EDA-Project

House prices of King County

by Marek Nowaczewski

Given data

Variables	Number of entities: 21597

id	sqft_basement
date	yr_built
price	yr_renovated
bedrooms	zipcode
bathrooms	lat
sqft_living	long
sqft_lot	sqft_living15
floors	sqft_lot15
waterfront	spec_price
view	lo_li_rat
condition	waterfront_str
grade	ziporder
sqft_above	

Calculated columns for analysis

Variables

- **specific price**
- **ratio between lot size and living size**

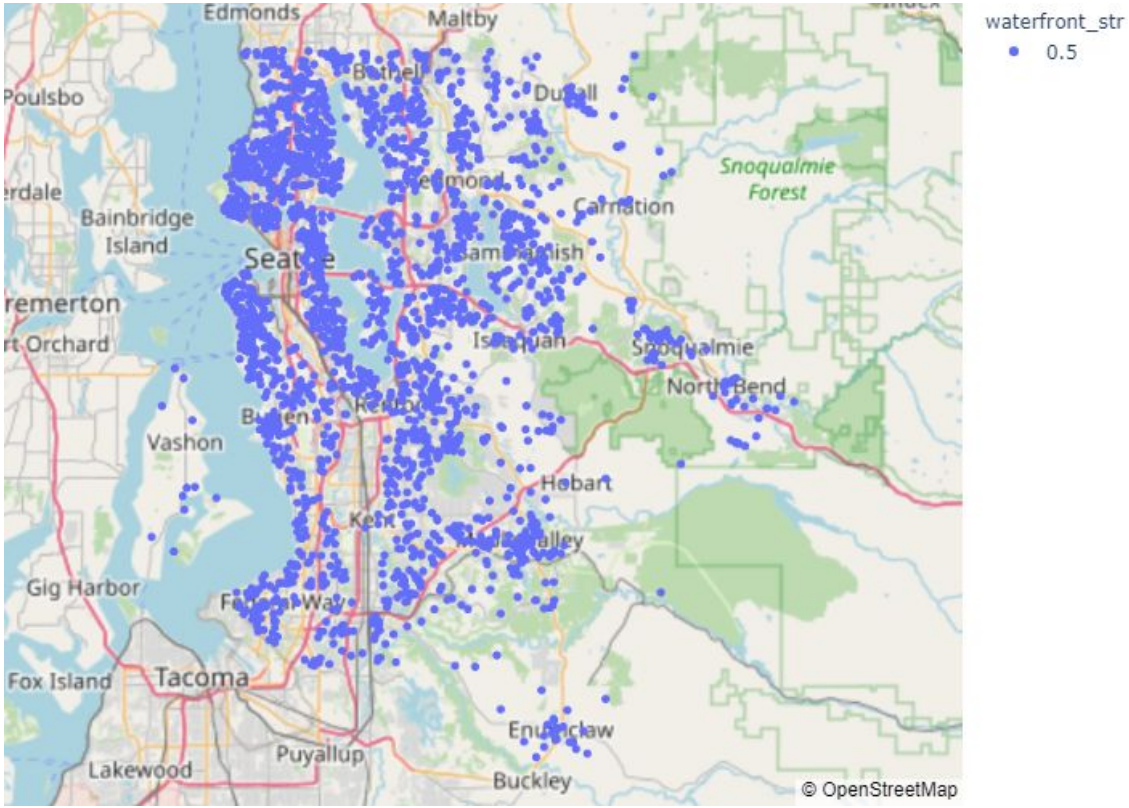
Derivate / calculated data

```
1 df['spec_price'] = df.price / df.sqft_living
2 df['lo_li_rat'] = df.sqft_lot / df.sqft_living
```

Missing values

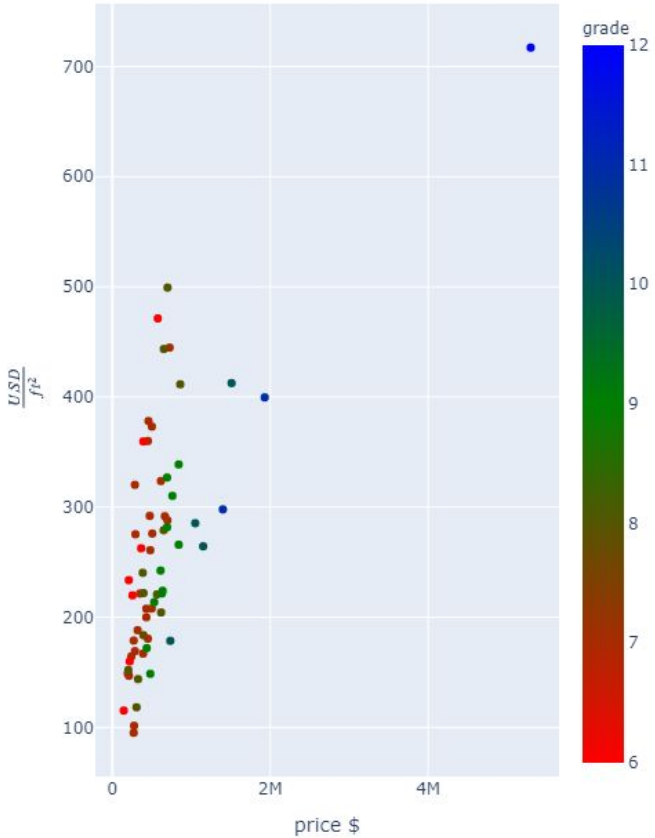
	counts	percentage
waterfront	2376	11.00
view	63	0.29
yr_renovated	3842	17.79

Missing values *waterfront*

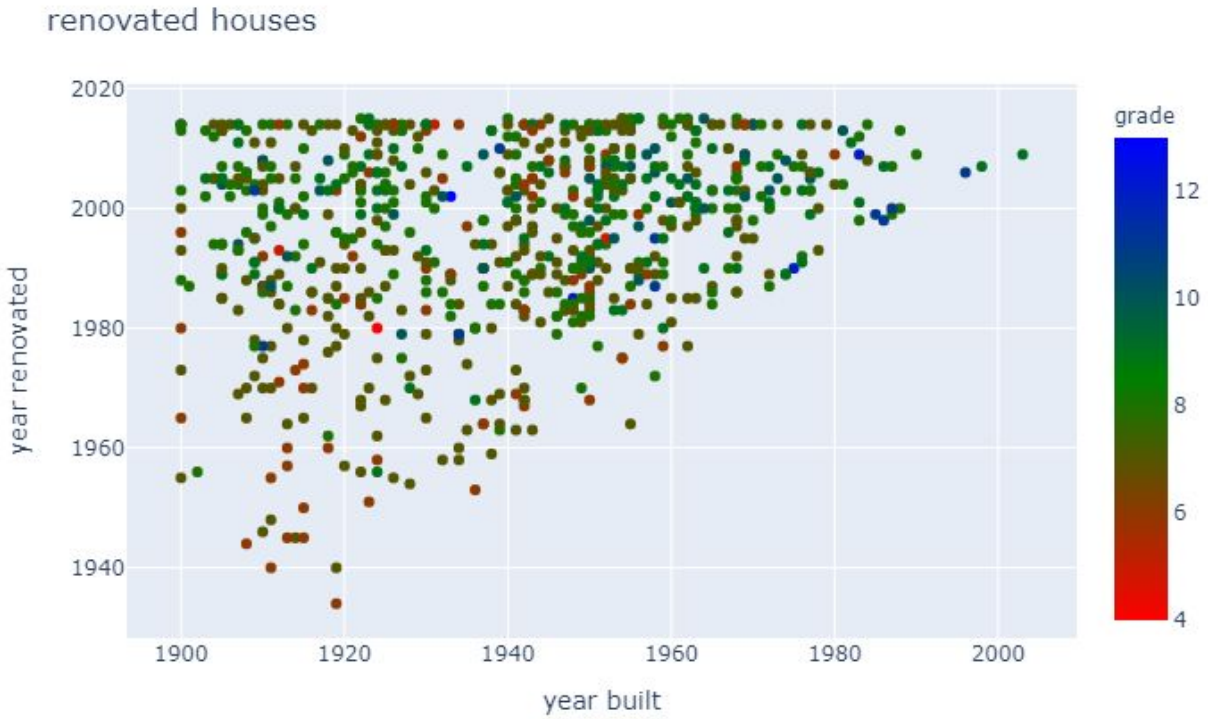


Missing values *view*

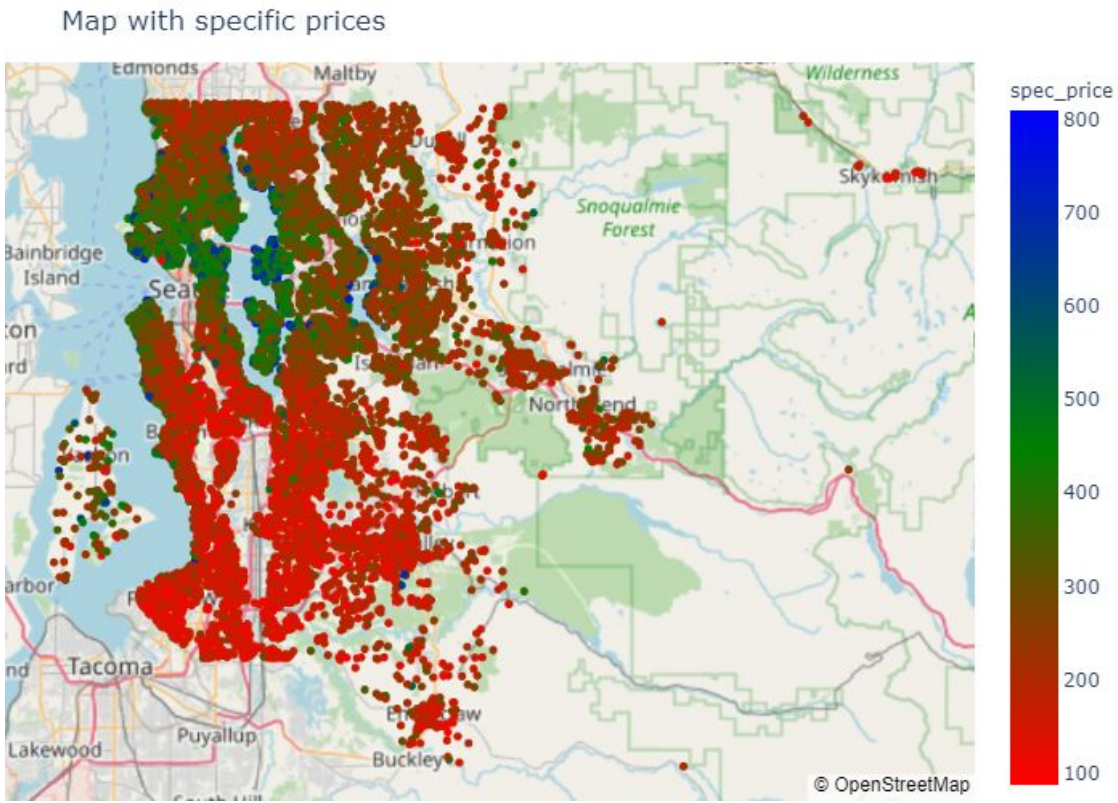
view	price	grade	spec_price	counts
0.0	4.968061e+05	7.566214	256.894200	19422
1.0	8.133733e+05	8.115152	320.076130	330
2.0	7.913904e+05	8.315569	304.420288	957
3.0	9.732852e+05	8.730315	323.027575	508
4.0	1.452466e+06	9.063091	434.540453	317



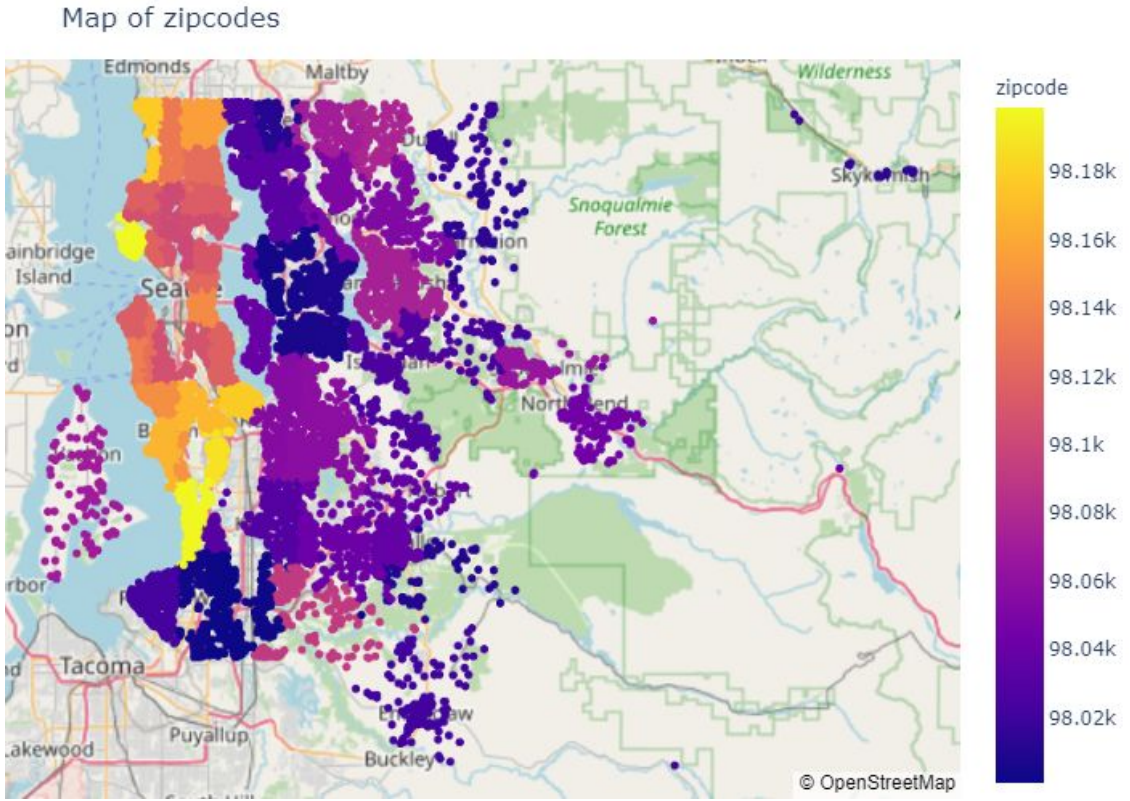
Missing values *yr_renovated*



Location, location and location….

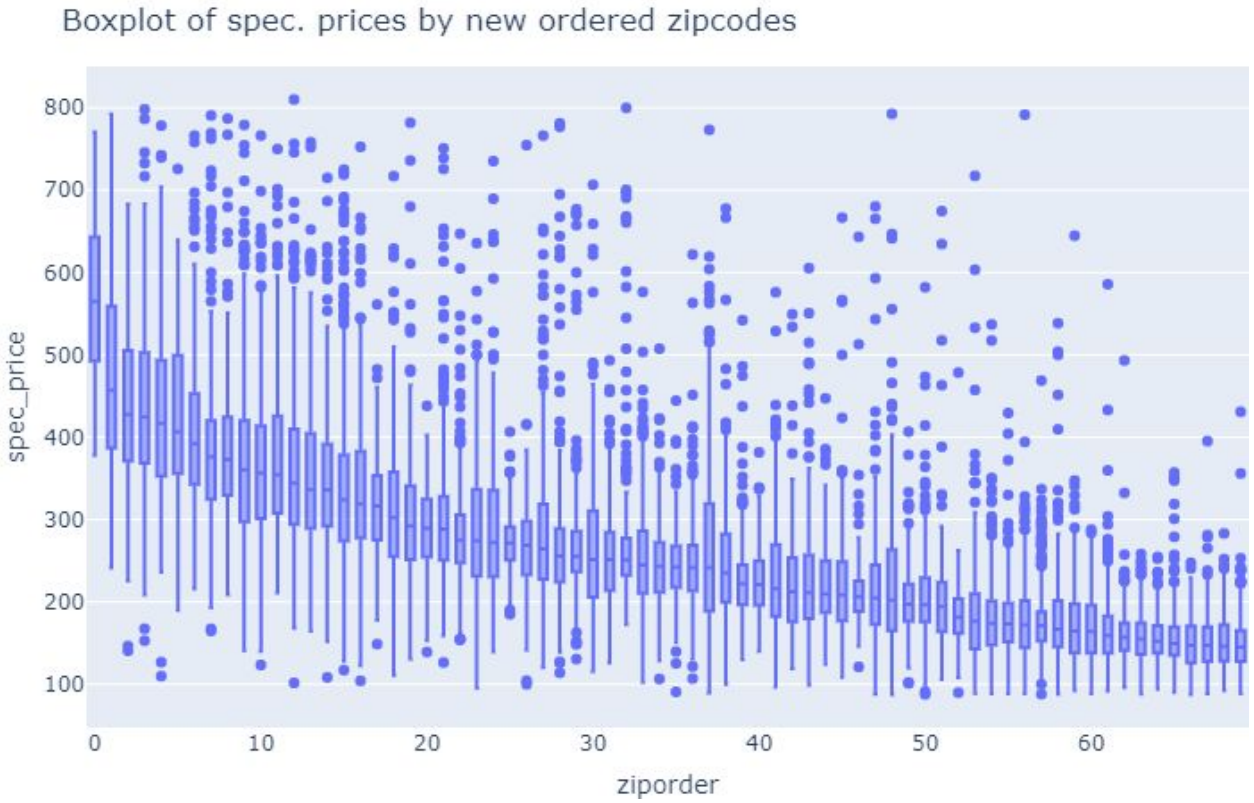


Location, location and location…. #2



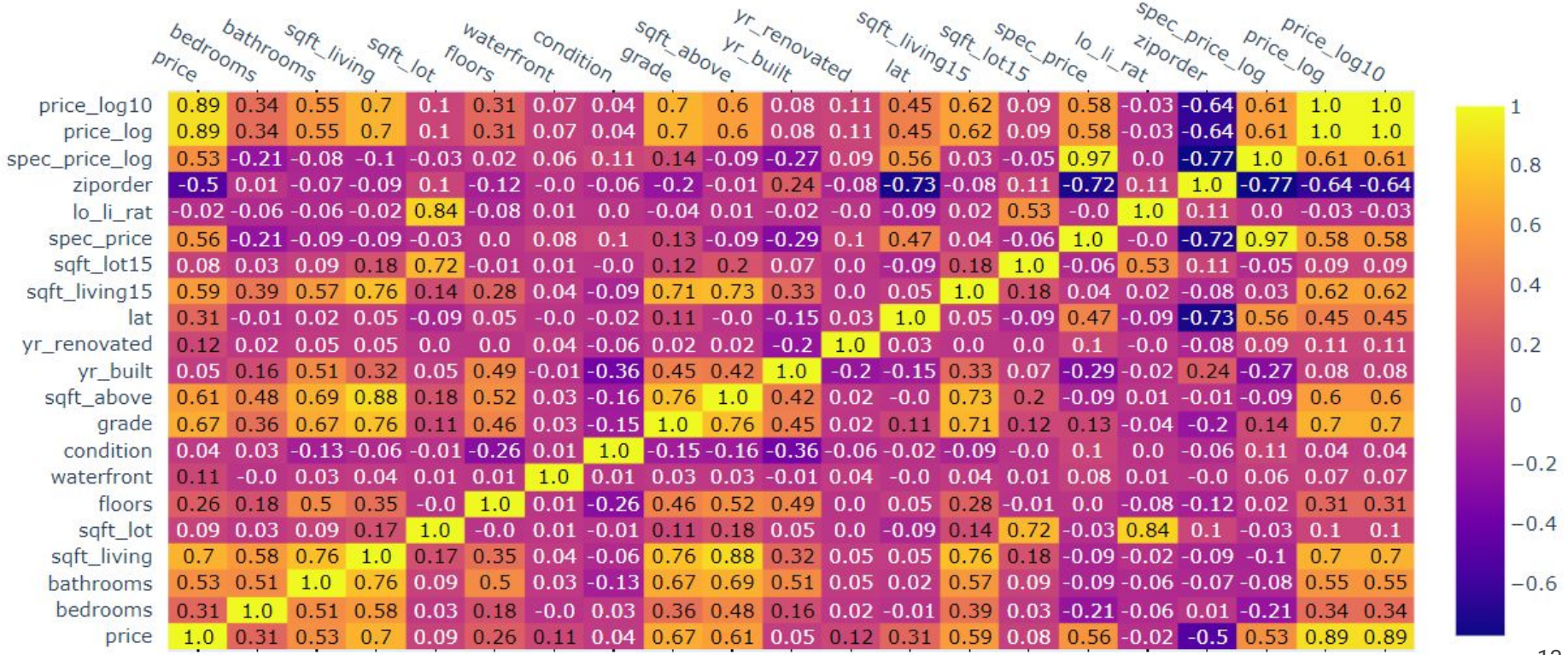
Location, location and location…. #3

Order of zipcodes



Order of zipcodes

Correlations



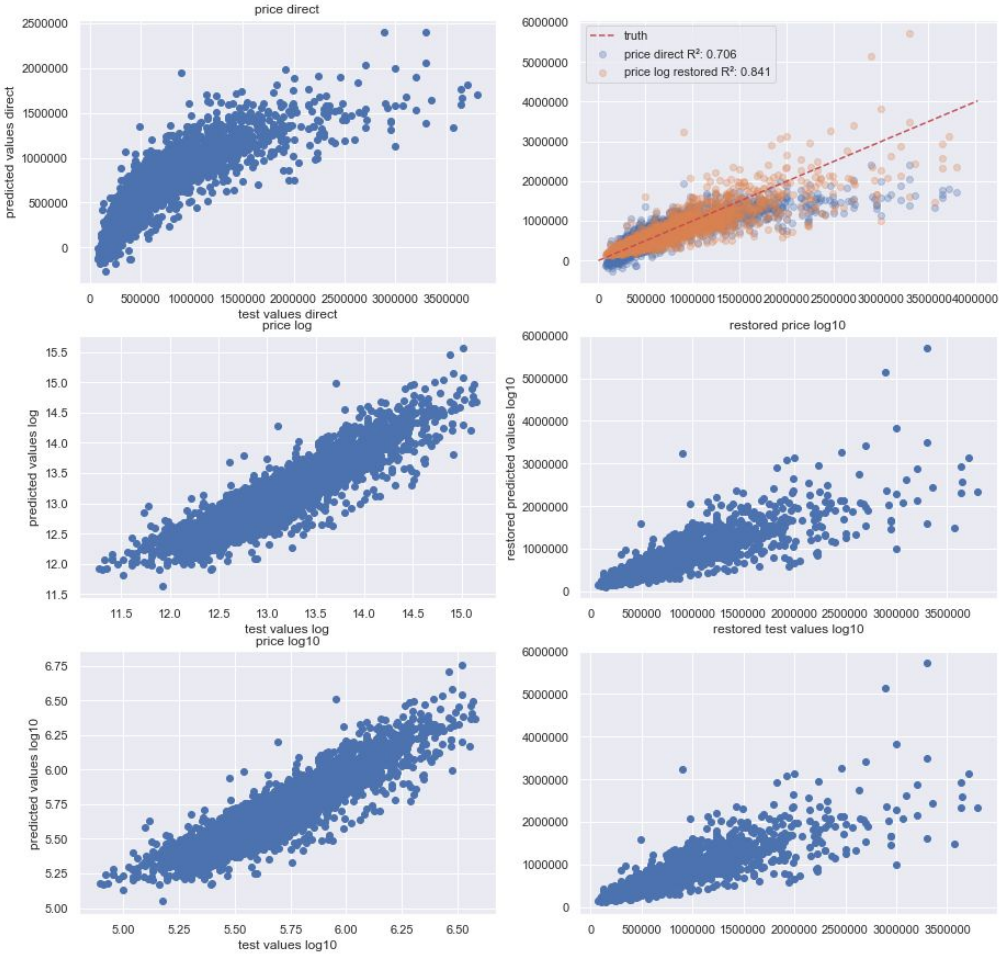
Linear Regression

here used

- ziporder
- sqft_living
- grade
- yr_built
- long
- condition

tragets

- price
- ln(price)
- log10(price)



Predictions

price

OLS Regression Results

Dep. Variable:	y	R-squared:	0.706
Model:	OLS	Adj. R-squared:	0.705
Method:	Least Squares	F-statistic:	5775.
Date:	Wed, 10 Jun 2020	Prob (F-statistic):	0.00
Time:	10:07:57	Log-Likelihood:	-1.9710e+05
No. Observations:	14469	AIC:	3.942e+05
Df Residuals:	14462	BIC:	3.943e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.721e+05	1.73e+06	-0.215	0.830	-3.76e+06	3.02e+06
x1	-6784.3794	96.683	-70.171	0.000	-6973.891	-6594.868
x2	204.0673	2.878	70.902	0.000	198.426	209.709
x3	8.325e+04	2454.583	33.915	0.000	7.84e+04	8.81e+04
x4	-1577.6081	81.014	-19.473	0.000	-1736.406	-1418.811
x5	-2.555e+04	1.36e+04	-1.872	0.061	-5.23e+04	1200.643
x6	2.029e+04	2633.930	7.704	0.000	1.51e+04	2.55e+04

log(price)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.841
Method:	Least Squares	F-statistic:	1.274e+04
Date:	Wed, 10 Jun 2020	Prob (F-statistic):	0.00
Time:	10:07:57	Log-Likelihood:	2060.4
No. Observations:	14469	AIC:	-4107.
Df Residuals:	14462	BIC:	-4054.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	32.0037	1.820	17.583	0.000	28.436	35.571
x1	-0.0144	0.000	-142.061	0.000	-0.015	-0.014
x2	0.0002	3.03e-06	78.986	0.000	0.000	0.000
x3	0.1347	0.003	52.171	0.000	0.130	0.140
x4	-0.0010	8.52e-05	-11.179	0.000	-0.001	-0.001
x5	0.1496	0.014	10.419	0.000	0.121	0.178
x6	0.0497	0.003	17.933	0.000	0.044	0.055

Predictions

Outlook

ToDo

- search / optimization / selection of input data
- filtering outliers
- usage of a squared regression?.
- other handling of zipcodes other categorical data
- usage of dummies
- interactive plots in presentation

Thank you!