

Infographic Style

Observations

Each row is a different house bought during the period in Kings county

No missing values

Data dimensions

21,000 rows X 21 columns

Added predictors

- city
- vintage



Exploratory Analysis of Data

Based on our correlation matrix

- bathrooms
- sqft_living
- grade
- sqft_above
- sqft_living15

are highly correlated with price

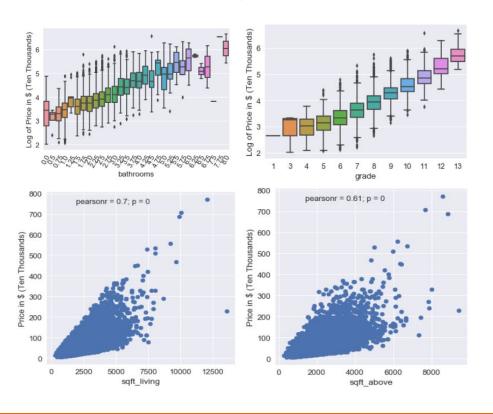
Pearson Correlation Matrix																			
price :	1.00	0.70	0.67									0.26	0.13	0.09	0.08	0.05	0.04	0.02	-0.05
sqft_living	0.70	1.00	0.76	0.88	0.76	0.75				0.05	0.10		0.06	0.17	0.18		-0.06		-0.20
grade '	0.67	0.76	1.00	0.76	0.71	0.66	0.25	0.17		0.11	0.08		0.01	0.11	0.12		-0.14		-0.18
sqft_above	061	0.88	0.76	1.00	0.73	0.69	0.17	-0.05	0.48	-0.00	0.07		0.02	0.18	0.19		-0.16		-0.26
sqft_living15	0.59	0.76	0.71	0.73	1.00			0.20		0.05	0.09	0.28	-0.00	0.14	0.18		-0.09		-0.28
bathrooms *		0.75	0.66	0.69	0.57	1.00	0.19			0.02	0.06		0.05	0.09	0.09		0.12		-0.20
view ·	0.40	0.28	0.25	0.17	0.28	0.19	1.00	0.28	0.08	0.01		0.03	0.10	0.07	0.07	-0.05	0.05	-0.08	0.08
sqft_basement	0.32			-0.05		0.28	0.28	1.00	0.30		80.0	-0.25	0.07	0.02	0.02	-0.13	0.17	-0.14	0.07
bedrooms	0.31		0.36				80.0	0.30	1.00	-0.01	-0.01	0.18	0.02	0.03	0.03	0.15	0.03	0.13	-0.15
lat ·	0.31	0.05	0.11	-0.00	0.05	0.02	0.01	0.11	-0.01	1.00	-0.01	0.05	0.03	-0.09	-0.09	-0.15	-0.01	0.14	0.27
waterfront -	0.27	0.10	0.08	0.07	0.09	0.06		0.08	-0.01	-0.01	1.00	0.02	0.09	0.02	0.03	-0.03	0.02	-0.04	0.03
foors	0.26				0.28		0.03	-0.25	0.18	0.05	0.02	1.00	0.01	-0.01	-0.01		0.26	0.13	-0.06
yr_renovated	0.13	0.06	0.01	0.02	0.00	0.05	0.10	0.07	0.02	0.03	0.09	0.01	1.00	0.01	0.01	-0.22	-0.06	-0.07	0.06
sqft_lot -	0.09		0.11	0.18	0.14	0.09	0.07	0.02	0.03	-0.09	0.02	-0.01	0.01	1.00	0.72	0.05	-0.01		-0.13
sqft_lot15	80.0	0.18	0.12	0.19	0.18	0.09	0.07	0.02	0.03	-0.09	0.03	0.01	0.01	0.72	1.00	0.07	-0.00		-0.15
yr_built	0.05					0.51	-0.05	-0.13	0.15	-0.15	-0.03	0.49	-0.22	0.05	0.07	1.00	-0.36		-0.35
condition	0.04	-0.06	0.14	-0.16	-0.09	-0.12	0.05	0.17	0.03	-0.01	0.02	-0.26	-0.06	-0.01	-0.00	-0.36	1.00	-0.11	0.00
long	0.02	0.24	0.20			0.22	-0.08	-0.14	0.13	-0.14	-0.04	0.13	-0.07		0.25	0.41	-0.11	1.00	-0.56
zipcode '	-0.05	-0.20	0.18	-0.26	-0.28	-0.20	80.0	0.07	-0.15		0.03	-0.06	0.06	-0.13	-0.15	-0.35	0.00	-0.56	1.00
	price -	sqft_living -	grade -	adif_above	sqfi_living15 -	bathrooms	wew.	sqft_basement -	bedrooms -	- ti	waterfront	foors -	y_renovated	sqft_lot -	sqft_lot15	y_bulk	condition -	- bug	zpcode -

Dograph Correlation Matrix

0.9

Exploratory Analysis

of highly correlated variables



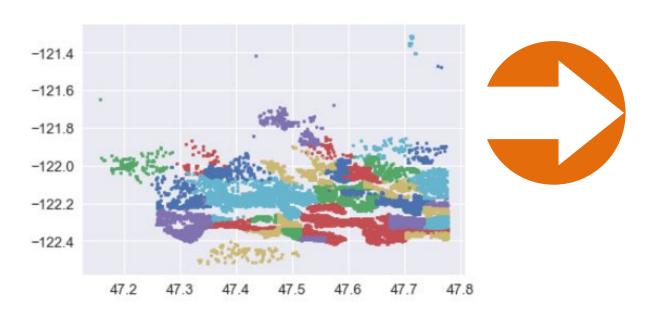
Bathroom & Grade

A general linear and upward relationship between log of price(in ten thousands)

sqft_living & sqft_above

a positive relationship

Adding New Features



Based on the zipcode we have (70 unique values). We create a column called 'city' so that houses can be classified by city

'City' column has 36 unique values

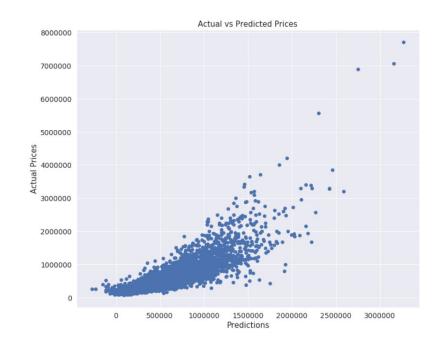
Data Modeling

Multiple Linear Regression is used to predict house prices based on all the other variables excluding date and ID number

For prediction on testing data, 70% of observations are used to set up the model

OLS Regression Results

Dep. Variable:	price	R-squared:	0.702		
Model:	OLS	Adj. R-squared:	0.702		
Method:	Least Squares	F-statistic:	2992.		
Date:	Sun, 09 Dec 2018	Prob (F-statistic):	0.00		
Time:	23:48:01	Log-Likelihood:	-2.9452e+05		
No. Observations:	21613	AIC:	5.891e+05		
Df Residuals:	21595	BIC:	5.892e+05		
Df Model:	17				
Covariance Type:	nonrobust				





Results

- Using our multiple linear regression, we were able to correctly predict 70% of the training data
- Our model successfully explains 70% of the variations in the testing data
- All variables significant
- Additional variables could help further research