

The background is an aerial photograph of a city skyline, likely New York City, with numerous skyscrapers and a body of water visible. A large, semi-transparent orange circle is centered over the image. Inside this circle, the text "STAT 131" and "FINAL PROJECT" is written in white, bold, sans-serif font. Below this, the text "PREDICTING HOUSE SALES IN KING COUNTY, USA" is written in a smaller, white, sans-serif font. At the bottom of the circle, the text "Python_Boiz.append(Girl)™" is written in a white, sans-serif font. Below the circle, the names of the team members are listed in a white, sans-serif font.

STAT 131 FINAL PROJECT

PREDICTING HOUSE SALES
IN KING COUNTY, USA

Python_Boiz.append(Girl)™

Annie Choi, Seungwoo Hong, Larny Lopez, Tyler “Python Boi” Rodriguez

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

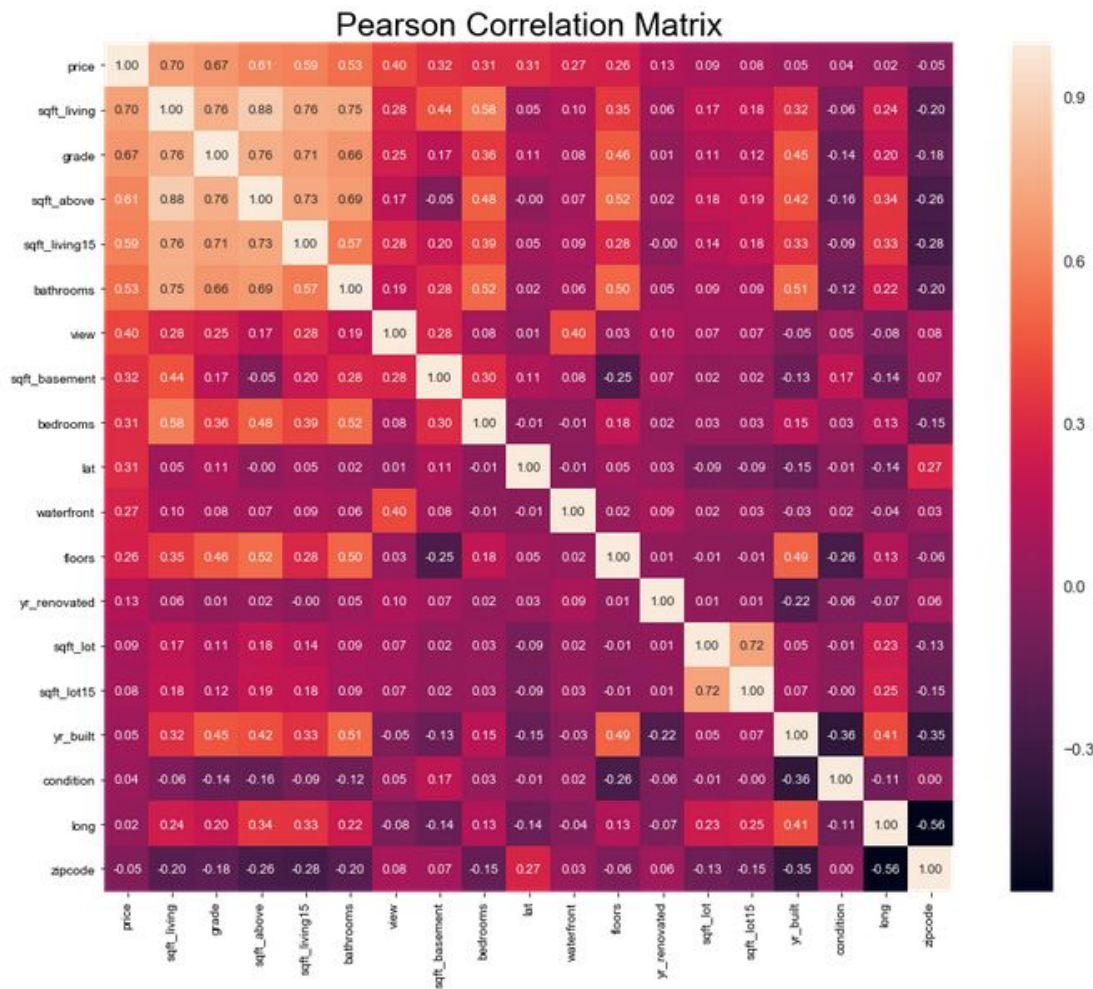


Exploratory Analysis of Data

Based on our correlation
matrix

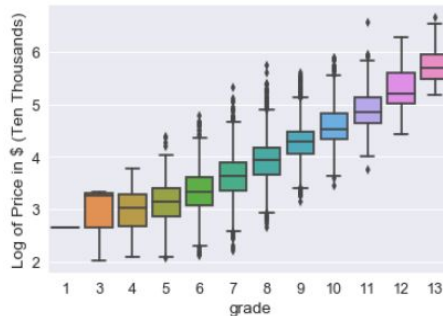
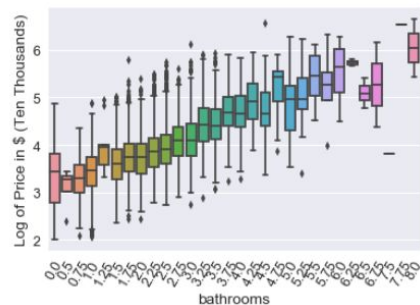
- bathrooms
- sqft_living
- grade
- sqft_above
- sqft_living15

are highly correlated with
price



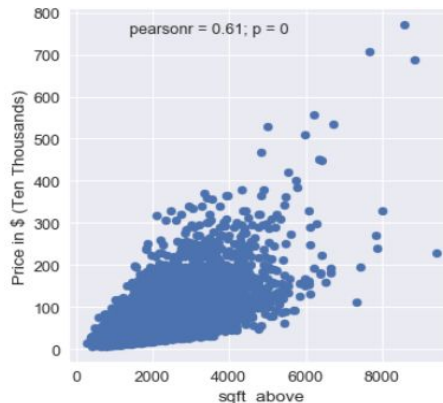
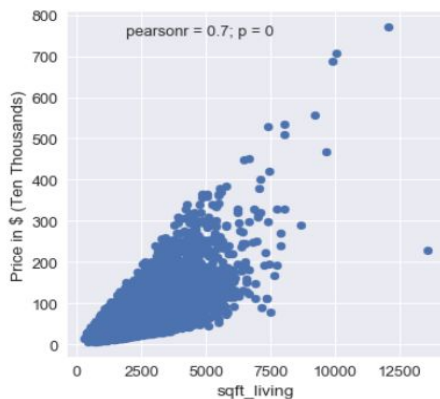
Exploratory Analysis

of highly correlated variables



Bathroom & Grade

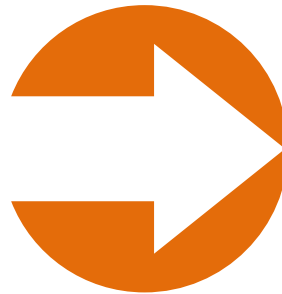
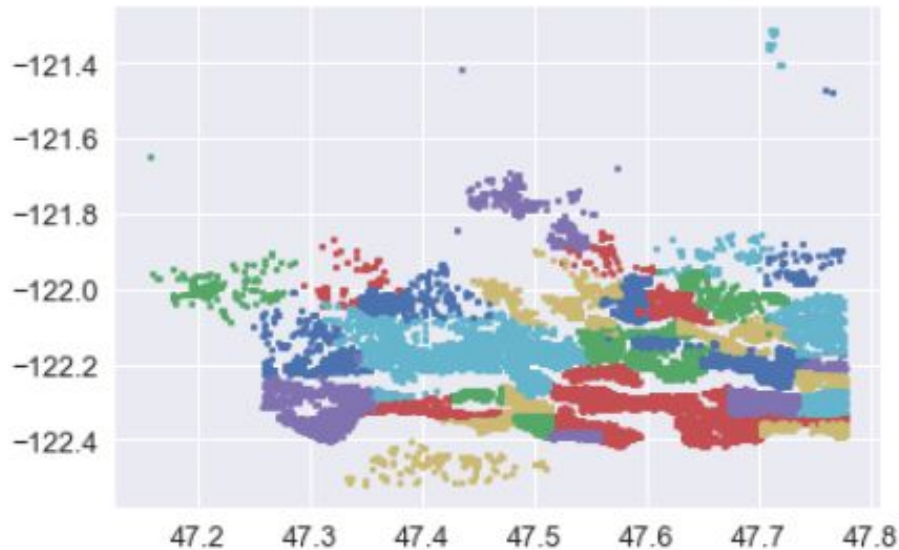
A general linear and upward relationship between log of price(in ten thousands)



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

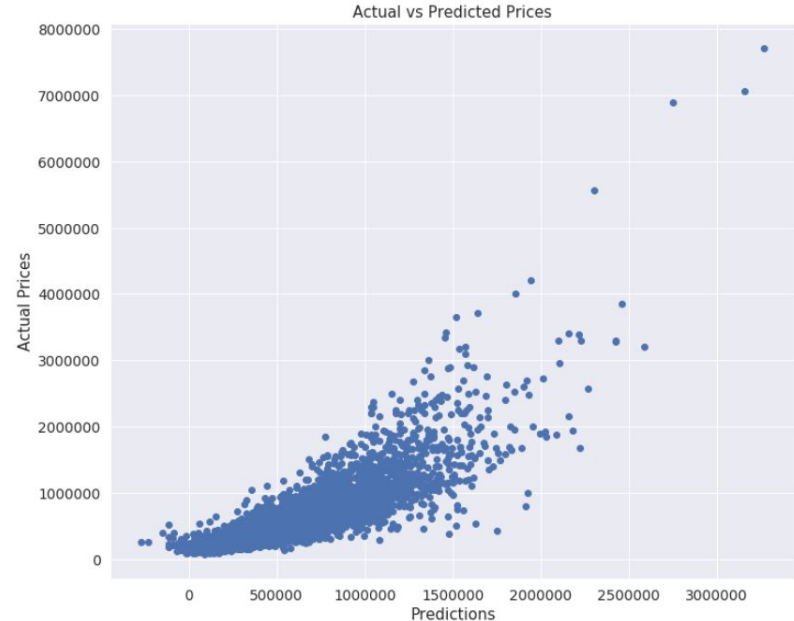
Data Modeling

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