King County House Sales Project

Nicholas Muchiri Kinyua

Introduction

The following project analyzes the housing sales of King county. It will try to compare how prices are influenced by different features e.g., bathrooms, bedrooms and such. This features will be tested side by side with the price to see their proportion against each other i.e does the increase in feature A lead to increase/decrease in the price. Does it even have an effect to start with. This will enable us to give the stakeholder the visuals and the recommendations to enable him or her to make the necessary conclusions.

Problem Statement

The problem statement is to give the stakeholders real estate data which will enable them know where to invest, how to renovate based upon previous patterns which have already occured in the northwestern county real estate business. There is need to identify the trends there in the past and their consequences e.g when the house was put up in front of a water body(waterfront), how fast did it find a customer. Does putting it in front of a waterfront have any positive impact?

Main Objective

The main objective is to select the features which influence the price of a house mostly and then visualize them and put them on a regression scale based on the past data to construct a predictive model to identify how in the future, what will be influencing the real estate business.

Specific Objectives

To identify the most correlated features with the price to create a multilinear regreesion to help us in knowing how the features influence the prices and even influence the customers.

To build a baseline model to identify the most relevant data to be used as the starting point of the analysis.

To use the linear regression metrics to have the appopriate coefficients which will enable us to come up with the relevant recommendations for the stakeholders of the real estates

Notebook Structure

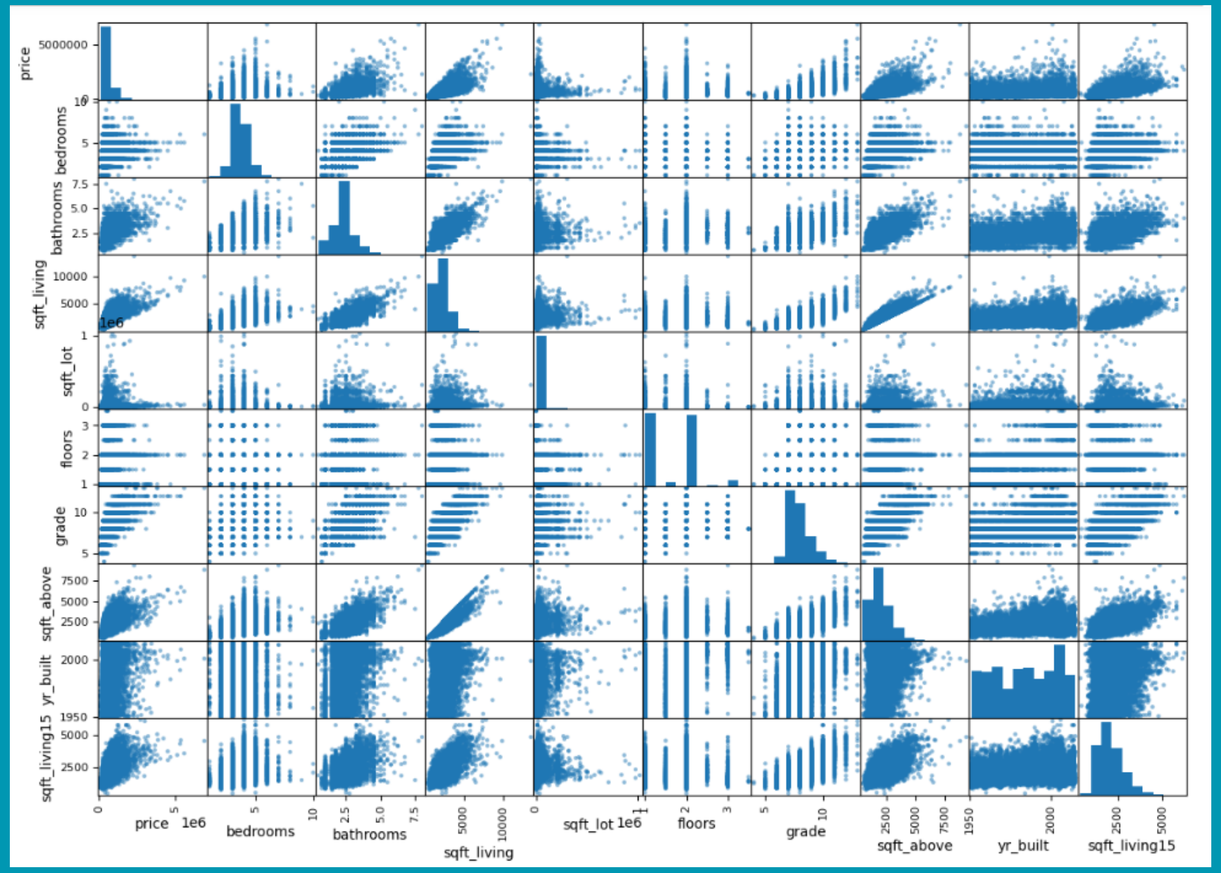
Introduction
Problem Statement
Main Objective
Specific Objectives
Importing Libraries
Data Understanding
Data Cleaning
Modelling
Regression Results
Data Visualizations
Interpretations
Recommendations and Conclusions

Data Understanding

'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

Data Visualizations

How different features correlate with price



Modelling

Multi-linear regression

Dep. Variable:	price	**R-squared:**	0.611
Model:	OLS	**Adj. R-squared:**	0.611
Method:	Least Squares	**F-statistic:**	5160.
Date:	Sun, 09 Jul 2023	**Prob (F-statistic):**	0.00
Time:	17:57:50	**Log-Likelihood:**	-42487.
No. Observations:	16413	**AIC:**	8.499e+04
Df Residuals:	16407	**BIC:**	8.503e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	57.7890	0.661	87.468	0.000	56.494	59.084
bathrooms	0.0410	0.012	3.431	0.001	0.018	0.064
sqft_living	0.4743	0.012	39.068	0.000	0.450	0.498
sqft_lot	-0.0132	0.003	-4.110	0.000	-0.019	-0.007
floors	-0.0765	0.009	-8.687	0.000	-0.094	-0.059
grade	1.8358	0.028	64.873	0.000	1.780	1.891

Omnibus:	232.746	**Durbin-Watson:**	1.970
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	267.674
Skew:	0.243	**Prob(JB):**	7.50e-59
Kurtosis:	3.393	**Cond. No.**	3.17e+03

Modelling

Recursive feature selection

selected features: Index(['bedrooms', 'sqft_living', 'grade'], dtype='object')

Regression model evaluation

Mean Squared Error (MSE): 10.3957
R-squared: 0.6285
Mean Absolute Error (MAE): 2.5568

Regression Interpretation

Adjusted rsquared of 0.614 means that the features can explain the houses prices by approximately 0.62 units. The better since the more the figure is near one the better.

F-statistic of 4354 means there is a strong relationship between the features and the prices. The model is wholly significant.

coefficient of bathrooms is 0.0563 means that the addition of a single bathroom (while the other features are constant) will lead to the increase of price by 0.0563 unit.

coefficient of sqft_living is 0.5434: Just like bathrooms, when there is an increase of one unit of sqft_luving, then there is an increase of 0.5434 unit of the price.

coefficient of grade is 1.7646 means one unit change for grade e.g good to better, will mean an increase of 1.7646 unit of the price.

Selected features are bedrooms, square foot living area and the grade hence they are the most important

Recommendations

Most features have been noted to have statistical significance on the price of the houses. In particular, these three features are very important and will play a very crucial role in the future in determining the price of the houses: Bedrooms, Grade and Living area

Also, the stakeholders should keep a close eye on the: bedroom, living area and grade of a house while making kings county housing investments later on.

a. The increase in the number of bathrooms means the increase in price of a house.

b. The better the grade of the house i.e average > good > excellent better means the the more expensive a house will be.

c.The more the bedrooms means the higher the price of a house becomes e.g a three bedroom house is more expensive than a two bedroom house.

Finally, the assumptions being correct means the data provided above is reliable and can be used to predict future house prices.

Next Steps

Things may change in the future so continously updating the model from time to time can be important too. e.g

1. Data drift - to ensure data remains accurate
2. Model performance - Incorporating fresh data to improve performance of the models
3. Adapting new business requirements because they may change with time

Thank You!