

KING COUNTY HOUSE SALES - LINEAR REGRESSION MODEL



KING COUNTY HOUSE SALES - LINEAR REGRESSION MODEL

Data Overview

DATA OVERVIEW - PREPROCESSED

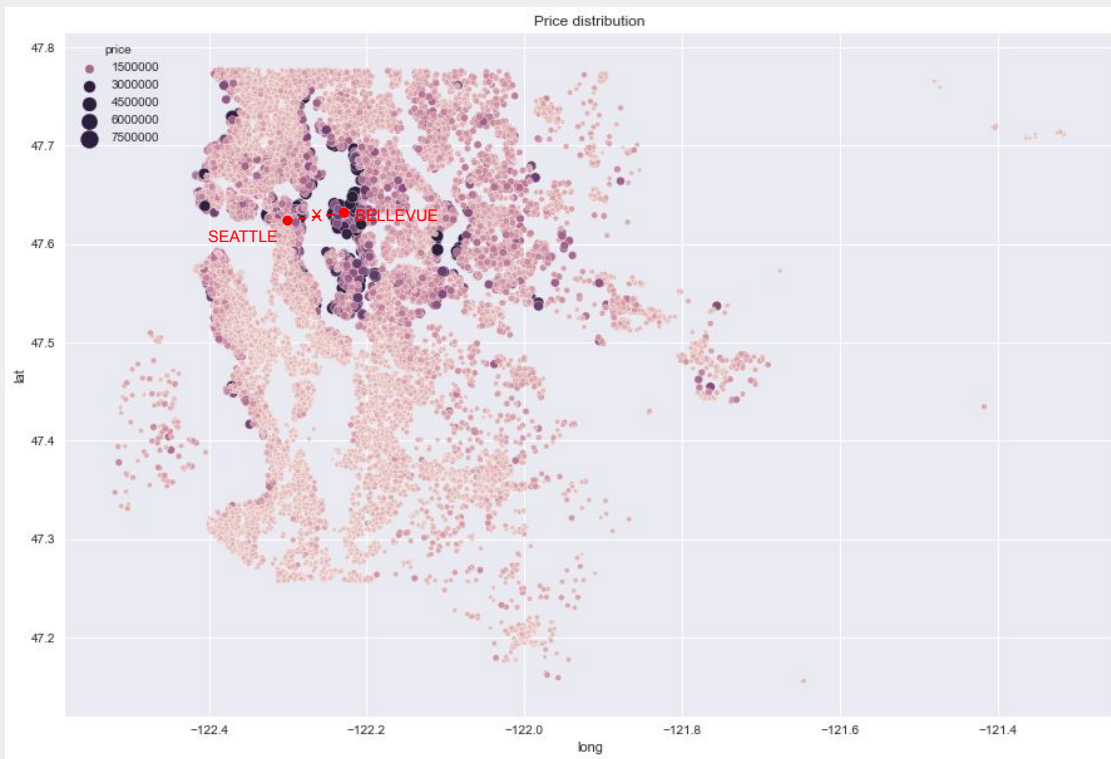
	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
0	2014	221900.0	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
1	2014	538000.0	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951	1	98125	47.7210	-122.319	1690	7639
2	2015	180000.0	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
3	2014	604000.0	4	3.00	1960	5000	1.0	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
4	2015	510000.0	3	2.00	1680	8080	1.0	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503

DATA OVERVIEW - NEW FEATURES TO INCREASE MODEL PERFORMANCE

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	...	zipcode	lat	long	sqft_living15	sqft_lot15	has_basement	dist_epicenter	zip_price_cat	above_lot_ratio	is_residential
0	2014	221900.0	3	1.00	1180	5650	1.0	0	0	3	...	98178	47.5112	-122.257	1340	5650	0	10.497043	1	4.79	1
1	2014	538000.0	3	2.25	2570	7242	2.0	0	0	3	...	98125	47.7210	-122.319	1690	7639	1	13.583780	1	3.34	1
2	2015	180000.0	2	1.00	770	10000	1.0	0	0	3	...	98028	47.7379	-122.233	2720	8062	0	14.846308	1	12.99	0
3	2014	604000.0	4	3.00	1960	5000	1.0	0	0	5	...	98136	47.5208	-122.393	1360	5000	1	13.753042	1	4.76	1
4	2015	510000.0	3	2.00	1680	8080	1.0	0	0	3	...	98074	47.6168	-122.045	1800	7503	0	16.131475	2	4.81	1

KING COUNTY HOUSE SALES - LINEAR REGRESSION MODEL

Price distribution



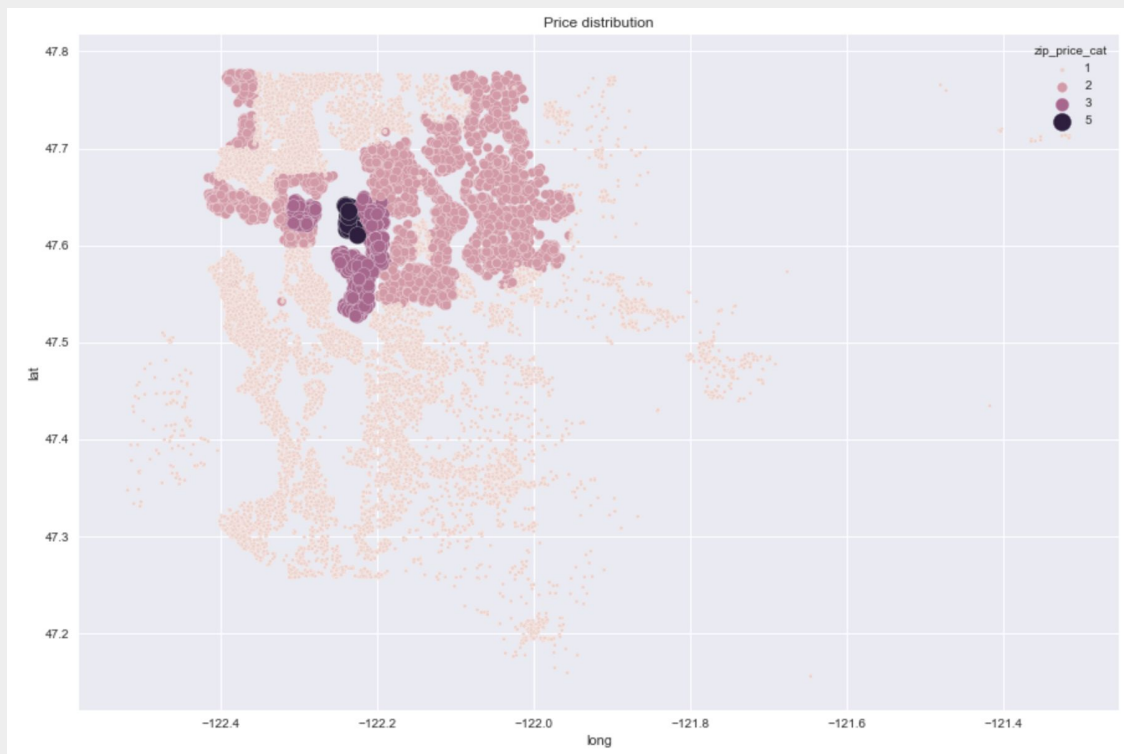
Key observations based on location:

- the house prices are strongly related to the distance from Seattle and Bellevue
- the house prices increase significantly if the house is a waterfront property
- the house prices vary based on the zip code

Scatter plot with house prices distribution across King County

KING COUNTY HOUSE SALES - LINEAR REGRESSION MODEL

Zip codes and mean house prices

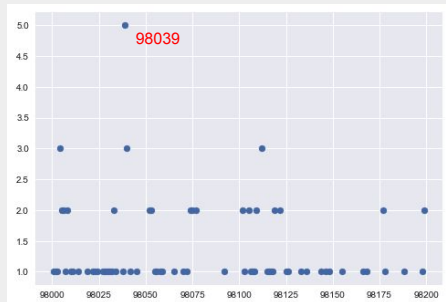


Scatter plot with mean house price distribution by zip code across King County

Zip codes categories:

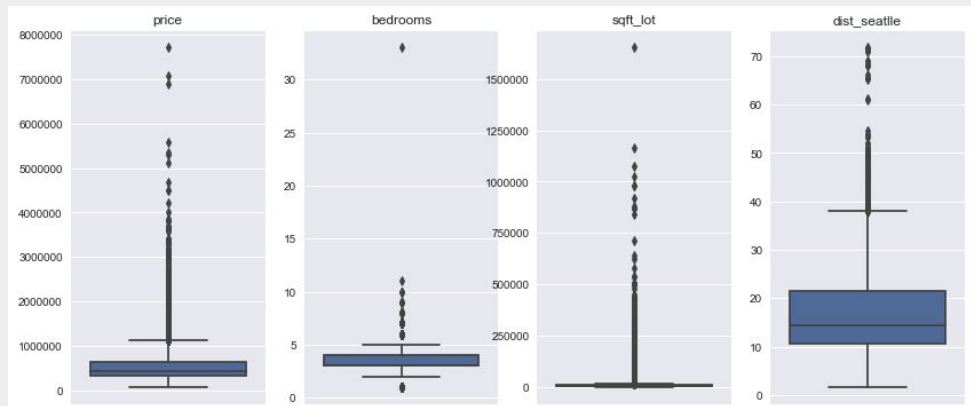
- Zip codes were categorized in four categories based on the mean house price in each zip code
- The most expensive zip code 98039 where the mean house price is above \$2M

	price	zip_price_cat
0	4.252814e+05	1
1	7.622701e+05	2
2	1.215532e+06	3
3	2.186224e+06	5



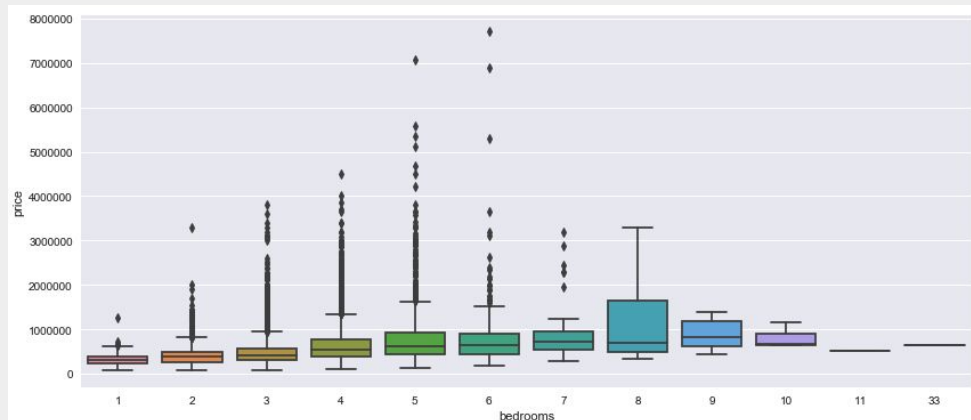
KING COUNTY HOUSE SALES - LINEAR REGRESSION MODEL

Outliers



Identify extreme values:

- Based on the observation of the dataset, we can identify some extreme values that can affect the accuracy of the model. Some of them were removed i.e 11 and 33 bedroom properties with suspiciously low price or exceptionally large lots.



Boxplots showing the value distribution. The values represented by points are outliers.

KING COUNTY HOUSE SALES - LINEAR REGRESSION MODEL

Results and interpretation

	coef	std err	t	P> t	[0.025	0.975]
const	7.6701	0.039	196.885	0.000	7.594	7.746
has_basement	0.2757	0.004	63.032	0.000	0.267	0.284
zip_price_cat	0.1953	0.004	47.813	0.000	0.187	0.203
sqft_above_log	0.8130	0.006	136.061	0.000	0.801	0.825
bedrooms	-0.0429	0.003	-16.161	0.000	-0.048	-0.038
waterfront	0.7267	0.024	30.889	0.000	0.681	0.773
dist_epicenter_log	-0.3209	0.004	-82.955	0.000	-0.329	-0.313

Key factors that has the most significant impact on the house price:

- **Square footage** - a 10% increase in the house square footage leads to a 8% increase in price
- **Waterfront properties** - the price of the house increases by 107% if waterfront equals 1
- **Distance from Seattle/Bellevue** - a 10% increase in distance from the epicenter leads to a 3% decrease in price
- **Basement** - the price of the house increases by 32% if the house has a basement (equals 1)
- **Zip code** - the price of the house increases by 22% when increasing the zip code category by 1