1. BUSINESS OVERVIEW

1.1 Introduction

As a data scientist analyzing the King County housing market, our business understanding is that the real estate industry is a crucial sector that plays a significant role in the economy. The success of a real estate transaction depends on several factors, including the location, the size of the property, the condition of the property, the amenities, and the current market conditions. The housing market is subject to various external factors such as interest rates, economic conditions, and government policies. Through analysis of the data, we can create a documentation that can be used to make informed decisions.

1.2 Challenges

The main challenge in this project is the quality of the data used. There may be missing or inconsistent data that could affect the accuracy of the analysis. Additionally, there may be outliers in the data that needs to be addressed. It is important to clean and preprocess the data before analyzing it to ensure the accuracy of the results.

1.3 Proposed Solution

To overcome these challenges, we need to use a combination of quantitative and qualitative analysis techniques and incorporate domain knowledge and expertise. By understanding the King County housing market's complexities and using data-driven insights, we can help real estate agents and property owners make informed decisions about pricing, marketing, and selling their properties, ultimately leading to more successful real estate transactions and a more robust housing market. The proposed solution is to use feature engineering techniques such as normalization and one-hot encoding to preprocess the data and multiple linear regression to model the relationship between the housing prices and the various features in the dataset. We will also explore other relevant models to see which model is the most accurate. The metrics of success will be the accuracy, RMSE, mean absolute error, R-squared and classification report of the models.

1.4 Conclusion

By analyzing the King County housing prices data, we have created this documentation that will help industry professionals make informed decisions. We will be able to identify patterns in the data and model the relationship between the housing prices and the various features in the dataset. This will allow us to make accurate predictions about the future trends in the house ng market.

1.5 PROBLEM STATEMENT

1.5.1 Objectives

Develop a pricing model: Create a predictive pricing model that incorporates the factors identified in the regression analysis, such as the number of bathrooms, living area, lot size, and condition and grade ratings. This model can help the agency more accurately price their properties, particularly those with unique features such as waterfront views.

Refine marketing strategies: Use the insight gained from the regression analysis to refine the agency's marketing strategies. For example, the agency can create targeted marketing campaigns that emphasize the features that have the greatest impact on price, such as the number of bathrooms, living area, and condition and grade ratings.

Identify investment opportunities: Use the regression analysis to identify properties that are undervalued based on their features, such as a high number of bathrooms or a waterfront view. The agency can use this information to identify potential investment opportunities that may offer a higher return on investment.

Analyze seasonal trends: Analyze the seasonal trends in home prices and develop strategies for pricing and marketing homes throughout the year. For example, the agency can adjust their pricing and marketing strategies to take advantage of the higher prices in the spring, when homes tend to sell for more.

2. Data Understanding

In this project, we are analyzing the King County housing market to build a multiple linear regression model that predicts the prices of houses in King County, Seattle. The dataset has 21 variables, including the price, the number of bedrooms, bathrooms, living area square footage, lot square footage, number of floors, whether the house is on a waterfront, quality of view from the house, overall condition of the house, overall grade of the house, square footage of the house apart from the basement, square footage of the basement, year when the house was built, year when the house was renovated, ZIP code, latitude coordinate, longitude coordinate, the square footage of interior housing living space for the nearest 15 neighbors, and the square footage of the land lots of the nearest 15 neighbors.

To better understand the data, we identified the categorical and numerical variables in the dataset that is relevant to our business problem.

Numerical Variables:

"Living area" (sqft_living) is a measure of the total interior living space of the property, which is a key factor in determining its value. "Bedrooms" and "bathrooms" are important because they directly affect the functional capacity of the house, which is an important consideration for prospective buyers. "Location" (zip code) is a critical factor because different zip codes have different levels of desirability and attractiveness to buyers, and this can significantly affect the value of the property. "Age" (yr_built) is important because it provides an indication of the property's overall condition and potential maintenance costs, which can also impact its value.

Categorical variables:

Condition: The condition variable describes the overall condition of the house, which is related to maintenance. This variable can help to identify how the condition of a house affects its price. It has five categories, ranging from 1 to 5, with 1 being the poorest condition and 5 being the best. The condition variable is a categorical variable as it has a limited number of values and cannot be measured quantitatively.

3. Data Cleaning

To make our data suitable, we will do some cleaning by elimination duplicates in the datasets, filling of some of the missing values with the relevant central measures and dealing with outliers to ensure we have a standard normal distribution.

4. Data analysis and Modelling

Based on the values of the independent variables, regression is a supervised learning process used to forecast the value of a dependent variable. In this instance, we're attempting to estimate the impact that various property characteristics have on our dependent variable, the homes' prices. As a result, we will be able to offer our stakeholders a model that can foretell the key characteristics of homes that will have the most effects on their prices. We will also use multiple linear regression because we are working with numerous features. Contrary to linear regression, which only employs one independent variable, multiple linear regression uses the values of many independent variables to predict the value of a dependent variable.

We will first start by building a baseline model. The baseline model will be used to compare the performance of the other models that we will be building. After that, we will build our multiple linear regression model. The target variable is price. Therefore, we look at the correlation coefficients for all of the predictor variables to find the one with the highest correlation with price. We will now iterate the baseline model by building a multiple linear regression model that will have more than one independent variable.

We will use a nonlinear transformation technique Log transformations are one of several different techniques that fundamentally reshape the modeled relationship between the variables The reason to apply this kind of transformation is that we believe that the underlying relationship is not linear. Then by applying these techniques, we may be able to model a linear relationship between the transformed variables

5. CONCLUSION

We will choose one of our models that is the best model and used in our recommendations

5.1 Recommendations

Bathrooms: The number of bathrooms has a positive effect on the price of the house, meaning that houses with more bathrooms tend to be priced higher. The agency may want to consider this factor when pricing and marketing homes with more bathrooms.

Living Area and Lot Size: The size of the living area has a positive effect on the price of the house, while the size of the lot has a negative effect. The agency may want to consider emphasizing the living area in their marketing efforts, while also being mindful of the lot size.

Floors: Houses with more floors tend to be priced higher. The agency may want to consider this factor when pricing and marketing multi-story homes.

Condition and Grade: Houses with higher condition and grade ratings tend to be priced higher. The agency may want to emphasize these ratings in their marketing efforts and pricing strategy.

Age and Renovated: The age of the house and whether or not it has been renovated both have significant effects on the price of the house. The agency may want to consider these factors when pricing and marketing homes, particularly when comparing newer, renovated homes to older ones.

Waterfront View: Houses with a waterfront view are priced significantly higher than those without. The agency may want to emphasize this factor in their marketing efforts for waterfront properties.

Season: The season in which a house is sold can also affect the price, with spring selling for higher prices than fall. The agency may want to consider this factor when planning their marketing and pricing strategies throughout the year.

5.2 Limitation

Some limitations of this model and analysis could include:

Limited variables: While this model includes many important variables that are known to impact house prices, there may be other factors that were not included in the analysis that could also have an effect on house prices.

Assumptions: The model assumes a linear relationship between the independent variables and the target variable. This may not always be the case, and there could be more complex, non-linear relationships between the variables that are not captured in this analysis.

Generalizability: The dataset used for this analysis was limited to a specific geographic area and time period. It may not be representative of other locations or time periods, which could limit the generalizability of the results. The data in the dataset is from 2014 and 2015. Therefore, it may not be able to account for changes in the housing market since then. As a result, the model may not be able to predict the value of a house in 2022.

Causality: While the model can identify relationships between variables, it cannot prove causality. Therefore, it's important to be cautious about making causal claims based solely on the results of this model.

In order to improve the value of a house, we would need to understand the market (i.e., what buyers are looking for). Therefore, by not having this information, we are unable to advise our clients on the best renovations to make. It is possible to build the most expensive house in the world, but if it is not what buyers are looking for, then it will not be sold. There is no value in that.

Next steps

The next steps in this project are to clean and preprocess this data, explore the data through data analysis and build multiple models to predict the housing prices. We will then evaluate the models based on the accuracy, RMSE and classification report to determine the most accurate model. Finally, we will provide recommendations based on the analysis to help in the decision making.

THANK YOU!!