King County House Price Analysis

Margaret Mitey

Overview

This project uses linear regression analysis to understand how certain variables impact housing prices in King County. Several models will be genereted in order to create the most accurate coefficients and the highest R-squared value.

Business Understanding

- The real estate market is a critical driver of economic growth in many countries globally. It is a dynamic and ever-changing industry where key stakeholders (buyers and sellers) have to make accurate prediction of home prices in order to make informed decisions through use of reliable and comprehensive data.

- Real estate is dramatically affected by its location and factors such as employment rates, economy, crime rates, transport facilities and land rates.

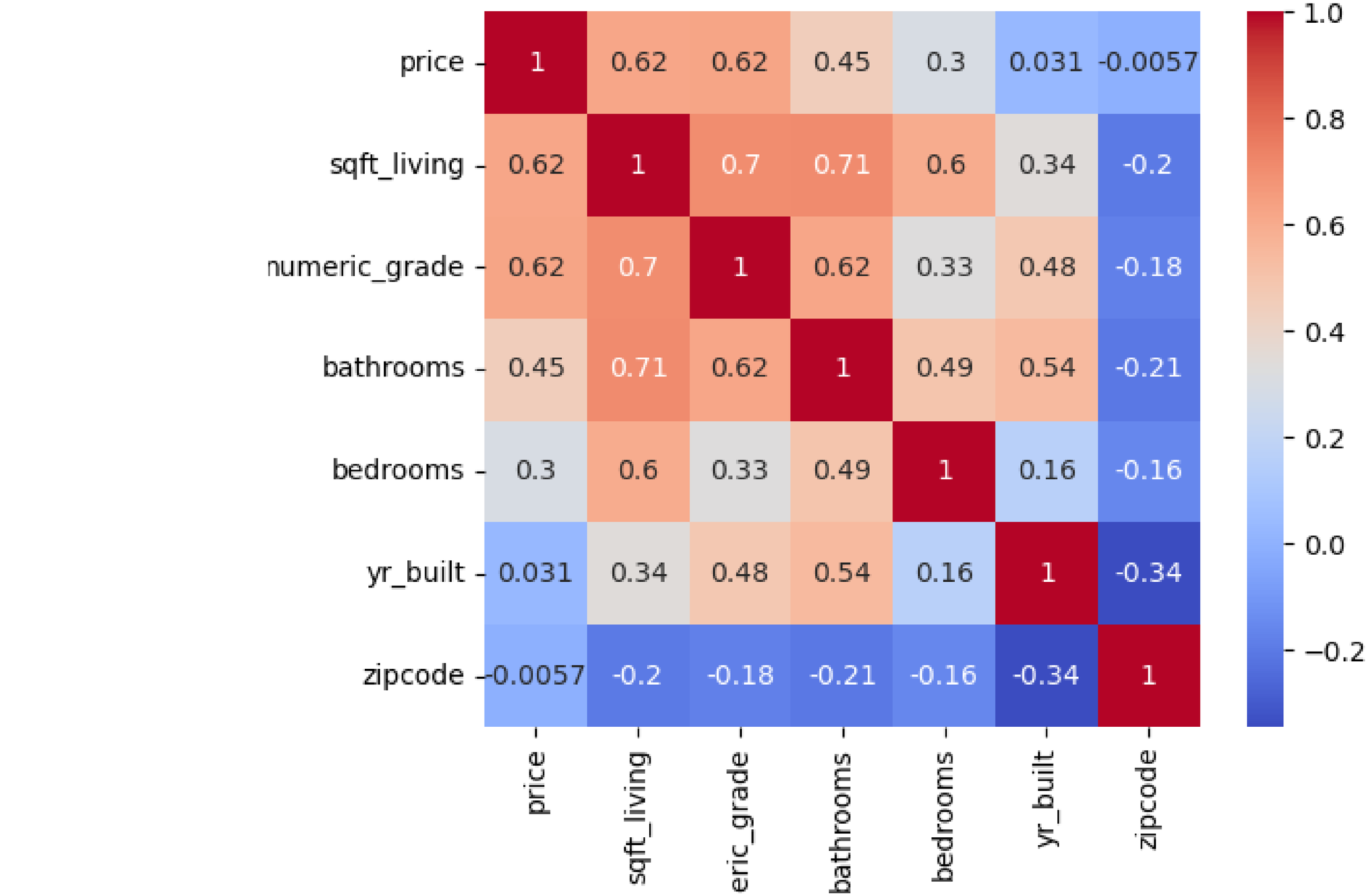
- The King County House Sale dataset is a valuable source of information for analysing house sales in the Northwestern county. The dataset has several variables that can act as key indicators in price of houses in this area such as Zip code, number of bedrooms, number of bath rooms, square feet, numeric grade and the year the house was built.
- The dataset will be used for analysis and modelling to better understand the factor that affects house pricing

Data Exploration

The King County House Sale dataset was used kc_house_data.csv

Modelling

The graph shows that there is a positive linear relationship between square feet of the house and price


price
sqft_living
numeric_grade
bathrooms
bedrooms
yr_built
zipcode
1 0.62 0.62 0.45 0.3 0.031 -0.0057
0.62 1 0.7 0.71 0.6 0.34 -0.2
0.62 0.7 1 0.62 0.33 0.48 -0.18
0.45 0.71 0.62 1 0.49 0.54 -0.21
0.3 0.6 0.33 0.49 1 0.16 -0.16
0.031 0.34 0.48 0.54 0.16 1 -0.34
-0.0057 -0.2 -0.18 -0.21 -0.16 -0.34 1
price
sqft_living
eric_grade
bathrooms
bedrooms
yr_built
zipcode
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

From our heatmap in the previous page, we can see that the square feet living, numeric grade and bathrooms had higher figure/collinearity with price because of the darker shade in those variables.

Interpretation of model with 6 predictors

1. The model explains a 56% of the variance in price which shows an increase compared to the first model which had a variance of 38% with only one predictor (sqft_living).

2. The model is statistically significant overally, with an F-statistic p-value below 0.05.

3. The model coefficients (const, sqft_living, numeric_grade, bathrooms, bedrooms, yr_built and zipcode) are all statistically significant, with t-statistic p-values of 0.000 which is below 0.05

4. For each increase of 1 sqft in sqft_living, there is an increase in price of about $103

Interpretation of model with 6 predictors

5. For each increase in numeric grade, there is an increase in price of about $114400

5. For each increase of 1 bathroom, we see an associated increase in price of about $45730

6. For each increase of 1 bedroom , we see an associated decrease in price of about $20580

7. For a change in location zipcode, there is an increase in price by $167

RECOMMENDATIONS

1. The real estate agency should advice their clients that the pricing of houses increases with each sqft increase of the house and they should showcase this.
2. Houses with higher numerical grade fall in high price category. Home owners should be advised to improve the grade of their houses so at to increase their value.
3. There are certain zipcodes that are more expensive than others as the price of houses increases by $167. The real estate agency should highlight this areas to home buyers and sellers.
4. Buyers who are looking for high-end properties may not necessarily prioritize the number of bedrooms. Therefore, sellers should be aware that adding additional bedrooms may not necessarily increase the value of the property as houses with more bedrooms have low prices.

CONCLUSION

This project is aimed to develop a model to predict housing prices in King County based on various features such as square footage, number of bedrooms and bathrooms, grading and zipcode.

Firstly, exploratory data analysis (EDA) was performed and found that the price o houses was positively correlated with the square footage. The location of the house also had a significant impact on the price.

Secondly, multiple linear regression models was used to predict the price of the house. The model included square footage, grade, bedrooms, bathrooms and zipcode as the predictors performed reasonably well, explaining about 56% of the variance in price.

Based on the results, we can conclude that square footage, the numeric grade and zipcode of the house are significant predictors of housing prices in King County.

A multiple linear regression was suitable for this project as it evaluated the dependent variable and several independent variables so as to access their strength of association.

Next Step

In this project, we were only looking at the coefficients in order to give our recommendations. Further research and indepth analysis can be done to improve the accuracy of the models and to gain a better understanding of the factors that influence housing prices in King County.