# Law and Order SQL: Technical Report

Albert Prouty, Chris Kusha, Clara McGrath, Jassleen Bhullar

## I. Introduction

It is known that crime affects the communities around it, and creeps into many facets of life. We explore crime in Chicago and focus on its connections to the school system. We aim to answer questions such as whether crime has increased or decreased in the last decade? Do students feel safe in Chicago public schools? How has crime in Chicago affected education and students? What effect has crime had on the safety, academic performance and behavior of students? We also aim to explore whether there are correlations between crime reported and behavioral misconduct within the school system, and if there are any connections between students' demographics and reported misconducts in districts.

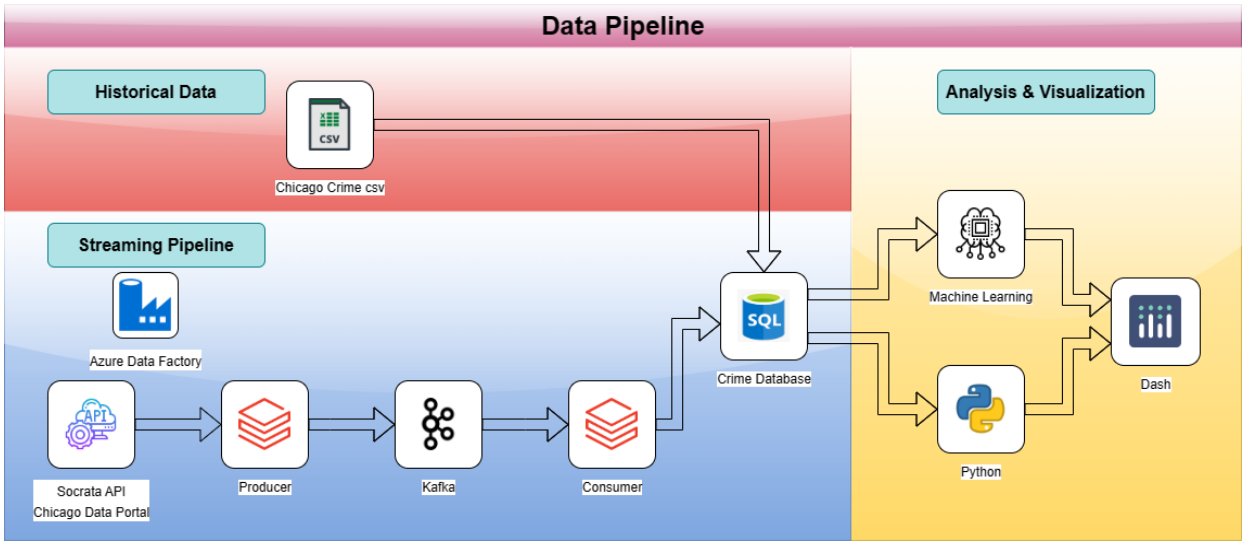
## II. Data Sources and ETL

We used a total of six datasets for our research in this project, with the majority of them coming from the Chicago Data Portal, which is open for public use. The first being Public School Profile information for 2022-2023, which contains demographic information and performance data for schools across Chicago. The next two are School Progress Reports, reporting on year over year growth for each school, safety survey results, awards and certifications as well as disciplinary statistics. An older progress report from 2011-2012 was also used to help connect the schools with their district, community area, and wards, which were no longer included in later progress reports, so we could make geographic connections between crime levels and school performance. The next two were misconduct reports on both the school and district level from the Chicago Public School website, that are used to gauge the misconduct statistics of schools and areas and correlate them together with outside crime information.

For our last dataset, we used a subset of data from crimes 2001 to present. At first we tried using the whole dataset, but with 7.72 million rows and 500-800 more being added daily, it was a little too large to use with our available storage space and tools. We decided to focus from 2013 to the present, which proved to be more manageable at roughly 2.5 million rows of data, which were put together through yearly views created from the main crime set. These gave information about crimes in Chicago with information such as what the crime was, where and when it had occurred, and whether an arrest was made.

All of this data came in the form of CSV (comma separated value) files available through the data portal. The crimes however are updated daily with new information on a week delay. Using an API through Socrata, we were able to obtain credentials and make daily calls to the API

to retrieve the latest data without having to download new CSVs. First we created a Databrick to pull in new information and send it out as messages through Kafka. We created a second Databrick to read the messages and write them onto an Azure Data Lake hosted SQL Database. We then automated this process through Azure Data Factory by creating a pipeline.

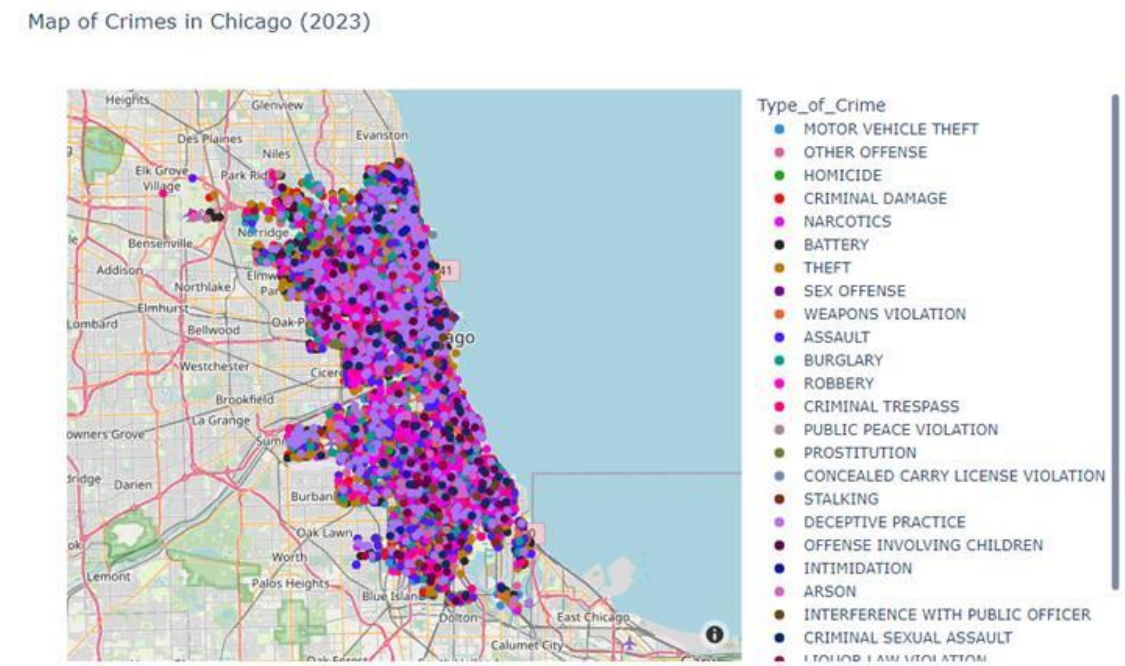



With some cleaning to get rid of some irrelevant columns and changing data types, this data was now ready to be used for exploratory data analysis (EDA), visualizations, and machine learning models.

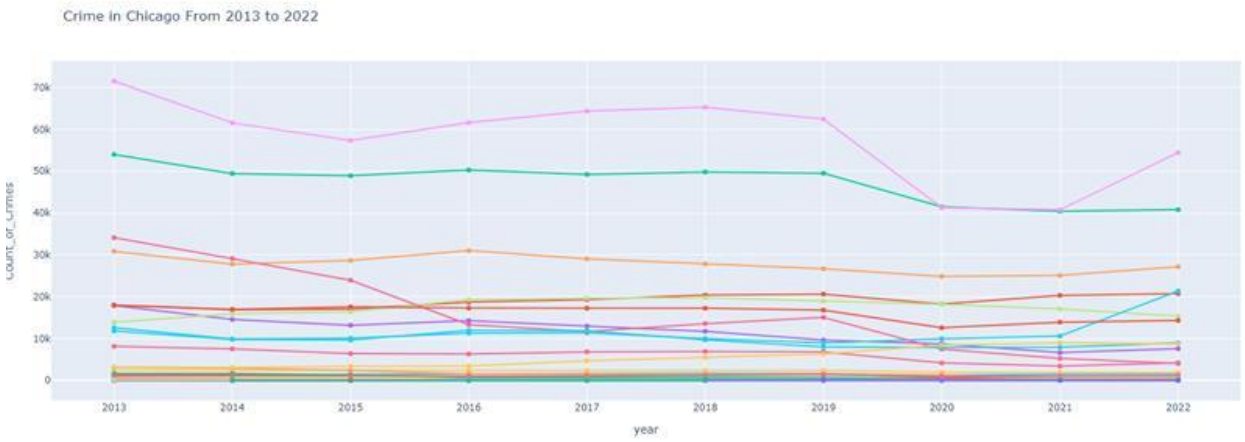
Data cleaning for the other datasets were done during the EDA process in Python, using libraries such as pandas and numpy to organize the data, deal with null values, merge data together to connect district and ward information to the school reports, and filter out information we weren't interested in (some of these sets had 100-200 columns!).

## III. Visualizations

We decided to use Dash with Plotly for our project dashboard. Visualizations were made in notebooks using Plotly express then written into the code for the Dash port. The dashboard contains a total of ten visualizations that help answer our research questions. We decided to create multiple pages, including one for crime in Chicago in general, one for the visualizations that involve Chicago education, and one for just the machine learning visuals.



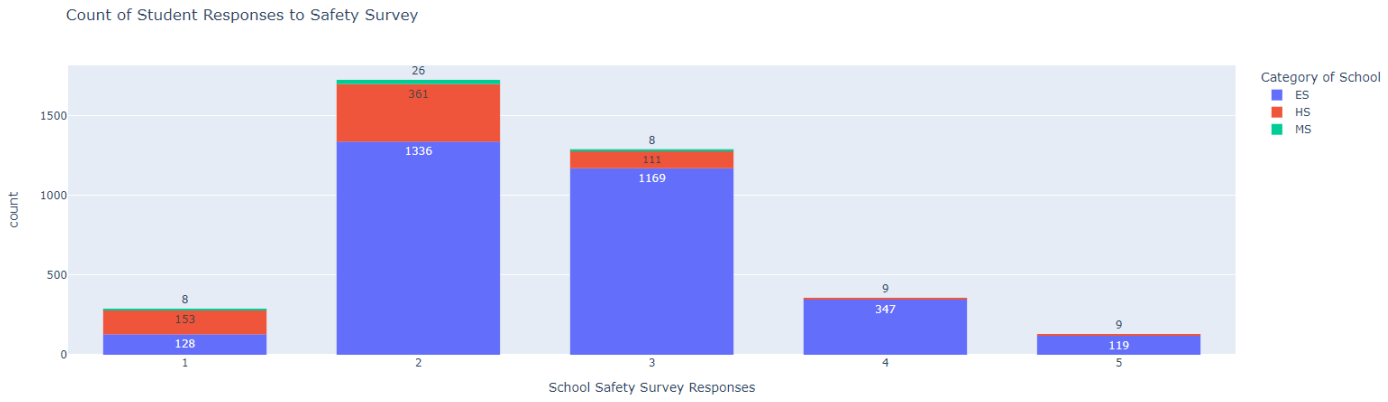

First, we see an interactive map that acts as an introduction to the crime data. This map is updated daily with new records of reported crimes, along with the crime's location. The data is pulled from the live dataset from the Chicago Data Portal, which is being streamed through a Kafka pipeline and stored in a secure SQL database. The data is filtered by type of crime, plotted throughout the Chicago area according to the longitude and latitude coordinates.




Next, there is a line graph that answers the question of whether crime in Chicago has increased or decreased over the last decade, where the different colored lines represent different crime categories. The following is another line graph that answers the same question, but is limited to only domestic crimes. The data has a separate, boolean column for domestic crimes, which requires a separate graph.

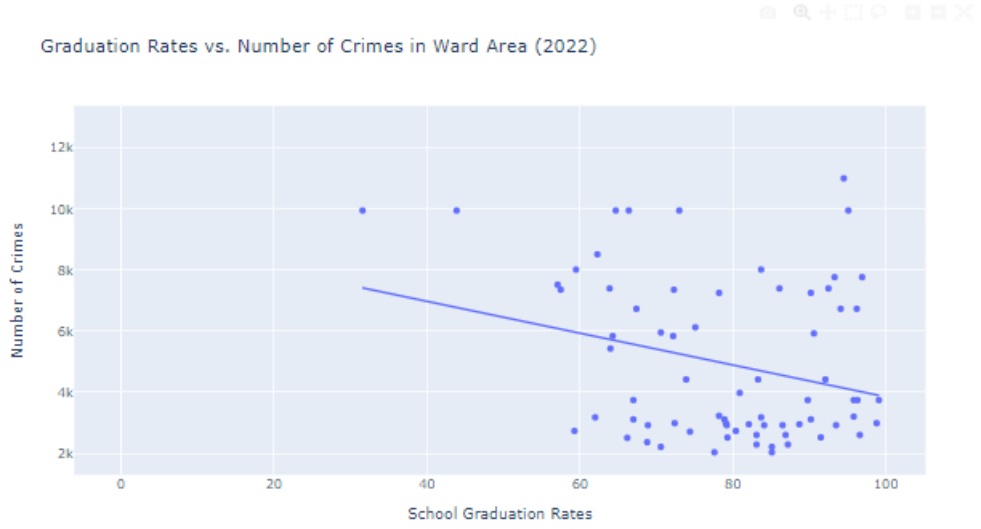

Both these charts demonstrate that overall, crime is decreasing in Chicago. In particular, violent crime rates are lower in 2022 than in 2013. The most commonly perpetrated crimes in Chicago are theft and battery, and both types of crime are experiencing a downwards trend. Unsurprisingly, there was a sharp decrease in nearly every type of crime in 2020. Considering the 2020 COVID-19 pandemic and the ensuing limitations of people in public places, the decline makes sense.

Concerningly, homicide (non-domestic) is the only crime that had increased numbers of occurrence in 2020. Homicide rates did decrease in 2021 and 2022, however. Reports of motor vehicle theft (non-domestic) and assault (domestic and non-domestic) also increased sharply in 2021. There are approximately 3,000 more reports of assault (non-domestic) in 2022 than in 2013. Motor vehicle theft, specifically, has increased dramatically, with nearly 8,000 more reports of motor vehicle theft in 2022 than in 2013. While the increasing rates of motor vehicle theft and assault are concerning, it should still be stated and understood that nearly all other types of crime are experiencing a decline in number of occurrences when compared to rates from a decade ago.



Using the data from the school safety survey, this bar chart shows the student responses. The responses are on a scale of 1 to 5, with 1 meaning very weak feelings of safety, 2 meaning weak, 3 neutral, 4 strong, and 5 very strong. The chart shows that there was a substantial amount of responses indicating a weak feeling of safety at school, with the second most populated category being neutral - this leaves us with the answer that the students overwhelmingly do not feel safe.

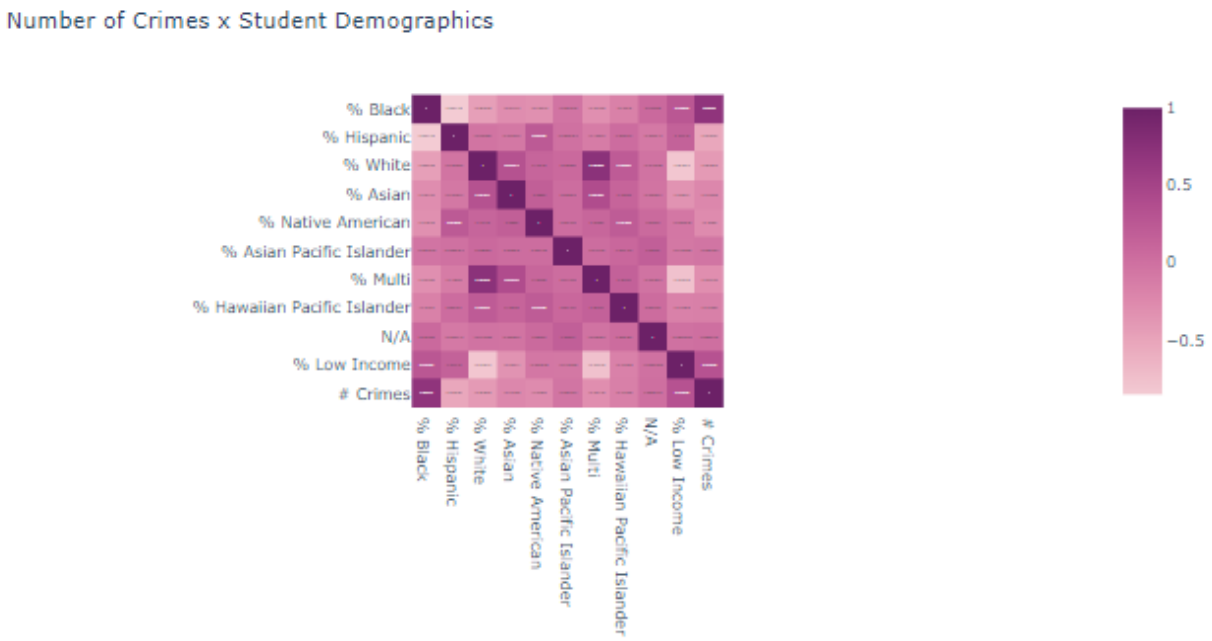
We do not have enough data to conclude that students feeling unsafe at schools is correlated to crime rates in Chicago. The survey only asked students to rate how safe or unsafe they feel, but it did not ask students the reason why they feel that way. The safety score was also aggregated by the owner of the dataset. Instead of receiving scores from each individual student who participated in the survey, the dataset presented an aggregated safety score for each school. The data also overwhelmingly represents elementary schools, with very little high schools included. High schoolers would be a better population to survey because they have a better understanding of safety levels. However, the fact that the large majority of elementary students feel unsafe in schools is definitely an interesting and concerning point that needs to be considered further.




We used a scatterplot with a trendline to demonstrate the relationship between graduation rates and number of crimes in the ward areas of Chicago Public Schools - the crime count was calculated based on ward, then matched to the ward of each school. Though in the visual we can see a negative relationship, the actual correlation is not very strong. When calculated, the value comes out to -0.28.

Although it is not a strong correlation, the finding of a negative relationship between crime and school graduation rates is still cause for concern that demands greater focus. It suggests that as crime rates increase in Chicago, graduation rates drop. This correlation, although weak, creates another incentive to keep crime rates down. Thankfully, crime rates in Chicago are

decreasing over the year, and Chicago officials should be further incentivized to maintain a downward crime rate to increase graduation rates of students from Chicago public schools.




To examine the trends in school demographics and reported crime, a correlational matrix was created, but percentages of the student body were used rather than numerical counts for each demographic group. The group that had the strongest correlation with the number of crimes was the low income group, meaning that the larger the percentage of low income students, the higher the number of crimes in that corresponding police ward area.

## IV. Machine Learning

The data for the machine learning model consisted of the Chicago Public School district and school level reports merged with the Chicago Police Crime data. A python environment within Visual Studio Code was used to drop non-relevant columns, create dummy variables for categorical data, and scale the resulting arrays with StandardScaler. For the sake of brevity, the steps for ETL and initial cleaning will be skipped over, but the resulting data frame consists of 15,093,883 rows and 37 columns. Every 1000th row was sampled and put into a test data frame which was subsequently split with a test size of 0.25 and random state of 42 before the model was instantiated.

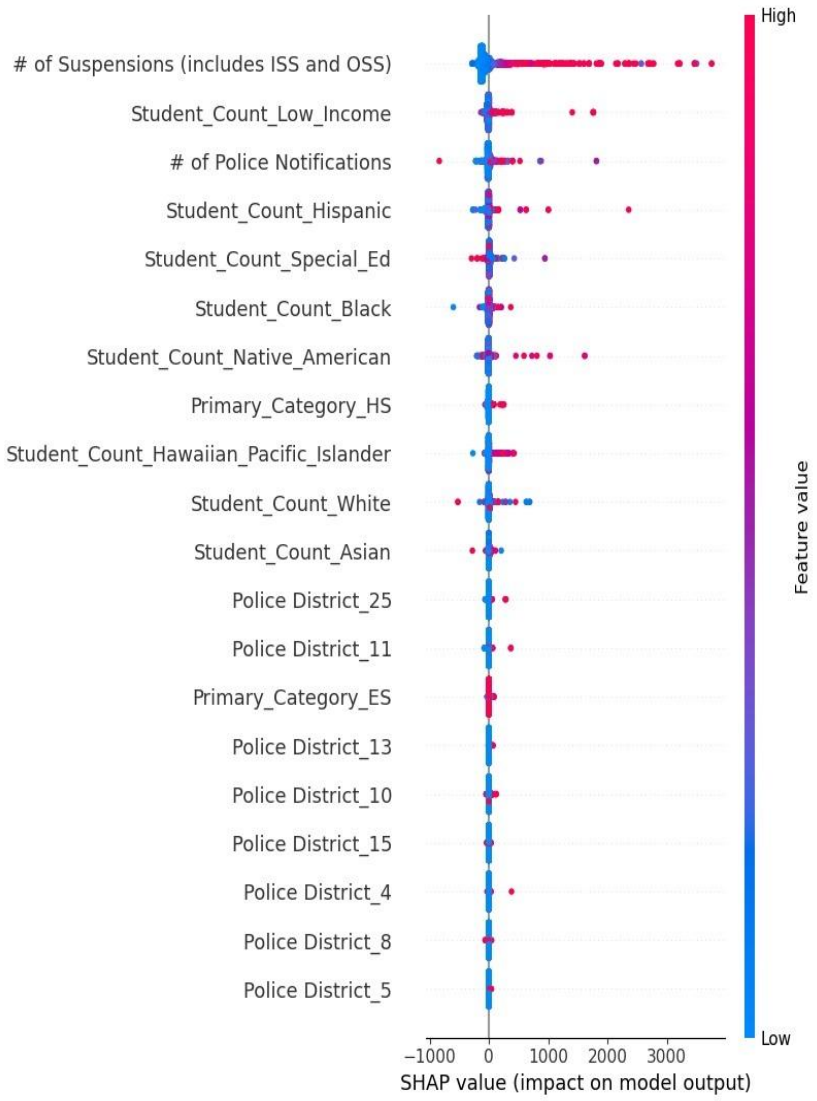
The goal of the model was to reliably predict the number of reported misconducts within the Chicago Public School system when presented with police district, reported incident, and student demographic/behavioral data.

With a single predictor variable (the number of reported misconducts), the initial model choice was the basic Linear Regression model which provided the following scores when using the default parameters: Training accuracy 0.472, testing accuracy 0.453, mean cross-validation score: 0.49, mean square error: 168093.31, root mean square error: 409.99. Having significantly few parameters to tune, it's no surprise the tuned model performed in an equally underwhelming fashion.

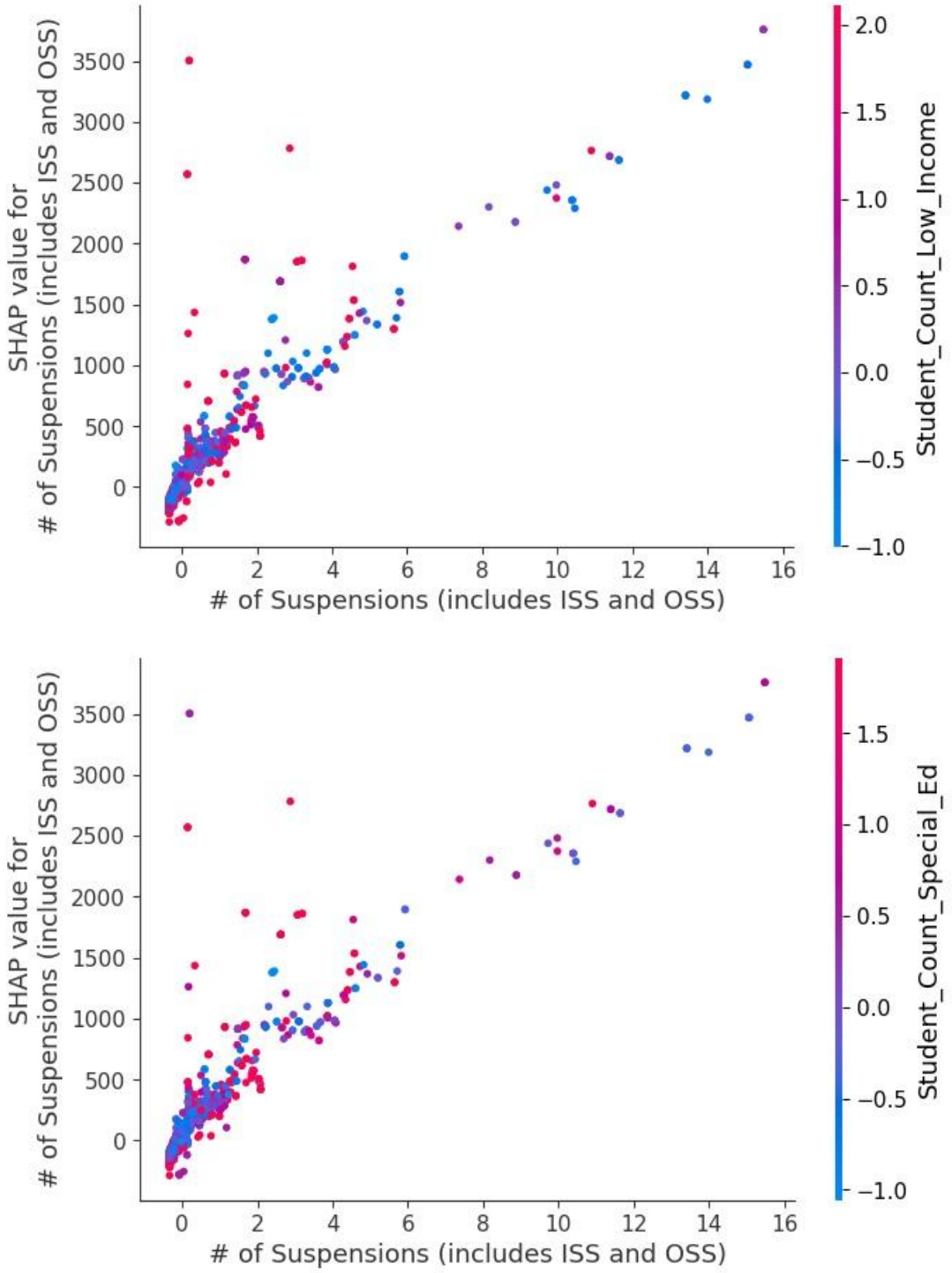
In an effort to locate a better model, the following algorithms were tested using default parameters for each: Ridge, Lasso, KNeighborsRegressor, RandomForestRegressor, XGBRegressor, and DecisionTreeRegressor. XGBRegressor provided the most ideal initial scores, so it was chosen as the new model.

After tuning a handful of the model's parameters (max depth, minimum child weight, gamma, subsample, and colsample by tree) using GridSearchCV, the following scores were achieved with the tuned model: Training accuracy 0.997, testing accuracy 0.978, mean cross-validation score: 0.89, mean square error: 6776.98, root mean square error: 82.32. The tuned model performed marginally better than the baseline model with the following changes: Testing accuracy increase of 0.009, mean cross-validation score decrease of 0.01, mean square error decrease of 2638.07, and a root mean square error decrease of 14.71.

Shapley Additive Explanations (SHAP) was used to delve into the inner relationships of the model and how the features impact the model's generalization process.




The above SHAP Summary Plot was created using the scaled X_test values and shows just how dominating of a feature the number of suspensions is within this model, but it's not much of a surprise that there's a connection between the number of suspensions and the number of reported misconducts. Similarly, it's not much of a surprise that the count of low income students has an effect on the model, but it is interesting to note how important certain outliers are to their respective rows.




The above SHAP Dependence Plots show the relationship between the actual value of a particular feature (in this case the number of suspensions) with how that value affected the prediction. The upward slope implies that as the number of suspensions increases, the higher the model's prediction is for having misconducts being reported–which makes sense. The spread

implies that other features seemingly interact with the feature (number of suspensions), while the color shows how they relate with the interaction feature (low income in the top plot and special ed in the bottom).

While a model was successfully created to reliably predict the number of reported misconducts within the Chicago Public School system when presented with police district, reported incident, and student demographic/behavioral data for the test set (with ~15k rows), the model was unable to be ran on the full data set (~15 million rows) due to insufficient memory.

The results of the model show a need for three further analysis projects:

1. A computer with more memory to run the model on the full data set to see if it performs as well on the full set as it did on the test set.
2. A qualitative analysis regarding the connections and biases between the number of misconducts and how they relate to low income, special education, and racial demographics while being especially careful not to draw incorrect conclusions.
3. A quantitative analysis exploring the connection between the number of misconducts and how they relate specifically to the police districts that had the highest impact on the model: 25, 11, 13, 10, 15, 4, 8, 5 (ordered by relevance/impact to model).

## V. Conclusions

Crime in Chicago is decreasing. Rates of nearly every crime have declined from 2013. Motor vehicle theft and assault (non-domestic) are the only crimes that are increasing in frequency. Homicide rates are up from 2013, but homicide appears to be decreasing in frequency after a rise in 2021. Despite Chicago's reputation as a dangerous, crime-infested city, things are looking up for Chicago in terms of frequency of crime. However, students overwhelmingly do not feel safe in Chicago public schools. There is a small, negative correlation between school graduation rates and crime in Chicago. There is a moderate, positive correlation between the number of crimes and the number of misconducts/suspensions/expulsions in public schools in Chicago. These correlations suggest that crime rates do affect education and specifically, the feelings of safety, performance, and behaviors of students. More data and further analysis is required.

This capstone project was successful in its goals of creating a data streaming pipeline that would record recent reports of crime and analyze past records to understand crime trends in Chicago. The project also effectively completes a correlation analysis of crime statistics to students' behavior, performance, and safety in Chicago public schools to understand if and how crime affects education in Chicago.

The project identified several trends in regards to crime in Chicago, as well as some interesting correlations between crime and students' safety, behavior, and performance. Although these correlations were not statistically significant, they still presented interesting ideas to consider and raised more questions about how students are affected by the presence of crime around them. The capstone group experienced several limitations in terms of data acquisition. The group found a company that conducted surveys in Chicago public schools that would have greatly supplanted the quality and quantity of data available regarding how students feel about safety and crime, but the company only provides those responses to the Chicago public school system. Elementary schools are also overrepresented in the datasets that are publicly available, as elementary schools make up a large portion of schools in the public school system. High schoolers would have provided better data for the topics the capstone focused on, but they make up a much smaller portion of public schools. Despite these limitations, this capstone group accomplishes all its goals while presenting interesting and relevant findings that can potentially guide public policy in Chicago.

## VI. References

Crimes 2013-2023 City of Chicago Data Portal

https://data.cityofchicago.org/Public-Safety/Crimes-2013/a95h-gwzm
https://data.cityofchicago.org/Public-Safety/Crimes-2014/qnmj-8ku6
https://data.cityofchicago.org/Public-Safety/Crimes-2015/vwwp-7yr9
https://data.cityofchicago.org/Public-Safety/Crimes-2016/kf95-mnd6
https://data.cityofchicago.org/Public-Safety/Crimes-2017/d62x-nvdr
https://data.cityofchicago.org/Public-Safety/Crimes-2018/3i3m-jwuy
https://data.cityofchicago.org/Public-Safety/Crimes-2019/w98m-zvie
https://data.cityofchicago.org/Public-Safety/Crimes-2020/qzdf-xmn8
https://data.cityofchicago.org/Public-Safety/Crimes-2021/dwme-t96c
https://data.cityofchicago.org/Public-Safety/Crimes-2022/9hwr-2zxp
https://data.cityofchicago.org/Public-Safety/Crimes-2023/xguy-4ndq

Chicago Public Schools Profile Information - School Year 2022-2023

https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Profile-Information-/9a5f-2r4p

Chicago Public Schools - Progress Report Cards (2011-12, 2022-23).

https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t

https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY2/d7as-muwj

Chicago Public Schools - Metrics

https://www.cps.edu/globalassets/cps-pages/about-cps/district-data/metrics/misconduct_report_police_and_expulsion_thru_eoy_2022_school_level.xlsx

https://www.cps.edu/globalassets/cps-pages/about-cps/district-data/metrics/misconduct_report_police_and_expulsion_thru_eoy_2022_district_level.xlsx

SHAP - Basic SHAP Interaction Value Example in XGBoost

https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Basic%20SHAP%20Interaction%20Value%20Example%20in%20XGBoost.html

Sci-kit learn - Documentation

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html