This project aims to analyze housing prices in Lisbon, Portugal, using applied regression analysis. The objective is to build a predictive model for estimating housing prices based on property attributes like condition, type, bedrooms, bathrooms, area, parking, and location.

Lisbon’s real estate market is dynamic and competitive, attracting diverse buyers and investors. Understanding the factors influencing housing prices is crucial for making informed decisions in this market.

The analysis utilizes a comprehensive dataset of residential properties in Lisbon, containing essential variables like property condition, type, bedrooms, bathrooms, area, parking availability, geographical coordinates, and price of each square meter.

By employing applied regression techniques, particularly linear regression, we will construct a predictive model that provides valuable insights into the factors driving housing prices in Lisbon, benefiting stakeholders, including buyers, sellers, investors, and policymakers.

(点)

Now I will introduce the methods of our analysis.

First of all, process the categorical and quantitative variables. Then find the variables which have collinearity, and delete one of them.

After that, do linear regression on all variables. Then analyze the results and delete the variables which are not significant.

And we will do linear regression on the remaining variables and continue to see if more variables need to be deleted. Then Use F-Test to decide if we will accept the reduced model.

Subsequently, analyze the Residuals vs Fitted values plot and see if the residuals satisfy our requests. Then List all the points with high leverage to ignore them and do the regression.

After that, analyze the Normal Q-Q plot and see the points with high Standardized Residuals. Then Analyze the Leverage plot and see the points with high Cook’s Distance.

And we will exclude points with high Standardized Residuals and Cook’s Distance and do the regression again. Then repeat the above three steps until there are no outliers. And plot the heat map of variables.

In model 2, it is very similar to the method of model 1, but more focus on the polynomial of variables X like X^2, X^3, log(X) ... and the multiplication of variables like X1\*X2, X1\*X2\*X3 ...

(点)

Firstly, we do the regression on all the variables and get the following result. Since there are too many variables, I just show some part of the results.

(点)

Then we find that the variable AreaGross is not included in the regression model since it has collinearity with other variables. Therefore, we do the scatter plot of AreaGross versus each variables to check the collinearity. Then we can find that the variable AreaNet has strong linear relationship with AreaGross. Since there are too many plots, I just choose the one which can show the collinearity.

(点)

Then we further do another regression deleting some NA variables and variables with t-test value greater than 0.1 shown in the first regression.

We can see that the r squared dropped a little, so we further use F-Test to decide if we will accept the reduced model. The result shows that we can accept the reduced model.

(点)

According to the Residual Plot below, we can notice that the variance of residuals increases as Fitted value getting large, with some outliers.

(点)

Then we do the Q-Q Plot to check if the standardized residual obeys normal distribution. Then we find 3 outliers: point 150, point 196, point 145. The other residuals basically obey the normal distribution since they are almost on a straight line and ranges from (-2,2).

To determine whether these points will have great influence on the regression, we do the leverage plot. We can see from the plot that point 145, point 196 and point 246 will have great influence on the regression result.

(点)

Combine the results of Q-Q plot and leverage plot, we decided to delete the outliers: point 145 and point 196. Then we do a new regression to see the results.

(点)

Using the similar methods mentioned above, we delete new outliers: points ”23”.

(点)

Using the similar methods mentioned above (See Figure 13 and Figure 14), we cannot find more outliers.

(点)

Finally, we do a new regression to see the results. And we use the heat map to visualize the degree of variable influence —- with deeper red colors indicating greater influence.