STATS 140 Final Project
Predicting House Sales in King County, USA

Introduction - Data Background

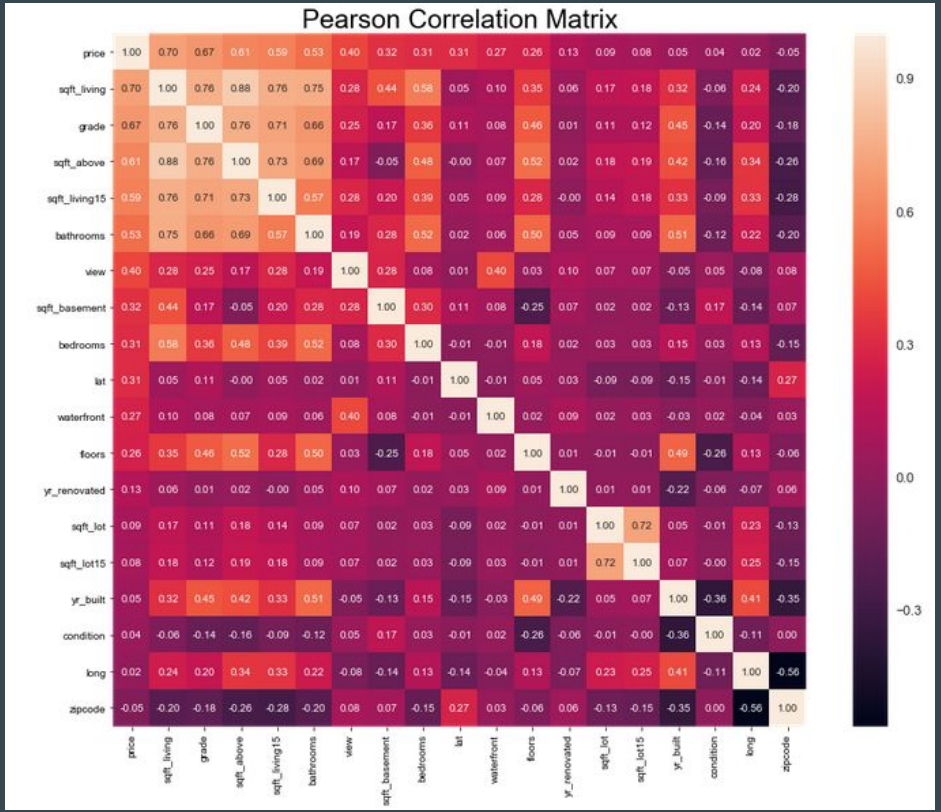
- Each observation is a different house bought during the period in Kings County
- 21,000 Rows, 21 columns
- No missing values
- Added city and vintage columns, because city and how old the house is are generally important predictors for price

Exploratory Analysis of Data

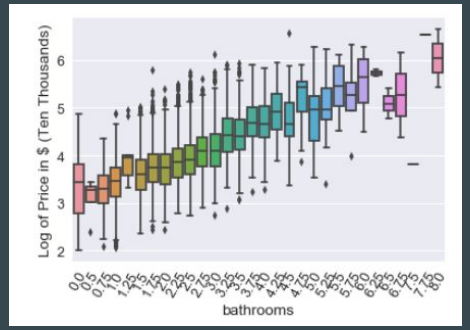
Based on the correlation matrix,

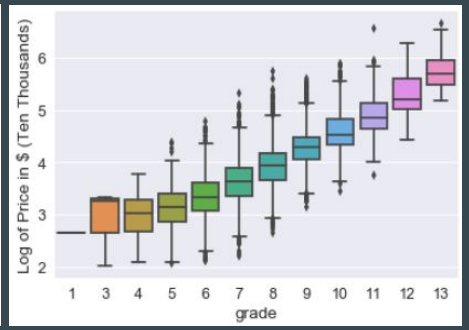
bathrooms
sqft_living
grade
sqft_above
sqft_living15

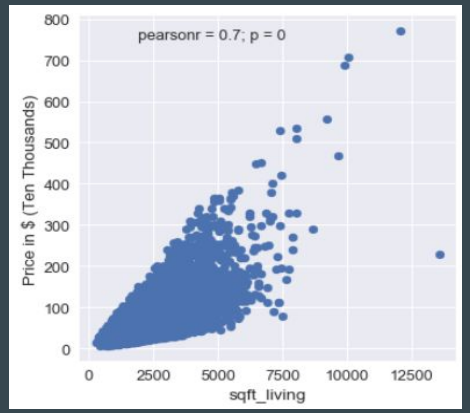
following variables are highly correlated to price.

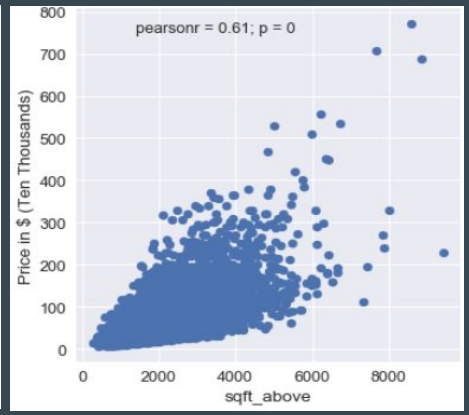


Exploratory Analysis of Data with highly correlated variables



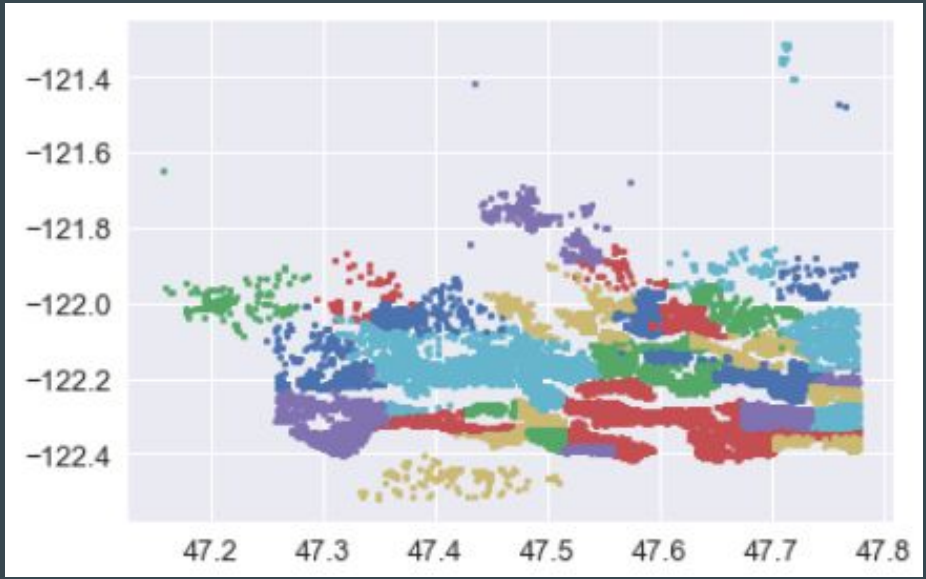






- From our graph, we see a general linear and upward relationship between log of price (in ten thousands) and bathrooms.
- We also see positive relationship between grade, sqft_living and sqft_above

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.

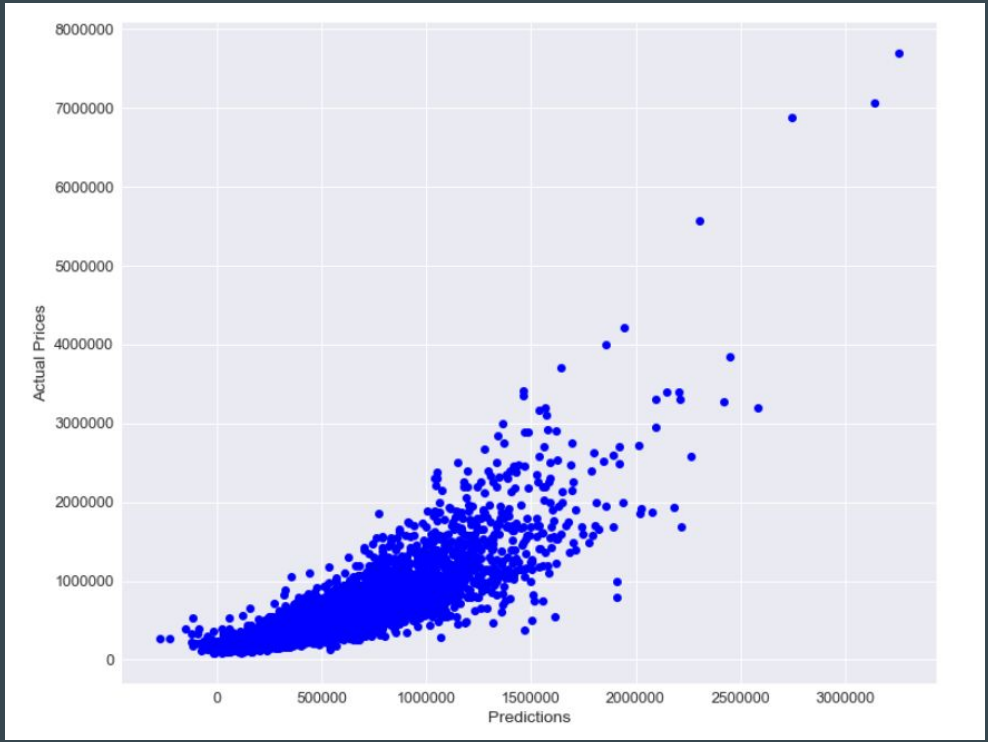
Data Modeling

- Multiple Linear Regression is used to predict house prices based on all the other variables excluding date and ID number.
- 70% of observations are used for training data, and 30% are used for testing data.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.762
Model:	OLS	Adj. R-squared:	0.762
Method:	Least Squares	F-statistic:	1355.
Date:	Mon, 10 Dec 2018	Prob (F-statistic):	0.00
Time:	22:58:19	Log-Likelihood:	-2.9208e+05
No. Observations:	21613	AIC:	5.843e+05
Df Residuals:	21561	BIC:	5.847e+05
Df Model:	51		
Covariance Type:	nonrobust		

Fitting predictions on Testing Data



The linear model predicts the house prices on testing data pretty well, but it is likely to overestimate the prices since there are lots of observations spotted on the bottom-right side.