



Phase Two Project

By John Mark Ang'awa

Overview

- Introduction
- Business understanding
- Project goal
- Data understanding
- Methods
- Data preparation
- Rationale of the analysis
- Limitations of the analysis
- The data analysis process
- Modeling
- Regression results
- Conclusion
- Visualization
- Recommendations





Introduction

The real estate agency operating in northwestern county wants to do an analysis on the types of properties in the county to assess their value. The business is very competitive since the quality of houses ranges with many factors involved. However they need a way forward on what type of properties to choose and the price ranges in order to make firm decisions that will generate profit for them and make them outstanding.




Business understanding

The real estate agency lacks enough knowledge on how to perform best in this field. The business problem is to develop a predictive model that can accurately estimate house prices based on various features provided in the King County House Sales dataset. The agency aims to use this model to guide their clients in setting competitive prices for their properties, optimize their marketing strategies, and improve their overall business performance in the real estate market.



Project goal


To develop a predictive model that can accurately estimate house prices that is profit yielding in various features such as location, size, condition, and amenities.





Data understanding

To achieve this, we will build multiple models starting with a basic model and then iteratively refining it to improve predictive accuracy and interpretability. By analyzing the dataset, performing data visualization, and evaluating different model performance, we can gain insights into the key factors influencing house prices in the area and provide actionable recommendations to the real estate agency.





Methods

The size of the dataset is wide allowing for a robust analysis and provides a comprehensive view of the houses' information. I will perform a statistical data analysis using various data science modeling analytical tools and come up with visualizations that can describe the data better to the stakeholders. This will help identify key trends and characteristics of the houses such as location, size, condition, and amenities.

However, it is important to acknowledge that the datasets may have some limitations. They might not capture the entire universe of houses, and there could be missing or incomplete data for certain houses. Additionally, the data might be subject to biases or limitations inherent in the sources themselves. Despite these limitations, the datasets provide valuable insights into the movie industry and are suitable for addressing the business problem at hand.



Data preparation

Here, I used the dataset from `kc_house_data`. This data provides a wide range of data collected overtime that can be a good starting point. I viewed the data, performed appropriate data cleaning in preparation for modeling and analysis.



Rationale of the analysis

Statistical analyses is used here, specifically multiple linear regression modeling, to capture the relationships between the vast features in a quantitative manner. Regression coefficients provide us with the magnitudes and directions of these relationships, enabling us to estimate the impact of each feature on house prices. Using regression coefficients goes beyond the visual exploration of data and provides precise numerical measures of the feature contributions.

The problem of analyzing house sales in a northwestern county is suitable for multiple linear regression because it involves multiple independent features that may collectively influence the house prices. This form of analysis allows us to estimate the relative importance of each feature and understand how changes in these features affect the predicted house prices.



Limitations of the analysis

- The assumptions of linear regression involved in the statistical modeling process. Violations of these assumptions can affect the validity of the model.
- The presence of high correlation among independent variables can lead to unstable coefficient estimates.
- Missing data and outliers should also be considered. Missing data handling is crucial but not explicitly addressed in this analysis. Dropping missing values might lead to a loss of information, and imputation techniques could be explored.



The data analysis process

Here, I used linear regression modeling in analysing the dataset. I performed data exploration and cleaning on the data, built a baseline model, iterate through it and report on its metrics. Then used visualizations and finally come up with recommendations on the findings of the analysis.



Modeling

Building baseline model.

Iterating through the model.

Choosing a final model and reporting on it.



Regression results:

1. Baseline model interpretation

- The model is statistically significant overall, with F-statistic and p-value of well above 0.05.
- The model explains about 96% of the variance in price. This might be good.
- Each time the reference category changes, the const and the other coefficients change. Const changes because it represents the value when all predictors are 0, and this means that const represents when the reference category is true.
- The model coefficients are overall statistically significant with t-statistic p-value well above 0.05.



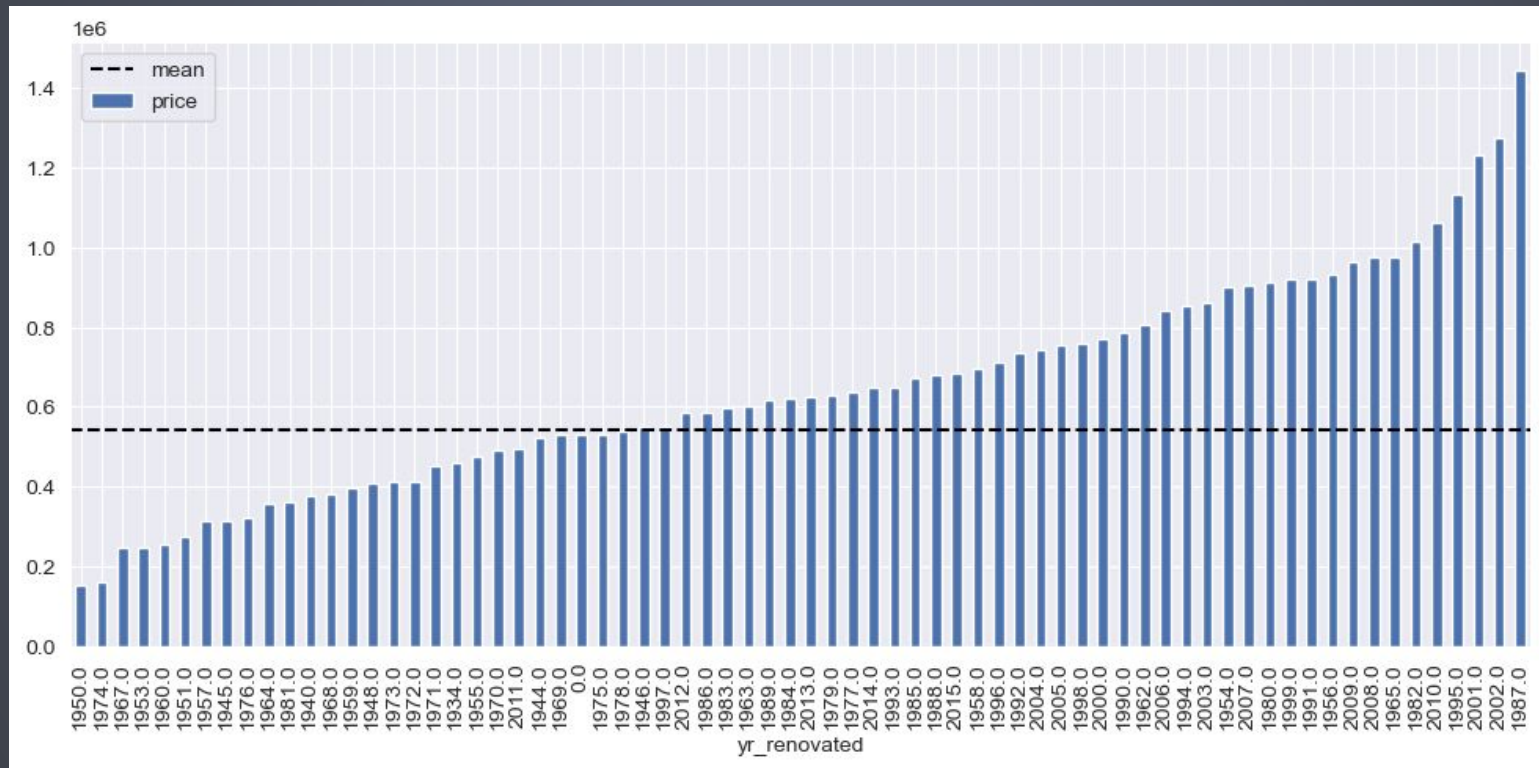
2. Iterated model results:

- The model's r-squared and adjusted r-squared are generally statistically significant, with p-values of about 0.05.
- Each time the reference category changes, the const and the other coefficients change. Const changes because it represents the value when all predictors are 0, and this means that const represents when the reference category is true.
- For each increase in the model coefficients (sqft_living, bedrooms, bathrooms, floors) there is an associated increase in price of houses.

Conclusion

The aim of this project is to provide valuable insights for the real estate agency through exploring the dataset. Through this comprehensive house dataset analysis, I can identify that the types of houses to focus on should be based on the year built and year renovated. The findings of this analysis will enable stakeholders to make informed decisions concerning the types of houses to invest on. Through these insights, the real estate agency can become successful in the real estate industry and establish a strong presence as a best genuine business people. The visualization shown next shows the different values of the houses progressively over the years and can be used as benchmark.

Visualisation of final results





Recommendations

Based on the model results and limitations, the real estate agency should consider the following actions:

1. The real estate agency should use the model as a tool to estimate house prices accurately, guide pricing strategies, and make informed business decisions.
2. The coefficient values can be a guide in understanding the relative importance of each feature on house prices. For example, a positive coefficient indicates an increase in price with an increase in the corresponding feature, while a negative coefficient indicates a decrease in price.
3. It should be aware of the limitations and uncertainties in the analysis, considering potential model assumptions and the excluded factors that could impact house prices. They should use the models as a supportive tool alongside domain knowledge and market expertise.

By leveraging the multiple linear regression models, stakeholders can make more accurate price estimations and gain valuable insights into the factors driving house prices in the Northwestern county.



Thank

you