REAL ESTATE AND CRIMINALITY

Amelia Maria Acuna Rodriguez - *910105*

Università degli Studi di Milano-Bicocca

Introduction

This project is intended to analyze the association between real estate prices and crime rates in various U.S states and cities. In this report the process of data collection, cleaning, merging, visualization, and correlation analysis, which are all essential steps in exploring such a relationship will be detailed. This type of study may be useful for prospective homebuyers, real estate investors, and urban planners.

The research question "How do crime rates influence real estate values in different U.S states and cities?" directs the inquiry, attempting to measure the effect of crime on property prices.

Data Acquisition

Web Scraping

Web Scraping is a technique used to collect data from websites that can be done through manual selection or it can involve automated crawling using pre-programmed scraping softwares and primarily serves to collect textual data (Research Data Service, University of Wisconsin-Madison 2019). Depending on the website and its content one should choose between static or dynamic scraping. For the former, the content is fixed and does not change based on user interaction, for the later one their content is generated or changed in response to user interaction (AndH 2023)

For this study, automatic and dynamic scraping using Python targeting real estate from Zillow is used, capturing information for each property on: *id_location, price, state, bedrooms, bathrooms, and sq_ft*.

Table 1: Libraries needed in Python to perform Web Scraping

Library	Package/Module	Description
undetected_chromedrive	N/A	Mimicking regular browser's behavior to evade bot-detection.
Selenium	WebDriver	Automates web browser interactions, locates elements and navigates through the listings.
	common.keys (keys)	
	common.by (by)	
	support.wait (WebDriverWait)	Wait for the web page to be valid before proceeding, e.g elements should be loaded before interacting with them.
	support (expected_conditions)	
	webdriver.chrome.options (Options)	Conditions were Chrome should wait,
time	N/A	Built-in Python module for handling time-related functions. Here, ensure the page is loaded and pace actions.
re	N/A	Built-in Python module for handling regular expressions. It finds patterns such as pricing zipcodes, and other information.
mysql.connector		Connects to MySQL to perform database operation (CRUD)
sys	N/A	Exit the script in case of failure.
traceback		Act like a diagnostic tool, that provides information on errors.

Application Programming Interface (API)

API allows the user to request data from third-party providers. To retrieve data related to crime in the United States, the API used for this study is "FBI Crime Data API', collected under the Uniform Crime Reporting program'. It is a RESTful API (Representational State Transfer) which means that it provides a set of URLs called endpoints that the user accesses to obtain the information from the web service, and that it could be included into any programming language that could make HTTP requests, in this case Python.

An example of the JSON document resulting from consuming this API can be found in Appendix 1. In this JSON object we can find different sections, but only a bunch has been selected for this analysis:

State	Two-letter state abbreviation	Ex: AK, FL
Burglary	Breaking into a building to commit theft or another crime	Counts of arrest on the type of crime for 2022
Stolen property	Buying, receiving, possessing, selling, concealing, or transporting any property with the knowledge that it was unlawfully obtained, such as embezzlement, fraud, larceny, or robbery.	
Vandalism	Wide range of actions from graffiti to destruction of property	

From the web scraping, the Zipcode for each property was extracted and used in the API to search with the zip code and save the criminality information that was detailed above.

Table 3: Libraries needed in Python to interact with API

Library	**Description**
requests	Sends HTTP requests to the specified API to fetch crime data for each zip code in the scraped data.
mysql.connector	Connects to MySQL to perform database operation (CRUD)
traceback	If an exception occurs, it provides error information and exit the script immediately
sys	

When evaluating crime data in this project, we face a granularity challenge provided that the main data source, the FBI API, only offers state-level information. While this state-specific data is valuable, it creates constraints to in-depth study by aggregating crime data across wide geographic areas.

This broad approach certainly masks more subtle differences and trends in crime on smaller scales, such as specific cities or neighborhoods. As per simplicity, and using the resources available, an initial approach to increase granulary for future analysis is given by including the following:

United States Crime Rates by City Population Dataset:
The dataset obtained from Kaggle that contains crime rates for cities in United States are formed of four CSV files divided by the population ranges of the cities:

- crime_40 _60.csv (40,000 to 60,000)
- crime_60 _100.csv (60,000 to 100,000)
- crime_100 _250.csv (100,000 to 250,000)
- crime _250_plus.csv (greater than 250,000)

From some search online it is confirmed that at least the table for cities with a population greater than 250,000 is based on the Federal Bureau of Investigation Uniform Crime Report, from 2019. As presenting data from a trust source is valuable in this study, it was decided to do a first approach only with the mentioned table.

The following variables will be taken into account: state, city, population, violent crime, property crime, and burglary. The yearly crime rates are presented per 100,000 people.

Data Storage

For the storage of the data MySQL a relational database model (RDBM) was chosen. Here are some consideration on this decision:

Structured data: From the web scraping the extracted real estate listings contains attributes like price, number of bedrooms, bathrooms, square footage which is well-structured and fits naturally a fixed schema. In the case of the API, although the result is semi-structured, the information needed for this analysis has been picked and organized along with the web scraping.

ACID over BASE: Based on the need for reliable, consistent and accurate analysis it seems to be more appropriate to choose an ACID approach which ensures that every transaction in the database is completed and free of errors (Atomicity), keeps data in a correct and valid state across all relations (Consistency), allows for safe and concurrent processing of multiple transactions (Isolation), and ensures that, once committed, a transaction is permanently recorded (Durability). This method is critical for tasks involving sensitive real estate and criminal data, because users rely on the information's integrity, trustworthiness, and permanence for informed decision-making and analysis.

Data relationships: It handles multiple related data fields efficiently and in a straightforward manner.

Complex queries: MongoDB was considered in a first instance to store the data thanks to the ease in connecting through Python and working with JSON structures. However, MySQL was preferred as it handles joins efficiently due to their relational nature and is less cumbersome when working with operations that require calculation over multiple rows.

Data Processing and Integration/Enrichment

This project incorporates data integration by combining multiple data sources into a cohesive dataset, and data enrichment by enhancing the real estate data with crime rate statistics to bring value and context to the research.

At first a database named *web_scrapper_api_property_data* is created. Within this database, the table *web_scrapper_table* is created to store the extracted information from Zillow with an additional auto-incrementing *id* column. In order to avoid duplicates, the *id_location* is checked before including a new property. If the elements that the script intends to extract are not present, the values are set to *None*. Finally, after each operation of inserting a new record, the transaction is committed as a safety measure to prevent data loss in case of an error.

A new table *api_table* is created in the database to store crime data for a distinct list of states retrieved previously from the web scraping. For preventing duplication, it checks if an entry for each state is already existent. If this is not the case, it sends a HTTP GET request to the API through a function to fetch the information previously mentioned.

The JSON resulting from consuming the API is inspected to find the specific data points mentioned previously in the Data Acquisition section. To handle missing data a default value of 0 is assigned in case the specific key is not found.

Placeholders (%s) are used in the INSERT statement to mark where the data values should be inserted. Later, a tuple that holds the actual and ordered values from the JSON file is made, so when the statement is executed the right information fills the right column in the *api_table.*

In the jupyter notebook, the two tables are merged in the database through MySQL and imported into a Pandas DataFrame, producing a structured data format for analysis.

Additional Data Formatting and cleaning:

- Mapping function is used to convert state abbreviations to full names.
- City names are taken from a location field using regular expressions.
- The dataset from Kaggle on crime rates by city is loaded and integrated with the original DataFrame.
- Missing Data: Rows with missing values (Null) are removed as it is not a considerable
- Unnecessary columns: Irrelevant columns such as *'id'* to streamline the dataset.
- Data Type: The population data from the Kaggle dataset is cleaned up by eliminating commas and converting it to floats, critical for numerical computations that use demographic data.

Data Quality

The quality of the data was assessed mainly through a completeness check and consistency verification.

<u>Completeness:</u>
The percentage of non-null values for each column was estimated with this line of code:

```
100 - web_scrapper_api_property_data_df. \
notnull().mean() * 100
```

id_location	0.000000
price	0.000000
state	0.000000
bedrooms	0.231750
bathrooms	0.231750
sq_ft	4.171495
count_burglary	0.000000
count_stolen_property	0.000000
count_vandalism	0.000000

Key columns such as price and state, among others, are fully completed, promising that the wanted analysis of these factors will not be hindered for the available data.

Both bedrooms and bathrooms columns show a minor amount of missing data (0.23%), which is around 2

entries for the total dataset. The highest percentage of missing values (4.17%) is found in the values of sq_ft that would be equivalent to around 36 records.

It is plausible to conclude that dropping these missing values is a viable approach. as the impact on the analysis of the dataset will be minimal, and simplicity and clarity are prioritized.

Consistency:
The *web_scrapper_api_property_data_df.dtypes* function checks the data types of each column in the main DataFrame. This phase ensures that each column contains the expected type of data, which is required for accurate analysis.

id_location	object
price	int64
state	object
bedrooms	float64
bathrooms	float64
sq_ft	float64
count_burglary	int64
count_stolen_property	int64
count_vandalism	int64

Categorical variables quality: To find the frequency distribution of properties across different states the following line of code was use:

web_scrapper_api_property_data_df['state'].value_counts()

The results confirm that the dataset includes listings from an extensive variety of states, the highest being Florida (FL) with 99, Texas (TX) with 65, and Georgia (GA) with 56 entries.

The lowest are Wyoming (WY), South Dakota (SD) and Maine (ME) each with only one, this would restrict the analysis and require cautious interpretation while studying the trends.

Uniqueness:
As a preventive and necessary measure before inserting data into the tables, there are some instances that uniqueness is verified:

- web_scrapper_table: Should first verify whether id_location already exists before inserting a new property.
- api_script.py: Checks if state is already present before fetching and adding new data.

print(df.duplicated().sum()) was 0
The results of these two lines are the same : *web_scrapper_api_property_data_df.nunique()* *len(web_scrapper_api_property_data_df)*

Accuracy:
Range Checks: Verification that the data related to price falls within expected ranges (e.g., no negative values, prices equal to 0 or below 10,000 USD)
11 rows were dropped with a DELETE query in the corresponding table.

web_scrapper_api_property_data_df[web_scrapper_api_property_data_df['price'] <= 10000]['price'].count()

Exploratory Analysis:

To study, in broad terms, the type of real estate that was extracted from the web, SQL queries were used:

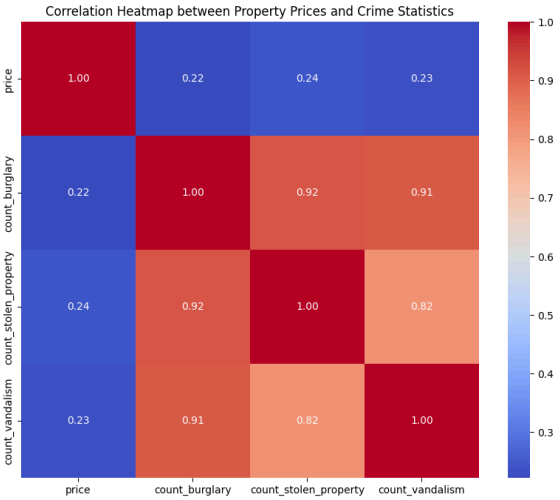
1. Top most expensive and cheapest properties from the real estate listing were retrieved.

	ID	Location	Price	State	Bedrooms	Bathrooms	Sq Ft
0	8	8828 Thrasher Ave, Los Angeles, CA 90069 TRULI...	6499000	CA	4	5	4159
1	583	400 New Salem Rd, Kingston, NY 12401 LISTING B...	5500000	NY	7	4	5540
2	807	179 Indian Mound Trl, Plantation Key, FL 33070...	4500000	FL	4	3	1960
3	531	433 Acacia Dr, Stockbridge, GA 30281 MLS ID #7...	50000	GA	3	2	1881
4	458	4705 Virginia Ave, Saint Louis, MO 63111 GARCI...	50000	MO	3	2	1820
5	257	607 Parkview Dr, Chattanooga, TN 37411 $61,750	61750	TN	3	2	1104

2. Average price of property depending on the number of bedroom (excepting 7 as it seems to be affected by outliers)

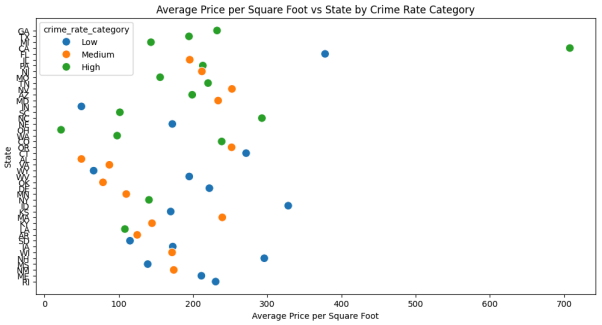
3. Distribution of property prices

Continuing with the research question, correlation was studied to determine if there is a statistical relationship between property prices and various types of crime extracted from the API.

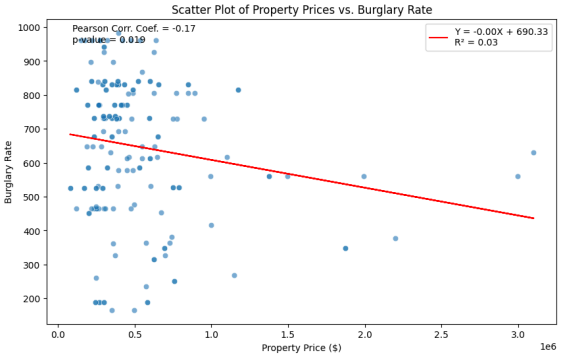


The values suggested a weak positive correlation (between 0.22 and 0.24), meaning that as the crime counts increase there is a slight tendency to get higher prices. This could be counterintuitive, but due to the lack of more granular data one may suggest that higher property values are frequently seen in cities, which also have higher absolute crime rates. The correlation between the types of crimes studied is high (between 0.82 and 0.952) indicating that high rates of one crime represent high rates of others too.

To answer the question "How does the average price per square foot vary among states with different crime levels?", the states were categorized into 'Low', 'Medium', and 'High' crime rates based on totals of crime count. The scatter plot below shows that states with greater crime rates do not always have lower average property values per square foot. There is not a clear, consistent pattern where states with lower crime rates have higher property values.

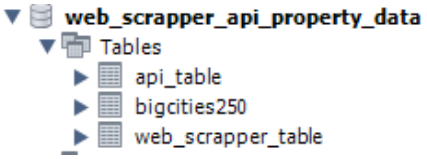


As mentioned before, a more granular dataset was used in order to add a bit more specificity. In this case, criminal information from a city-wise perspective was used to find if there was a correlation between property prices and the rate of violent crimes in the area. Please remember that this was done only for cities whose populations exceed 250 000 people.



Overall, while the graphs show a statistically significant association, the actual strength of the relationship is relatively low and it has minimal practical implications due to its low R^2 value. It implies that variables other than violent crimes and burglary rates have a substantially larger role in influencing the property price.

Finally the last used data frame (bigcities250) that allowed for the previous calculation was added to the database in MySQL.



Conclusions and Improvements

The exploratory analysis undertaken in this study shows that, while property values have a relation with crime rates, the correlation is minimal. Other factors that might impact real estate values include socioeconomic status, the quality of local schools, proximity to public facilities, and a region's overall economic health.

The fact that crime rates are presented at the state and later at city level whereas property values are provided at a more granular level, such as neighborhoods or specific addresses, is a crucial factor in interpreting the results of this study.

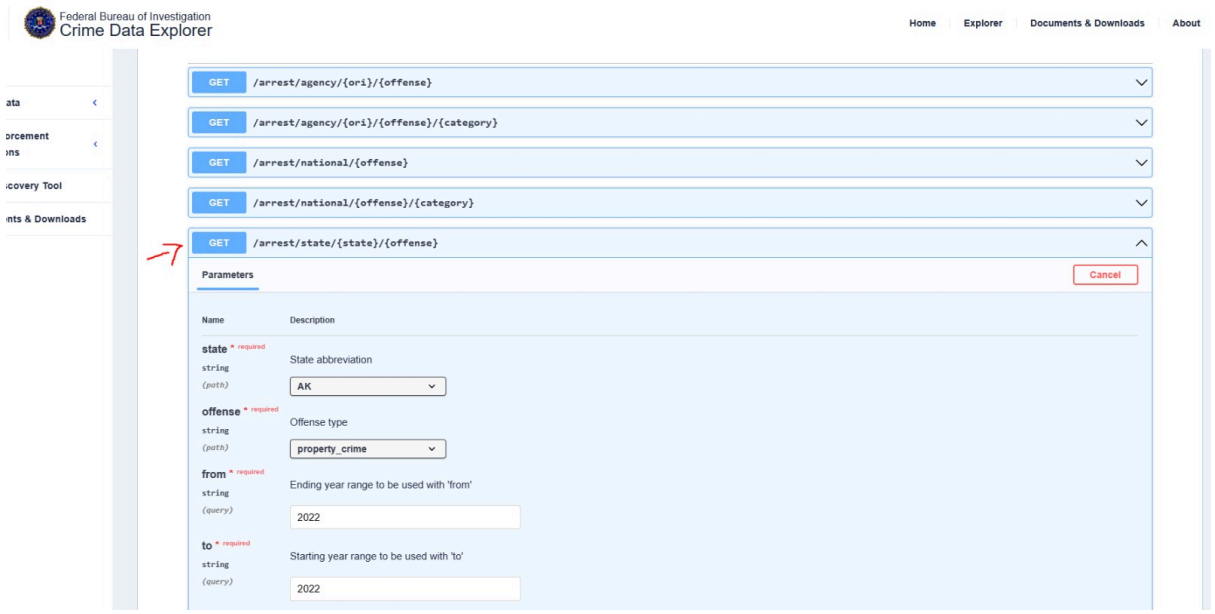
The current used crime statistics may not reflect adequately the crime situation in localized places. For example, a home in an unsafe area in a city with generally low crime rates may nevertheless be costly.

Further investigation, hopefully with more granular data or controlling for additional variables, would be required to determine the real nature of the association between property values and crime rates.

Technical Documentation (Separate Files)

Code: Provide all the code developed with comments.
Operational Guide: Include instructions for reproducibility.
Presentation: A summary PowerPoint presentation.

Appendix 1



Example for AK in 2022


```
{
 "title": "Arrests for Crime Against Property",
 "keys": [
  "Motor Vehicle Theft",
  "Arson",
  "Burglary",
  "Embezzlement",
  "Forgery and Counterfeiting",
  "Fraud",
  "Larceny - Theft",
  "Vandalism",
  "Stolen Property: Buying, Receiving, Possessing"
 ],
 "data": [
  {
   "data_year": 2022,
   "Motor Vehicle Theft": 323,
   "Arson": 42,
   "Burglary": 428,
   "Embezzlement": 30,
   "Forgery and Counterfeiting": 44,
   "Fraud": 104,
   "Larceny - Theft": 1063,
   "Vandalism": 988,
   "Stolen Property: Buying, Receiving, Possessing": 70
  }
 ]
}
```