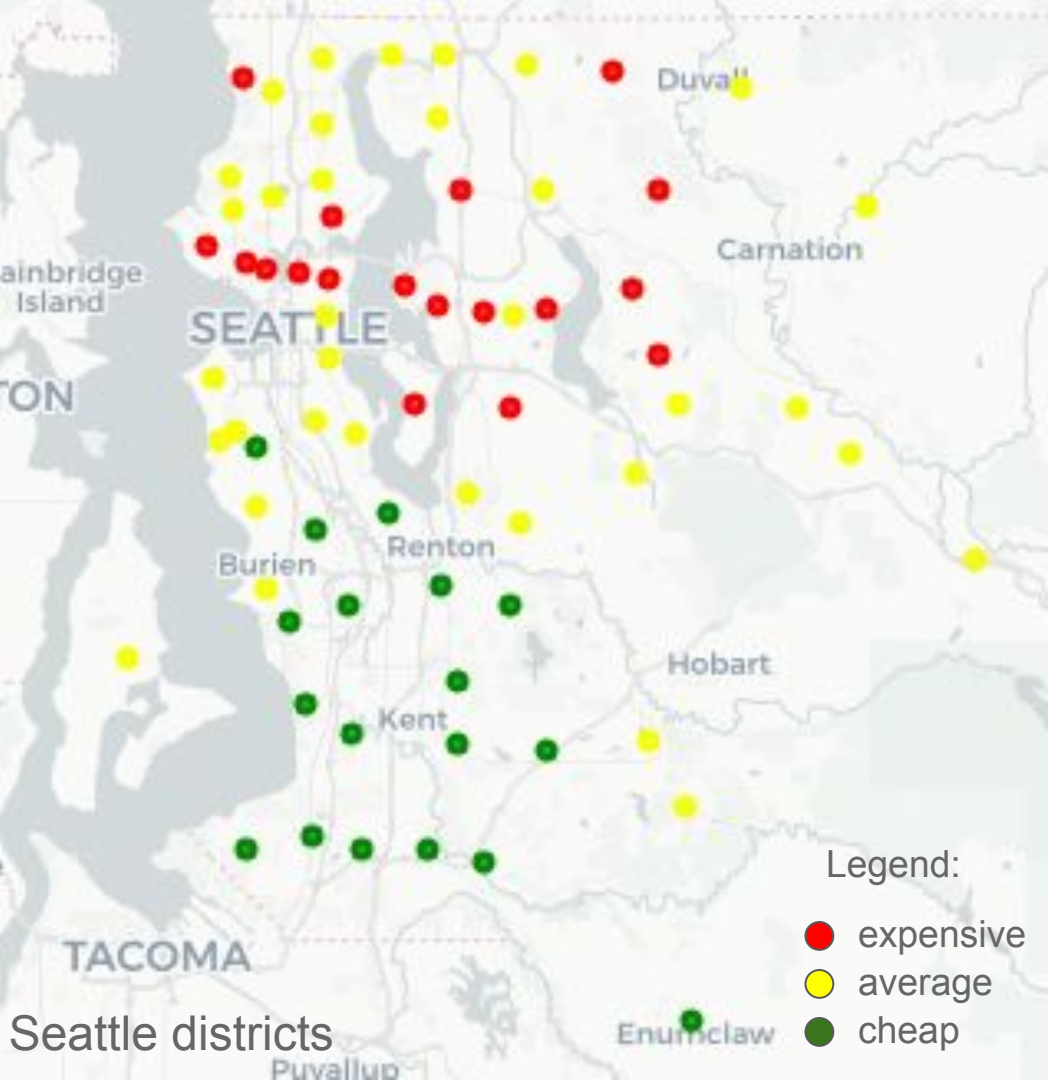
The workflow for the Housing Prices Project for Module 1

Cleaning

Analysing

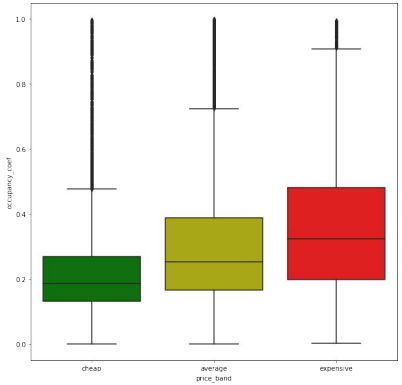
Modelling

ıqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
5650	1.0	NaN	0.0	...	7	1180	0.0	1955	0.0	98178	47.5112	-122.257	1340	5650
7242	2.0	0.0	0.0	...	7	2170	400.0	1951	1991.0	98125	47.7210	-122.319	1690	7639
10000	1.0	0.0	0.0	...	6	770	0.0	1933	NaN	98028	47.7379	-122.233	2720	8062
5000	1.0	0.0	0.0	...	7	1050	910.0	1965	0.0	98136	47.5208	-122.393	1360	5000
8080	1.0	0.0	0.0	...	8	1680	0.0	1987	0.0	98074	47.6168	-122.045	1800	7503
101930	1.0	0.0	0.0	...	11	3890	1530.0	2001	0.0	98053	47.6561	-122.005	4760	101930
6819	2.0	0.0	0.0	...	7	1715	?	1995	0.0	98003	47.3097	-122.327	2238	6819
9711	1.0	0.0	NaN	...	7	1060	0.0	1963	0.0	98198	47.4095	-122.315	1650	9711
7470	1.0	0.0	0.0	...	7	1050	730.0	1960	0.0	98146	47.5123	-122.337	1780	8113
6560	2.0	0.0	0.0	...	7	1890	0.0	2003	0.0	98038	47.3684	-122.031	2390	7570

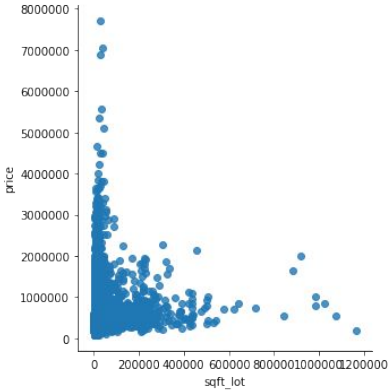


Seattle districts

Analysing data



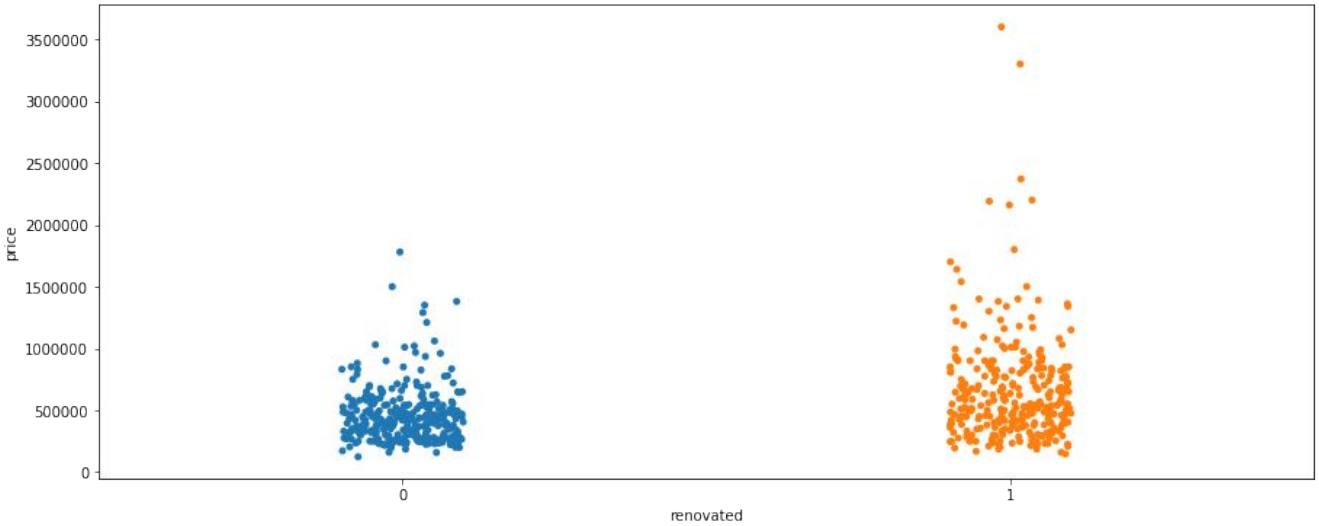
Districts density

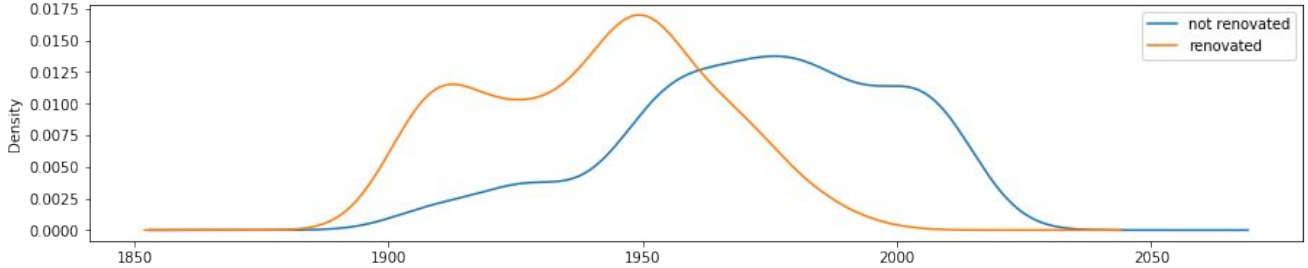


Prices for lots

How dense are cheap, average, and expensive districts?

How are districts distributed in the city according to their prices?

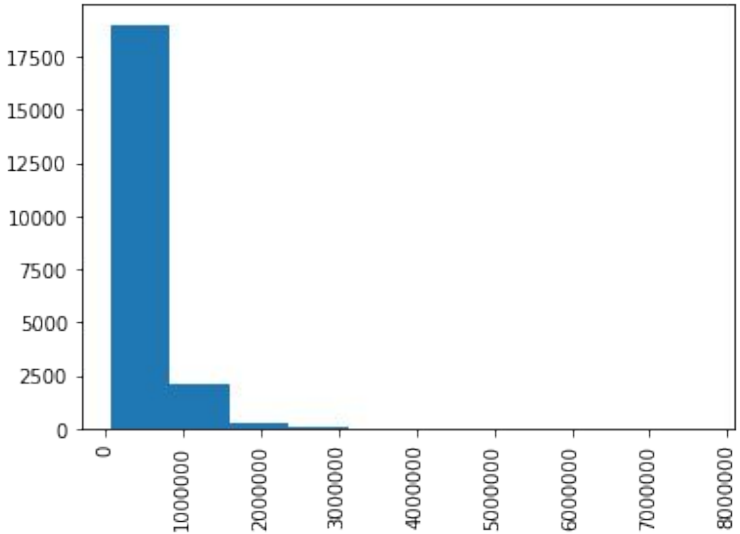


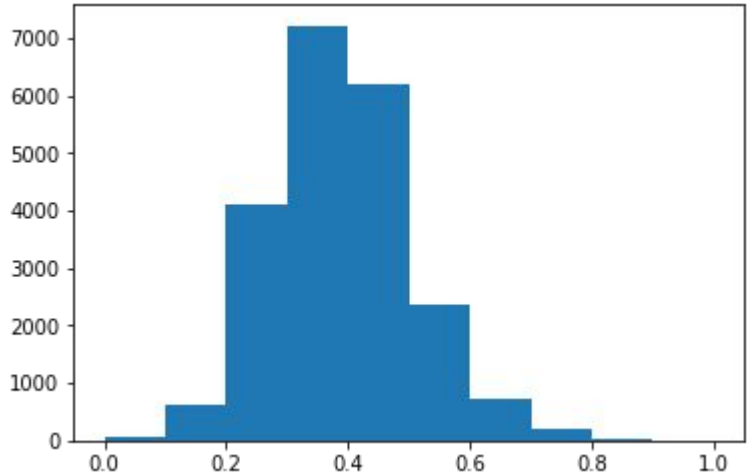


Are renovated houses more expensive than not renovated?

Transforming the response

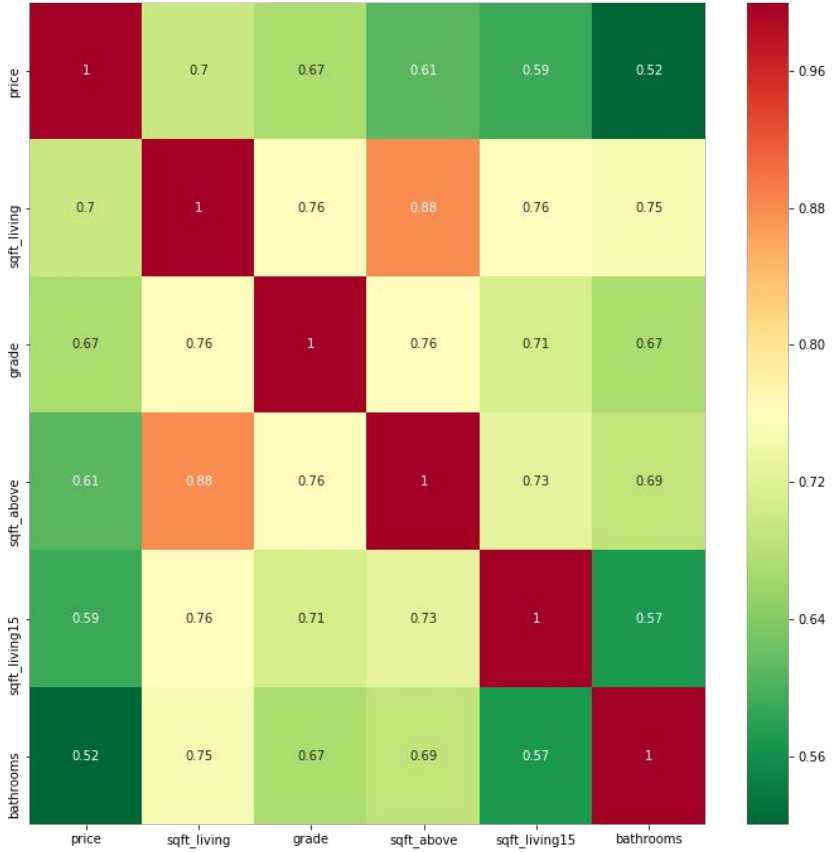
Log transformation





Adding features and dealing with collinearity

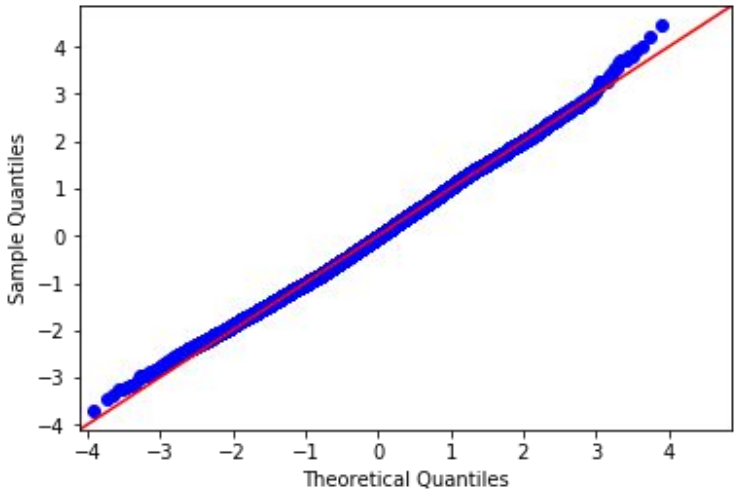
	correlation
log_price	1.00
scaled_grade	0.70
grade	0.70
sqft_living	0.69
scaled_sqft_living15	0.62



The final model

price = exp(0.68 * grade + 0.17 * sqft_living15 + 0.09 * bedrooms)

R^2 = 0.96



Possible extensions

1. Create additional models for the different price bands
2. Feature engineering
3. Split the data into a training set and a test set