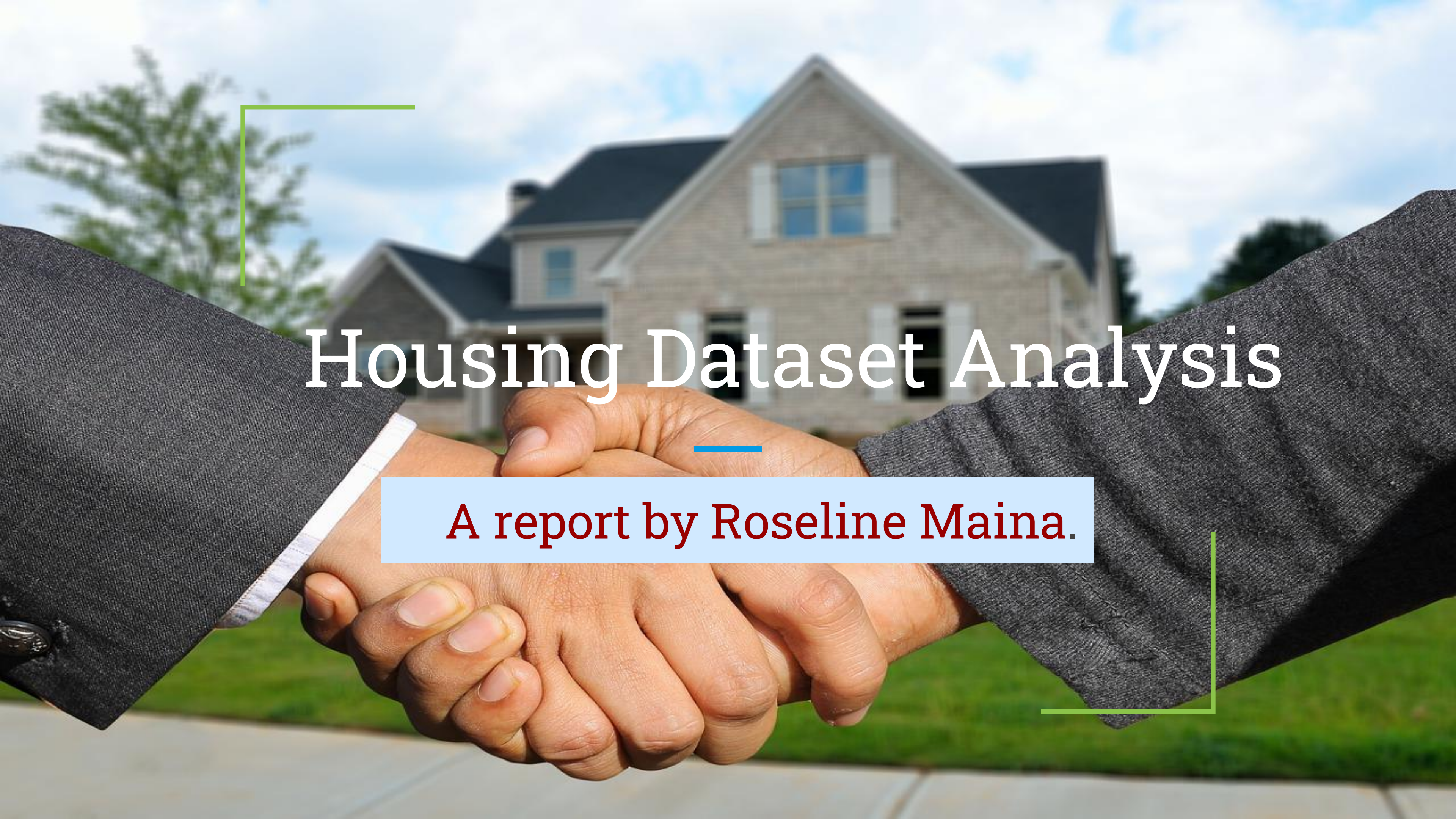


Housing Dataset Analysis

A report by Roseline Maina.

Table Of Content

Business Problem

Data Understanding

Data Preparation

Data Exploration

Modeling

Evaluation

Business Problem

- Our client is a real estate agency situated in King County, Washington that helps homeowners buy / or sell homes.
- They are looking to get a better understanding on what features about a house are the most important when trying to estimate a home's price in that area so as to know how a shift in the features may affect the pricing of the houses.
- They also what you to come up with a pricing algorithm to give an estimate of a home's price based on different features to assist buyers and sellers evaluate their homes.

Data Understanding

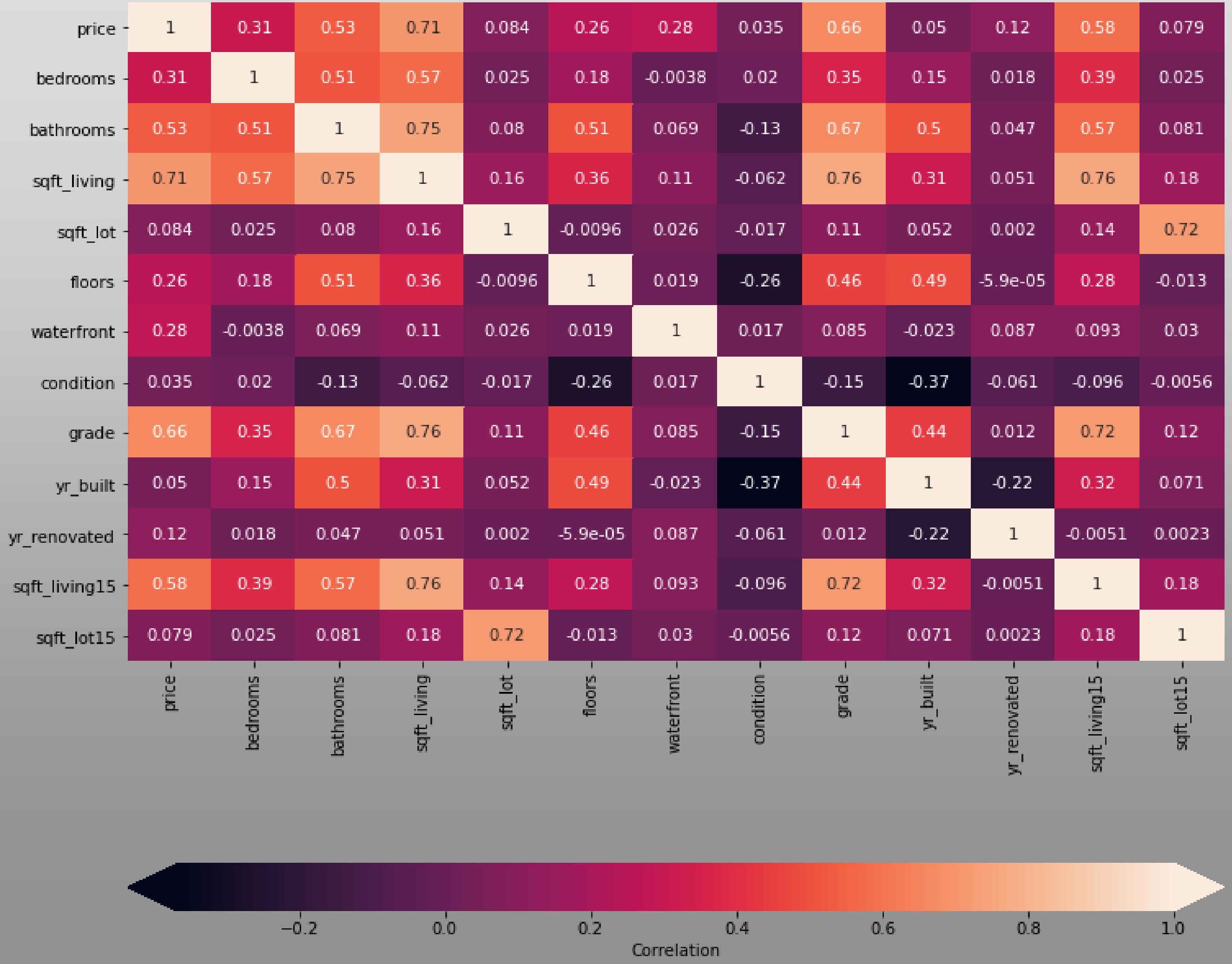
- We have been provided access to data containing information on over 10,000 homes together with their respective features.
- The data used in the analysis contains information on the King County homes pricing together with the different features that make up the homes. The data is contained in the Data folder where:
  - kc_house_data.csv contains data on the different houses together with their features
  - column_names.md contains a breakdown on the Column Names and their descriptions for Kings County Data Set
- The data modeling will be broken down in two parts where the first bit will focus on the inferential modeling and the second bit will be on predictive modeling.

Data Preparation

- After loading the data and familiarizing ourselves with the data in the previous stage, at this stage we are going to process our data and prepare it for analysis.
- Some of the tasks that i carried out in this stage was the data cleaning, this was done by checking the data for missing values, outliers, checking if the data columns were assigned the right data type and finally checking the data for multicollinearity.
- Below is a brief overview on how i went over checking data for multicollinearity:

- Multicollinearity tells us the relationship between two predictors. In our case, the ‘price of a home was our target’ was the target and the the homes features were the predictors.
- If two features are highly correlated with each other and also highly correlated with the target it will be hard to distinguish the effects of one feature on the target and the other feature on the target.
- This is a problem as it reduces the performance of our model.
- I used correlation to show the relationship between the features. The heatmap on the side provides a visual summary of the same:

Heatmap of Correlation Between Attributes



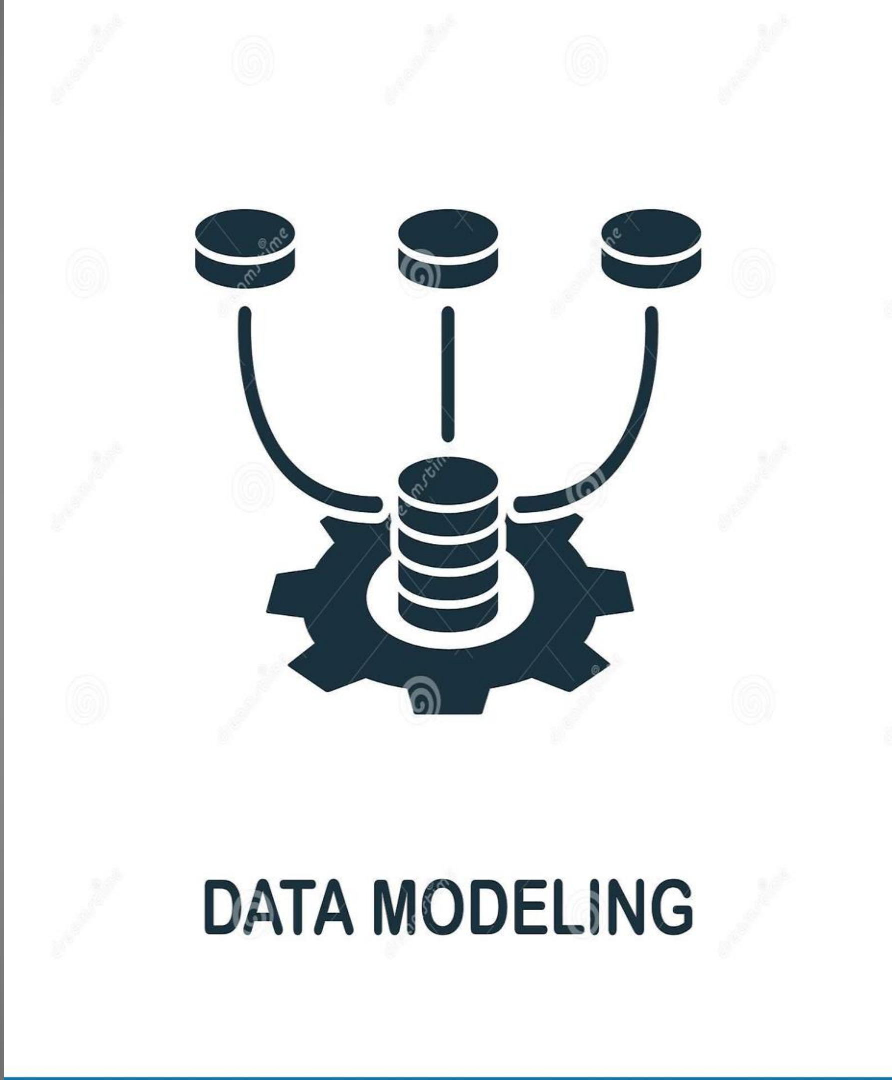
	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	condition	grade	yr_built	yr_renovated	sqft_living15	sqft_lot15
price	1	0.31	0.53	0.71	0.084	0.26	0.28	0.035	0.66	0.05	0.12	0.58	0.079
bedrooms	0.31	1	0.51	0.57	0.025	0.18	-0.0038	0.02	0.35	0.15	0.018	0.39	0.025
bathrooms	0.53	0.51	1	0.75	0.08	0.51	0.069	-0.13	0.67	0.5	0.047	0.57	0.081
sqft_living	0.71	0.57	0.75	1	0.16	0.36	0.11	-0.062	0.76	0.31	0.051	0.76	0.18
sqft_lot	0.084	0.025	0.08	0.16	1	-0.0096	0.026	-0.017	0.11	0.052	0.002	0.14	0.72
floors	0.26	0.18	0.51	0.36	-0.0096	1	0.019	-0.26	0.46	0.49	-5.9e-05	0.28	-0.013
waterfront	0.28	-0.0038	0.069	0.11	0.026	0.019	1	0.017	0.085	-0.023	0.087	0.093	0.03
condition	0.035	0.02	-0.13	-0.062	-0.017	-0.26	0.017	1	-0.15	-0.37	-0.061	-0.096	-0.0056
grade	0.66	0.35	0.67	0.76	0.11	0.46	0.085	-0.15	1	0.44	0.012	0.72	0.12
yr_built	0.05	0.15	0.5	0.31	0.052	0.49	-0.023	-0.37	0.44	1	-0.22	0.32	0.071
yr_renovated	0.12	0.018	0.047	0.051	0.002	-5.9e-05	0.087	-0.061	0.012	-0.22	1	-0.0051	0.0023
sqft_living15	0.58	0.39	0.57	0.76	0.14	0.28	0.093	-0.096	0.72	0.32	-0.0051	1	0.18
sqft_lot15	0.079	0.025	0.081	0.18	0.72	-0.013	0.03	-0.0056	0.12	0.071	0.0023	0.18	1

Data Exploration

- In this stage of our analysis, we are going to explore and visualize data to uncover insights.
- The first stop in our data exploration was identifying which two features are the most correlated with the price. Based on my analysis the two most correlated features were:
  - Sqft_living with 0.7065
  - Grade with 0.6636
- Next, i checked the distribution of the prices and also checked the features to identify out categorical and numerical variables.

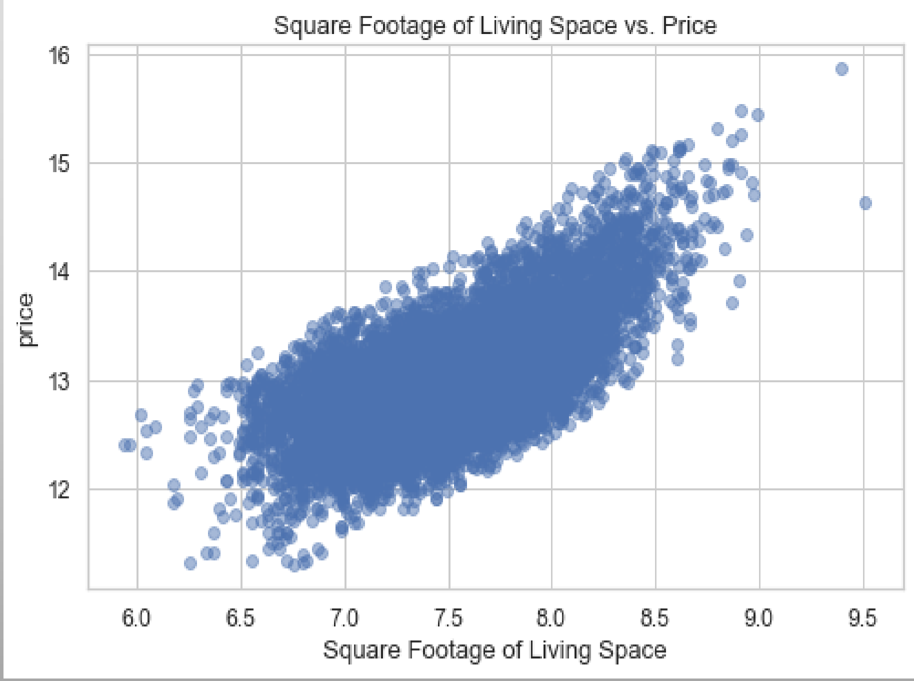
Modeling

- This is the main focus of our analysis.
- Here, we get to build the linear regression models that will help us in finding the solution to our business problem.
- We are going to cover two kinds of data modeling which is:
    - Inferential Modeling
    - Predictive Modeling
- In the coming slides we'll get to discuss both in depth.



➢ Inferential Modeling

- Inferential models help us understand the relationship between the target(price) and the predictors(home features).
- Using the preprocessed data, i started off by finding the home feature that shares the strongest relationship with the price this was sqft_living.
- The figure on the side shows the relationship between the two. It shows a positive relationship between the two variables.
- Next, i fitted a linear model between the two variables and got a r-squared value of 0.453 which meant that about 45.3% variation in the price is explained by the model.



- I followed it up by fitting a linear regression model between the target(price) and the five most correlated features(sqft_living, bathrooms, grade_7, bedrooms, grade_10).
- This model provided a r-squared value of 0.492 showing thet about 49.2% variation in the price is explained by the model. This is a better score than the one provided by the first model.

➢ Predictive Modeling

- After building our inferential models, i followed it up by fitting the predictive models.
- Predictive models helps us in making future prediction and in our case it will help clients estimate the prices of homes based on the homes features in situations where by they want to sell or buy homes.
- In this section i came up with two predictive models. One was the Linear regression model which when tested on the data showed that about 70% of the variation is the price is explained by the homes features
- The other was the polynomial model which when tested on the data also showed that about 80% of the price is explained by the homes features.
- Therefore, the polynomial model is better when it comes to predicting a homes price.

Evaluation

- In this section we get to go over the the results of our analysis and give recommendations.
- Based on the results we can conclude that the most correlated feature with the price is the sqft_living with a coefficient of 0.8339 this means that a one unit change in the a home’s square footage increases the price by 0.8339.
- The second most correlated feature with the price of a home is the home grade with homes of grade_10 and grade_7 having the highest correlation with the price.Homes with a grade of 10 had a coefficient of 0.2583 this means that a home simply being of grade 10 increases the price by 0.2583. However, homes of grade 7 had a negative coefficient of about -0.1114 this shows that a home simply being of grade 7 reduces the price by 0.1114 therefore i believe further investigation needs to be done to discover the underlying reasons behind it.
- We also discovered that a polynomial model will be the better to use when we want to makes predictions on a homes price based on its features as it shows that 80% of the variation in the prices is explained by the model hence it will give a better price estimate.
- In conclusion, i recommend that stakeholders pay key attention to the grade and square footage of a home when it comes to determining the price of a home as the two features are the most correlated with the price.

THANK
YOU

Q&A