STATS 520 - Applied Multivariate Analysis

“What does the data on Housing in the United States and London tell us about future global economic trends and housing affordability?”

By Brandon Miner, Diego Angulo Nevarez, Jihyun Do, Xinnan Li

Abstract

This study analyzes the housing markets in multiple U.S. cities and compares them to the London housing market to identify trends and factors influencing home prices and affordability. Using data from Arnav Kulkarni's Kaggle dataset on London housing and Febin Philips' Kaggle dataset on US housing, the analysis applies Principal Component Analysis (PCA) for visualization of multidimensional clustering, and cluster analysis was employed to identify patterns across regions. Null hypothesis Significance tests identified significant differences between U.S. and London markets. The analysis is conducted using R, Python and SAS, utilizing libraries from the "tidyverse" package in R and "sklearn" in Python. Findings reveal that average housing prices in London are significantly higher than in the U.S. (London: $1,864,172.54 vs U.S.: $532,439.91), reflecting different market dynamics.

Introduction

The housing market plays a crucial role in economic stability, personal wealth, and societal well-being. In both the United States and London, property values and rental rates significantly impact individual finances and national economies, making housing costs a central focus of socio-economic research. This study aims to compare housing markets across various U.S. cities with those in London to gain a broader understanding of regional and international housing dynamics. By identifying geographic similarities and differences, the research seeks to identify patterns, pricing trends, and potential socio-economic impacts that influence housing affordability in these distinct markets.

The primary research questions guiding this study include: Which U.S. cities exhibit housing market characteristics similar to London? How do factors such as property size, location, and neighborhood characteristics differently influence housing prices in the U.S. and London? Are there significant differences in average home prices and the factors affecting them between these regions? The study will focus on key variables such as average prices, geographic location, and property size, while adjusting for currency conversion and inflation. Additionally, differences in market indicators between datasets present challenges in achieving direct comparability. The key distinction between our datasets besides the geography is the number and type of variables captured.

This research is based on the hypotheses that home prices in London are generally higher than those in the U.S. and that significant differences in housing market trends exist due to the unique economic and geographic contexts of the two regions. By addressing these questions and hypotheses, the study aims to provide valuable insights for policymakers, investors, and urban planners seeking to better understand and navigate the complexities of housing markets in a global context.

Methodologies

This study analyzes housing markets in the United States and London using data sourced from Kaggle for various demographic and economic factors. Key variables include RegionName, price, income levels, population density, and unemployment rates. These variables provide insights into the factors affecting housing prices and affordability.

Cluster Analysis was employed using k-means with the interest of identifying types of US cities and what type London identifies as. This would give us a unique perspective on how London compares to the United States in terms of housing. Cluster analysis was performed using three datasets: (1) the full numeric U.S. dataset, (2) a subset containing only numeric variables shared between the U.S. and London datasets, and (3) a combined U.S. and London dataset. K-means was chosen due to its computational efficiency and conception cost.

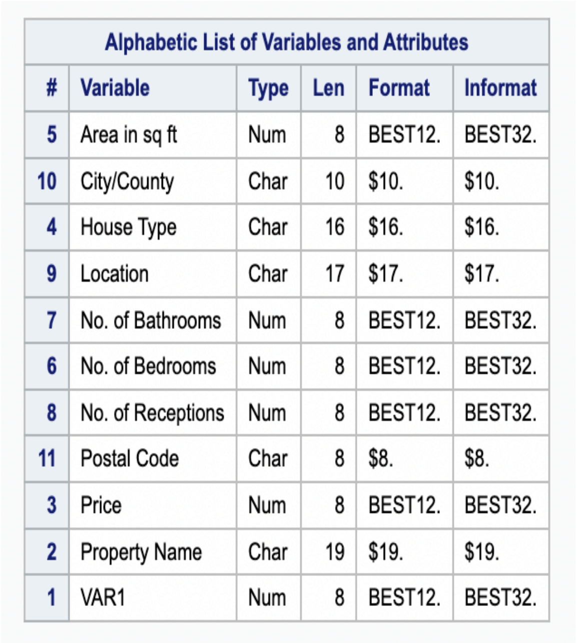
Exploratory Data Analysis (EDA) is performed first using summary statistics and visualizations using histograms and scatter plots to identify trends and outliers in the datasets. Cluster analysis is used to identify cities with similar housing market characteristics and correlations with London housing, while t-tests and ANOVA are conducted to assess statistically significant differences in pricing factors between U.S. cities and London.

Regression models, including multiple linear regression, are employed to predict the influence of factors such as income, neighborhood characteristics, and geographic location on housing prices. Decision tree algorithms using Random Forest (RF) are implemented to classify the London dataset (test set) based on patterns learned from the U.S. housing data (training set). These machine learning models allow for the identification of non-linear relationships and the prediction of housing price trends in both regions.

Statistical analysis and machine learning techniques are implemented using R, Python, and SAS, with libraries like tidyverse for data manipulation and visualization, and stat for statistical testing. This approach provides a comprehensive framework to understand the key economic and geographic factors affecting housing affordability in both the U.S. and London.

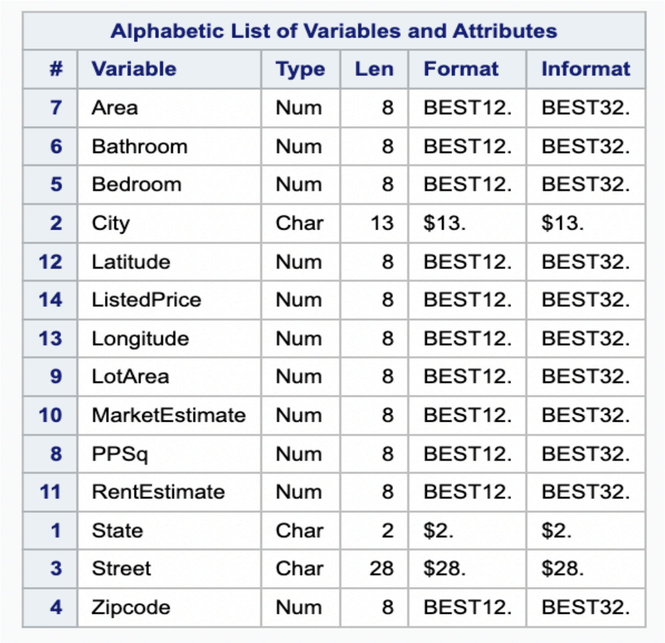
Exploratory Data Analysis (EDA)

To begin the analysis, both the US housing and London housing datasets were imported into the SAS studio under the MYDATA library. This served as the foundation to compare the Area (in square feet) and Price variables, as they share a consistent length across both datasets.



Alphabetic List of Variables and Attributes

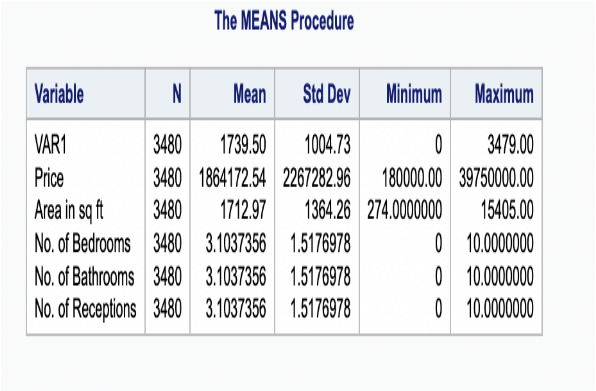
#	Variable	Type	Len	Format	Informat
5	Area in sq ft	Num	8	BEST12.	BEST32.
10	City/County	Char	10	$10.	$10.
4	House Type	Char	16	$16.	$16.
9	Location	Char	17	$17.	$17.
7	No. of Bathrooms	Num	8	BEST12.	BEST32.
6	No. of Bedrooms	Num	8	BEST12.	BEST32.
8	No. of Receptions	Num	8	BEST12.	BEST32.
11	Postal Code	Char	8	$8.	$8.
3	Price	Num	8	BEST12.	BEST32.
2	Property Name	Char	19	$19.	$19.
1	VAR1	Num	8	BEST12.	BEST32.



Alphabetic List of Variables and Attributes

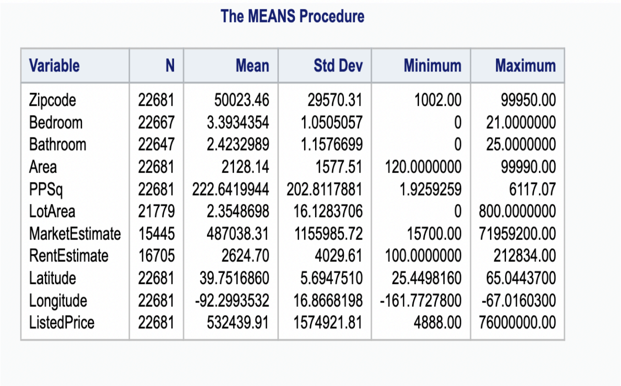
#	Variable	Type	Len	Format	Informat
7	Area	Num	8	BEST12.	BEST32.
6	Bathroom	Num	8	BEST12.	BEST32.
5	Bedroom	Num	8	BEST12.	BEST32.
2	City	Char	13	$13.	$13.
12	Latitude	Num	8	BEST12.	BEST32.
14	ListedPrice	Num	8	BEST12.	BEST32.
13	Longitude	Num	8	BEST12.	BEST32.
9	LotArea	Num	8	BEST12.	BEST32.
10	MarketEstimate	Num	8	BEST12.	BEST32.
8	PPSq	Num	8	BEST12.	BEST32.
11	RentEstimate	Num	8	BEST12.	BEST32.
1	State	Char	2	$2.	$2.
3	Street	Char	28	$28.	$28.
4	Zipcode	Num	8	BEST12.	BEST32.

The Proc Means procedure was utilized to compare various variables between the U.S. and London housing datasets, allowing for a direct comparison of average housing prices in both markets.



The MEANS Procedure

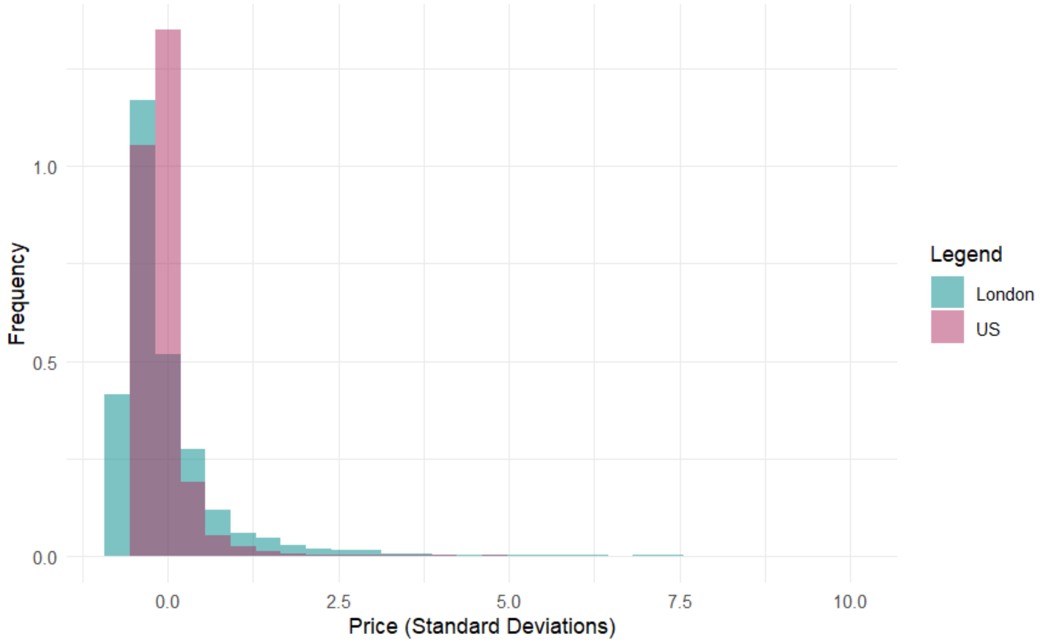
Variable	N	Mean	Std Dev	Minimum	Maximum
VAR1	3480	1739.50	1004.73	0	3479.00
Price	3480	1864172.54	2267282.96	180000.00	39750000.00
Area in sq ft	3480	1712.97	1364.26	274.0000000	15405.00
No. of Bedrooms	3480	3.1037356	1.5176978	0	10.0000000
No. of Bathrooms	3480	3.1037356	1.5176978	0	10.0000000
No. of Receptions	3480	3.1037356	1.5176978	0	10.0000000



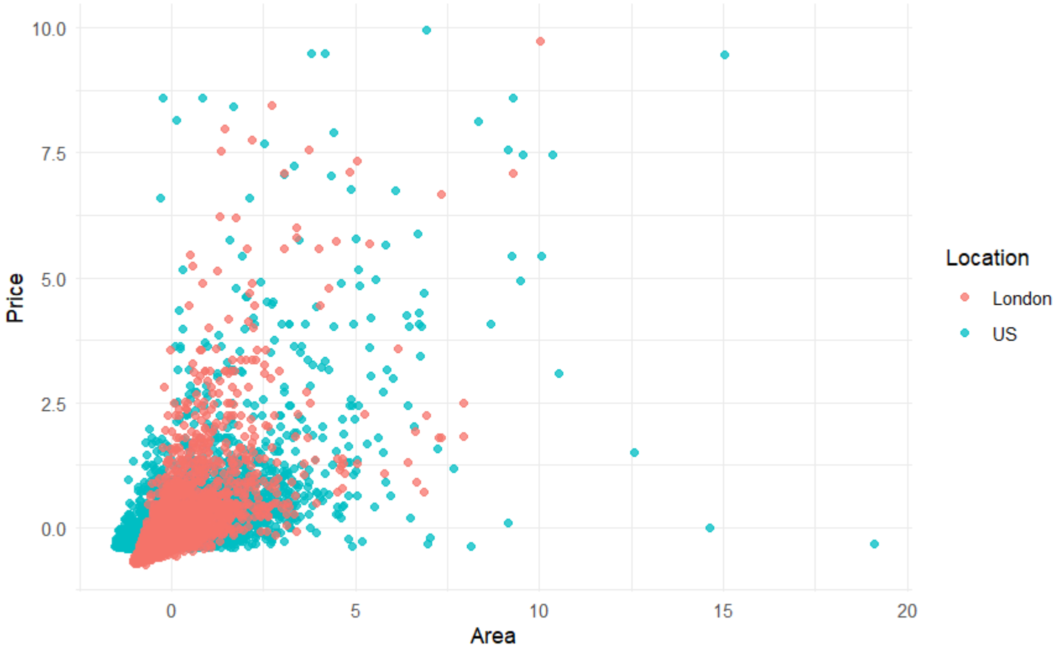
The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Zipcode	22681	50023.46	29570.31	1002.00	99950.00
Bedroom	22667	3.3934354	1.0505057	0	21.0000000
Bathroom	22647	2.4232989	1.1576699	0	25.0000000
Area	22681	2128.14	1577.51	120.0000000	99990.00
PPSq	22681	222.6419944	202.8117881	1.9259259	6117.07
LotArea	21779	2.3548698	16.1283706	0	800.0000000
MarketEstimate	15445	487038.31	1155985.72	15700.00	71959200.00
RentEstimate	16705	2624.70	4029.61	100.0000000	212834.00
Latitude	22681	39.7516860	5.6947510	25.4498160	65.0443700
Longitude	22681	-92.2993532	16.8668198	-161.7727800	-67.0160300
ListedPrice	22681	532439.91	1574921.81	4888.00	76000000.00

We used a histogram to further understand the distribution of housing prices.



From the histogram, we see that the data is right skewed and unimodal. It is important to note that London's peak being less than the US's does not directly correlate with lower prices. This is because the histogram is in terms of standard deviations. The mean price of London housing is larger than that of US housing. The most interesting variable to compare to price is area.

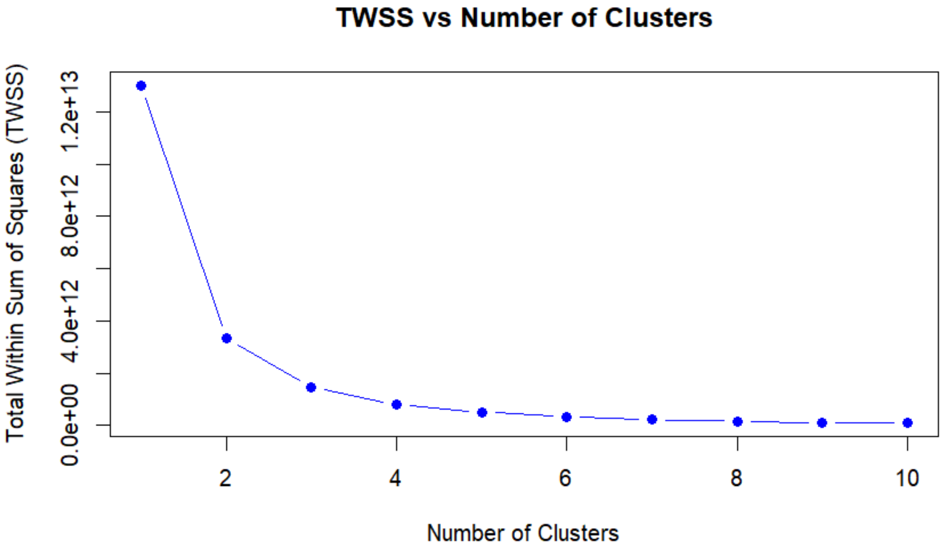


We can see US housing has a larger spread of pricing vs area. This makes sense given the large variety of houses in the United states from large to small cities of varied economic climates.

Cluster Analysis

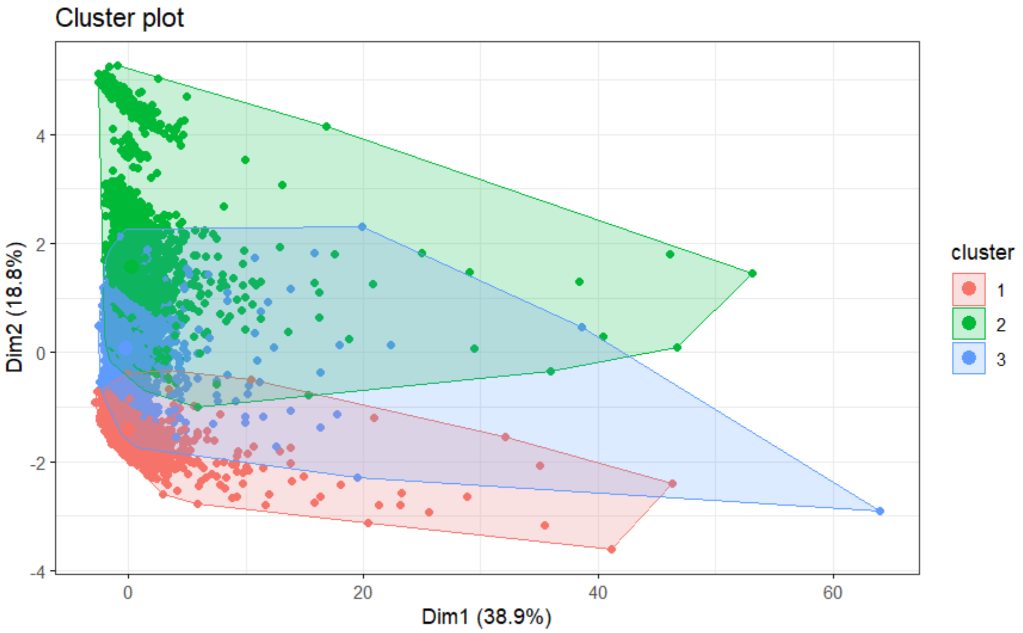
One approach taken was to run a statistical clustering of the US housing data and compare it to the London housing data. There were 3 different datasets that were clustered: The US dataset with all numeric data types, the US dataset with only numeric variables shared with the London dataset, and the US & London datasets combined.

The method used was k-means. The first dataset clustering served as a basis for understanding our dataset. We started by identifying the best number of clusters using a Total Within Sum of Squares vs number of clusters plot.

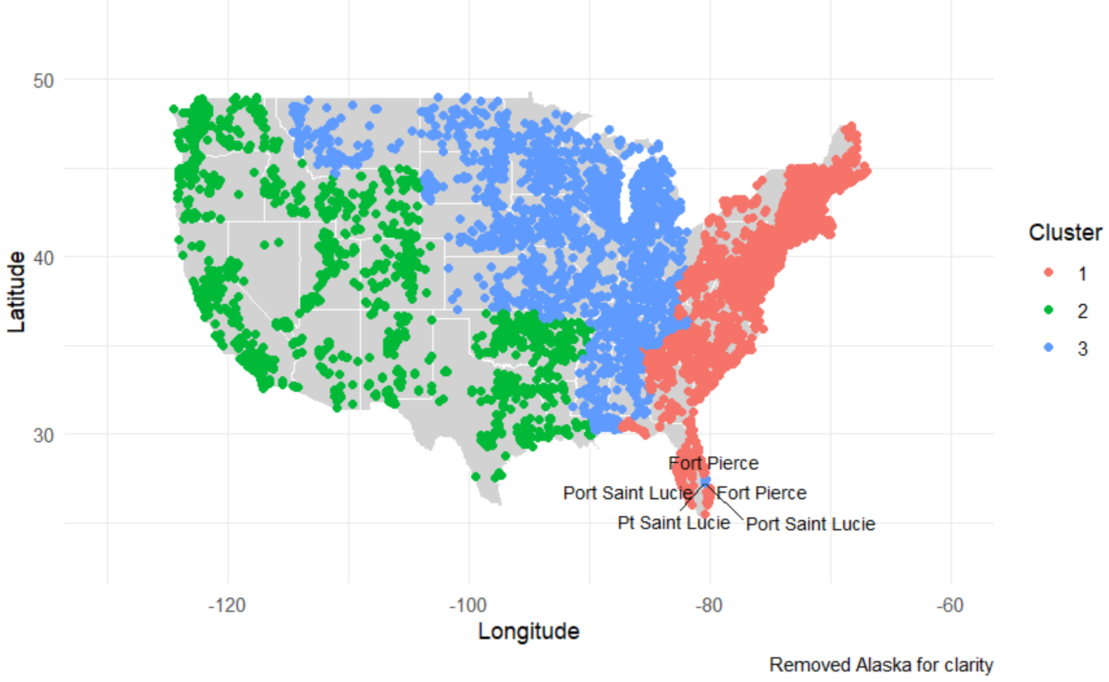


The optimal number of clusters was determined using the Total Within Sum of Squares (TWSS) plot, where the ‘elbow’ point indicated that three clusters best captured the data’s variability. An important characteristic of the graph is the extremely large TWSS. The range is on a scale of 10^12. This will be explored later.

The clustering of the US dataset with 3 clusters is visualized below in a cluster plot. This plot uses PCA to restructure the variables into principal components to plot our multivariate data on a two-dimensional plot using the first two principal components.

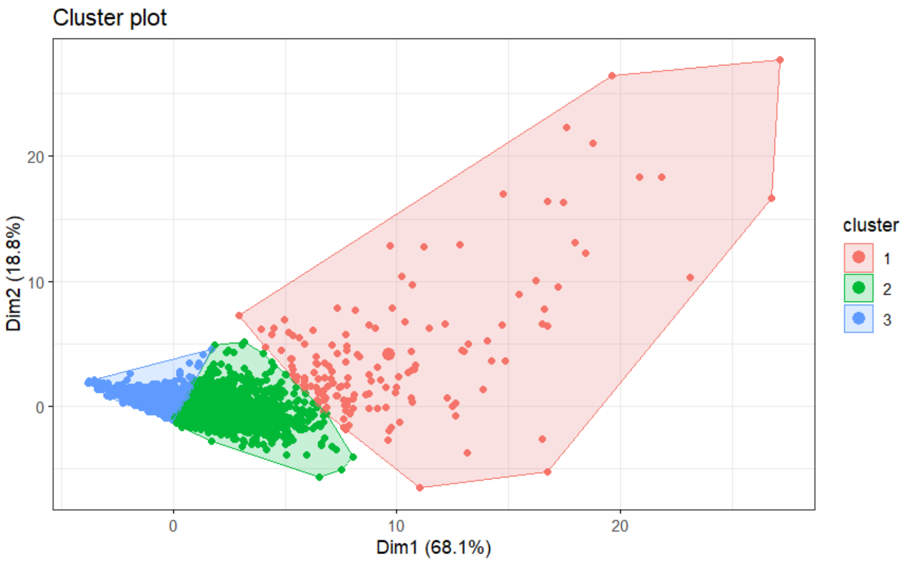


Because we have Longitude and Latitude as variables in the clustering, we can actually plot the data on a map of the United States.



There are two characteristics of note. The first is that the points for Alaska were removed for clarity. They were all in the second cluster. The second characteristic is the outlying points in Fort Pierce and Port Saint Lucie in Florida. These points are clustered into the third cluster whereas the eastern coast is in the first cluster.

The second clustering done was on the US and London datasets together.



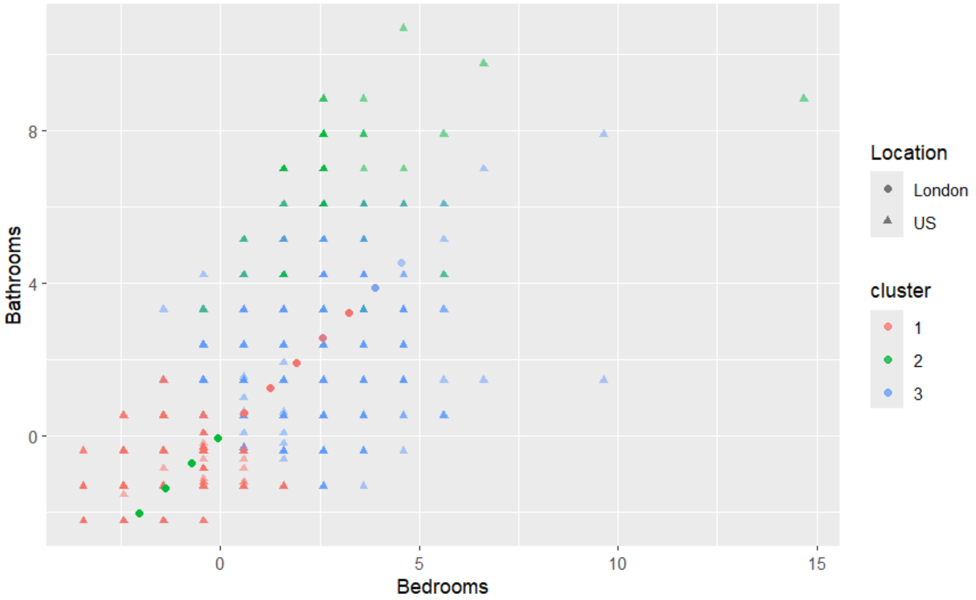
We can actually see that the clusters are much more distinct than the previous cluster plot. Theory as to why is that there is more data(rows) and less variables which leads to much more refined principal components. If the first cluster plot were plotted in three dimensions, it might look more distinct.

Lastly, using the cluster of the US housing data with only variables which are shared between datasets, and projecting the London housing dataset onto the clusters we found the following results.

Cluster 1	36.638%
Cluster 2	62.989%
Cluster 3	0.374%

The results indicate that London housing data predominantly maps to Clusters 1 and 2 of the U.S. dataset. Due to the strong influence of geographic location in clustering, the comparison between London and U.S. cities was limited. Future work could consider aggregating data by city to improve interpretability. Another limitation identified is that every city has a distribution of types of houses. This makes identifying similar cities difficult because they are so varied within their own housing economies.

The distinction between clusters is not distinct across two dimensions. The clearest example is across Bathrooms vs Bedrooms.



The plot shows that the distribution is strictly linear for our London dataset. We can then surmise that the 62% of London data points are closer than -5 standard deviations away from the mean or at the mean.

Regression Analysis

To examine the relationship between price and square footage of a lot and three variables of a property: number of bathrooms, square footage of living area, and square footage above, we utilized multivariate regression. In preliminary analysis, we found that our response variables, price and square footage of the lot were highly correlated with one another. By using multivariate regression, we are able to account for the correlation among our two outcome variables and conduct meaningful multivariate tests. Our multivariate regression model in matrix form can be seen in Figure 1 below.

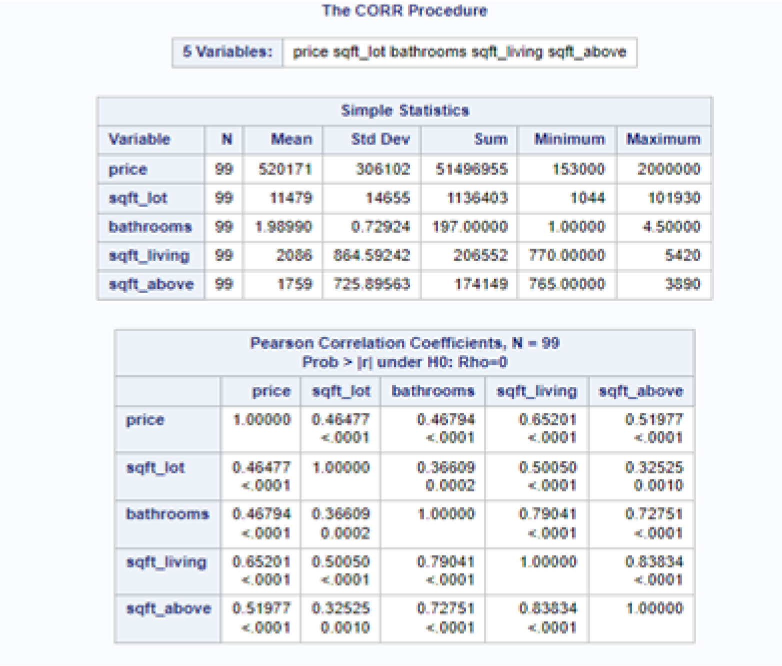
$n = 99$ observations
$p = 3$ predictors (Bathrooms, Sqft of Living Area, Sqft of Above)
$k = 2$ response variables (Price, Sqft of Lot)

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} Y_{1,1} & Y_{1,2} \\ \vdots & \vdots \\ Y_{99,1} & Y_{99,2} \end{bmatrix}_{99 \times 2} \quad X = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,3} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{99,1} & \cdots & X_{99,2} \end{bmatrix}_{99\times(3+1)} \quad \beta = \begin{bmatrix} \beta_{01} & \beta_{02} \\ \vdots & \vdots \\ \beta_{3,1} & \beta_{3,2} \end{bmatrix}_{(3+1)\times 2} \quad \varepsilon = \begin{bmatrix} \varepsilon_{1,1} & \varepsilon_{1,2} \\ \vdots & \vdots \\ \varepsilon_{99,1} & \varepsilon_{99,2} \end{bmatrix}_{99 \times 2}$$

Statistics Descriptive:

From the following tables, we can obtain a summary of the statistics for the two response variables and three predictors. It is evident that we can derive basic information such as mean, standard deviation, etc, for each variable. For instance, the mean value for each house is $520,171, and the maximum value for a house is $2 million. Additionally, from the second table, we observe a strong correlation between ‘price’ and ‘sqft_lot’, as well as those two response variables and the other three predictors. We hypothesize the existence of a multilinear relationship between these two response variables and the other three predictors. We will validate our hypothesis in the next section.



The CORR Procedure

5 Variables:	price sqft_lot bathrooms sqft_living sqft_above

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
price	99	520171	306102	51496955	153000	2000000
sqft_lot	99	11479	14655	1136403	1044	101930
bathrooms	99	1.98990	0.72924	197.00000	1.00000	4.50000
sqft_living	99	2086	864.59242	206552	770.00000	5420
sqft_above	99	1759	725.89563	174149	765.00000	3890

Pearson Correlation Coefficients, N = 99
Prob > |r| under H0: Rho=0

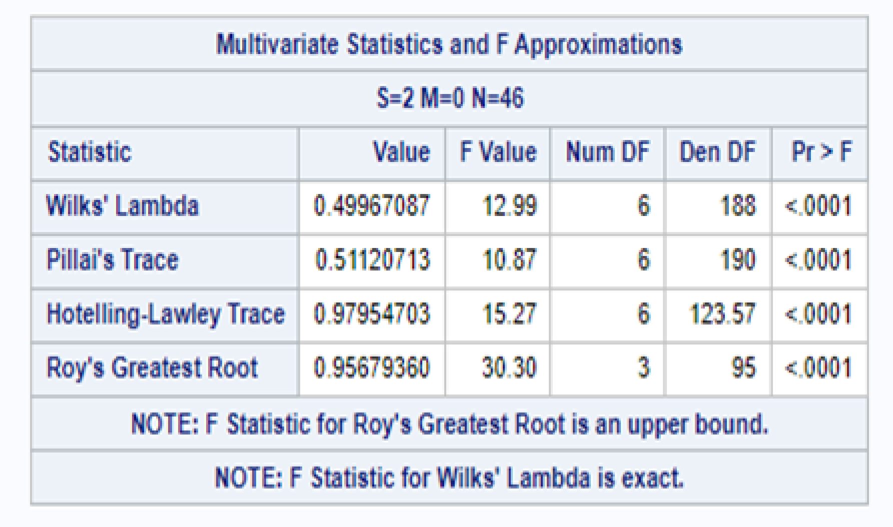
	price	sqft_lot	bathrooms	sqft_living	sqft_above
price	1.00000	0.46477 <.0001	0.46794 <.0001	0.65201 <.0001	0.51977 <.0001
sqft_lot	0.46477 <.0001	1.00000	0.36609 0.0002	0.50050 <.0001	0.32525 0.0010
bathrooms	0.46794 <.0001	0.36609 0.0002	1.00000	0.79041 <.0001	0.72751 <.0001
sqft_living	0.65201 <.0001	0.50050 <.0001	0.79041 <.0001	1.00000	0.83834 <.0001
sqft_above	0.51977 <.0001	0.32525 0.0010	0.72751 <.0001	0.83834 <.0001	1.00000

Testing Linearity:

Since we are testing whether there exists a multivariate linear relationship between those two response variables and three predictors, our null hypothesis is β1 = β2 = β3 =0, and the alternative hypothesis is that at least one of the coefficients is not equal to 0. From the following table, we observe that Wilk's Lambda is around 0.5. This means that approximately 50% of the variance can be explained by those three predictors. Additionally, since the p-value is so small, we can reject the null hypothesis. This suggests that there exists a multilinear relationship between the two response variables and three predictors.

Coefficients Explanation:

From the following table, we observe the four coefficients: β0 = 75472, β1 =-48304, β2 =282.38, β3 =-27.47, respectively. If we assume the predictors 'bathrooms', 'sqft_above' remain constant, this suggests that for every single one-square-foot increase in the predictor 'sqft_living', the response variable 'price' will increase by $282,38. The formula to calculate the estimated price for a house is 'price' = 75472 + (-48304)* 'bathrooms' + 282.38 * 'sqft_living' + (-27.47) * 'sqft_above'.



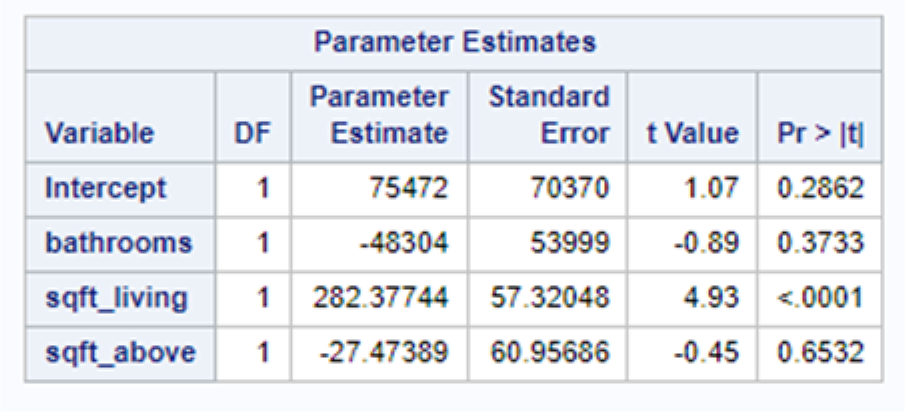
Multivariate Statistics and F Approximations

S=2 M=0 N=46

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.49967087	12.99	6	188	<.0001
Pillai's Trace	0.51120713	10.87	6	190	<.0001
Hotelling-Lawley Trace	0.97954703	15.27	6	123.57	<.0001
Roy's Greatest Root	0.95679360	30.30	3	95	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

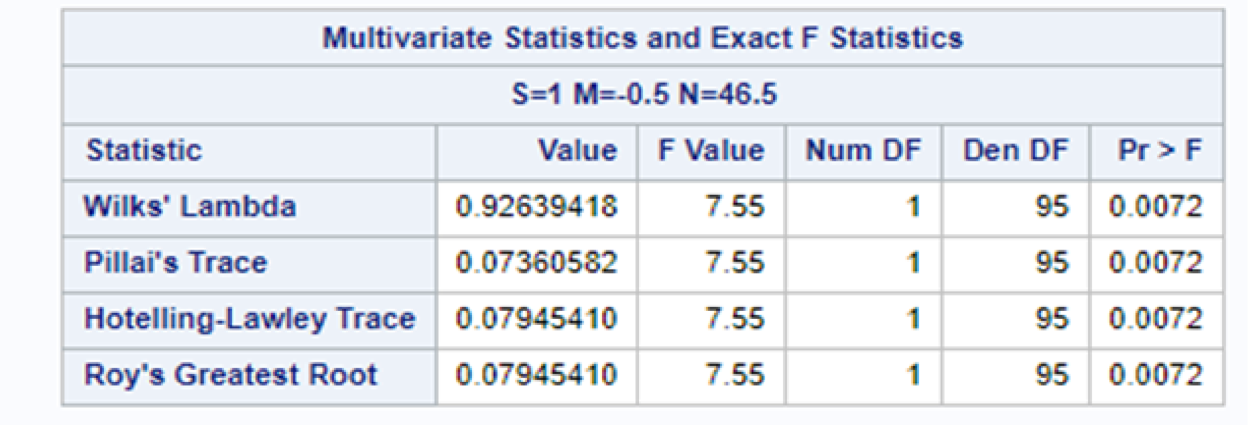
NOTE: F Statistic for Wilks' Lambda is exact.



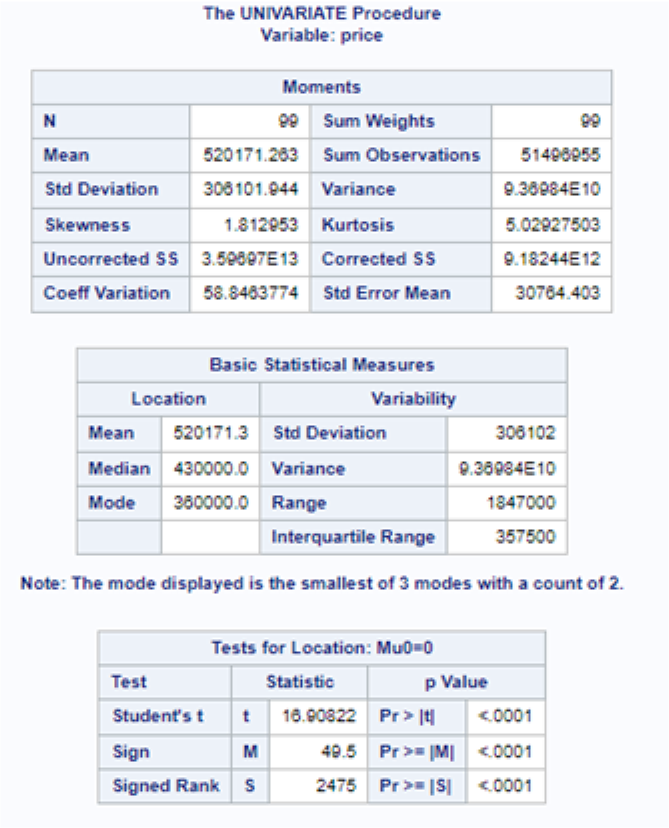
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	75472	70370	1.07	0.2862
bathrooms	1	-48304	53999	-0.89	0.3733
sqft_living	1	282.37744	57.32048	4.93	<.0001
sqft_above	1	-27.47389	60.95686	-0.45	0.6532

Additional Test:

We are particularly interested in whether the difference between ‘sqft_living’ and ‘sqft_above’ would be equal among the two response variables, ‘price’ and ‘sqft_lot’. Our null hypothesis is: β21 - β22= β31 – β32. After a minor adjustment to the alternative hypothesis, the alternative hypothesis is now: (β21 – β31)- (β22 – β32) is not equal to 0. To test this null hypothesis, the necessary matrices L and M are [0 0 1 -1] and [1 -1], respectively. By running the SAS code, we can obtain the following table as the result. According to Wilks' Lambda, the p-value is very small, so we can reject the null hypothesis. This suggests that the difference between ‘sqft_living’ and ‘sqft_above’ regarding the ‘price’ response variable is different from the difference between ‘sqft_living’ and ‘sqft_above’ regarding the response variable ‘sqft_lot’.



Multivariate Statistics and Exact F Statistics					
S=1 M=-0.5 N=46.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.92639418	7.55	1	95	0.0072
Pillai's Trace	0.07360582	7.55	1	95	0.0072
Hotelling-Lawley Trace	0.07945410	7.55	1	95	0.0072
Roy's Greatest Root	0.07945410	7.55	1	95	0.0072



The UNIVARIATE Procedure
Variable: price

Moments			
N	99	Sum Weights	99
Mean	520171.263	Sum Observations	51496955
Std Deviation	306101.944	Variance	9.36984E10
Skewness	1.812953	Kurtosis	5.02927503
Uncorrected SS	3.59697E13	Corrected SS	9.18244E12
Coeff Variation	58.8463774	Std Error Mean	30764.403

Basic Statistical Measures			
Location		Variability	
Mean	520171.3	Std Deviation	306102
Median	430000.0	Variance	9.36984E10
Mode	360000.0	Range	1847000
		Interquartile Range	357500

Note: The mode displayed is the smallest of 3 modes with a count of 2.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	16.90822	Pr > |t|	<.0001
Sign	M	49.5	Pr >= |M|	<.0001
Signed Rank	S	2475	Pr >= |S|	<.0001

Normality Tests:

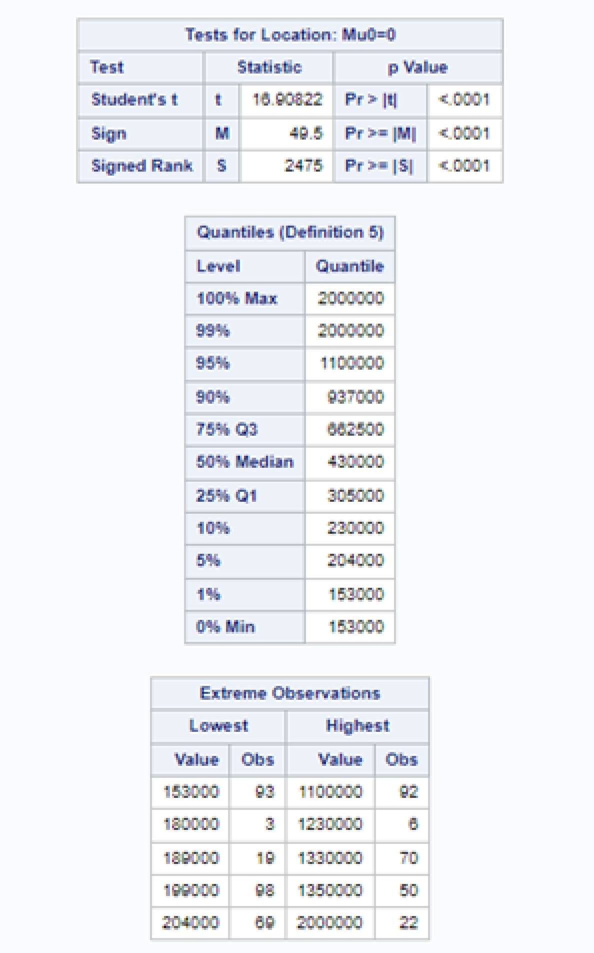
The normality of the price distribution was rigorously tested using several methods:

Student's t-test: The test statistic of 16.0822 with a p-value <0.0001 strongly rejects the null hypothesis of the mean being zero, indicating significant variation from normal expectations.

Sign and Signed Rank Tests: Both tests showed extremely small p-values (<0.0001), which also reject the hypothesis of median equality to zero, further affirming the data's deviation from normality.

These statistical tests, alongside the calculated skewness, suggest that the distribution of housing prices in the dataset is not normal. This conclusion is critical as it affects the choice of statistical methods for further analysis. For instance, non-parametric methods might be better suited for analyzing this data due to the lack of normality.

The histogram shows the distribution of housing prices, overlaid with a theoretical normal distribution curve. The data has been summarized into a histogram with the normal curve fitted using the calculated mean (μ = $520,171.3) and standard deviation (σ = $306,102).



Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	16.90822	Pr > |t|	<.0001
Sign	M	49.5	Pr >= |M|	<.0001
Signed Rank	S	2475	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	2000000
99%	2000000
95%	1100000
90%	937000
75% Q3	662500
50% Median	430000
25% Q1	305000
10%	230000
5%	204000
1%	153000
0% Min	153000

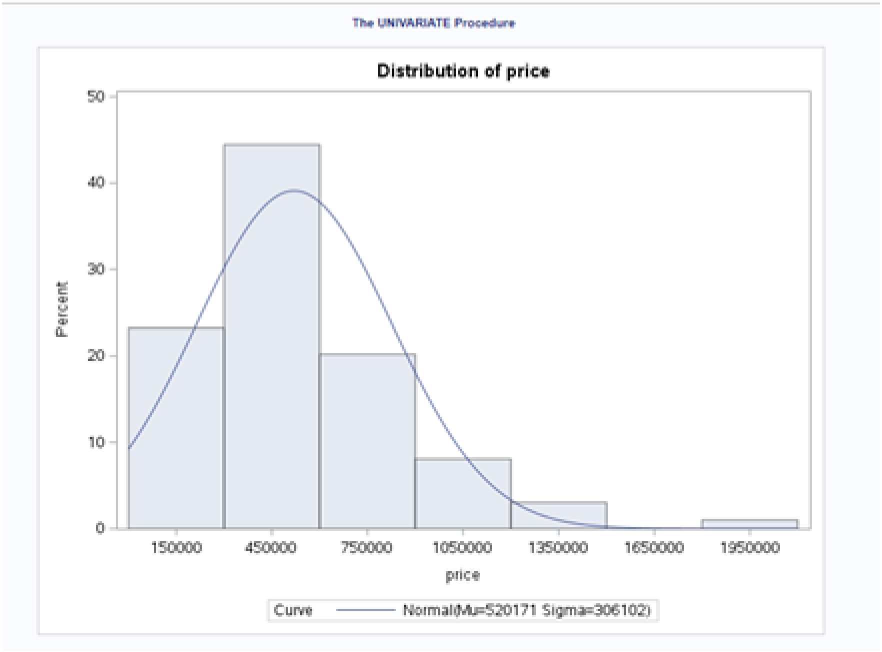
Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
153000	93	1100000	92
180000	3	1230000	6
189000	19	1330000	70
199000	98	1350000	50
204000	69	2000000	22

Analysis of the Distribution

The histogram illustrates that the distribution of housing prices is right-skewed, with a majority of data points lying below the mean and a long tail extending towards the higher prices. This is visually evident as the bars representing the frequency of prices are more concentrated on the left side of the mean and decrease in height as the price increases.

The normal curve, superimposed on the histogram, does not align well with the distribution of the observed data, especially in the tails. The peak of the observed distribution is to the left of the mean, and there are several outliers or extreme values on the right, which the normal curve fails to encapsulate effectively. This discrepancy highlights the presence of skewness and kurtosis beyond what is expected in a normal distribution.

The histogram shows the distribution of housing prices, overlaid with a theoretical normal distribution curve. The data has been summarized into a histogram with the normal curve fitted using the calculated mean (μ = $520,171.3) and standard deviation (σ = $306,102).



Random Forest

In addition to Regression Models and Cluster analysis, a Random Forest regression model was employed to predict housing prices and evaluate the importance of various predictors for the US housing data by Phillips, F. This machine learning technique is particularly effective for capturing complex non-linear relationships and identifying key drivers of price variability.

The Random Forest model was trained to predict U.S. housing prices (`ListedPrice`) based on 13 independent variables, including geographic features (State, City, Zipcode, Latitude, Longitude), property characteristics (Bedroom, Bathroom, Area, PPSq, LotArea), and market-related estimates (MarketEstimate, RentEstimate). The dataset consisted of 11,882 samples for training, and the model used 500 trees with 4 variables considered at each split.

Model Performance:

Mean Squared Error (MSE): The MSE of approximately **$43.93 billion** reflects the average squared difference between the predicted and actual housing prices during model validation.

Root Mean Squared Error (RMSE): The model yielded an RMSE of **$336,698.10** representing the average magnitude of error in predicting housing prices. On average, the model's predictions deviate by around $336,698 from actual housing prices.

Explained Variance (% Var explained): The model achieved a **96.1%** of the variance in housing prices that is explained by the model during training validation, demonstrating high accuracy.

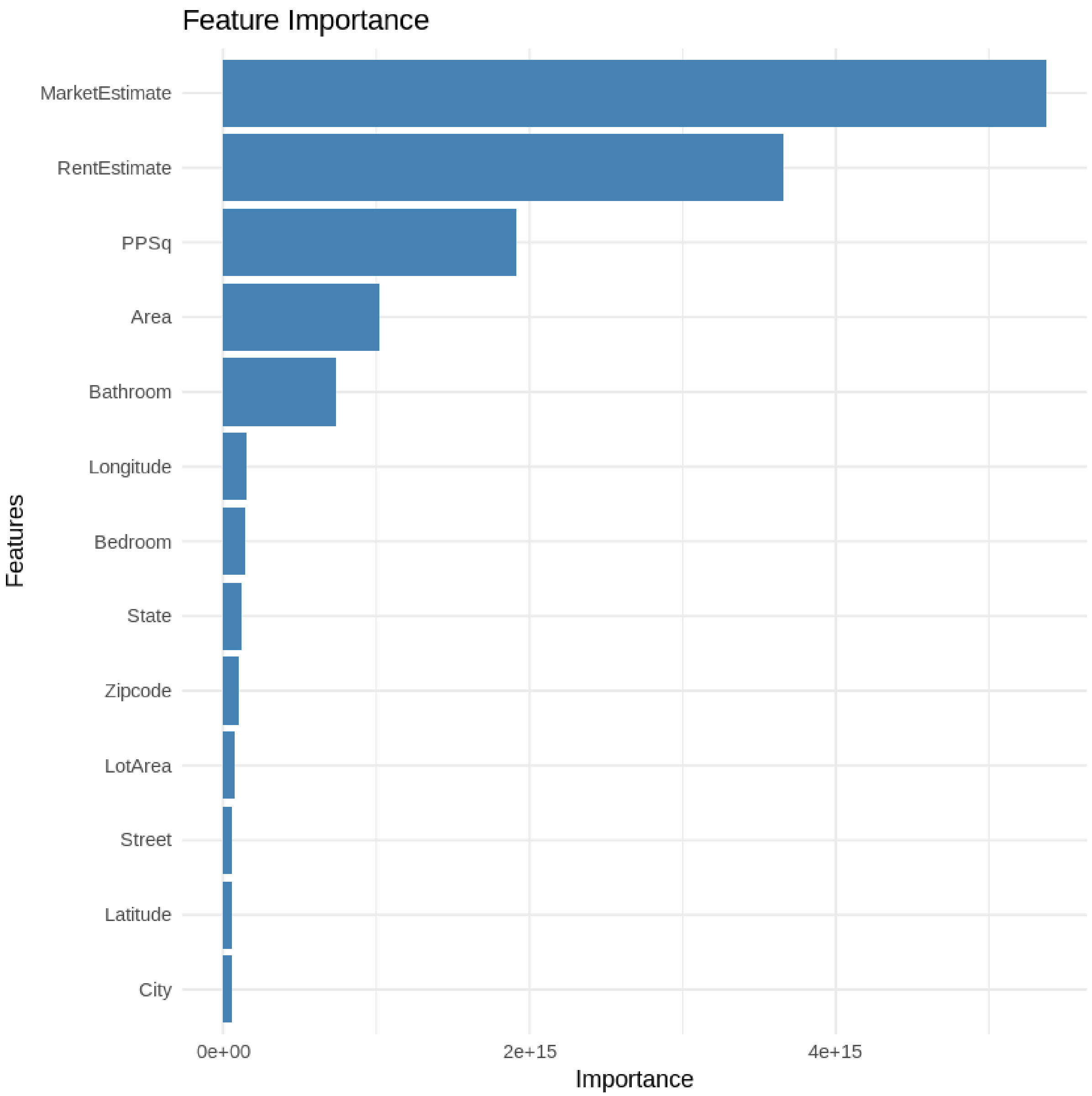
R^2 (Coefficient of Determination): The model resulted in a R^2 of 0.928 implies that **92.8%** of the variability in housing prices in the test data is captured by the model, confirming its strong predictive capability.

The strong R^2 suggests the model can handle variability across different U.S. housing markets effectively, capturing key price determinants such as property features, market trends, and location-based factors.

The model achieves high accuracy, with over 92% of variability in housing prices explained by the predictors. This level of performance is excellent given the complexity and variability in housing markets.

The RMSE value of **$336,698** suggests that the model may still have limitations in predicting extremely high or low housing prices, which are more prone to outliers or market specific nuances. However, this error is reasonable when compared to the wide range of housing prices in the dataset, going from hundreds of thousands to millions of dollars in listed housing prices. This also indicates room for improvement, especially for predicting outliers or highly volatile price ranges.

Furthermore, a following feature importance plot highlights which variables had the most influence on the Random Forest model's predictions for U.S. housing prices.



Significant Influential Variables.- The most important predictors for housing prices included:

MarketEstimate and *RentEstimate*: These two features have the highest importance scores by a significant margin. This suggests that housing price predictions rely heavily on these variables.

PPSq (Price Per Square Foot): This metric is another strong predictor, as it reflects the cost per unit area and can be a good indicator of property value relative to size.

Area and *Bathroom*: These features also contribute meaningfully to the model, highlighting the importance of property size and utility in determining prices.

Lower Influential Predictors.- Variables with minimal contributions to the model's predictive power included:

Latitude, *City*, *Street*, *Zipcode*, and *LotArea*: These variables have minimal contributions to the model's predictive power.

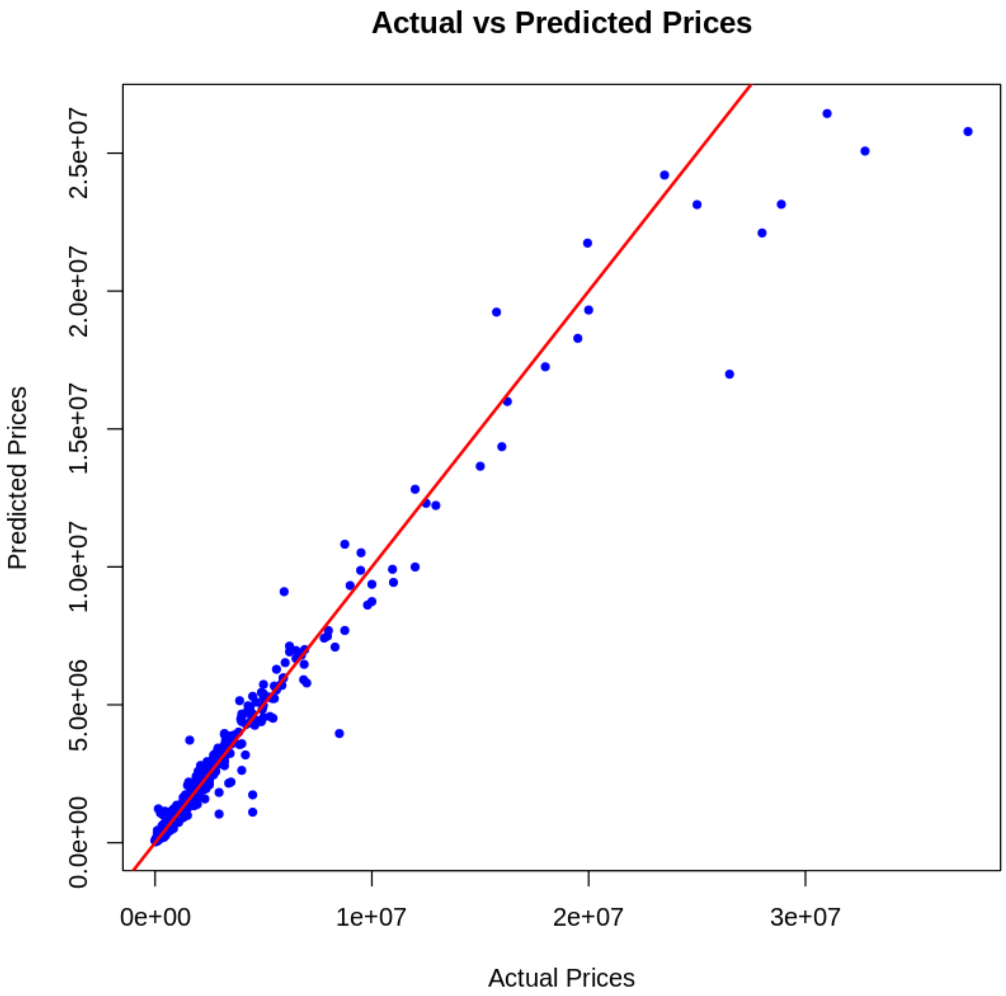
City and *Street* being low in importance may indicate uniform pricing trends within cities or streets, reducing their need in differentiating prices.

Zip code might have less importance if housing prices in the data are less location dependent or already well captured by other features like MarketEstimate.

Hence, it's interesting how this plot reflects that market-based metrics (MarketEstimate, RentEstimate) and key property features (Area, PPSq, Bathroom) drive housing prices more than detailed location-specific information (City, Zipcode).

Moreover, while geographic features play a role, they seem to be secondary to market and property specific variables in explaining price variability.

An additional graph was plotted to reflect the predictive power the random forest model yielded by showing the Actual vs Predicted Prices (listed) of housing.



At lower actual prices, from the low end of tens of thousands to millions of dollars (less than 1e+07), most of the data points are closely aligned along the diagonal line, indicating that the model's predictions are very accurate for properties with lower prices. This suggests that for typical housing within this price range, the model is making reliable and consistent predictions.

As the actual prices increase beyond $20 million (2e+07) and approach $30 million (3e+07), the data points start to spread more significantly to the right of the diagonal line (predicted values). This spread indicates that the model's accuracy begins to decrease for very high-priced properties. The model tends to slightly underestimate these higher actual prices, as reflected by the positions of the predicted prices being slightly lower than the actual ones.

Results

SAS analysis highlighted key differences in housing prices between the two markets. The average U.S. housing price was $532,439.91, compared to $1,864,172.54 in London. Further analysis observed significant differences in price ranges: in the US, housing prices ranged from a minimum of $4,888 to a maximum of $76,000,000, whereas in London, prices ranged from $180,000 to $39,750,000. These findings suggest that the US housing market offers a broader price distribution, with more affordable options and a wider spectrum of property values. In contrast, London's housing market has higher starting prices and a narrower range of high-end properties, emphasizing its higher overall cost compared to the US.

The preliminary analysis revealed a strong correlation between the two response variables, price and square footage of the lot, as well as between these variables and the predictors: number of bathrooms, square footage of living area, and square footage above ground. Our regression model confirmed the existence of a significant multilinear relationship, supported by Wilks' Lambda and a small p-value, allowing us to reject the null hypothesis. The coefficients in the model suggest that square footage of the living area has the most substantial positive impact on housing price, while bathrooms and square footage above ground exhibit negative and relatively weaker influences. This finding highlights the critical role of living space in determining property value.

Additional testing demonstrated that the relationship between the predictors and the two response variables is significantly different, suggesting distinct factors influence price and lot size.

Finally, normality tests indicated that the distribution of housing prices is right-skewed and deviates significantly from a normal distribution. This is evident from the skewness observed in the histogram, where prices are concentrated below the mean, with a long tail of higher-priced properties. Given the non-normality of the data, future analyses may benefit from non-parametric methods to provide more robust insights.

The random forest regression model showed a high level of predictive accuracy, with an R-squared value of 0.928, indicating that approximately 93% of the variability in the actual property prices is explained by the model's predictions. The Mean Squared Error (MSE) of $43.9 billion and Root Mean Squared Error (RMSE) of $336,698 suggest that while the model provides relatively accurate predictions, there is still some degree of error, especially for properties with very high prices. The high R-squared value further supports the model's strong predictive power across a wide range of listed prices.

For the Actual vs Predicted Prices in the Random Forest model, the plotted results of actual versus predicted prices reveal that for lower-priced properties, the model's predictions align closely with actual values. However, as actual prices rise beyond the $20 million mark, a noticeable spread develops, with predicted prices constantly falling below actual prices for high-value properties. This widening gap suggests that while the model performs well for the majority of properties, its accuracy diminishes for higher-end real estate, potentially due to the more complex and less predictable nature of high-priced markets. This limitation should be considered when interpreting predictions for properties at the upper end of the price spectrum.

In terms of feature importance, the results highlight the significant role of MarketEstimate and RentEstimate, which stand out as the dominant predictors of housing prices. These variables are closely tied to broader market trends and potential rental income estimates, both of which are essential drivers of property valuation. Other important predictors include PPSq (Price Per Square Foot), Area, and Bathroom, which reflect key aspects of a property's size and utility. This reinforces the idea that housing prices are strongly influenced by both market trends and the physical characteristics of properties.

Concluding Remarks

A comparative analysis reveals that housing in London is more expensive than in the US, driven by factors such as market conditions, location, and cost of living. The US housing market exhibits a much broader price range, with properties available for under $5,000 and luxury homes reaching up to $76 million. In contrast, London's housing market has a higher minimum entry point, with prices starting around $180,000, but its maximum prices are lower compared to the US.

This wider price range in the US suggests a more diverse and accessible market, offering options for buyers across various income levels, including affordable housing. On the other hand, London's higher starting prices indicate limited availability of low-cost housing, making it less accessible to average buyers or renters. Factors such as limited land availability, higher population density, and global demand may contribute to the higher cost and limited affordability in London's housing market.

The motivation behind the cluster analysis clashed with the reality of our dataset. Perhaps a future approach could be done using aggregated and averaged data for each city. This would decrease our data points, but could lead to a more intuitive comparison between London and other US cities. Interpretation of the clusters computed is also dubious in meaning to an extent. Again, aggregated and averaged data could serve to increase interpretability given London would also in turn have to become a single data point to then compare. Future analyses might explore hierarchical clustering or DBSCAN to identify more nuanced patterns in housing market data.

The analysis highlights the nuanced relationship between housing prices and key variables. Quality (grade) and space (sqft_living and sqft_above) emerged as the most critical factors, while location metrics (lat and long) reinforced the importance of geographic desirability. This aligns with findings from regression modeling, where similar variables showed strong significance.

Overall, regarding the Random Forest model, the findings from the analysis emphasize the importance of both macro-level market factors and property-specific features in predicting housing prices. While geographic data provides context, it is the broader market conditions and the characteristics of the properties themselves that most significantly drive pricing trends. This insight is valuable for developers, investors, and policymakers seeking to understand and predict housing market dynamics, particularly in diverse urban environments. Future research may explore ways to refine the model's accuracy for high-value properties and examine how additional location-specific features could improve predictive performance in varying markets.

An important consideration in our research process was the exclusion of certain datasets. Although initially identified as potential resources, the datasets from Zillow Home Value Index and United States Census Bureau were ultimately not incorporated because they were not essential to our specific research objectives. These datasets were excluded because the scope of our analysis did not require additional data on US housing trends.

References

TheDataSquad (2024). *TheDataSquad*. https://github.com/Branflakes333/TheDataSquad

Kulkarni, A. (2020). *Housing prices in London* [Dataset]. Kaggle. Retrieved November 12, 2024, from https://www.kaggle.com/datasets/arnavkulkarni/housing-prices-in-london

Philips, F. (2023). *US house listings 2023* [Data set]. Kaggle. https://www.kaggle.com/datasets/febinphilips/us-house-listings-2023

UC Irvine Machine Learning Repository housing data july 4th, 2023, from https://archive.ics.uci.edu/datasets

Authors' Bios

Brandon Miner is a fifth-year undergraduate student at San Diego State University, majoring in Statistics with an emphasis in Data Science and minoring in Mathematics. He has developed strong skills across the data analysis pipeline, from data wrangling to predictive modeling, as well as proficiency in Python and R. Currently in his final semester, Brandon has professional experience working with public policy data as an intern at the San Diego County Taxpayers Association. As he prepares for graduation, Brandon is eager to apply his skills in a data science career across a variety of industries.

Diego Angulo Nevarez is a fourth-year student at San Diego State University, majoring in Statistics with a focus in Data Science. His interests lie in applying statistical analysis and predictive modeling to better understand mental health and human behavior. Diego has experience in college law enforcement, where he analyzed local crime data to support initiatives aimed at improving campus safety. He's now shifting his focus to educational data, exploring areas like transfer student demographics, retention rates, and academic success. Proficient in R, SAS, and Python, Diego hopes to be well-equipped for a career in data-driven analysis. v

Jihyun Do is a fifth-year undergraduate student majoring in General Biology with a minor in Statistics at San Diego State University. Her academic interests focus on ecological data analysis and statistical modeling, particularly using programming languages such as R and SAS. With a strong foundation in both biology and quantitative methods, Jihyun has gained hands-on experience in analyzing ecological datasets, conducting statistical tests, and applying advanced modeling techniques to explore patterns in biodiversity and ecosystem dynamics.

Xinnan Li is a graduate student in biostatistics at San Diego State University. His interests lie in biostatistical analysis and statistical models to better understand the laws of bioinformatics. He focuses on visualizing data and is proficient in R and SAS statistical software. Currently in his second year, Xinnan has interned at ABB.inc in San Jose and has experience working with multiple sets of data. His hope is that he learns to better promote data development.

Contact Information

Brandon Miner: Email: [bminer5476@sdsu.edu], LinkedIn: [linkedin.com/in/brandonminer], GitHub: [https://github.com/Branflakes333]

Diego Angulo Nevarez: Email: [dangulonevarez5690@sdsu.edu], LinkedIn: [linkedin.com/in/diego-angulo-654625328]

Jihyun Do: Email: [jdo4550@sdsu.edu], LinkedIn: [linkedin.com/in/jihyundo]

Xinnan Li: Email: [xli1479@sdsu.edu], LinkedIn: [linkedin.com/in/xinnan-li-010533180]