House Price Regression

Timothy Yeh
Tej Mahale
Liam Reilly
Matthew Yoon
Ali Muhammad Lalani

READ ME: This dataset is different from the outlined in our Project Proposal document that was due last October 4th. Due to a lot of issues that we ran into with the dataset, we talked to Dr. Poliak and she allowed us to change and run our models with this new dataset

Introduction (Timothy)

The house price regression dataset provides a detailed snapshot of housing characteristics, including variables such as `Square_Footage`, `Num_Bedrooms`, `Num_Bathrooms`, `Year_Built`, `Lot_Size`, `Garage_Size`, `Neighborhood_Quality`, and the target variable, `House_Price`. These features represent both structural and locational aspects that are critical in determining property values. Studies have consistently highlighted the relevance of these factors in price-prediction models. For instance, Cheng and Yu (2017) emphasized the role of location and structural features in real estate price modeling, aligning with the dataset's focus on attributes like `Neighborhood_Quality` and `Garage_Size`. By capturing these nuances, the dataset becomes a robust resource for exploring market dynamics and predicting house prices.

The variability in the dataset is another critical strength. House prices range from $111,626 to $1,108,237, while `Square_Footage` spans 503 to 4,999 square feet, reflecting a mix of small, medium, and large properties. A recent report by Shinde et al. (2021) highlighted the importance of such diversity in data for training machine learning models that can generalize across different market segments. Moreover, the inclusion of temporal data, such as `Year_Built`, enables analysis of historical trends in real estate pricing. This aligns with the findings of Shinde et al. (2021), who demonstrated the value of integrating historical and structural data for more accurate predictions. By providing these features, the dataset supports both advanced modeling and traditional econometric analyses.

The dataset's clean structure and comprehensive features make it particularly suitable for machine learning applications, including regression and ensemble models. Cheng and Yu (2017) noted the effectiveness of gradient-boosting techniques in real estate modeling due to their ability to handle complex, non-linear relationships between variables. With features like `Lot_Size`, `Garage_Size`, and `Neighborhood_Quality`, this dataset offers ample opportunities for

researchers to investigate how individual attributes influence overall property value. Furthermore, its application can extend to policy-making and market analysis, providing insights for real estate developers and buyers alike.

The dataset's utility is not limited to straightforward prediction models; it also enables exploratory data analysis (EDA) and hypothesis testing to uncover hidden relationships between features. For example, correlating `Square_Footage` with `House_Price` can reveal the extent to which property size drives value in various neighborhoods. Similarly, examining interactions between `Neighborhood_Quality` and `Year_Built` may uncover trends in urban development and gentrification over time. Shinde et al. (2021) emphasized the role of EDA in identifying non-linear relationships and outlier behavior, which can guide the feature engineering process for machine learning models. By leveraging these insights, researchers can optimize models for predictive accuracy and interpretability, ensuring they are both effective and actionable.

Methodology and Analysis

Principal Components (Tej and Matthew)

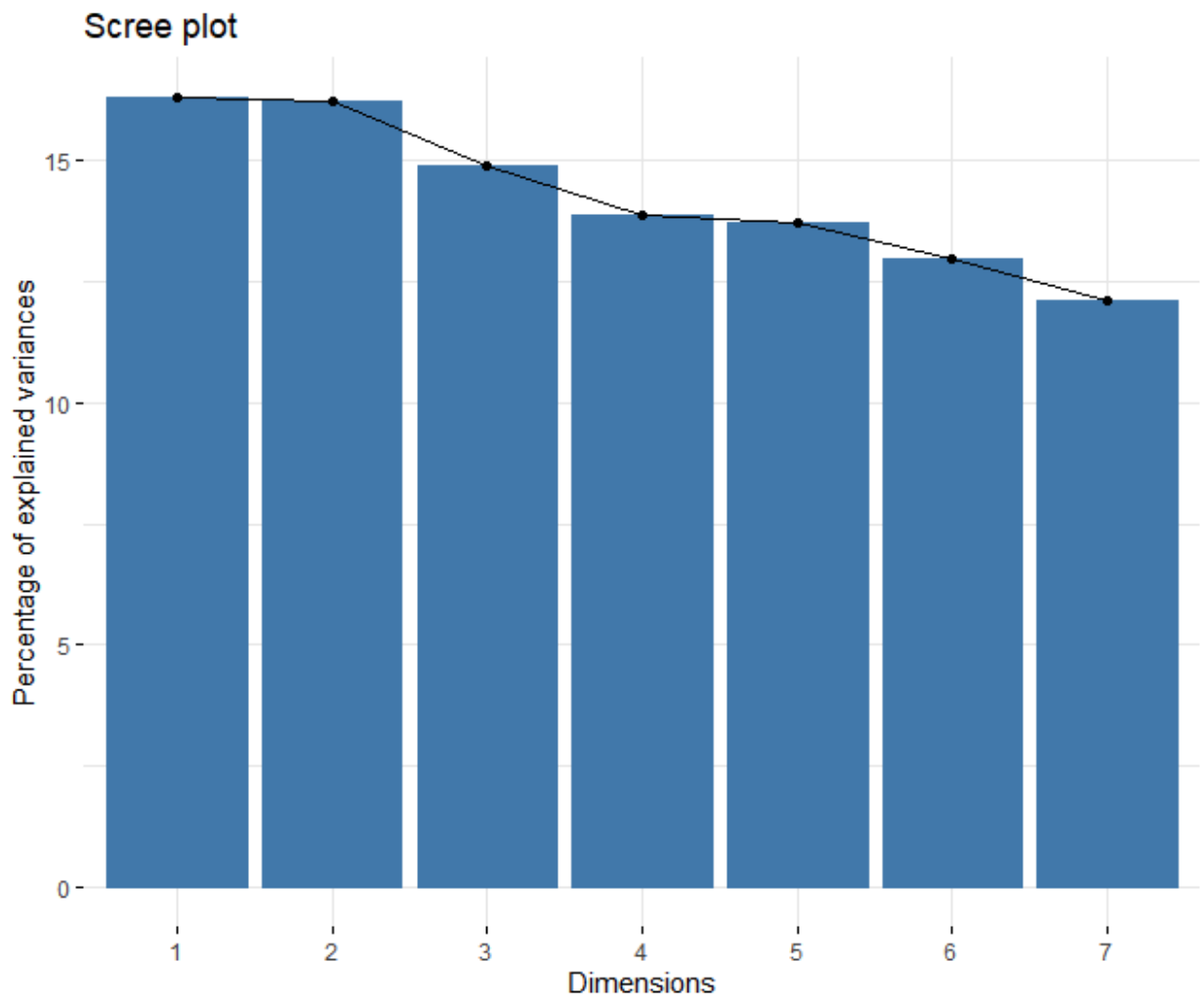
One of the ways we studied the house price is by using Principal Component Analysis and Principal Component Regression.

```
> housing <- house_price_regression_dataset
> pred_vars <- housing[, c("Square_Footage", "Num_Bedrooms", "Num_Bathrooms",
+                          "Year_Built", "Lot_Size", "Garage_Size",
"Neighborhood_Quality")]
> sum(is.na(pred_vars))
[1] 0
```

Loading the dataset and checking if there are any missing or nan values.

```
> scaled_housing <- scale(pred_vars)
> pca_result <- prcomp(scaled_housing, center = TRUE)
> summary(pca_result)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7
Standard deviation     1.0680 1.0653 1.0201 0.9850 0.9801 0.9524 0.9197
Proportion of Variance 0.1629 0.1621 0.1487 0.1386 0.1372 0.1296 0.1208
Cumulative Proportion  0.1629 0.3251 0.4737 0.6123 0.7496 0.8792 1.0000
```

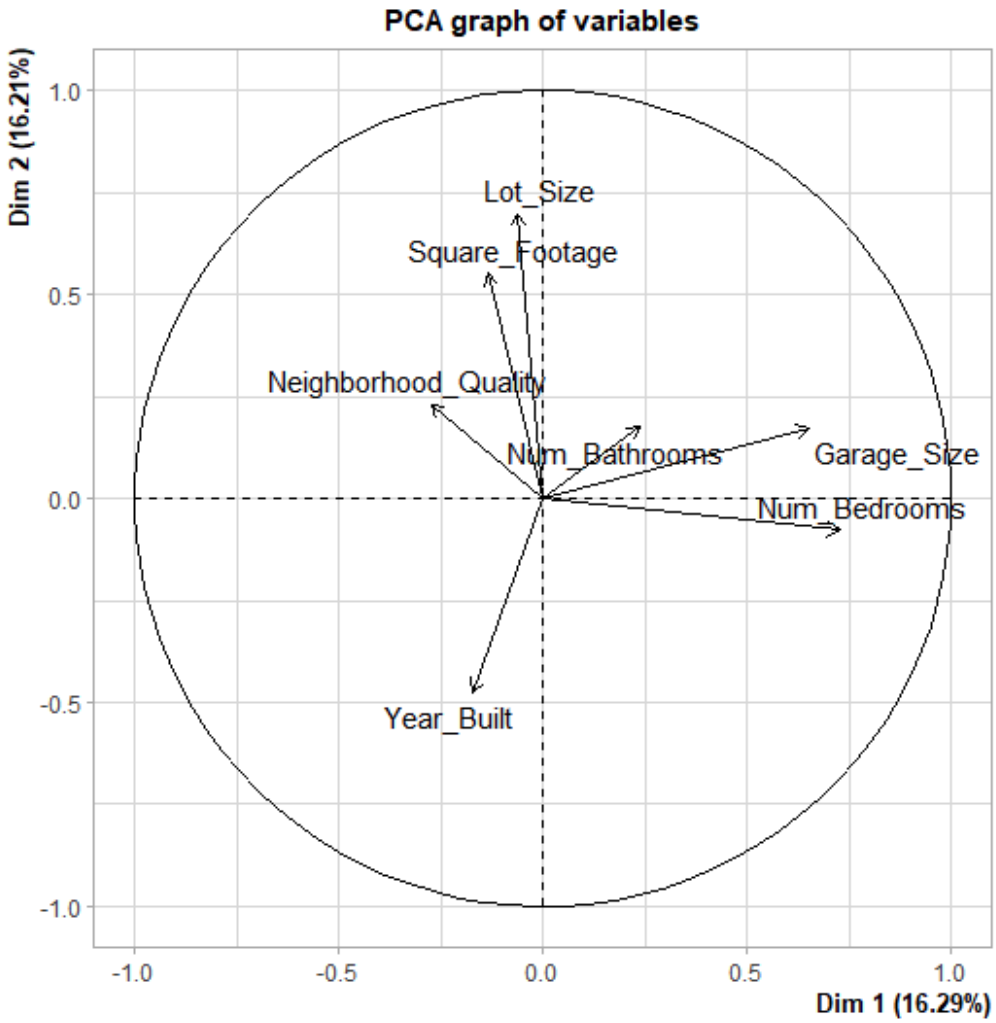
We scaled the predictor variables because the dataset had pretty uneven values when looking at the summary.



Based on the cumulative proportion, using 5 Principal Components explains about 75%

```
> pca_result2 = PCA(pred_vars,scale.unit = TRUE)
> pca_result2$var$cor
                        Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
Square_Footage      -0.13291000  0.55081155 -0.54508024  0.17904828  0.24257399
Num_Bedrooms         0.72803192 -0.07827446 -0.02200324  0.06224725 -0.08061036
Num_Bathrooms        0.23976741  0.17443796  0.65574091 -0.16438866  0.59185036
Year_Built          -0.16997343 -0.47418356 -0.12654039  0.55284941  0.60131752
Lot_Size            -0.06150513  0.69768596  0.03782540 -0.03781763  0.19882506
Garage_Size          0.65421032  0.17384104 -0.13594758  0.43528128 -0.05956700
Neighborhood_Quality -0.27340205  0.23059746  0.52631865  0.64091473 -0.37463914
```

Based on the variation results, Number of Bedrooms and Garage Size are heavily influences on PC1, for PC2 the Square Footage and Lot Size are dominant. For PC3, the Square footage, number of bathrooms are very good with the square footage having negative correlation. PC4, Neighborhood quality is quite strong and for PC5 the Year Built is really strong.



Bi-plot is also shown here to show the dimensions, which are similar to what we observed in the variations before.

After performing regular Principal component of ONLY the predictor variables, we also performed a Principal Component Regression on our response variable: **House_Price**

```
> house.z =
as.data.frame(cbind(house_price_regression_dataset$House_Price,pca_result$x[,1:
5]))
> colnames(house.z) = c("house_price","PC1","PC2","PC3","PC4","PC5")
> house.lm.pca = lm(house_price~.,data=house.z)
> summary(house.lm.pca)

Call:
lm(formula = house_price ~ ., data = house.z)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-370922  -86342     -53   92482  471686

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   618861       3958  156.37  < 2e-16 ***
PC1            21651       3708    5.84 7.08e-09 ***
PC2          -134022       3717  -36.06  < 2e-16 ***
PC3          -132241       3882  -34.07  < 2e-16 ***
PC4            58124       4020   14.46  < 2e-16 ***
PC5           -81095       4040  -20.07  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 125200 on 994 degrees of freedom
Multiple R-squared:  0.7576,   Adjusted R-squared:  0.7564
F-statistic: 621.4 on 5 and 994 DF,  p-value: < 2.2e-16
```

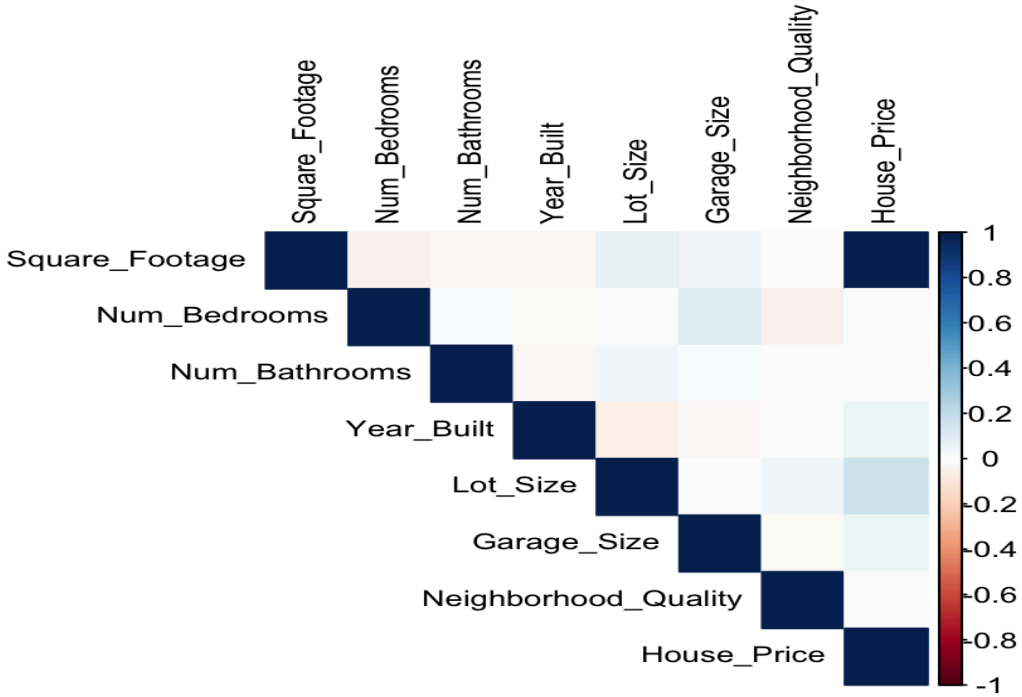
With 5 Principal components, we can see that all of them are actually very significant in predicting the House Price for our dataset. With a P-value of very close to 0 and adj R^2 value of around 75%, explaining 75% of the variation in our model, this is actually a pretty decent model to predict the House Price.

Multivariate Linear Regression (Liam and Ali)

The second way we studied housing prices was using a Multivariate Linear Regression model. After loading the data set, we checked for any null values in the data. Cleaning the dataset here is essential to make sure the rest of our code works properly without any problems.

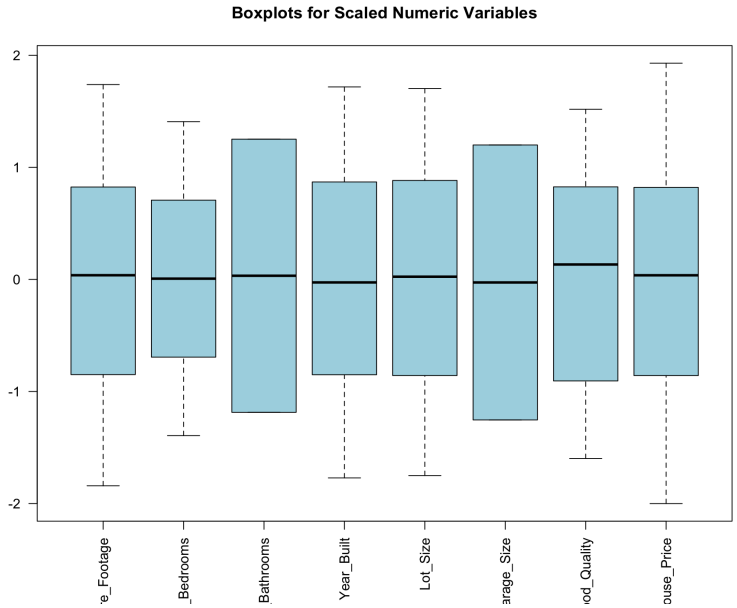
```
# Load the data
data <-
read.csv("/Users/liamreilly/Desktop/house_price_regression_dataset.csv", header = TRUE)
# Check for null values
sum(is.na(data))  # Output should be 0 if there are no missing values
> sum(is.na(data))  # Output should be 0 if there are no missing values
[1] 0
```

Seeing that there were no null values in the data, we then went to see if there were any variables in the data set with high correlation so that we could then remove them to help optimize the model.



As we can see in the correlation matrix, no variables have a high correlation (that is, greater than .7). This is good for our model since we wanted the variables to not be strongly related to each other.

We also created a box plot for the variables to check if there were any outliers. In case there were any outliers we would have to remove them to optimize our model. Before creating the boxplot, we also scaled the data to help with visualizations since the variables had different units.



From the boxplot, we can see that there are no outliers in the data. All the values for all the variables lie within the 1.5*IQR range.

After checking all the necessary conditions we are running our Multivariate Linear Regression Model now, and we can analyze the outputs.

```
> model <- lm(House_Price ~ Square_Footage + Num_Bedrooms +
Num_Bathrooms +
+               Year_Built + Lot_Size + Garage_Size +
Neighborhood_Quality, data = scaled_data)
> summary(model)

Call:
lm(formula = House_Price ~ Square_Footage + Num_Bedrooms +
Num_Bathrooms +
    Year_Built + Lot_Size + Garage_Size + Neighborhood_Quality,
    data = scaled_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.126314 -0.025802  0.000128  0.026569  0.126771

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -5.242e-16  1.222e-03   0.000    1.000
Square_Footage        9.891e-01  1.231e-03 803.789   <2e-16 ***
Num_Bedrooms          5.726e-02  1.234e-03  46.409   <2e-16 ***
Num_Bathrooms         2.667e-02  1.225e-03  21.771   <2e-16 ***
Year_Built            8.068e-02  1.226e-03  65.812   <2e-16 ***
Lot_Size              7.637e-02  1.231e-03  62.013   <2e-16 ***
Garage_Size           1.658e-02  1.232e-03  13.455   <2e-16 ***
Neighborhood_Quality  9.179e-04  1.225e-03   0.749    0.454
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03864 on 992 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9985
F-statistic: 9.543e+04 on 7 and 992 DF,  p-value: < 2.2e-16
```

First of all, looking at the residuals, we can observe that the minimum is **-0.126314** and the maximum is **0.126771**. This is a small range, which tells us that the model fits the data well. Analyzing the estimate column, we can see that square footage has an extremely high correlation with housing price, followed by the number of bedrooms, and the year built. The variables of lot size, garage size, and neighborhood quality show minimal correlation with housing price.

Another way our model tells us that it is a strong fit is that it has an Adjusted R-Squared value of **0.9985** which tells us that approximately 99.85% of the variation can be explained by the predictors that we have selected for our model. The RSE for the model is extremely small, being **0.03864**, which again tells us that the model is accurate. The small F-statistic of **9.543e+04** and p-value < **2.2e-16** tell us that the model is statistically significant. **Square_Footage** has the highest impact on the house prices. It influences the house prices more than any other predictor that we have selected. Other predictors like Num_Bedrooms, Num_Bathrooms, Year_Built, Lot_Size, and Garage_Size have statistically significant coefficients highlighting the significance of their relationship with house prices. However, **Neighborhood_Quality** does not show a significant relationship with house prices ($p = 0.454$). Overall, the information tells us that the model does a good job of indicating the significance of variables that correlate with Housing Prices, and the strength of those variables.

Summary and Conclusion

Based on the results of both Principal Components (Analysis and Regression) as well as Multivariate Linear Regression, the predictor variables we were given in the dataset were all major factors in predicting the house price, with Neighborhood_Quality playing the smallest role in the model. The Principal Component that was heavily influenced by Neighborhood Quality was shown to be significant in the Principal Component Regression, looking more closely at the variation correlation data. There could also be hints of multicollinearity in that 4th Principal Component. More specifically, the year built and garage sizes are also heavily influenced by that 4th principal component.

Of note is that while Cheng and Yu in their research emphasize the significance of location, Neighborhood_Quality actually play the weakest roles in the model, showing us that in spite of a neighborhood's quality, the overall size and structure of the house remain to be the most significant factors in a house's price.

Overall, our findings highlight that our predictor variables are accurate in predicting the house price in some way, with some stronger than others. These variables can be used to further explore the housing market dynamics and shifts in correlation to predicting the price of a house. The models and analysis techniques used here can be valuable for other people, such as stakeholders, buyers, investors, etc. so that it allows them to make informed decisions and identify potential opportunities for a certain house.

References:

1. Cheng, T., & Yu, L. (2017). Data-driven approaches to real estate price modeling. *Journal of Property Research, 34*(1), 1-26.
2. Shinde, M., Patel, R., & Rao, K. (2021). Machine learning in real estate: Predicting property prices using multi-factor analysis. *International Journal of Advanced Computing Research, 15*(3), 45-56.
3. https://www.kaggle.com/datasets/prokshitha/home-value-insights