**INCOME RENT AND VIOLENT CRIME**

Mathew Cobb, Tyer Megill, William Robinson

Department of Computer Science, University of West Florida

CAP 4770: Data Mining

Dr. Shashi Bhushan Jha

April 25, 2025

**INCOME RENT AND VIOLENT CRIME**

**Problem Statement and Background Information**

Crime rates significantly impact community safety and resource allocation for law enforcement. Violent crime rates are a concern for urban safety, influenced by factors such as income and cost of living. This project aims to analyze and predict violent crime rates in Aspen, Colorado, and Atlanta, Georgia, from 2013 to 2019, using median individual income and rent indices as predictors. Understanding how individual income and rent indices correlate with crime rates, law enforcement and city officials could develop strategies to reduce the crime rate. This analysis provides insights into the relationship between economic conditions and crime, enabling data-driven decisions for community safety.

**Data Sources and Collection**

For our data sources we gathered our data for Median individual income from Data Commons - who in turn sourced their data from the U.S. Census Bureau. The data came in the form of Excel spreadsheets, one file for each of Aspen, Colorado and Atlanta, Georgia. These xls files each contained data points for varying years, and we focused on rows denoted by the year column for 2013 - 2019. This time period was selected due to our other data sources being restricted in viable data points which will be described later. The Median individual income xls files supplied by Data Commons had already been pre calculated by the Census.gov original data source and Data Commons presented already processed and cleaned data for those years we selected.

For Historical rent index data, we sourced from FRED - Federal Reserve Bank of St. Louis - Economic Data. These source files were also formatted in Excel spreadsheets, one for each of Aspen and Atlanta. This Dataset contained monthly average rent values for all years between 1917-present. They sourced the initial data from the U.S. Bureau of Labor Statistics - FRED and preprocessed for xls format.

Our last data source was the most complicated as it had the least amount of preprocessing made by the initial sourcing agency- U.S. Department of Justice, Federal Bureau of Investigation, Criminal Justice Information Division (FBI.UCR). The interactive portal allows searching of all cities in the US and retrieving historical crime data from all DOJ agencies in the pertaining jurisdiction. To gather this data for our project, we had to search for both Atlanta and Aspen historical crime data independently and further source for the individual years independently. This did not utilize an API for easy data querying. We had to download individual csv files for each year for each city (14 different source files).

**Data Preprocessing**

As some of our source datasets from Data Commons (for rent and income) had already been preprocessed by the sourcing agency, we could easily convert those csv datasets into a SQLite database in Python. We read these csv files into a Python datastream for automatic conversion into dataframe using Pandas dataframe conversion method. As the source files only contained the columnal fields we needed for our following analysis, cleaning and parsing the rows was straightforward and only required slightly altering column headers for cross file matching. Once this initial Pandas dataframe was created, we used Pythons built in sqlite3 conversion tool to upload the pandas dataframe into a locally stored Database for further querying.

The multitude of xls files we retrieved from the FBI.UCR for historical crime required a more methodical approach to processing and conversion. These base files each contained rows for all municipalities of all cities in the United States for the individual year of collection. This required initial cleaning and reformatting of the source spreadsheet for the column headers. As each individual agency of every municipality in the US had independently added data to the files, the layout of the data points we collected varied by column names, formatting, column indexing and just general styling. For us to later generate a Database for these sources, it required cleaning of all column headers across all files and redefining for acceptable column identification. We did this both by manually altering the headers of the xls files directly and once read into python, further regex parsing of fields for data type merging.

Once all the source files for Crime Data were clean enough to upload to the python notebook, and further preprocessed, we staged them for merging by converting each years data into a pandas dataframe and merging these frames by filtering for our desired city names (row primary keys) and generating a new dataframe for both Atlanta and Aspen only sorted by year. Once this merged frame was finished we could convert into a local SQLite database as well for further querying and analysis.

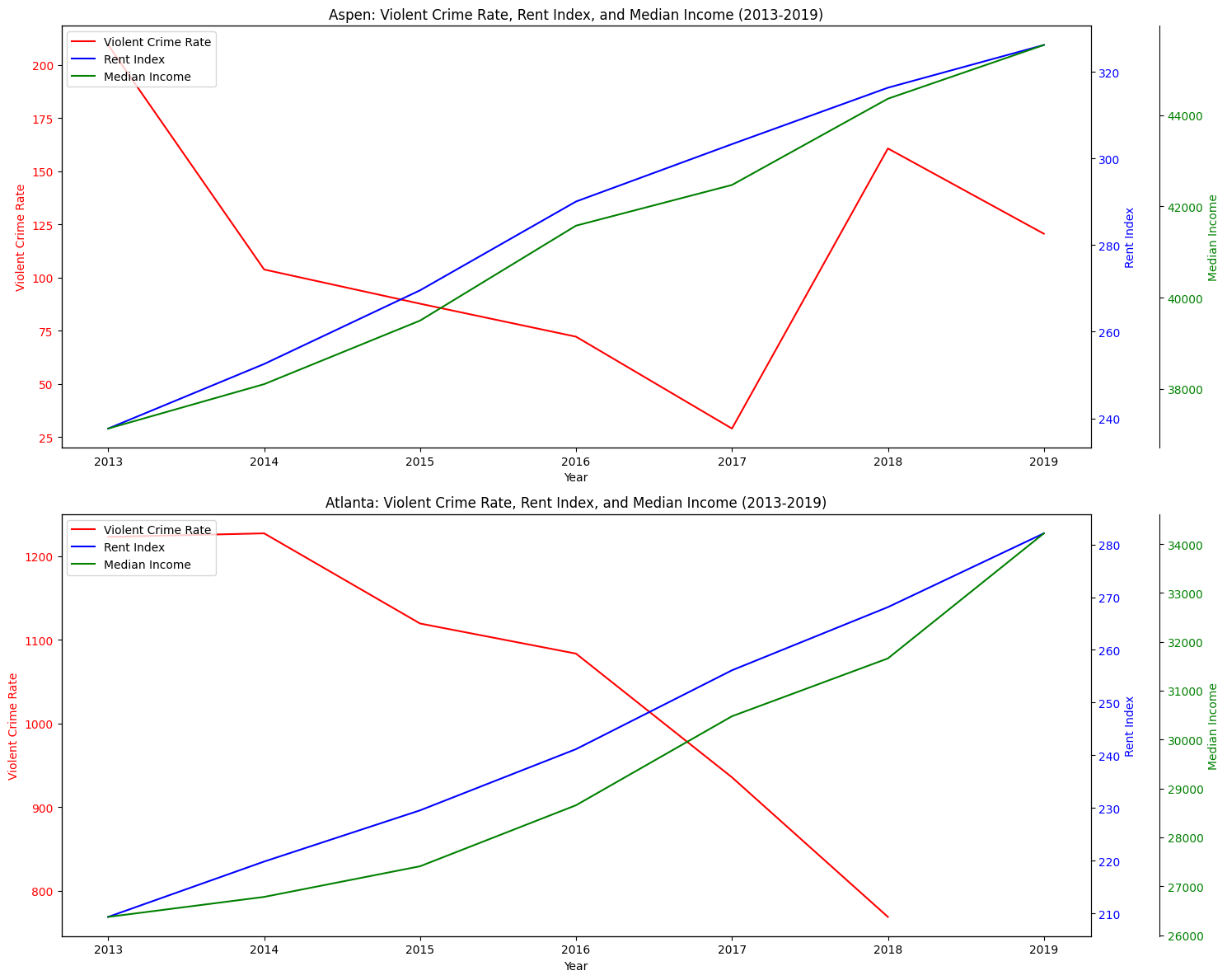
**Data Analysis and Exploration**

Once we had all three data sources and variables for each of our defined cities, we could analyze and plot these data points for initial model analysis. Taking the sqlite Databases, we generated datatables for each variable and queried by the selected years (2013-2019) because these years presented the most concurrent data points from all tables for both cities. This prevented data loss from being entered into our analysis. With each variable we plotted their historic data points on Numpy line plots and generated a Pearson Line of Best fit to analyze the yearly trends for these factors.

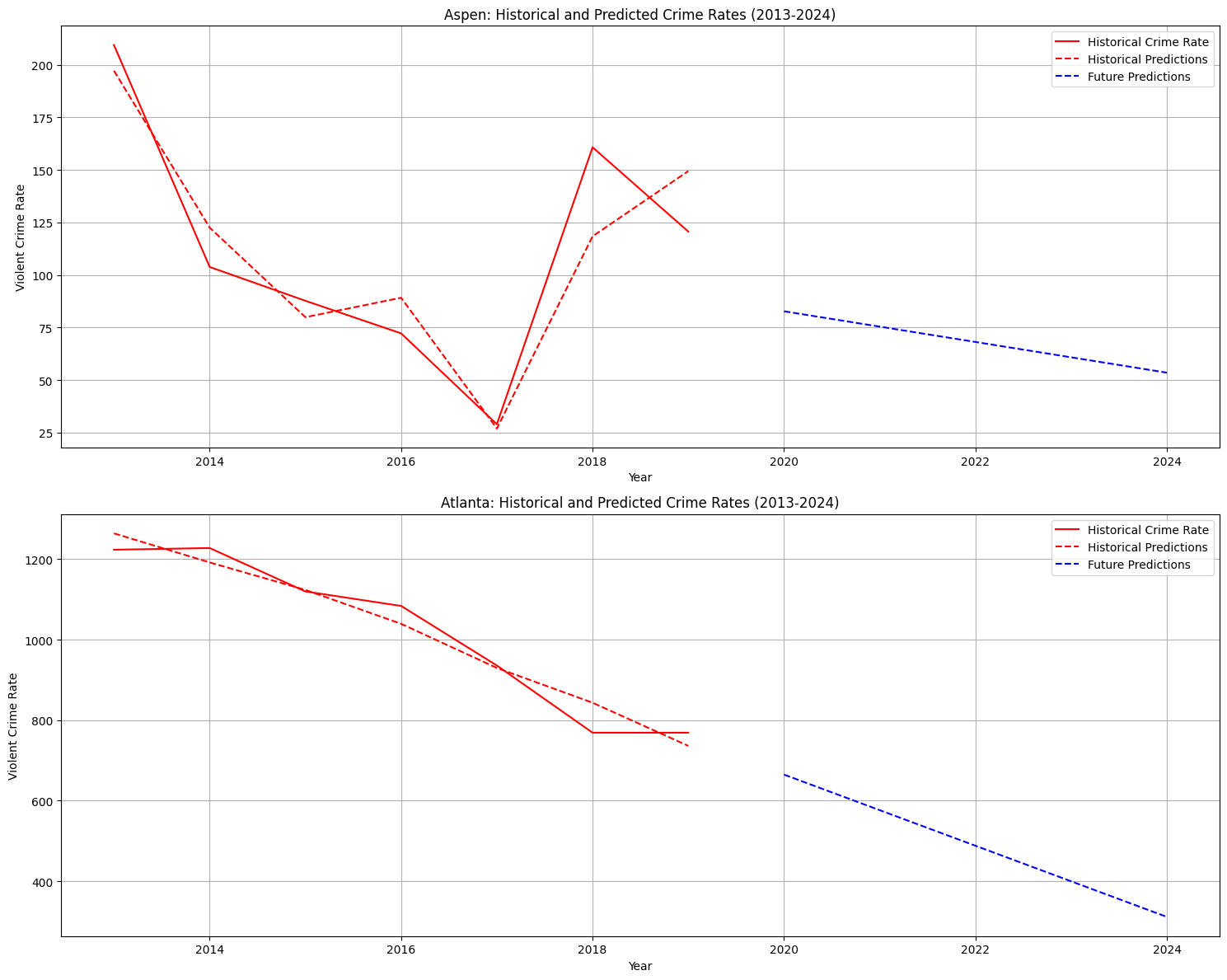
Once a plot was generated for all factors for each city, we merged the plots to produce an overall data visualization for our included datasets which allowed generalized analysis. This initial visual model showed promising results for our hypothesis but still required further modelling and statistical analysis.

**Data Modelling and Statistical Summaries**

To model the relationship between violent crime, median income, and rent index, we employed linear regression analysis using the scikit-learn library in Python. We also used matplotlib for our correlation graphs showing the relationship between all three factors. Our primary objective was to evaluate how well median individual income and rent index can predict the frequency of violent crimes in Aspen CO and Atlanta GA from 2013 to 2019. The correlation graph as shown below gives a great visual on showing how violent crime continues to decrease as economic positions rise. These graphs show a very real correlation between rent, median income, and violent crime, our data shows close to a 98% match from how much median income is to the rate of violent crime.



Our predicted crime rates continue to show that if we have continued economic growth we will continue to see declining violent crimes.



**Results and Interpretation**

Our analysis suggests that median income and rent index are both strong predictors of violent crime rates in Aspen and Atlanta from 2013 to 2019. In both cities higher rent index and median income levels were associated with lower violent crime rates confirming our initial hypothesis.

The linear regression models we developed not only revealed a statistically significant negative relationship between economic success and violent crime but also offered a predictive view into future trends. Based on our graphs, if the economic indicators continue their upward trend we can expect a gradual decline in violent crime over the next five years in both locations.

These findings align with existing research on urban economics and crime, suggesting that improved living conditions and economic opportunity may contribute to reductions in crime. However it’s important to note that crime is influenced by a wide array of social political and structural factors not included in our model.

**References and Citations**

Palmer, C. J., & Pathak, P. A. (2017). *Gentrification and the amenity value of crime reductions: Evidence from rent deregulation* (No. w23914). National Bureau of Economic Research.

Zhang, Z., & Barr, A. (2024). Gentrification and crime in Buffalo, New York. PloS one, 19(6), e0302832. https://doi.org/10.1371/journal.pone.0302832

Tyler Megill:

I played a critical role in our data mining project by leading the effort in sourcing, cleaning, and aligning key datasets across multiple variables including violent crime rates, rent index, and median income for two cities: Aspen and Atlanta. I ensured that our datasets were consistent across timeframes by developing an interpolation-based preprocessing pipeline. This allowed us to align the data temporally and fill in any missing values, which was essential for accurate analysis. I also created a comprehensive visualization using Matplotlib that compared all three metrics for each city in a single view. This helped our team easily observe trends and potential relationships over time.

To support our predictive analysis, I helped developed a linear regression model to examine how rent index and median income influence violent crime rates. Linear regression was chosen because it provides a straightforward yet powerful way to quantify relationships between continuous variables, which is ideal for understanding socio-economic trends over time. This model allowed us to both evaluate historical data with R² scoring and project future crime rates based on observed trends in income and rent. By generating clear coefficient interpretations and visual forecasts through my code, we could make data-driven conclusions about how economic changes may impact public safety in these cities over the next five years.

**Matthew Cobb**

As the writer for the proposal my duties were to outline the goals of the project from conception and define what our data sources would be referenced from. I chose where we sourced our data from, as well as initializing the steps for data collection and preprocessing of my sources. Once I had all the files cleaned to the point of database retrieval, I wrote the code for database and dataframe creation for such. These sqlite databases I created where the primary roots for all of the team’s data analysis.

As well as initializing the core framework for data sourcing and processing and cleaning, I also designed the extrapolation algorithms for our datatable merges to parse through the multiple databases and collect only applicable data for our projects needs. This was a critical step in data cleaning as it allowed the other team members to focus on analysis without having to worry about false, missing, and inadequate data entries.

These steps I took to ensure we had a clean and full database preprocessed for easy data retrieval meant that we could formulate our correlation and modelling algorithms to train only upon the most viable of datapoints and allowed all of our dataframe generation to work across 20+ source files without the need of further parsing and header manipulation. This spread up the process for all members of the team to focus on analysis, modelling, and summarization of the transformed data only after my work had completed on the sourcing and processing of all source data into usable and retrievable databases.

**William Robinson**

In this group project we were tasked with selecting a problem statement of which would interest us. We picked a statement in which we believed that we could investigate and address with the knowledge obtained in this class. To start our project our group gathered to discuss potential topics. We reached the conclusion of taking a look at the effects of gentrification on the crime rate of two different cities in the United States. After discussing a potential hypnosis, we then each searched the web for any publications to support our claim. Once we found our supporting documentation we started the search for usable data sets. I helped look for the data sets for Atlanta crime data. For the database portion of the project, we decided to use SQLite 3. Once the data was uploaded, we transferred all the excel files to database files. Next, we worked on the coding part. Once our coding part was done, I got started working on a little power point presentation to go along with our live data walkthrough. Once the presentation was complete we started to tackle the final report part of the project. For this part we broke the report into different sections and worked on the section in which we had the most knowledge for. It was fun taking a look at the results of our analysis and how it compares to our original hypothesis. In this project we learned about the basics of data mining with python. We also learned about data cleaning and clustering.