Chicago’s open data site. It consisted of 3 datasets: 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 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 a few interesting points. First, when graphing the trend of daily ridership over 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 stations (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 stop for DePaul. Both Midway and O’Hare stops were also in the top 20. More interestingly, 95th St. and 79th St. stations were also in the list of top stations. Both stations are on the far south of the Red Line. 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 Red Line 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.

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

Figure 1: Trend of Daily Total Rides Over Time

A graph showing a blue line

Description automatically generated

Figure 2: Top 20 Stations by Total Rides

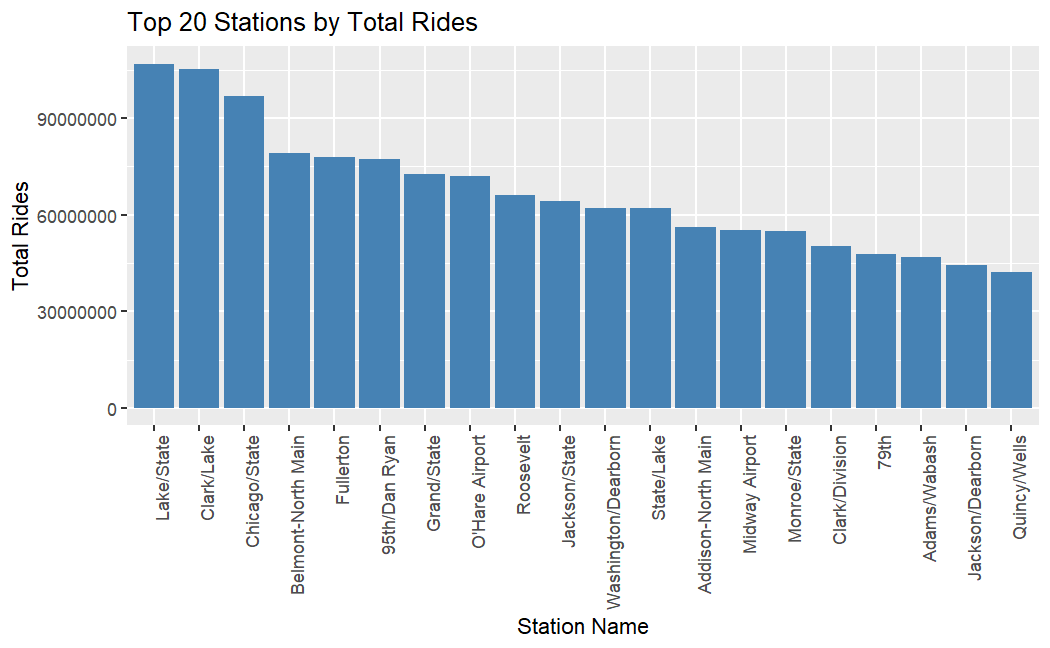


Figure 3: Top 10 Daily Rides at CTA Stations

A graph of a number of rides

Description automatically generated

Figure 4: Top Stations for Weekday Rides

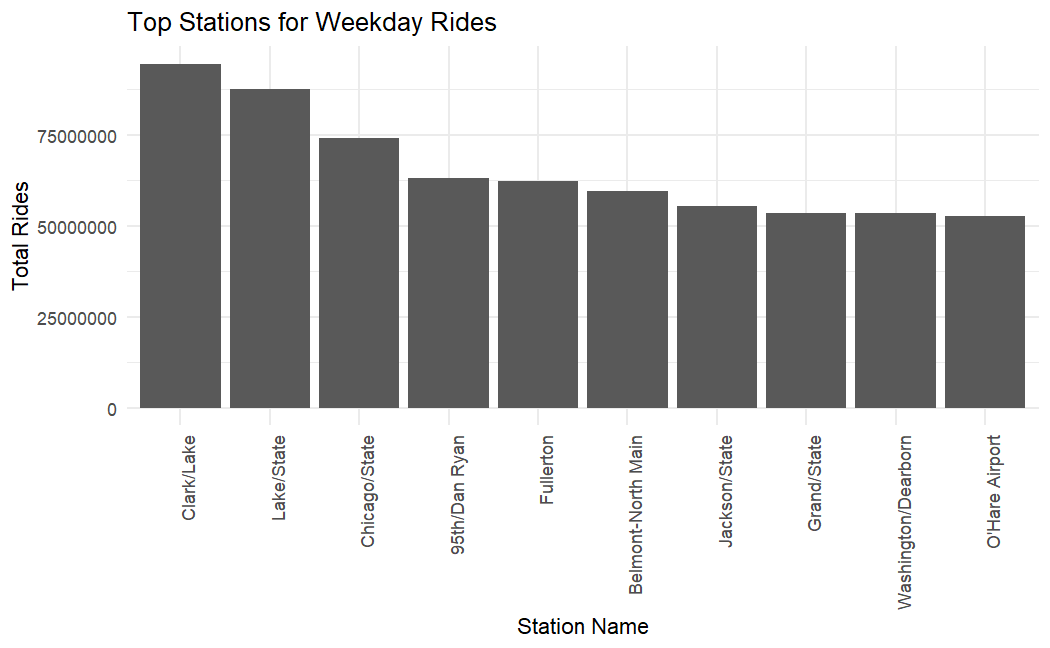
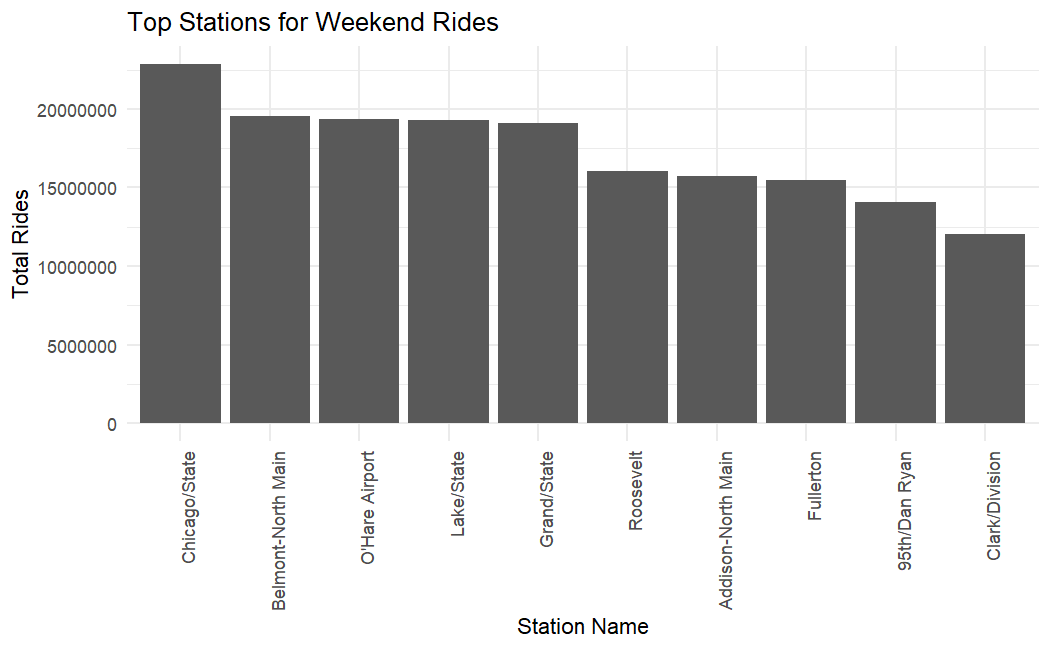
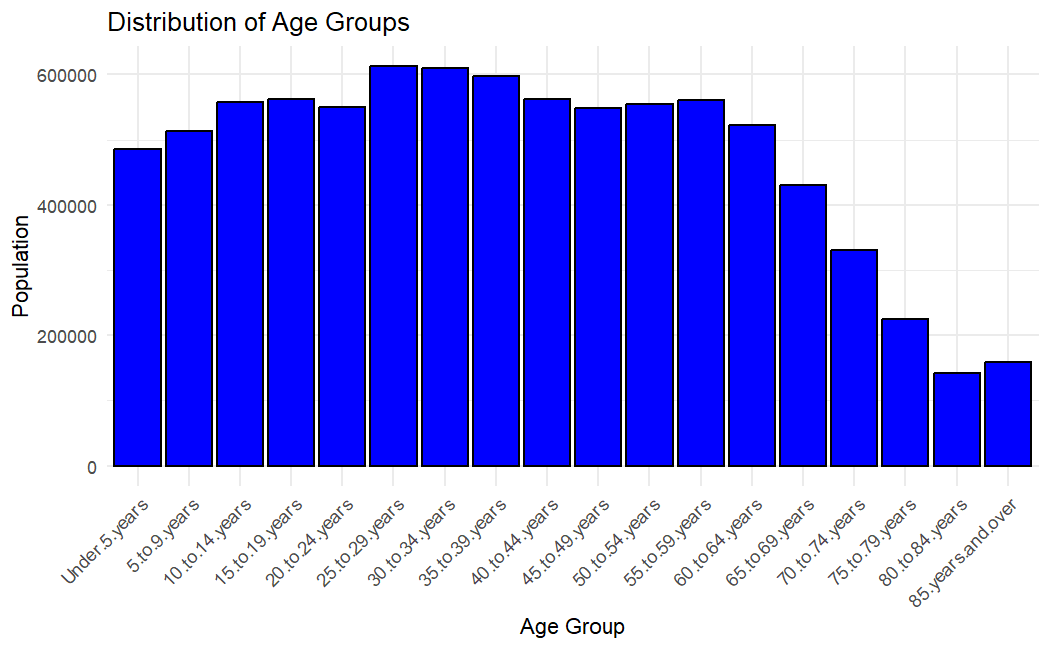
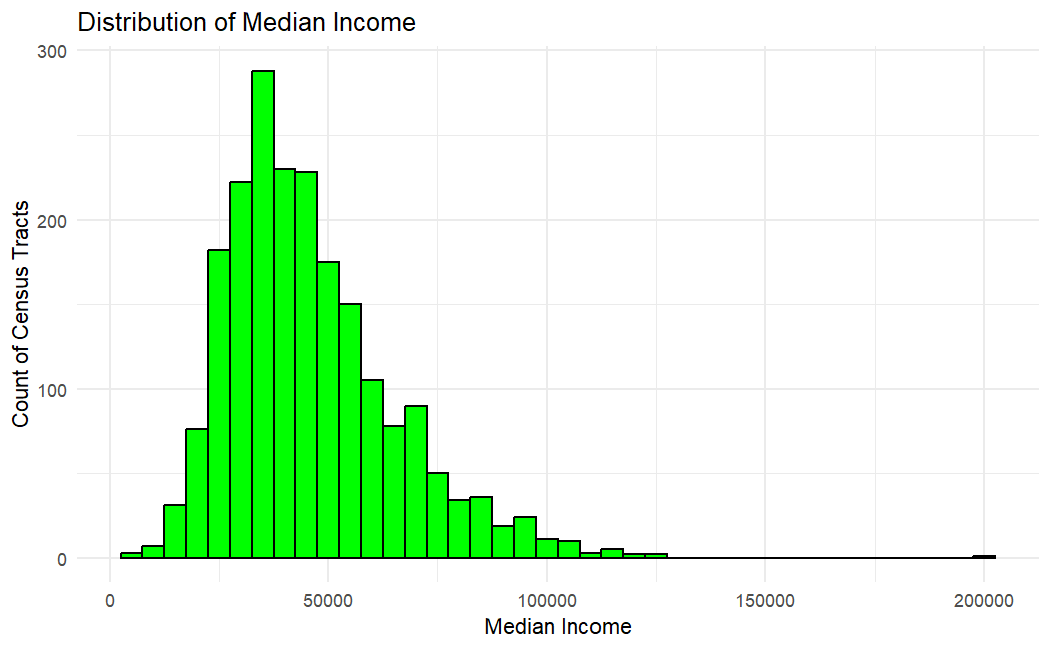


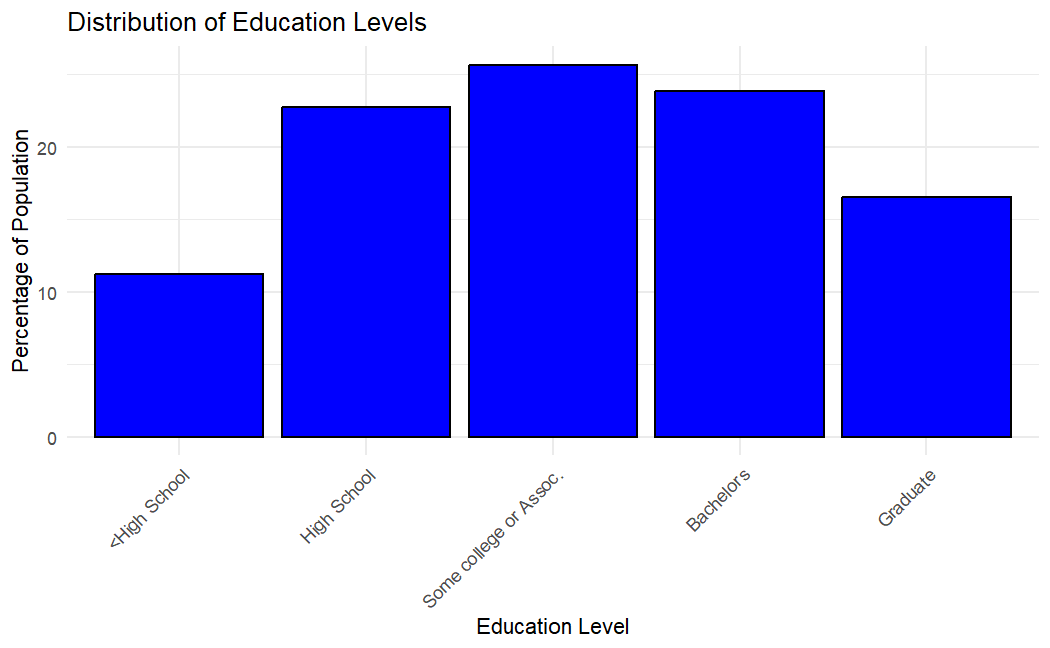
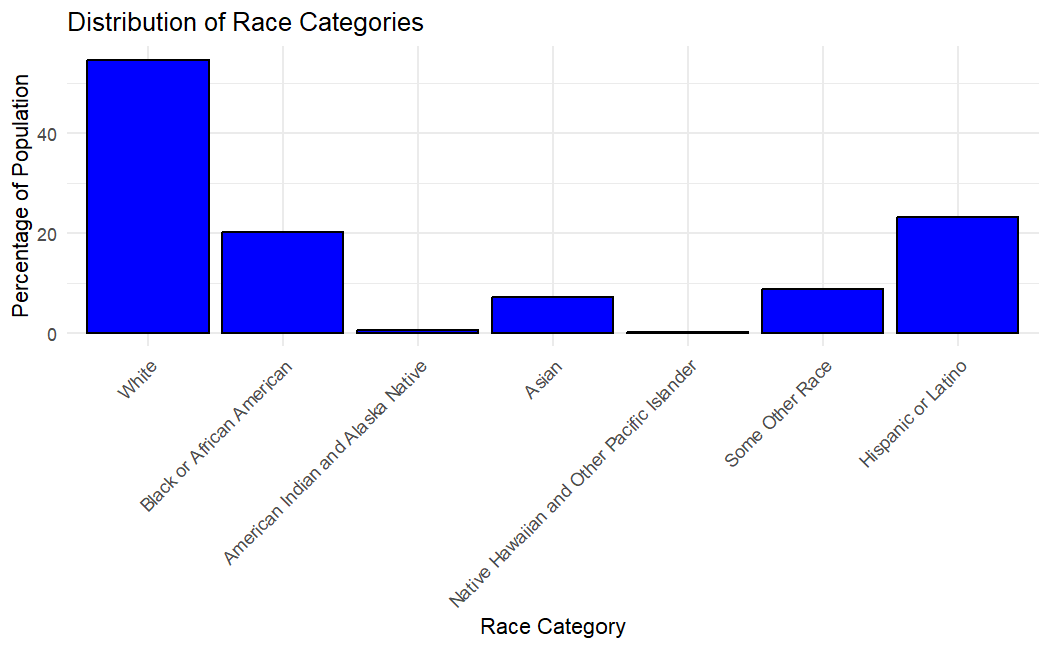
Figure 5: Top Stations for Weekend Rides

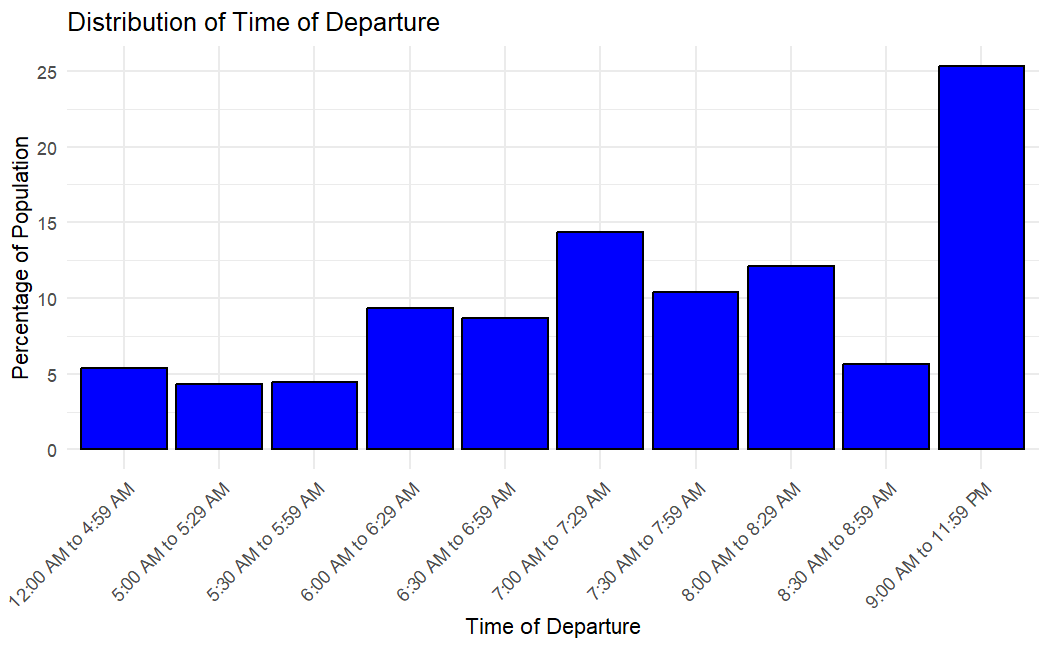
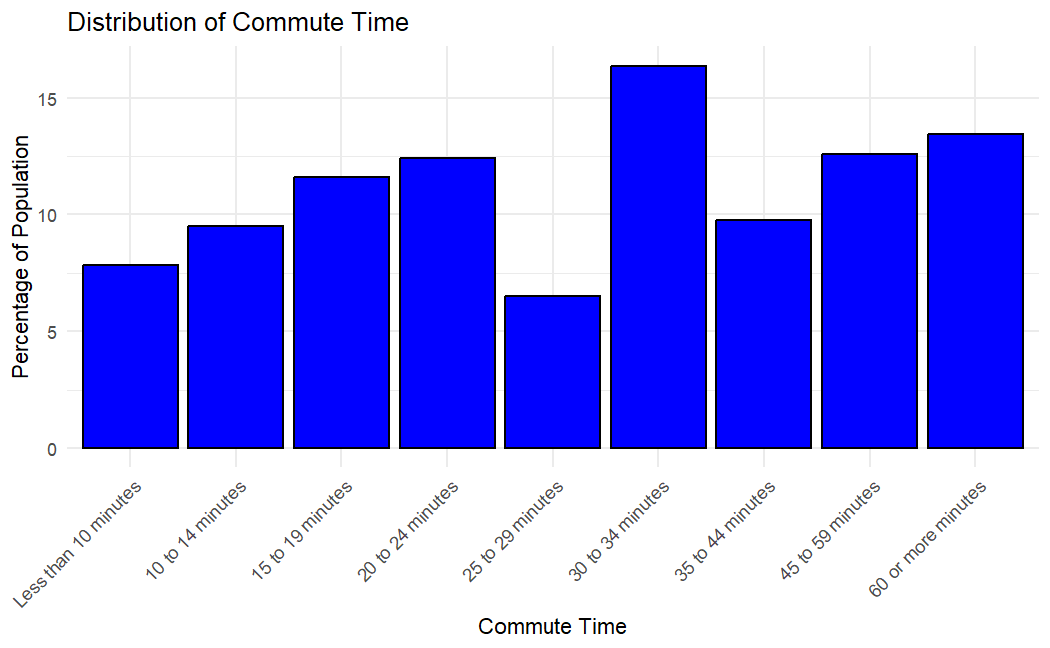


The census data we used was taken from the Census Bureau's publicly available datasets and was downloaded in several different datasets and then combined into one full dataset. The data was taken from the 10-year census as well as the American Community Survey, which collects additional demographic information. 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 Illinois). This dataset 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. (Figures 6-11)

Figures 6-11: Distribution of Age Groups, Distribution of Median Income, Distribution of Education Levels, Distribution of Race Categories, Distribution of Time of Departure, Distribution of Commute Time

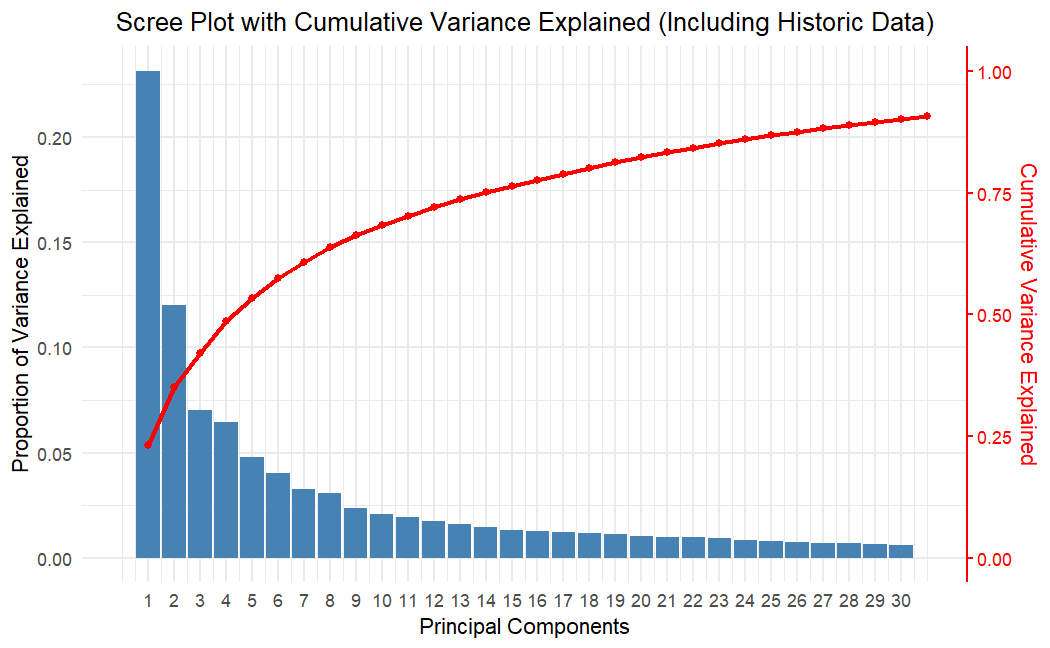
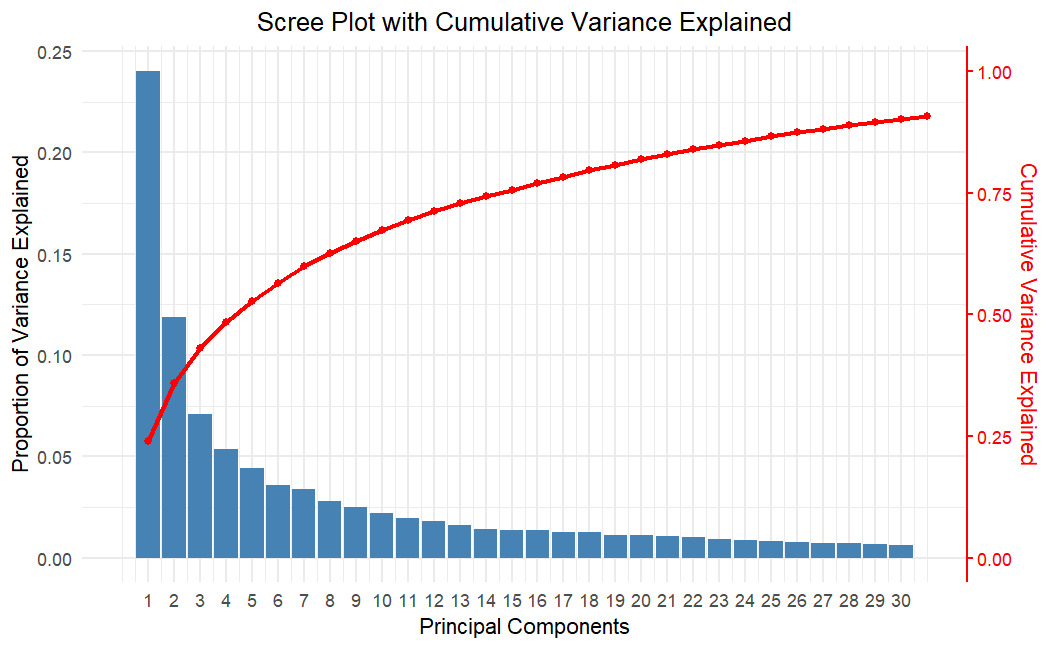
 

Model Development

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 components, 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 historical data, so the final models would include the PCs created with the historical ridership data. (Figures 12 & 13)

Figures 12 & 13: Scree Plot with Cumulative Variance Explained, Scree Plot with Cumulative Variance Explained (Including Historic 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 Model

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. (Figure 14)

Figure 14: Predicted vs Actual Values for A\_2023

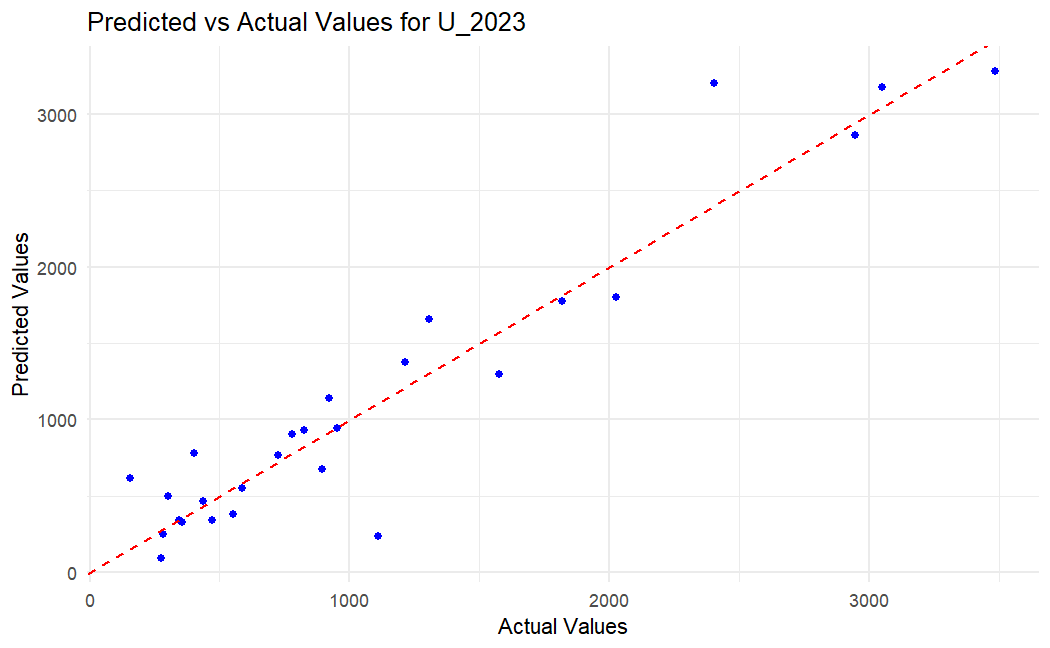
A graph with blue dots and red lines

Description automatically generated

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. (Figures 15 & 16)

Figures 15 & 16: Predicted vs Actual Values for W\_2023, Predicted vs Actual Values for U\_2023

A graph with blue dots and red lines

Description automatically generated

**Housing and Crime Data**

Exploratory Analysis

Figure 17: Variations in Crime Severity Across Chicago

A map of a city

Description automatically generated

The visualization depicted above (Figure 17) highlights distinct variations in crime severity across different areas of Chicago. Specifically, it reveals that the South Side, Downtown, and West Side exhibit noticeably higher levels of severe crimes compared to the North Side. This spatial disparity underscores significant differences in crime dynamics throughout the city, reflecting varying socioeconomic factors, population density, and possibly differing levels of law enforcement presence and community resources. Understanding these geographical patterns is crucial for local policymakers, law enforcement agencies, and community organizations aiming to implement targeted interventions and resource allocations to address crime prevention and community safety effectively.

The figure below (Figure 18) illustrates that neighborhoods on the north side of Chicago generally exhibit higher sale prices. This trend can be attributed to several factors, including lower crime rates and better access to transportation infrastructure, particularly robust transit networks. These areas benefit from a combination of factors that enhance desirability and livability, such as safer environments and convenient commuting options. This spatial disparity underscores how urban amenities and safety considerations significantly influence real estate values, highlighting the importance of location-specific factors in shaping property markets.

Figure 18: Median Sale Prices by Neighborhood

A graph showing the number of cities in the united states

Description automatically generated

From the plot below (Figure 19), it's evident that crime rates are notably higher on the south side of Chicago, contributing to lower average housing prices in those areas. Moving northward, we observe a gradual decrease in crime rates, corresponding with an increase in average housing prices. However, the downtown Loop area stands out as an exception: despite higher crime rates compared to some northern neighborhoods, it maintains high average housing prices. This anomaly suggests that other factors, such as urban amenities, job opportunities, and historical significance, play significant roles in determining property values in the downtown core.

Figure 19: Housing Prices vs. Crime Rate for Each Chicago Neighborhood (Ridge)

A graph showing the average price of a neighborhood

Description automatically generated

Model Development

The primary objective of our analysis was to understand how the introduction of a new CTA (Chicago Transit Authority) line in a neighborhood affects property pricing and crime rates in that area. This requires a comprehensive examination of various factors influencing both property values and crime incidents, necessitating the use of advanced machine learning techniques to predict these outcomes accurately.

Our problem can be framed as a regression task where the goal is to predict the average house price per census block group. This involves examining the relationships between a diverse set of features derived from demographic data, crime statistics, transportation accessibility, public school availability, housing sales attributes, and neighborhood characteristics. Specifically, we aimed to answer the following questions:

1. How does proximity to a new CTA line impact property prices in Chicago's neighborhoods?

2. What is the effect of a new CTA line on crime rates in these neighborhoods?

Model Selection

To address these questions, we employed a variety of regression models, conducting a grid search to identify the best-performing algorithms. The models used include:

Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor

These models were chosen due to their suitability for predicting quantitative outcomes with high levels of accuracy, which we measured using Root Mean Squared Error (RMSE).

Feature Selection

Our feature set includes a comprehensive array of variables to capture the multifaceted influences on property prices and crime rates:

1. Demographic Data: Total Population, Median Age, Median Household Income, Total Housing 6 Units, Median Number of Rooms, Median Year Built, Median Gross Rent, Mean Household Size, Percent White Population, Percent Black Population, Percent Household with Children, and Percent Housing Vacant.

2. Crime Data: Total Number of Crimes (all types of crime) and Total Number of Crimes per Capita (Total Number of Crimes / Total Population from ACS).

3. CTA Data: Distance to the nearest L train station (distance is calculated in miles). All distance data is calculated for the census block group level from the original data set.

4. Sales & Housing Attributes Data: Number of Sales, Average Age of House Sold, Price per Square Feet, Price per House, and Average Square Feet of House Sold. Sales data were calculated for the census block group level and were 3 years lagged. After filtering the original dataset, we focused on approximately 220,000 housing sales within the City of Chicago.

5. Neighborhood and Side Indicators: Dummy variables indicating which neighborhood (i.e., Hyde Park, Woodland, etc.) and which side (i.e., North, West, etc.) the census block is in.

Feature Engineering and Correlation Analysis

During the analysis, we examined the correlations and distributions of all our features. No additional feature transformations were found necessary beyond those initially included. We ensured that our features were relevant and adequately captured the variability needed for accurate prediction of property prices and crime rates. By integrating these diverse data sources and employing a rigorous model selection process, we aimed to create robust predictive models to inform urban planning and policy decisions regarding the impact of new CTA lines on Chicago neighborhoods.

Temporal Data Splitting

Given our objective to use past features data (i.e., 3 years lagged) to predict future average house prices per census block group, we employed a temporal splitting strategy for our train, validation, and test sets. This approach ensures that the models are trained and evaluated in a manner that mimics real-world forecasting scenarios. The splitting was done as follows:

Figure 20: Train, Validation, and Test Sets

A chart with text overlay

Description automatically generated

The years mentioned above (Figure 20) indicate the target variable years and not the feature years. For example, the "2016 Train" set includes average house prices per census block group for 2016 and features data from 2013. This lagging of feature data is crucial because the oldest available data for all features is from 2013, making 2016 the earliest year for which we can train our models to predict.

Due to this limitation, we have two validation sets and one test set. To further validate our models and ensure they do not overfit, we performed non-temporal 5-fold cross-validation. This involved splitting each training set (before 2017 and before 2018) into five random folds and reporting the average test error (measured as RMSE) of the validation fold. This cross-validation process helps corroborate the robustness of our final model.

Models and Hyperparameters Grid Search

We initiated our modeling approach with a simple Linear Regression to establish a foundational baseline, which inherently does not require tuning hyperparameters. Subsequently, our strategy involved delving into more sophisticated models where we systematically explored various sets of hyperparameters to optimize their predictive performance. These models included Ridge Regression, Lasso Regression, Elastic Net, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor. The objective of our grid search was to methodically evaluate different combinations of hyperparameters across these models to identify the optimal settings that would maximize predictive accuracy and enhance the generalizability of our results.

Our grid search encompassed essential hyperparameters such as regularization strength for Ridge and Lasso, the balance between L1 and L2 regularization for Elastic Net, maximum tree depth and features for Decision Tree and Random Forest, and boosting parameters like number of estimators, learning rate, and subsample size for Gradient Boosting. Each model underwent rigorous evaluation to determine which hyperparameter settings yielded the most promising outcomes in predicting the impact of new CTA lines on property prices and crime rates within Chicago neighborhoods.

For further insights into our grid search results and the final model selection process, detailed documentation is available in our Git repository under the "train-models-final" notebook. Our comprehensive approach, coupled with a robust feature set, aims to uncover actionable insights that illuminate the complex interplay between transportation infrastructure developments and urban dynamics, particularly focusing on housing markets and crime trends in Chicago.

Evaluation Metric and Model Selection

For evaluating the performance of our regression models, we opted for Root Mean Squared Error (RMSE) as our primary metric. RMSE is widely used in regression tasks because it provides a clear measure of prediction accuracy that aligns with the scale of the target variable. This choice allows for straightforward interpretation of how well our models predict average house prices per Census block group.

Given the limited number of temporal validate sets (2017 and 2018), we acknowledged the potential challenge of assessing model generalizability. Therefore, in addition to evaluating RMSE on these temporal sets, we also conducted non-temporal 5-fold cross-validation. This approach helped us validate the robustness of our models across different splits of the data.

Model Performance

Figures 21 & 22: RMSE Performance

A screenshot of a graph

Description automatically generated

As illustrated in the charts (Figures 21 & 22), our initial exploration revealed that the simple Linear Regression model performed inadequately, suggesting the need for more sophisticated techniques. Among the models evaluated, the Random Forest Regressor consistently demonstrated superior performance based on RMSE. Consequently, we selected it as our final model class for its robust predictive capabilities.

Through a rigorous grid search process, we identified optimal hyperparameter configurations for the Random Forest Regressor. The top-performing sets of hyperparameters were determined based on their average rank of temporal validate RMSEs across 2017 and 2018 datasets. This approach ensured that our chosen configurations were consistently effective across different temporal splits.

The table below summarizes the top three sets of hyperparameters selected based on their average temporal validate RMSE ranks:

Table 1: Hyperparameters

A screenshot of a computer

Description automatically generated

These configurations not only performed well in temporal validation but also showed promising results in non-temporal 5-fold cross-validation, with RMSEs within a negligible margin of the best obtained results. This consistency across different validation methods provided confidence in the robustness of our chosen hyperparameters.

Final Model Selection

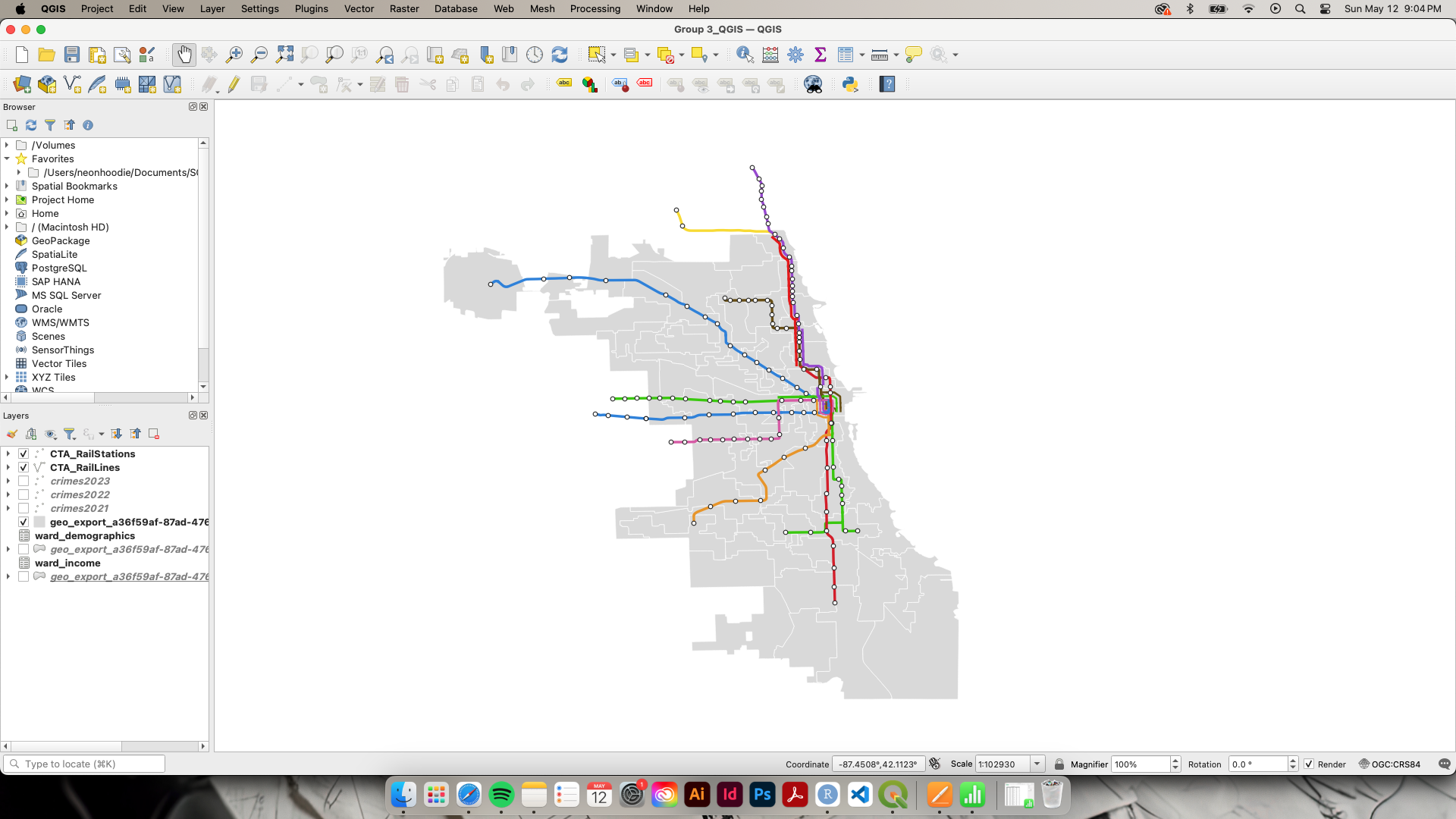
Considering concerns about potential overfitting, we opted for the set of hyperparameters that offered a balance of complexity and performance: {'max\_depth': 40, 'max\_features': 'sqrt', 'min\_samples\_split': 2, 'n\_estimators': 100}. This decision was reinforced by observing that these parameters achieved competitive RMSEs on the temporal test set (2019 data) and maintained favorable rankings across various evaluation metrics.

**QGIS Geospatial Visualizations**

Exploratory Analysis

To visualize various dataset in QGIS, a map of Chicago separated by ward was created, overlayed with the CTA train lines map to serve as a geographic framework.

Figure 23: CTA Train Lines in Chicago Wards



The data for this project was sourced from the City of Chicago and included CTA ridership statistics, crimes rates, demographic distributions, and income levels across different wards. By importing these datasets into QGIS, we were able to geospatially represent the information, allowing for a visual analysis of patterns and correlations. Layers were created for each dataset and appropriate symbology and color schemes were applied to differentiate the data visually. The goal was to utilize these visualizations to gain insights into how these variables might inform transit planning and infrastructure development.

Below are the different variables explored, utilizing the map of Chicago and CTA train line overlay as the foundation:

Figure 24: Ridership

Ridership

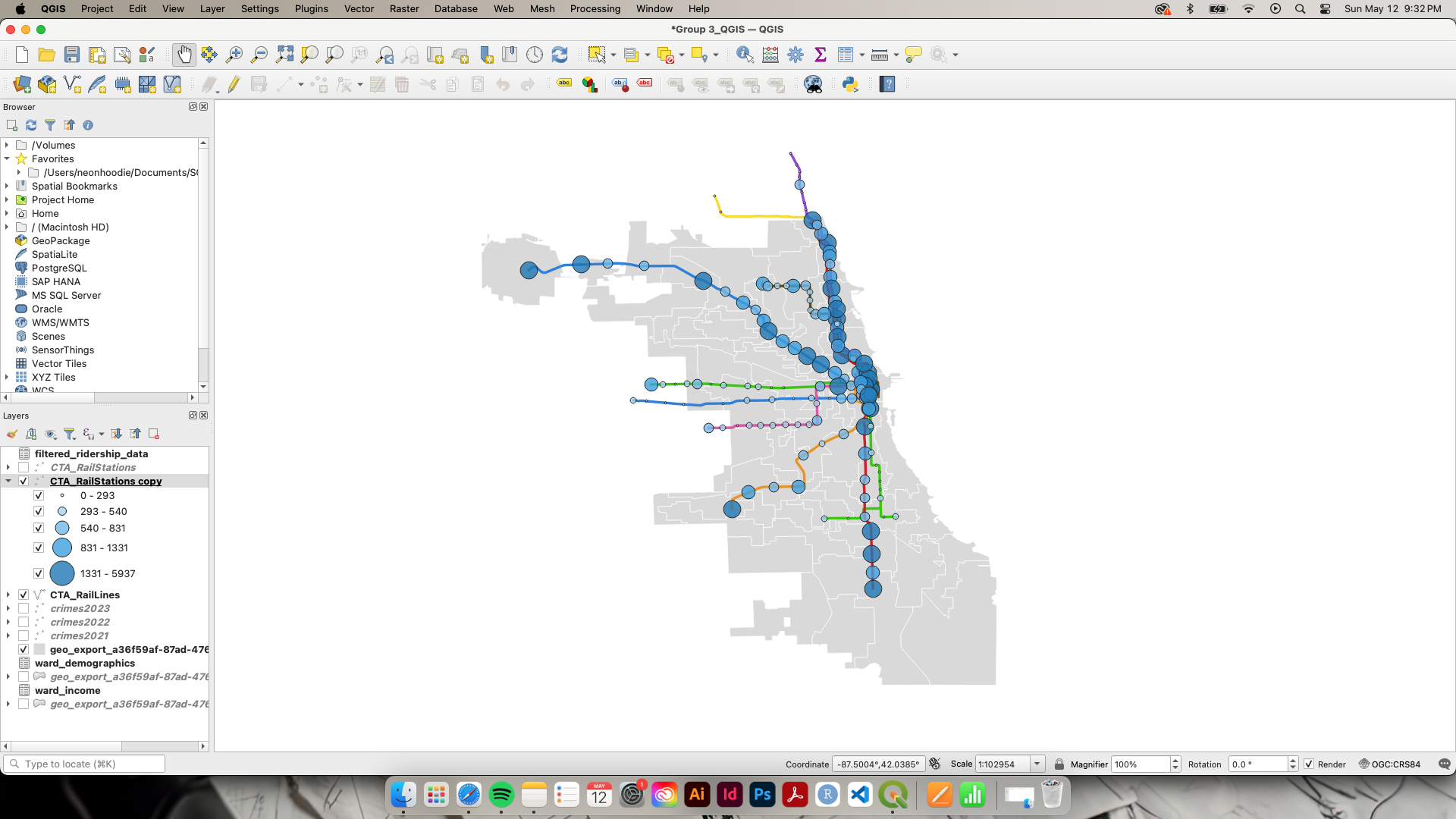


Figure 25: Crime

Crime

A map of a city

Description automatically generated

Figure 26: Population by Race/Ethnicity

A map of the population

Description automatically generated with medium confidence

Figure 27: Population by Income

A map of a city

Description automatically generated with medium confidence

**Hardship Index Score Data**

Exploratory Analysis

The hardship index is used to examine the conditions of economic hardship, or the difficulty resulting from not having enough collective economic resources available, within Chicago community areas. The score is the average of six standardized variables: Percent of Crowded Housing (housing units with more than one person per room), Percent of Households with Income Below Poverty Level (below the federal poverty level), Unemployment Rate for Population Age 16 and Over, Percent Aged 25 and Over with No High School Diploma, Percent of Population Under Age 18 and Over Age 64 (dependency), and Per Capita Income.

There are 77 community areas within Chicago. For 2016 to 2020, the hardship index score ranged from 9.4 to 76.5 with a mean of 44.66. The five community areas with the highest hardship index scores were Riverdale (76.5), Fuller Park (75.1), West Garfield Park (71.6), Englewood (68.2), and New City (67.0) The five community areas with the lowest hardship index scores were Near North Side (9.4), Near South Side (10.2), Lincoln Park (11.4), Loop (12.9), and Lakeview (13.5).

Table 2: Hardship Index Score Data Sample

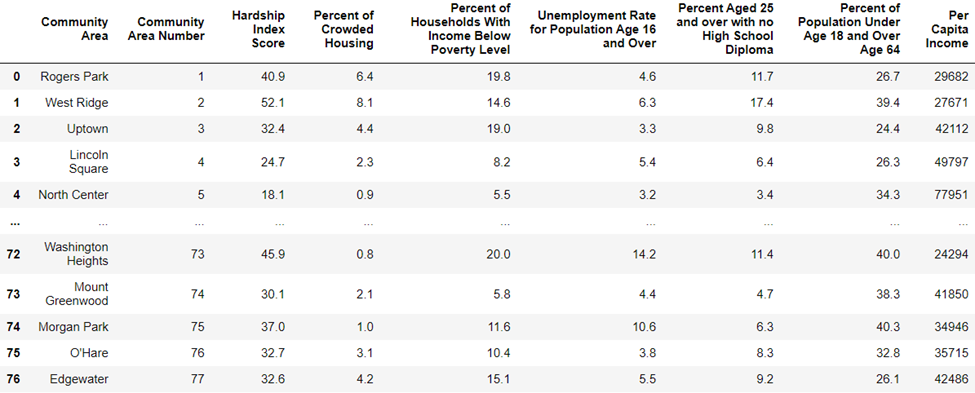
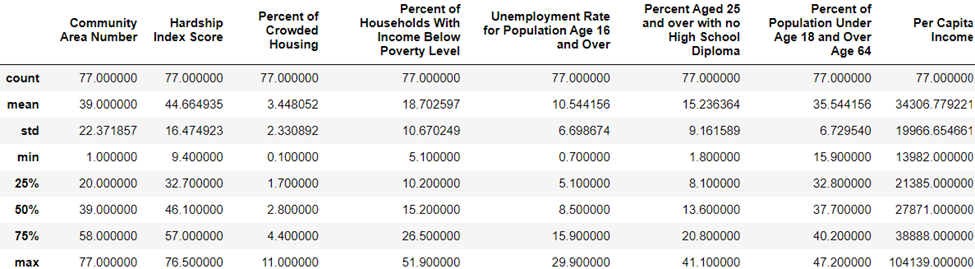


Table 3: Hardship Index Score Data Summarization



We used these variables to create a correlation heatmap (Figure 28). It makes sense that Hardship Index Score is highly correlated with the other variables because the score is created using those variables. Per Capita Income is almost perfectly correlated with Hardship Index Score. The higher a person's income, the lower their hardship index score will be. Percent of Households With Income Below Poverty Level and Unemployment Rate for Population Age 16 and Over have the next strongest correlation. Again, this makes sense. If a person is unemployed and doesn't have an income, their household is more likely to have an income below the poverty level. The least correlated variables are Unemployment Rate for Population Age 16 and Over and Percent of Crowded Housing.

Figure 28: Correlation Heatmap Using Hardship Index Score Variables

*A screenshot of a computer screen

Description automatically generated*

**Hardship Index Score Data and CTA Ridership Data**

Exploratory Analysis

After combining the two datasets (CTA\_HIS), we wanted to determine if there was a correlation between the number of stations in a community area (representing access to public transportation) and the hardship index score. We used the variables STATION\_DESCRIPTIVE\_NAME and stationname from the CTA dataset and information on the train lines to determine the community areas where each station was located.

Six of the stations were located on the borders of two community areas. We updated the dataset to include these stations twice, once for each of the community areas. This means that there were 129 stations instead of 123 and that 43 community areas had stations instead of 41. We thought it was more important to have the data accurately reflect all the community areas with stations than the exact total number of stations.

Of the 77 community areas, 43 had CTA train stations.

Figure 29: Community Areas with CTA Train Stations

A list of cities with black text

Description automatically generated   A list of cities with black text

Description automatically generated with medium confidence

We created a simplified dataset (CTA\_HIS\_simple) with the original variables that make up the hardship index score plus the number of stations in each community area. The Loop had the most stations (17). The mean number of stations per community area was 1.67.

Table 4: Hardship Index Score and Number of Stations Data Sample

A screenshot of a computer

Description automatically generated

Table 5: Hardship Index Score and Number of Stations Data Summarization

A table with numbers and a number of objects

Description automatically generated

We created another simplified dataset (CTA\_HIS\_updated) using CTA\_HIS plus a column with the number of stations and columns from the CTA dataset related to race, education, income, and population. In addition, we included the data on the stations and the associated hardship index score data.

The goal was to expand on the hardship index data and get a better understanding of how these variables (race, education, income, and population) relate to the community areas and the availability of public transportation.

Model Development

We split the data into training and test sets with a test size of 0.25. We created a linear regression model and a ridge regression model to predict Number of Stations. The dependent variable was made up of STATION\_DESCRIPTIVE\_NAME, stationname, Community Area, Other Community Area, Number of Stations, Community Area Number, Hardship Index Score, Percent of Crowded Housing, Percent of Households With Income Below Poverty Level, Unemployment Rate for Population Age 16 and Over, Percent Aged 25 and over with no High School Diploma, Percent of Population Under Age 18 and Over Age 64, Per Capita Income, RACE.White, RACE.Black.or.African.American, RACE.American.Indian.and.Alaska.Native, RACE.Asian, RACE.Native.Hawaiian.and.Other.Pacific.Islander, RACE.Some.other.race, RACE.Hispanic.or.Latino.origin..of.any.race., EDUCATION.Less.than.high.school.graduate, EDUCATION.Less.than.high.school.graduate, EDUCATION.Some.college.or.associate.s.degree, EDUCATION.Bachelor.s.degree, EDUCATION.Graduate.or.professional.degree, Median.income..dollars., Total.population, Median.age..years., Sex.ratio..males.per.100.females., Male.Total.population, Male.Median.age..years., Female.Total.population, and Female.Median.age..years.

RESULTS/FINDINGS

**CTA Ridership Data**

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:

Table 6: Model Results

|  |  |
| --- | --- |
| 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:

Table 7: Predicted Average Daily Riders

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 |

Housing and Crime Data

Based on the feature importance scores obtained from our model, it is evident that certain variables significantly influence the average house prices per Census block group in Chicago neighborhoods. Notably, variables such as 'prev\_year\_price\_p\_house', 'prev\_year\_price\_p\_sf', 'Median HH Income', 'Percent White Population', and 'Median Gross Rent' emerged as key determinants. These features underscore the socioeconomic dynamics and housing market characteristics that play pivotal roles in shaping property values across different neighborhoods.

Moreover, the inclusion of 'Norm distance\_miles' among the influential factors suggests that proximity to CTA lines indeed impacts property pricing and crime rates. As anticipated, our analysis revealed that areas closer to new CTA lines experience observable changes in both housing prices and localized crime rates. This finding aligns with our initial hypothesis, indicating that transportation infrastructure developments can have significant socio-economic implications on urban neighborhoods.

Table 8

A table with numbers and letters

Description automatically generated

Figure 30: Top 15 Important Features

A graph of a number of people

Description automatically generated with medium confidence

Based on the feature importance scores obtained from our model, it is evident that certain variables significantly influence the average house prices per Census block group in Chicago neighborhoods. Notably, variables such as 'prev\_year\_price\_p\_house', 'prev\_year\_price\_p\_sf', 'Median HH Income', 'Percent White Population', and 'Median Gross Rent' emerged as key determinants. These features underscore the socioeconomic dynamics and housing market characteristics that play pivotal roles in shaping property values across different neighborhoods.

Moreover, the inclusion of 'Norm distance\_miles' among the influential factors suggests that proximity to CTA lines indeed impacts property pricing and crime rates. As anticipated, our analysis revealed that areas closer to new CTA lines experience observable changes in both housing prices and localized crime rates. This finding aligns with our initial hypothesis, indicating that transportation infrastructure developments can have significant socio-economic implications on urban neighborhoods.

QGIS Geospatial Visualizations

Upon analyzing the visualized data, several key insights emerged. Crime data, despite its detail, did not reveal significant correlations with the CTA transit lines, suggesting that crime rates might not directly influence transit usage patterns. However, the ridership data provided valuable information, highlighting which stations and stops experience the highest traffic. This information is crucial for understanding peak usage areas and potential needs for service adjustments or enhancements.

The most significant findings came from comparing population data by income and demographic groups. The visualizations showed that white residents were more evenly distributed throughout Chicago, while Hispanic and Black communities were more concentrated in specific geographic areas. By combining this demographic information with income data, we could see the socioeconomic disparities across different neighborhoods. This dual-lens analysis provided a clearer picture of where different demographic groups reside, their economic conditions, and how these factors might influence their use of public transit. Understanding these patterns is essential for developing targeted transit policies that address the needs of diverse populations and promote equitable access to transportation.

Having this information and gaining the demographic and monetary understanding of the groups will greatly aid in predicting needs and ridership for new station additions. By leveraging the geographic knowledge obtained from the visualizations, new stations can be strategically placed in areas with high potential ridership or a demonstrated need for transit access, taking into account the socioeconomic characteristics of the surrounding communities. This would ensure that future transit expansions are not only efficient but also inclusive.

Hardship Index Scores

The models performed much better on the training data than the test data. In order to better accurately predict the number of stations in a community area, we could try different models that have more hyperparameter tuning involved. We could also try using more or different variables from the original CTA dataset. We think these R^2 values (0.69 and 0.71) show that there is potential here. We also acknowledge that there are many other factors that could affect the number of stations in an area, such as access to other public transportation options.

Figure 31: Linear Regression and Ridge Regression Model Results





OPPORTUNITIES DISCUSSED AND CLOSING REMARKS

**Efforts Continued**

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.

**AUTHOR CONTRIBUTIONS**

Lucy Chavez: Data cleaning and exploratory analysis CTA rail line data, QGIS visualizations and analyses, transit and urban development research

Denvir Gama: Analyzing the impact of introducing a new CTA line on housing prices in the neighborhood and its effect on crime rates.

Zach Hollis: Data cleaning and exploratory analysis for CTA and Census data, development of Elastic Net, SVM models and predictions for new Red Line Stations

Danielle Martin: Exploratory analysis and model development for hardship index data

REFERENCES

Ahangari, S., Chavis, C., & Jeihani, M. (2020). Public transit ridership analysis during the COVID-19 pandemic. *medRxiv*. <https://doi.org/10.1101/2020.10.25.20219105>

Alexandre, T., Bernardini, F., & Pantoja, C. E. (2023). Machine Learning Applied to Public Transportation by Bus: A Systematic Literature Review. *Transportation Research Record, 2677*(7), 639-660. <https://doi.org/10.1177/03611981231155189>

Aston, L., Currie, G., Kamruzzaman, M., Delbosc, A., Fournier, N., & Teller, D. (2020). Addressing transit mode location bias in built environment-transit mode use research. *Journal of Transport Geography, 87,* 102786. <https://doi.org/10.1016/j.jtrangeo.2020.102786>

Branda, F., Marozzo, F., & Talia, D. (2020). Ticket sales prediction and dynamic pricing strategies in public transport. *Big Data and Cognitive Computing, 4*(4), 36-. <https://doi.org/10.3390/bdcc4040036>

Caliwag, A., Han, S.-Y., Park, K.-J., & Lim, W. (2022). Deep-Learning-Based Fault Occurrence Prediction of Public Trains in South Korea. *Transportation Research Record*, *2676*(4), 710–718. <https://doi.org/10.1177/03611981211064893>

Dong, N., Li, T., Liu, T., Tu, R., Lin, F., Liu, H., & Bo, Y. (2023). A method for short-term passenger flow prediction in urban rail transit based on deep learning. *Multimedia Tools and Applications.* <https://link.springer.com/article/10.1007/s11042-023-14388-z>

Griffin, G. P., & Sener, I. N. (2016). Public Transit Equity Analysis at Metropolitan and Local Scales: A Focus on Nine Large Cities in the US. *Journal of Public Transportation, 19*(4)*,* 126-143. <https://doi.org/10.5038/2375-0901.19.4.8>

Kuo, Y.-H., Leung, J. M. Y., & Yan, Y. (2023). Public transport for smart cities: Recent innovations and future challenges. *European Journal of Operational Research, 306*(3)*,* 1001-1026. <https://doi.org/10.1016/j.ejor.2022.06.057>

Lee, J., & Miller, H. J. (2018). Measuring the impacts of new public transit services on space-time accessibility: An analysis of transit system redesign and new bus rapid transit in Columbus, Ohio, USA. *Applied Geography, 93,* 47-63. <https://doi.org/10.1016/j.apgeog.2018.02.012>

Li, Y., Liu, G., Zhang, Z.-L., Luo, J., & Zhang, F. (2019). CityLines: Designing Hybrid Hub-and-Spoke Transit System with Urban Big Data. *IEEE Transactions on Big Data, 5*(4)*,* 576-587. <https://doi.org/10.1109/TBDATA.2018.2840222>

Liu, X., Li, Y., Zhang, S., & Niu, Q. (2023). Spatiotemporal patterns, driving mechanism, and multi-scenario simulation of urban expansion in Min Delta Region, China. *Ecological Indicators, 158.* <https://doi.org/10.1016/j.ecolind.2023.111312>

Lock, Oliver, et al. “The visual analytics of big, open public transport data – A framework and pipeline for Monitoring System Performance in greater sydney.” *Big Earth Data*, vol. 5, no. 1, 8 July 2020, pp. 134–159. <https://doi.org/10.1080/20964471.2020.1758537>.

Mocanu, T., Joshi, J., & Winkler, C. (2021). A data-driven analysis of the potential of public transport for German commuters using accessibility indicators. *European Transport Research Review, 13*.<https://doi.org/10.1186/s12544-021-00507-0>

Ray, R. S., Garrick, N., & Atkinson-Palombo, C. (2022). Transit-Oriented Data: The Importance of Data and Coordination to Transit-Oriented Urban Transformation. *Frontiers in Sustainable Cities, 4.* <https://doi.org/10.3389/frsc.2022.869532>

Santanam, T., Trasatti, A., Hentenryck, P. V., & Zhang, H. (2024). Public Transit for Special Events: Ridership Prediction and Train Scheduling. *IEEE Transactions on Intelligent Transportation Systems*, 1-17.<https://doi.org/10.1109/tits.2024.3373634>

Sathyan, Sigma, and Jagadeesha S.N. “Traffic Flow Prediction Using Machine Learning Techniques - A Systematic Literature Review.” *ResearchGate*, 2022.

SMU City Perspectives Team. “Using Real-Time Data Analytics to Make Public Transportation Smarter: City Perspectives.” *Using Real-Time Data Analytics to Make Public Transportation Smarter | City Perspectives*, cityperspectives.smu.edu.sg/article/using-real-time-data-analytics-make-public-transportation-smarter-0. Accessed 2 June 2024.

Sun, F., Dubey, A., White, J., & Gokhale, A. (2019). Transit-Hub: a smart public transportation decision support system with multi-timescale analytical services. *Cluster Computing*,22, 2239-2254. <https://doi.org/10.1007/s10586-018-1708-z>

Wang, C-H., & Chen, N. (2015). A GIS-based spatial statistical approach to modeling job accessibility by transportation mode: case study of Columbus, Ohio. *Journal of Transport Geography, 45,* 1-11. <https://doi.org/10.1016/j.jtrangeo.2015.03.015>

Wey, W.-M. (2015). Smart growth and transit-oriented development planning in site selection for a new Metro Transit station in Taipei, Taiwan.” *Habitat International*, *47,* 158–168. <https://doi.org/10.1016/j.habitatint.2015.01.020>