King County Housing Prices

Lia Kapanadze
April 5th, 2019

Objective

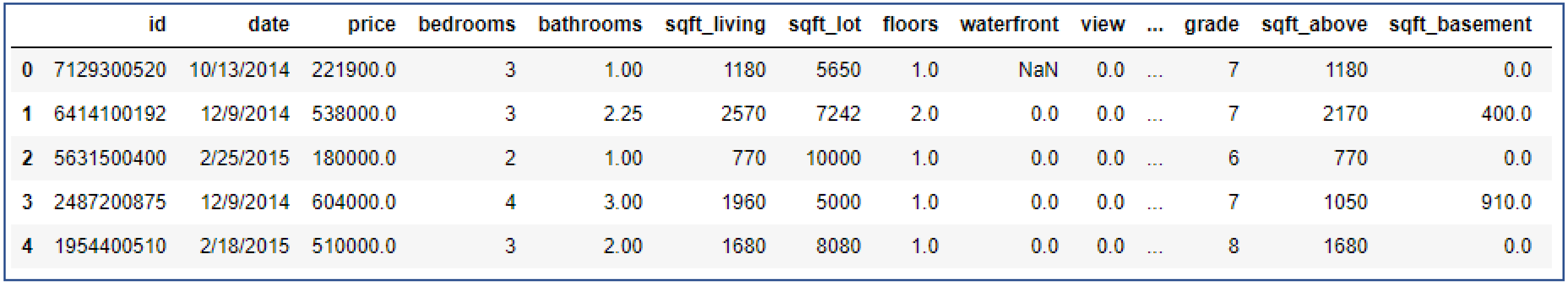
Client Inc. is looking to invest in King County real estate and wants to know what influences prices of properties in the area. They provided us with a data set containing various metrics around apartment/house sales in the County. Using this data we aim to answer the following questions:

- What are the strongest indicators of the selling price of a house?
- Is waterfront real estate more expensive?
- Can I influence the price by choosing to sell during a specific time of the year?
- Should I renovate before selling?

Data Overview

Raw data was provided in csv format containing 21 columns and 21,597 rows

RAW DATA SNAPSHOT

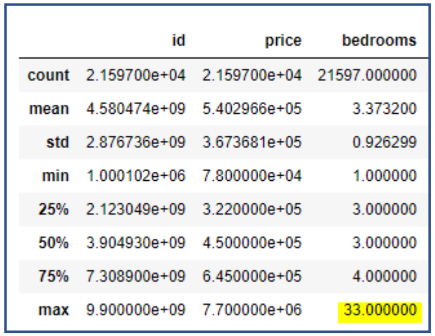


	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement
0	7129300520	10/13/2014	221900.0	3	1.00	1180	5650	1.0	NaN	0.0	...	7	1180	0.0
1	6414100192	12/9/2014	538000.0	3	2.25	2570	7242	2.0	0.0	0.0	...	7	2170	400.0
2	5631500400	2/25/2015	180000.0	2	1.00	770	10000	1.0	0.0	0.0	...	6	770	0.0
3	2487200875	12/9/2014	604000.0	4	3.00	1960	5000	1.0	0.0	0.0	...	7	1050	910.0
4	1954400510	2/18/2015	510000.0	3	2.00	1680	8080	1.0	0.0	0.0	...	8	1680	0.0

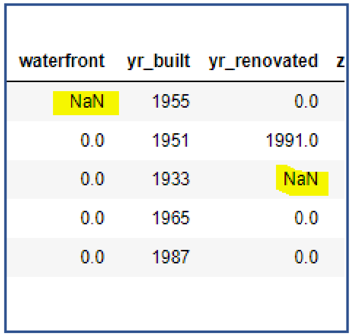
CLEAN UP

Preliminary clean up identified and eliminated the following flaws in the dataset:

- Missing data
- Wrong formatting of variables
- Typos
- Outliers



	id	price	bedrooms
count	2.159700e+04	2.159700e+04	21597.000000
mean	4.580474e+09	5.402966e+05	3.373200
std	2.876736e+09	3.673681e+05	0.926299
min	1.000102e+06	7.800000e+04	1.000000
25%	2.123049e+09	3.220000e+05	3.000000
50%	3.904930e+09	4.500000e+05	3.000000
75%	7.308900e+09	6.450000e+05	4.000000
max	9.900000e+09	7.700000e+06	33.000000



waterfront	yr_built	yr_renovated	z
NaN	1955	0.0	
0.0	1951	1991.0	
0.0	1933	NaN	
0.0	1965	0.0	
0.0	1987	0.0	

DERIVED FIELDS

- We added fields derived from existing columns
- This helped make practical sense of variables that were otherwise just numbers
- It also helped consolidate categorical variables with too many categories

date **season**

Seasons are believed to influence real estate prices

Sqft_basement **Basement Y/N**

Distinguish between having a basement or not

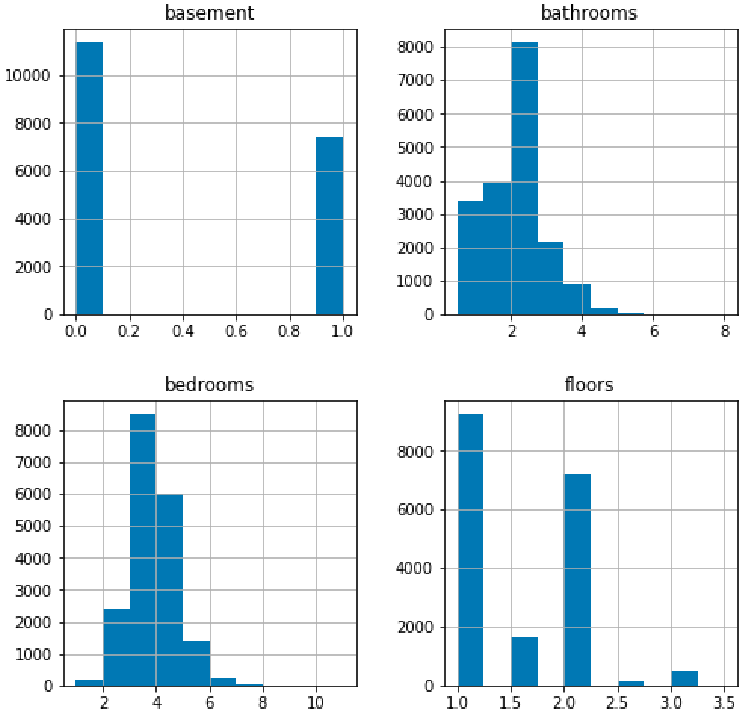
zipcode **Zip means**

Prices of neighboring houses

Preliminary Observations

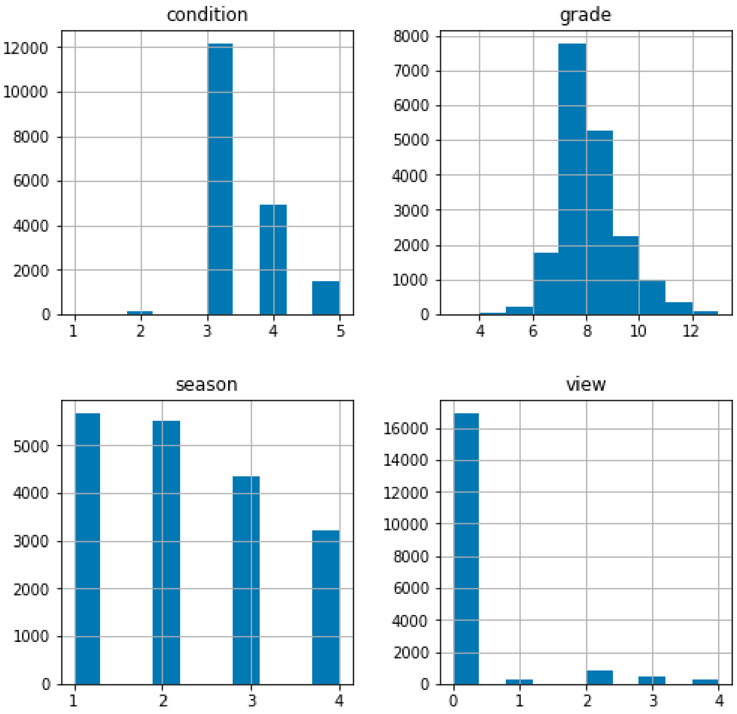
After initial clean up we have a DataFrame containing 26 columns and 18,748 rows; here are visualizations for some of the categorical descriptions of the properties

APARTMENT TYPES



- Number of bathrooms skewed
- Bedrooms normal
- Floors mostly 1 and 2
- Most apartments have no basements

GRADING & SALE

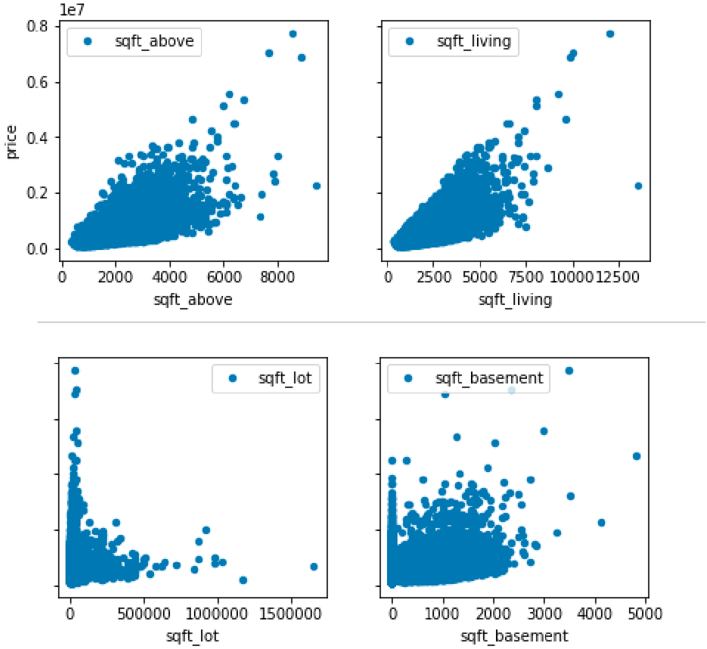


- Condition
- Grades are normally distributed, mostly around 7-8
- Spring sees more sales, but no significant surges
- Views are mostly 0

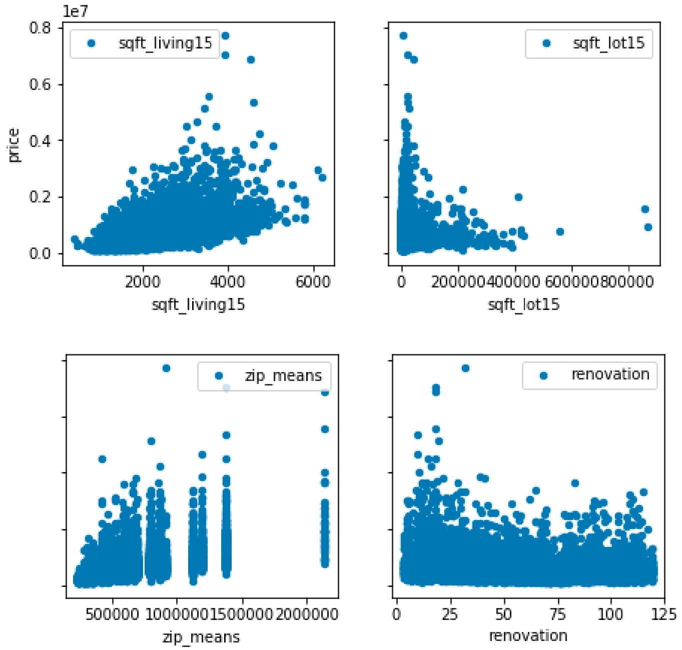
Preliminary Observations (cont.)

After initial clean up we have a DataFrame containing 26 columns and 18,748 rows; here are the visualizations of measurements for property size and other characteristics

APARTMENT SIZES



NEIGHBORS & RENOVATION



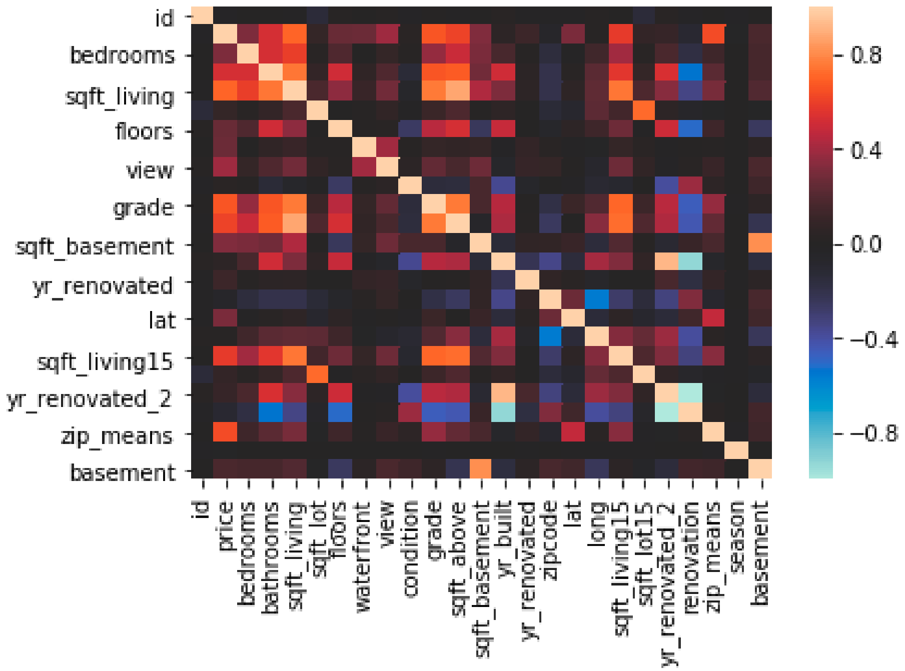
- Sqft_living has most observable relationship with the price
- All dimensional metrics are similar, except sqft_lot which shows a hint of inverse correlation

- Observable relationship between neighbors' living space and the price
- Prices of houses within same zips are indicators
- There is a hint of reverse relationship between prices and "age" of renovation

Shortlisting Relevant Predictors

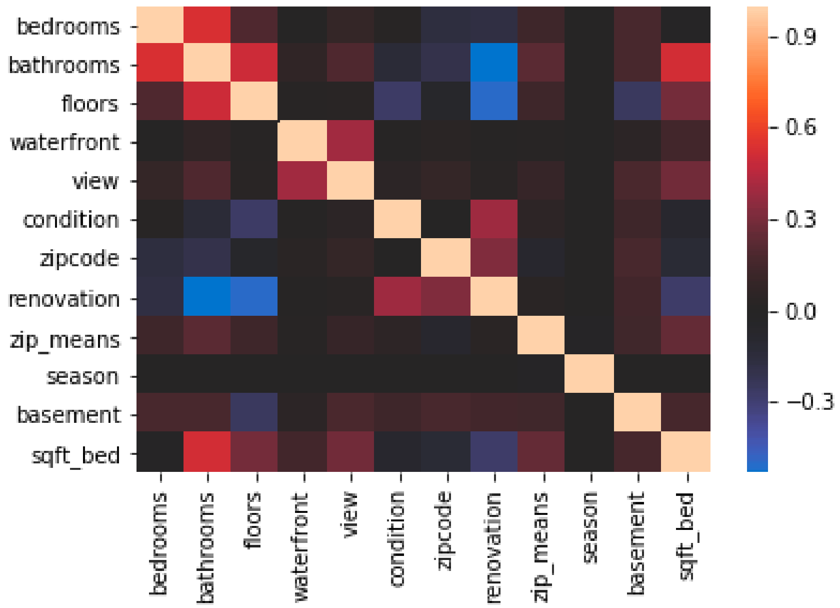
All available variables individually can serve as good inputs for predicting the direction of the price, but they overlap and may cause to distort the results of our model

ALL VARIABLES



- Sqft_X fields seem to move together, which makes intuitive sense and we also saw on scatter plots; they also overlap with # beds, # baths, # floors
- Replaced sqft_X by living sqft per bedroom to eliminate this overlap but keep the sense of size
- Removed other overlapping variables like yr_X, date
- Decided to keep condition instead of grade because move together, but grade overlaps more with others

RELEVANT PREDICTORS



- This heatmap is darker, indicating less “noise” and better chances of reliable model
- Notice how sqft_bed is not correlated with beds and bathrooms as much as sqft_living
- All dimensional metrics are similar, except sqft_lot which shows a hint of inverse correlation

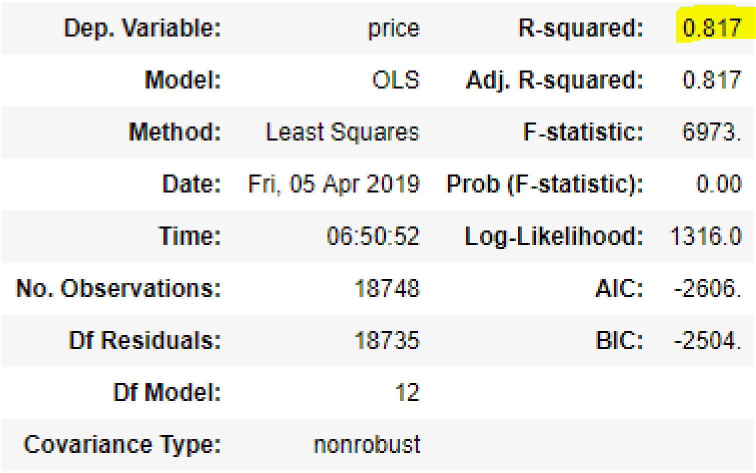
Model Output

The model gave us three good indicators of the price movement : sqft of living area, waterfront, and the avg. prices in given zipcodes

OVERVIEW

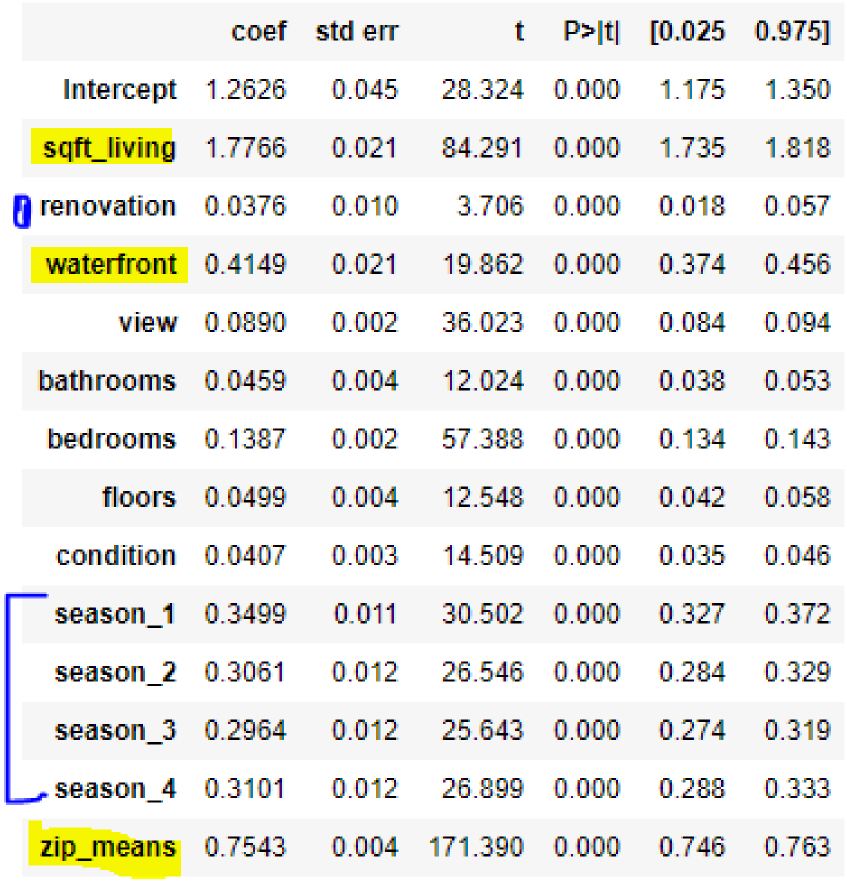
- R-squared of 0.81 means that the model accounts for 81% variance, which means it's a good model
- Higher coefficients indicate higher correlation; in this case, most likely indicators of price are sqft_living, waterfront, and zip_means (highlighted yellow)
- Less accurate predictors, despite my intuition, are the "age" of renovation and seasons when the sale happened

MODEL ACCURACY



Dep. Variable:	price	R-squared:	0.817
Model:	OLS	Adj. R-squared:	0.817
Method:	Least Squares	F-statistic:	6973.
Date:	Fri, 05 Apr 2019	Prob (F-statistic):	0.00
Time:	06:50:52	Log-Likelihood:	1316.0
No. Observations:	18748	AIC:	-2606.
Df Residuals:	18735	BIC:	-2504.
Df Model:	12		
Covariance Type:	nonrobust		

COEFFICIENTS



	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.2626	0.045	28.324	0.000	1.175	1.350
sqft_living	1.7766	0.021	84.291	0.000	1.735	1.818
renovation	0.0376	0.010	3.706	0.000	0.018	0.057
waterfront	0.4149	0.021	19.862	0.000	0.374	0.456
view	0.0890	0.002	36.023	0.000	0.084	0.094
bathrooms	0.0459	0.004	12.024	0.000	0.038	0.053
bedrooms	0.1387	0.002	57.388	0.000	0.134	0.143
floors	0.0499	0.004	12.548	0.000	0.042	0.058
condition	0.0407	0.003	14.509	0.000	0.035	0.046
season_1	0.3499	0.011	30.502	0.000	0.327	0.372
season_2	0.3061	0.012	26.546	0.000	0.284	0.329
season_3	0.2964	0.012	25.643	0.000	0.274	0.319
season_4	0.3101	0.012	26.899	0.000	0.288	0.333
zip_means	0.7543	0.004	171.390	0.000	0.746	0.763

Conclusion

FINDINGS

- The strongest indicators of a selling price are:
  - the size of the property, particularly the living area
  - whether or not it is a waterfront property
  - average prices of the properties in the same zipcode (5-digit)

- Being a waterfront property does have a positive influence on the price

- Seasonality of the sale does not indicate the direction of the price

- Did not see any evidence that renovation would increase the selling price (may be because older houses tend to be in more prestigious areas)

NEXT STEPS

- Conduct a cross-validation to further assess the reliability of the model
- Use price per sqft (instead of total) and map out expensiveness of neighborhoods
- Try to separate out renovation/built age from expensive neighborhoods
- Understand how the grading system works and revisit the variable, maybe it’s better to use it instead