# Regressin' with Real Estate

**Author names and Institutional Affiliations**

Lana Mai Huynh California Polytechnic State University
Andrew Davis Kerr California Polytechnic State University

## Acknowledgements

We cannot express enough thanks to the individuals that made the dataset we used in our project. We offer our sincere appreciation for those that built the very awesome homes.

Thank you to Dr. Bret Holladay for providing us with the Real Estate Dataset.
Thank you to Discord for providing us the means to communicate and work on this project.


## Dedication

We dedicate this work to Dr. Andrew Schaffner.

## Group Picture

# Table of Contents

## Abstract

How can we predict house prices when there are a myriad of factors that influence this cost? We constructed a linear regression model through trial and error, utilizing the best subsets method. In the end, we determined that the best model included the predictor variables square footage, number of bedrooms, number of bathrooms, age of house, number of garages, and quality of house. The strength of our model's predictive power was confirmed by internal and external validation techniques.


## 1. Introduction

It is often difficult to determine what price to list houses at on the market. There are multiple variables to consider, from square footage and number of bedrooms, to whether or not the house has air conditioning. Based on a previous study conducted by Karabuk University, the variables that have the most impact on the price of a house are the size of the real estate, the distance to the city center, the popularity, and the age of the building. In our report, we will utilize some of the same variables -- square footage and age of house -- as well as new variables -- number of bedrooms, number of bathrooms, number of garages, quality of the house, whether or not the house has air conditioning, and whether or not the house has a pool -- to predict the sale price of the house.

We hope to find a model that accurately predicts the prices of houses so we can better understand how the housing market works when we are ready to become homebuyers. Buying a house is one of the greatest financial milestones in a person's life, and it would be unfortunate to spend more money than necessary. We wish to examine some common key factors that influence the price of a house, so people can know what to look for when the time has come!

## 2. Materials and Methods

The housing data we used for this investigation represents 522 recently sold homes in a city. Each observational unit is a house, with the sale price (thousands of dollars) recorded when the house was sold as the response variable. Our explanatory variables include the interior size of the home (square feet), number of bedrooms, number of bathrooms, number of garages, age of the home at the time of the sale (years), quality of the home (low, medium, high), whether or not the home has air conditioning, and whether or not the home has a pool.

## 3. Data Splitting

Our data set consists of 522 houses, and therefore, 418 houses will be used in the training data and 104 in the testing data. Using the seed number '420', we randomly sampled 418 houses from our data set and saved them as training data, with the remaining houses saved as testing data. In the end, one observation was removed for being unusual, so a total of 521 houses were used with 417 in the training data and 104 in the testing data.

We are splitting the observations so we can evaluate how well our final model predicts on data that it has not seen. Using our model, created from the training data, we will predict the sale prices of the test data and compare the predicted prices to the actual prices using mean squared prediction error (MSPE).

## 4. Data Visualization

### 4.1 Matrix Scatter Plot and Correlation Matrix

As shown in Figure 1, the variable that appears to be the most strongly associated with price is the square footage of the house. This intuitively makes sense

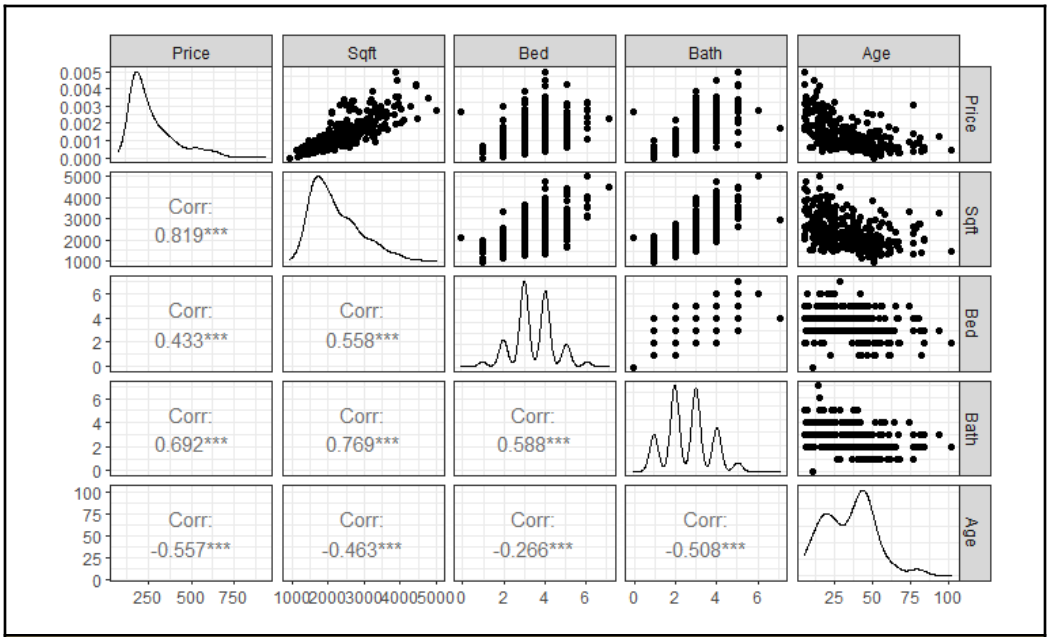


Figure 1. Matrix scatter plot and correlation matrix for Price v. Sqft, Bed, Bath, Age

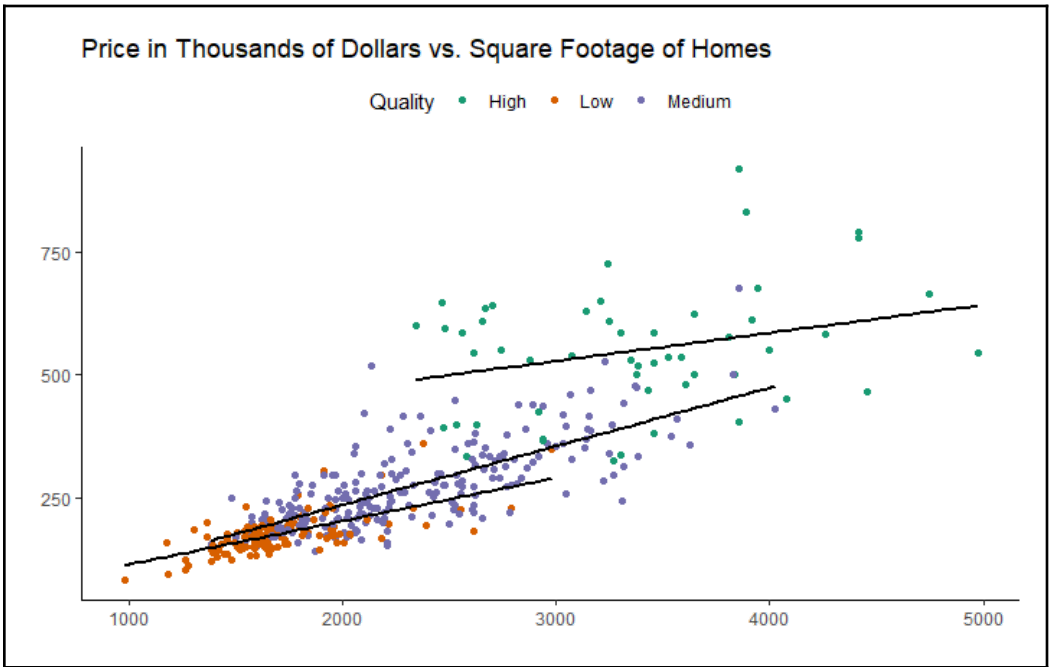


Figure 2. Coded scatter plot of Price v. Sqft by Quality

because the bigger the house is, the more money a buyer should pay. Figure 1 reveals that we will most likely perform transformations on Sqft, and Age.

We noticed that there was an unusual observation in the matrix scatter plot for Bed, and Bath (observation 294). Upon further examination, this observation had 0 baths, 0 beds, and 3 garages. This observation is unusual enough that we decided to omit it from the data.

Figure 1 also reveals a strong, positive, linear relationship between Price and Saft, with a correlation coefficient (r) of 0.819. There is also a moderate, positive, linear relationship between Price and Bed (r = 0.437), but on the other hand a moderate, negative, linear relationship between Price and Age (r = -0.557).

### 4.2 Investigating Interaction

We decided to investigate the interaction between Sqft and Quality on the Price of a house. We chose these predictors because we were interested in determining whether higher quality houses required less square feet to have higher prices.

Based on Figure 2, we suspect that the interaction between Sqft and Quality will be significant. The slopes for each level of Quality differ in steepness. Since all of the slopes are positive, as the square footage of the house increases, the price of the house increases. However, the rate at which the price increases depends on the quality of the house.

## 5. Model Selection

### 5.1 Best Subsets

We used the best subsets technique to select our model. This method fits all possible models, then we compared the summary statistics (e.g. $C_p$, AIC, BIC) and selected three models to check assumptions for using the residual plots.

#### 5.1.1 First Iteration

After running our training data through best subsets, we organized the results in order of ascending Schwarz's Bayesian Information Criterion (BIC), which penalizes larger models.

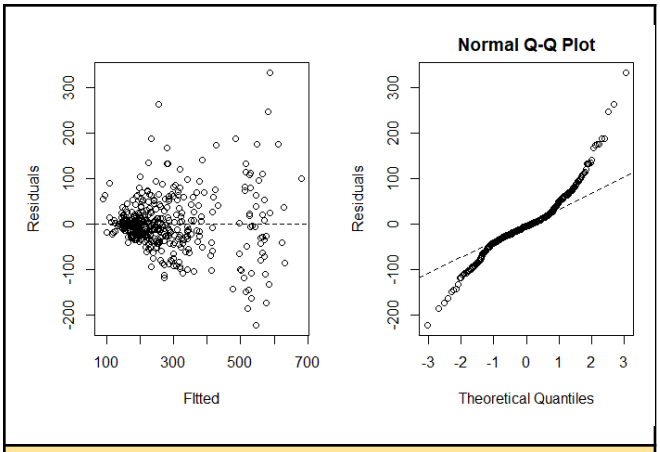


Figure 3. Fitted v. Residuals and Q-Q Plots for model 1 (All residual plots in Appendix B)

We selected the top three models for further investigation. The residual plots of these three models all looked similar, and shown in Figure 3, the residuals v. fitted displays equal variance being violated and the Q-Q Plot normality being violated.

Due to these findings, we decided that we needed to transform Price. In Figure 1, the shape of the Price v. Age suggested decreasing the power; therefore, we applied a natural log transformation to Price.

#### 5.1.2 Second Iteration

The next three models were, once again, selected according to BIC. However, the interaction term between Sqft and Quality was selected, but Sqft itself was not. To be consistent, we decided to keep Sqft in the model.

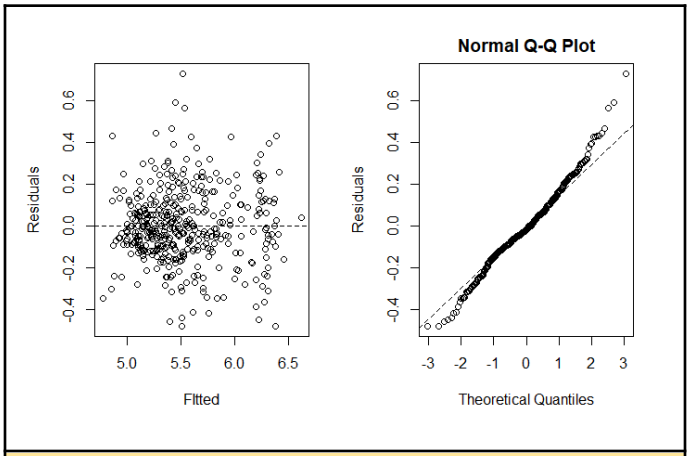


Figure 4. Fitted v. Residuals and Q-Q Plots for model 1 (All residual plots in Appendix B)

Once again all three models had similar residual plots, and Figure 4 reveals that these plots were similar to those in the first iteration. The fitted v. residuals slightly improved, although equal variance still appears violated. The Q-Q Plot has also improved; however, normality is still violated.

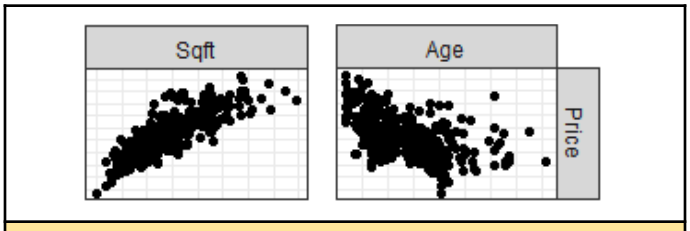


Figure 5. Sections from matrix scatter plot for log(Price) v. Sqft, Bed, Bath, Age (Full matrix scatter plot in Appendix B)

Since equal variance and normality were still violated, we decided to transform some of the predictors. In Figure 5, Sqft and Age both suggested a decrease in power, so we applied a natural log transformation to both variables.

### 5.1.3 Third Iteration

We selected the top two models based off of BIC, as well as the model with the smallest amount of predictors (p), where Mallows $C_p$ ($C_p$) roughly equaled p. Once again, we forced Sqft to be in each model.

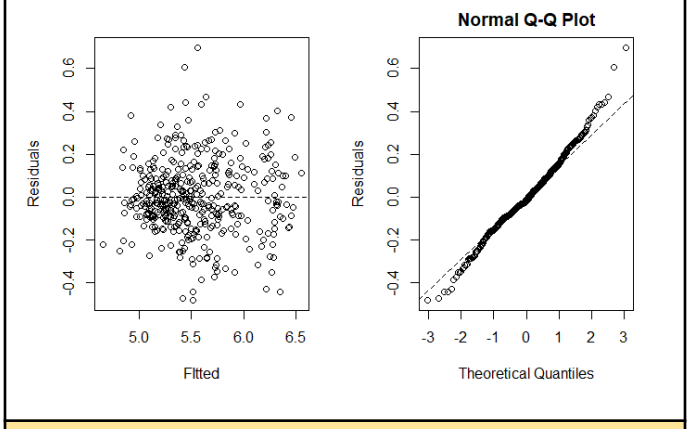


Figure 6. Fitted v. Residuals and Q-Q Plots for model 3 (All residual plots in Appendix B)

Table 1. Breusch-Pagan and Shapiro-Wilk Tests

| | Breusch-Pagan Test p-value | Shapiro-Wilk Test p-value |
|---|---|---|
| Model 1 | 0.064 | 0.0012 |
| Model 2 | 0.073 | 0.00052 |
| Model 3 | 0.062 | 0.0023 |

All of the fitted v. residual plots and Q-Q Plots slightly improved, although there is still a possibility equal variance and normality are violated in each model. We decided to select the model with the best combination of p-values from the Breusch-Pagan (BP) and Shapiro-Wilk tests (SW). Model three ended up having the smallest p-value for BP. With a p-value of 0.06 from BP, equal variance is borderline violated. Therefore, we decided to select the model with the largest p-value for SW, which was model three.

### 5.2 Model Fitting

Our model contextually makes sense. As the size of the house, number of bathrooms, number of bedrooms, and number of garages increases the predicted price of the house does as well. Meanwhile, if the quality is low or medium the price will decrease compared to the reference level of high, which makes sense since higher quality houses should cost more. Furthermore, our interaction terms show that as the size of the house increases, the low and medium qualies reduce the price at slower rates. This makes sense since larger plots houses require larger plots of land, which inherently increases the value of the property.

In Table 2, we see that a high quality house with zero bedrooms, zero bathrooms, zero garages, zero square feet and is zero years old has a predicted price of $e^{5.608}$ = \$272,609. This is the y-intercept for our model; however, this information is not very useful since no house would be zero square feet and this is extrapolation. Furthermore, we see that an increase in one bedroom increases the predicted price by a factor of $e^{0.0262}$ = 1.027, after adjusting for all other variables in the model. Meanwhile, we estimate that the mean price for low quality houses is $e^{5.43}$ = \$228.15 less than high quality houses, after adjusting for all other variables in the mode

Table 3 reveals that our model explains 83.1% of the variability in predicted Price ($R^2$ = 0.831), and the typical prediction error is \$172 (s = 0.172). With a large $R^2$ and small s, we have evidence that the data fits our model.

Table 4 displays the variance inflation factors (VIFs) and generalized variance inflation factors (GVIFs) of our model. For our quantitative variables -- Sqft, Bed, Bath, Age -- we examine the VIFs', while for our categorical variables -- Quality, Garage, Sqft:Quality -- we examine $GVIF^2$. These values are important to check since high VIFs' or $GVIF^2$s' signify multicollinearity, which leads to untrustworthy results from a model. Since all of our predictors have large VIF or $GVIF^2$ values (> 10), except for Garage, our model has severe multicollinearity issues. Thus, our earlier interpretations of slope coefficients should not be trusted because the values are linearly related to each other. However, since the primary purpose of our model is prediction, it should be fine to keep all of the variables in and take extra care not to extrapolate.

## 6. Statistical Inference

### 6.1 Model Utility Test

Based on the small p-value in Table 5, we have strong evidence that at least one of the predictor variables significantly reduces the Sum of Squared Errors (SSE) for $\log(\beta_{\text{Price}})$. Therefore, our model is useful.

| Final Model Regression Equation |
|---|
| log(Price)-hat = 5.608 + 0.07[log(Sqft)] + 0.0262(Bed) + 0.044(Bath) - 0.115[log(Age)] + 0.053(Garage) - 5.43(QualityLow) - 5.95(QualityMedium) + 0.63[log(Sqft):QualityLow] + 0.71[log(Sqft):QualityMedium] |

Table 2. Summary of fitted model

| | **Coefficient** | **Standard Error** | **t-statistic** | **P-value** |
|---|---|---|---|---|
| Intercept | 5.608 | 1.067 | 5.26 | < 0.001 |
| log(Sqft) | 0.07 | 0.13 | 0.49 | 0.622 |
| Bed | 0.0262 | 0.011 | 2.38 | 0.018 |
| Bath | 0.044 | 0.014 | 3.03 | 0.003 |
| log(Age) | -0.115 | 0.019 | -6.00 | < 0.001 |
| Garage | 0.0530 | 0.016 | 3.36 | < 0.001 |
| QualityLow | -5.43 | 1.23 | -4.39 | < 0.001 |
| QualityMedium | -5.95 | 1.11 | -5.35 | < 0.001 |
| log(Sqft):Quality Low | 0.63 | 0.16 | 4.02 | < 0.001 |
| log(Sqft):Quality Medium | 0.71 | 0.14 | 5.11 | < 0.001 |

Table 3. Regression statistics of fitted model

| | |
|---|---|
| Multiple $R^2$ | 0.831 |
| Adjusted $R^2$ | 0.8340 |
| Residual Standard Error | 0.172 |
| Observations | 417 |

Table 4. VIF and $GVIF^2$

| | VIF or $GVIF^2$ |
|---|---|
| log(Sqft) | 21.12 |
| Bed | 16.88 |
| Bath | 33.15 |
| log(Age) | 18.38 |
| Garage | 1.56 |
| Quality | 1764.074 |
| log(Sqft):Quality | 1683.19 |

| Table 5. Model Utility Test | |
|---|---|
| $H_0$ | $\beta_i = 0$ for all i = 1, …, 9 |
| $H_A$ | At least one $\beta_i \neq 0$ |
| F | 243 |
| DF(s) | 9, 407 |
| p-value | < 0.001 |

### 6.2 Single Coefficient Test

In Table 6, the small p-value indicates that we have enough evidence to conclude that $\log(\beta_{Sqft})$ significantly improves the model containing the other 8 variables.

| Table 6. Single Coefficient Test | |
|---|---|
| $H_0$ | $\log(\beta_{Sqft}) = 0$ |
| $H_A$ | $\log(\beta_{Sqft}) \neq 0$ |
| F | 35.101 |
| DF(s) | 8 |
| p-value | < 0.001 |

### 6.3 Interaction

The small p-value in Table 7 implies that we have enough evidence to conclude that the interaction between log(Sqft) and Quality significantly improves the model containing the other 8 variables. This matches what we learned from the coded scatter plot in Section 4.2; the effect of Sqft on predicted Price changes depending on the level of Quality of the house.

| Table 7. Formal Interaction Test | |
|---|---|
| $H_0$ | $\log(\beta_{Sqft})$:Quality = 0 |
| $H_A$ | $\log(\beta_{Sqft})$:Quality $\neq$ 0 |
| F | 13.11 |
| DF(s) | 2 |
| p-value | < 0.001 |

### 6.4 Prediction Inference

For the confidence and prediction intervals, we wanted to predict a price that would correspond to our ideal home. That is, a house that has the following characteristics:

- 2500 square feet
- 4 bedrooms
- 3 bathrooms
- 20 years old
- 2 garages
- “high” quality

#### 6.4.1 Confidence Interval

We are 95% confident that the mean price of our ideal house with the characteristics listed above is between $418,071 to $498,823.

#### 6.4.2 Prediction Interval

We predict with 95% confidence that the price for our ideal house with the characteristics listed above is between $321,802 to $648,057.

## 7. Model Validation

We want to validate our model to make sure its predictive power is still strong. Internal validation allows us to assess the validity of our fitted regression model, meanwhile external validation allows us to know if the predictive ability of our model is acceptable.

### 7.1 Internal Validation

As shown in Table 8, the PRESS and SSE are close to each other and the $R^2_{pred}$ and $R^2$ values are reasonably close. Therefore, our fitted regression model is considered to be internally valid.

Table 8. Testing PRESS and $R^2$ predicted

| | |
|---|---|
| PRESS | 12.851 |
| SSE | 12.082 |
| $R^2_{pred}$ | 0.833 |
| $R^2$ | 0.84 |

### 7.2 External Validation

In Table 9, we see that the MSE and MSPE values are close to each other. Thus, the MSE from the training data is a reasonably valid indicator of the predictive ability of the fitted model.

Table 9. Testing MSPE

| | |
|---|---|
| MSE | 0.03 |
| MSPE | 0.032 |

### 7.3 Combined Data Model

Based on the fitted regression model, utilizing the whole data set, from Appendix B under the "Model Validation" section, we observe that not much has changed from our original regression model. All of the coefficients still have the same signs, and their values have roughly the same values. Additionally, the $R^2$ and s values are approximately the same. Since there is not much difference between our original and final regression model, we can reasonably conclude that our linear regression model is a good predictor of housing prices.

## 8. Conclusion

The linear regression model that we decided was best for predicting included the predictor variables square footage, number of bedrooms, number of bathrooms, age of house, number of garages, and quality of house. This is confirmed by the small p-values of our predictor coefficients, except $\log(\beta_{Sqft})$. However, our model does not pass the Shapiro-Wilk test for normality and barely passes the Breusch-Pagan test for equal variance.

Looking at the coefficient values from our standardized model in Appendix B under "Standardized Model", we notice that the largest magnitude values include QualityLow and QualityMedium. This means that the quality of the house has the greatest impact on the house price. Furthermore, we confirmed that our model is able to accurately predict the prices of homes using both internal and external validation tests.

Our model is easy to use since the predictor variables are straightforward and utilized for real life applications. Additionally, these variables are usually the most important considerations when homeowners are in the process of buying a house. Typically, the number of bedrooms, bathrooms, and square footage are listed on flyers or advertisements of homes, so this information is easy to access. However, we have nine predictor variables, making our model large and overly complicated. Having this many predictors is not necessarily a bad thing; however, in the future we would like to simplify our model. One possible method to mitigate this problem is by finding a dataset with more observations. Another weakness of our model is its multicollinearity. We have to be careful about extrapolation, but if we are making predictions of values within the range of our data, our model is fine to use.

In the future, we would like to make similar models for different real estate datasets. For example, we can use datasets from different cities from all across the world to see how city, state, country, or even content changes our model. Datasets from different cities may also affect what predictors are significant, which would provide insight into what a home buyer is looking for in said city. For instance, we might think that square footage is a more significant predictor in a city like New York compared to rural Texas because homes in New York, on average, have a small square footage; therefore, each increase in square feet would have a more significant impact on the price. On the other hand, since homes in

Texas are typically large, increases in square footage would not make as much of a difference on price.

If we were to do our study again, we would choose a dataset where we have more information about how the data was collected. We chose the dataset that Dr. Holladay has provided us with, but we have no information on the geographical location of the homes, time period of data collection, method of data collection, etc. Knowing where our data comes from would make our interpretations and reports more meaningful.

## Appendix

### Appendix A

Data for 522 recently sold homes in a city.

Variables

- Price: Sale price of the home in thousands of dollars
- Sqft: Interior size of the home in square feet
- Bed: Number of bedrooms
- Bath: Number of bathrooms
- Age: Age of the home in years at time of sale
- Garage; Number of cars that will fit in the garage
- Quality: Quality of the home (high, medium, low)
- AC: Does the house have air conditioning? (yes or no)
- Pool: Does the house have a swimming pool? (yes or no)

### Appendix B

See attached R Markdown file.

## References

Ersoz, F., Ersoz, T., Soydan, M. (2018, December). Research on factors affecting real estate values by Data Mining. Research Gate. Retrieved May 7, 2022, from https://www.researchgate.net/publication/329778823_Research_on_Factors_Affecting_Real_Estate_Values_by_Data_Mining