Analyzing Crime Data for Campus Safety Enhancement

Kansas City Public Schools

A Fictional BI Project

COSC 6510 – Data Intelligence

Jennifer Sailor

Introduction

Ensuring the safety and security of our educational community is important at Kansas City Public Schools (KCPS). “Our main focus and goals are to ensure the safety and security of our students, teachers, administration, parents, and visitors of KCPS” (1). However, despite our commitment, we face a significant challenge: the ability to meaningfully interact with the amount of data available to us.

While data is abundant, its value in informing decision-making within our school system is limited. Until now, the engagement with data has yet to expand to enhancing safety and security measures. Over the years, the Kansas City Police Department has collected and maintained crime datasets that are available to the public (2). Yet accessing and visualizing this data is filled with complexities. Consequently, KCPS deployed a business intelligence (BI) task force for assistance in finding interventions rooted in statistical evidence that yield safety strategies. Through this initiative, the BI project will facilitate informed decisions regarding security allocation, evaluate the effectiveness of security enhancements implemented over recent years and determine optimal locations for summer school programs based on crime risk assessments.

In essence, the BI project represents a pivotal step towards realizing KCPS overarching goal of fostering safe learning environments. By leveraging data-driven insights, we can proactively address safety concerns, ultimately ensuring the well-being and success of all members of our educational community.

Business Intelligence Methods

Transitioning into the implementation phase of the project, it’s crucial to assess our readiness and clearly define a plan to reach our goals. Throughout the 3-month project I (the team) utilized R Studio 4.3.2 and Tableau. Our timeline starts in February, the focus lies on data collection, cleaning, integration, and collaboration with local law enforcement and the KC public school board. March is dedicated to visualizations, trend identification, and statistical evaluations. April will see the development of statistical models, dashboard creation, presentations, and stakeholder engagement. Anticipated outcomes include comprehensive

crime trend insights, geospatial visualizations, and a dashboard for displaying results to business constituencies.

Success in business analytics hinges on planning, motivation, and a commitment to excellence. The team possesses significant expertise in R programming, data cleaning, visualization techniques, regression analysis, and statistical modeling, boosted by three years of experience in the field. While our team has strong analytical skills, there is room for improvement in specific areas such as crime data analysis, handling big data sets, and maximizing the functionalities of Tableau. To bridge these gaps effectively, we will embrace a culture of continuous learning and mentorship from seasoned advisors. Additionally, we remain committed to staying organized and adhering to the roadmap outlined for our BI project.

When analyzing data, it is important to have a solid grasp of the content and structure. There are two datasets used in this project. The first being the Kansas City Crime data set. The dataset originates from 10 individual data sets spanning a little over nine years from January 2015 – March 2024 (2). After preprocessing which details are explained below the final crime data set had 10 columns: Reported_Date (of the crime), Description (describing the type of crime), Age (of criminal), Address, City, Zip.Code, Rep_Dist, Area, Latitude, and Longitude (describing location of crime) and 1.04 million rows. The second data set is department generated of all the KC Schools within the KCPS system (Appendix A). It contains 8 columns: Level (whether school is primary or secondary education level), SchoolName, Address, City, State, ZipCode, Latitude, and Longitude (describing location of school), and 36 rows. When obtaining open-source data like the KC Crime data poses no obstacles as it is readily available, facilitating easy utilization. While generating our own KC School data was imperative as it was not available online, resulting us to undertake the time extensive task of creating the data file found in Appendix A. Dealing with open-source versus closed-source faces different challenges. While closed-source takes time in creation, open-source takes time in validation of correctness and format.

When preparing the KC Crime individual datasets for combination difficulty arose when it came to the column names and types of data. Many columns contained the same information

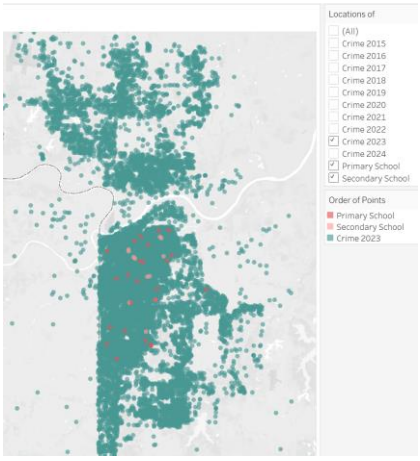
but were a different data type or the column name was different. The most difficult discrepancy was the latitude and longitude columns because of the three varying formats found across the datasets. The explicit code to get the coordinates in the same format can be found in Appendix B starting on page 2. Once all formats were correct the individual data sets were easily merged and cleaning took place. Cleaning the data involved standardizing missing values, standardizing date values, correcting implausible age entries, and validating geospatial coordinates within predefined bounds around KC.

The now preprocessed data holds significant importance for the BI campaign, specifically for its relevance to crimes occurring in proximity to schools, as illustrated in Figure 1. This figure, generated in Tableau, visually shows the spatial relationship between 2022 Crime data and school locations, forming the foundation of our project. With this groundwork established, the next step involves conducting descriptive statistics to gain insights. Additionally, confirming the presence of the seasonal trends within the crime data is pivotal. Figure 1 showcases these trends, indicating clear patterns on both a monthly and a weekly basis. With this knowledge, inferential statistics, and regression analysis, can now be pursued with confidence.

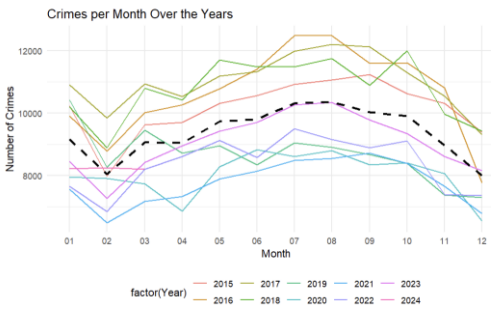
To conduct the results three types of methodologies were used. Initially descriptive statistics were used to comprehensively summarize key data features, including measures such as total, average, and standard deviations. The use of these can be seen in Figures 1,2, and 3. These metrics give a clear picture of central tendencies and dispersion, which is essential for subsequent analyses. Moreover, inferential statistics were used to get broader conclusions. Using confidence intervals to evaluate hypothesis testing. The use of this can be seen in Figure 3. Lastly, regression analysis plays a crucial role in finding temporal trends and seasonality patterns within our dataset. By examining the relationship between variables like time and frequency of crime, we gain nuanced insights into the temporal dynamics as seen in Figure 4. Although time series analysis offers deeper insights, it falls beyond the scope of the project. Hence, we rely on polynomial regression which is known to work well with seasonal data and was used over linear regression due to our data's nonlinear pattern, shown in Figure 4. Evaluating model performance through metric such as R-squared and mean squared error

(MSE) ensures the accuracy of our predictions, shown in Table 1. Through these methodologies, we empower business constituencies with actionable insights.

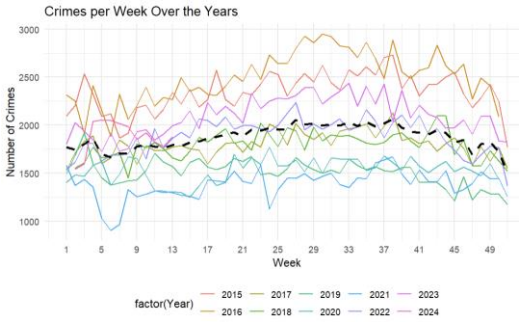
Figure 1:



Appendix E: Tableau: Map of all the locations of crime selected for 2023 (Teal) and locations of KC Schools (Red and Pink).



Appendix C – Page 7: Amount of crime per month grouped by year which crime occurred. The dashed line is the average of all the years.

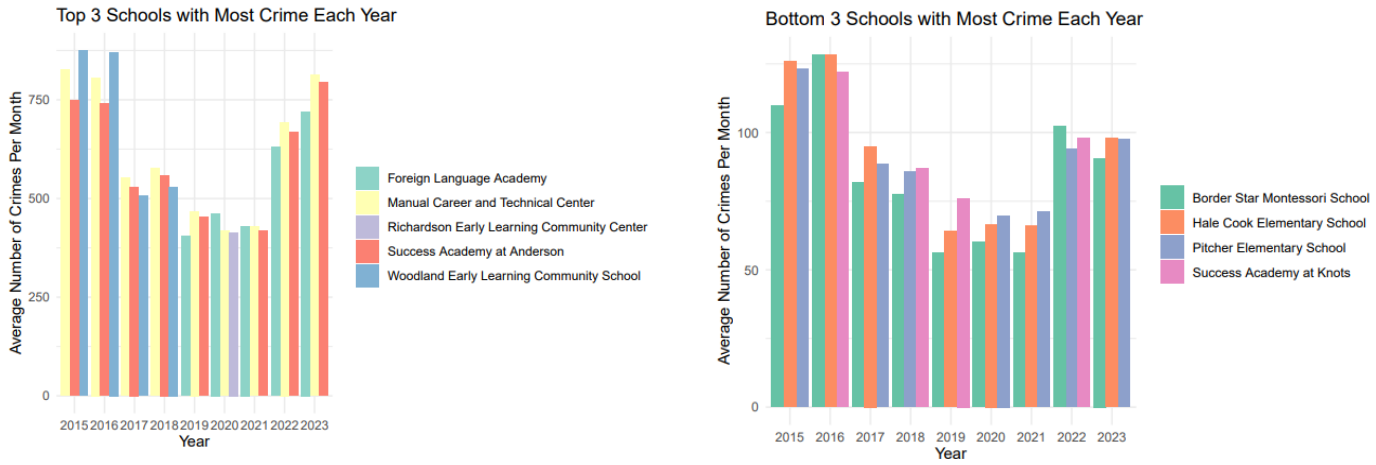


Appendix C – Page 19: Amount of crime per week grouped by year which crime occurred. The dashed line is the average of all the years.

Results

In this section, we use descriptive statistics to analyze crime risk associated with schools within our dataset. The focus is on identifying schools with the highest and lowest levels of crime risk. To achieve this, we calculated the total number of crimes occurring within a one-mile radius of each school for every month of available data, subsequently deriving an average monthly crime value for each school across each year. The analysis reveals the top three schools with the highest crime risk and the bottom three with the lowest risk for each year, as depicted in Figure 2. Notably, certain schools consistently rank among the top or bottom three over the nine-year period. Those consistently ranking among the top suggest the importance of focusing on these schools for targeted crime prevention efforts. Conversely, schools with consistently low crime risk levels present opportunities to maintain current safety protocols effectively. These findings offer valuable insights for strategic decision-making and resource allocation in enhancing campus safety measures.

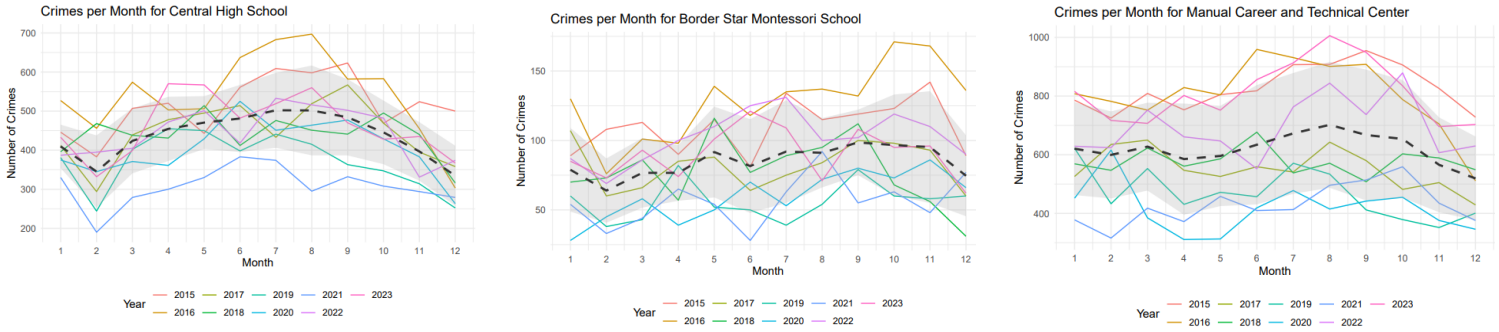
Figure 2:



Appendix D – Pages 20-23: A bar plot of the year (X) versus the average number of crimes per month (Y). The graph displays either the "Top" or "Bottom" 3 schools for each year. Top is defined as the highest and bottom defined as the lowest based off there Y value.

Utilizing inferential statistics, our aim was to find any significant changes in the level of crime surrounding specific schools. For this analysis, three schools were selected for analysis, though this selection can be tailored as necessary: Central High School, serving as the primary high school for KCPS; Border Star Montessori School, consistently exhibiting low crime rates year after year (Figure 2); and Manual Career and Technical Center, frequently experiencing higher crime rates (Figure 2). Figure 3 illustrates the monthly count of crimes grouped by year for each school, with a black dotted line denoting the average and a grey banded area representing the confidence interval. Notably, across all three charts, instances where data points fall outside the confidence interval bounds are observed, particularly in 2016, a year marked by elevated citywide crime rates, and 2021, a period characterized by notably lower crime rates. While this information proves insightful in identifying potential abnormalities in crime trends near specific schools, its utility in hypothesis testing regarding the effectiveness of safety protocols is limited. Since the safety measures implemented by schools do not universally impact all crimes within a one-mile radius, adjustments such as reducing the distance from the school or devising alternative metrics are necessary for more accurate assessments.

Figure 3:

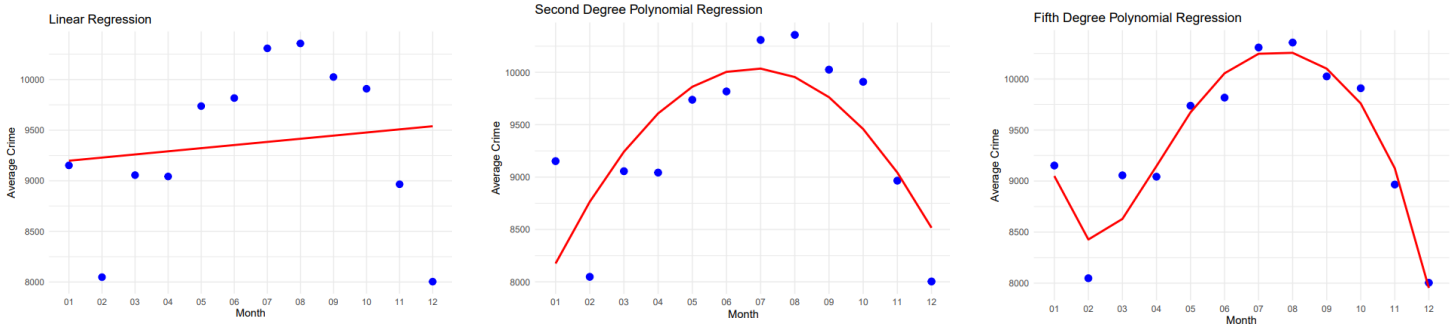


Appendix D – Pages 14-17: The line graph shows the month (X) versus the total number of crimes within 1 mile (Y) grouped by year. Each graph shows a different school and is in the school's name is in the title of the plot. The black stripped line is the average, and the grey band is the confidence interval.

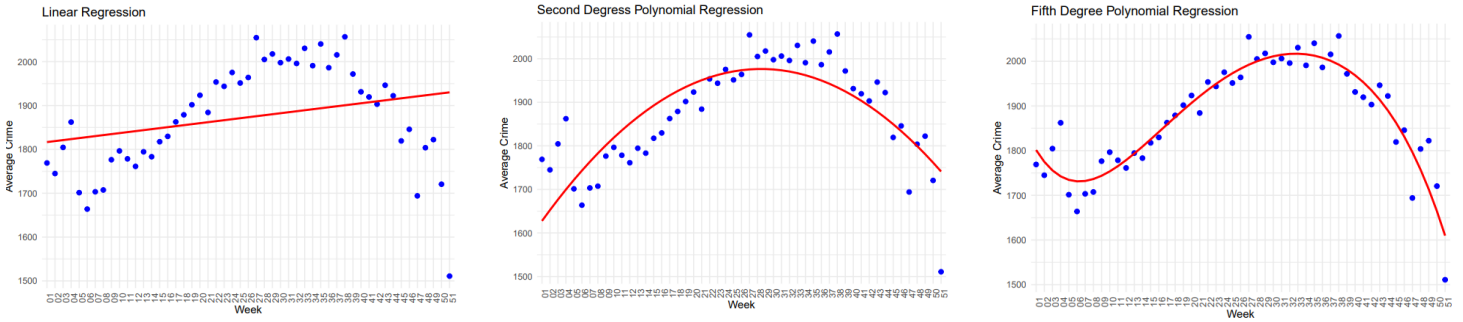
Lastly, I utilized regression analysis to model the seasonal patterns in crime data. By considering both monthly and weekly frequencies of crime as independent variables and the average crime rate in KC as the dependent variable, I compared three regression models visually and numerically, as depicted in Figure 4 and Table 1. It became evident that linear regression was unsuitable due to the data's nonlinear pattern, as indicated by poor MSE and adjusted R-squared values in Table 1. Consequently, I explored second and fifth-degree polynomial regression models, which exceeded those of linear regression. Despite the increased complexity of the fifth-degree model, it outperformed the second-degree model both visually and numerically, boasting lower MSE and higher adjusted R-squared values exceeding 0.8, indicative of a well-fitted model.

Fitting this regression model gives valuable insights for business decision-making. For instance, identifying peaks and valleys in crime occurrence throughout the year allows for strategic planning, such as scheduling summer school sessions at low-risk locations, using suggestions from Figure 2. Moreover, these regression analysis results pave the way for future endeavors in machine learning applications, potentially enabling the prediction of future crime rates, and extending the analytical capabilities of the company beyond the current project.

Figure 4:



Appendix C – Pages 13-17: The charts show the month (X) versus the average crime count (Y). The blue points are the actual values, and the red line is the regression line found using the technique described in the title of the plot.



Appendix C – Pages 21-24: The charts show the week (X) versus the average crime count (Y). The blue points are the actual values, and the red line is the regression line found using the technique described in the title of the plot.

Table 1:

	Linear Regression	Polynomial Regression – Degree 2	Polynomial Regression – Degree 5
Complexity	$y = \beta_0 + \beta_1 x$	$y = \beta_0 + \beta_1 x + \beta_2 x^2$	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_5 x^5$
Frequency of Crime is Monthly			
Adjusted R^2	-0.07825	0.5337	0.8736
MSE	567,082.8	220,713.3	39,899.7
Frequency of Crime is Weekly			
Adjusted R^2	0.05886	0.6225	0.8637
MSE	13,245.3	5204.8	1761.8

Appendix C – Pages 13-17: The table shows the results of the 3 types of regression used on both the weekly and monthly crime data.

Conclusions

In reflection, the analysis revealed certain weaknesses coming from model assumptions and data analysis limitations. The selected analyses were justified by the need to address seasonal patterns evident in the data, with polynomial regression chosen for its suitability with nonlinear data. While polynomial regression was utilized to capture seasonal patterns, the

project could have benefited from a more nuanced approach such as time series analysis, offering deeper insights into temporal trends. Although time series analysis would have offered deeper insights, its scope fell beyond the project's constraints. Another weakness comes from the lack of knowledge regarding the optimal radius threshold to use for the proximity of crimes to schools. Then lastly, implementing R Studio visualizations in Tableau was rather difficult therefore only Figure 1 was created on the dashboard. However, despite these limitations, the project has provided valuable insights into enhancing safety measures within Kansas City Public Schools.

Moving forward, these insights will inform decision-making processes, facilitating informed security allocation and assessing the effectiveness of past security enhancements. Furthermore, the project sets the stage for future advancements in safety and security measures, positioning KCPS to continually refine protocols and create a secure learning environment. Looking beyond the educational community, the project's insights hold potential impact to surrounding neighborhoods and the city as a whole. By leveraging data-driven approaches, these initiatives can serve as models for collaborative efforts to address safety concerns across diverse communities. Ultimately, the project highlights the connection of safety initiatives within the societal context, emphasizing the importance of action in creating safer environments for all.

References

1. "Safety & Security." Kansas City Public Schools, https://www.kcpublicschools.org/about/departments/safety-security

2. "KCPD Crime Datasets." DataKC, https://data.kcmo.org/browse?limitTo=datasets&q=crime&sortBy=relevance .

Appendices

A read me is attached within folder to help with navigation of all files.

Appendix A – KC Schools Data

Appendix B - Preprocessing File

Appendix C - Analysis 1 File

Appendix D - Analysis 2 File

Appendix E - Tableau File