Predicting Housing Prices

MATH7243 Final Project Report
Northeastern University

Sam Delfino, Katherine Chapkis, Megan Chin, Lichen Yang

April 23 2023

1 Abstract

Housing in Boston is notoriously expensive and unpredictable, especially for the 152,000 students living in Boston. Our project attempts to predict housing prices and categorize housing as affordable or not affordable for our prospective college student: a Northeastern University student with two roommates, each able to pay 1,727 dollars/month for housing. To achieve this goal, this project applied Linear Regression, Lasso Regression, Ridge Regression, Logistic Regression, LDA, QDA, and DNN. This project will help future students and potential renters find affordable housing based on the measures defined in our project. This gives more information and ownership of the search for housing to the renter.

2 Introduction

There are many factors that go into choosing the price of a house, such as: location, age and size of the property, distance from public transportation, etc. In addition to features of the house itself that affect the price, there are also economic phenomena that affect the market, including the interest rate and demand. We aim to construct a model that allows for the prediction of housing prices based on varying factors, giving potential homeowners/renters a more precise estimate on the price of the house they are interested in. In the end, we hope to give students, renters, and homeowners the opportunity to input their parameters and get an accurate prediction of the price of the apartment or home they are looking at. This would eliminate the guessing behind if potential renters or homeowners are receiving the correct estimates for the home/apartment they are looking for.

The most important questions we are trying to answer are: what is the most accurate price of a home in various Boston neighborhoods based on multiple key factors? What factors affect the price of a home? Are there areas where the price is overly inflated, and why?

3 Related Work

There is a lot of published research about predicting housing prices, with a lot acknowledging how difficult it is to predict these prices. Many research papers used different metrics in their attempt to predict housing prices, from economic to socioeconomic to geographic factors, as well as the specifics of the house. One metric that stood out was the FHFA Housing Price Index, which "is a comprehensive collection of public, freely available house price indexes that measure changes in single-family home values based on data from all 50 states and over 400 American cities that extend back to the mid-1970s" (Federal Housing Finance Agency, 2023). This is a broad value that can be used as a basis for our research and analysis. In one research paper, X. Q. Guan and H. Burton introduce the statistical theory of modern machine learning algorithms and propose three equations meant to guide parameter selection (2022). This criterion outlines the relationship between model complexity and generalizability via an objective function, which is then minimized to obtain an optimal complexity. We aim to use this criteria in our project to close the gap between training and testing error.

The reading of published research regarding predicting housing prices showed us that finding the right factors is not the only important question we should be answering, but what model will best predict the housing price. Muralidharan and Sharmila used Decision Trees and Artificial Neural Networks to predict the price of house properties (2018). "Housing Price Prediction via Improved Machine Learning Techniques" used different regression models, including Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine, Hybrid Regression, and Stacked Generalization (Truong et al., 2020) in an attempt to predict housing prices.

Another study used the Hedonic Pricing Model. Researchers applied various models, including a Correlation Analysis, Stepwise Regression Analysis, and Subsets Regression Analysis. The Subsets Regression Analysis provided the highest R-Squared value, leading it to be the chosen model to conduct the Hedonic Pricing Model. The researchers stated in their conclusion that including variables such as "inflation rate, the interest rates, the supply and demand for housing, and the unemployment rates" could increase the accuracy of the model (Jafari et al., 2019). The different approaches provide insights into many possibilities, for a basis to build our own approach.

4 Methods

Our initial research began with the publicly available Massachusetts Government property dataset. The datasets show specific geographical locations of all the neighborhoods that were surveyed. For the purpose of our analysis, we cleaned the dataset to only include zip codes in the greater Boston area. We also added property assessment from Boston Building Inventory as well as economic data from FRED for the Boston area to see if there are any additional

correlations between these variables.

After initial data cleaning and concatenation of the relevant datasets, we performed feature selection using a correlation matrix to determine which features from the Boston government property datasets were most correlated with the average total price of the documented properties. From here, we selected the eight zip codes closest to Northeastern University, and filtered the dataset accordingly. We constructed various regression models, including Linear, Ridge and Lasso Regression for each zipcode with the aim of finding the best model to approximate an average property value for that region. With these average property values, we were able to conclude which Boston zip codes are most affordable for our potential renter. Individual regression models, Linear, Ridge and Lasso Regression, were constructed for the cheapest zip code to demonstrate the correlation between the property values and property features, thereby reinforcing the validity and applicability of our findings.

From here, current property listings in the four cheapest zip codes were pulled from Zillow.com and filtered according to the needs of our renter: 3 bedrooms and a total rent maximum of 5200 dollars/month. At this point, the price feature was isolated and one-hot encoded to represent properties that are 0: not affordable for our renter, or 1: affordable for our renter. Then, we constructed various regression and classification models, including Linear Regression, Lasso Regression, Ridge Regression, Linear Discriminant Analysis, Logistic Regression, and DNN, to obtain one with high accuracy. Thus, we can determine a model to accurately discern a property as affordable or not.

5 Results and Analysis

To be able to prepare the data, to maximize results, we conducted a correlation matrix, shown in Figure 1, to find the most significant columns in predicting the most affordable neighborhoods and affect the price the most. The columns found to be the most significant were

`GROSS_TAX, LAND_VALUE, and BLDG_VALUE.`

These results are consistent with general assumptions that the value of the building, the value of the land and the taxes would affect the overall price of the property.

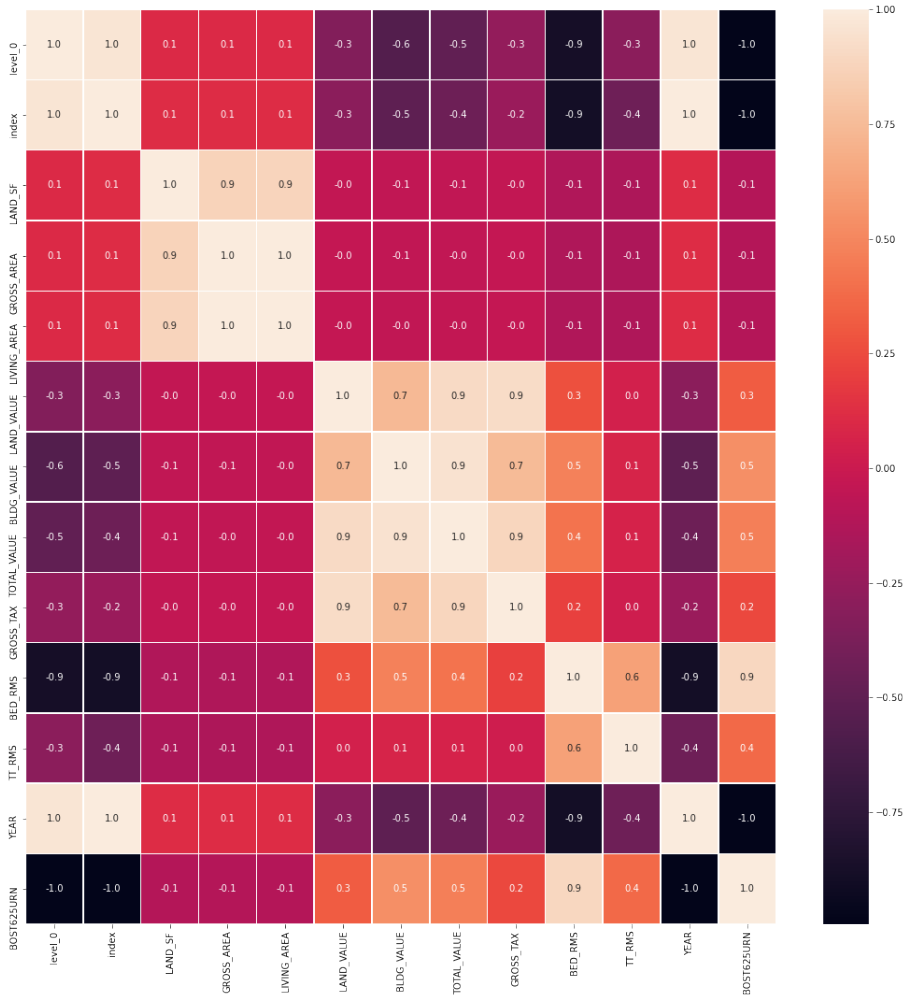


Figure 1 - Correlation Heatmap

We conducted Metric analysis for the eight nearest zip codes: 02108, 02111, 02115, 02116, 02120, 02125, 02127, 02215. Using the Boston government property dataset, we fitted Linear, Ridge, and Lasso Regressions to determine the best method for each zip code in finding the most affordable area of living. Based on the result of these fitting methods we found that 02115 is the most affordable zip code in Boston with an average value of 526,719 dollars per property. This result was found with the OLS model, which was the best fitting model for the 02115 neighborhood. All of the models used: Linear, Lasso, and Ridge Regression were all very accurate in predicting the most affordable zip code to live in, getting accuracy scores of 0.999, but OLS was still the best of three models.

With the knowledge of 02115 being the most affordable zip code for our col-

lege students, we attempted to determine which factors were the best predictors of price for the 02115 neighborhood. We first ran an analysis to find if there exists a relationship between the unemployment rate and the housing price index in Boston with Linear, Ridge and Lasso Regression. The Linear Regression model received an R-Squared score of -0.01497, the Ridge Regression model received a score of -0.01497, and a Lasso Regression model received a score of -0.01478. These scores indicate that there is no significant correlation between unemployment rate and housing price index. This is interesting because it is commonly assumed that raises in unemployment affect the demand for buying houses, affecting the price of housing.

To find what other factors were significant to the average housing price, we plotted a correlation heatmap for all the property factors in the government property dataset, and we found that gross tax, land value, and building value are the three factors that affect the total value of a property the most. To confirm the effects of these factors to the property value, we fitted a Linear, Ridge, and Lasso Regression model for these factors on the dataset with zip code 02115. We found a Linear Regression score of 0.999926, a Ridge Regression score of 0.999925, and a Lasso Regression score of 0.999905. Thus, indicating that we can very accurately predict, with all of the regression models, the property value with gross tax, land value, and building value for zip code 02115. It can also be concluded that all regression models have the same accuracy with negligible differences.

To determine if a property was affordable or not for our model student, we ran multiple classification models, attempting to find the most accurate model. First, we used Logistic Regression, which resulted in an accuracy score of 0.95, correctly classifying a property 95 percent of the time. This is an extremely accurate result. The ROC curve in Figure 2 reflects the level of accuracy in predicting the True Positives and True Negatives, with the area under the curve being 0.842. The model is not perfect, but it predicts the True Positives and True Negatives majority of the time.

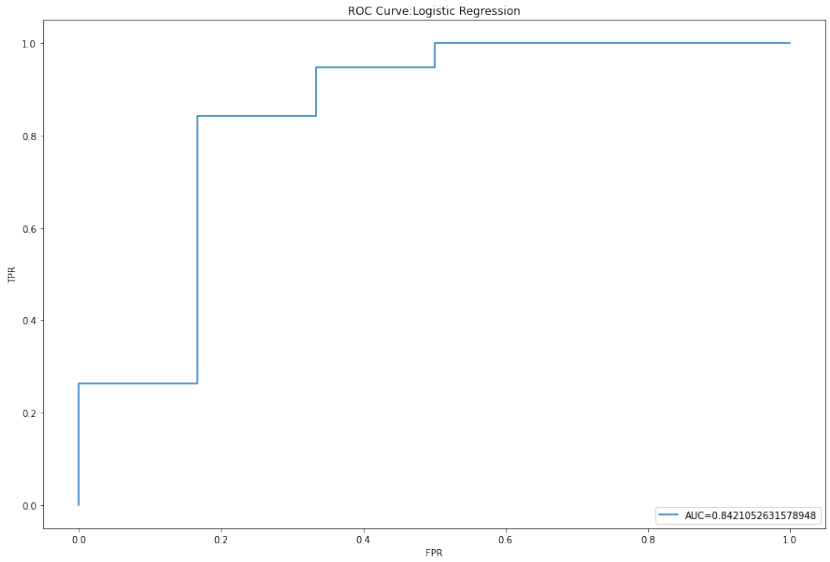


Figure 2 - ROC Curve from Logistic Regression Model

Linear Discriminant Analysis produced a similar result as the Logistic Regression model with an accuracy score of 0.95. Figure 3 shows that the ROC Curve was identical to the ROC curve produced from the Logistic Regression model, with an area under the curve 0.842, showing that both the Linear Discriminant Analysis and Logistic Regression models are very accurate in predicting the affordability of the listing.

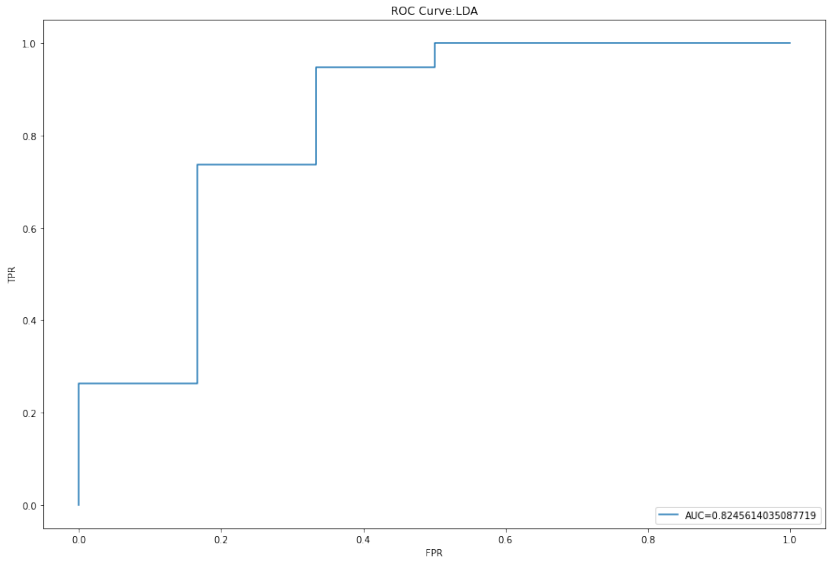


Figure 3 - ROC Curve from Linear Discriminant Analysis

While Linear Discriminant Analysis and Logistic Regression models produced extremely similar results that were very good, we attempted to achieve even better accuracy by applying a Sequential Neural Network.

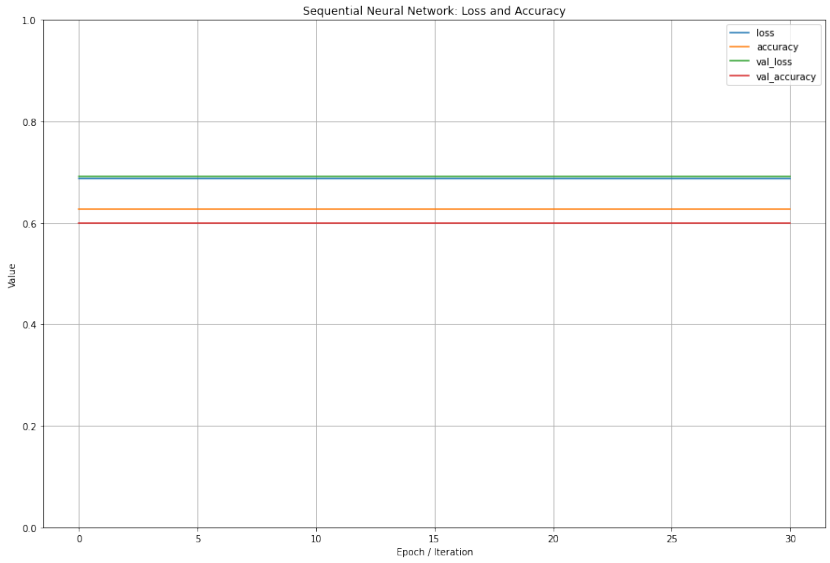


Figure 4 - Accuracy Scores from Sequential Neural Network Model

The Sequential Neural Network did worse than the Linear Discriminant Analysis and Logistic Regression models. The Sequential Neural Network achieved an accuracy score of 0.6267 and val-accuracy score of 0.6. Despite the accuracy score being lower than the Linear Discriminant Analysis and Logistic Regression models, the loss and val-loss scores were very similar and the val-accuracy and accuracy scores were very similar, displayed in Figure 4. This means the accuracy of the model is consistent and the model is not overfitting, correctly classifying properties in both the test and training data with similar precision.

Ultimately, the best model to predict whether an apartment is affordable or not are the Linear Discriminant Analysis and Logistic Regression models, accurately predicting both affordability and unaffordability 95 percent of the time.

6 Conclusion

Through many different approaches and analysis, this project was able to accurately predict and find apartments for rent for the model student and their three roommates. The project was able to find the factors that most contributed

to the price of a house/apartment, narrow down the most affordable neighborhood that the student could live in in the city of Boston, then predict whether current Zillow listings were affordable or not for the student. The factors that most influenced the price of the house/apartment were its building value, land value, and gross tax. The neighborhood that was most affordable for the model student was 02115. The optimal models to predict the affordability of a current apartment were the Linear Discriminant Analysis and Logistic Regression models.

It's worth noting that a drawback to our model could be the loss of data when we were cleaning it. In the cleaning stage, we got rid of properties where one or more data points were missing and this could impact the validity of our results. Furthermore, while the scores for the classification models were not terrible, achieving an accuracy of over 60 percent is hard with our dataset since Zillow is constantly being updated with new listings so training a very accurate model is difficult.

To improve the result of our analysis we could run the regressions on other zip codes other than 02115 to see the accuracy of our result. Furthermore, we could certainly include more fields for analysis. For example, we could include crime rate, proximity to campus, and other factors that might affect the decision of renters. In addition to using other zip codes and other features, we could also try and use more data instead of the 2020 - 2022 timeframe. The only drawback of this would be that we would have to consolidate a lot of data which might make it harder to run the models in an efficient amount of time.

7 References

"Boston by the Numbers Colleges and Universities." http://www.bostonplans.org/getattachment/1770c181-7878-47ab-892f-84baca828bf3.

"House Price Index." FHFA House Price Index | Federal Housing Finance Agency, Federal Housing Finance Agency, 2023, https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index.aspx.

Jafari, Amirhosein, and Reza Akhavian. "Driving Forces for the US Residen-tial Housing Price: A Predictive Analysis." Built Environment Project and Asset Management 9.4 (2019): 515-29. ProQuest. Web. 5 Feb. 2023.

"Most expensive places to live in the U.S. in 2022-2023." https://realestate.usnews.com/places/rankings/most-expensive-places-to-live

Muralidharan, Sharmila, et al. \Analysis and Prediction of Real Estate Prices: A Case of the Boston Housing Market." Issues In Information Systems 19.2 (2018): 109{118. Web 5 Feb. 2023.

"STUDENT HOUSING TRENDS 2018-2019 ACADEMIC YEAR." City of Boston.

Truong, Quang, et al. "Housing Price Prediction via Improved Machine Learning Techniques." Procedia Computer Science 174 (2020): 433{442. Science Direct. Web. 5 Feb. 2023.

X. Q. Guan and H. Burton, Bias-variance tradeoff in machine learning: Theoreti -cal formulation and implications to structural engineering applications, Structures 46 (2022), 17-30.

Appendix A: Data Tables

	index	ZIPCODE	LAND_SF	GROSS_AREA	LIVING_AREA	LAND_VALUE	BLDG_VALUE	TOTAL_VALUE	GROSS_TAX	BED_RMS	TT_RMS	YEAR	BOST625URI
0	0	2115.0	1563.0	3320.0	3058.0	908800.0	1512300.0	2421100.0	25566.82	11.0	7.0	2020	9.87
1	1	2115.0	1451.0	3332.0	3100.0	1085100.0	1591300.0	2676400.0	28262.78	8.0	4.0	2020	9.87
2	2	2115.0	1451.0	3452.0	3100.0	947400.0	2518000.0	3465400.0	36594.62	8.0	4.0	2020	9.87
3	3	2115.0	1451.0	3308.0	3076.0	1076600.0	1561400.0	2638000.0	27857.28	8.0	3.0	2020	9.87
4	4	2115.0	1648.0	3328.0	2694.0	1056100.0	1455700.0	2511800.0	26524.60	10.0	4.0	2020	9.87

Figure A.1: Excerpt of Property Data

	BOST625URN
YEAR	
2020	9.875
2021	5.575
2022	3.325

Figure A.2: Excerpt of Unemployment Data

	status_dttm	description	violation_city	violation_zip	ward
0	2023-03-13 15:19:03	Failed to comply w permit term	East Boston	02128	01
1	2023-03-13 13:38:48	Failure to Obtain Permit	East Boston	02128	01
2	2023-03-13 11:55:12	Unsafe Structures	Dorchester	02124	14
3	2023-03-13 11:54:38	Testing & Certification	Boston	02115	05
4	2023-03-13 10:35:08	Failure to Obtain Permit	Hyde Park	02136	18

Figure A.3: Excerpt of Building Violations Data

```
2     6727.0
4    10500.0
5     6698.0
7     4650.0
8     4000.0
Name: hdpData.homeInfo.price, dtype: float64
```

Figure A.4: Excerpt of Zillow Prices Data

	beds	baths	hdpData.homeInfo.zpid	hdpData.homeInfo.latitude	hdpData.homeInfo.longitude	hdpData.homeInfo.bathrooms	hdpData.homeInfo.bedrooms	hdpDa
2	3.0	2.0	2.061651e+09	42.352215	-71.059060	2.0	3.0	
4	3.0	3.0	2.077718e+09	42.362114	-71.059364	3.0	3.0	
5	3.0	3.0	2.080586e+09	42.351814	-71.062454	3.0	3.0	
7	3.0	1.0	2.063494e+09	42.344230	-71.089560	1.0	3.0	
8	3.0	1.0	2.058860e+09	42.345505	-71.089195	1.0	3.0	

Figure A.5: Excerpt of Zillow Listings Data

Appendix B: Results

```
Model fit with Linear Regression:
-0.01496812169296513
Mean Squared Error: 3028.0154716663883
Model fit with Ridge Regression:
-0.014967838653744625
Mean Squared Error: 3028.01462725845
Model fit with Lasso Regression (with CV):
-0.014784955858999638
Mean Squared Error: 3027.4690220121834
```

Appendix B: Regression Analysis Results for Correlation Between Housing Price Index and Unemployment Rate

```
 Metric Analysis for Zipcode 02108

Model fit with Linear Regression:
0.9997809920745001
Mean Squared Error: 577150827.0786208
Model fit with Ridge Regression:
0.9997812762850615
Mean Squared Error: 576401847.9712161
Model fit with Lasso Regression (with CV):
0.9994691555242071
Mean Squared Error: 1398932607.3689961
The regression method most accurate is: Ridge
Average total housing value with Ridge for zipcode 02108 (2020-2022): 1113000

 Metric Analysis for Zipcode 02111

Model fit with Linear Regression:
0.999987380752739
Mean Squared Error: 48046526.60588291
Model fit with Ridge Regression:
0.999987382116424
Mean Squared Error: 48041334.51078087
Model fit with Lasso Regression (with CV):
0.9999933509089132
Mean Squared Error: 25315751.819274098
The regression method most accurate is: Lasso
Average total housing value with Lasso for zipcode 02111 (2020-2022): 1278415
```

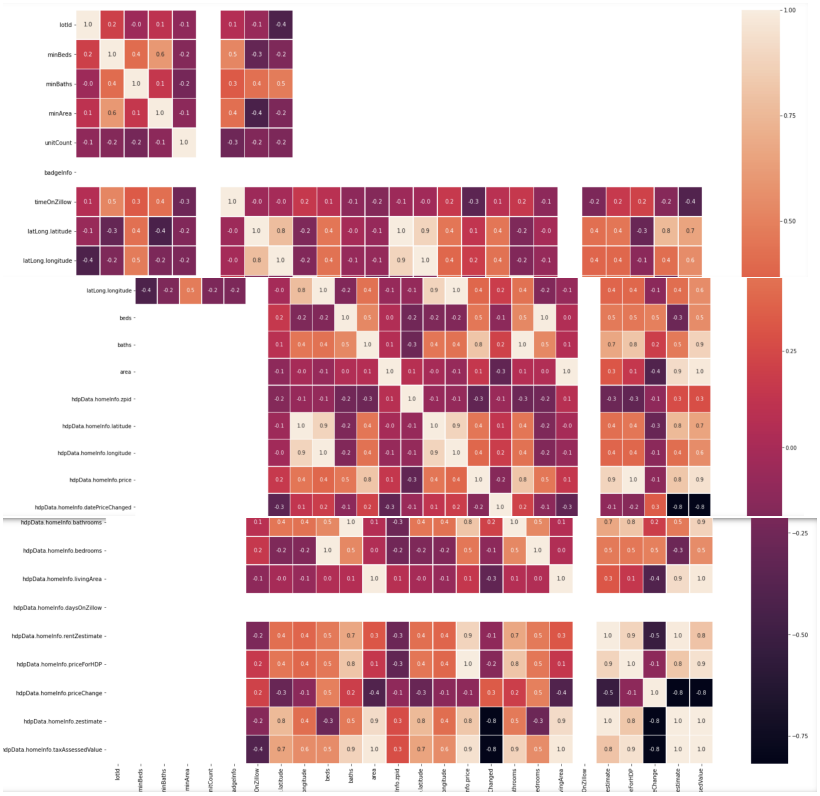
```
 Metric Analysis for Zipcode 02115

Model fit with Linear Regression:

/Users/sdelfino/opt/anaconda3/lib/python3.8/site-packages/sklearn/utils/validati
lumn-vector y was passed when a 1d array was expected. Please change the shape o
ng ravel().
  return f(**kwargs)

0.9990713562107881
Mean Squared Error: 76720144.6341697
Model fit with Ridge Regression:
0.9990708511946422
Mean Squared Error: 76761866.67249368
Model fit with Lasso Regression (with CV):
0.9988891083141916
Mean Squared Error: 91776601.3171251
The regression method most accurate is: OLS
Average total housing value with OLS for zipcode 02115 (2020-2022): 526701

 Metric Analysis for Zipcode 02116

Model fit with Linear Regression:
0.9999999178745604
Mean Squared Error: 24436.91387194663
Model fit with Ridge Regression:
0.9999999191186318
Mean Squared Error: 24066.7330144201
Model fit with Lasso Regression (with CV):
0.9999922891210966
Mean Squared Error: 2294417.962862132
The regression method most accurate is: Ridge
Average total housing value with Ridge for zipcode 02116 (2020-2022): 1197922

 Metric Analysis for Zipcode 02120

Model fit with Linear Regression:
0.9999172507593761
Mean Squared Error: 1310390449.050677
Model fit with Ridge Regression:
0.9999172507228338
Mean Squared Error: 1310391027.7245505
Model fit with Lasso Regression (with CV):
0.9999040022393149
Mean Squared Error: 1520189765.8959105
The regression method most accurate is: OLS
Average total housing value with OLS for zipcode 02120 (2020-2022): 1173301
```

```
 Metric Analysis for Zipcode 02125

Model fit with Linear Regression:
0.9999776343111962
Mean Squared Error: 297499174.94799906

/Users/sdelfino/opt/anaconda3/lib/python3.8/site-packages/sklearn/linear_model/_ri
itioned matrix (rcond=2.17491e-18): result may not be accurate.
  return linalg.solve(A, Xy, sym_pos=True,

Model fit with Ridge Regression:
0.9999776342083495
Mean Squared Error: 297500542.97277
Model fit with Lasso Regression (with CV):
0.9999713247182036
Mean Squared Error: 381426780.5789528
The regression method most accurate is: OLS
Average total housing value with OLS for zipcode 02125 (2020-2022): 1091188

 Metric Analysis for Zipcode 02127

Model fit with Linear Regression:
0.9999958243908091
Mean Squared Error: 24309811.236456733
Model fit with Ridge Regression:
0.9999958623660862
Mean Squared Error: 24088724.49832089
Model fit with Lasso Regression (with CV):
0.9999909703684429
Mean Squared Error: 52569248.85757052
The regression method most accurate is: Ridge
Average total housing value with Ridge for zipcode 02127 (2020-2022): 2034498

Metric Analysis for Zipcode 02215

Model fit with Linear Regression:
0.9999256856212774
Mean Squared Error: 2500238094.2289367
Model fit with Ridge Regression:
0.9999251499501555
Mean Squared Error: 2518260250.480517
Model fit with Lasso Regression (with CV):
0.9999048530687719
Mean Squared Error: 3201129930.6344295
The regression method most accurate is: OLS
Average total housing value with OLS for zipcode 02215 (2020-2022): 1498356
```

Appendix B.2: Metric Analysis for Top 8 Zip Codes

```
Model fit with Linear Regression:
0.999971810315485
Mean Squared Error: 26866188.399756532
Model fit with Ridge Regression:
0.9999718103154851
Mean Squared Error: 26866188.399605617
Model fit with Lasso Regression (with CV):
0.9999231844543632
Mean Squared Error: 73209081.85429057
```

Appendix B.3: Regression Analysis of the Three Cheapest Neighborhoods and the Three Most Significant Columns



Appendix B.4: Correlation Heatmap Between Zillow Property Features and Price

```
Shape of X_train, y_train:
(57, 8) (57,)
Shape of X_test, y_test:
(20, 8) (20,)

Classifying listings from Zillow with Logistic Regression:

[[ 6  0]
 [ 1 13]]
              precision    recall  f1-score   support

           0       0.86      1.00      0.92         6
           1       1.00      0.93      0.96        14

    accuracy                           0.95        20
   macro avg       0.93      0.96      0.94        20
weighted avg       0.96      0.95      0.95        20
```

Appendix B.5: Training and Testing Data Size and Logistic Regression

Accuracy Scores

```
Accuracy score of the Logistic Regression model: 0.95

Classifying listings from Zillow with Linear Discriminant Analysis:

[[ 5  1]
 [ 0 14]]
              precision    recall  f1-score   support

           0       1.00      0.83      0.91         6
           1       0.93      1.00      0.97        14

    accuracy                           0.95        20
   macro avg       0.97      0.92      0.94        20
weighted avg       0.95      0.95      0.95        20
```

Appendix B.6: Linear Discriminant Analysis Accuracy Scores