PHASE 2 PROJECT - GROUP 13

OVERVIEW

The idea of this project is to explore the data and find the degree of association between different variables. This is approached by creating visuals/graphs and statistical metrics like correlation coefficients. Finally, using various regression methods to try to identify the locations with the highest sales prices, to identify how seasonal trends affect sales, and to predict prices of houses depending on the features. There are numerous ways to fulfill the purpose and each has its pros and cons. In this project, we chose Multiple Regression analysis because its straightforward, gives more insight about the data and makes easier to interpret the relationship between the response and the explanatory variables.

BUSINESS STAKEHOLDER

The real estate agency

BUSINESS PROBLEM

A real estate agency wants to analyze the factors that influence the prices of houses in order to provide accurate pricing estimates to their clients. The agency aims to understand the relationship between various features of a house, such as the number of rooms, living area, basement area, overall quality, and other relevant factors, and how they affect the sale price.

The clients being, homeowners and potential house buyers have difficulty in making informed decisions regarding property investments, to make this decision, understanding the factors influencing housing prices in a specific area is necessary.

OBJECTIVES

REAL ESTATE AGENCY

- To identify the locations with the highest sales prices.
- To identify how seasonal trends affect sales.
- To predict prices of houses depending on the features.

BUSINESS AND DATA UNDERSTANDING

Business Understanding

The key audience for this analysis comprises real estate agents at the agency that provided the data. These agents work directly with home buyers and sellers. By understanding factors that drive pricing, agents can provide more informed guidance to clients on listing prices for sellers and reasonable offer prices for buyers. Equipping agents with data-driven pricing insights allows them to enhance their credibility as local market experts. This strengthens the agency's competitive positioning and enables them to better meet client needs around pricing. In a nutshell, the core audience is the real estate agents who will leverage these findings to offer more tailored pricing counsel to home buyers and sellers in their market

Data undertanding

Understanding the columns in our data set id - Unique identifier for a house date - Date house was sold price - Sale price (prediction target) bedrooms - Number of bedrooms bathrooms - Number of bathrooms sqft_living - Square footage of living space in the home sqft_lot - Square footage of the lot floors - Number of floors (levels) in house waterfront - Whether the house is on a waterfront view - Quality of view from house condition - How good the overall condition of the house is. Related to maintenance of house. grade - Overall grade of the house. Related to the construction and design of the house. sqft_above - Square footage of house apart from basement sqft_basement - Square footage of the basement yr_built - Year when house was built yr_renovated - Year when house was renovated zipcode - ZIP Code used by the United States Postal Service lat - Latitude coordinate long - Longitude coordinate sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors

MODELING

Several regression models were built to predict house prices and understand the relationships between features: Baseline Model A simple linear regression model was created with only sqft_living as the independent variable and price as the dependent variable. This baseline had an R-squared of 0.497, explaining around 50% of variance in price. Multiple Linear Regression A multiple linear regression model was built adding more features like bedrooms, bathrooms, living space above ground, grade, year built etc. This improved model

performance with an R-squared of 0.650, explaining 65% of variance. Polynomial Regression To capture any nonlinear relationships, polynomial terms up to degree 2 were added for all features. The polynomial model further improved R-squared to 0.719, explaining ~72% of variance in prices. Various regression metrics like R-squared, MSE, MAE were used to evaluate model performance. Target engineering steps like log transforming the price and polynomial transforms helped improve model fits.

REGRESSION RESULTS

The final polynomial regression model had the following key results:

```
R-squared: 0.719
Adjusted R-squared: 0.719
F-statistic: 815.8
P-value: 0.00
```

This indicates that the model explains about 72% of the variance in house prices. The F-statistic shows the overall model is statistically significant. The coefficients provide insight into how each feature impacts predicted price:

```
Living square footage had the highest positive coefficient, indicating it strongly increases predicted price.
Number of bathrooms also had a large positive coefficient.
Interestingly, being on the waterfront had a negative coefficient, suggesting it decreases predicted price after controlling for othe
Higher grades and view ratings increased predicted prices.
```

CONCLUSION

The model provides reasonably accurate estimates of house prices based on the provided features. Further improvements could potentially be made by incorporating neighborhood data or economic indicators. The regression analysis provided important insights into factors driving house prices in King County. Key takeaways include:

```
Square footage of living space, number of bedrooms and bathrooms, overall grade and view are strongly positively associated with pri
Interestingly, waterfront houses appear to sell for lower prices than non-waterfront homes when controlling for other features.
Location plays a major role, with cities like Seattle and Bellevue commanding higher prices.
Prices tend to be higher in spring compared to other seasons.
```

The polynomial models created provide reasonably accurate estimates of house prices based on a set of intrinsic and locational features. The real estate agency can use these models to advise clients on pricing for both buying and selling. They can use the same when building houses for clients.