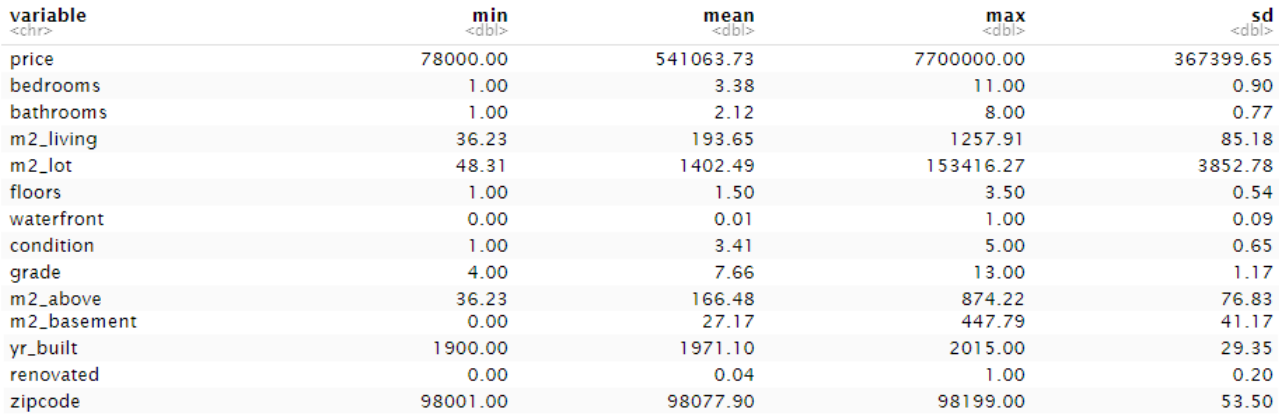
Google colab link:
https://colab.research.google.com/drive/1JeTp7sTLlDc9HUYlqRCO51T1suHK0QwC?usp=sharing

Group Assignment 1 Stakeholder Report

This report will analyze the housing sales in King County USA from May 2014 to May 2015. The data includes variables such as price, square feet of living space, number of bedrooms and bathrooms, condition etc. Using techniques from machine learning, the report will explore the characteristics of the data and model price determination. First of all, the report will present some short summaries of the data, this is followed by a section that clusters the estates into five groups, and lastly a model is constructed which is able to predict the house prices.

Summary Statistics

The full dataset consists of 21 variables, there the 14 most important are chosen for further work. This dataset is cleared of some few unreasonable observations, and only contain estates that at least consists of one bedroom and one bathroom. In the table below minimum-, mean- and maximum value with the addition of standard deviation is shown. Thereby it is possible to get a sense of the data structure.



variable	min	mean	max	sd
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
price	78000.00	541063.73	7700000.00	367399.65
bedrooms	1.00	3.38	11.00	0.90
bathrooms	1.00	2.12	8.00	0.77
m2_living	36.23	193.65	1257.91	85.18
m2_lot	48.31	1402.49	153416.27	3852.78
floors	1.00	1.50	3.50	0.54
waterfront	0.00	0.01	1.00	0.09
condition	1.00	3.41	5.00	0.65
grade	4.00	7.66	13.00	1.17
m2_above	36.23	166.48	874.22	76.83
m2_basement	0.00	27.17	447.79	41.17
yr_built	1900.00	1971.10	2015.00	29.35
renovated	0.00	0.04	1.00	0.20
zipcode	98001.00	98077.90	98199.00	53.50

Table 1: Summary statistics

The table shows a wide price range, with a fairly high variation. The square meters of living space, which is calculated from the square feet of living space, is more concentrated around the mean of 193,65 m2.

The data has been examined in regard to connections between variables. This shows some potential relations in the dataset:

- Waterfront estates have, on average, a higher price.
- A tendency towards more bathrooms is associated with a higher price.
- The most expensive estates do not have the highest number of bedrooms.
- Square meters of living space seem to be positively related to the price.

These relations are somewhat logical and will be investigated further in the last part of this report.

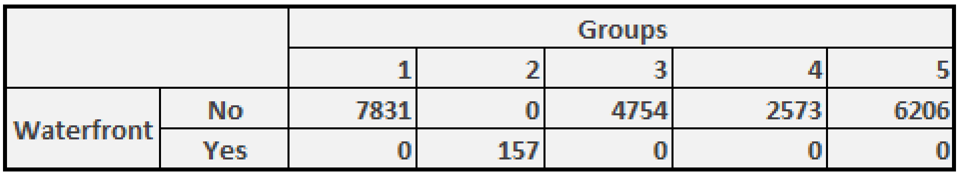
Google colab link:
https://colab.research.google.com/drive/1JeTp7sTLlDc9HUYlqRCO51T1suHK0QwC?usp=sharing

Cluster Analysis

This section will cluster the data into five groups. The number of groups is determined based on statistical properties. The clustering sorts the data, in such a way that there are uniform characteristics in each group. This analysis does not give definitive answers, but underlines tendencies in the data, which can be used to examine questions like; which neighborhoods are most expensive? is location influential on the price? how important are grades? etc. This analysis can however be used by real estate agents to define similar estates to customers that show interest in a particular kind of house. These five groups can be labeled in the following way:

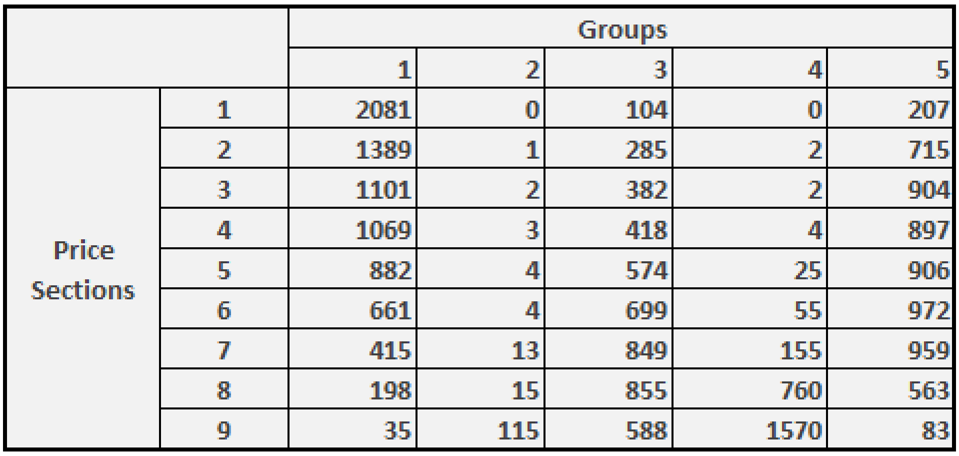
1. Cheaper estates
2. Waterfront estates
3. Expensive neighborhood estates
4. High-end estates
5. Middleclass estates

The labels highlight a defining feature of each group. To determine these labels, each group are examined according to the variables in the dataset. To do this tables are made which compares the group with some variable. Here there will be a couple of examples of these tables:



		Groups				
		1	2	3	4	5
Waterfront	No	7831	0	4754	2573	6206
	Yes	0	157	0	0	0

Table 2: Clusters compared to waterfront



		Groups				
		1	2	3	4	5
Price Sections	1	2081	0	104	0	207
	2	1389	1	285	2	715
	3	1101	2	382	2	904
	4	1069	3	418	4	897
	5	882	4	574	25	906
	6	661	4	699	55	972
	7	415	13	849	155	959
	8	198	15	855	760	563
	9	35	115	588	1570	83

Table 3: Clusters compared to price

Table 2 shows how the clusters clearly separates all waterfront estates. The next table present the clusters in according to house prices. Here the prices are split into nine sections, there the first section represent the lowest prices and the ninth section represent the highest prices. There is a clear pattern that shows that location increases price, group 2, 3 and 4 are generally high in price and based on the analysis they also generally lie in expensive neighborhoods or on the waterfront. Group 3 is interesting, because these houses typically fall into relatively lower grades and therefore give a strong indication of the location impact on prices.

Modelling House Prices

This section will model house prices in King County USA so that it is possible to predict housing prices if all variables in the dataset are known. The prediction will always be associated with some uncertainty. To ensure the predicted prices are somewhat consistent there will be used supervised machine learning. This procedure fits some models on a subset of the data, picks the best model and thereafter predicts the prices on the reminding data. This ensures that the predictions from the model are reasonable, even though there is no actual data of these predictions. In this case the models have been fitted on 75% of the dataset (training data) and tested on the remaining 25% (test data). The model will always perform the best when predicting the data from which it has been estimated. Because of the richness of the dataset, combined with the right type of model, the predictions will however be rather precise when utilized on other data, if the characteristics of the area are similar. This is shown in this visualization:

Google colab link:
https://colab.research.google.com/drive/1JeTp7sTLlDc9HUYlqRCO51T1suHK0QwC?usp=sharing

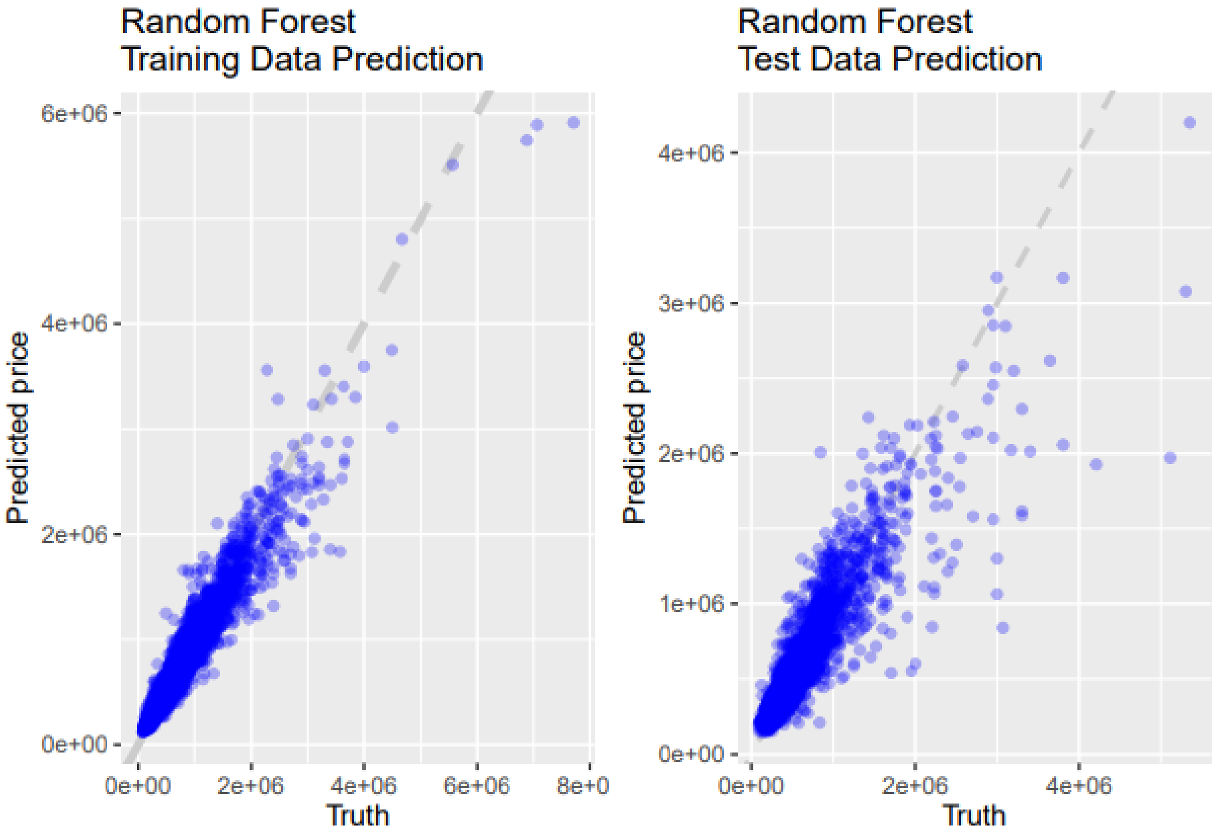


Figure 1: Predictions compared with true values of price

The left-hand side shows the predictions based on the training data. The right-hand side shows the predictions based on the test data. The grey dashed line shows perfect predictions. It can be observed that the model loses predictive power, when used on data outside of the training dataset, as expected. Even though there is a loss in the accuracy of the predictions, this loss is negligible. The results show that the model explains 82,4% of the variation in the training data. In comparison the model explains 81,3 % of the variation in the test data, this corresponds to a loss of 1,1% in explanatory power which is considered very low.

The prediction of price is based on the rest of the variables in the dataset, these variables are of different importance which is shown in the figure below:

Google colab link:
https://colab.research.google.com/drive/1JeTp7sTLlDc9HUYlqRCO51T1suHK0QwC?usp=sharing

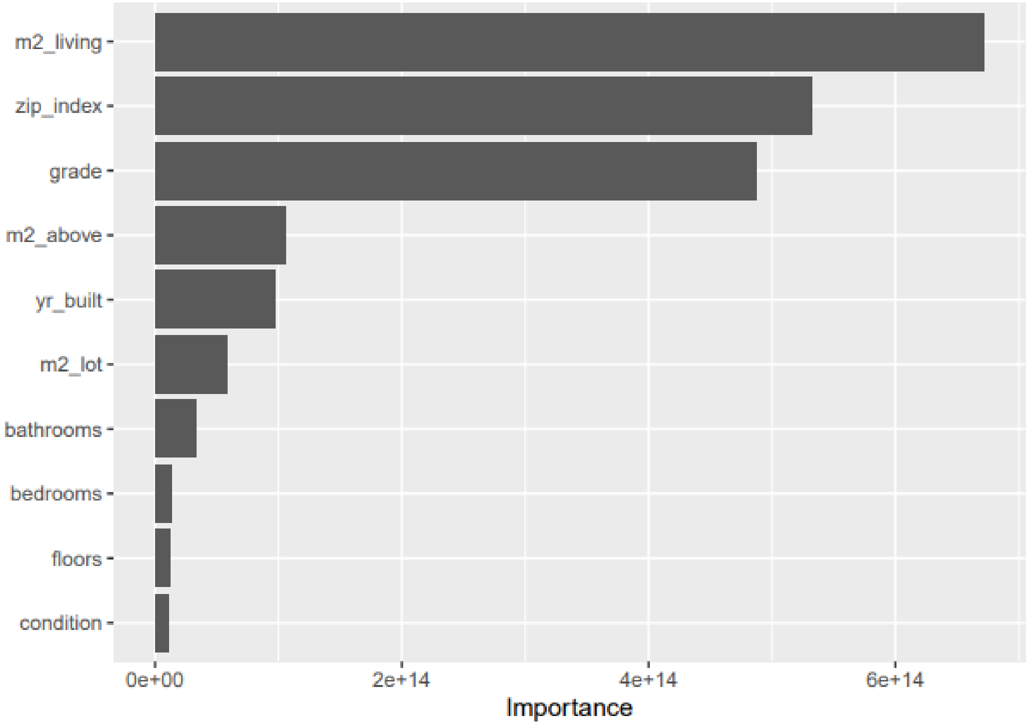


Figure 2: Variable importance

The analysis indicates that the square meters of living space is the most important variable in explaining the housing price. Furthermore the location (based on zipcode) and grade are the next most important variables and these three variables explains the majority of the housing price. Thereby the analysis confirms the previous statements in this report.