DSCI 510: Principles of Programming for Data Science

Name: Rafayel Mirijanyan

USC ID: 3487192016

GitHub: https://github.com/rafayelmirijanyan1997

Introduction

The **Section 8 Housing Program** is a federal assistance program designed to help low-income individuals and families afford safe and decent housing in the private rental market. This python program "Section 8 investment analyzer" is for real estate investors, it processes listings in 5 investor friendly cities to come up with best deals. It identifies profitable investment options with its justifications and ensures passive income for investors within boundaries of section 8 housing program. The program overall states whether it is worth investing in real estate in Detroit, Indianapolis, Memphis, Cleveland and Birmingham to have passive income from Section 8 or not.

Data Collection

Data for "Section 8 Investment Analyzer" comes from Zillow.com and huduser.gov (Department of Housing and Urban Development's). Zillow.com provides listings, while huuser.gov provides info about how much US government pays in private rental market based on zip code and number of bedrooms. I started scraping Zillow.com with beautifulsoup and requests library to get listings. Function extract_details helps to identify number of beds, baths and square feet for each listing. Then to be on the safe side and avoid paid APIs or rotating IPs, I exported the rest of the data from Zillow extension to csv files (https://zillow.scraper.plus/), as Zillow might block web-scraping in case of bigger amount of data. I have concentrated on 5 best cities at this point, which are investor friendly. Then I got best of the best investment options within those 5 states. Two different data sources were used Zillow.com and huduser.gov ending up with having ~ 3500 samples.

Data Cleaning

Raw data had a lot of null values, unnecessary columns and lot of outliers.

detect_outliers_iqr function uses the following logic to identify outliers.

The IQR is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1).

Lower Bound = Q1 - 1.5 * IQR

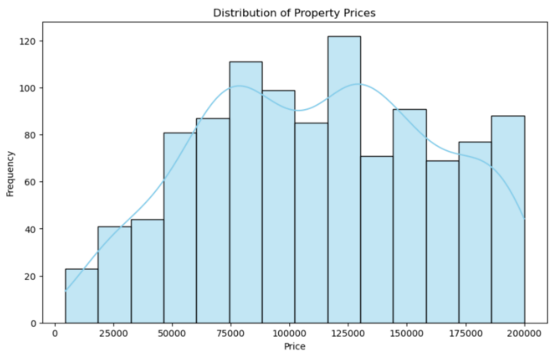
Upper Bound = Q3 + 1.5 * IQR

I concentrated on houses for sale only with specific filters on area, number of beds, price etc. Some of the values were in different units (acres instead of square feet). Initial data was in html format that needed to be scraped to get a data-frame and export it to csv file. After cleaning and filtering for listings, there were less columns, no null values.

Once the data was collected and cleared from both sources (Zillow.com and huduser.gov), the function get_fmr pulled correct fair market value from second data source based on zip code and number of beds.

Visualizations

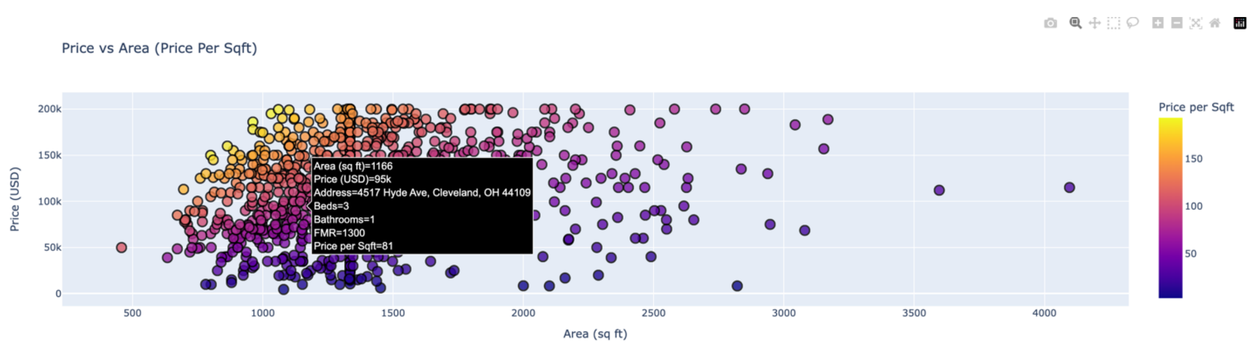
After identifying outliers, boxplot visualization helped to understand the extent of the problem and confirm if they are true anomalies or part of the valid data distribution. As my main focus was on price range up to 200 K, hist-plot of distribution of property prices helped to understand how the property prices are spread



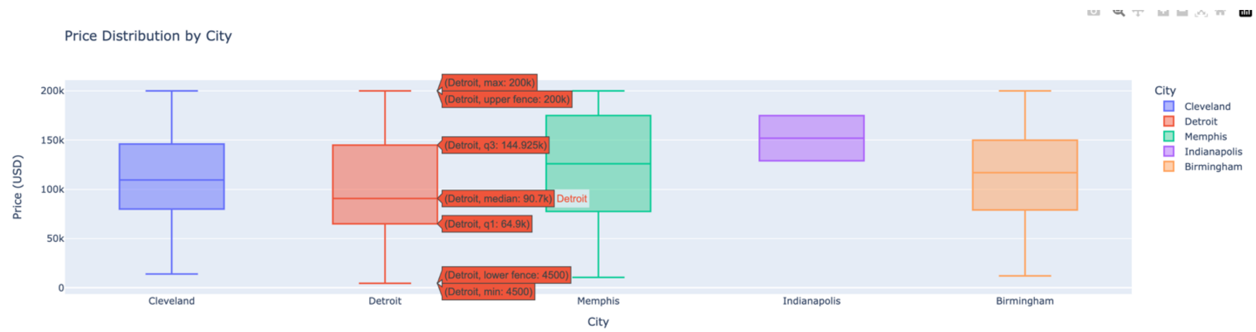
The **Price vs. Area scatter plot** visualizes the relationship between the **price** of a property and its **area** (typically measured in square feet or square meters). This is a key plot in real estate analysis, as it can reveal several important patterns and insights about property prices and sizes.

Violin plot helped to compare price distribution with number of bedrooms. I concluded that the price distribution within our data is almost the same for 3- and 4-bedroom houses.

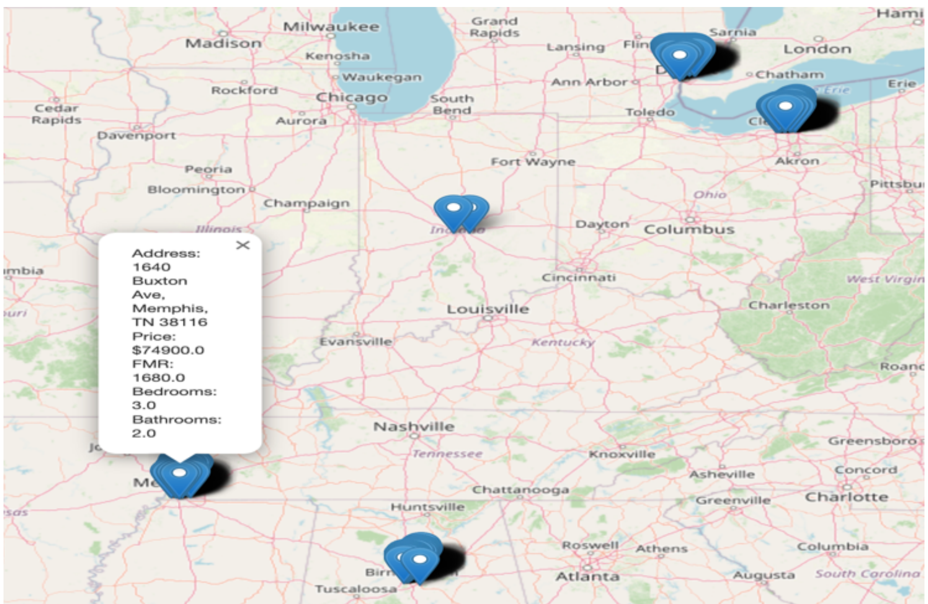
Typically, you would expect that larger properties have higher prices per square feet. Interactive scatter plot helped to confirm if that is the case in my dataset.



To be able to make price range comparisons and skewness of price data within 5 cities that were selected, I have used interactive box plot



To have more visual context about locations of those listings, I have used latitude and longitude within our data to create a map with markers for each property that will show details about each listing once you hover over into it. The map will be saved as html after program is completed.



Now that we have clear picture of what our data represents, lets integrate investment property cashflow calculator, to be able to calculate monthly cashflow for each house.

Cash flow = Rental Income - (Mortgage Payment + Operating Expenses).

The Fair Market Rent will be the Rental income for investors. To calculate the mortgage payment, we will use the following formula.

$$M = P\,\frac{r(1+r)^n}{(1+r)^n - 1}$$

Where:

M = Monthly mortgage payment

P = Loan principal (amount borrowed) (assume 90 % of total price)

r = Monthly interest rate (annual rate divided by 12) (6.7/12 = 0.56 % monthly for 30 year fixed rate)

n = Number of payments (loan term in years × 12 months) (12 x 30 = 360)

Now that we have a mortgage payment, we also need operating expenses to calculate monthly cash flow. To calculate monthly operating expenses, we will use the following market average numbers:

1 Homeowners Insurance: 0.5 % of house price annual -> 0.042 % monthly of house price

2 Property Management Fees: 10 % of monthly rent

3 Maintenance/Repairs: Estimate for monthly upkeep. A common rule of thumb is to budget $1 per square foot per year, which can be divided by 12 for monthly expenses.

4 Utilities: for moderate climates (Memphis, Indianapolis, Birmingham) we would use approximate 325 dollar for average household, for non-moderate climates (Detroit and Cleveland) we will use 450 dollar for monthly utilities.

5 HOA Fees: for single family homes without extensive amenities we will calculate 75$ monthly

After calculating all the required metrics, I have merged that data into the main dataframe.

FMR	monthly_mortgage_payment	Maintenance	HOA_fees	Monthly_Insurance	Monthly_property_mgmt	Monthly_Utilities
1410.0	1158.069888	147.833333	75	82.916667	141.0	450
1540.0	843.819768	91.000000	75	60.416667	154.0	450
1410.0	436.458501	136.083333	75	31.250000	141.0	450
1300.0	378.264034	133.083333	75	27.083333	130.0	450
1300.0	343.347354	95.666667	75	24.583333	130.0	450

After having all the required info I have calculated monthly cashflow after all the expenses which as stated above is:

Rental Income - (Mortgage payment + Operating Expenses)

We are only interested in positive cashflow houses with min of $500 per month.

```
df_combined[df_combined['Monthly_cashflow'] > 500].value_counts().sum()
```

```
82
```

As we only have 82 houses out of 1089 that is worth paying attention, the last step we need to do is to get rid of demolished, burned properties that need new construction. To do that let's get rid of all the houses that are cheaper than 50,000.

Conclusion

Here is the list of all properties that could be considered as great investment opportunities with very little investment of ~ 10K and > 500$ monthly cashflow based on guaranteed rents paid by US government, which is approximately 60 % of Return on Investment.

This project shows that although Indianapolis is an investor friendly state, it is not very beneficial for section 8 investment portfolios. Out of 3500 listings, only 24 of them will ensure guaranteed passive income with minimum of $500 monthly cashflow. Depending on the lease agreement, utilities might be paid by tenant, which means that the list of houses with >$500 will increase.

Once the program is run, please find the results folder to check out all the plots and 3 html files for interactive visualizations.

The final list that is extracted is called passive_income_properties.csv which is also saved in the same results folder.

Future Work

Given more time, I would run this project for all other states, would also integrate crime rates into it, and renovation expenses to pass inspection if any. I would still use Zillow.com as it is the largest database for real estate listings, while huduser.gov website will provide detailed info about how much US Government pays per zip code and per number of bedrooms for Section 8 Investors.