Supervised Assignment

“Linear Regression”

EXECUTIVE SUMMARY

The use of linear regression analysis has been widespread to estimate the effects of independent variables on a dependent variable. This comprehensive report presents a detailed linear regression analysis aimed at developing a robust model for helping you estimating rental prices. This model serves as a vital tool for the real estate agency, enabling the setting of competitive prices and identification of lucrative market opportunities.

We employed a methodical approach, using logarithmic transformation of the dependent variable, rent, to stabilize variance and reduce skewness, thus enhancing data interpretability.

The model demonstrates strong explanatory power with an R^2 score of 58.78% and exhibits high predictive accuracy, indicated by a Mean Absolute Percentage Error (MAPE) of 4.07%. This performance is achieved through careful data splitting into training and testing subsets, ensuring robustness against unseen data and preventing overfitting.

Our analysis, utilizing a stepwise method for variable selection, identifies property size (Sq.Mt) and type (Cottage, Semidetached, Penthouse) as significant predictors of rental prices, crucial for strategic decision-making in pricing and investment. Sq.Mt, with the highest T-test value of 42.24, is the most influential, where each additional unit increases rent by 0.2%, ceteris paribus. Conversely, Cottages, negatively impacting rent, decrease prices by 29% per unit. Penthouses and Semidetached properties boost rent by 9% and 17.8% per unit, respectively, ceteris paribus, showcasing the varied impact of different property types on rental pricing.

Our validation strategy involved partitioning the dataset into 80% for model training and 20% for testing, crucial for avoiding overfitting and assessing performance on new data. Post-training, we analyzed prediction errors, confirming the model's consistency and reliability.

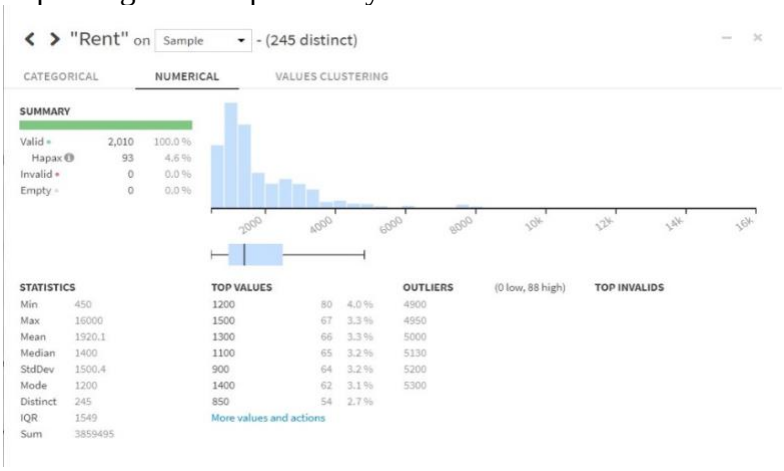
We recommend integrating this model into the agency’s pricing toolkit, regularly updating it with the latest market data. This integration will not only refine pricing decisions but also enable the agency to swiftly identify and exploit profitable real estate investment opportunities.

Our model exemplifies the effectiveness of data-driven strategies in the dynamic economic environment of real estate, highlighting the importance of analytical approaches in decision-making.

Hence, the linear regression model developed in this report offers a powerful tool for understanding and predicting rental prices. It underscores the value of analytical methodologies in real estate, guiding our agency towards informed, data-driven decisions in a competitive market landscape.

TECHNICAL APPROACH TO REGRESSION AND VARIABLE SELECTION

In our technical approach to regression and variable selection, the model specification began with a comprehensive analysis of various variables, focusing on rent as the variable of interest. To enhance the model's effectiveness, the logarithm of rent was chosen as the dependent variable. This transformation stabilizes the variance, addresses the skewness seen in the image below in the data, and makes the relationships more linear. It also mitigates the impact of outliers, thereby improving the interpretability of the model.



Running the stepwise selection method yielded the following relevant variables: Cottage, Semidetached, Penthouse, and Sq.Mt. The Cottage variable has a negative impact on rental prices, whereas the remaining three exhibit a positive influence. Square meters (Sq.Mt) are the most crucial in elucidating rental prices, as evidenced by the highest T-test value of 42.24. This underscores a strong relationship between a property's size and its rental value. It's reasonable to expect that larger properties generally command higher rents. Furthermore, the size of a property affects rental prices universally, unlike property type, which, being a categorical variable, influences rent in a binary fashion—either contributing to the rent when present (1) or not at all when absent (0). Precisely, an increase of one square meter generates a 0.2% rise in rent, ceteris paribus. Conversely, the Cottage variable is associated with a 29% reduction in rent per unit. In contrast, additional units of the Penthouse and Semidetached variables are linked to an average rent increase of 9% and 17.8% respectively, all else being equal. The regression model generated a set of regression coefficients for each relevant variable, and was therefore formulated as follows:

$$\text{Log (Rent)} = 0.002 \times \text{Sq.Mt} - 0.29 \times \text{Cottage} + 0.09 \times \text{Penthouse} + 0.18 \times \text{Semidetached} + 2.94$$

The Ordinary Least Squares (OLS) method was employed to determine the best fitting line, aiming to minimize the sum of squared errors between the observed and predicted values. This method is crucial for finding the most accurate representation of the data. The successful application of OLS resulted in a fitting line that aligns well with our data points, as reflected in the R^2 value of 58.78%. This value indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, with a higher value denoting a better model fit. Therefore, ~59% of the rent could be explained by our four independent variables selected.

The Mean Absolute Percentage Error (MAPE) at 4.07% falls well below the acceptable threshold of 6%, underscoring the precision of our predictions. Furthermore, the Mean Squared Error (MSE)

stands at 0.028, confirming that the average squared deviation is low, hence our model's predictions are consistently close to the actual values.

MODEL EVALUATION

To evaluate our model, we have conducted a primary analysis which consisted of checking the variables selected one by one and verifying the logic behind the coefficient's signs and variable importance. It made sense that the most important variable would be the Sq.Mt since it is usually the factor that affects the rent the most. The bigger the house the higher the rent.
Second, we have checked the individual significance of the variables by assessing the t-student test and the p-values of each of the variables following the stepwise method, just to select which variable is the most important in explaining the target variable.
Third, we have checked the global significance of the model by assessing the R^2 and the Adjusted R^2 which were high enough for us to conclude that the model was satisfactory, and the independent variables are indeed able to explain the target variable. The errors were also checked; the Mean Squared Error (MSE) was 0.028 and the Mean Absolute Percentage Error (MAPE) was 4.07%.
Finally, the outliers were observed and no particular outliers had an extreme influence on our results, hence they were kept as they were. Also, statistical assumptions before running the analysis about homoscedacity in residuals, lack of multicollinearity between explanatories, lack of correlation between residuals and the normality in residuals were checked before running the anlaysis. Additionally, null values were removed by running the python code for linear regression, making sure that data quality was good enough for selecting the important variables before proceeding with the analysis.

VALIDATION STRATEGY

In our robust validation strategy, we meticulously partitioned the dataset into training and testing sets, with 80% allocated for model training (rent_model) and the remaining 20% reserved for testing (rent_reserved). This careful division is critical to prevent overfitting and to evaluate the model's performance on unseen data. After training on the larger subset, we conducted a comparative analysis of prediction errors from both sets, thereby affirming the model's consistency and reliability across different data samples.

To enhance our understanding of the model's predictive accuracy, we exported the test set predictions to Excel. This step allowed for an in-depth analysis of the error distribution, offering valuable insights into the model's performance nuances. The error % calculated was found to be around the range of 1-10% between actual and predicted values, indicating a strong accuracy in our model. The error distribution analysis revealed a bell-shaped curve centered around zero, a clear indication of the model's unbiased nature in its predictions. This distribution pattern is particularly encouraging as it suggests that the model is equally likely to overestimate and underestimate the rental prices, thus maintaining a balanced prediction approach.

Moreover, the very low average error percentage observed in this distribution underscores the model's precision. Such a low error margin is indicative of the model's ability to closely mirror the actual rental prices, enhancing its practical utility in real-world applications. This thorough

examination not only validates our model but also provides a clear demonstration of its potential as a reliable tool for predicting rental prices with a high degree of accuracy.

CONCLUSION AND RECOMMENDATIONS

After conducting a linear regression analysis to explain house rental prices, the following key findings and conclusions emerge for you to take into consideration:

-Square Meters: The analysis indicates a positive and significant relationship between the size of the property (square meters) and rental prices. Larger properties tend to command higher rental rates, reflecting the demand for more spacious living spaces.

-Cottage Type: The presence of a cottage shows a negative relationship with rental prices. This suggests that, all else being equal, houses with cottages may have lower rental values, possibly due to factors like reduced overall living space or other characteristics associated with cottage-style properties.

-Penthouse or Semidetached: The analysis reveals a positive impact on rental prices for properties categorized as penthouses or semidetached. This implies that these types of properties are associated with higher rental values.

As for the recommendation for you in the real estate agency rental estimation: you should utilize the developed linear regression model to estimate rental prices for houses in the market. This will provide you with a quantitative tool to assess and propose competitive rental rates based on the property's characteristics. Additionally, you must consider the impact of square meters, cottages, and property type when estimating rental prices for clients and tailor the estimates to reflect the unique features and characteristics of each property.

Hence, you should focus on Square Meters when looking for potential investment opportunities or properties for clients by prioritizing larger properties, identifying flats with ample square meters that may be underpriced in the market, offering potential for competitive rental returns. Also, you should cautiously Consider Cottages given the negative relationship with rental prices. Assessing the overall market demand is a must for such properties and adjusting pricing strategies accordingly. You should explore opportunities in the market for penthouses or semidetached properties, as the analysis suggests a positive impact on rental prices. Identify the properties with these characteristics that may be undervalued in the current market.

For monitoring and adaptation, you should:

- Regularly update and refine the linear regression model as new data becomes available to ensure its continued accuracy in reflecting market trends.
- Stay informed about changes in the real estate market, economic conditions, and evolving preferences of renters to adapt the model and recommendations accordingly.

By incorporating these recommendations, you can enhance your ability to estimate competitive rental prices and identify potentially lucrative investment opportunities in the dynamic housing market.

TECHNICAL ANNEX

Metrics and Assertions:

Metrics and assertions

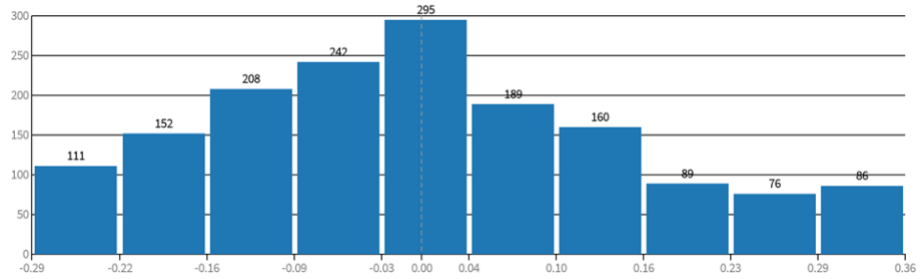
Detailed metrics

Explained Variance Score	**0.5878**
Mean Absolute Error (MAE)	**0.1300**
Mean Absolute Percentage Error	**4.07%**
Mean Squared Error (MSE)	**0.02861**
Root Mean Squared Error (RMSE)	**0.1691**
Root Mean Squared Logarithmic Error (RMSLE)	**0.03956**
Pearson coefficient	**0.7667**
R2 Score	**0.5878**

Variable importance:

Variable	Coefficient	Std. Err	T stat	p-value	Confidence
Cottage	-0.2919	0.0323	-9.0285	< 1e-4	★★★
Semidetached	0.1784	0.0483	3.6926	0.0001	★★★
Penthouse	0.0902	0.0156	5.7679	< 1e-4	★★★
Sq.Mt	0.0020	0.0000	42.2383	< 1e-4	★★★
Intercept	2.9407	2.9407	1.0000		

Residuals:



Predicted VS Actual values: