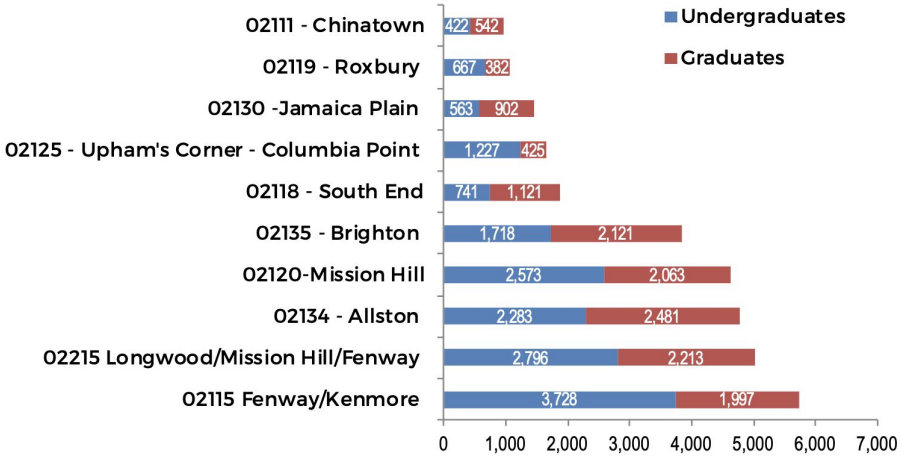
Predicting Housing Prices

Sam Delfino, Katherine Chapkis, Megan Chin, Lichen Yang

Background

- Housing is expensive in Boston, especially for students
- Boston ranked 25th most expensive city by US News
- 250,000 students attend college in Boston
- Almost 37,000 students are living in private housing in Boston
- Factors that affect price: location, size of property, accessibility, and transportation



Our Target Student

- Northeastern student
- Three bed apartment (with 2 roomates)
- Budget of $1,727 per month/student = $5,181 per month

Problems to address

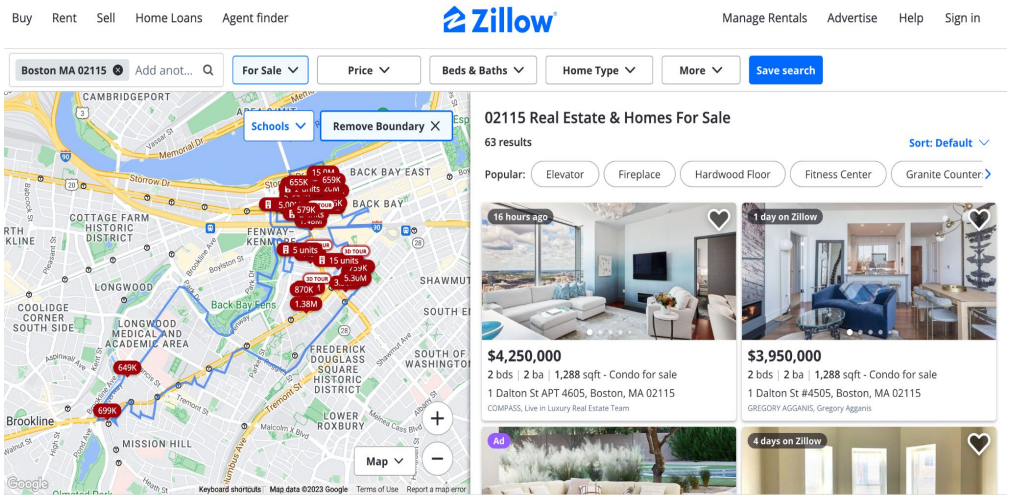
- What is the most accurate average price of a home in various Boston zip codes based on property features?
- What factors affect the price of a home?
- Are there areas where the price is overly inflated, and why?
- How many current Zillow listings are suitable for our target student, and how many of them would be in their price range?

Related works

- Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications(X. Q. Guan and H. Burton)
- Analysis and Prediction of Real Estate Prices: A Case of the Boston Housing Market(Muralidharan et al., 2018).
  - Decision tree and Neural network for property price prediction
- Housing Price Prediction via Improved Machine Learning Techniques(Truong et al., 2020)
  - Random Forest, Extreme Gradient Boosting, Light Gradient Boosting Machine, Hybrid Regression, and Stacked Generalization for price prediction.

Data

- Massachusetts Government property data (2020-2022)
  - Zip code, value of property, area of property
- FRED (Federal Reserve Economic Data)
  - Unemployment rate in Boston
  - Housing Price Index (HPI) in Boston
- Zillow
  - House listings
  - Prices



Methods + Results

Methods

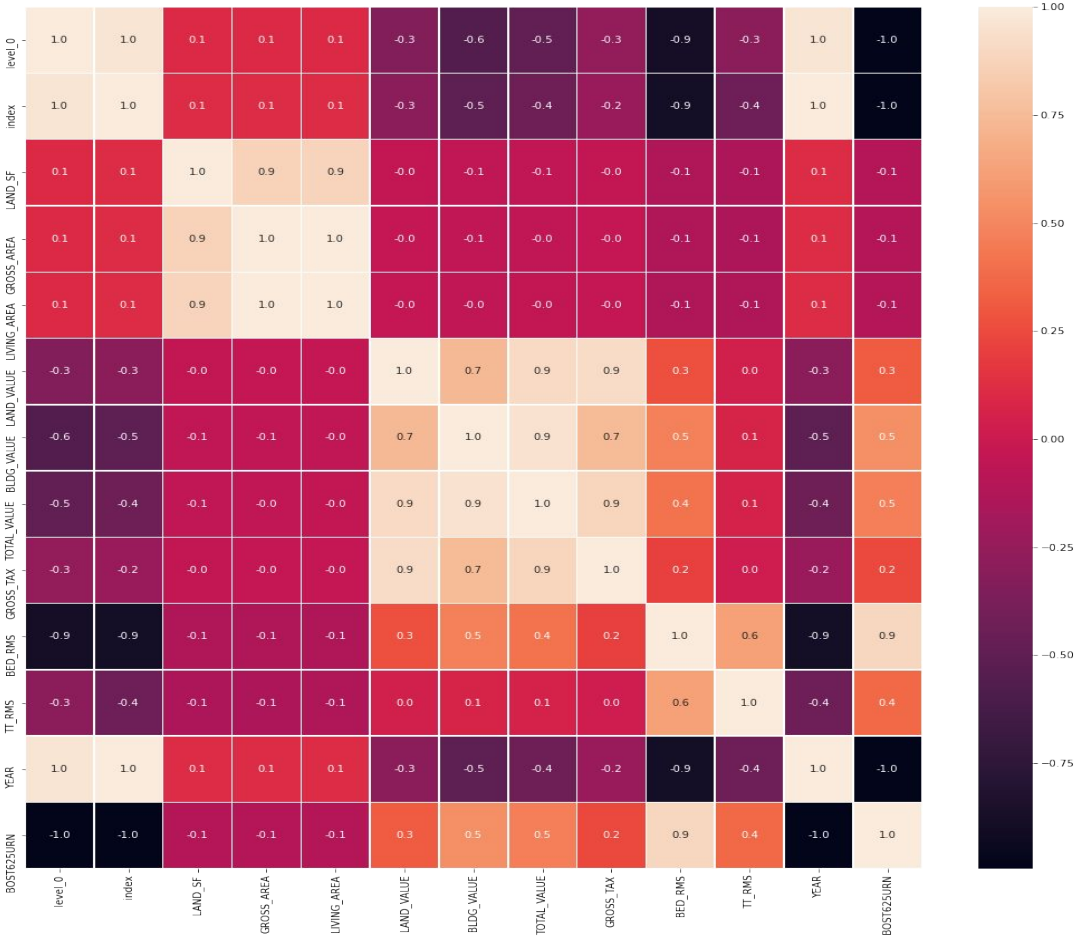
- REGRESSION
  - Data from Massachusetts Government and FRED
  - 8 zip codes closest to Northeastern
  - Goal is to find which zip code is the cheapest and which features are the best for our model
- CLASSIFICATION
  - Data from Zillow
  - 4 zip codes closest to Northeastern
  - 0 for not affordable and 1 for affordable
  - Purpose is to find how many properties satisfy our budget

Initial Zip Code Analysis and Fitting

- Purpose: to find which of the 8 nearest zip code in Boston has the cheapest housing prices

	02108	**02111**	**02115**	**02116**	**02120**	**02125**	**02127**	**02215**
Best Method	Ridge	Lasso	Ridge	Ridge	OLS	OLS	Ridge	OLS
Average Total Housing Value ($)	1,113,002	1,278,415	526,719	1,197,922	1,173,301	1,091,188	2,034,498	1,498,356

Feature Selection: Property Data



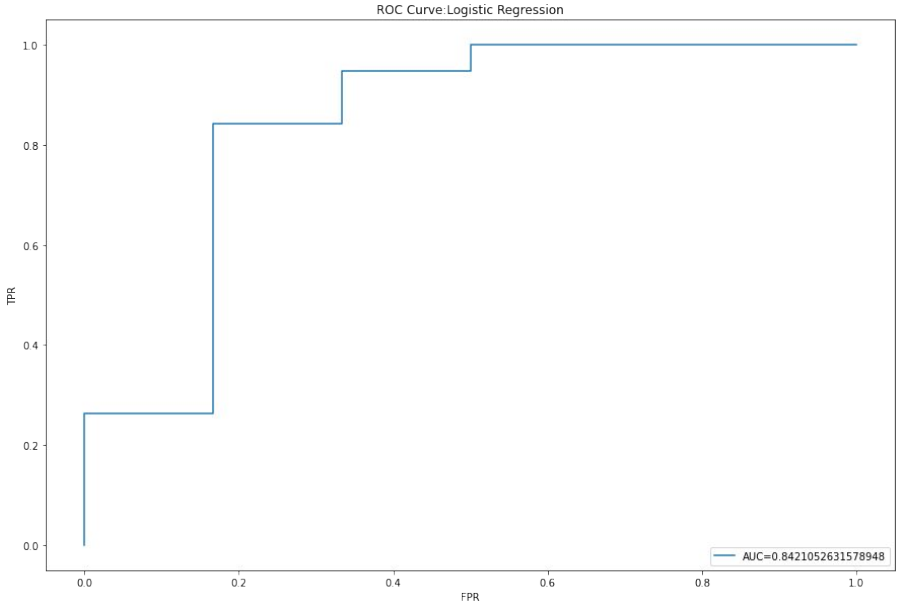
- Factors that affect total value the most are:
  - GROSS_TAX
  - LAND_VALUE
  - BLDG_VALUE
- Regressions all had >99% accuracy
- Model fit with Linear Regression:
  - 0.9998997524032356
- Model fit with Ridge Regression:
  - 0.9998997524032361
- Model fit with Lasso Regression (with CV):
  - 0.9999250309716614

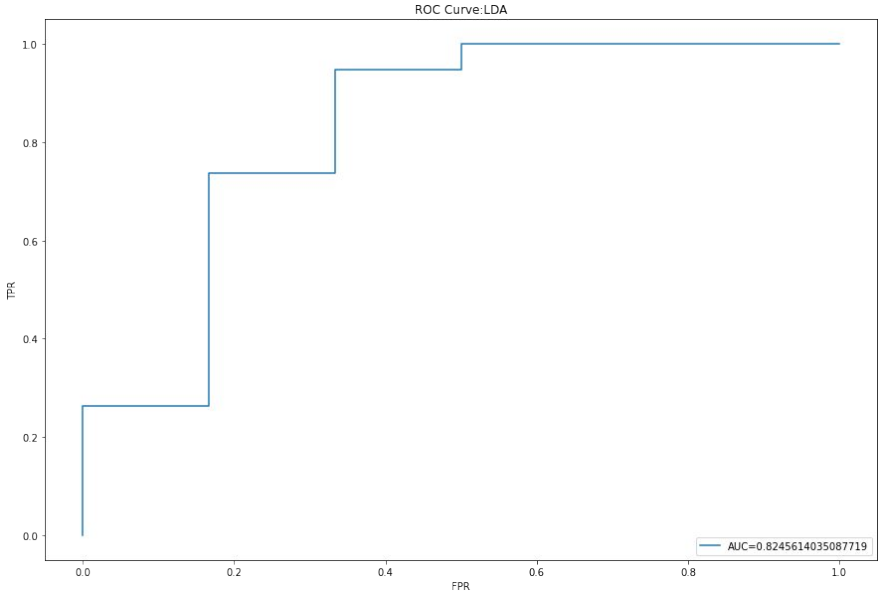
EDA: Unemployment Rates with HPI

- Tried to see if there was a correlation between unemployment rate and housing price index
- Didn't prove to be helpful
  - R^2 values were all low for linear, lasso, and ridge regressions
- Could've skewed our regressions if we used it

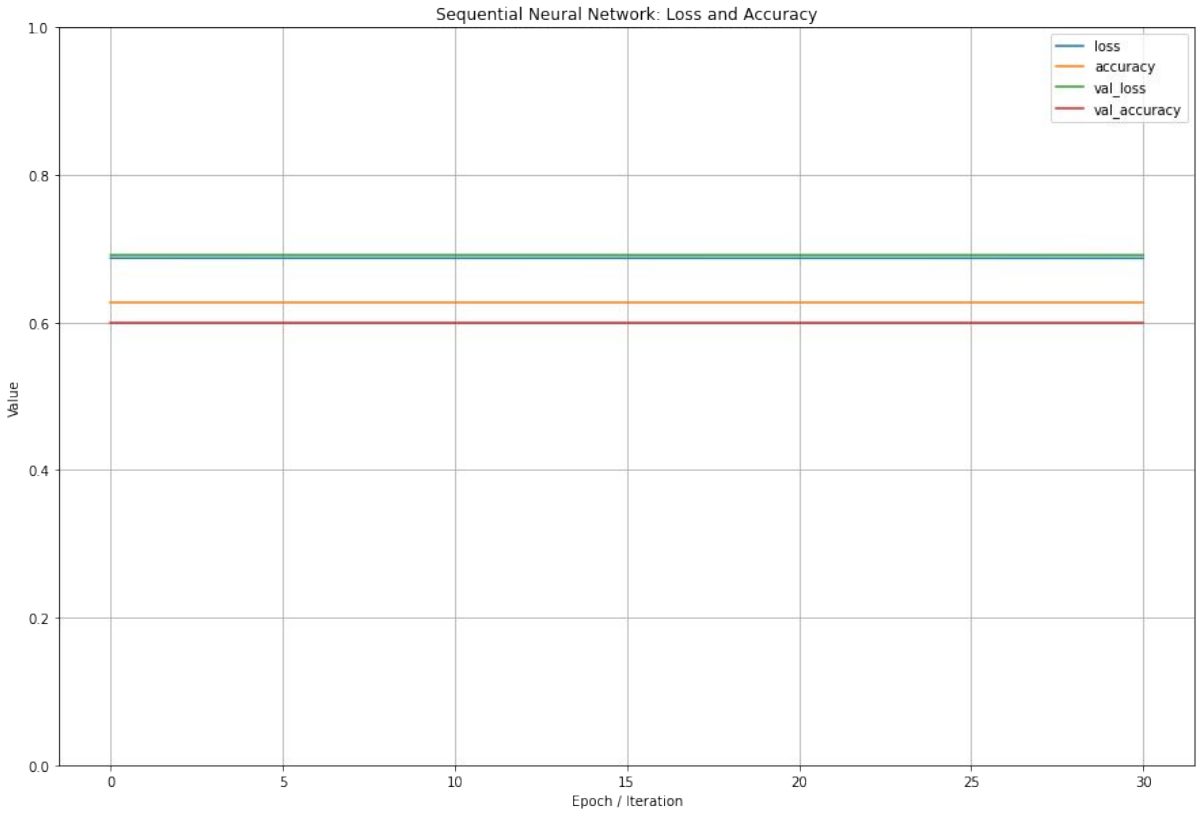
Logistic Regression and LDA

Logistic Regression and LDA: Accuracy Score: 0.95 AUC: 0.842





Deep Neural Network



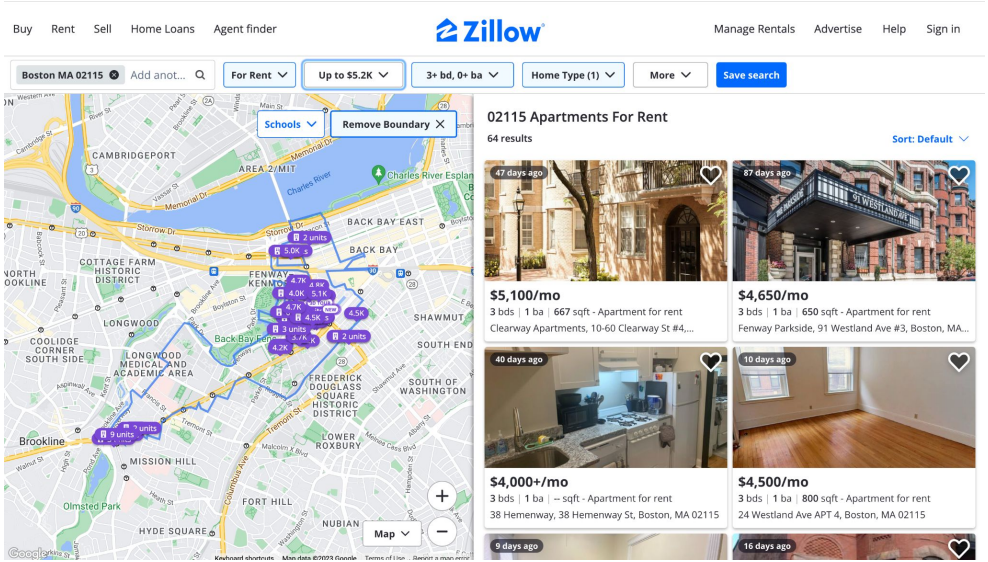
- **Accuracy score:** 0.6
- **Validation Accuracy:** 0.6267

Conclusion

Best Model: Linear Discriminant Analysis & Logistic Regression

- Accuracy Score: 0.95
- Area Under the Curve: 0.842

Most Affordable Neighborhood: 02115



Next Steps...

- Inclusion of more analysis fields
  - Crime rate
  - Proximity to campus
  - Other tangible factors that might affect a renter's decision
- Applying model to different/more zip codes to continually test the models

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 Residential 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: Theoretical formulation and implications to structural engineering applications, Structures 46 (2022), 17-30.

Q&A

Appendix

Figure A.1: Excerpt of Property Data

	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.2: Excerpt of Unemployment Data

YEAR	BOST625URN
2020	9.875
2021	5.575
2022	3.325

Figure A.3: Excerpt of Building Violations 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.4: Excerpt of Zillow Prices 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 Listings 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	

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

```
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.2.i: Metric Analysis for Top 8 Zip Codes

```
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
```

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

```
 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
```

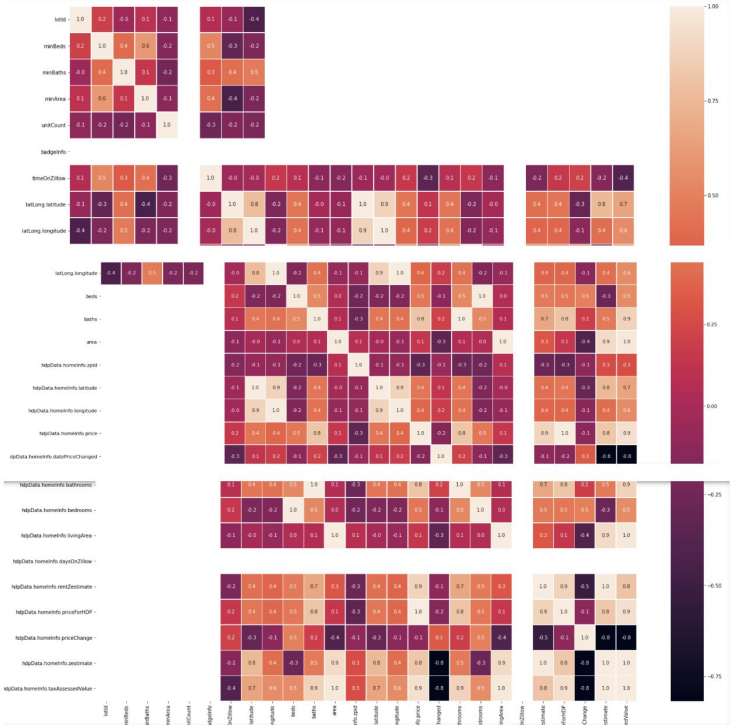
```
 Metric Analysis for Zipcode 02215

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
```

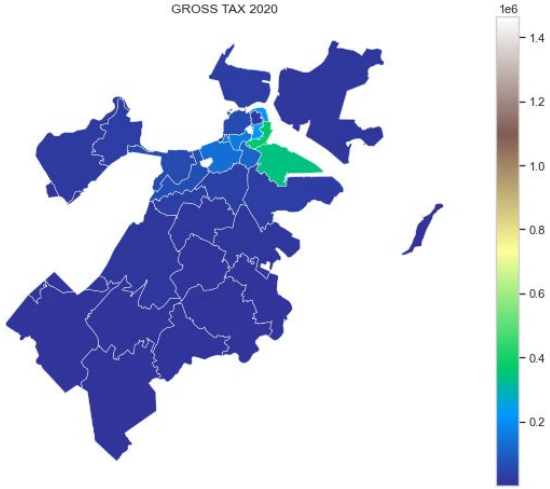
Appendix B.3: Regression Analysis of the Three Cheapest Neighborhoods and the Three Most Significant Columns

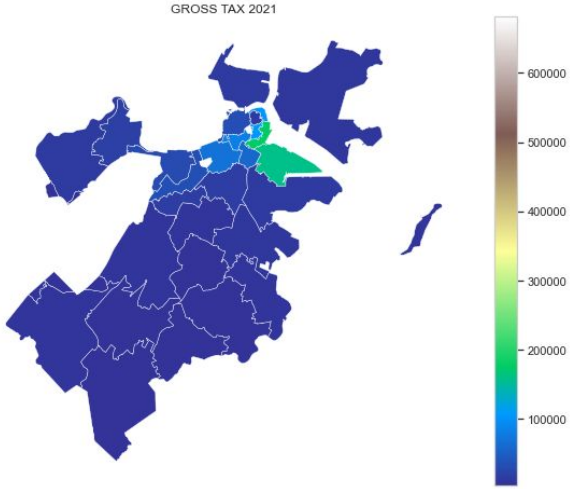
```
Model fit with Linear Regression:
0.999971810315485
Mean Squared Error: 26866188.399756532
Model fit with Ridge Regression:
0.9999718103154851
Mean Squared Error: 26866188.399605617
Model fit with Lasso Regression (with CV):
0.9999231844543632
Mean Squared Error: 73209081.85429057
```

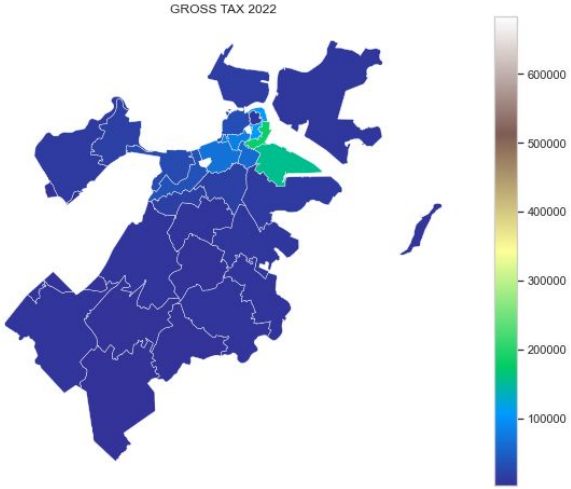
Appendix B.4: Correlation Heatmap Between Zillow Property Features and Price



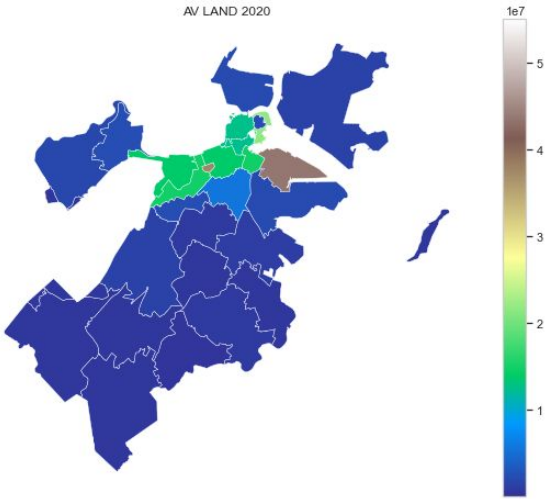
Appendix B.5: Average Gross Tax Visualization

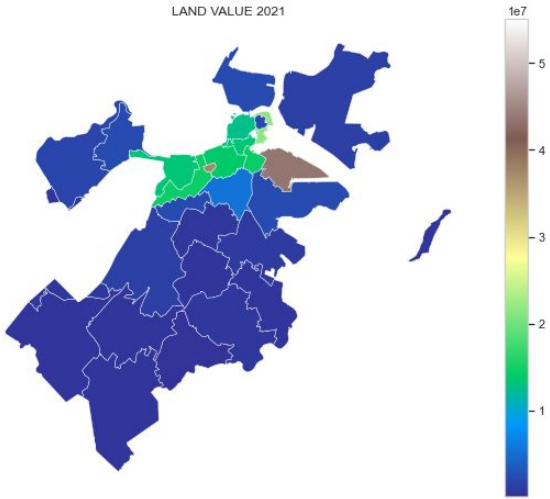


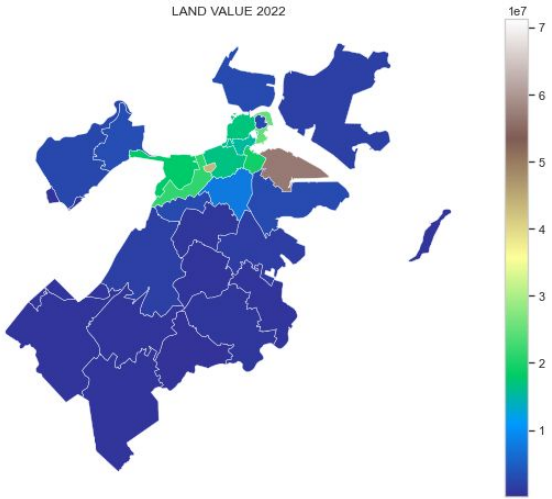




Appendix B.6:Average Land Value Visualization







Appendix B.7: Training and Testing Data Size and Logistic Regression Accuracy Scores

```
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.8: Linear Discriminant Analysis Accuracy Scores

```
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
```