Northeastern University

Research Project Final Report

Boston House Sell Price Analysis & Prediction

Weijia Xiao, Yining Tao, Yihan Luo, Lang Shao

Professor: Eric Gerber

Class: DS 3000 SEC 04 Foundation of Data science

Date: 28 November 2023

Abstract:

Our project uses Redfin's data to interpret the Boston housing market, focusing on how property details and timing affect sales outcomes. Our findings reveal geographical and temporal price variations, offering strategic insights for market participants. We employed and compared multiple regression and polynomial multiple regression analyses, and chose the simpler multiple regression as our final model. Although our model captures notable trends, its predictive precision is limited, indicating the potential influence of external factors.

This research serves as a foundational step for stakeholders aiming to optimize their engagement with the Boston real estate market.

Research Question:

- What is the average sold price of houses in different ZIP Codes of Boston ?
- Is there a particular trend in recent months for the house sale price?
- How do ZIP Codes and the number of bedrooms and bathrooms affect the sold price?

Introduction:

Redfin, one of the largest real estate marketplace in the U.S. and Canada, boasts a database of millions of properties, offering not only sales records but also estimates of current market prices based on nearby sales data. This is incredibly valuable for sellers and buyers looking to gauge property values. However, individuals aiming to sell at the most opportune time face a challenge. While Redfin's estimates provide current market values, they don't predict optimal future selling times. Thus, from a Boston seller's standpoint, our project will focus on analyzing Redfin's Boston housing sales data to maximize a seller's profit. This approach, conversely, can aid buyers in minimizing their expenditure when purchasing a house.

Data Description:

For this project, we web scraped the Redfin website to collect data on the most recently sold houses in each ZIP Code of the Boston area that have records on Redfin. The dataset comprises the sold date, sold price, number of bedrooms and bathrooms, square footage, street address, city, state, and the ZIP Code of each house.

```
boston_houses.head(10)
```

	Sold Date	Sold Price	Bedroom Count	Bathroom Count	Area (sq ft)	Street Address	City	State	Zipcode
0	2023-11-15	1475000.0	2.0	1.0	1010.0	40-42 Mt Vernon St Unit 4A	Boston	MA	02108
1	2023-11-08	3600000.0	3.0	2.5	2234.0	28 Mount Vernon St #1	Boston	MA	02108
2	2023-11-03	569000.0	1.0	1.0	362.0	80 Mount Vernon St #12	Boston	MA	02108
3	2023-10-31	3900000.0	3.0	2.5	2303.0	73 Mount Vernon St	Boston	MA	02108
4	2023-10-30	500000.0	0.0	0.0	0.0	70 Brimmer	Boston	MA	02108
5	2023-10-27	2234000.0	2.0	2.0	1389.0	98 Chestnut St #4	Boston	MA	02108
6	2023-10-27	2800000.0	2.0	3.0	2172.0	29 Brimmer St #1	Boston	MA	02108
7	2023-10-27	17250000.0	4.0	5.0	5300.0	73 Beacon St	Boston	MA	02108
8	2023-10-13	3200000.0	2.0	2.0	1832.0	21 Branch	Boston	MA	02108
9	2023-10-06	3550000.0	3.0	3.5	2395.0	4 Acorn St	Boston	MA	02108

Usage of Data:

After the data scraped from Redfin is cleaned and visualized, we can use the data to answer the questions listed above:

• The first question is answered with a bar graph that shows the average housing price per square foot in each area of Boston separated by ZIP Code.

• The second question is answered with a line plot that directly shows us if there is a particular pattern for housing prices in recent months.

• For the third question, we built multiple regression to discover the relationship between the number of bedrooms and bathrooms and the sold price per square foot.

Pipeline Overview:

For scraping the Redfin website:

- isNumber(str)
  - The isNumber function is designed to ascertain if a string represents a numerical value, accounting for floating-point numbers by permitting a single decimal point. It returns a boolean value, confirming True for a numeric string, otherwise False. This utility is pivotal for data validation in numerical datasets.
- get_recent_sold(zipcode)
  - builds a dataframe of relevant information for all recently sold houses on Redfin in the given ZIP Code.

get_recent_sold(zipcode) function will be utilized with a list of all ZIP Codes in Boston and use a for-loop to create a dataframe for all of the recently sold house information for all areas in Boston. The code is designed to

handle multiple pages of listings by first determining the total number of pages available for the given ZIP code. It then iterates through each page, pulling data for individual properties, such as the date sold, sale price, bedroom and bathroom count, square footage, and address details. This extracted information is then processed, with numerical values being sanitized and converted to the appropriate data type. Textual data, such as dates and addresses, are also parsed and formatted. These processed pieces of data are accumulated into a dictionary, which subsequently gets converted into a pandas DataFrame for ease of manipulation and analysis.

Data Cleaning:

We looked into each individual houses in the dataset we scraped that have an area of 300 square feet or less, which is a fairly small area for a house, and found that houses that marked to have an area with 200 square feet or less are all either have their area marked incorrectly or is actually a parking space that shouldn't count as a house, so we decided to clean the dataset by eliminating those incorrect data. We also found that the house with 99 bedrooms is an entire apartment building rather than a residential house, so we decided to eliminate that as well. A few of the ZIP Codes have only a part of them included in the city of Boston, so we also make sure we only include data for houses that label their city as Boston.

```
# cleaned dataset after eliminating abnormal data that either have incorrect area or isn't a house
boston_houses_cleaned = boston_houses.loc[(boston_houses['Area (sq ft)'] > 200)
                        & (boston_houses['Bedroom Count'] <= 30)
                        & (boston_houses['Bathroom Count'] <= 20)
                        & (boston_houses['City'] == 'Boston')]
```

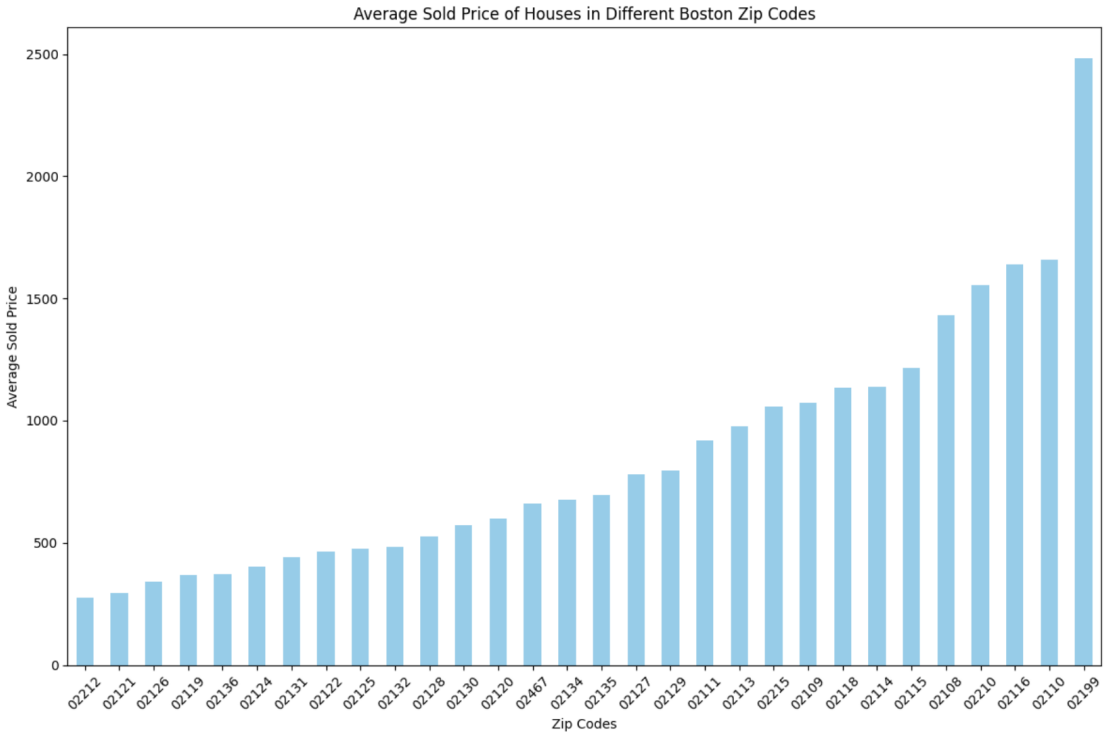
```
boston_houses_cleaned
```

	Sold Date	Sold Price	Bedroom Count	Bathroom Count	Area (sq ft)	Street Address	City	State	Zipcode
0	2023-11-15	1475000.0	2.0	1.0	1010.0	40-42 Mt Vernon St Unit 4A	Boston	MA	02108
1	2023-11-08	3600000.0	3.0	2.5	2234.0	28 Mount Vernon St #1	Boston	MA	02108
2	2023-11-03	569000.0	1.0	1.0	362.0	80 Mount Vernon St #12	Boston	MA	02108
3	2023-10-31	3900000.0	3.0	2.5	2303.0	73 Mount Vernon St	Boston	MA	02108
5	2023-10-27	2234000.0	2.0	2.0	1389.0	98 Chestnut St #4	Boston	MA	02108
...	...	...	...	...	...	...	...	...	...
10446	2023-10-31	2295000.0	3.0	2.5	2042.0	2400 Beacon #401	Boston	MA	02467
10455	2023-10-20	965000.0	2.0	1.5	1206.0	2400 Beacon St #211	Boston	MA	02467
10456	2023-10-20	1372500.0	2.0	2.5	2717.0	228 Allandale Rd Unit 2B	Boston	MA	02467
10459	2023-10-16	361000.0	1.0	1.0	762.0	57 Broadlawn Park Unit 17A	Boston	MA	02467
10476	2023-09-20	323000.0	2.0	1.0	786.0	52 Bryon Rd #4	Boston	MA	02467

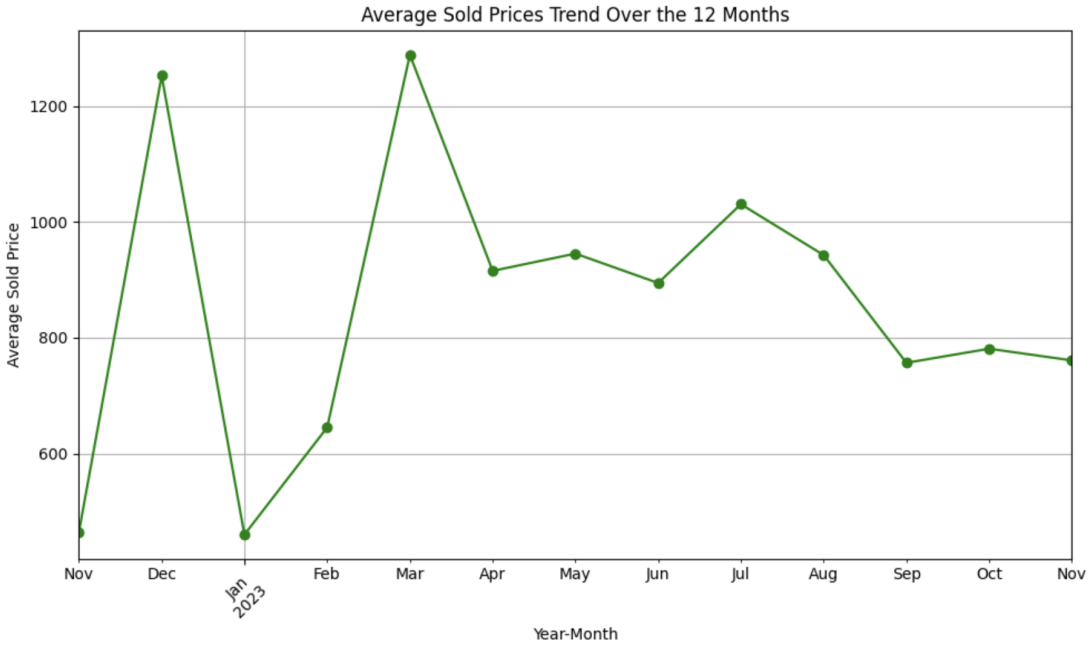
9682 rows × 9 columns

Visualizations:

The following bar plot visualizes the average housing prices per square foot in each area of Boston (separated by ZIP Code provided by cityofboston.gov).



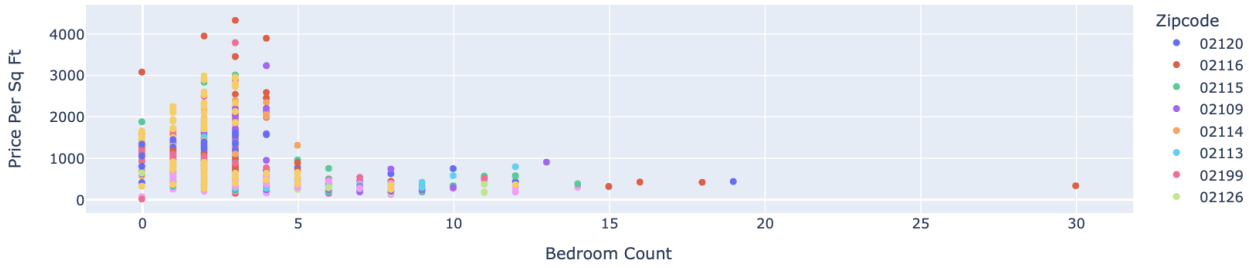
The following line chart visualizes the trend of average sold price per square foot over the 12 months (Nov 2022 to Nov 2023).



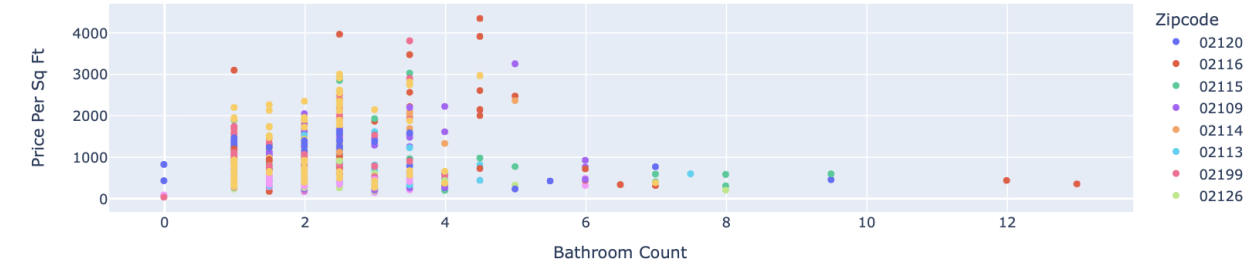
The following scatter plot visualizes housing prices per square every month for houses in each area of Boston (separated by ZIP Code provided by cityofboston.gov).

The following 2 scatter plots visualize the relationship between number of bedrooms and bathrooms with the average housing price per square foot (separated by ZIP Code provided by cityofboston.gov)

Relationship Between Number of Bedroom and Price per Sq. Ft.



Relationship Between Number of Bathroom and Price per Sq. Ft.



Machine Learning Method & Results

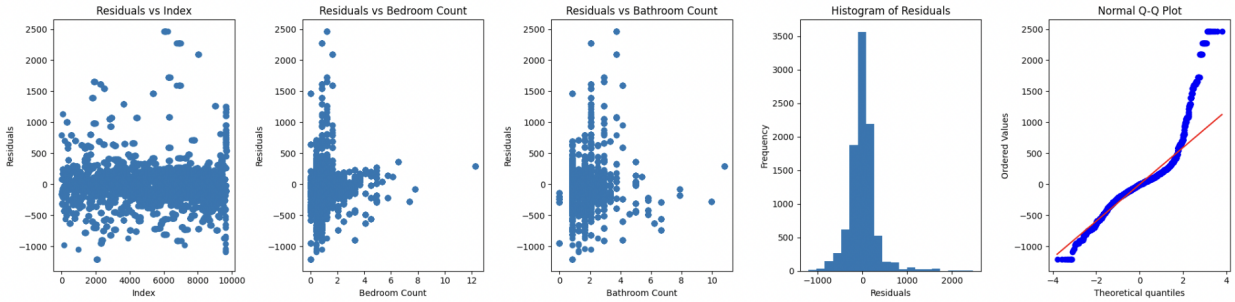
We will use multiple regression to answer the question: How do the ZIP Code and the number of bedrooms and bathrooms affect the sold price per square foot?

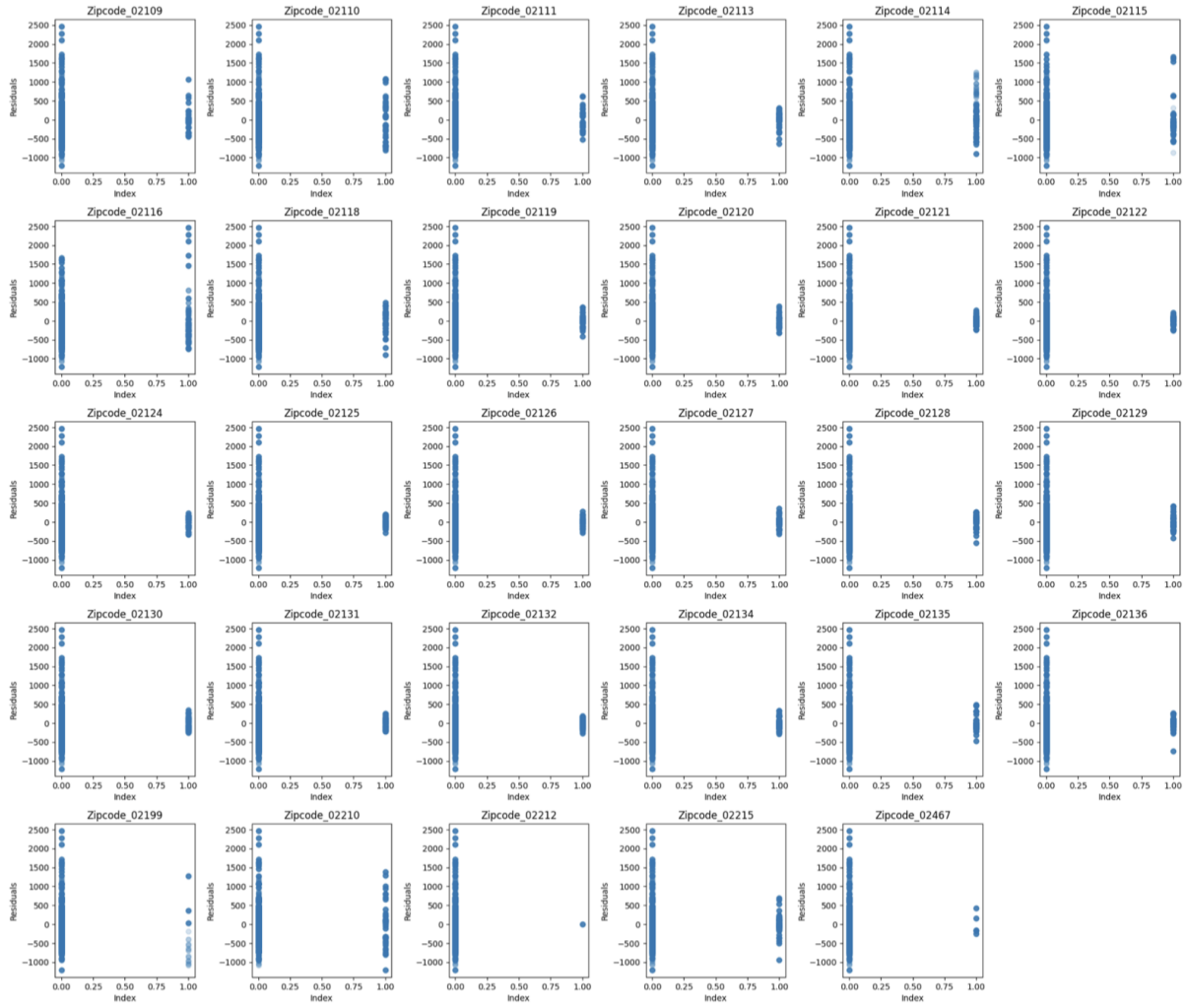
```
X = np.column_stack((bedroom_count_scaled,
                     bathroom_count_scaled,
                     zipcode_dummy))
y = price_sqft.to_numpy()
```

Single Cross Validated MSE and R^2:
MSE = 109764.59927220162
R^2 = 0.6467155645274544

MSE and R^2 of final trained model:
MSE = 103449.85868839844
R^2 = 0.6572528743442002

Slopes for the final trained model:
[1288.80155151, -158.07225631, 151.94578556, -318.41349924, 230.95378989, -485.91647633, -364.54303818, -246.02559726, -164.87180142, 217.99413479, -267.59055158, -925.74626135, -721.91771029, -962.9666055 , -844.38588545, -880.76530122, -831.63732255, -962.19174535, -595.0408778 , -822.45305118, -578.67139213, -760.24923172, -876.05545703, -838.83047295, -647.3713361 , -631.8308087 , -935.00789761, 1004.6686511, 140.69568088, -807.882315 , -304.55126416, -711.2048951]





The model has a R^2 value of 0.657, suggesting the model explains approximately 65.7% of the variance in the house sold prices. The MSE is 103449.85, indicating there is an average error of 321 dollars between the predicted and actual unit price per square foot, which is a little high. The Residuals vs. Index plot shows a random scatter, which is a positive indicator of no apparent autocorrelation. However, the Residuals vs. Bedroom Count and Residuals vs. Bathroom Count plots show a much higher variation in the middle than in tails, suggesting that we might want to include in other features that can better explain the variation with houses with a medium level of bedroom and bathroom count. The histogram of residuals indicates a skewed distribution, leaning towards the left, suggesting the presence of outliers or a non-normal distribution of errors. The Normal Q-Q Plot confirms this, with residuals deviating from the line in the tails. These patterns suggest the model may benefit from incorporating more complex variables or employing transformation techniques to improve its accuracy and handle the non-normality of residuals.

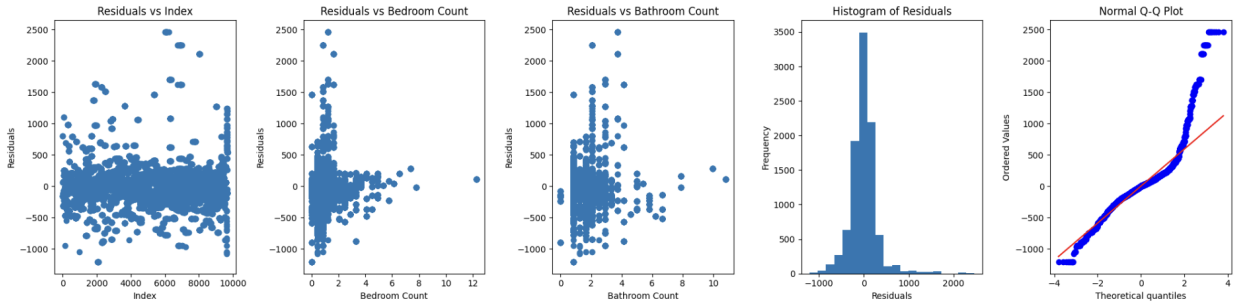
Because there are slightly curvy patterns in both residual plots, we then tried adding in polynomial terms to see if a polynomial multiple model could fit the data better.

```
X = np.column_stack((bedroom_count_scaled, bedroom_count_scaled ** 2,
                     bathroom_count_scaled, bathroom_count_scaled ** 2,
                     zipcode_dummy))
y = price_sqft.to_numpy()
```

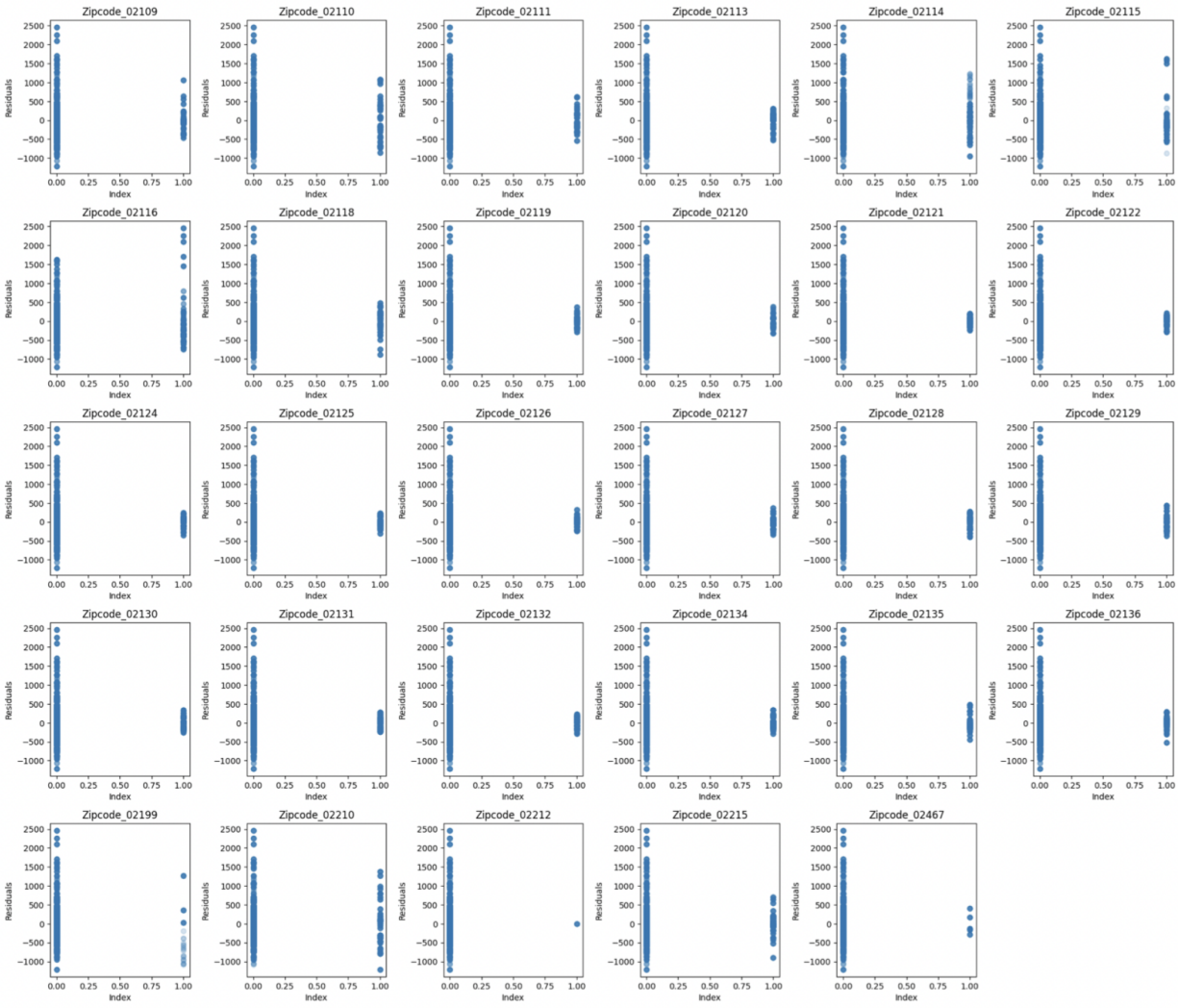
Single Cross Validated MSE and R^2:
MSE = 109150.82069026845
R^2 = 0.648691050442415

MSE and R^2 of final trained model:
MSE = 102406.57349707292
R^2 = 0.6607094571283605

Slopes for the final trained model:
[1219.35212112, -223.65291354, 13.26802951, 253.15633085, -17.6224271 , -312.70038692, 229.81086636, -484.77470293, -339.37840104, -236.12963765, -151.1311113 , 220.84226041, -259.0712244 , -903.90536137, -709.35576271, -956.86436305, -825.03046692, -864.6138177 , -813.76783917, -943.42099457, -591.78444569, -808.71887165, -564.82055719, -741.14921599, -857.55950287, -819.6176682 , -617.99866571, -604.84225908, -899.32372113, 991.96658593, 139.51414093, -819.65820235, -287.55649192, -700.04346295]



The regression diagnostics reveal that our model has moderate explanatory power with an R^2 of 0.661, indicating it captures around 66.1% of the variance in housing prices, and the MSE is 102406.57. Both R^2 and the MSE didn't improve significantly from the first simpler multiple regression model. The Residuals vs. Index plot does not show apparent patterns, which is promising for the independence assumption. Yet, the plots for Residuals vs. Bedroom and Bathroom Counts, again, suggest heteroscedasticity, showing a much higher variation in the middle than in tails. The Histogram of Residuals still shows a skewed distribution, and the Normal Q-Q Plot confirms that residuals deviate from normality, particularly at the extremes.



These residual plots of residuals vs. each ZIP Code look very similar to that of the first model. They display distinct patterns of prediction errors across various ZIP Codes for a housing price model. Most plots show an almost uniform distribution above and below 0, suggesting the model is fairly good at predicting housing prices for those areas, without systematic over- or under-predictions. However, because the ZIP Codes we're using have a wide distribution, each ZIP Code might be located in an area far from each other, which makes the difference of housing price become larger or smaller.This variation could be due to the distinct market dynamics and property characteristics in each ZIP Code. The analysis underscores the importance of location in housing prices and the challenge of creating a one-size-fits-all model for a diverse real estate market.

Discussion:

In this comprehensive analysis of the Boston real estate market using Redfin datasets, we address three key questions.

1. Price differences by ZIP Code: Our analysis first explored the differences in average selling price per square foot across Boston's ZIP Codes. The bar chart analysis highlights significant differences in house prices, with certain ZIP Codes having much higher prices. This change shows that location is one of the main determinants of property value. While the reasons for this might relate to socioeconomic status, the data provides a clear map for potential sellers to identify high-value areas and a more affordable one for buyers.

2. Seasonal sales patterns: The seasonality of the housing market is evident through our line chart analysis, which shows clear peaks in sales during certain months. This seasonal trend can guide sellers to effectively time their entry into the market, for example by listing during periods of high demand to maximize returns. For buyers, this pattern could signal a time to buy, perhaps during the off-season when potential competition is waning. Notice that because Redfin only provides the most recent sold houses information for each ZIP Codes, so average sold price for the months further away are less trustworthy because of the lack of data, but one takeaway from the line chart is that the average house sold price have a continuing drop since July 2023.

3. The impact of property characteristics on price: We choose the multiple regression model as our final model, because it has very similar R^2 and MSE as the polynomial multiple regression model, but can be more general and better fit future incoming data. The model has an R^2 value of 0.657, which is relatively high, and also indicates that bedrooms and bathrooms do have an impact on pricing. From the slope of details, the bedroom is negatively correlated with price, while the bathroom is positively correlated, which means the more the bathroom or the less the bedroom, the price is higher. However, the MSE is still high, suggesting that our model did not capture the role played by other factors. The inclusion of more data on property conditions or renovations could improve the accuracy of these forecasts.

In conclusion, while our models provide valuable insights into the Boston real estate market, they also highlight the complexity of real estate economics, influenced by a range of factors outside the scope of our current dataset. To build a more reliable forecasting model, future research should integrate a wider range of variables, possibly including macroeconomic indicators, local real estate policies, and demographic changes. This comprehensive modeling will enhance our understanding and provide more reliable tools for real estate decisions.

Citation:

Our data source is the Redfin website: redfin.com.

We reference the City of Boston website for specific ZIP Codes in Boston area: cityofboston.gov