Kiryl Pashkouski ds042219_mod1_proj1

Presentation_1


Kingsgate
Bainbridge Island
Clyde Hill
Medina
Bellevue
Seattle
Lake Sammamish
Eastgate
Newcastle
Ames Lak
Leaflet

1. Objectives:

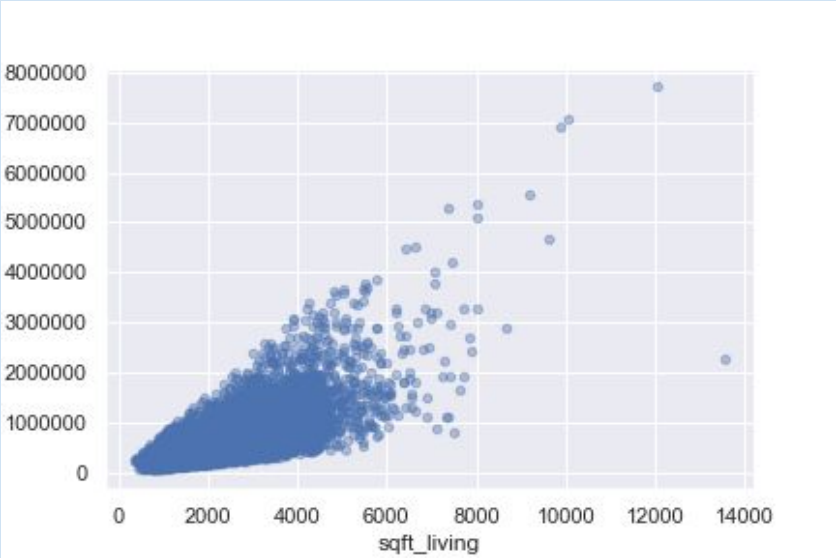
- Explore data set
- Clean data set
- Chose features as predictors
- Run model
- Interpret outcome
- Test model validation

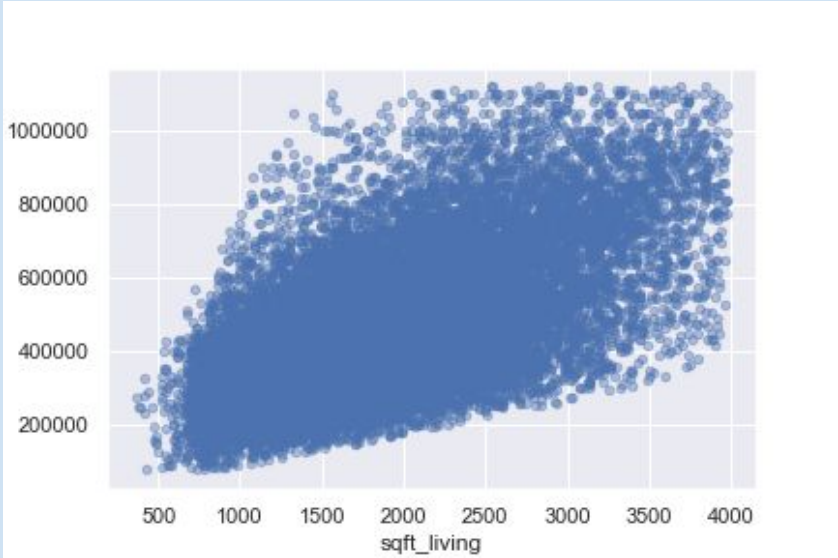
2. Quick characteristics about data set:

1. Prices of sold houses in King county, Washington in 2014-2015
2. 20 features
3. Contains over 21k entries

3. Step1: Data Cleaning Process

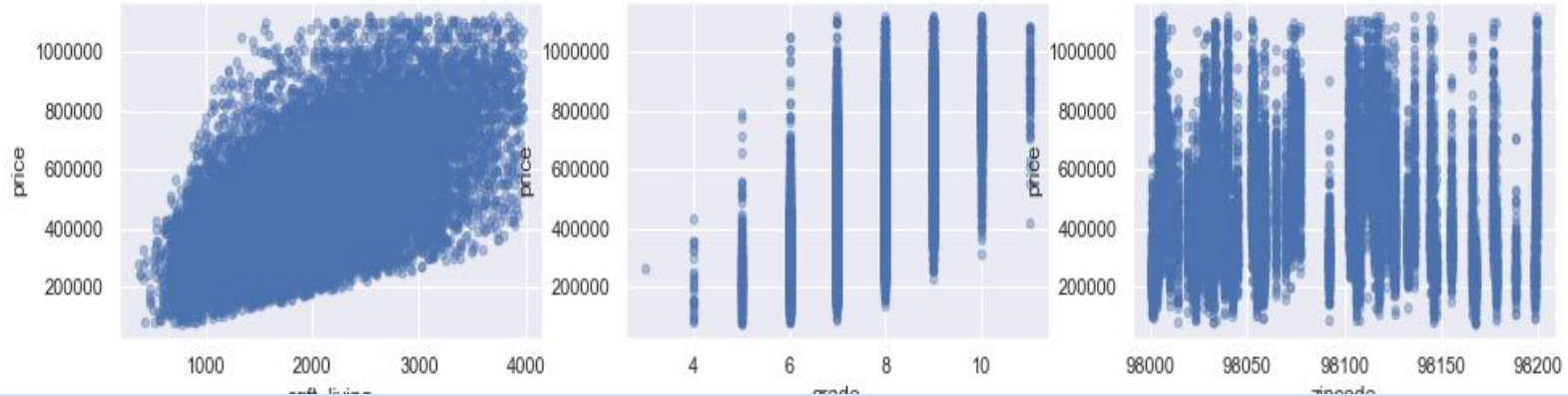
1. Kept duplicates;
2. Removed such features as ‘waterfront’, ‘view’, ‘yr_renovated’, ‘id’, ‘lat’, ‘long’, ‘date’;
3. Cut outliers





4. Step 2: Categorical variables

We use in our model such categorical variables as : **grade** and **zip codes**. And we group all zip codes whether it is in Seattle or not



5. Step 3: Multivariable Linear Regression

Run Multivariable Linear Regression:

Variables: footage of a house; what grade it has; where it located: in city or out

[85]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.492
Method:	Least Squares	**F-statistic:**	1952.
Date:	Wed, 08 May 2019	**Prob (F-statistic):**	0.00
Time:	10:52:06	**Log-Likelihood:**	-2.6759e+05
No. Observations:	20120	**AIC:**	5.352e+05
Df Residuals:	20109	**BIC:**	5.353e+05
Df Model:	10		
Covariance Type:	nonrobust		

5a Coefficients

P-value of some coefficients in our model is more than 0.05. Thus they don't have a significant impact on our model and we remove them

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.829e+05	1.05e+04	17.428	0.000	1.62e+05	2.03e+05
gr_3	-2.089e+04	1.3e+05	-0.161	0.872	-2.75e+05	2.33e+05
gr_4	-1.184e+05	2.92e+04	-4.050	0.000	-1.76e+05	-6.11e+04
gr_5	-1.282e+05	1.75e+04	-7.318	0.000	-1.63e+05	-9.39e+04
gr_6	-1.219e+05	1.57e+04	-7.777	0.000	-1.53e+05	-9.12e+04
gr_7	-5.997e+04	1.54e+04	-3.885	0.000	-9.02e+04	-2.97e+04
gr_8	2.109e+04	1.55e+04	1.361	0.173	-9278.044	5.15e+04
gr_9	1.276e+05	1.57e+04	8.104	0.000	9.68e+04	1.58e+05
gr_10	2.072e+05	1.64e+04	12.609	0.000	1.75e+05	2.39e+05
gr_11	2.763e+05	2.16e+04	12.769	0.000	2.34e+05	3.19e+05
sqft_living	112.0436	2.007	55.840	0.000	108.111	115.977
city_0	4.173e+04	5370.516	7.770	0.000	3.12e+04	5.23e+04
city_1	1.412e+05	5345.095	26.410	0.000	1.31e+05	1.52e+05

Omnibus:	1426.809	Durbin-Watson:	1.967
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1963.311
Skew:	0.616	Prob(JB):	0.00
Kurtosis:	3.907	Cond. No.	1.77e+19

6. Step 4: Final Model, Interpret Results

OUR FINAL MODEL:

Y_HAT = INTERCEPT +

+ (GR_3, GR_4, GR_5, GR_6, GR_7, GR_8, GR_9, GR_10, GR_11)*X1* +
+ *(CITY_0 + CITY_1)*X2 +
+ SQFT_LIVING*X3

Y_HAT = 182900 +

+ ((0), (-118400), (-128200), (- 121900), (-59970), (0), (127600), (207200), (276300))*X1 +
+ ((41730), (141200))*X2 +
+ 112.04*X3

where:

X1 - what grade is given to a house;

X2 - whether a house is located in city limits or not;

X3 - footage of the home, sq feet

6a Example:

If YOU’re going to sell a house which is:

1100 sq.ft, located outside the city, and graded as “9”, so predicted price will be=:

Intercept + GR_9 + CITY_0 + Sq.Liv =

182900 + 127600 + 41730 + 112.04*1100 =

475474 $

7. Step 5: Model validation

1. Perform 80/20 train/test split
2. Run model on train set
3. Run model on test set
4. Calculate and compare root mean squared errors

Train Root Mean Squared Error,$: 144427.0

Test Root Mean Squared Error, $: 144781.0

Ratio test MSE/train MSE, %: 100.25

There is no significant difference between train and test sets

THANK YOU FOR YOUR PATIENCE!