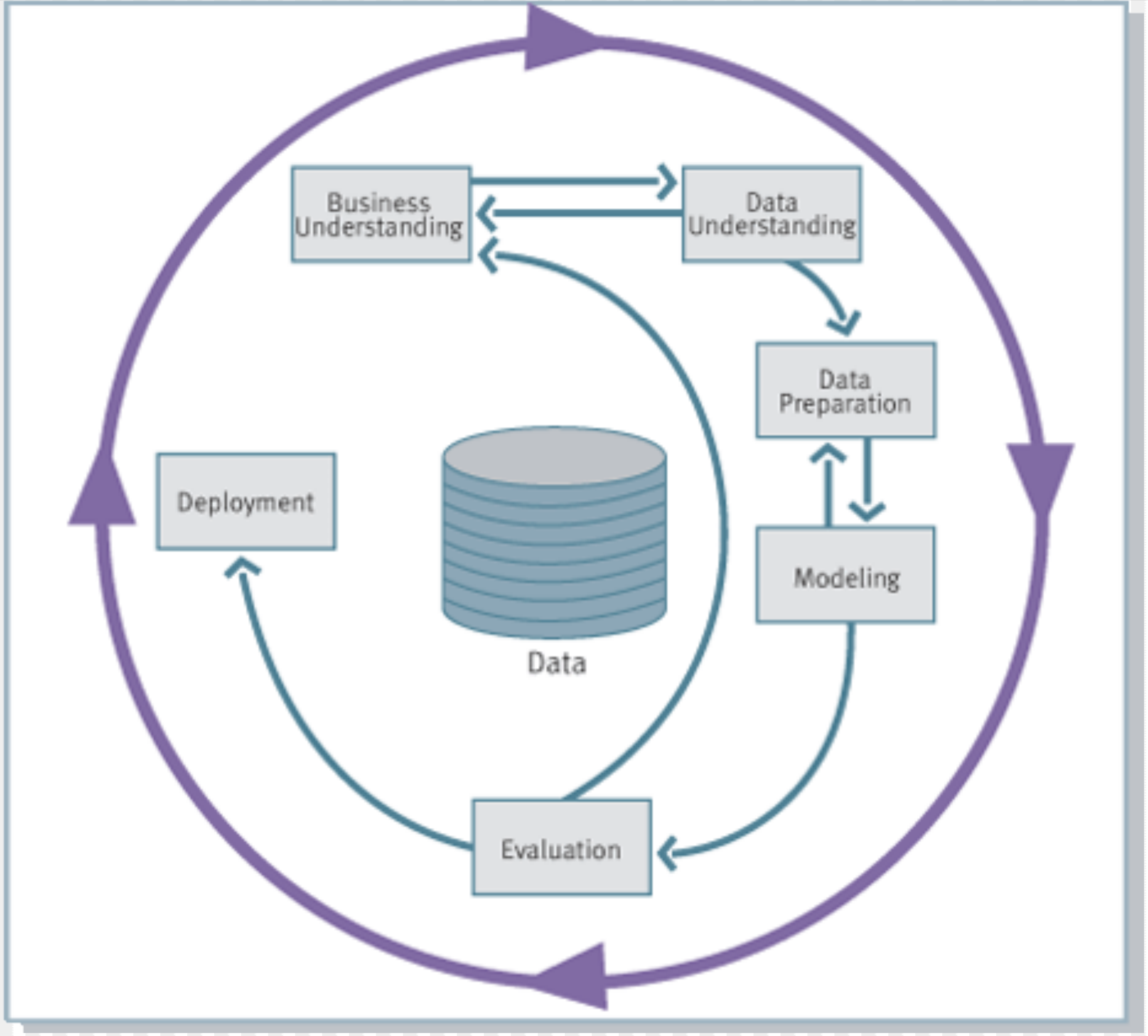
Kings County Housing Prices Prediction

Overview

- This project created a model that can predict the prices of homes sold in the Seattle, WA area. All kinds of features related to the houses are provided. Linear and non-linear machine learning methods in scikit-learn library are used to construct the models with selected data. Houseing prices are predicted within a reasonable error range.

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Deployment
Data
Modeling
Evaluation

Business Understanding

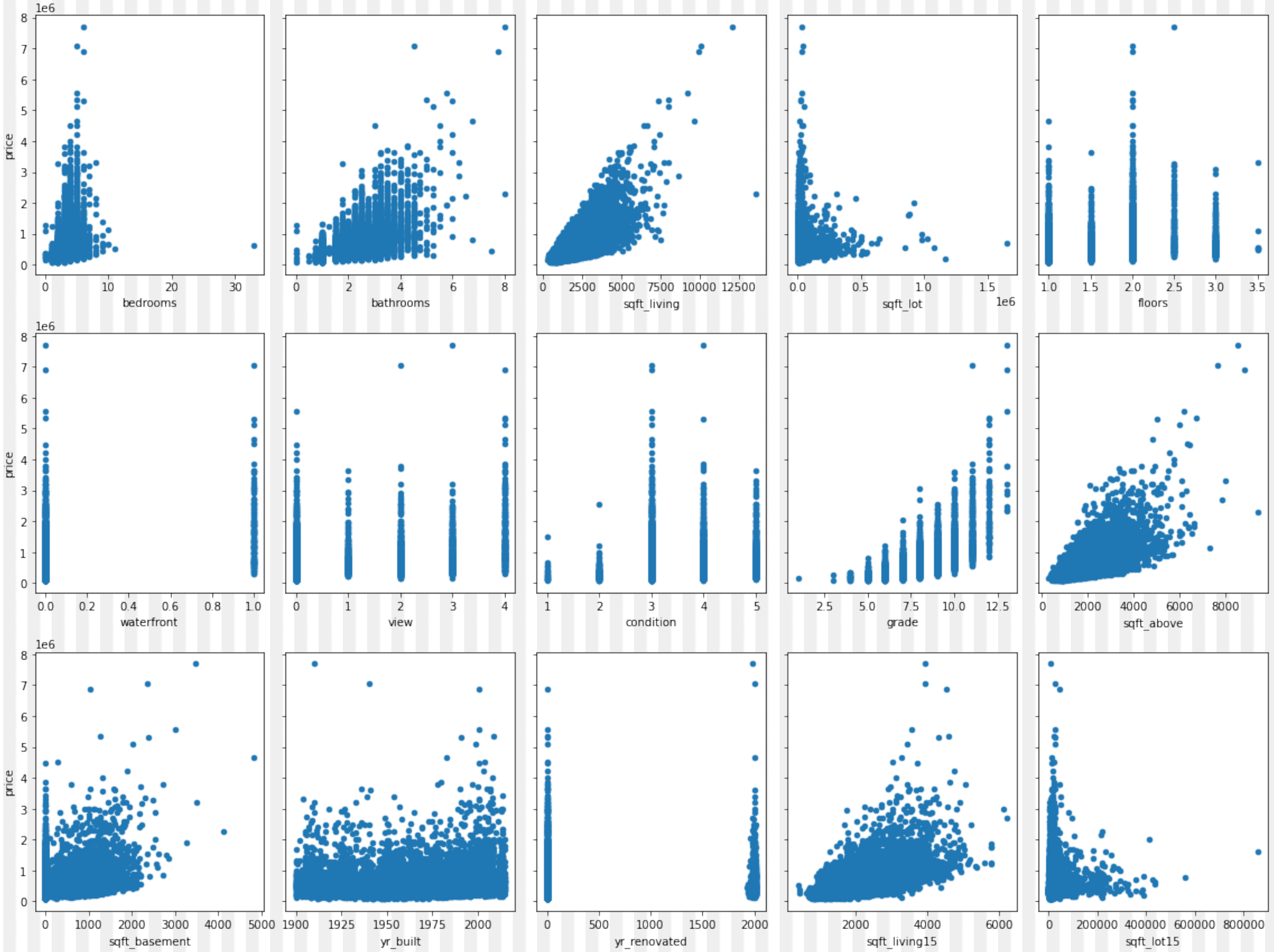
- Initial phase
- Focuses on:
  - Understanding the project objectives and requirements from a business perspective
  - Converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives

A real estate agency in Seattle WA wants to predict the housing prices in order to develop market strategies and target potential customers. The prices of houses with various features are provided for year 2014-2015. Geographical data can also be fetched to enrich the pool of features. The study of population structure of residents is also an important factor for the prediction, but it is outside the scope of this project.

Data Understanding

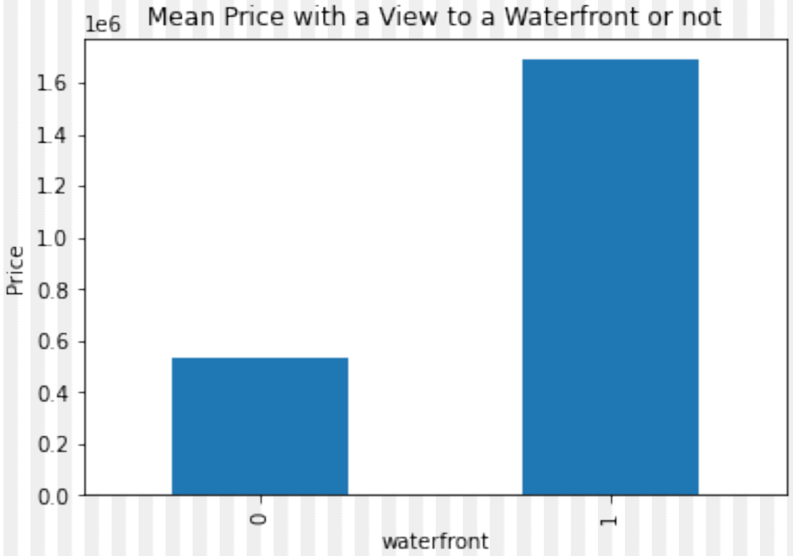
- Starts with an initial data collection
- Proceeds with activities aimed at:
  - Getting familiar with the data
  - Identifying data quality problems
  - Discovering first insights into the data
  - Detecting interesting subsets to form hypotheses for hidden information

Data Understanding



Statistical Test

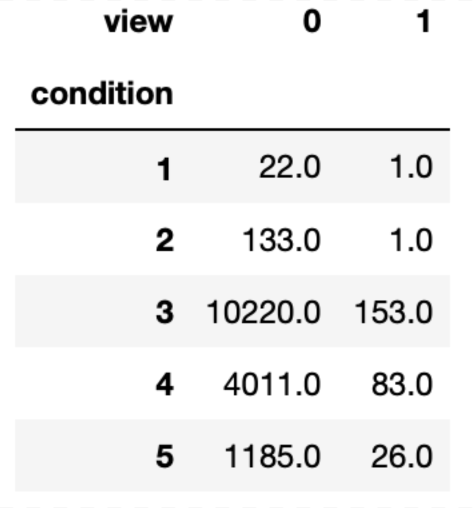
Two-Sample T-test
H0: The sample mean of prices of the houses with a view to a waterfront or not are the same.
Ha: The sample mean of prices of the houses with a view to a waterfront or not are different.



p=2.465e-299

Statistical Test

Chi-square Test
H0: Different conditions of the houses has been viewed the same time.
Ha: Different conditions of the houses has been viewed different time.

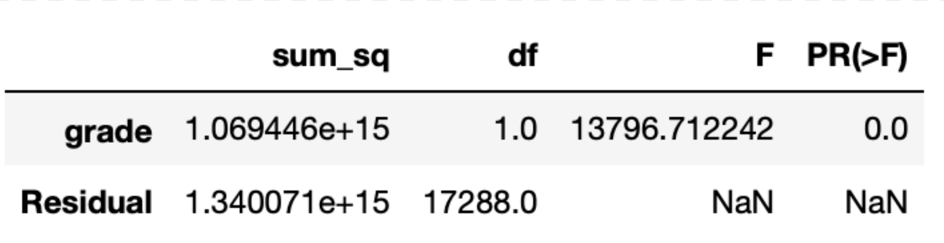


view	0	1
condition		
1	22.0	1.0
2	133.0	1.0
3	10220.0	153.0
4	4011.0	83.0
5	1185.0	26.0

p= 0.0616

Statistical Test

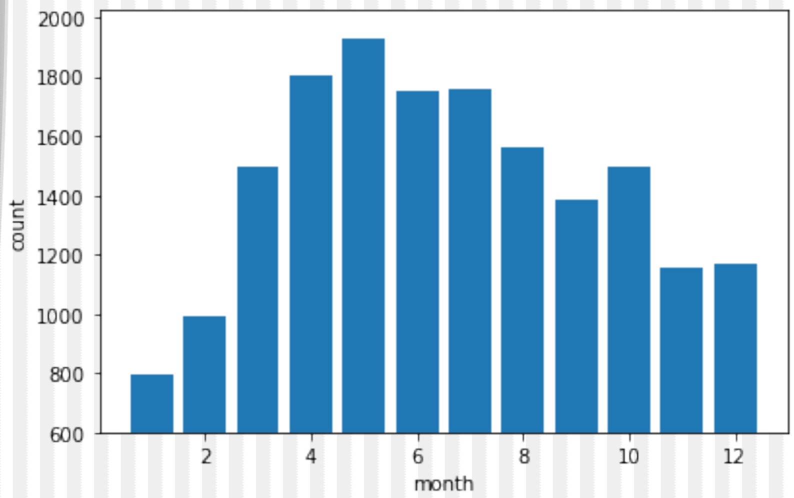
ANOVA
H0: The sample mean of prices of the houses with different grading are the same.
Ha: The sample mean of prices of the houses with different grading are different.

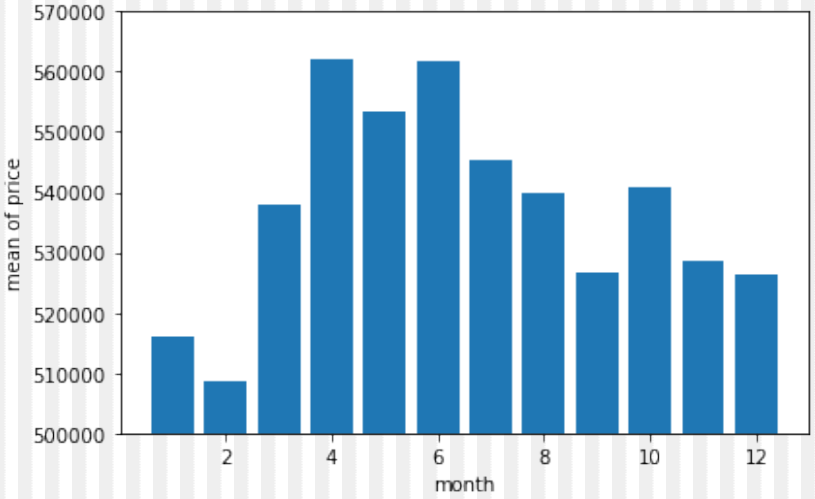


	sum_sq	df	F	PR(>F)
grade	1.069446e+15	1.0	13796.712242	0.0
Residual	1.340071e+15	17288.0	NaN	NaN

Data Preparation

- Covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data
- Data preparation tasks are likely to be performed multiple times, and not in any prescribed order
- Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools





Data Preparation

Identify categorical variables in the data set and create dummy columns to numeric format through one-hot encoding

Bedrooms: <=1, 2, 3, 4, 5, >=6

Condition: 1, 2, 3, 4, 5

Grade: <=5, 6, 7, 8, 9, 10, >=11

Sqft_basement: 0, <1000, <2000, else

Zipcode

Data Preparation

Renovate

not renovated:

house_age = yr_sold - yr_built

renovated:

house_age = yr_sold - yr_rebuilt

yr_rebuilt = yr_renovated * r + yr_built * (1-r)

Data Preparation

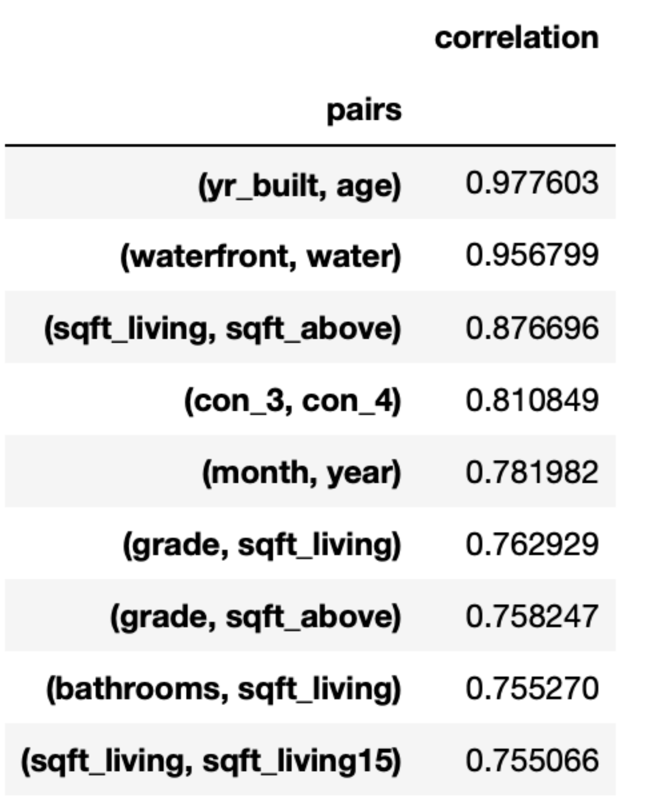
Interaction Features

bath = bathrooms / sqft_living

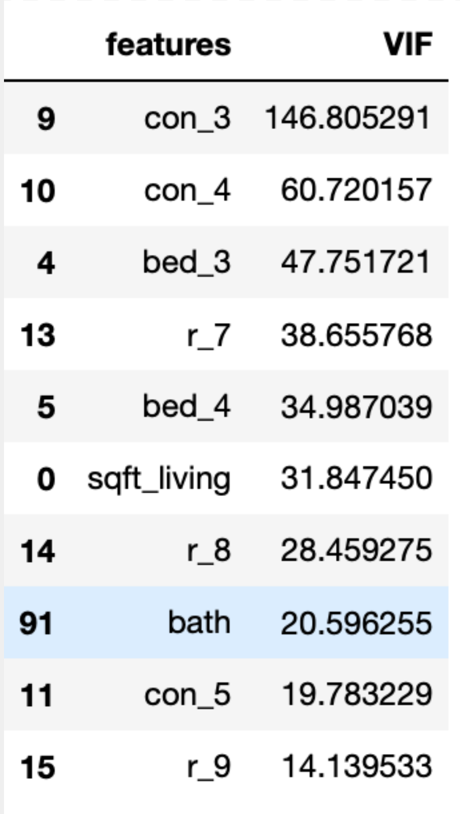
water = waterfront * df.sqft_living15

Data Preparation

Multicollinearity



	correlation
pairs	
(yr_built, age)	0.977603
(waterfront, water)	0.956799
(sqft_living, sqft_above)	0.876696
(con_3, con_4)	0.810849
(month, year)	0.781982
(grade, sqft_living)	0.762929
(grade, sqft_above)	0.758247
(bathrooms, sqft_living)	0.755270
(sqft_living, sqft_living15)	0.755066



	features	VIF
9	con_3	146.805291
10	con_4	60.720157
4	bed_3	47.751721
13	r_7	38.655768
5	bed_4	34.987039
0	sqft_living	31.847450
14	r_8	28.459275
91	bath	20.596255
11	con_5	19.783229
15	r_9	14.139533

Modeling

- Various modeling techniques are selected and applied, and their parameters are calibrated to optimal values
- Typically, there are several techniques for the same data mining problem type
- Some techniques have specific requirements on the form of data, therefore, stepping back to the data preparation phase is often needed

Linear Regression

Training: R2=0.8186858117745891,
RMSE=158449.45380438215 Testing:
R2=0.8026151053998818,
RMSE=167958.07072984657

Modeling

Non-linear transformations

Training: degree=1, R2=0.8165061898786338, RMSE=160803.4853785666

Testing: degree=1, R2=0.8120041137522926, RMSE=159820.86003841623

Training: degree=2, R2=0.9199531039571069, RMSE=106207.94271508754

Testing: degree=2, R2=-9.009923421698971e+18, RMSE=1106419548120385.4

Modeling

Kbest

```
Training: degree=1, R2=0.7591979059264619, RMSE=182601.769260877

Testing: degree=1, R2=0.747065486564507, RMSE=190128.8263162377

Training: degree=2, R2=0.8285844629325447, RMSE=154063.56508682124

Testing: degree=2, R2=-1.2163530063239503e+18, RMSE=416940167772755.6
```

Forward Selection

Modeling

RFECV

Training Root Mean Squared Error: 182601.769260877
Testing Root Mean Squared Error: 190128.8263162377

Lasso

Training Error: 94797.28262323063 Testing Error: 159920.2578841084

Ridge

Training Error: 94760.53244108071 Testing Error: 159819.7450336178

Evaluation

- At this stage, a model (or models) that appears to have high quality, from a data analysis perspective, has been built
- Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives
- A key objective is to determine if there is some important business issue that has not been sufficiently considered
- At the end of this phase, a decision on the use of the data mining results should be reached

Evaluation

- The housing prices is predicted for Kings County in Seattle WA. The model solve the problems for pridiction satisfactorily. New features are created to understand the questions. In any of these cases, it is totally encouraged to revisit the earlier steps.