Linear Regression Analysis in Miami Housing

Chongchong Jiang, Jain, Sunayana, Boateng, Bernard

1. Introduction

1.1 Background

Miami, a metropolis renowned for its breathtaking beaches, cultural richness, and dynamic lifestyle, boasts a real estate scene as diverse as its neighborhoods. Each district possesses distinctive characteristics and amenities, contributing to the city's multifaceted identity. The real estate landscape in Miami is not only a reflection of its vibrant atmosphere but also subject to fluctuations driven by economic factors, prevailing trends, and external events. The market's volatility adds an extra layer of intrigue to the city's thriving property sector, making it an ever-evolving and dynamic investment landscape.

1.2 Statement of the Problem

The project aims to investigate and to identify the variables influencing housing prices using the Miami Housing Dataset. This dataset encompasses various features such as sale price, land area, floor area, distance to amenities, structure characteristics, and geographical coordinates. We hypothesize that collecting the data of multiple factors affecting the house sold price. Moreover, the house price can be expressed as a formula by multiple influencing factors. And how well these variables can predict house prices.

The analysis will use R programming language and involve a combination of descriptive statistics, correlation analysis, and data visualization to identify patterns and relationships between housing prices and potential predictor variables.

1.3 Significance and Applications

The problem of accurately modeling and predicting house prices, particularly in a dynamic city like Miami, holds immense significance for various stakeholders due to its far-reaching implications. Precise price prediction is crucial for several aspects of the real estate industry and has broader implications for the economy, urban planning, and society as a whole. For prospective homebuyers, accurate house price predictions can aid in making informed purchasing decisions, ensuring that they acquire a property that aligns with their financial capabilities and aspirations. For real estate agents, reliable house price predictions can enhance their ability to guide clients and optimize pricing strategies. Furthermore, house price predictions play a pivotal role in urban planning and economic development, enabling policymakers to make informed decisions regarding infrastructure investments, zoning regulations, and tax policies.

2. Dataset

The project's dataset was sourced from Kaggle, a reliable source of data recognized for its expertise in offering a broad variety of datasets, including real estate data. This

dataset, which is especially suited to Miami, provides useful information on 13,932 single-family homes in 2016. The dataset consists of a substantial amount of data, with 13,932 rows and 17 columns without missing data.

This dataset primarily includes information in three aspects. Firstly, there is building-related information, encompassing the geographical coordinates (latitude and longitude) of the house, its age, land and floor area, and structural quality. Secondly, there is location information concerning the proximity to commercial centers, transportation hubs and the ocean and the body of water. Lastly, there is information regarding the sale of the house, including the selling price, the month of sale, and the value of special equipment within the house. These data can be utilized for tasks such as predicting house prices or conducting related analyses.

3. Method

3.1 Variables Visualization

Upon reviewing the dataset, it is evident that, apart from the ID numbers, the dataset comprises three types of variables: spatial, continuous, and categorical. we create scatterplots and boxplots individually to obtain an overview of the relationships between the response variable and other predictors, as well as the general distribution of predictors. From the results of data visualization, we can see that total living area is the most highly and positively correlated with sale price. And it is clearly evident that structure quality is directly proportional to sale price.

3.2 Correlation Coefficients

The correlation matrix provides insights into the relationship between various variables in the dataset including predictors collinearity checking. If we employ a cutoff of ±0.8, our conclusion is that no predictors are highly correlated. Additionally, given the correlation coefficient of 0 between SALE_PRC and month_sold, we have opted to exclude month_sold from the predictors at this stage.

	SALE_PRC	LND_SQFOOT	TOT_LVG_AREA	SPEC_FEAT_VAL	RAIL_DIST	OCEAN_DIST	WATER_DIST
SALE_PRC	1.00	0.36	0.67	0.50	-0.08	-0.27	-0.13
LND_SQFOOT	0.36	1.00	0.44	0.39	-0.08	-0.16	-0.06
TOT_LVG_AREA	0.67	0.44	1.00	0.51	0.08	-0.05	0.15
SPEC_FEAT_VAL	0.50	0.39	0.51	1.00	-0.02	-0.06	0.01
RAIL_DIST	-0.08	-0.08	0.08	-0.02	1.00	0.26	0.16
OCEAN_DIST	-0.27	-0.16	-0.05	-0.06	0.26	1.00	0.49
WATER_DIST	-0.13	-0.06	0.15	0.01	0.16	0.49	1.00
CNTR_DIST	-0.27	-0.02	0.14	-0.05	0.44	0.25	0.53
SUBCNTR_DI	-0.37	-0.16	-0.04	-0.15	0.49	0.43	0.20
HWY_DIST	0.23	0.13	0.23	0.15	-0.09	0.09	0.40
age	-0.12	0.10	-0.34	-0.10	-0.23	-0.16	-0.33
avno60plus	-0.03	-0.01	-0.06	-0.01	-0.12	0.04	-0.10
month_sold	0.00	0.01	0.00	-0.01	0.01	-0.01	0.01
structure_quality	0.38	-0.01	0.17	0.19	-0.07	0.21	-0.03

	CNTR_DIST	SUBCNTR_DI	HWY_DIST	age	avno60plus	month_sold	structure_quality
SALE_PRC	-0.27	-0.37	0.23	-0.12	-0.03	0.00	0.38
LND_SQFOOT	-0.02	-0.16	0.13	0.10	-0.01	0.01	-0.01
TOT_LVG_AREA	0.14	-0.04	0.23	-0.34	-0.06	0.00	0.17
SPEC_FEAT_VAL	-0.05	-0.15	0.15	-0.10	-0.01	-0.01	0.19
RAIL_DIST	0.44	0.49	-0.09	-0.23	-0.12	0.01	-0.07
OCEAN_DIST	0.25	0.43	0.09	-0.16	0.04	-0.01	0.21
WATER_DIST	0.53	0.20	0.40	-0.33	-0.10	0.01	-0.03
CNTR_DIST	1.00	0.77	0.08	-0.55	-0.13	0.02	-0.33
SUBCNTR_DI	0.77	1.00	-0.09	-0.39	-0.07	0.02	-0.25
HWY_DIST	0.08	-0.09	1.00	-0.12	-0.02	0.00	0.19
age	-0.55	-0.39	-0.12	1.00	0.11	-0.04	0.01
avno60plus	-0.13	-0.07	-0.02	0.11	1.00	-0.02	0.10
month_sold	0.02	0.02	0.00	-0.04	-0.02	1.00	-0.01
structure_quality	-0.33	-0.25	0.19	0.01	0.10	-0.01	1.00

3.3 Data preprocessing

The dataset comprises diverse units of predictors ranging from categorical to numerical values. To accurately gauge the impact on each category within the categorical variables and to unbiasedly compare the significance of predictors, we conduct normalization for continuous variables and create dummy variables for categorical variables separately.

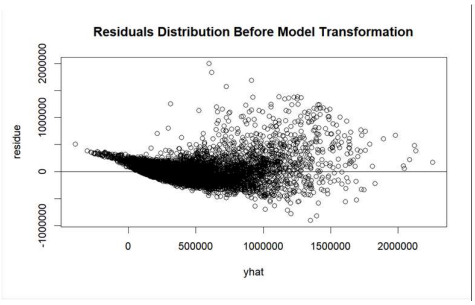
Z-score was used to adjust the scale of numerical variables to a standard range, typically with a mean of zero and a standard deviation of one.

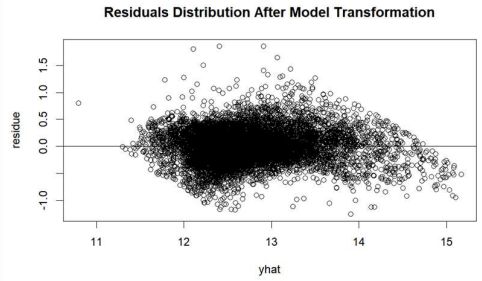
$$z\text{-score} = (x - \mu) \div \sigma$$

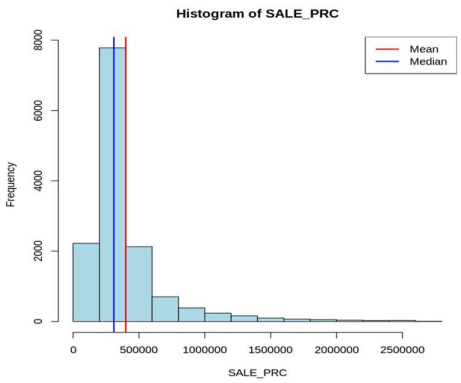
In the dataset, the categorical variables "structure_quality" and "avno60plus" have been encoded to dummy variables. As the result, “avno60plus” was converted to binary variables (taking values of 0 or 1), and “structure_quality” was used as one-hot encoding converted to multiple dummy variables.

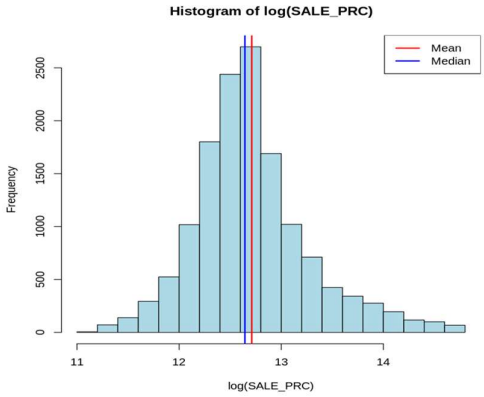
3.4 Model transformation

After data normalization and encoding of variables, the first model incorporated all the variables. Based on the residual checking output, the plot depicting residuals against fitted response variables does not exhibit a random distribution around 0. Instead, they appear clustered together on the right side. the normal Q-Q plot was observed to have a long right upper tail suggesting presence of outliers or skewness toward higher values. To enhance the distribution of residuals, we apply a logarithmic transformation to the response variable. It's evident that after applying the model transformation, the residuals exhibit a distribution closer to normal.









3.5 Predictor's selection

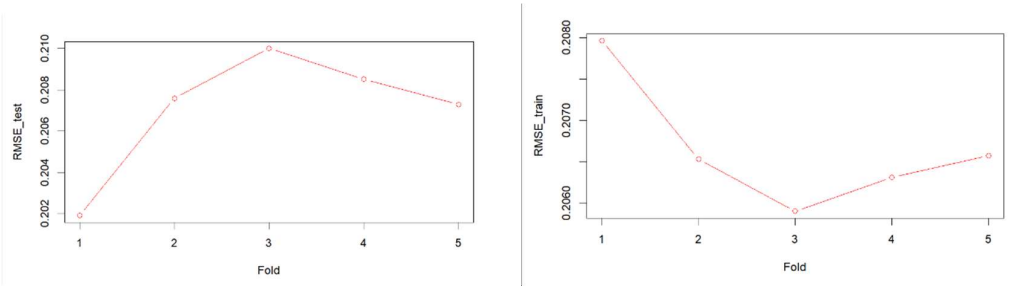
We conducted a Backward model selection using both the Akaike Information Criterion (AIC) and p-values in the T-test (with a significance cutoff of 0.05). Both methods converge to the same conclusion, identifying WATER_DIST as the sole predictor that is not significantly affecting the response variables.

3.6 Outliers Identification

Studentized residuals were utilized to detect outliers significantly impacting the model's performance, and a subsequent 5% cutoff was applied to eliminate all such values. This step was taken to facilitate the development of a more robust model. Ultimately, a total of 808 outliers were identified and removed.

3.7 Cross Validation

After thorough scrutinization and tuning of the log-transformed regression model, a final model without outliers and the WATER_DIST variable was settled upon. The model was subjected to a cross-validation process. The dataset was divided into five folds and the model was trained and tested iteratively. The Root Mean Squared Error (RMSE) is calculated for both the training and test sets, providing a comprehensive evaluation of the model's predictive performance across different subsets of the data.



4. Conclusion

Assessing the distribution of residuals, we transformed the model using a logarithmic function for the response variable. Based on the T-statistic test and AIC values, we selected variables highly significant to house price data and conducted a multiple linear regression analysis to construct a predictive model. This adjustment aimed to enhance the model's fitting effect and improve prediction accuracy, thereby enabling us to select the most optimal predictive model.

Through the research process described above, the model derived in this paper for forecasting house prices in Miami and beyond is as follows:

$$Y = 12.08 + 0.04 * X1 + 0.28 * X2 + 0.05*X3 + 0.03*X4 - 0.04* X5 - 0.07 * X6 - 0.14 * X7 + 0.04 * X8 - 0.10 * X9 - 0.06 * X10 + 0.47 * X11 + 1.27 * X12 + 0.65 * X13 + 0.91 * X14$$

In the equation, X1 expresses land area (square feet), X2 expresses floor area (square feet), X3 expresses value of special features, X4 expresses distance to the nearest rail

line, X5 expresses distance to the ocean , X6 expresses distance to the Miami central business district, X7 expresses distance to the nearest subcenter, X8 expresses distance to the nearest highway, X9 expresses age of the structure, X10 expresses airplane noise exceeding an acceptable level, X11-X14 expresses different class of quality of the structure.

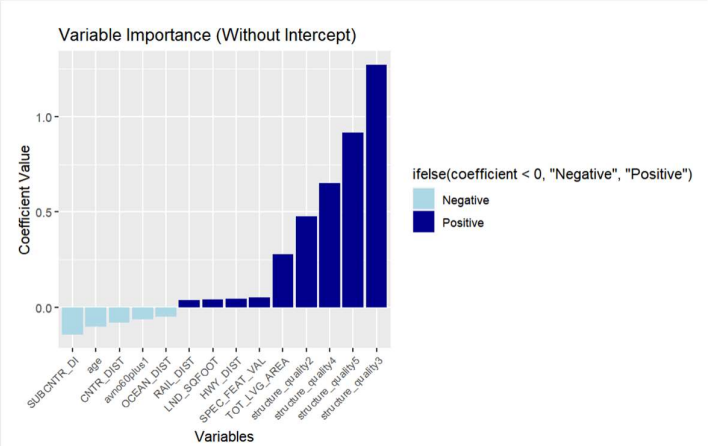
The regression model's performance in explaining housing sale price variability in the Miami dataset is noteworthy. It demonstrated a strong fit, as evidenced by an Adjusted R-squared value of 0.8425, indicating that approximately 84.25% of the variation in the logarithm of sale prices was accounted for by the independent variables.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.58268 -0.12866 -0.00324  0.12797  0.65122

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       12.082775   0.016877 715.945  < 2e-16 ***
avno60plus1       -0.062249   0.015061  -4.133  3.6e-05 ***
structure_quality2  0.478320   0.017428  27.446  < 2e-16 ***
structure_quality3  1.270483   0.054771  23.196  < 2e-16 ***
structure_quality4  0.650053   0.017087  38.045  < 2e-16 ***
structure_quality5  0.916811   0.017655  51.928  < 2e-16 ***
LND_SQFOOT          0.041874   0.002312  18.115  < 2e-16 ***
TOT_LVG_AREA        0.280342   0.002550 109.925  < 2e-16 ***
SPEC_FEAT_VAL       0.052058   0.002215  23.506  < 2e-16 ***
RAIL_DIST           0.037960   0.002146  17.686  < 2e-16 ***
OCEAN_DIST         -0.047933   0.002260 -21.206  < 2e-16 ***
CNTR_DIST          -0.078439   0.003617 -21.686  < 2e-16 ***
SUBCNTR_DI         -0.141532   0.003385 -41.813  < 2e-16 ***
HWY_DIST            0.045126   0.002020  22.335  < 2e-16 ***
age                -0.100018   0.002470 -40.494  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2068 on 13109 degrees of freedom
Multiple R-squared:  0.8426,    Adjusted R-squared:  0.8425
F-statistic:  5013 on 14 and 13109 DF,  p-value: < 2.2e-16
```

The model highlights several influential factors affecting housing prices in Miami. As considering the importance of predictors, varying levels of structural quality significantly influence property values, underscoring the critical role of a building's quality in determining its worth. Larger land areas and increased living spaces, along with specific distinctive features, contribute to higher property prices. Conversely, proximity to the ocean or city center, as well as older properties, often correlates with lower prices.



The closeness between the RMSE values in 5-folds cross-validation indicates the model's robustness, as it performs comparably well on new, unseen data, suggesting it doesn't overfit and maintains predictive accuracy consistently across different datasets.

Overall, these assessments and validations affirm the regression model's reliability in explaining housing price variations in Miami. The model not only comprehensively captures influential factors but also demonstrates consistent predictive performance across various datasets, emphasizing its generalizability and accuracy in predicting housing prices.

5. Limitations

The study faced a few constraints and highlighted potential areas for future exploration. Firstly, the data was confined to the year 2016, possibly limiting the comprehensive understanding as unable to track changes over time. This might have obscured broader trends. Additionally, crucial temporal aspects such as interest rates and the Consumer Price Index (CPI) were absent from our dataset, which are fundamental in understanding real estate market behavior. Furthermore, the analysis didn't delve into other influential factors on housing, such as demographic statistics, educational districts, and governmental policies, which could substantially enhance the depth of the investigation.

Furthermore, the model was intricate due to its numerous variables and complex transformations, posing challenges in visualization and interpretation. As for future endeavors, simplifying the model or exploring alternative methodologies could make it more accessible and understandable for a wider audience.

6. References

1.Ningyan Chen. (2022).House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis.Hindawi.House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis (hindawi.com)
2. Ashish. (2018).Housing Price Prediction (Linear Regression).Kaggle.Housing Price Prediction (Linear Regression) | Kaggle

7 Appendix

Dataset Resource: https://www.kaggle.com/datasets/deepcontractor/miami-housing-dataset
Code: PDF file in a compressed document