Real Estate Project

By Ibraheem Qureshi, Jordan Sutherland, Robel Mamo, and Karthik Turimella

Outline

- Aim
- User stories
- Software Architecture
- Getting set up
- Collecting Data
- Running the program
- Important commands
- Future Work and Questions

Aim of the Project

- Conduct an Exploratory Data Analysis on real estate data
- Develop a machine learning model for predicting real estate prices
- Identify potentially profitable investment opportunities in the housing market
- Utilize historical and current data of the real estate market

User Stories

Initial Requirements:

- Use zillow api or historic data on real estate to store house information in database
- Clean data of missing and null values
- Analyze database to give top recommendations for investment

1st Sprint:

- Allow for ability to produce graphs that can be saved and viewed later of real estate information
- Give factors for top recommendations
- Build model to give those recommendations based on information in database

2nd Sprint:

- More variance on graphs
- Give more information on why the top recommendations are made
- Provide in depth information on selected property

Software Architecture

Our goal:

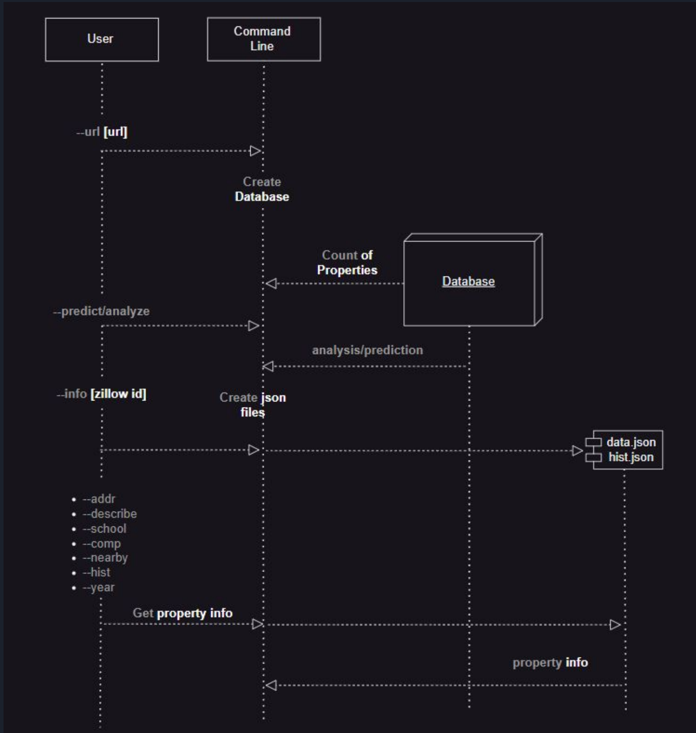
- Create an architecture easy for you, our customer, to use with the minimal amount of request for the analysis and research you are conducting.

Main features:

- Perform analysis on the requested data.
- Make predictions for best investment options.
- Provide detailed information on a selected property.

The architecture:

- A database - created upon first request and used for analysis/prediction
- Two json files - created when requesting information on a selected property and used to get various property details.



Getting Started and Clone GitHub Repository

- Install the latest version of python
- In a terminal window, run the following command lines to install the necessary dependencies for the program: sqlite3, pandas, plotly, seaborn, matplotlib, folium, requests, scikit-learn
- Clone the repository onto your local machine.
- Choose a directory that you plan to have the program in, open a terminal window on that directory, and run the following command line to clone: git clone https://github.com/dussec/real-estate-price-analysis.git

Data Collection

- Go to the Zillow website (website) and search for homes (for sale) in the Denver.
- Try to include a variety of areas (Aurora, Centennial, etc) and a broad criterion for a large dataset.
- Once you're done with the search, copy the link in the address bar.
- If you wish, you can use this link already provided in the `zillow_example.txt`file under the 'Examples' folder.

Running the Program

- Open a new terminal window on the folder `Main`.
- Now you can use various command line arguments to execute different commands…

Command: --url

Command: python main.py --url [zillow search url]

Purpose: It will create a SQLite database of listings from the search url you obtained previously. The url is a required argument for the command.

Expected output: 'zillow_listings.db' should be created in the same directory (Main folder). And, a list of keys and the number of properties retrieved should be be shown. For example:

-
  ```
  dict_keys(['user', 'mapState', 'regionState', 'searchPageSeoObject', 'requestId', 'cat1', 'categoryTotals'])
  Count of properties: 1969
  ```

Note: In order to run the other two commands, it is important to have zillow_listings.db in the Main folder. Otherwise, the following 2 commands will not work.

Command: --analyze

Command: python main.py --analyze

Purpose: It will conduct Exploratory Data Analysis (EDA) on the properties off the database.

Expected output:

- A summary table to the terminal.
- A variety of graphs (should be 6) that will appear in a separate popup window (the next graph will not be shown unless you exit out of the window of the current graph).
- A heat geographic heat map (should appear as heatmap.html) of current prices that will be saved to the current directory (inside Main). You can open it in your browser to interact with it.

Command --predict

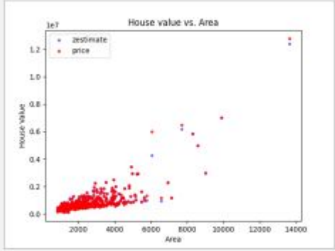
Command: `python main.py --predict`

Purpose: It will use a linear regression machine learning model create predicted prices of the properties off the database and compare it with the actual price to determine which properties are undervalued (ideal for investment).

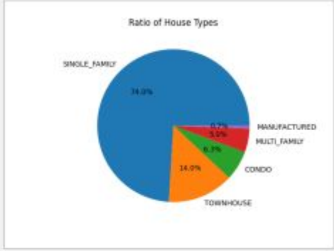
Expected output:

- a list of 5 properties with the best value. It will have their Zillow ID, actual price, predicted price, and the difference between the two. It will be shown in the terminal.
- a variety of graphs (should be 5) that will appear in a separate popup window (the next graph will not be shown unless you exit out of the window of the current graph).
- a heat geographic heat map (should appear as price_difference_heatmap.html) that will be saved to the current directory (inside Main). You can open it in your browser to interact with it.

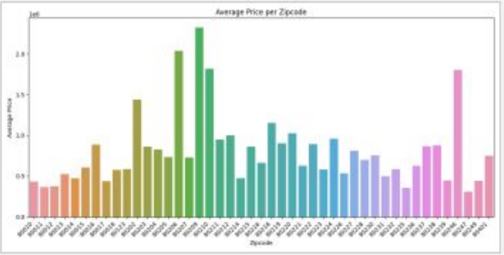
Interpreting the output from --analyze



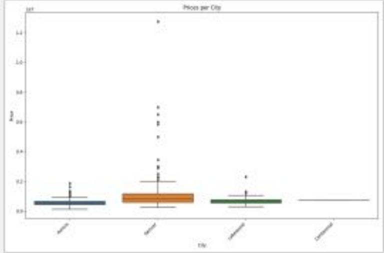
A scatter plot that compares the area of houses with their corresponding zestimate and price values, providing insights into the relationship between house value and size.



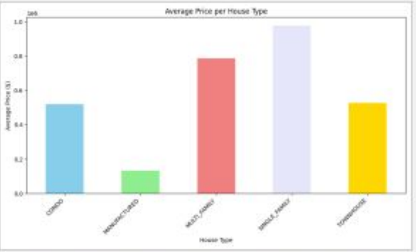
A pie chart that illustrates the distribution and proportions of different house types in the dataset, enabling an understanding of the relative prevalence of each type.



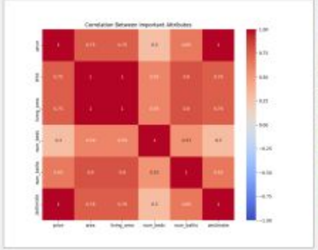
A bar chart displaying the average price per zipcode, helping identify variations in housing prices across different areas.



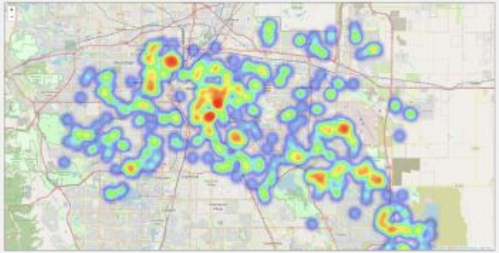
A boxplot that visualizes the distribution of prices for each city, allowing for comparisons of price ranges and identification of potential outliers.



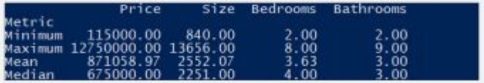
A bar chart that displays the average price for each house type, enabling comparisons of price levels among different types.



A correlation matrix heatmap that shows the correlations between important attributes such as price, area, living area, number of bedrooms, number of bathrooms, and zestimate, allowing for an analysis of the relationships between variables.



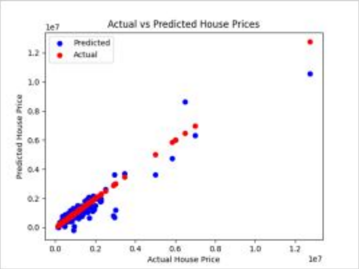
A heatmap that visualizes the spatial distribution of house prices, providing insights into areas with higher or lower property values.



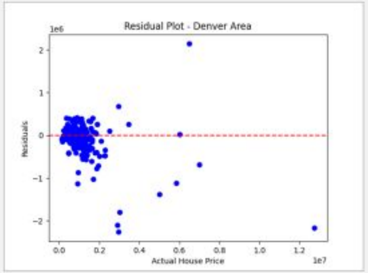
```
                 Price      Size  Bedrooms  Bathrooms
Metric
Minimum    115000.00    840.00      2.00       2.00
Maximum 12750000.00 13656.00      8.00       9.00
Mean       871058.97   2552.07      3.63       3.00
Median     675000.00   2251.00      4.00       3.00
```

A summary table that provides key metrics such as the minimum, maximum, mean, and median values for price, size, bedrooms, and bathrooms, allowing for a quick overview of the dataset.

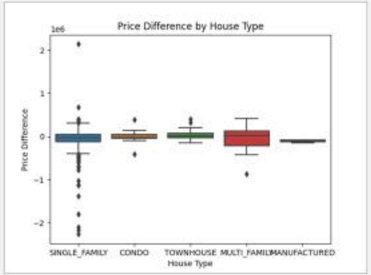
Interpreting the output from --predict



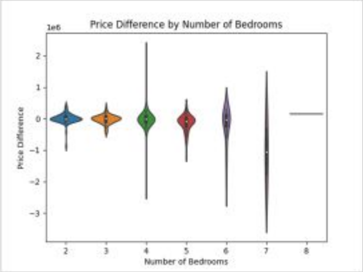
A scatter plot that compares the actual house prices with the predicted house prices, providing a visual representation of the model's performance.



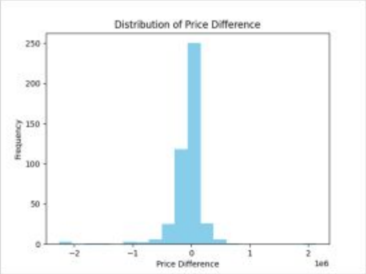
A scatter plot that shows the residuals (the differences between predicted and actual prices) against the actual house prices, helping to identify any patterns or trends in the model's errors.



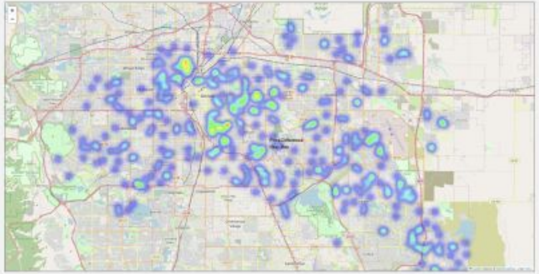
A box plot that displays the distribution of price differences for different house types, allowing for comparisons and analysis of how house types affect the prediction errors.



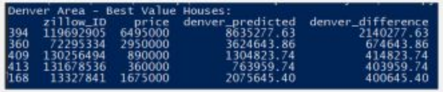
A violin plot that illustrates the distribution of price differences for different numbers of bedrooms, providing insights into the relationship between the number of bedrooms and prediction errors.



A histogram that visualizes the distribution of price differences, allowing for an analysis of the frequency and spread of the prediction errors.

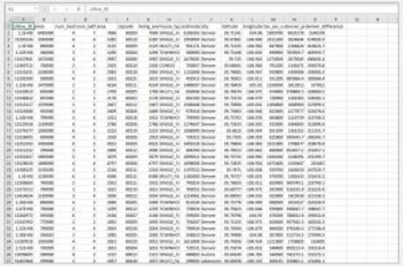


A heatmap that visualizes the spatial distribution of price differences, highlighting areas where the model's predictions deviate from the actual prices.



```
Denver Area - Best Value Houses:
      zillow_ID    price  denver_predicted  denver_difference
394  119692905  6495000        8635277.63         2140277.63
360   72295334  2950000        3624643.86          674643.86
409  130256494   890000        1304823.74          414823.74
413  131678536   360000         763959.74          403959.74
168   13327841  1675000        2075645.40          400645.40
```

A table that displays the 5 “best” investments based on the difference between actual and predicted.



A spreadsheet of all the properties along with their predicted prices, differences, and other important attributes.

Command --info

And sub-commands (--addr, --describe, --school, --comp, --nearby, --hist, --year)

Command: python main.py --info **[zillow id]**

Purpose: It will request house data for the provided zillow id and save two formatted json files. One of the files, hist.json, contains the historical prices of the property. The other file, data.json, contains the data for all the other information about the property.

Sub-commands: python main.py **[sub-command]**

- data.json
  - **--addr**: outputs street address and zip code of the house.
  - **--describe**: outputs a block of text description for the house.
  - **--school**: outputs primary-, middle- and high- school, distance from the house, name of school, and its rating.
  - **--comp**: outputs a link to all the comparable houses.
  - **--nearby**: outputs a list of nearby cities.
  - **--year**: outputs the start to completion dates of the house.
- Hist.json
  - **--hist**: outputs the historical prices, date of transaction, price change rate in percentage, price per square footage at the time of the transaction, and event type (sold, listed, etc.)

Prediction Model

How?

- The prediction model was trained on Denver data and Los Angeles data separately. This resulted in having two different models: Denver Model & Los Angeles Model.

Usage

- The model is given house attributes and it returns a prediction. The price differences between the predicted and actual price are calculated. Houses with positive price difference are ranked higher in investment suggestion list.

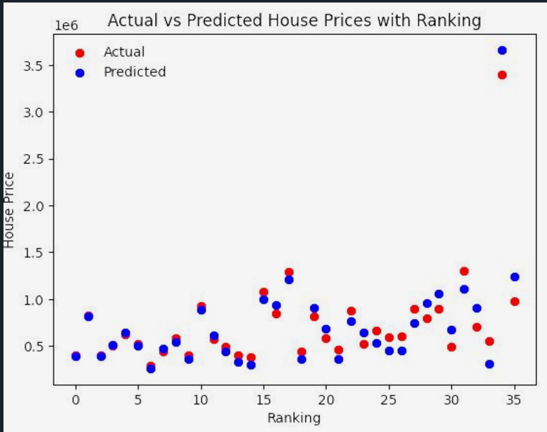
Denver Model:

- Correlation: 0.95

Los Angeles Model:

- Correlation: 0.7

Influential attributes: zip code, lat/long, num beds/baths, tax assessed value



Future Work and Questions

- GUI
- Access to zillow developers api to get better model
- Gather more data to train the model better (e.g., crime, etc)
- Look at more factors to take into account for suggestion
- More graphs to explore and describe additional data gathered
- Get more financial data (i.e., loans, etc)
- Expand database

Let’s demo!

Conclusions

1. This project focuses on real estate analysis and prediction, aiming to provide users with valuable insights for investment decisions and exploratory data analysis (EDA) in the real estate market. It consists of three main goals:

- Real-time Data Retrieval: The project utilizes the Zillow API to pull real-time data of property listings, ensuring up-to-date information for analysis and prediction.
- Exploratory Data Analysis (EDA): Various graphs and visualizations are generated to perform EDA on the acquired data. These graphs help identify trends, patterns, and potential investment opportunities in the real estate market.
- Predictive Modeling: The project employs linear regression models to make predictions for the listed properties. By comparing predicted prices with actual prices, the program identifies properties with the potential for investment. The predictions are further analyzed and visualized through different graphs.

2. This project serves as a stepping stone for users interested in real estate analysis, providing them with direction for further research and analysis. It offers flexibility, allowing users to use their preferred development environment and easily integrate the program into their workflow.