Phase 2 Project Description

Project Overview

For this project, I used linear regression to show how certain house characteristics impact the price of a house in King County. I also used t-tests and ANOVA to analyse the differences in prices for specific groups of houses.

Business Problem & Stakeholder

A real estate firm in King County wants to win new customers by increasing the transparency when it comes to the sales price of houses. This applies on the one hand to customers who want to sell their homes to give them an early indication for how much money houses with comparable characteristics sell. On the other hand, it will give customers who plan to buy a home an indication how much budget they need to plan with. The real estate firm specifically wants to look at the following things:

What are the most important characteristics of a house that impact the sales price? Are renovated houses more expensive than houses without renovation? Do houses with a waterfront sell for more money? Do houses with a basement sell for more money?

The Data

This project uses a subset of the King County House Sales dataset. As there are objects in the dataset that are outside of King County, we exclude those. Also, we excluded some outliers and detailed information. Each row represents a house that was sold in King County and includes information about the sale and the object. In our model, we will look at the following information:

- Age of the property
- Construction grade
- Sqft above
- Basement (yes/no)
- Renovated (yes/no)
- Waterfront (yes/no)
- Greenbelt (yes/no)
- Location of the object
- Sales price

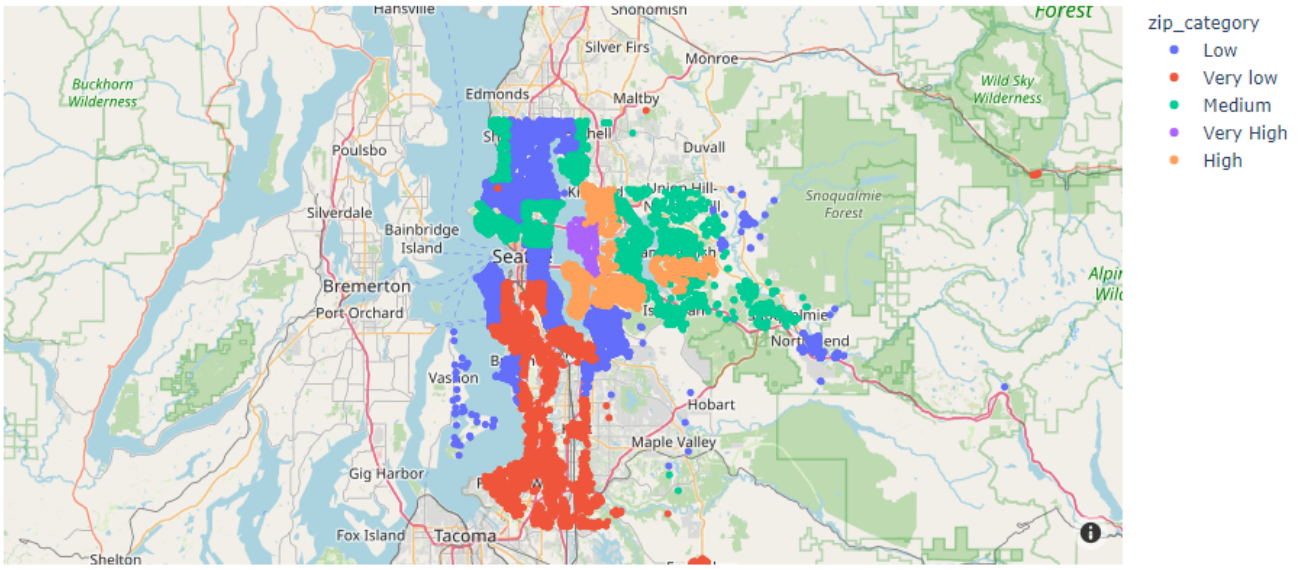
Approach

In this project, I took the following steps:

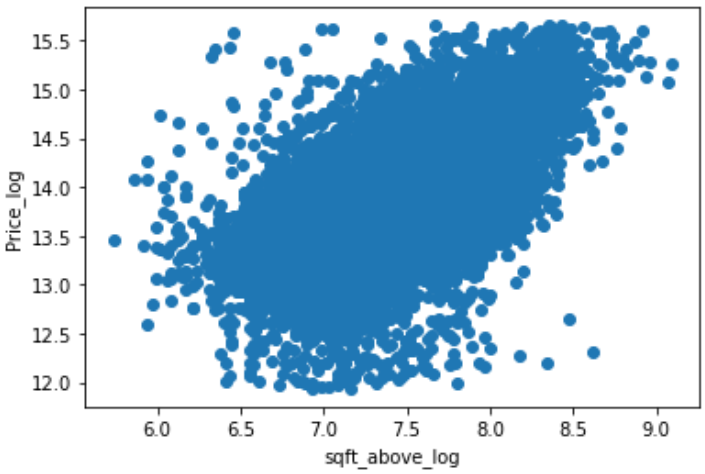
- Loading and combining data
- Check for data completeness and overall structure
- Inspect all variables visually to determine relevance for future model
- Prepare data for modelling (one hot encoding, transformation)
- Build and improve model
- Interpret model
- Confirm results with t-tests and ANOVA

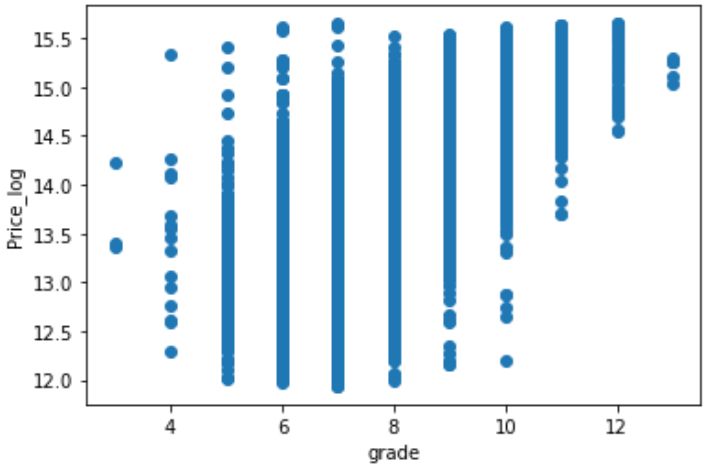
Analysis

Usually, the location of the property and the neighbourhood are of high importance when it comes to house prices. Therefore, I first looked at the different prices per zip codes on a map. After removing some heavy outliers and transforming the price variable, we can see that the neighbourhood around Bellevue and Mercer Island seem to have the highest pricing houses. As there seem to be differences in the zip codes, I clustered them in 5 price categories.

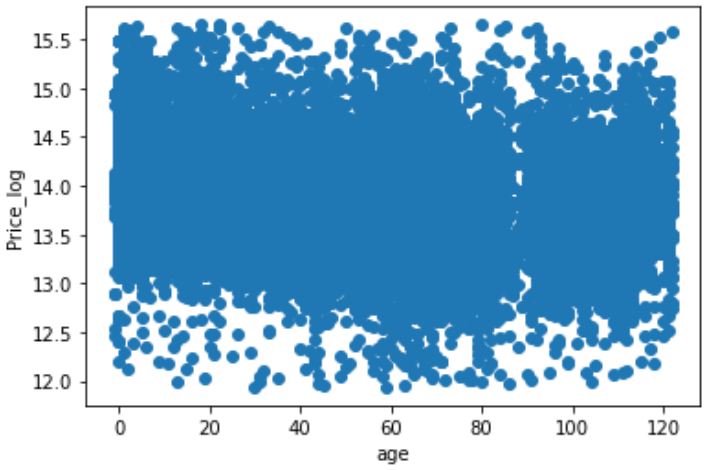


Looking at the histogram of my numeric variables, we can see that after log transforming, the price and squarefoot living/above look normally distributed. Bedrooms and bathrooms have several outliers even though the majority have around 3 bedrooms and 3 bathrooms. Floors, condition and age of the property are not normally distributed. Further looking at the scatter plot, the most promising variables seem to be grade, sqft above and sqft living.





Surprisingly, age of the object does not seem to be correlated with the price.



I dropped sqft living after detecting an issue with multicollinearity between sqft living and sqft above.

Modeling

For my customer, I want to be able to predict house prices. A common statistical technique is linear regression. As I have many variables that seem to be promising, I will build a multiple linear regression model with both categorical and continuous variables. The benefit of linear regression is that it is easy to implement and to interpret. We will also look at correlations but with our regression we can give our customer a clearer indication of how the price changes depending of our independent variables.

Based on the scatter plots and correlations with log price, I excluded the following variables from the start: Nuisance, Age, condition, floors and sqft living (due to multicollinearity). My baseline model includes bedrooms, bathrooms, grade, sqft above (log) and the categorical dummy variables for waterfront, basement, garage, patio, greenbelt, view, heat source, renovated and zip categories. The dependent variable is price (log) The baseline model is significant and explains roughly 70% of the variance. Several variables do not make sense (negative price development for having a garage) or are not significant. Therefore, I've iterated through several models and ended up with the 6th version of it:

OLS Regression Results

Dep. Variable:	price_log	R-squared:	0.687
Model:	OLS	Adj. R-squared:	0.687
Method:	Least Squares	F-statistic:	3947.
Date:	Tue, 21 Feb 2023	Prob (F-statistic):	0.00
Time:	08:30:51	Log-Likelihood:	-4186.7
No. Observations:	16207	AIC:	8393.
Df Residuals:	16197	BIC:	8470.
Df Model:	9		
Covariance Type:	nonrobust		

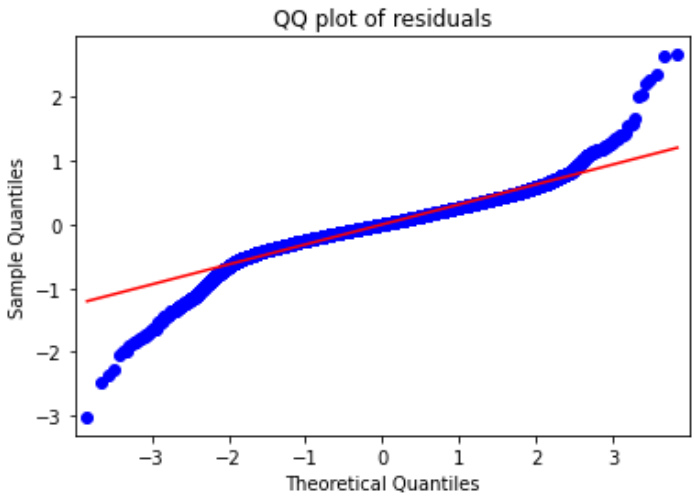
	coef	std err	t	P>|t|	[0.025	0.975]
const	10.0961	0.049	204.468	0.000	9.999	10.193
grade	0.1083	0.003	34.162	0.000	0.102	0.115
sqft_above_log	0.4044	0.008	49.478	0.000	0.388	0.420
waterfront_YES	0.4280	0.023	18.663	0.000	0.383	0.473
basement_YES	0.1893	0.005	35.002	0.000	0.179	0.200
renovated_YES	0.0985	0.011	9.136	0.000	0.077	0.120
zip_category_High	0.2230	0.010	21.271	0.000	0.202	0.244
zip_category_Low	-0.2222	0.007	-33.703	0.000	-0.235	-0.209
zip_category_Very High	0.5922	0.020	29.844	0.000	0.553	0.631
zip_category_Very low	-0.6086	0.008	-77.555	0.000	-0.624	-0.593

Omnibus:	2560.250	Durbin-Watson:	1.946
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24737.794
Skew:	-0.462	Prob(JB):	0.00
Kurtosis:	8.981	Cond. No.	218.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The model is significant and explains 68% of the variance. My target variable is in log dollars so if all features are 0, the house would sell for around 22000 USD (exp(10)). For my categorical values, my reference values are: zip_category medium, basement no, waterfront no. For the most impactful numerical values, we can say that with each grade increase, the price will increase by 1.1%. Also, for 1% increase in sqft above, the price will increase by 1.49%. Exp(0.4) The location of a property depending on the zip code also has an impact on price. Compared to zip codes with medium average house prices, the prices drop significantly for low and very low zip code categories and rise significantly for high and higher zip codes. This fits to what we have seen in the zip code map where there are significant differences based on the region.

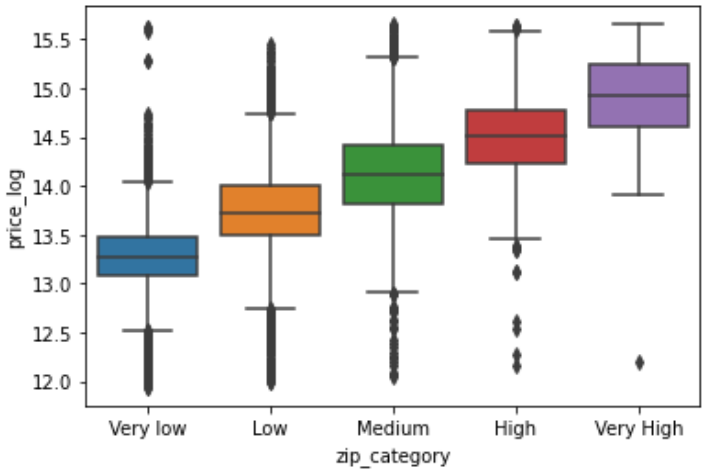
A limitation of the model is that the residuals do not seem to be normally distributed. The qq-plot and histogram of the residuals show this as well as the Shapiro-Wilk test which had a p-value of 0.00



The peak is quite high in the middle and I have long tails in the distribution. This could mean that the linear regression may not be the best model to use. When I calculate the Mean Absolute Error which is more appropriate for log transformed dependent variables, it indicates that my model is off by 23% of the geometric mean of the actual and predicted values.

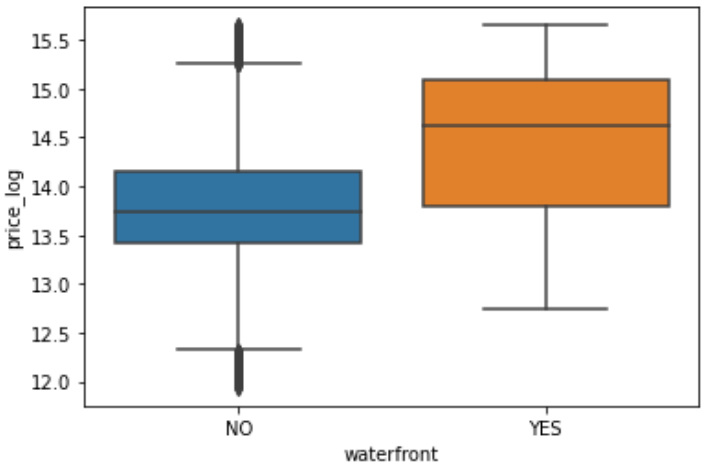
Further analyses

In this section I will test different hypotheses with t-test or anova: H1: The price is different depending on the neighbourhood you live in. H0: There is no difference depending on where you live.



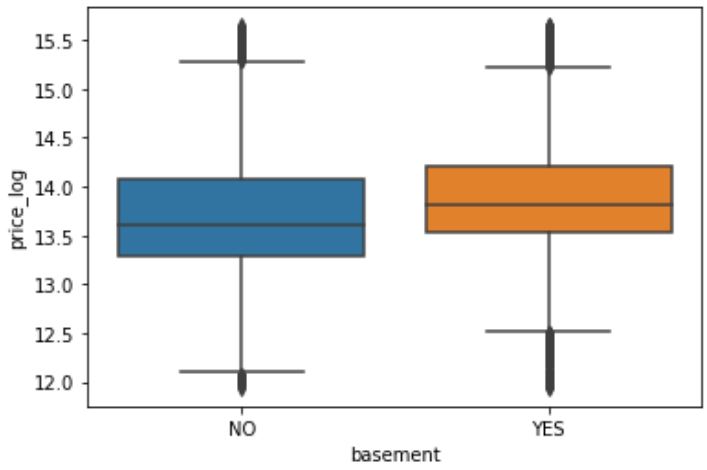
ANOVA shows that there is indeed a price difference depending on the neighbourhood you are in. Therefore, rejecting the null hypothesis. The prices vary from a mean of 580.0000 USD in the very low zipcodes and 2.9 mio USD in the very high zipcodes

Hypothesis 2: The price of a house is higher if you have a waterfront H0: The price is less or equal when you have a waterfront



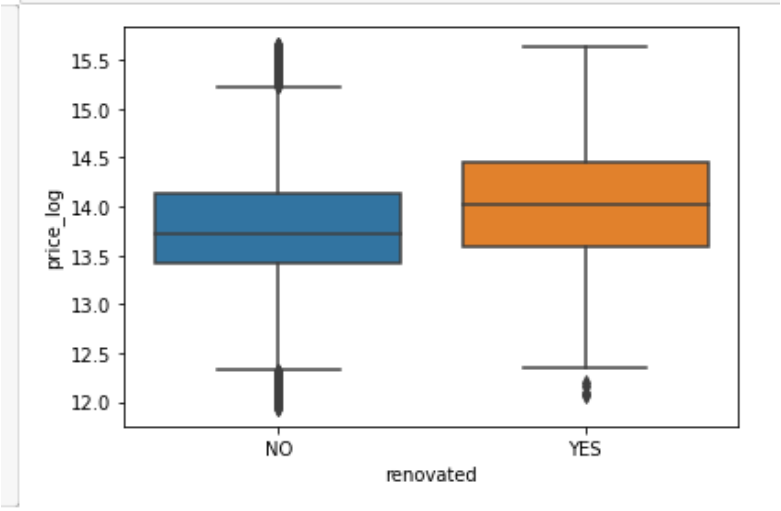
Rejecting the null hypothesis. If you have no waterfront, the mean price is exp(13.79) which are 974000 USD. If you have a waterfront, the mean price is exp(14.45) which is 1.88 Mio USD. In the boxplot you can see, that even with no waterfront, you have a lot of outliers on both sides.

Hypothesis 3: The price of a house is higher if you have a basement H0: The price is less or equal when you have a basement



Rejecting the null hypothesis. If you have no basement, the mean price is exp(13.69) which are 882000 USD. If you have a basement, the mean price is exp(13.88) which is 1.06 Mio USD. Compared to the waterfront, the price difference is not that large which makes sense.

Hypothesis 4: The price of a house is higher if you have renovated H0: The price is less or equal when you have renovated



Rejecting the null hypothesis. If you have not renovated, the mean price is exp(13.79) which are 974000 USD. If you have renovated, the mean price is exp(14.01) which is 1.2 Mio USD.

Summary

With the model, I have identified several factors that are important in determining the price of a house. I found out that houses with a waterfront, basement and which have been renovated sell for more than houses without those attributes. Nevertheless, my model has limitations and should be further improved. It is better in predicting houses with average prices and is off significantly for comparably low and high house prices.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%