Report Title*

(COMP3125 Individual Project)

David Arzumanyan
DATA SCIENCE FUNDAMENTALS - COMP3125

I. Introduction (*Heading 1*)

The topic explores analysis between housing and crime in Massachusetts state. Specifically, we explore questions whether we have any changes in housing prices due to crime rise or fall, whether we can classify cities safe or unsafe, and also trying to forecast how changes in crime rates can affect housing prices in future. These are all very interesting and informative problems to explore, especially when we are working with the actual live data of Massachusetts state, both for market and crime data. These answers can be very helpful in many niches and can come handy for an everyday consumer when approaching the topic of buying a new house.

II. Datasets

A. *Source of dataset (Heading 2)*

Two datasets were used in this project. First, the market live dataset that includes a ten-year nationwide market housing information, which was obtained from Kaggle.com. The dataset is being updated monthly and was last updated on November 11, 2025, created on December 1, 2024. It is considered a credible source since its actual source comes from Redfin.com data.

The second dataset is the crime data that was obtained from a very credible source, specifically, mass.gov. This dataset includes 2015-2024 crime rates across all cities in Massachusetts state. The dataset includes information such as the total number of crimes as well as all different types of crimes counted individually. While the government does not state a specific interval that this dataset is being updated, we can assume quarterly or annually since every dataset is per year.

B. *Character of the datasets*

The market dataset was in a 'tsv000' format which is a simple, plain-text file format used to store tabular data, such as a spreadsheet or database table. The size of the original data was 2.93GB, but I cut the dataset to have information only above year 2015 since we are only exploring the past 10-year information. The slim dataset was converted into 'csv' format for easy data access. The original slim dataset contains 51 different columns which can be explored via the link to the actual source of the dataset from the "References" section. Some of the key features are `

PERIOD_BEGIN	object
PERIOD_END	object
PERIOD_DURATION	int64
REGION	object
CITY	object
STATE	object
STATE_CODE	object
PROPERTY_TYPE	object
MEDIAN_LIST_PRICE	float64
Year	int64

As for data cleaning, the very first step was to filter the market data to only Massachusetts state listings. The following columns were dropped, since we didn't need those for our analysis.

"PERIOD_DURATION","REGION_TYPE","TABLE_ID","PARENT_METRO_REGION_METRO_CODE","PENDING_SALES","PENDING_SALES_MOM","PENDING_SALES_YOY","NEW_LISTINGS","NEW_LISTINGS_MOM","NEW_LISTINGS_YOY","MONTHS_OF_SUPPLY","MONTHS_OF_SUPPLY_MOM","MONTHS_OF_SUPPLY_YOY","MEDIAN_DOM_MOM","MEDIAN_DOM_YOY","AVG_SALE_TO_LIST_MOM","AVG_SALE_TO_LIST_YOY","SOLD_ABOVE_LIST_MOM","SOLD_ABOVE_LIST_YOY"

The "Year" column was renamed to "LISTING_YEAR" so there will be no confusion with Crime data's "Year" column later when we merge both datasets. The market dataset was also filtered to be in the range between 2015 and end of 2024 for the "LISTING_YEAR" column.

The crime datasets were all individually downloaded per year from the mass.gov website. These datasets came in "csv", all in total 9 datasets each per year from 2015 to 2024. Due to multi-dimensionality of these datasets I had to fix couple of things, such as` filtering the crime data to skip the first column to fix the multi-dimensional panel that was caused by the "Jurisdiction by Geography" column with all NaN values. I also had to rename the "Offense Type" column to "CITY" for the same above reason. Replaced empty strings, empty spaces, NaN, nan, None values to np.nan values. Dropped rows where ALL offenses were NaN values. Dropped the following unnecessary columns` "Unnamed: 59", "Missing". Filled NaN values with 0, since any np.nan values that we previously replaced represent 0 offenses for a specific category of offense. Replaced commas in offense rows, such as 4,321 -> 4321. Added "Year" column to crime datasets for concatenating. Afterwards, concatenated all crime data into single crime

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

data with different years. The above changes, data cleaning, and processing for crime datasets were applied to all individual crime datasets from 2015 to 2024.

Finally, market data and the single crime data were inner merged together based on left [“LISTING_YEAR”, “CITY”] and right [“Year”, “CITY”] keys.

METHODOLOGY

C. Method - Correlation

For the question “How does the crime rate in a city relate to its median house price?” correlation was applied to “All Offense Types” and “MEDIAN_LIST_PRICE” columns. This wasn’t applied to the actual dataset, but rather a new sub-data was selected with less columns.

For “Over the past 5–10 years, has the rise or fall in crime rates affected housing prices?” Data selection was performed on the dataset based on years of records` 2015-2020 and 2020-2024. Correlation was performed on both datasets for the same offense types and median list price columns. The correlation approach is appropriate in this case since it can show us the correlation between all offense types and median list price in yearly intervals such as 2015-2020 and 2020-2024. By this we can see if the rise/fall of the housing prices was any different between years. Crime trends were visualized for both yearly intervals.
Afterwards the data was grouped by city, listing year and all offense types columns combined for the median list price’ median value. This grouped dataset gives us a sight to explore whether the median list price of houses had significantly changed due rise/fall of offenses in a specific city.

D. Method - RandomForestClassifier

For “Can we classify whether a city is “Safe” or “Unsafe” based on housing and economic indicators?”
I approached this problem by first thinking which ML model would be a good fit, and obviously it should be a classification model since we are classifying Safe/Unsafe labels. I decided to go with RandomForestClassifier and the reason why I chose RandomForestClassifier is because it handles mixed numerical and one-hot encoded data extremely well compared to Logistic Regression. It automatically captures non-linear relationships and, in our case ` crime, housing and safety labels are nonlinear.
A safety label column was added in our dataset by the following rule`

Unsafe > Median
Safe < Median

Then I chose categorical and numeric features as follows`

["MEDIAN_LIST_PRICE","MEDIAN_PPSF","HOMES_SOLD", "INVENTORY","MEDIAN_DOM","AVG_SALE_TO_LIST","SOLD_ABOVE_LIST","PRICE_DROPS", "CITY", "REGION", "PROPERTY_TYPE"]

Categorical features were one-hot encoded, and X was chosen for features combining both numeric and categorical features. The target (Y) in this case was our earlier created “SAFETY_LABEL” column. Since our target is not numerical, I encoded it with LabelEncoder. The reason I went with LabelEncoder vs One-Hot Encoder is because LabelEncoded works best for binary columns, which is the case with Safe/Unsafe. Afterwards, data was split into 20% testing and 80% training chunks.

Finally, the model was fitted with no issues.

E. Method – LinearRegression with Time series lags

For “Can we forecast how changes in crime rates might affect future housing prices?”
My assumption was that a LinearRegression model itself would be a good fit to predict future housing prices, but after careful considerations it was understood that we have to predict values based on history of crime and housing prices. For that reason, I added time series lag columns to the original dataset with 1 and 3 shifts per column. Any null values in the new lag columns were filled with the mean of those columns. Train and test datasets were filtered as

Train < 2023-01-01
Test >= 2023-01-01

For features I chose the following columns`

'price_lag_1', 'price_lag_3', 'All Offense Types', 'crime_lag_1', 'crime_lag_3', 'Crimes Against Person', 'Crimes Against Property'

Our target (Y) was the train[“MEDIAN_LIST_PRICE”].
Finally, the LinearRegression model was fitted.

III. RESULTS

A. Result A

How does the crime rate in a city relate to its median house price?

We have the following correlation for the above question:

	All Offense Types	MEDIAN_LIST_PRICE
All Offense Types	1.000000	-0.056606
MEDIAN_LIST_PRICE	-0.056606	1.000000

The above correlation between All offense types and median house price tells us that there is almost no trend between them. House prices with a higher crime rate go down very slightly. There is not any significant trend, so we can surely state that the crime rate in a city does not direclty relate to the median price of a house.

B. Results B

Over the past 5–10 years, has the rise or fall in crime rates affected housing prices?

As a first step, I compared the correlations between 2015-2020 and 2020-2024 "All Offense Types" and "MEDIAN_LIST_PRICE" columns. The 2015-2020 range correlation showed a small negative number, which is most likely due to some outliers, so I wouldn't call that there is any significant trend here, nor do we have any trend in 2020-2024 range correlation, since the number was even smaller.

I also plotted 2 scatterplots to see if there is any pattern between crime and median house price for the above-mentioned yearly ranges, but nothing significant was found, except that the 2015-2020 range MEDIAN_LIST_PRICE does not go over $200,000 in cities with a crime rate above 4800. This was just an exploitative approach to this scatterplot. Again, there isn't any significant change on the plots for us to surely state that there is a trend between the crime rate and the median house price. All these plots can be found in Github's Pictures folder.

Lastly, I sub-selected an "annual_prices" dataframe, and that proves too, that there isn't any visually noticable trend. We can see random records with a high crime rate and a high median price. A very good example is the "Boston" city:

CITY	LISTING_YEAR	All Offense Types	MEDIAN_LIST_PRICE
Boston	2020	43189.0	739450.0
Boston	2021	40378.0	749000.0
Boston	2022	40073.0	799000.0
Boston	2023	42676.0	849450.0
Boston	2024	42566.0	861000.0

Notice that prices increase despite rising crime rates. Boston is downtown, and we barely see any fall in prices downtown no matter what. I would call these cities "desirable" and such desirable cities can really mix and puzzle our plots and make noise.

So, the short answer is no, crime rates go up and down, but prices increase steadily. So crime rates did not directly affect the median house price in the past 10 years.

C. Results C

Can we classify whether a city is "Safe" or "Unsafe" based on housing and economic indicators?

Luckily, I had many useful columns in the dataset that could potentially have an effect on this question. The results of my model prediction were based on "All Offense Types". The safe point was below the median, and the unsafe point was above the median.

The model successfully predicted the X train features with an accuracy of 0.957288.....
This is a good number of successes. To see the whole accuracy score and the confusion matrix, please check the code in GitHub.

The answer is yes; we can definitely classify a city safe or unsafe based on its features.

D. Result D

Can we forecast how changes in crime rates might affect future housing prices?

According to extracted coefficients of our features, we see that crimes against person and crimes against property have significant effects on the median list price. The model predicted with good accuracy according to the scatterplot I have in my code.

Since there is no direct relationship between crime rates and median house prices, it wouldn't be easy to state that such crime rates have effect on those prices, but we added time series lags and we made the data to be historical. With this, we can definitely forecast how changes in crime rates can affect future housing prices. The plots and predictions are available in the code.

IV. Discussion

I am satisfied with the results of the first, second and third questions' insights. The last problem might need some improvement, especially in building a proper relationship between crime rates, and housing prices and making it as historical events. This doesn't mean that my code doesn't work, but I believe I could improve it a little more to have more accurate insights. The reason behind this thinking is that I believe time series shifts alone are not enough to make data historical. Also, my insights were mainly based on the coefficients of my features, and I wouldn't call this enough solution to this question. For future, I may consider strengthening the relationship between crime and housing median price and bonding those with historical events.

V. Conclusion

In this project I was able to classify cities with safe and unsafe labels; it is not the hardest task to classify those, and I believe this can definitely help consumers make decisions prior to buying a house in Massachusetts. According to my results, Massachusetts is not a state where crime has big effects on the housing market; this was based on the past 5-10 year crime rate effects on housing. This can help people who are considering moving to Massachusetts or maybe investing into Massachusetts housing market, as it was obvious that houses here in Massachusetts do not go down in their values due any crimes; yes, some cities might be an exception and this can also be considered as a future work to explore those cities, but for the general part houses in Massachusetts hold their values steady and even go up each year.

Acknowledgment *(Heading 5)*

I would like to thank Dr. Weijie Pang for the support and knowledge provided for this project.

References

[1] V. Vaseghi, *US cities housing market data – live dataset*, Kaggle. Available: https://www.kaggle.com/datasets/vincentvaseghi/us-cities-housing-market-data. Updated 22 days ago. Sources: Redfin Public Data (https://redfin-public-data.s3.us-west-2.amazonaws.com/redfin_market_tracker/city_market_tracker.tsv000.gz). Direct download methodology: https://www.redfin.com/news/data-center.

[2] Massachusetts Government, *Crime statistics*. Available: https://www.mass.gov/crime-statistics.