Report Title*

(COMP3125 Individual Project)

*Note: Do not used sub-title

David Arzumanyan
dept. name of organization

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.

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

Afterwards, concatenated all crime data into single crime 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

In this section, present your findings using an appropriate method, such as equations, numerical summaries, or visualizations like charts and graphs. Clearly explain all results and provide guidance on how to interpret them. If any unexpected results arise, discuss possible reasons or contributing factors. To improve clarity and organization, consider using subsections (e.g., A, B) to separate different aspects of your results.

Example: After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Result A

Example: XXX

1) For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) For papers with less than six authors: To change the default, adjust the template as follows.

a) Selection: Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

c) Deletion: Delete the author and affiliation lines for the extra authors.

B. Results B

Example: Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

C. Results C

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I. TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

IV. Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

Example: xxx

V. Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what's the effect of this result on the real-world?

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

Example: xxx

Acknowledgment *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Use the IEEE format for the citation. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..." Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.