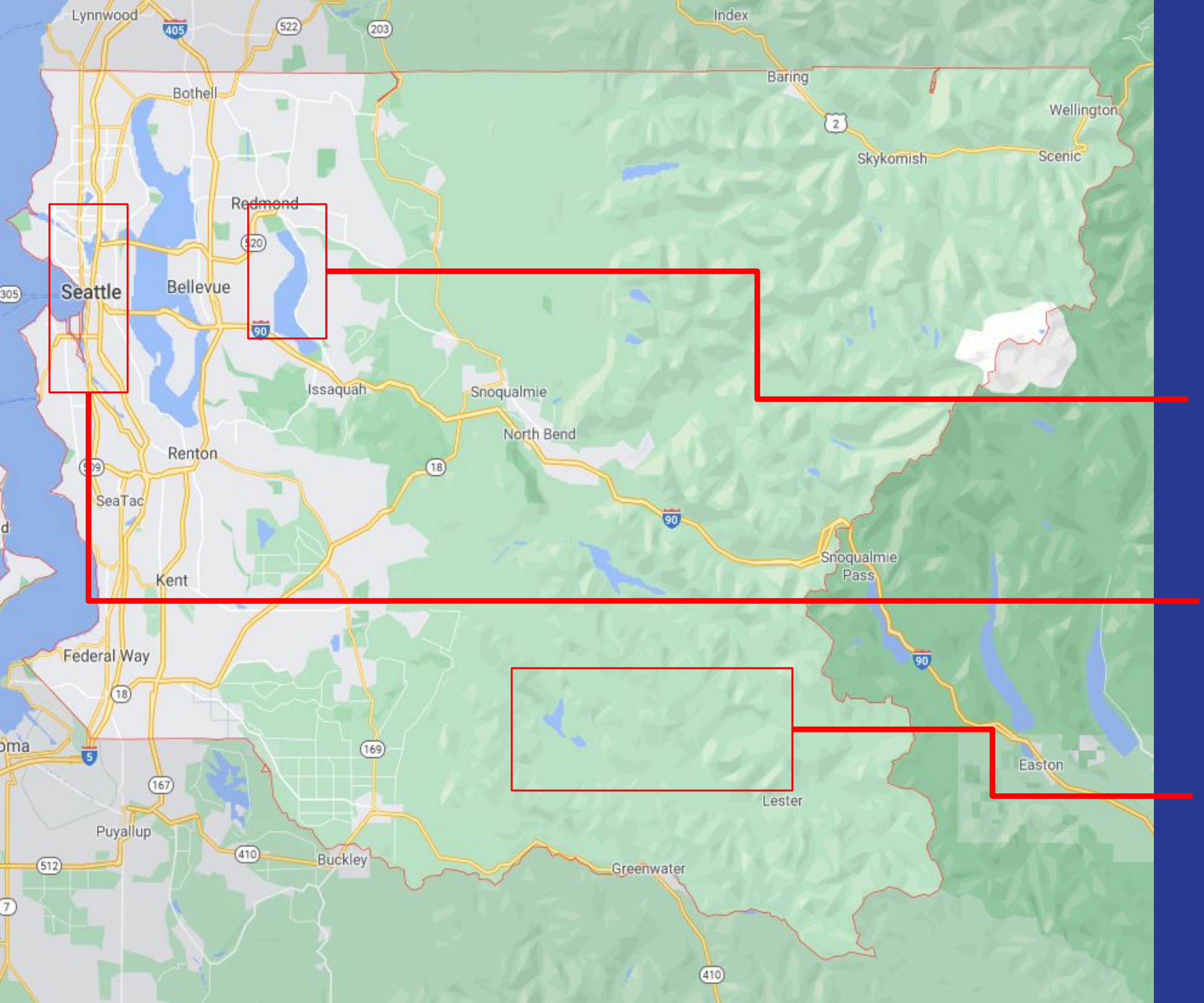
Predicting House Prices in King County, Washington

Leonard Boes

First look on map

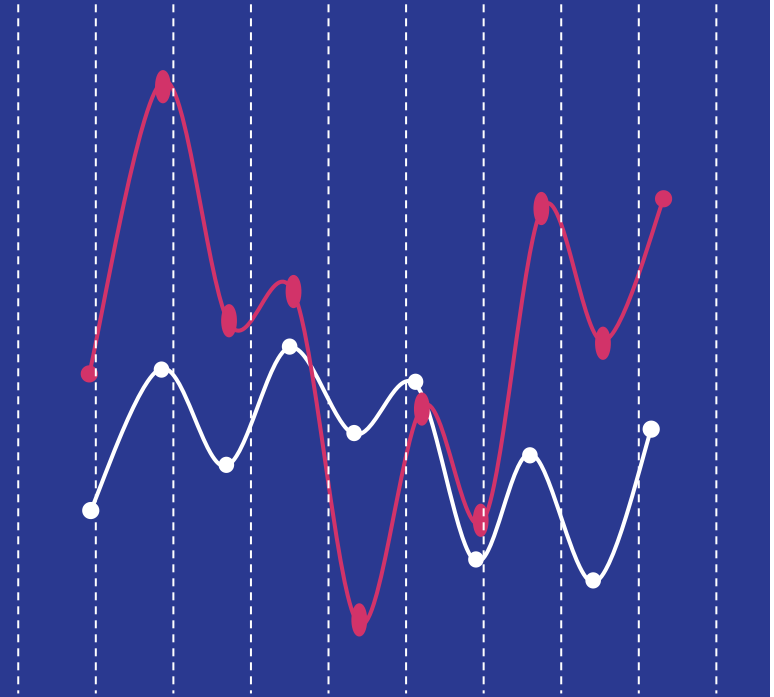


Unique areas like waterview

Dense Seattle Metropolis

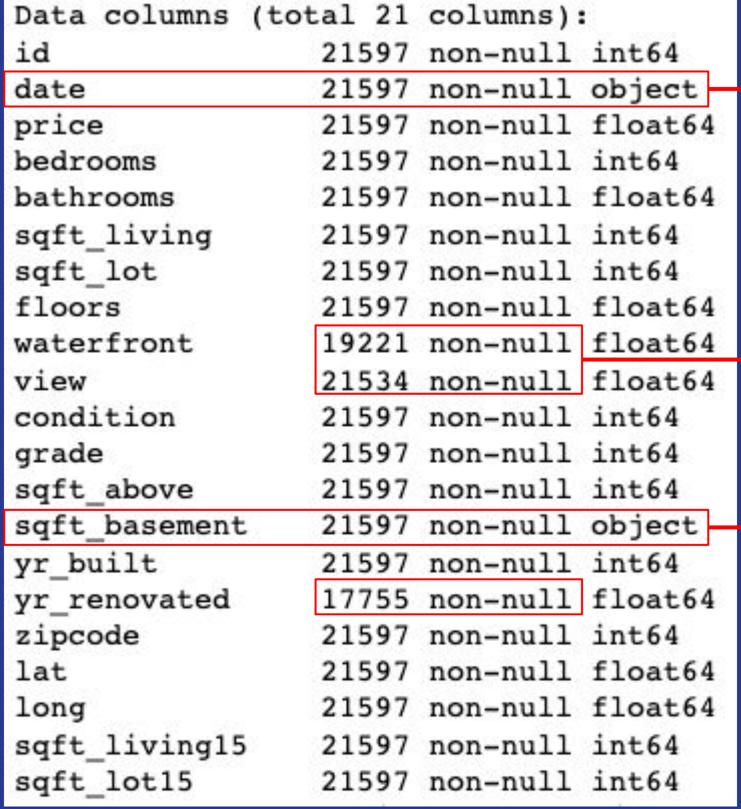
Rural areas far away from city and interstate

First look on the data



- Price and date of individual sales
- Number of bedrooms, bathrooms and floors
- Rank of condition and information about waterview
- Geographical location and zipcode
- Information about size of lot and living area
- Information about size of lot and living area of 15 neighboring houses
- Year of construction and renovation
- Roughly 21,600 observations

Condition of the data



```
Data columns (total 21 columns):
id               21597 non-null int64
date             21597 non-null object
price            21597 non-null float64
bedrooms         21597 non-null int64
bathrooms        21597 non-null float64
sqft_living      21597 non-null int64
sqft_lot         21597 non-null int64
floors           21597 non-null float64
waterfront       19221 non-null float64
view             21534 non-null float64
condition        21597 non-null int64
grade            21597 non-null int64
sqft_above       21597 non-null int64
sqft_basement    21597 non-null object
yr_built         21597 non-null int64
yr_renovated     17755 non-null float64
zipcode          21597 non-null int64
lat              21597 non-null float64
long             21597 non-null float64
sqft_living15    21597 non-null int64
sqft_lot15       21597 non-null int64
```

Date needed to be converted

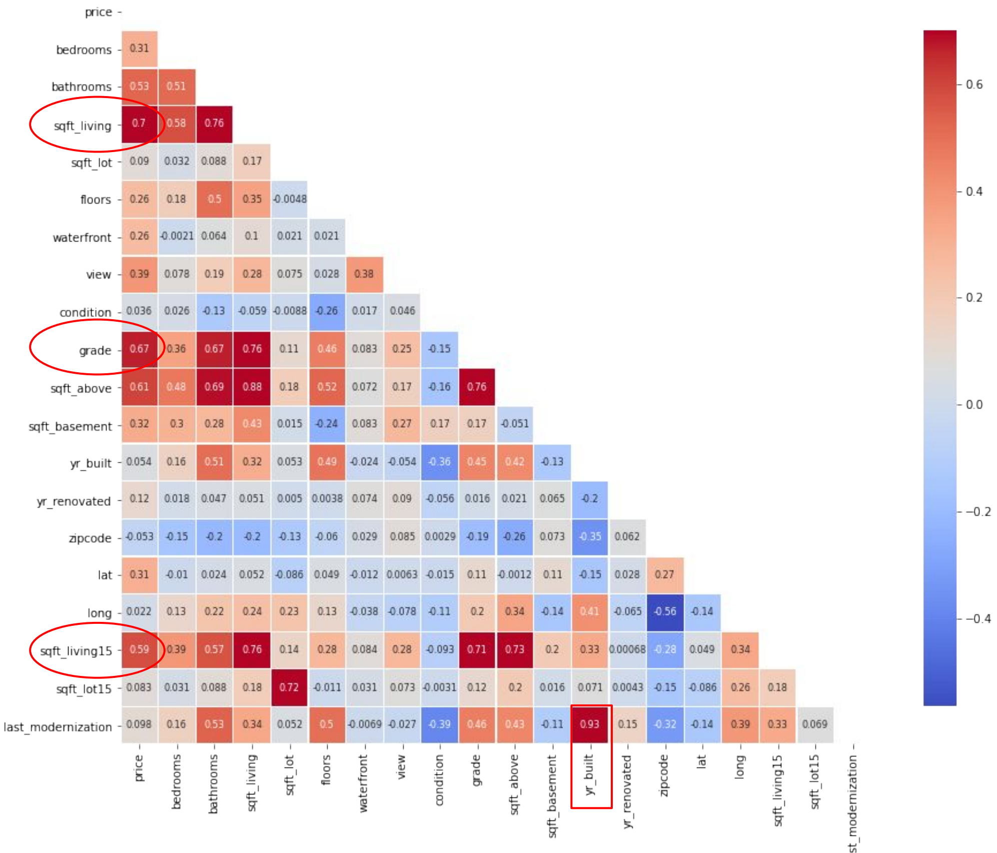
Missing values needed to be accounted for

Something is wrong with the basement measures

Fortunately, only few values were missing and could logically be replaced with zeros

Exploring the data

Exploring the data

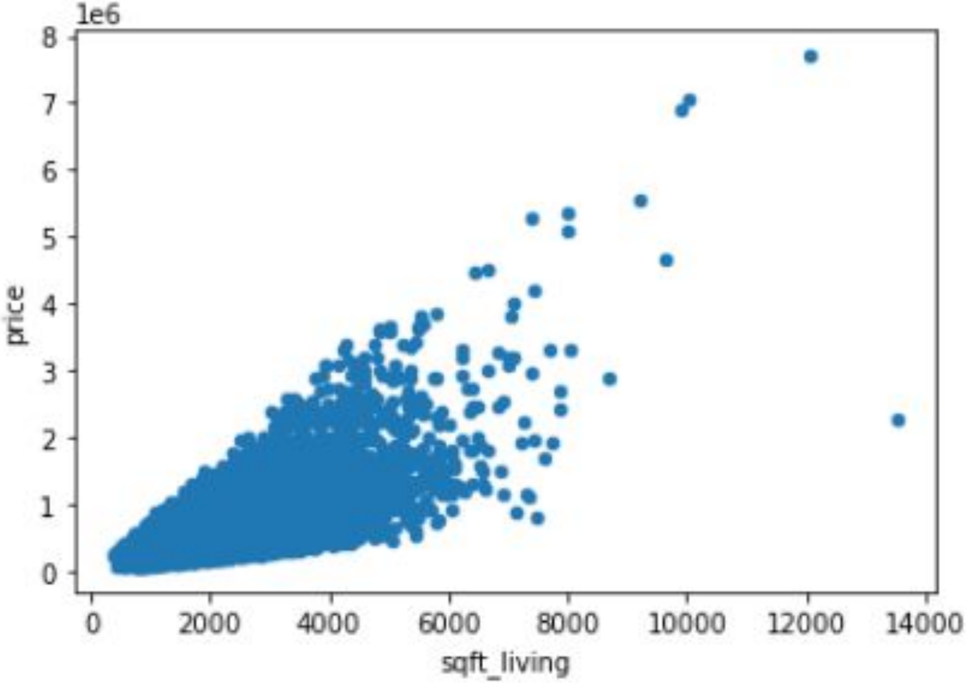


	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	st_modernization
price																				
bedrooms	0.31																			
bathrooms	0.53	0.51																		
sqft_living	0.7	0.58	0.76																	
sqft_lot	0.09	0.032	0.088	0.17																
floors	0.26	0.18	0.5	0.35	-0.0048															
waterfront	0.26	-0.0021	0.064	0.1	0.021	0.021														
view	0.39	0.078	0.19	0.28	0.075	0.028	0.38													
condition	0.036	0.026	-0.13	-0.059	-0.0088	-0.26	0.017	0.046												
grade	0.67	0.36	0.67	0.76	0.11	0.46	0.083	0.25	-0.15											
sqft_above	0.61	0.48	0.69	0.88	0.18	0.52	0.072	0.17	-0.16	0.76										
sqft_basement	0.32	0.3	0.28	0.43	0.015	-0.24	0.083	0.27	0.17	0.17	-0.051									
yr_built	0.054	0.16	0.51	0.32	0.053	0.49	-0.024	-0.054	-0.36	0.45	0.42	-0.13								
yr_renovated	0.12	0.018	0.047	0.051	0.005	0.0038	0.074	0.09	-0.056	0.016	0.021	0.065	-0.2							
zipcode	-0.053	-0.15	-0.2	-0.2	-0.13	-0.06	0.029	0.085	0.0029	-0.19	-0.26	0.073	-0.35	0.062						
lat	0.31	-0.01	0.024	0.052	-0.086	0.049	-0.012	0.0063	-0.015	0.11	-0.0012	0.11	-0.15	0.028	0.27					
long	0.022	0.13	0.22	0.24	0.23	0.13	-0.038	-0.078	-0.11	0.2	0.34	-0.14	0.41	-0.065	-0.56	-0.14				
sqft_living15	0.59	0.39	0.57	0.76	0.14	0.28	0.084	0.28	-0.093	0.71	0.73	0.2	0.33	0.00068	-0.28	0.049	0.34			
sqft_lot15	0.083	0.031	0.088	0.18	0.72	-0.011	0.031	0.073	-0.0031	0.12	0.2	0.016	0.071	0.0043	-0.15	-0.086	0.26	0.18		
last_modernization	0.098	0.16	0.53	0.34	0.052	0.5	-0.0069	-0.027	-0.39	0.46	0.43	-0.11	0.93	0.15	-0.32	-0.14	0.39	0.33	0.069	

Highest correlation with price can be observed at:

Size of living area,
grade
& neighbors size of living area

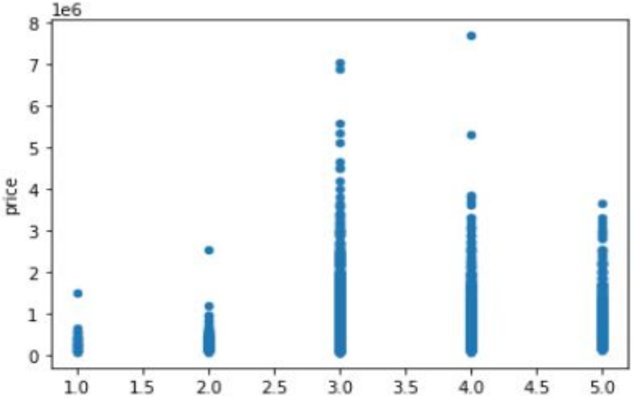
Exploring the data



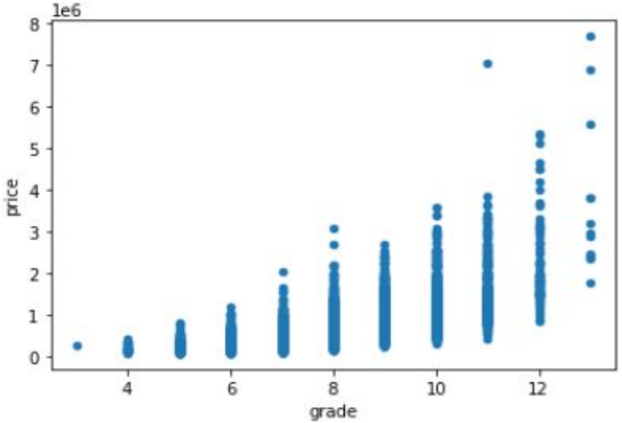
Highest correlation with price can be observed at:

Size of living area,
grade
& neighbors size of living area

Exploring the data

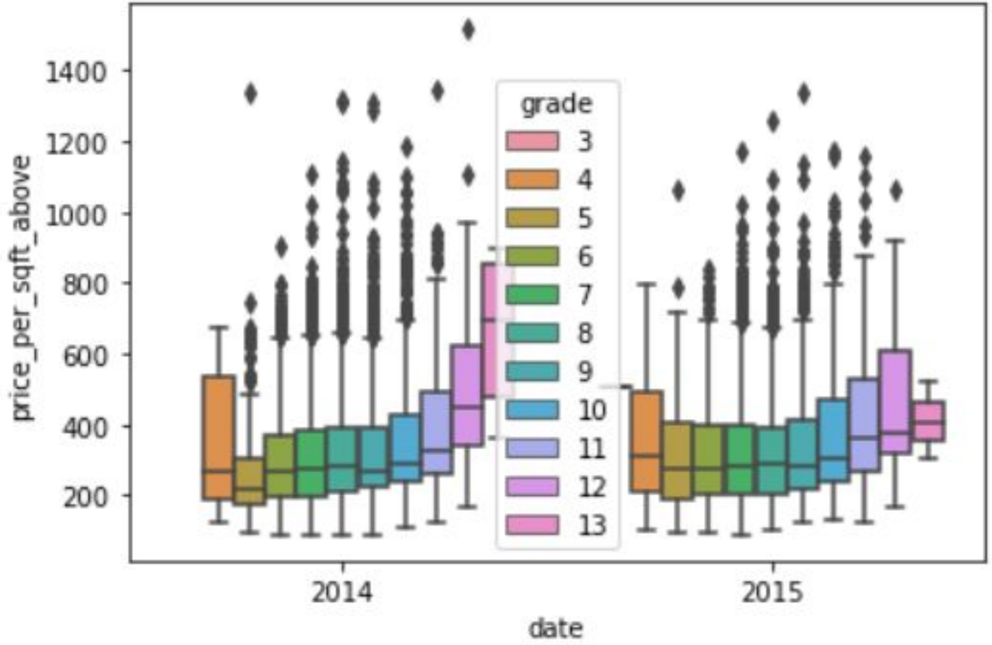


No linear relationship between price and condition rating. Condition just needs top be > 2



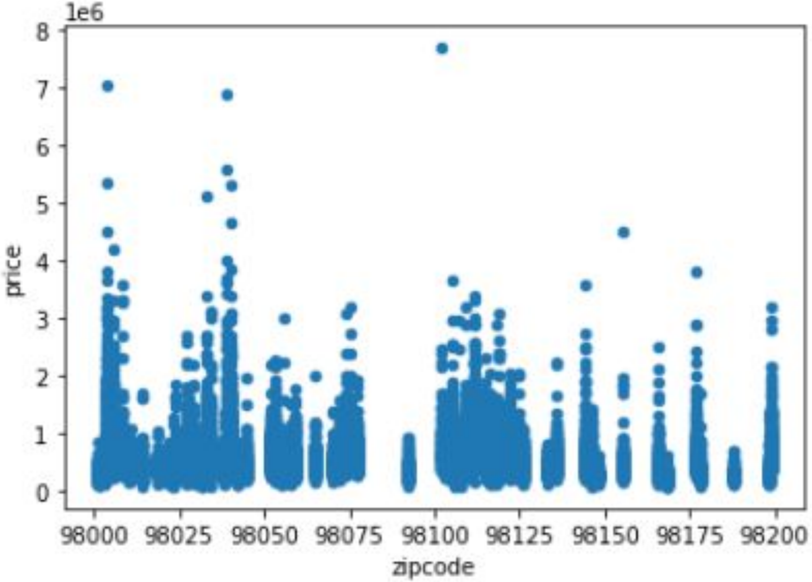
Positive correlation trend between grade and price

Exploring the data



If you want to buy luxury real estate, you might want to put a special emphasis on market timing

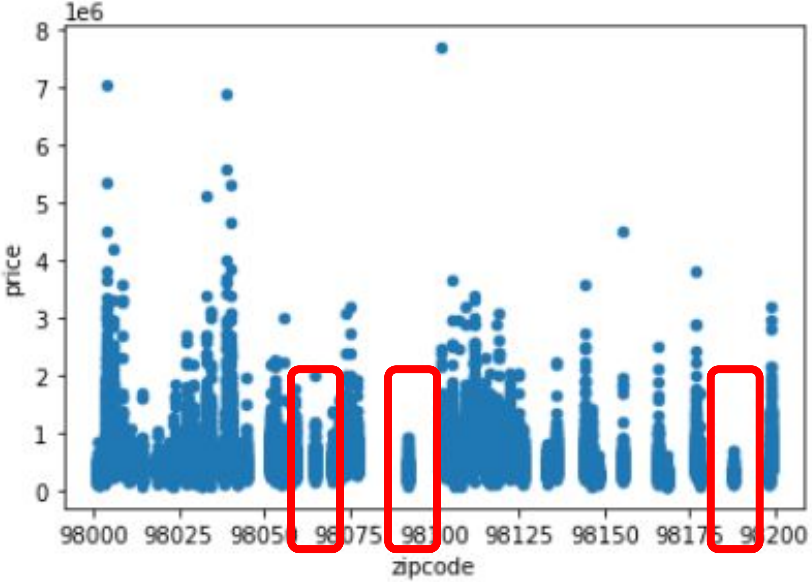
Exploring the data



Some zipcode areas seem to have really cheap house prices.

The zip code areas with best price to footage ratio are:
98092, 98002, 98030, 98001, 98023, 98042, 98003 & 98038,

Exploring the data



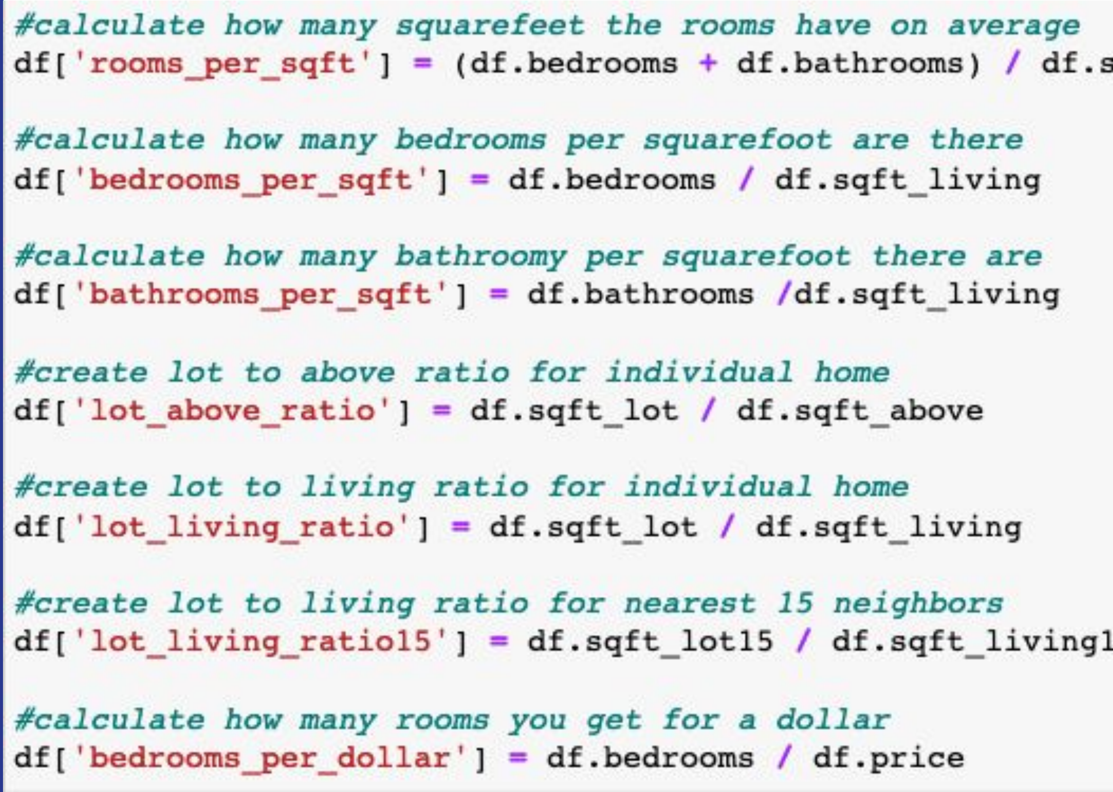
Some zipcode areas seem to have really cheap house prices.

The zip code areas with best price to footage ratio are:
98092, 98002, 98030, 98001, 98023, 98042, 98003 & 98038,

Building new features

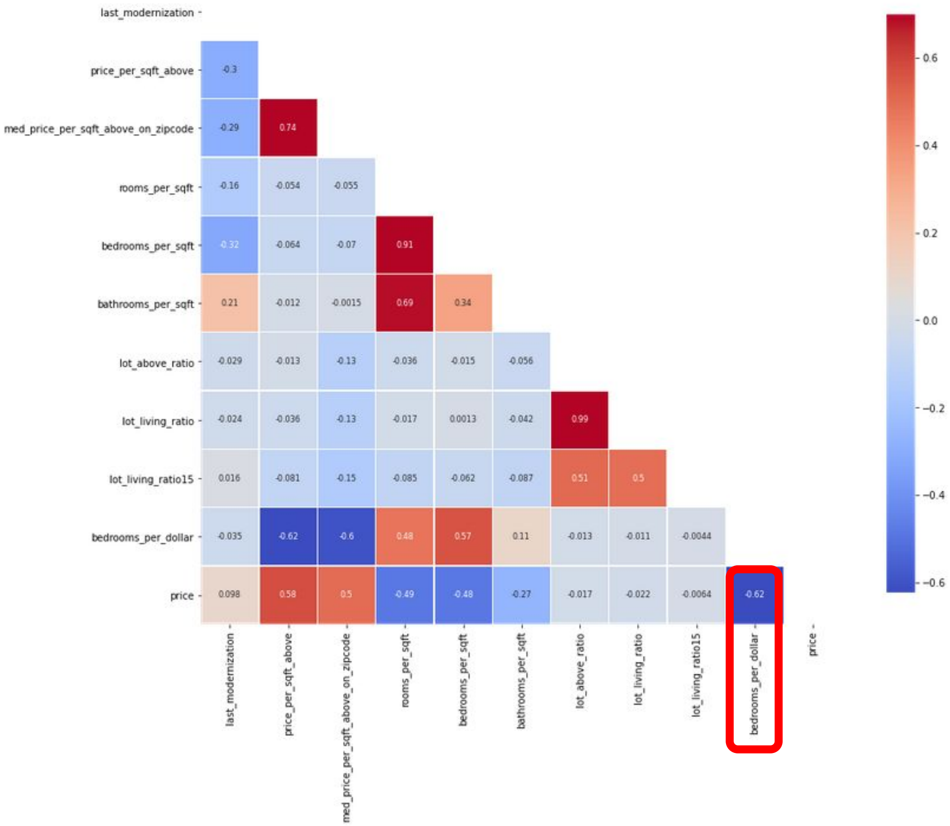
Building new features

- Size of living area seemed to be important
- I wanted to add geographical dimension as well



```
#calculate how many squarefeet the rooms have on average
df['rooms_per_sqft'] = (df.bedrooms + df.bathrooms) / df.s

#calculate how many bedrooms per squarefoot are there
df['bedrooms_per_sqft'] = df.bedrooms / df.sqft_living

#calculate how many bathroomy per squarefoot there are
df['bathrooms_per_sqft'] = df.bathrooms /df.sqft_living

#create lot to above ratio for individual home
df['lot_above_ratio'] = df.sqft_lot / df.sqft_above

#create lot to living ratio for individual home
df['lot_living_ratio'] = df.sqft_lot / df.sqft_living

#create lot to living ratio for nearest 15 neighbors
df['lot_living_ratio15'] = df.sqft_lot15 / df.sqft_living1

#calculate how many rooms you get for a dollar
df['bedrooms_per_dollar'] = df.bedrooms / df.price
```

Exploring the data



The more bedrooms you get for a dollar, the cheaper the house price.

Sellers should keep the number of rooms normal and not try to squeeze more rooms in a property that neccessary

Building a predictive model

Looking at different model qualities

Best Single Predictor

Square footage of living area

RSquared: 0.469
RMSE: $ 249,948

Price = -52,451 + 285 x area

Best Feature Combination

Square footage of living area, grade & year of construction

RSquared: 0.598
RMSE: $ 217,556

Price = 6319009
+ 182 x area
+ 144080 x grade
- 3684 x year of construction

Looking at different model qualities

Old Best Feature Combination

Square footage of living area, grade & year of construction

RSquared: 0.598
RMSE: $ 217,556

Price = 6319009
+ 182 x area
+ 144080 x grade
- 3684 x year of construction

New Best Feature Combination

Square footage of living area, median price per sqft in zip-area, waterfront

RSquared: 0.758
RMSE: $ 168,863

Price = -497856
+ 267 x area
+ 1594 x median price per sqft in zip area
+ 839,473 x waterfront

Looking at different model qualities

Old Best Feature Combination

Square footage of living area, grade & year of construction

RSquared: 0.598
RMSE: $ 217,556

Price = 6319009
+ 182 x area
+ 144080 x grade
- 3684 x year of construction

New Best Feature Combination

Square footage of living area, median price per sqft in zip-area, waterfront

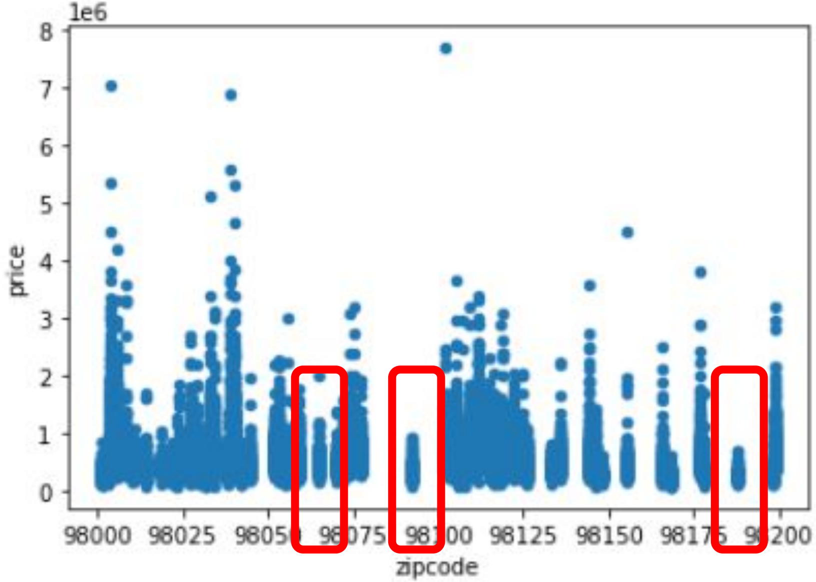
RSquared: 0.758
RMSE: $ 168,863

Not the margin of error a house buyer wants to see

Price = -497856
+ 267 x area
+ 1594 x median price per sqft in zip area
+ 839,473 x waterfront

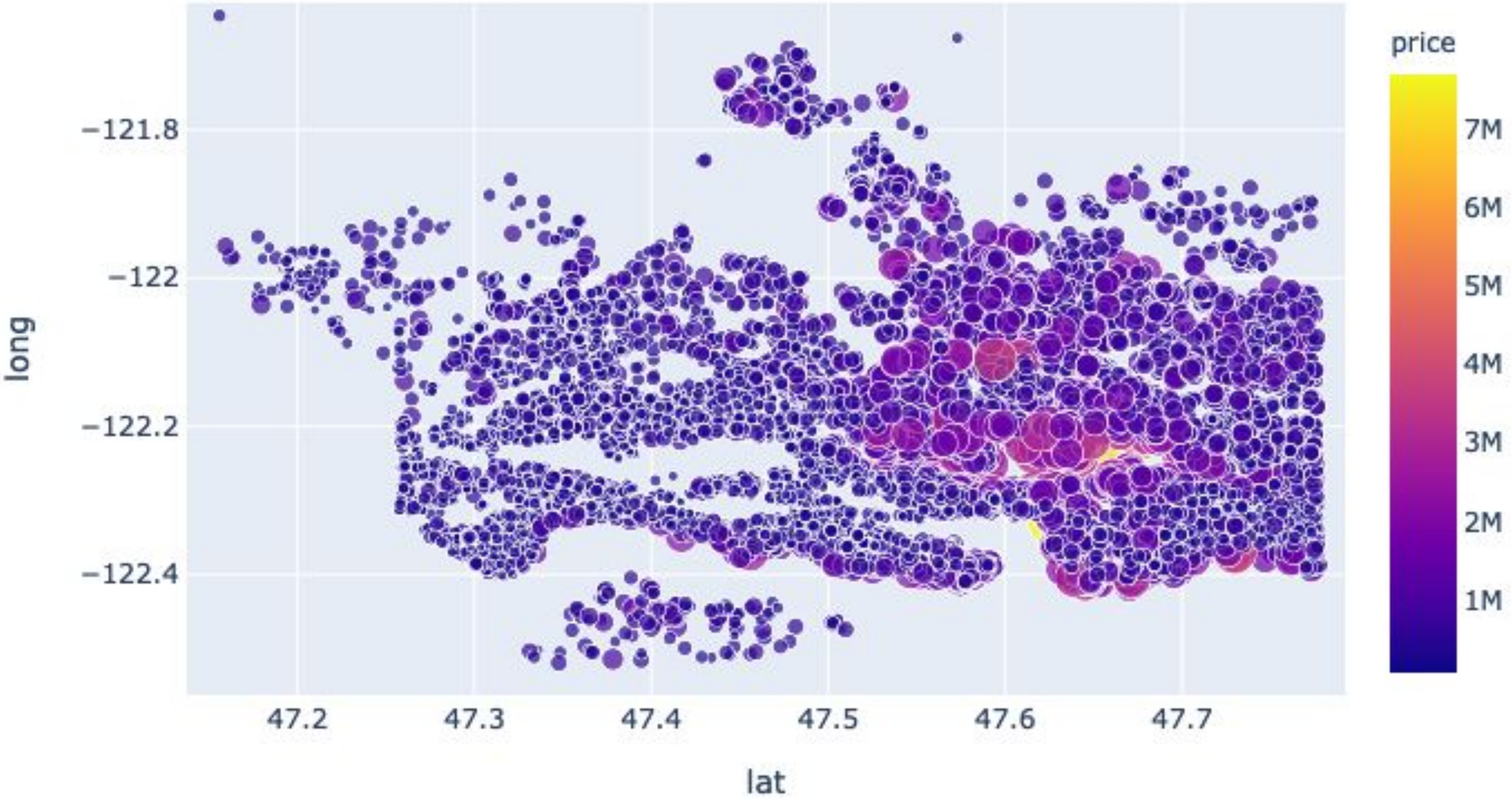
Outlook

Remember this chart?

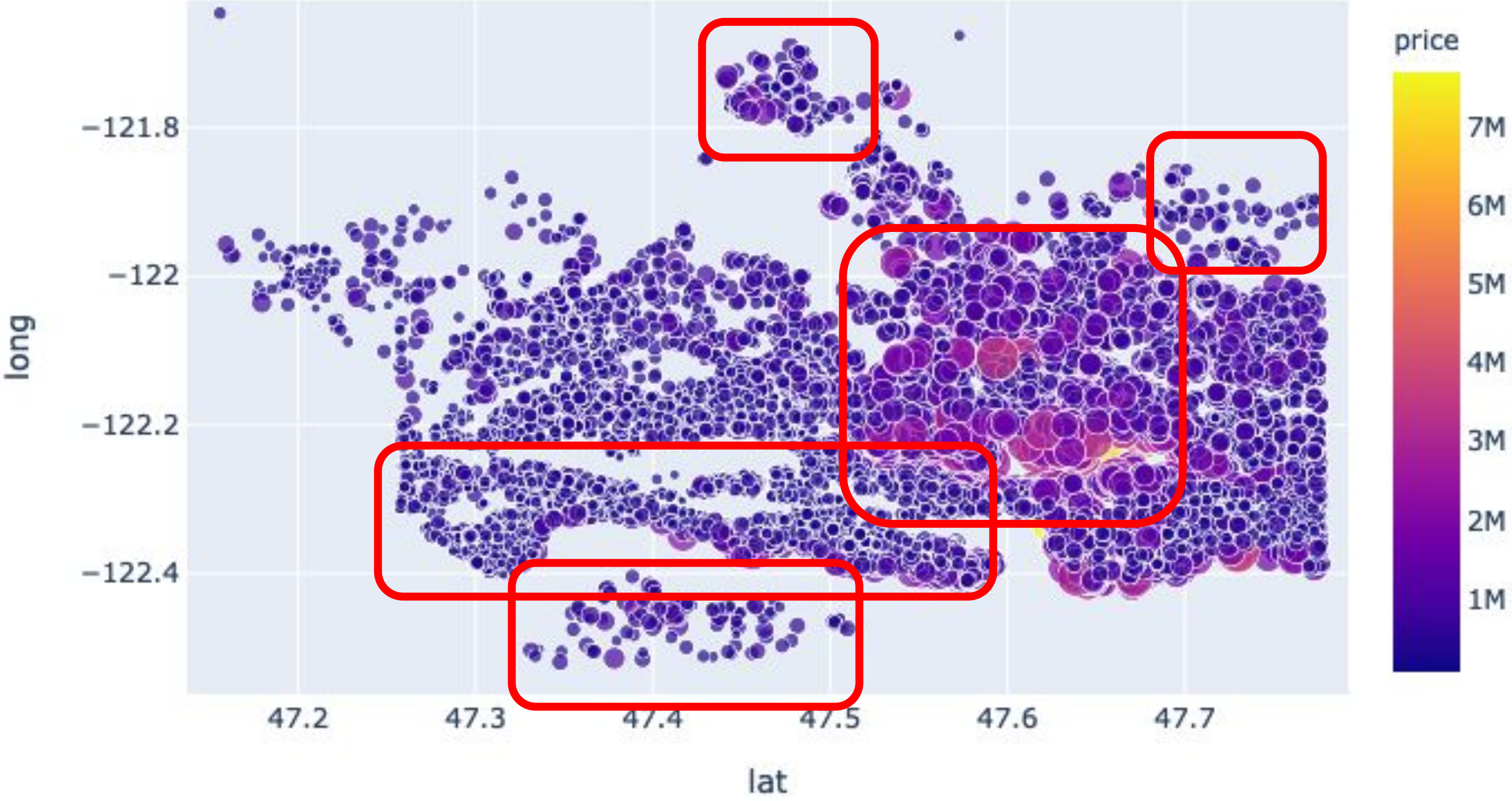


Some zipcode areas seem to have really cheap house prices.

Clustering by region could be even better feature



Clustering by region could be even better feature



Thanks for your attention