Chicago Public Transport Analysis Report

Members (Group 2 - CS 418, Spring 2025):

- Srijita Banerjee (Project Lead, sbane4@uic.edu)
- Amina Marin (amari22@uic.edu)
- Heer Patel (hpate328@uic.edu)
- Prince Sonani (psona2@uic.edu)
- Burak Simsek (bsims7@uic.edu)

Project Repository Link: https://github.com/p-rinceS/CS-418-Project

Introduction and Research Problem:

In this project, we investigated how external factors—including weather, crime, and public events—affect the Chicago Transit Authority (CTA) ridership patterns. Our goals were to build predictive tools for demand forecasting, identify correlations between security and ridership, and generate actionable insights for equitable and secure transit planning.

We hypothesized that extreme weather and public events would lead to spikes in ridership, while crime and safety concerns would lead to declines.

Data Sources and Preprocessing

We collected and cleaned data from the following sources:

- **CTA Ridership Data** (2001–2024): Daily and monthly ridership counts.
- **Weather Data** (NOAA): Daily weather indicators, including precipitation and temperature.
- **Chicago Events Calendar**: Event listings and public holidays.
- **Crime Reports** (Chicago Data Portal): Filtered for CTA-related incidents from 2010–2024.
- **Reddit Sentiment Data**: Extracted and cleaned from Reddit comments discussing CTA safety.

Data preprocessing included the following steps:

- Standardizing date formats across all datasets.
- Filtering for CTA-specific records in the crime and sentiment data.
- Normalizing weather indicators and event frequency by date.
- Merging datasets by day for aligned modeling and visualization.

Machine Learning and Statistical Analysis

Each team member explored one or more techniques to generate insights:

Burak Simsek

Data Preparation and Methods:

I worked with several datasets covering the years **2019 to 2023 (source: Chicago Health Atlas)**, including:

- **CTA Station & Line Data**: Geographic coordinates and line membership for all L stops.
- **Community Income & Demographics**: Per capita income, drive-alone-to-work rates, and public perception of safety.
- **Security Spending Records**: Official CTA vendor payments filtered through keyword matching to isolate safety-related expenditures.

Techniques Used:

- Spatial joins between station data and community polygons
- Folium for layered interactive maps
- Branca colormaps to represent socioeconomic variables
- `pandas` and matplotlib for time-series and categorical expenditure visualizations

Insights from Visualizations:

1. **Transit & Income**
   High-income neighborhoods tend to cluster around multiple L lines (e.g., the Loop, Lincoln Park), while underserved low-income regions such as West Garfield Park and Englewood show fewer transit nodes. This suggests geographic inequity in transit infrastructure.

2. **Transit & Perceived Safety**
   Areas with **lower safety perception scores** often lack transit density, while **higher-trust areas** (e.g., Lakeview, Hyde Park) overlap with strong L line connectivity. This visualized correlation implies a link between mobility access and feelings of security or community trust.

3. **Drive-Alone Rates vs Transit Access**
   Car dependency was **visibly higher** in neighborhoods with fewer CTA stations (e.g., far southwest and northwest). These areas, shaded in red on the map, contrast with green-shaded

transit-rich zones where public transportation seems more viable.

4. **Security Expenditure Trends**
   Between 2019–2023, CTA spent increasing amounts annually on safety-related services. Vendors like **Allpoints Security** and **Digby's Detective Agency** received the highest cumulative payments. However, this investment was **not always aligned** with neighborhoods exhibiting the lowest perceived safety or highest crime rates—highlighting potential mismatches in budget allocation.

Conclusion

Through these layered visualizations, I observed that transit access, community safety perception, and car dependency are all spatially patterned and influenced by underlying income disparities. Mapping these variables side by side provided a more holistic view of urban inequality.

These tools can:

- **Guide equity-focused policy decisions**
- **Inform future CTA station planning**
- **Support more data-driven budget transparency**

Ultimately, this work reaffirms the value of geographic data in uncovering unseen urban gaps, and offers a framework to ensure smarter and more just investment in public infrastructure.

Amina Marin

I decided to perform a statistical analysis technique. I performed the Chi-Squared Test of Independence to assess whether crime type is related to location type.

Data Sources & Cleaning

The data I used was collected from annual crime datasets covering the years 2019 to 2024. I merged all 6 yearly datasets into a single DataFrame using pd.concat(). Important columns I used were 'Primary Type', and 'Location Description'.

Steps Taken

- I used pd.crosstab() to create a contingency table of 'Primary Type' vs 'Location Description'
- I imported chi2_contingency() from scipy.stats to run the Chi-Square Test of Independence
- The resulting values where:

- Chi-square statistic: 30033.22148366497
- P-Value: 0.0
- Degrees of Freedom: 319

Exploratory Data Analysis

I used value_counts() and grouby() to identify the most common crime types and top locations where crimes occur. For example, as shown in one of my visualizations, theft, battery, and robbery were the most frequent crime types in the years 2019-2024. As shown in another visualization, the CTA bus, bus stop, platform, station, and train were the locations with the highest amount of crimes in the year 2019-2024.

Correlation & Trend Influence

The Chi-square test assessed categorical correlation, specifically whether crime types were independent of location types. The p-value resulting in 0.0 shows a strong association between the two variables. This implies that crime trends are not evenly distributed and that some crimes are clearly location-dependent.

Key Observations

Crime type and location are not independent, there are certain crimes that are more prevalent in specific locations. These patterns held consistently across all six years of data (2019-2024). I can conclude that ‘Primary Type' and 'Location Description' are not independent; certain types of crimes are significantly more likely to occur in specific types of locations. This insight can inform more focused crime prevention efforts and improve how law enforcement resources are deployed.

Heer Patel

Section: CTA Increase in Police Officers Hires

Dataset Used

For my analysis, I utilized the CTA Employee Salary Dataset (2025). This dataset contains detailed salary information, including base pay, overtime, and department affiliations. I focused on identifying salary trends and departmental priorities, particularly regarding CTA’s investments in transit policing.

Steps Taken

- Began by cleaning the dataset by removing missing or irrelevant rows.
- Then filtered the data to focus only on the top 10 highest paid departments.

- Used grouping techniques to calculate total salaries and overtime amounts by department.
- After that, created visualizations to compare regular and overtime earnings across these departments.
- Specifically highlighted the Chicago Police Department (CPD) in the visuals.

Techniques Used

- Used pandas to group and sort the data.
- Used matplotlib and seaborn to create bar plots for better comparison.\
- Separated regular pay and overtime pay to show how much each department earned from both.

Analysis

Using Python and pandas, I cleaned and filtered the dataset to isolate the top 10 highest-paid departments within the CTA. I created visualizations to compare regular earnings and overtime compensation across departments. In particular, I emphasized the Chicago Police Department (CPD), which has a dedicated transit unit funded through the CTA budget.

Findings

- CPD was consistently the highest paid department among all CTA departments in 2025.
- A significant portion of CPD's earnings came from overtime, indicating increased patrol shifts and response to security concerns.
- Visuals showed a clear spike in police compensation aligning with broader safety investments, suggesting increased reliance on enforcement-based strategies within the transit system.

Implications:

This increase in police officer hiring and pay reflects CTA's growing prioritization of security and public safety. When considered alongside ridership trends and sentiment analysis, our results suggest that visible investment in policing may be one strategy to rebuild rider trust post-2020. However, the financial emphasis on CPD also raises questions about the balance between enforcement and community-based safety efforts.

Prince Sonani

Before analyzing crime data related to CTA infrastructure, we first investigated public sentiment regarding safety at CTA stations. To capture how people perceive and discuss security on the CTA, we collected user comments from public forums, specifically Reddit, spanning the years 2020 to 2025. These comments provide valuable context by reflecting public expectations, concerns, and evolving

attitudes toward CTA safety over time. We used web scraping techniques to extract this data and analyzed it to uncover recurring themes and sentiment trends.

Data Sources & Cleaning:

Sentiment Data: Comments from Reddit users on CTA & Chicago related subreddits. Cleaned by removing stop words, and adding a word limit so that comments with 1-2 words will not be considered for positive or negative sentiment.

Steps Taken

- Web Scraped data via Beautiful Soup and json parsing libraries.
- Cleaned the data
    - Lowercase all text
    - Removed stop words
    - Removed special characters
- Used NLTK's sentiment analysis to determine positive and negative comments
- Created insightful visualizations using the "word cloud" library & pandas.

Exploratory Data Analysis:

- Tools used: *Pandas, NLTK, BeautifulSoup, json*
- We conducted exploratory analysis to understand the overall distribution and themes in public sentiment related to CTA safety.
    - **Word Clouds:** Created separate word clouds for negative comments to highlight the most frequent terms in the sentiment category. Negative words featured "unsafe," "homeless," "drugs," and "creepy."
    - **Bar Chart:** Used to show the proportion of positive vs. negative sentiment in the dataset. Results indicated a majority of comments were generally positive, but negative themes were more concentrated and emotionally charged.

Correlation & Trend Influence:

Despite a high number of positive general impressions about the CTA, negative sentiment frequently clustered around three recurring concerns: visible homelessness, violations of social etiquette, and drug use on or near CTA infrastructure. These findings suggest that while overall public opinion may remain favorable, underlying safety and social condition concerns persist.

Key Observations:

Among all negatively classified comments, 42% referenced homelessness, 31% referenced drug use, and 25% referenced behavior or etiquette concerns.

Srijita Banerjee

My analysis focused on investigating whether there is a visible relationship between **reported crimes associated with CTA infrastructure** and **daily ridership patterns** across the city of Chicago. The approach combined historical ridership logs from the CTA with filtered crime reports spanning 2010 to 2024.

Data Sources and Cleaning

Ridership Data: Daily turnstile entries and exits per CTA station, aggregated to calculate citywide daily ridership.

Crime Data: Public safety dataset from the City of Chicago, filtered using the keyword "CTA" across the location_description and primary_type fields.

Steps taken:

- Merged daily ridership and crime data on the date field.
- Removed entries with missing or erroneous date formats.

Exploratory Data Analysis

- Generated **time-series plots** to visualize fluctuations in crime volume and ridership from 2010 to 2024.
- Disaggregated the timeline into **pre-2020 (stable ridership)** and **post-2020 (COVID impact)** periods.
- Found that **crime spikes were more frequent on weekdays**, aligning with high-volume commuting patterns.

Correlation and Trend Influence

- Created **lagged variables** (e.g., crime on Day N compared to ridership on Day N-10, N, N+10) to test the effect of prior-day crime on next-day ridership.
- Used matplotlib and seaborn to plot:
    - Daily crime count vs. ridership trends
    - Scatterplots with regression lines for visual correlation assessment

Key observations

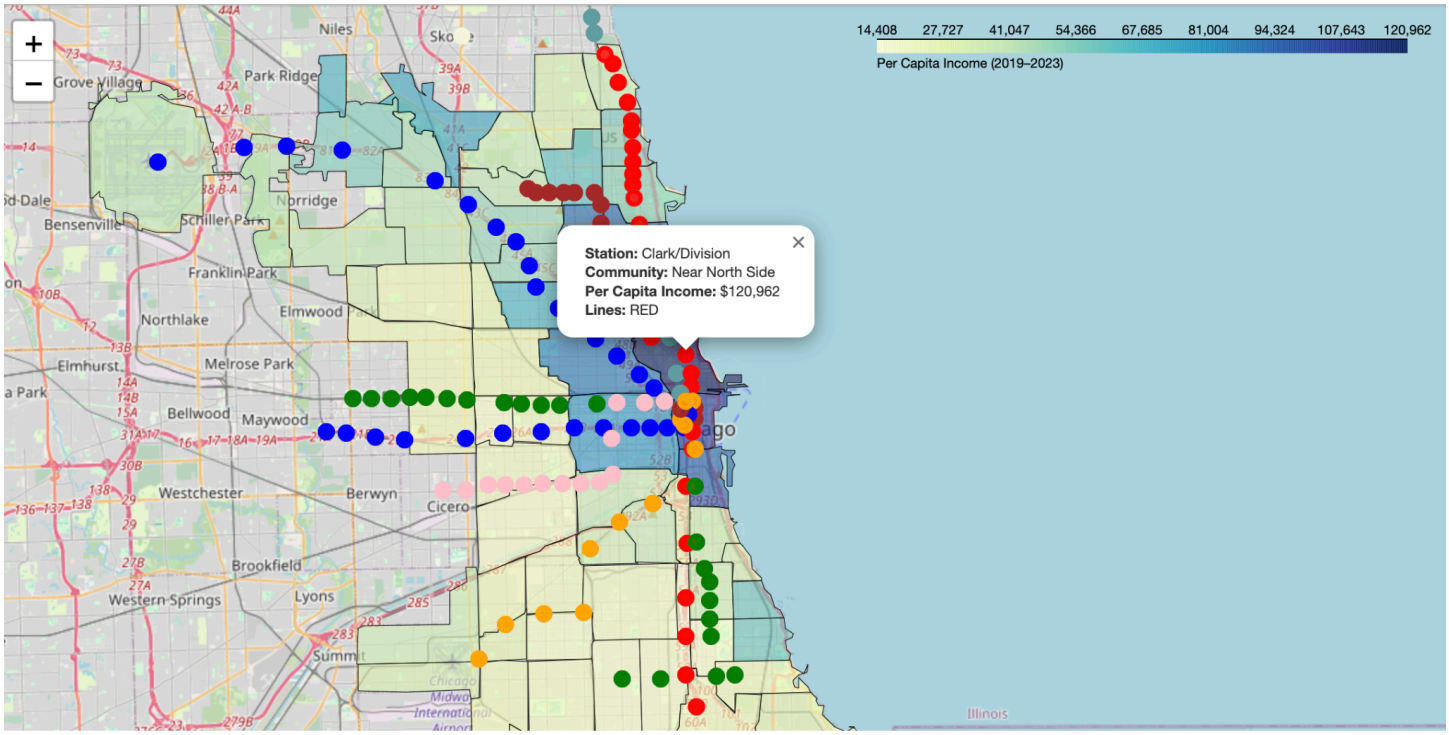
- In the **pre-2020 period**, ridership remained fairly stable even during high-crime days, suggesting rider desensitization or limited alternatives.
- **Post-2020**, spikes in CTA-related crime corresponded more visibly with dips in ridership, especially during weekday rush hours.
- This shift may reflect **increased rider sensitivity** to safety in the post-pandemic context.

Selected Visualizations

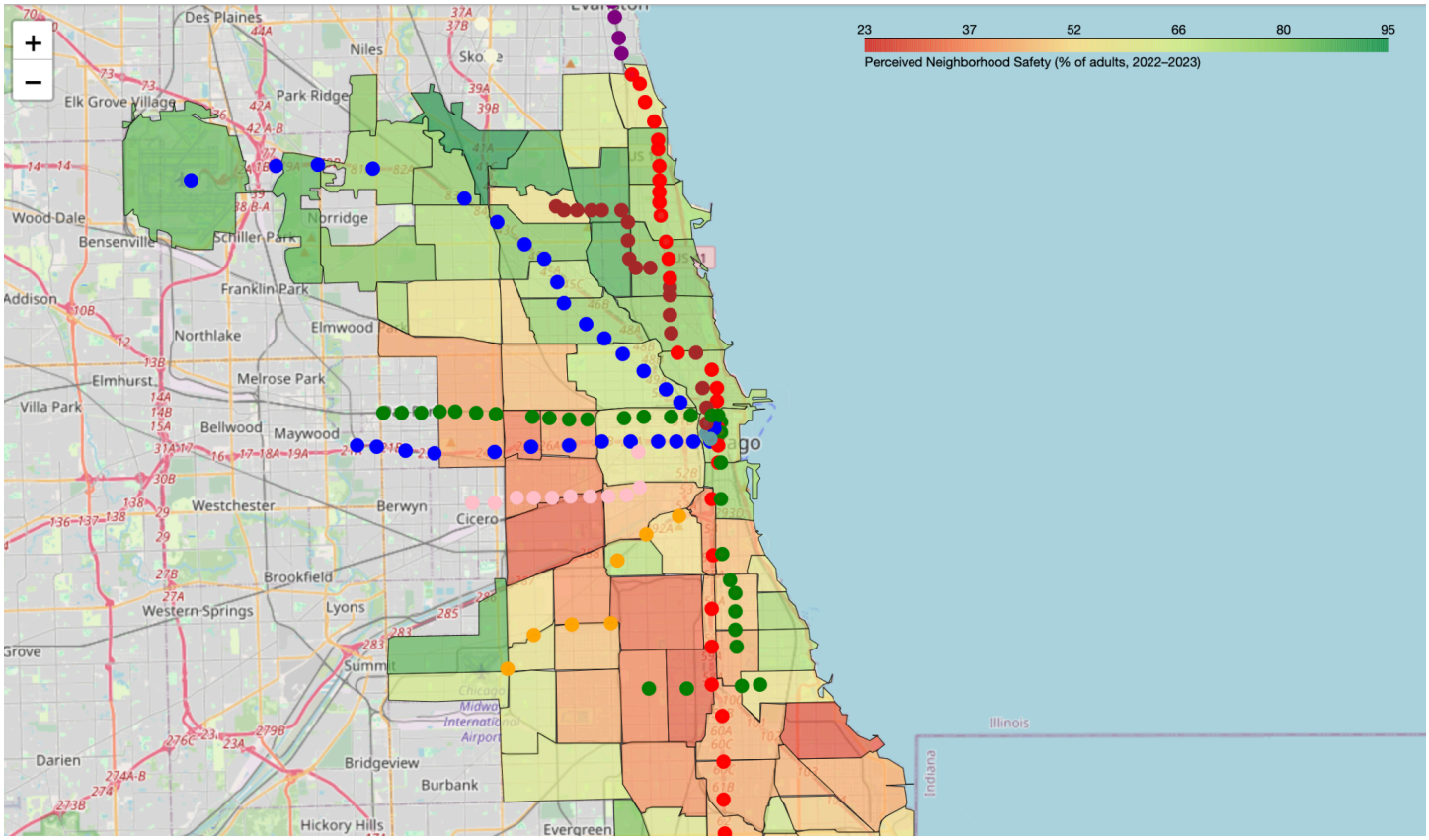
Each member of the group contributed with at least 2 visualizations.

Burak Simsek

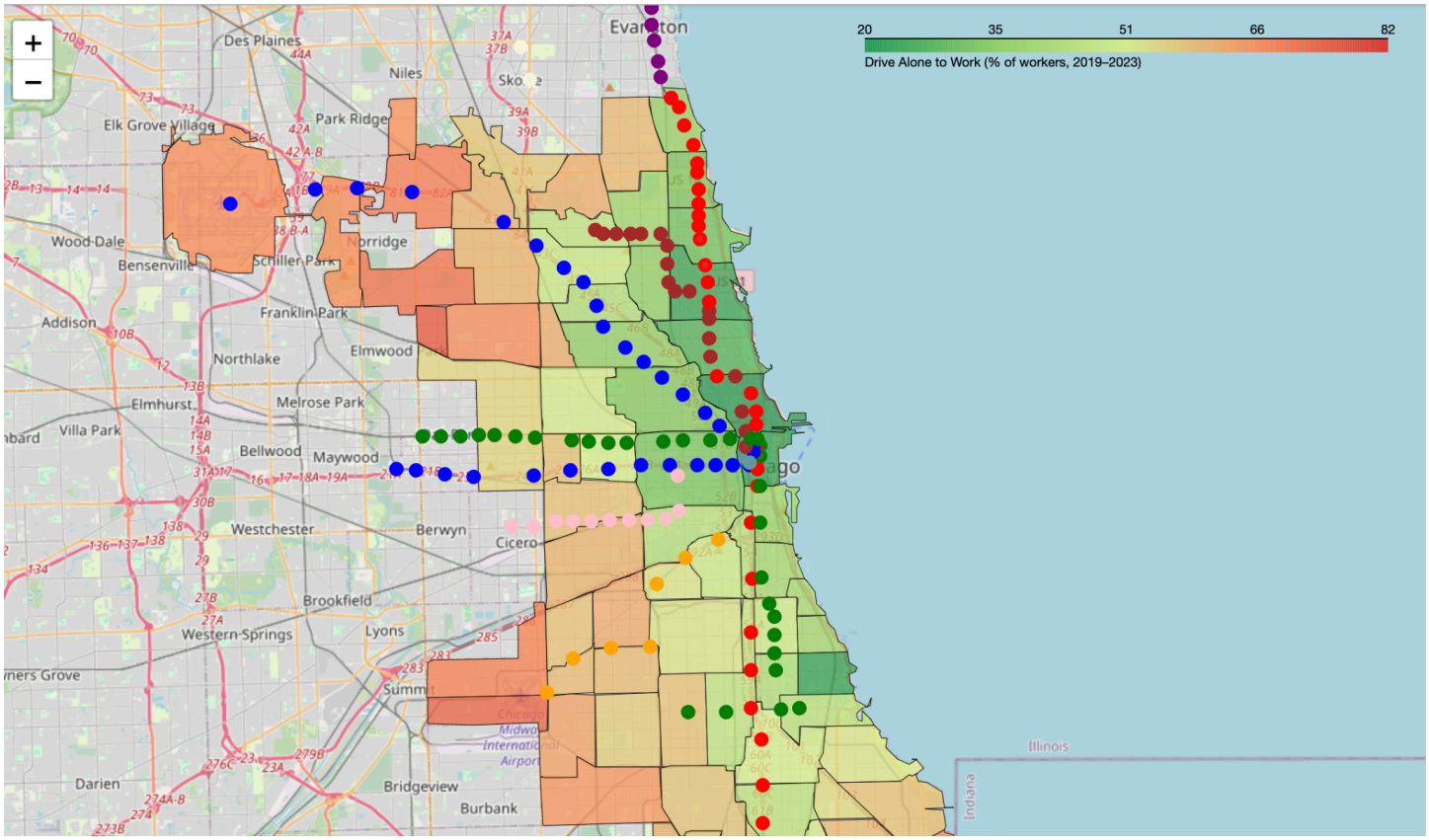
CTA Transit Access & Socioeconomic Disparity Map (2019–2023)



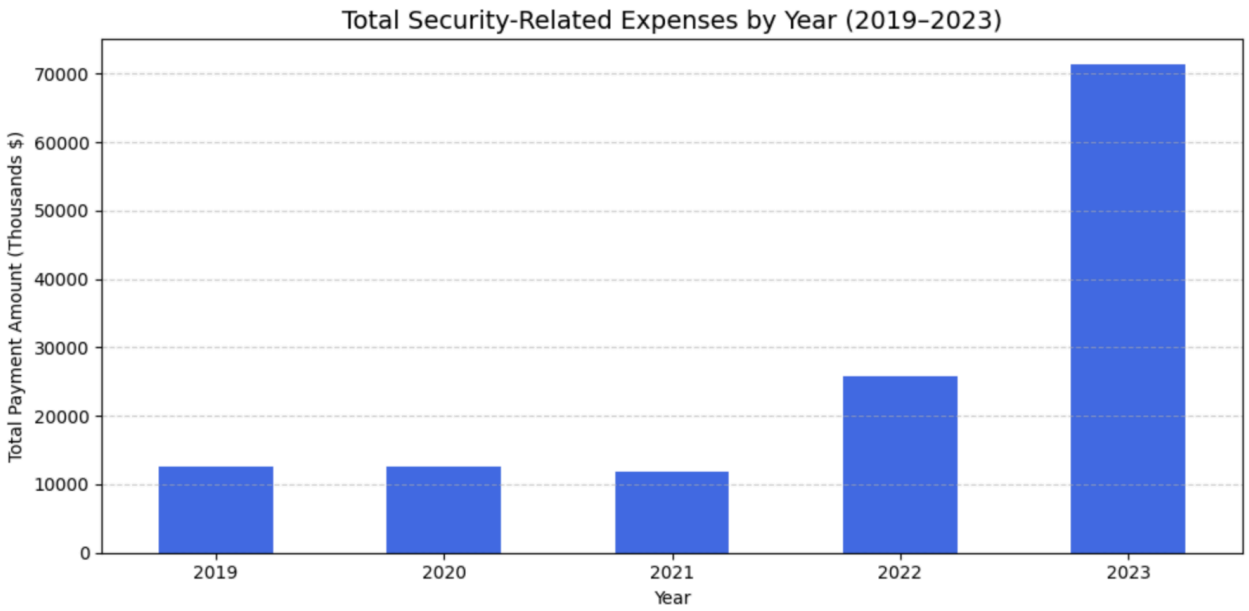
Interactive Safety and Transit Map of Chicago (2019-2023)

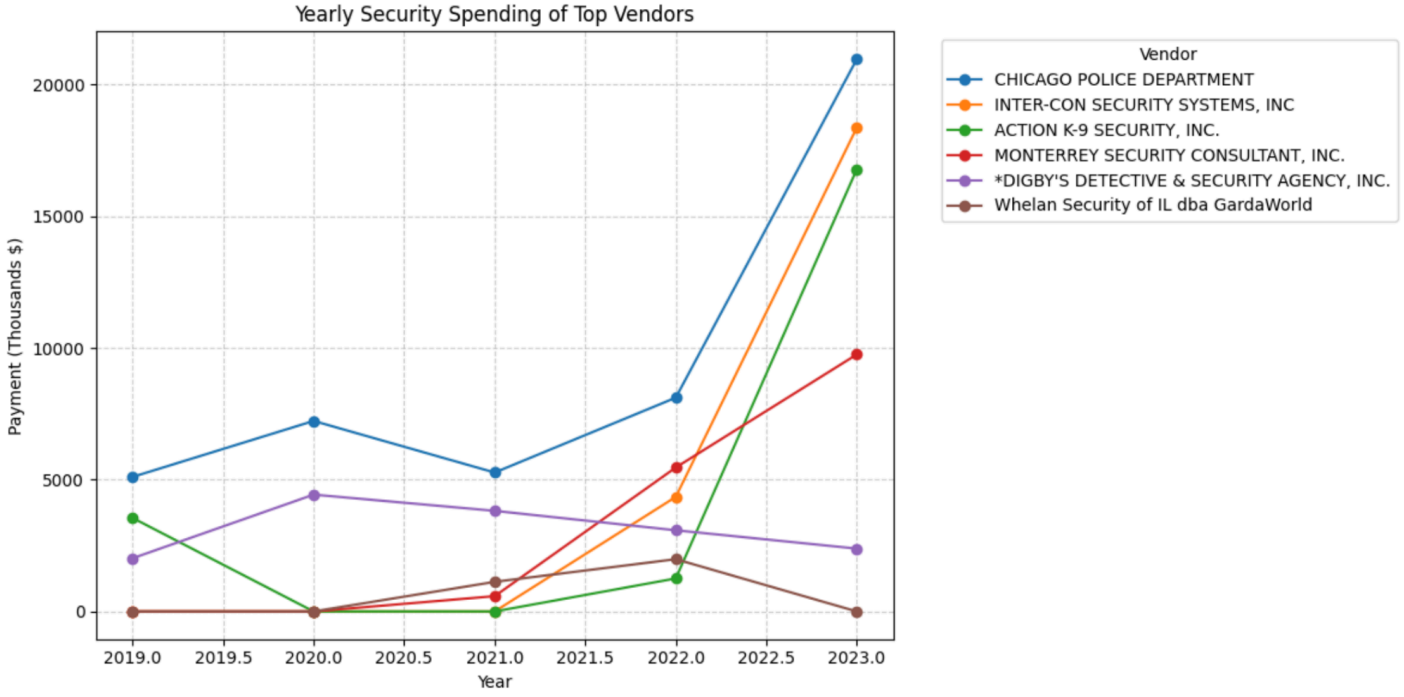


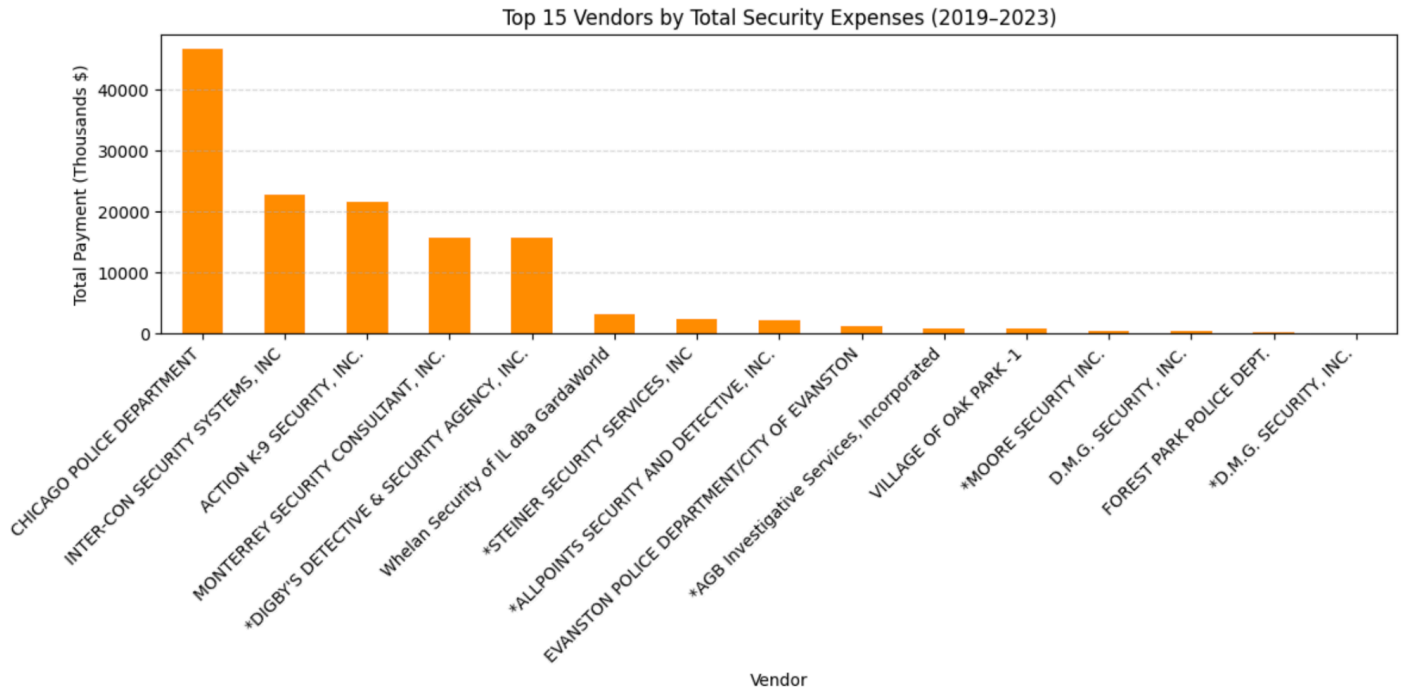
Drive Alone to Work vs. CTA L Train Accessibility in Chicago (2019–2023)



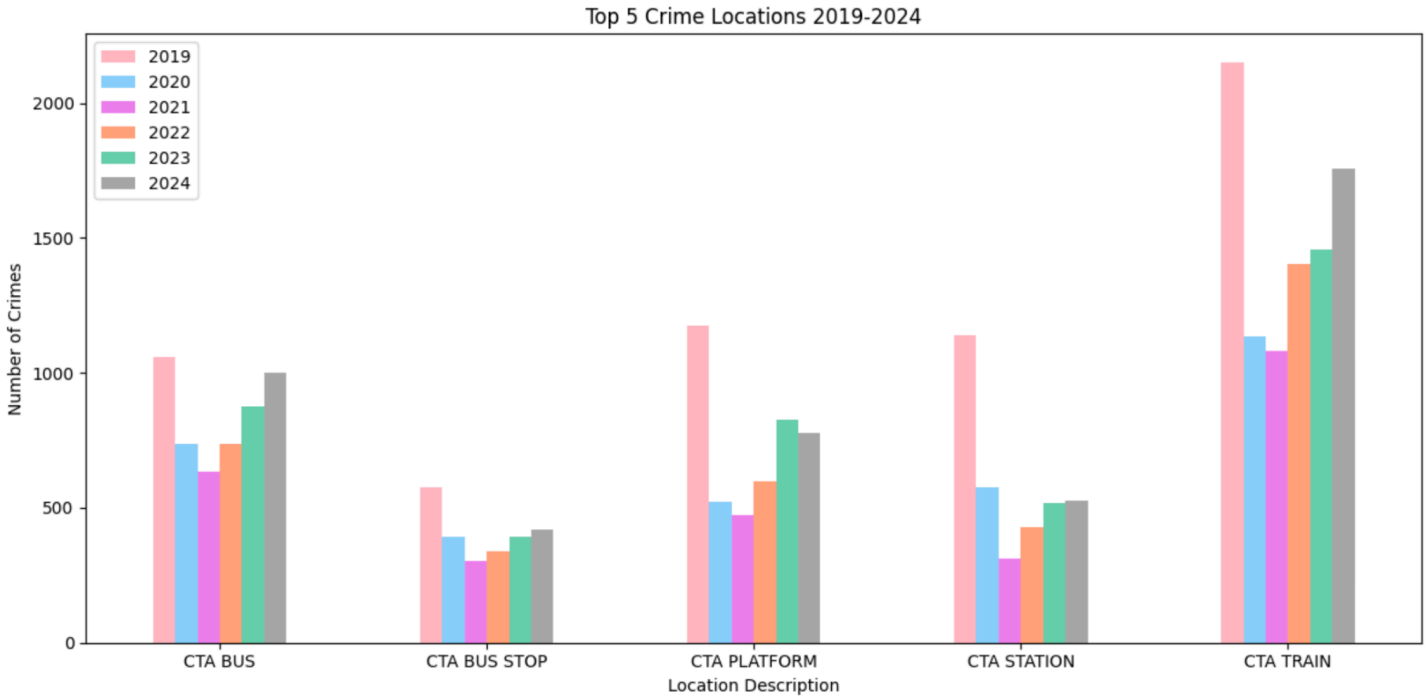
CTA Security-Related Expenditures (2019–2023)

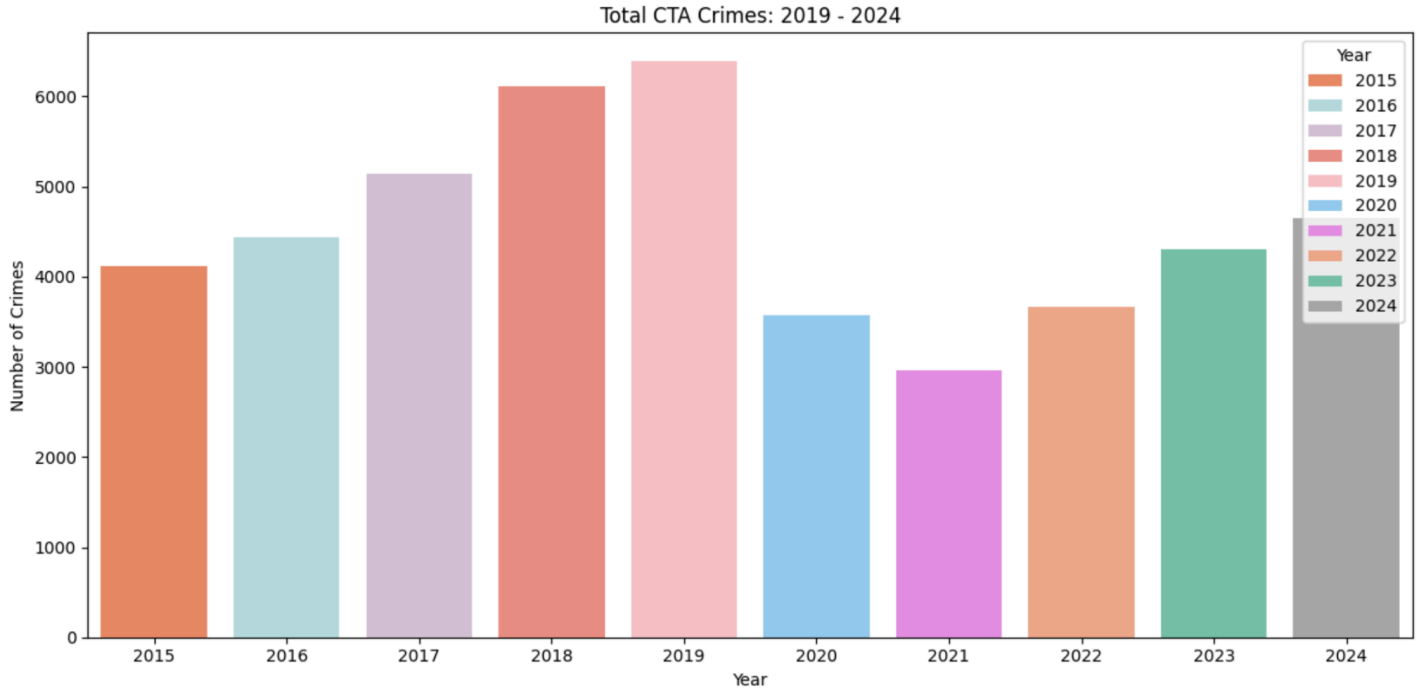


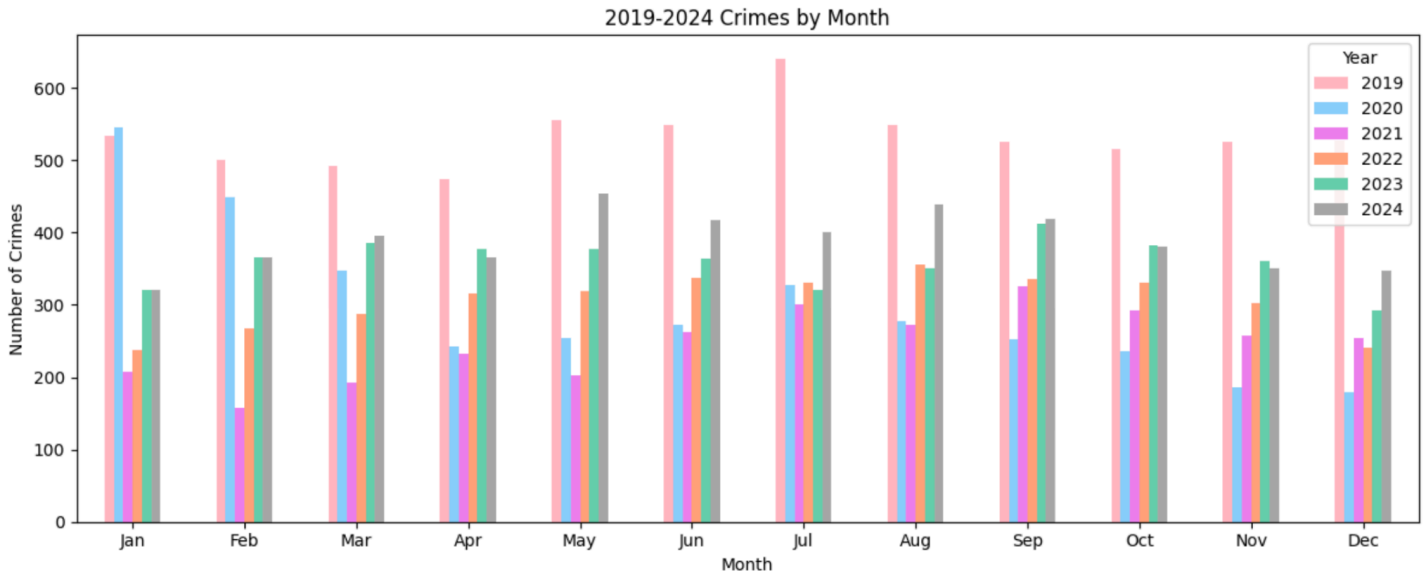

Yearly Security Spending of Top Vendors
Payment (Thousands $)
Year
Vendor
CHICAGO POLICE DEPARTMENT
INTER-CON SECURITY SYSTEMS, INC
ACTION K-9 SECURITY, INC.
MONTERREY SECURITY CONSULTANT, INC.
*DIGBY'S DETECTIVE & SECURITY AGENCY, INC.
Whelan Security of IL dba GardaWorld

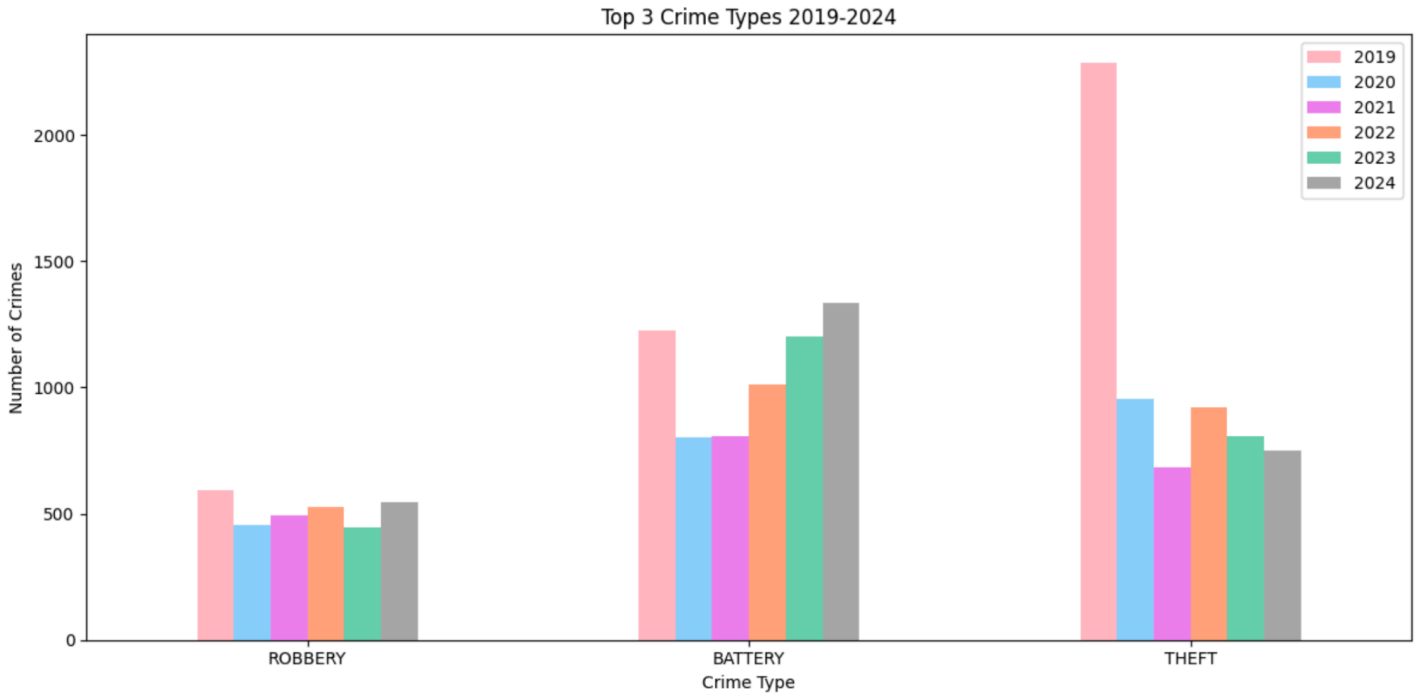

Top 15 Vendors by Total Security Expenses (2019–2023)
Total Payment (Thousands $)
Vendor
CHICAGO POLICE DEPARTMENT
INTER-CON SECURITY SYSTEMS, INC
ACTION K-9 SECURITY, INC.
MONTERREY SECURITY CONSULTANT, INC.
*DIGBY'S DETECTIVE & SECURITY AGENCY, INC.
Whelan Security of IL dba GardaWorld
*STEINER SECURITY SERVICES, INC
*ALLPOINTS SECURITY AND DETECTIVE, INC.
EVANSTON POLICE DEPARTMENT/CITY OF EVANSTON
*AGB Investigative Services, Incorporated
VILLAGE OF OAK PARK -1
*MOORE SECURITY INC.
D.M.G. SECURITY, INC.
FOREST PARK POLICE DEPT.
*D.M.G. SECURITY, INC.

Amina Marin


Top 5 Crime Locations 2019-2024
2019
2020
2021
2022
2023
2024
Number of Crimes
0
500
1000
1500
2000
CTA BUS
CTA BUS STOP
CTA PLATFORM
CTA STATION
CTA TRAIN
Location Description

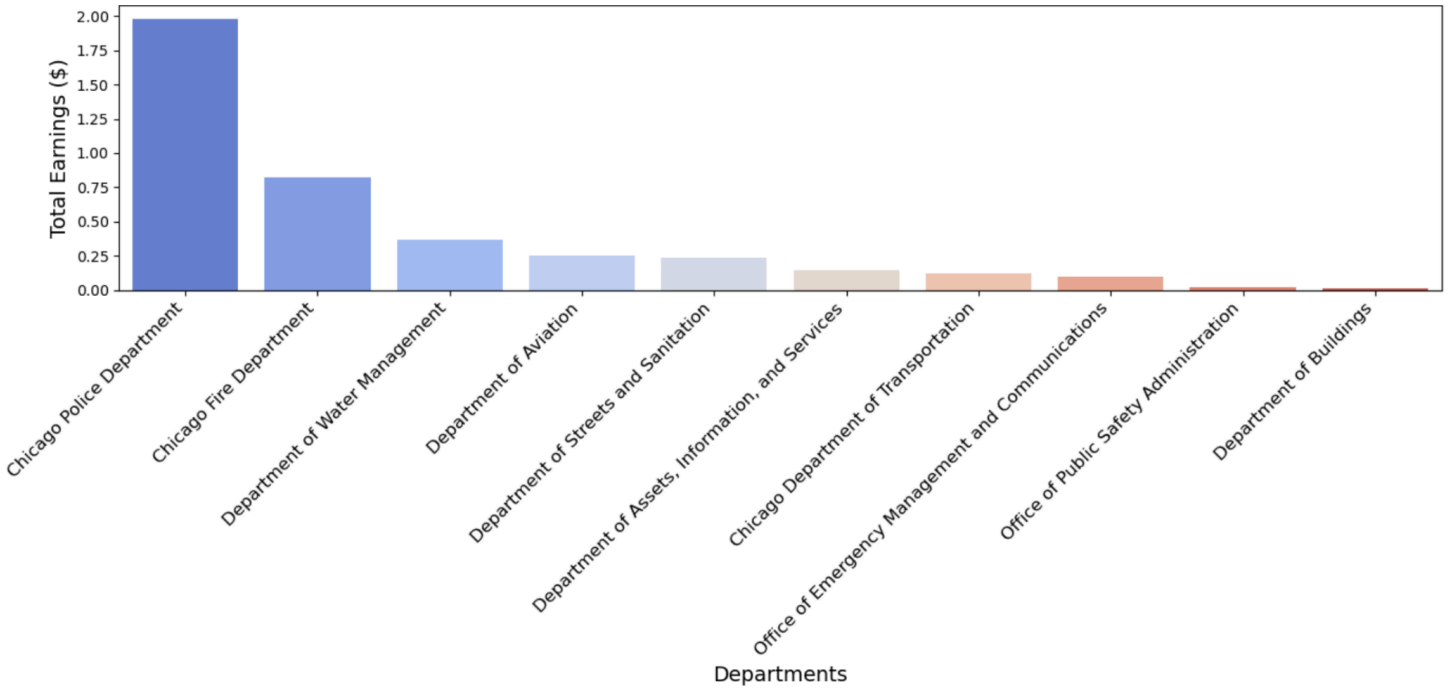

Total CTA Crimes: 2019 - 2024
Year
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Number of Crimes
0
1000
2000
3000
4000
5000
6000
Year



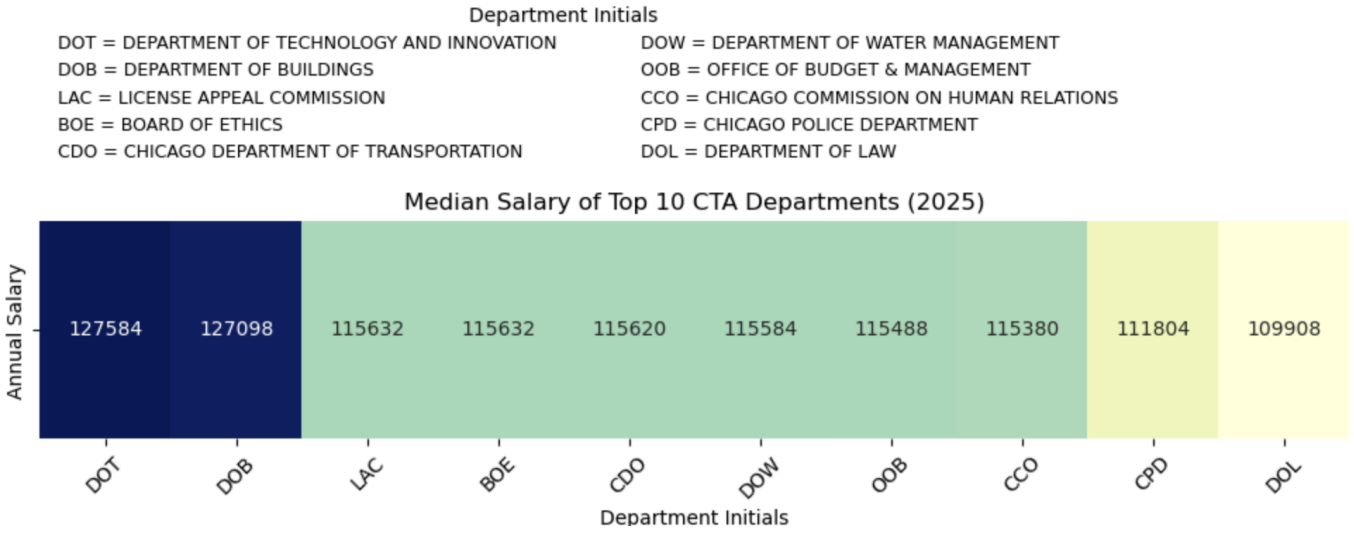


Heer Patel

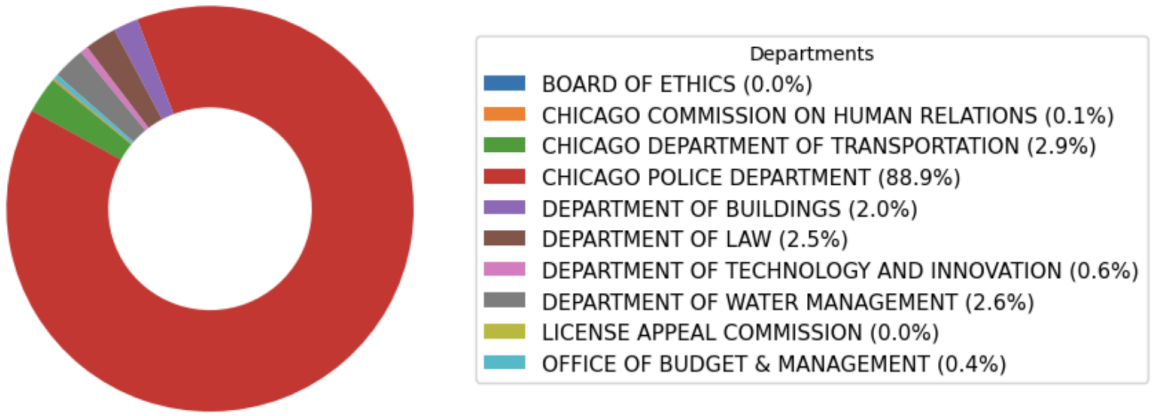
Top 10 Departments by Total Earnings (2022)



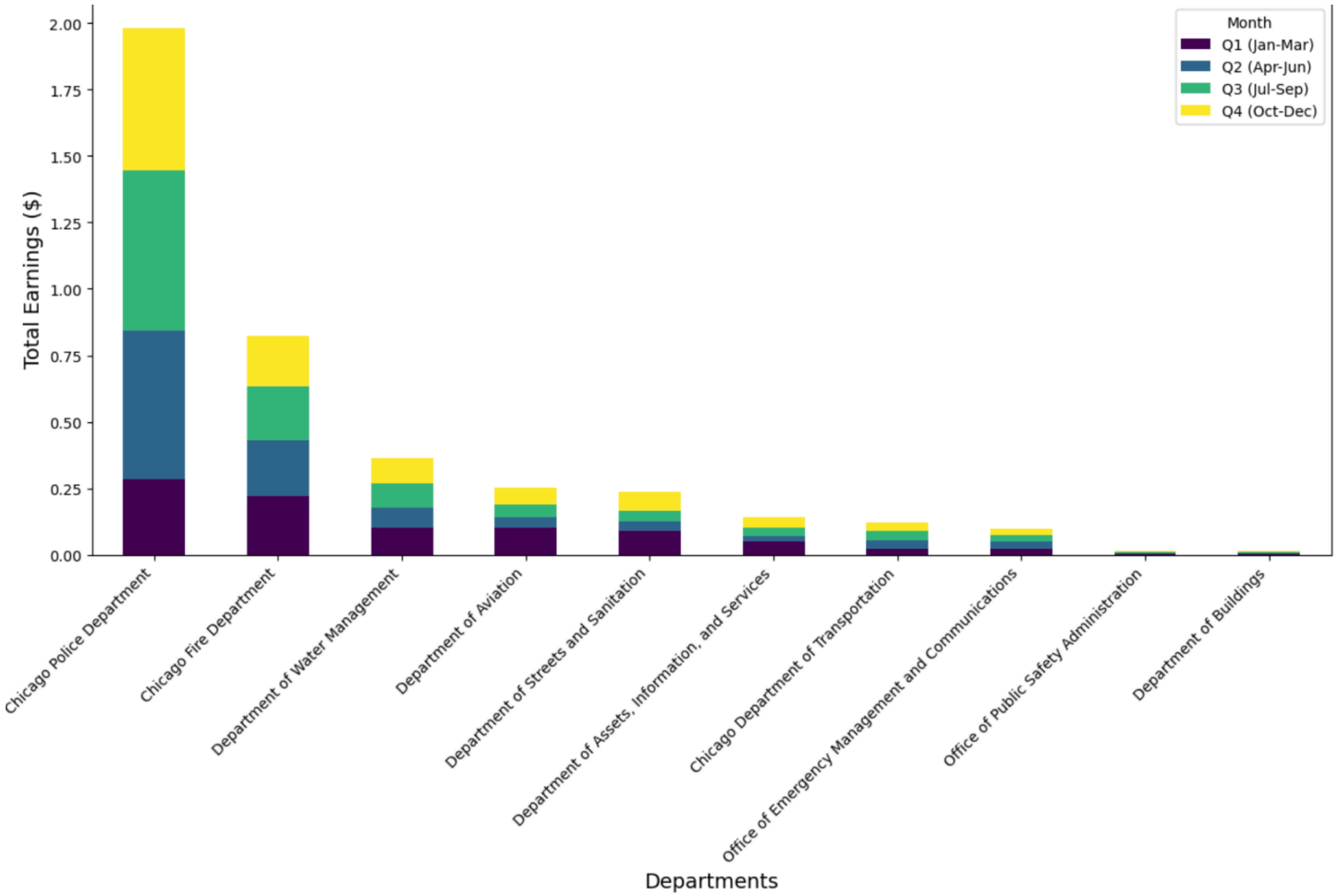
Median Salary of Top 10 CTA Departments (2025)



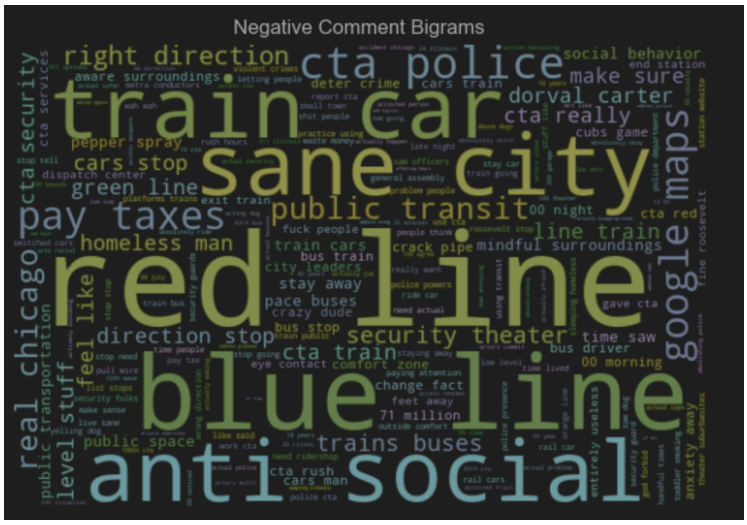
Share of Total Salaries by Top 10 Departments

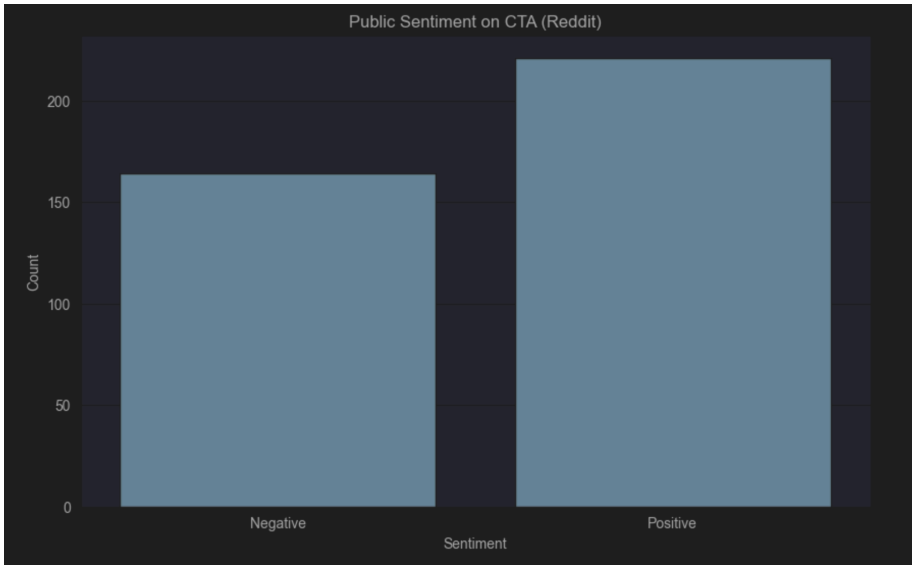


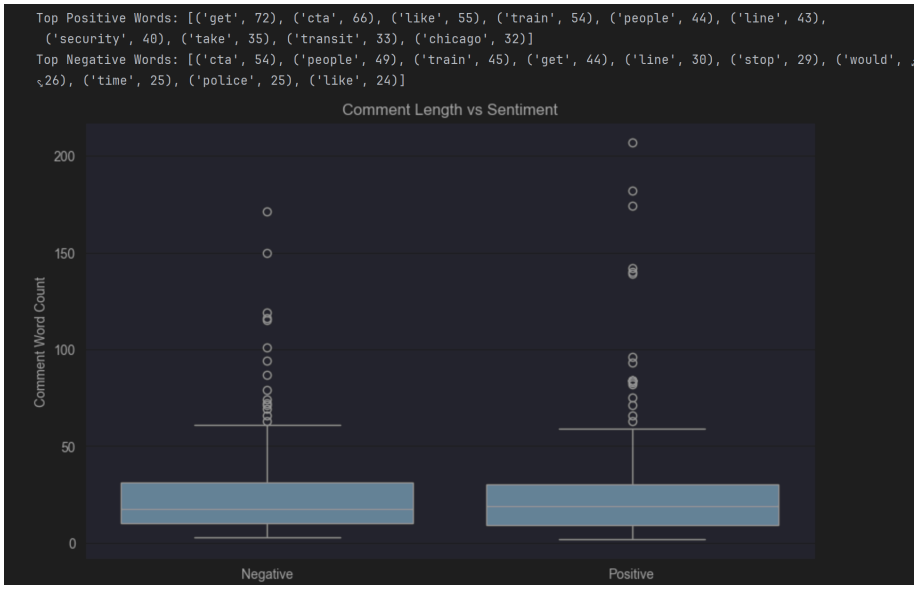
Top 10 Departments by Overtime and Regular Earnings (2020)



Prince Sonani







Srijita Banerjee

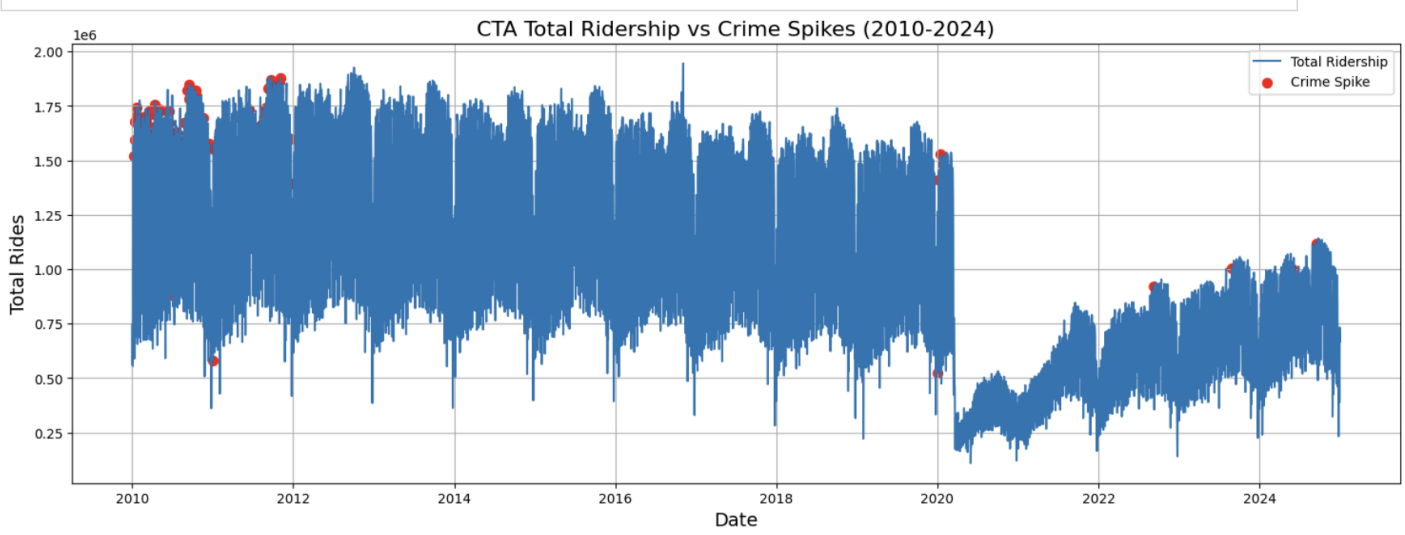


Figure: CTA Daily Ridership and Crime Spikes (2010–2024)
Daily ridership trends with red dots marking crime spike days. Highlights include the COVID-19 dip and uneven correlation between ridership and crime.

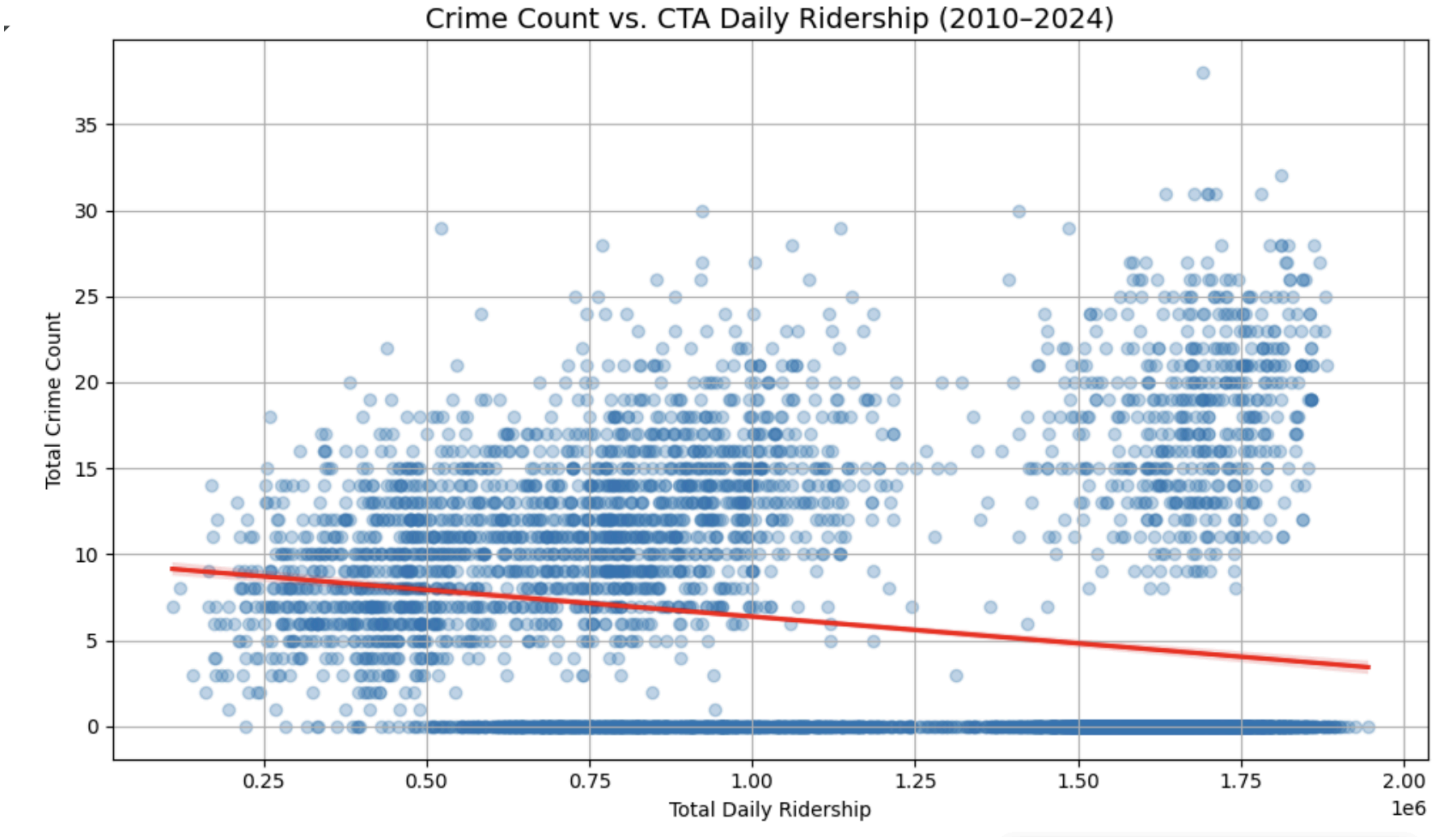


Figure: Crime Count vs. CTA Daily Ridership (2010–2024)
Scatterplot with regression line showing a weak negative trend between ridership and crime. High variance suggests limited linear correlation.

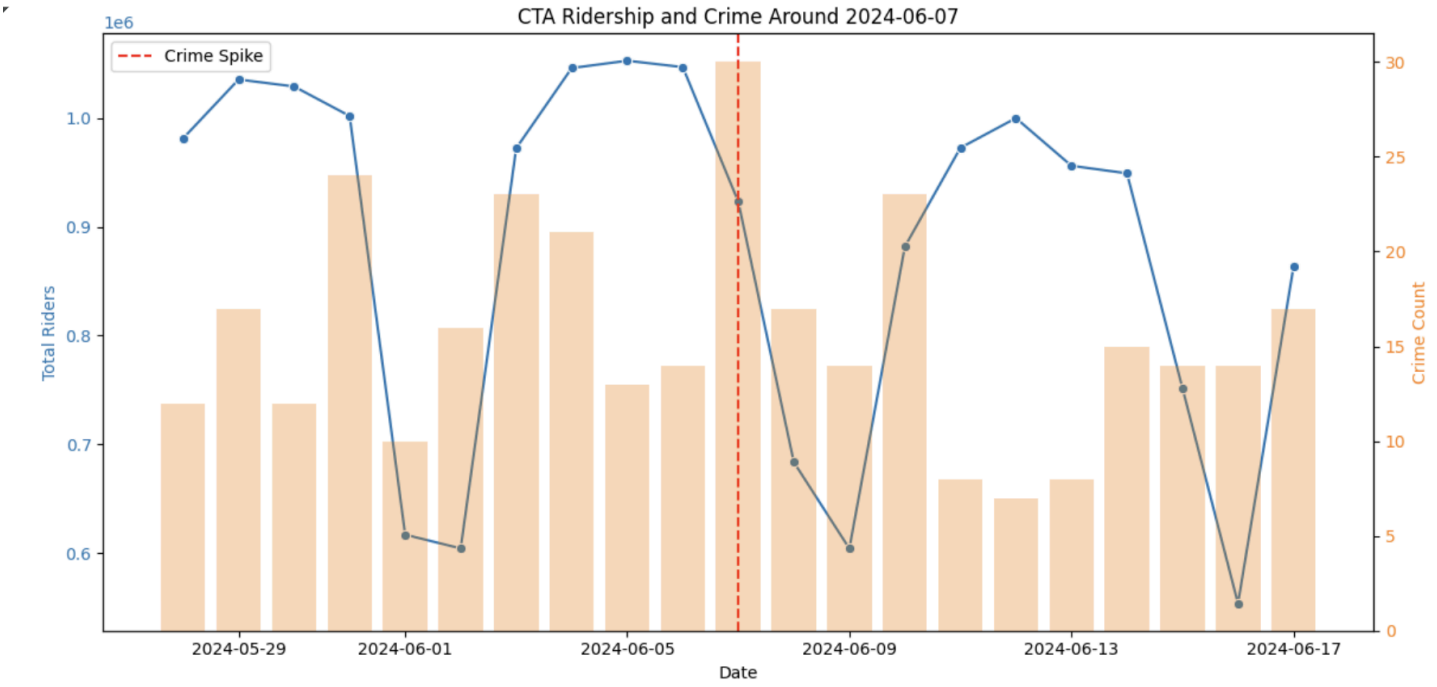


Figure: CTA Ridership and Crime Around June 7, 2024
Zoomed view of a crime spike event. Crime count peaks sharply on June 7 while ridership stays relatively stable, showing a localized anomaly without a clear inverse relationship.

Additional Work

As part of our extended analysis beyond the core deliverables, we developed a focused deep-dive on the relationship between reported crimes and CTA ridership from 2010 to 2024. This work is documented in the **crime_on_crime_ridership_levels.ipynb** notebook and involved several custom analytical approaches:

- **Merged multi-year crime datasets** across different formats and sources, filtered explicitly for CTA-relevant incidents (e.g., platform, train, and station-based crimes).
- Built a **daily crime aggregation** and defined statistical spike thresholds using the mean and standard deviation of daily crime counts.
- Merged the crime dataset with daily CTA ridership figures to enable day-level comparisons.
- Designed and implemented **custom dual-axis visualizations** that zoom in on high-crime days and show both crime count and ridership movement.
- Created an **overall time series overlay plot** showing red markers on crime spike dates over ridership trends, highlighting macro-scale alignment (or misalignment) between security concerns and travel behavior.
- Developed a **regression-based scatterplot** to assess whether there was any linear relationship between crime volume and ridership levels.

- Used this extended analysis to question simplistic assumptions about causality, and better understand temporal context (e.g., post-2020 behavior shifts).

This deeper investigation went beyond the basic assignment expectations and added new dimensions to the project's conclusions—particularly regarding the **interaction between perception of safety, timing of crime events, and systemwide ridership patterns**.

Overall Results and Insights

Key findings include:

- **Weather & Events**: Rain and extreme cold led to significant weekday ridership drops. Public events like parades increased traffic, particularly on the Blue and Red Lines.
- **Crime**: Theft and battery were the most frequent crimes. Although pre-2020 ridership remained high despite crime spikes, post-2020 ridership became more sensitive.
- **Public Sentiment**: A majority of Reddit users expressed security concerns about off-peak hours, influencing travel decisions, especially for women and seniors.
- **Equity**: Neighborhoods with low income had poorer train access and higher crime exposure, potentially compounding transit avoidance.

Conclusion

Through this project, we set out to explore how external factors—such as crime, weather, and citywide events—interact with and influence public transportation usage across Chicago. While our analyses were exploratory in nature, they revealed meaningful trends that transit agencies like the CTA could use to plan more equitable, responsive, and safe systems.

One of the central takeaways was that **perceptions of safety play a major role in shaping ridership behavior**. From crime datasets filtered specifically for CTA-related incidents, we observed visible fluctuations in ridership during and after periods with higher rates of theft, battery, and deceptive practices. These fluctuations were most noticeable during commuter-heavy weekday schedules and in certain neighborhoods with historically underfunded infrastructure. Although we did not apply formal causal inference methods, our merged dataset and visual analyses indicate a strong association between reported crime and shifts in ridership.

In parallel, **community sentiment—as reflected through public posts on Reddit—added an important qualitative dimension**. While not every comment was directly relevant, many reflected real concerns over station security, especially during early morning or late-night hours. Sentiment analysis helped us identify recurring themes such as fear, trust, and reliability, which quantitative data alone cannot capture. These insights reinforce the value of integrating public perception into transit planning.

Our weather and event data also revealed intuitive trends. Large-scale events led to increased ridership along specific lines, while extreme weather conditions generally correlated with decreased use. These findings underscore the need for dynamic planning—both in terms of fleet and staff allocation, as well as safety presence during high-demand periods.

Finally, our mapping of **socioeconomic disparity and transit access revealed spatial inequities** in CTA infrastructure. Low-income neighborhoods on the West and South sides of Chicago had both limited access to train stations and higher rates of crime near those that existed. This dual burden could discourage public transit use and reinforce existing mobility and opportunity gaps.

Final Thoughts

This project helped us demonstrate that **even basic data integration and exploratory modeling can produce powerful insights**. Public transportation systems are deeply intertwined with public safety, economic conditions, and environmental variables. We believe that agencies like the CTA can benefit from multidisciplinary approaches that go beyond operational data and incorporate social and environmental perspectives.

Looking forward, the foundation we built—especially through merged datasets and interactive visualizations—could support future applications in real-time prediction, equitable transit design, or even targeted policy recommendations. More importantly, we hope this work sparks conversation about how public data can be leveraged to build safer, more inclusive transit systems for all Chicagoans.