



Module 1 Project

Seeta Rajpara
Self-Paced, Online Data Science Program
Flatiron School

King County House Sales Data Analysis Strategy

Upon examination of the column names from this dataset, I wanted to explore the variables further to understand the data structure of each column

Using basic deduction, the following column names warranted a deeper investigation into the relationship between that variable and the price of the house:

- Bedrooms, bathrooms, sqft_(living, basement, above, lot) floors, grade, yr_built, and zip code

```
In [4]: df_housedata.columns
```

```
Out[4]: Index(['id', 'date', 'price', 'bedrooms', 'bathrooms', 'sqft_living',  
              'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade',  
              'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode',  
              'lat', 'long', 'sqft_living15', 'sqft_lot15'],  
             dtype='object')
```

Categorical Variables (boxplot visualization):

- Bedrooms, bathrooms, grade, year built, zip code, and floors

Quantitative Variables (scatterplot visualization):

- square footage (living, basement, above, lot)

Descriptive Statistics on Dataset

```
In [4]: df_housedata.describe()
```

Out[4]:

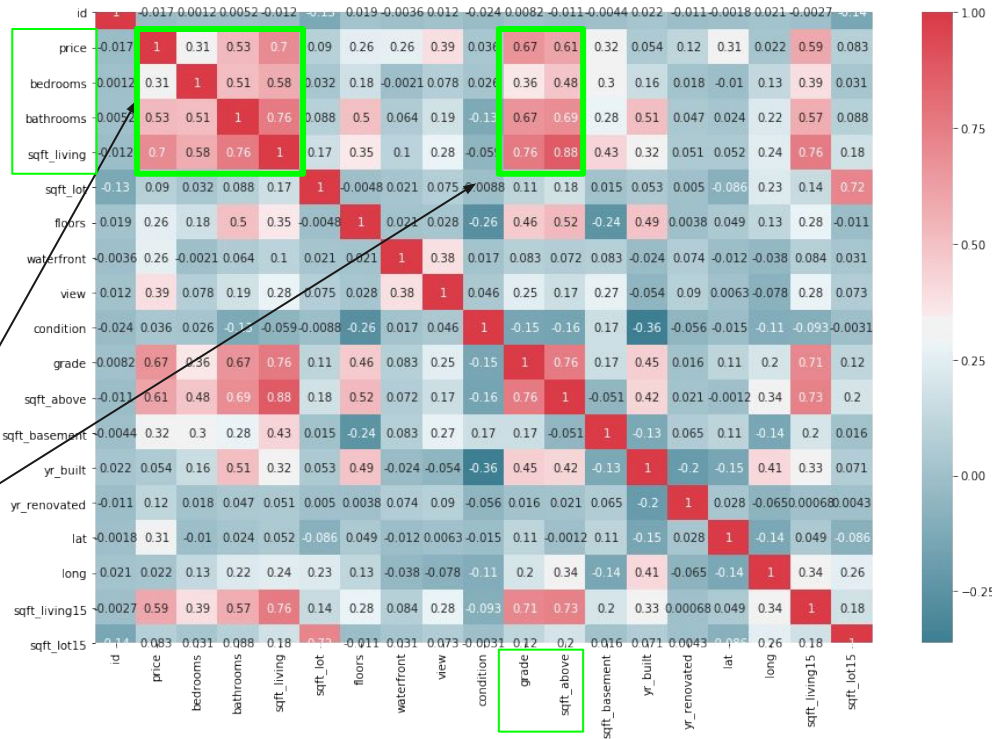
	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above
count	2.159700e+04	2.159700e+04	21597.000000	21597.000000	21597.000000	2.159700e+04	21597.000000	19221.000000	21534.000000	21597.000000	21597.000000
mean	4.580474e+09	5.402966e+05	3.373200	2.115826	2080.321850	1.509941e+04	1.494096	0.007596	0.233863	3.409825	7.650000
std	2.876736e+09	3.673681e+05	0.926299	0.768984	918.106125	4.141264e+04	0.539683	0.086825	0.765686	0.650546	1.170000
min	1.000102e+06	7.800000e+04	1.000000	0.500000	370.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000	3.000000
25%	2.123049e+09	3.220000e+05	3.000000	1.750000	1430.000000	5.040000e+03	1.000000	0.000000	0.000000	3.000000	7.000000
50%	3.904930e+09	4.500000e+05	3.000000	2.250000	1910.000000	7.618000e+03	1.500000	0.000000	0.000000	3.000000	7.000000
75%	7.308900e+09	6.450000e+05	4.000000	2.500000	2550.000000	1.068500e+04	2.000000	0.000000	0.000000	4.000000	8.000000
max	9.900000e+09	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	3.500000	1.000000	4.000000	5.000000	13.000000

- the most expensive property is \$7.7m
- 75% of the properties fall below \$645,000
- max number of bathrooms is 8
- max number of bedrooms is 33 (outlier)
- median number of bedrooms is 3
- 75% of the properties are at 2.5 bathrooms or below
- these values are of concern because they skew the mean

Pearson Correlation Heatmap

Pearson correlations show how variables relate to one another

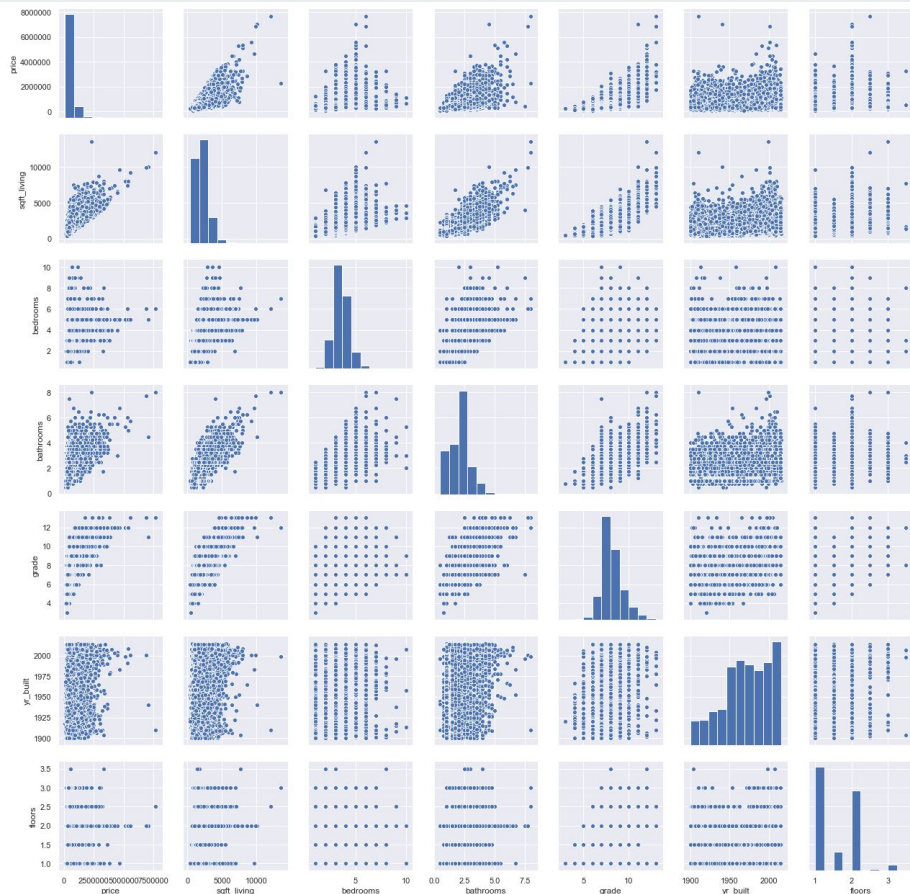
In this heatmap, it can be inferred that the variables from the previous slide have some relationship with the 'price' variable



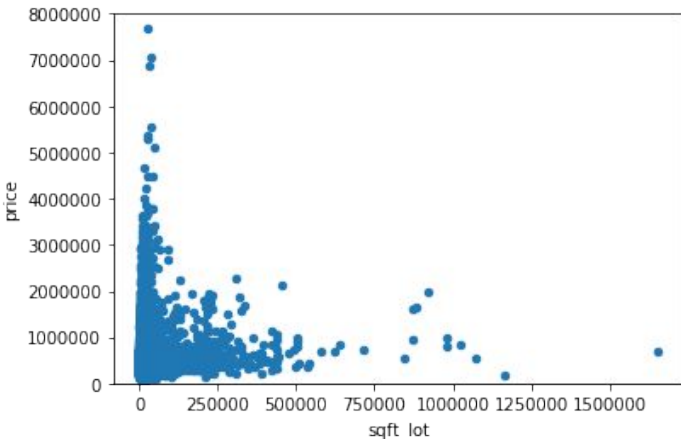
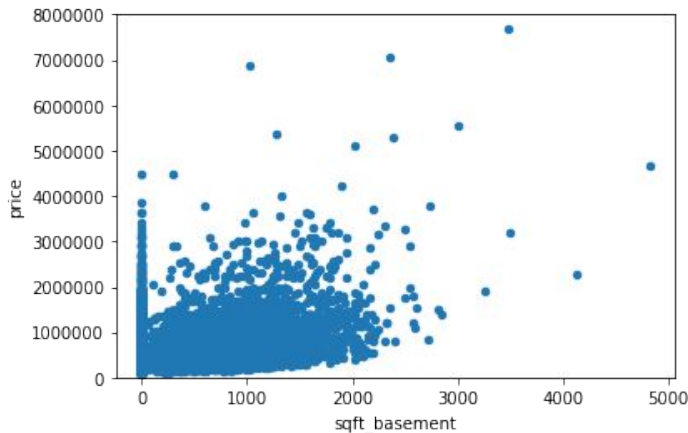
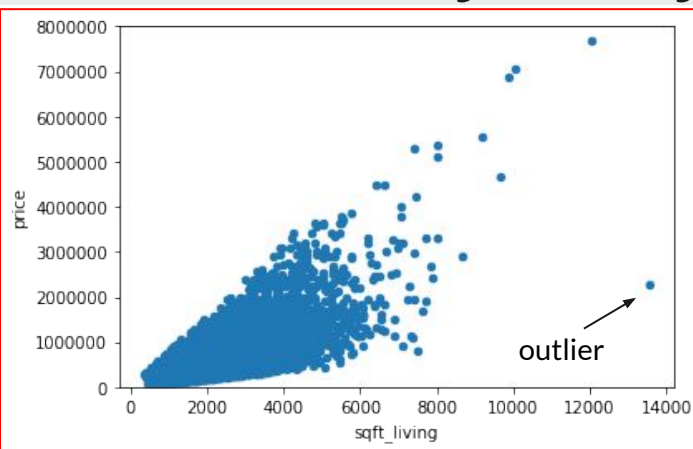
Correlation

Between Variables

- This plot set confirms the previous heatmap showing correlations between price and the following: sqft_living, bathrooms, and grade

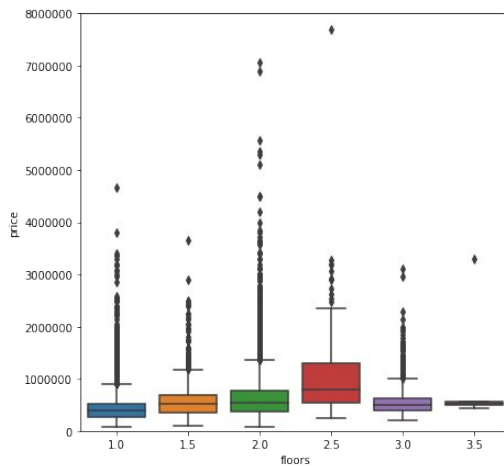
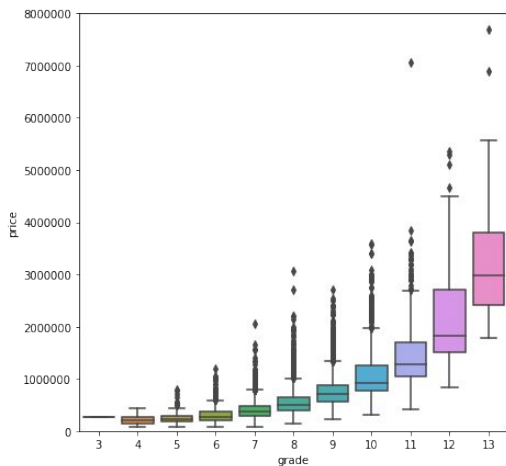


Summary: Analysis of Quantitative Variables

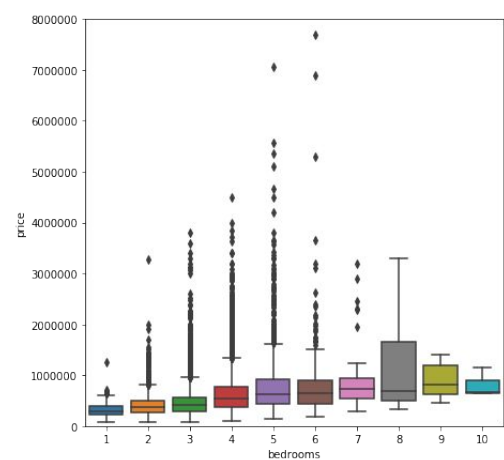
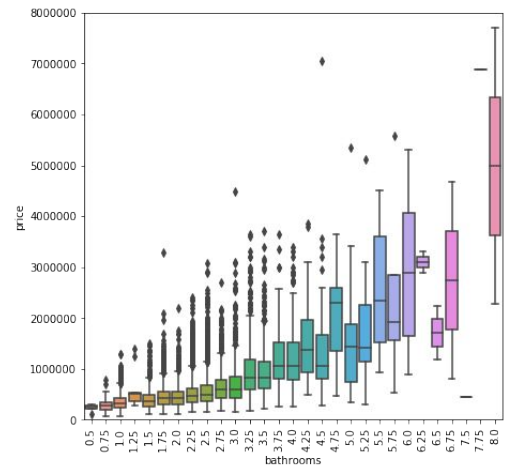


- The most meaningful plot here would be price vs. **sqft_living**: there's an obvious linear relationship with increasing square footage of living space, the price of the home increases
- The data from **sqft_lot** don't correlate at all, and this plot shows that this data might not be reported for each home
- Some homes do not have basements, and therefore the **sqft_basement** plot also can be misleading

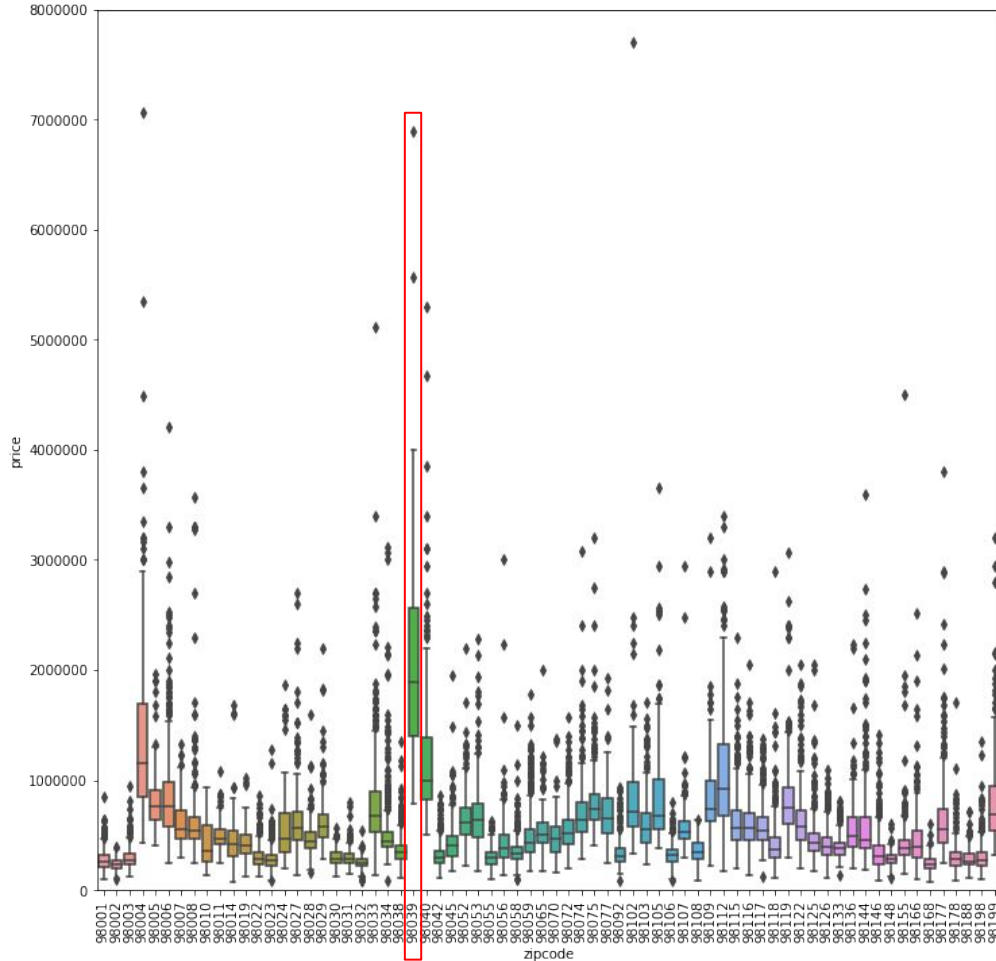
Summary: Analysis of Categorical Variables



- With increasing grade, price increases
- As the number of floors increase, up until 2.5, the price of the house also increases
- In general, the more bedrooms in the house, the higher the price; although this seems to plateau after 8 bedrooms
- With an increased number of bathrooms, the price, as expected, also increases



Zip Codes vs. Housing Prices in King County




The boxplot analysis for zip codes vs. home price in King County shows that one zipcode in particular seems to have the highest housing prices in this analysis: 98039¹

Upon further research, this zip code correlates with a Seattle suburb, Medina, which is home to billionaires like Bill Gates and Jeff Bezos²!

Other zip codes also seem to be home to expensive houses in King County.

- 1) https://en.wikipedia.org/wiki/Medina,_Washington
- 2) <https://www.businessinsider.com/where-billionaires-bill-gates-jeff-bezos-live-photos-of-medina-washington-2017-12>

Conclusions/Predictions: Regression Modeling (Ordinary Least Squares)



The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values - meaning that variables with values closer to 0.5 (ideally, closer to 1.0) are more reliable metrics for prediction than other variables.

0	ind_var	r_squared
1	price	1
2	bedrooms	0.123176
3	bathrooms	0.301547
4	sqft_living	0.444836
6	floors	0.0973044
7	waterfront	0.0269729
8	view	0.114352
9	condition	0.00149605
10	grade	0.489964
11	sqft_above	0.355566
12	sqft_basement	0.0964142
13	yr_built	0.00683272
14	yr_renovated	0.0110459