Relationship Between Home Attributes and Price in the Greater Boston Area

Sarah Katz '24, Data Science Major Capstone

Background

- The housing market in the greater Boston area has become highly competitive.
- It is economically and informationally beneficial for brokers, appraisers, lenders, buyers and/or sellers to understand specifically which characteristics of homes are the most valuable.

Question: What home characteristics are most important for explaining the variation in home prices in the greater Boston area?

Data

- A newly developed web scraping program was used to create a dataset from Zillow with recent housing transactions.
- Time: November 2020 to November 2023.
- Location: four counties in Massachusetts (Norfolk, Essex, Suffolk, Middlesex) with 192 total zip codes.
- Cleaning:
  - Data from four counties were merged.
  - Limited to residential home sales, so home type = Townhouse, Condo or Single Family.
  - Removing missingness, unusual and/or impossible values resulted in eliminating 1368 houses (1.5% of rows).
  - Created 3 new variables:
    - distance in miles from Boston city hall
    - urban level categorical variable: 1-5, with 5 indicating the most urban area:
      - 5: 0-3 mi; 4: 3-6 mi; 3: 6-12 mi; 2: 12-20 mi; 1: 20+ mi from Boston city hall
    - renovation binary variable, inferred by tracking renovation indicator words in the home descriptions
- Size: **88,412 houses**

Data Exploration

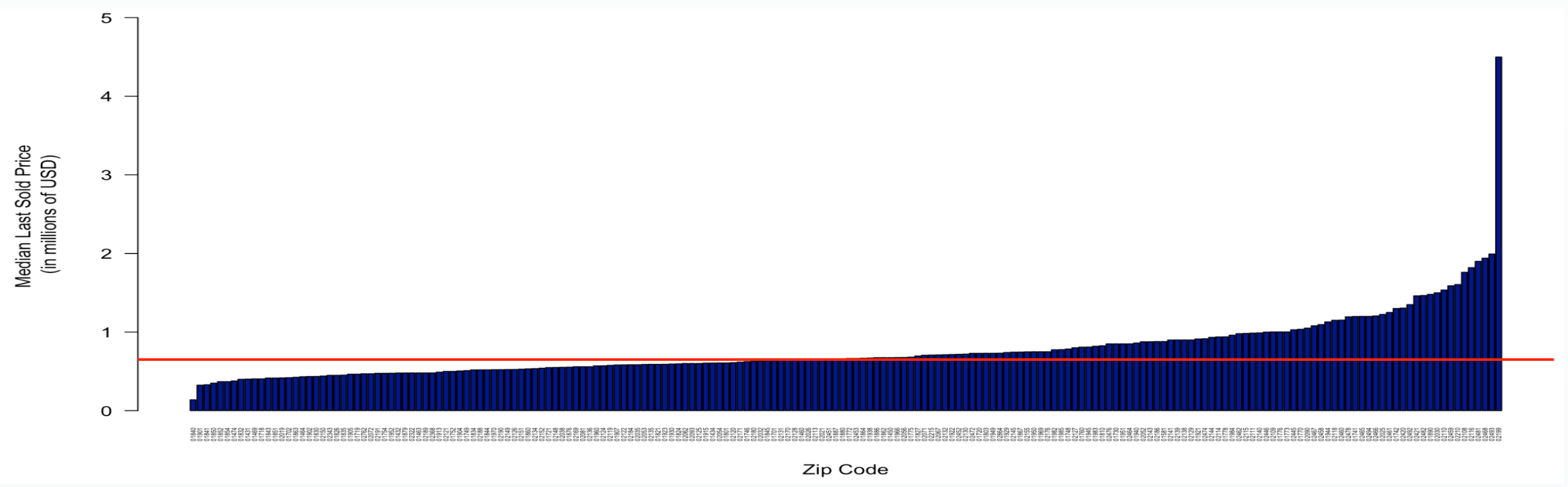


Figure 1: Median last sold prices for Boston area zip codes. Overall median sold price is $660,000. The zip code with the highest median sold price is 02199 (Back Bay) at $4.5 million.

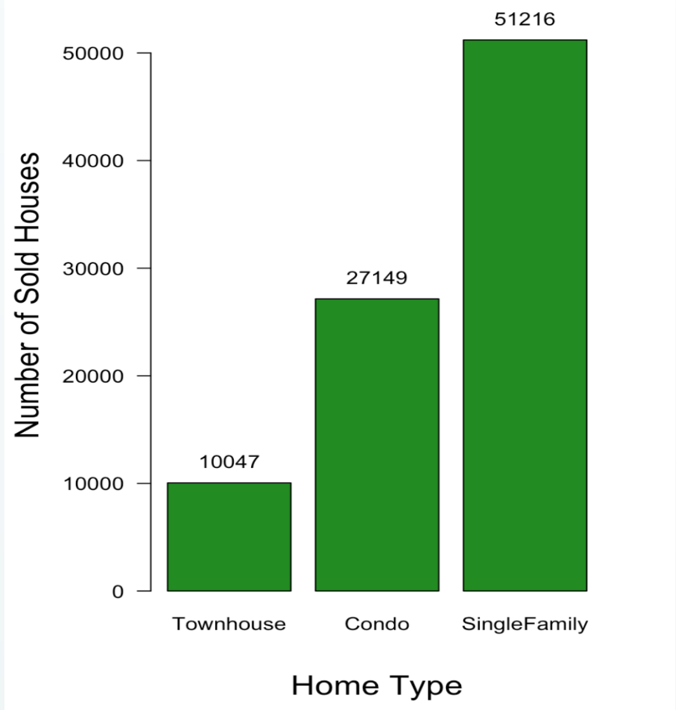


Figure 2: Distribution of home types.

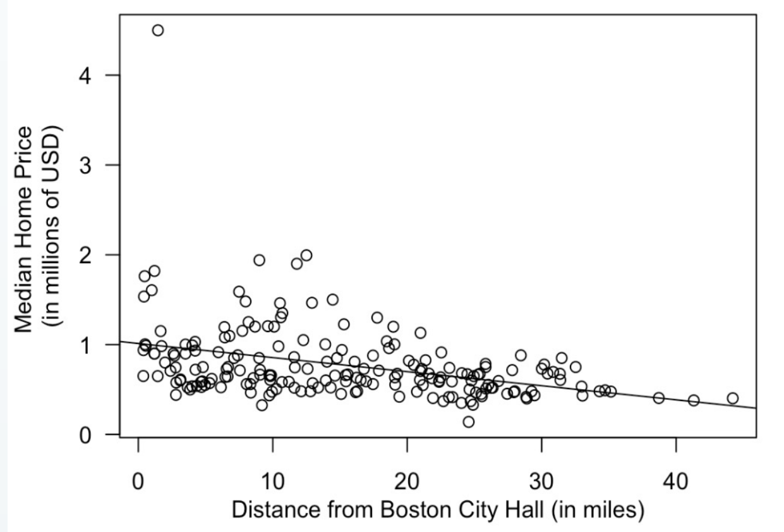


Figure 3: Zip codes by median home price and distance from city center. As expected, home prices decrease as distance from city center increases. Outlier: Back Bay.

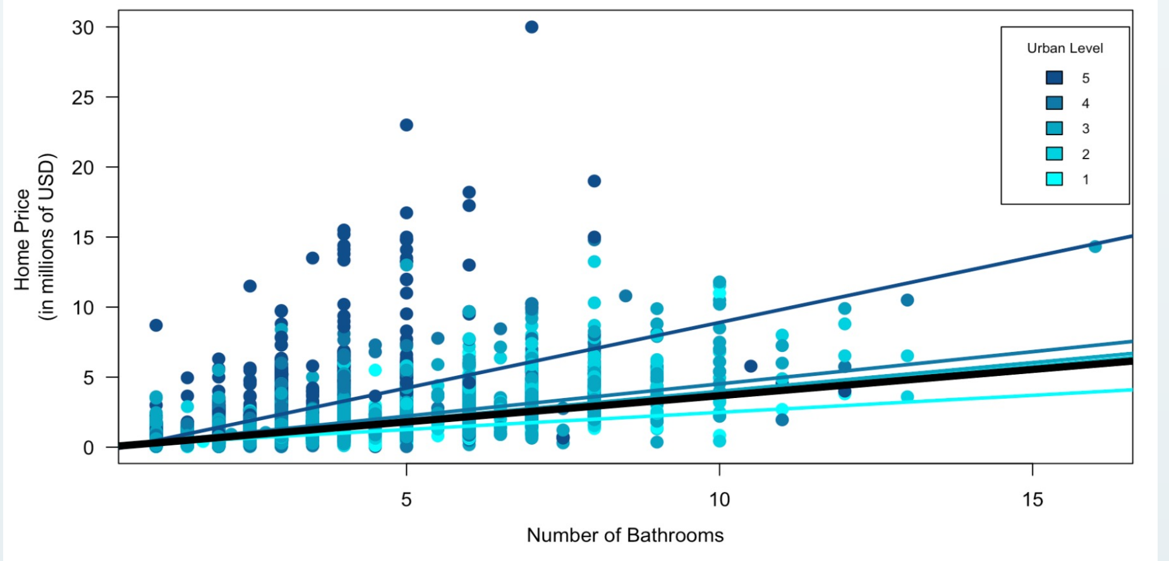


Figure 4: Home price by number of bathrooms, color coded by urban level. The most expensive homes have fewer bathrooms but are closer to the city.

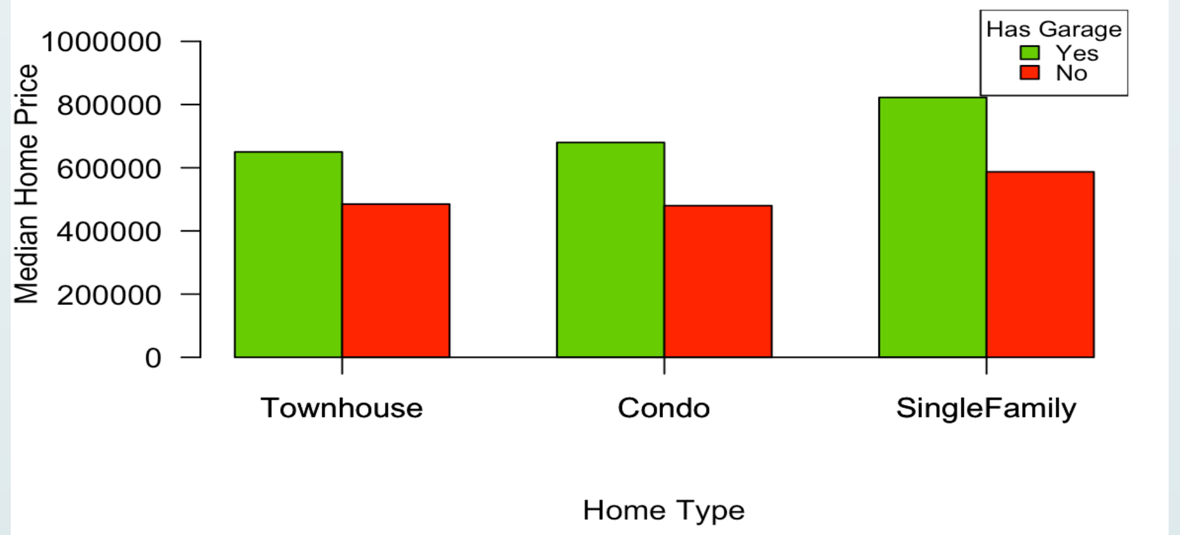


Figure 5: Median home price for different home types, based on whether the house has a garage. For all home types, having a garage increases the median home price.

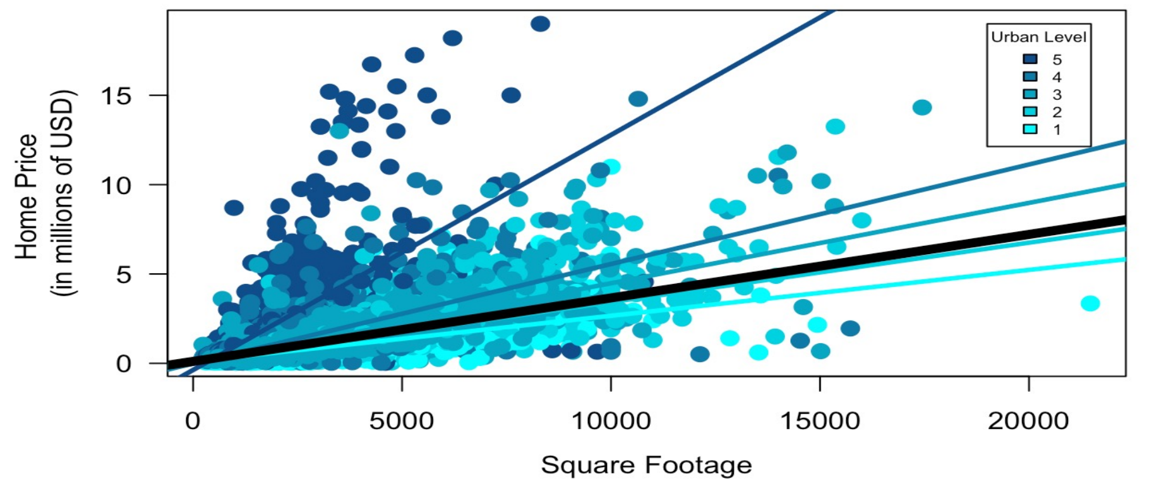


Figure 6: Home price by square footage, color coded by urban level. There is a general tradeoff between the size of a house and its proximity to city center.

Methods

- The home price variable on the original scale is severely right-skewed *(see Figure 7)*.
- A Box-Cox transformation showed optimal lambda = 0, therefore a log transformation on home price was performed *(see Figure 8)*.
- VIF analysis confirmed that there were no multicollinearity issues in the predictors.
- Using Cook's Distance, 2 influential observations were removed.

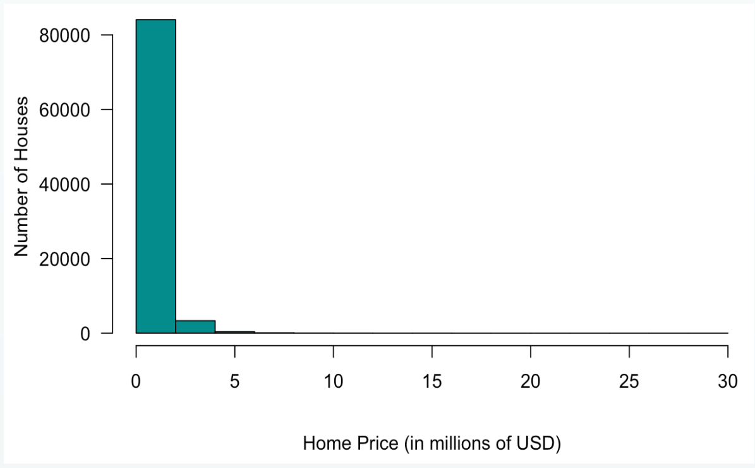


Figure 7: Histogram of home prices before log transformation.

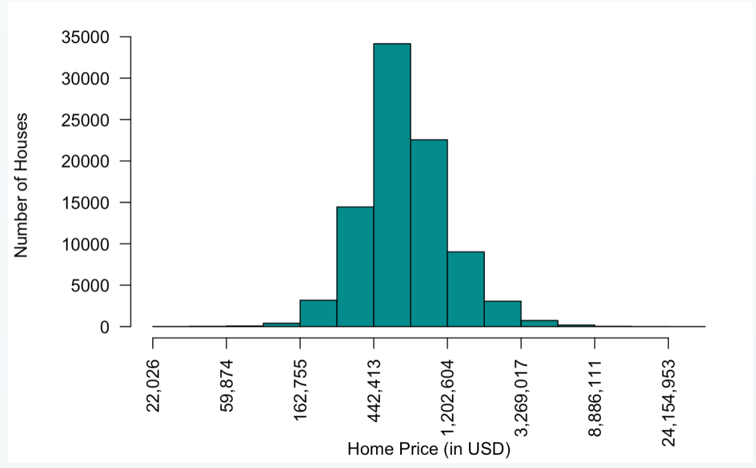
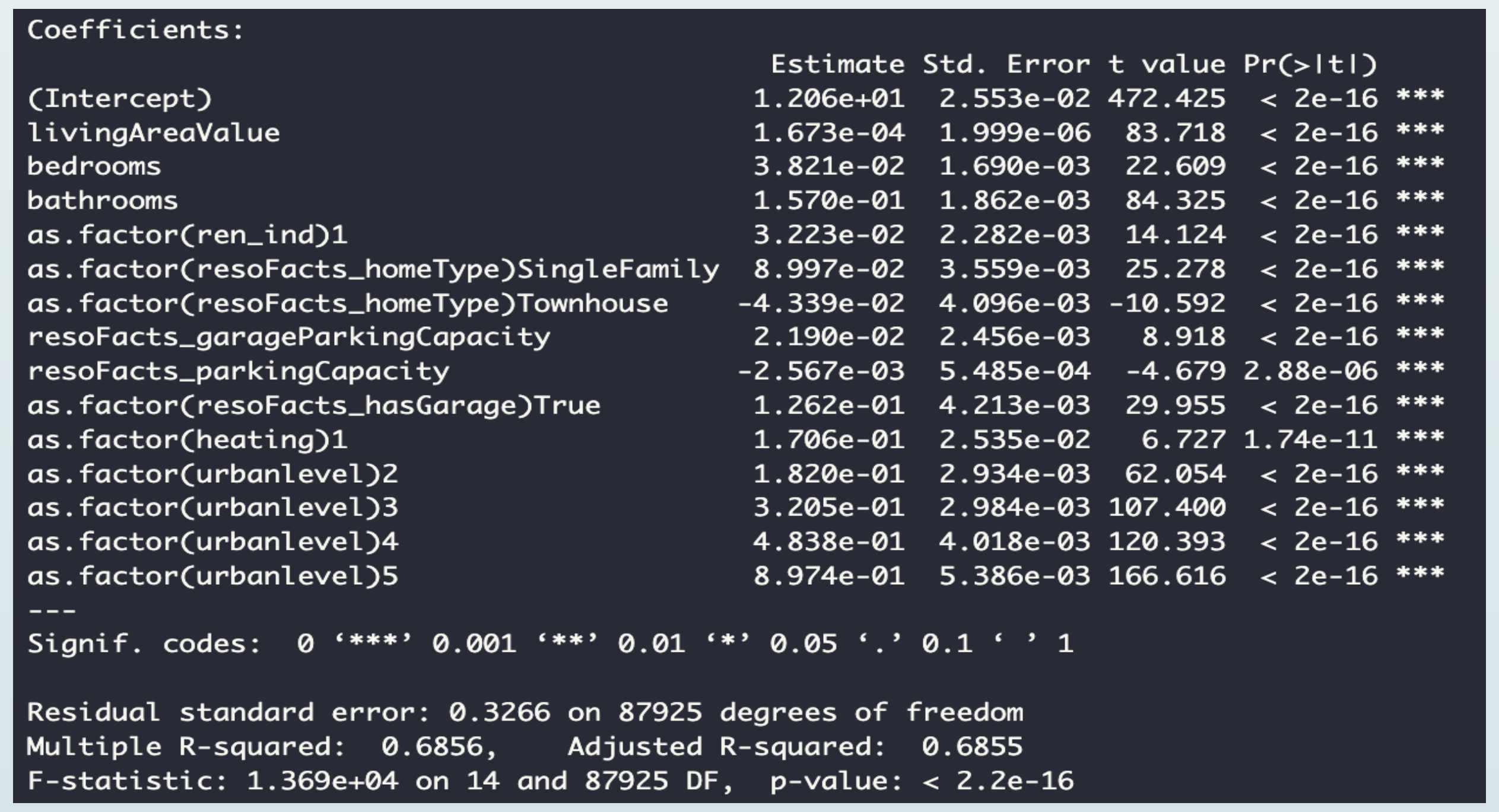


Figure 8: Histogram of home prices after log transformation. Skewness has been reduced.

Model

Predicted Log Sold Price = 12.06 + 0.0002 (Square Footage) + 0.0382 (Bedrooms) + 0.1570 (Bathrooms) + 0.0322 (Is Renovated) + 0.0900 (Is Single Family Home) - 0.0434 (Is Townhouse) + 0.0219 (Garage Parking Capacity) - 0.0026 (General Parking Capacity) + 0.1262 (Has Garage) + 0.1706 (Has Heating) + 0.1820 (Urban Level = 2) + 0.3205 (Urban Level = 3) + 0.4838 (Urban Level = 4) + 0.8974 (Urban Level = 5)

Adjusted R Squared: 68.55%



```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              1.206e+01  2.553e-02 472.425  < 2e-16 ***
livingAreaValue                          1.673e-04  1.999e-06  83.718  < 2e-16 ***
bedrooms                                 3.821e-02  1.690e-03  22.609  < 2e-16 ***
bathrooms                                1.570e-01  1.862e-03  84.325  < 2e-16 ***
as.factor(ren_ind)1                      3.223e-02  2.282e-03  14.124  < 2e-16 ***
as.factor(resoFacts_homeType)SingleFamily 8.997e-02  3.559e-03  25.278  < 2e-16 ***
as.factor(resoFacts_homeType)Townhouse  -4.339e-02  4.096e-03 -10.592  < 2e-16 ***
resoFacts_garageParkingCapacity          2.190e-02  2.456e-03   8.918  < 2e-16 ***
resoFacts_parkingCapacity               -2.567e-03  5.485e-04  -4.679 2.88e-06 ***
as.factor(resoFacts_hasGarage)True       1.262e-01  4.213e-03  29.955  < 2e-16 ***
as.factor(heating)1                      1.706e-01  2.535e-02   6.727 1.74e-11 ***
as.factor(urbanlevel)2                   1.820e-01  2.934e-03  62.054  < 2e-16 ***
as.factor(urbanlevel)3                   3.205e-01  2.984e-03 107.400  < 2e-16 ***
as.factor(urbanlevel)4                   4.838e-01  4.018e-03 120.393  < 2e-16 ***
as.factor(urbanlevel)5                   8.974e-01  5.386e-03 166.616  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3266 on 87925 degrees of freedom
Multiple R-squared:  0.6856,    Adjusted R-squared:  0.6855 
F-statistic: 1.369e+04 on 14 and 87925 DF,  p-value: < 2.2e-16
```

Model Diagnostics

- The residual plot *(see Figure 9)* shows linearity and constant variance only for homes between about $270,000 and $5 million (~12.5 to ~15.5 on the log scale).
- The Q-Q plot shows deviation from normality *(see Figure 10)*.

Figure 9: Residual Plot

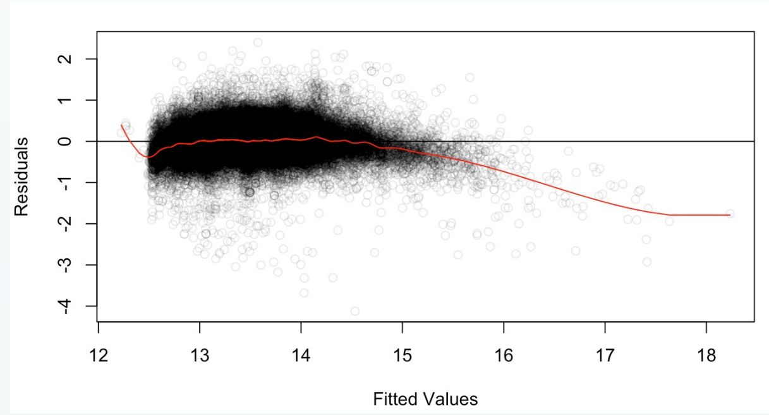
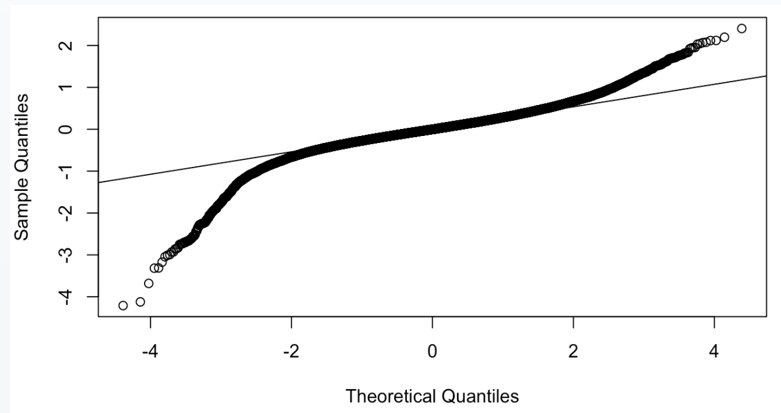


Figure 10: Q-Q Plot



Discussion

Results

- This multiple linear regression model with 14 predictors explains 68.55% of the variation in home prices in the greater Boston area.
- Square footage, bedrooms, bathrooms, home type, renovation status, parking capacity and type, heating and proximity to city center are the most significant home attributes for explaining home price variation.

Considerations

- This model performs best when predicting home prices within the range of $270,000 to $5 million. The model underestimates homes that are priced higher than $5 million.
- The web scraping program can only retrieve a certain number of recently sold homes per zip code, so there may be some zip codes with high turnover rates that do not have transactions going back to 2020.

Acknowledgements

Thanks to Wellesley College's Data Science Program, Professor Anny-Claude Joseph, Professor Cassandra Pattanayak, Professor Eni Mustafaraj, Professor Kyung Park and the CS 350H independent study group.