Data Housing Project

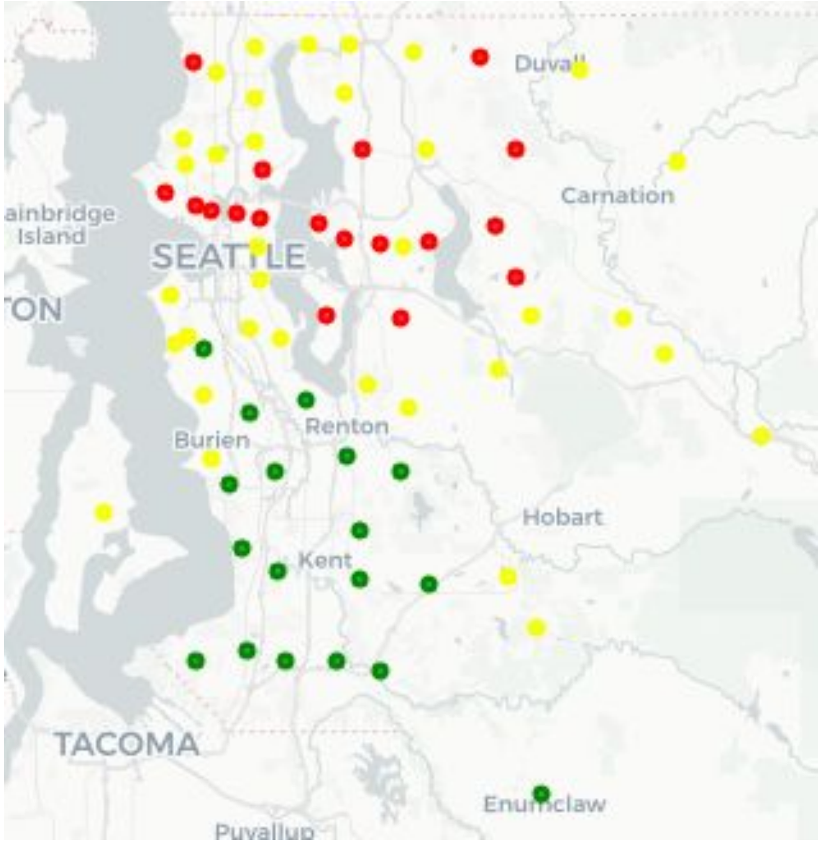
Mod 1 Project

1. Data Cleaning

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21597 entries, 0 to 21596
Data columns (total 21 columns):
id               21597 non-null int64
date             21597 non-null object
price            21597 non-null float64
bedrooms         21597 non-null int64
bathrooms        21597 non-null float64
sqft_living      21597 non-null int64
sqft_lot         21597 non-null int64
floors           21597 non-null float64
waterfront       19221 non-null float64
view             21534 non-null float64
condition        21597 non-null int64
grade            21597 non-null int64
sqft_above       21597 non-null int64
sqft_basement    21597 non-null object
yr_built         21597 non-null int64
yr_renovated     17755 non-null float64
zipcode          21597 non-null int64
lat              21597 non-null float64
long             21597 non-null float64
sqft_living15    21597 non-null int64
sqft_lot15       21597 non-null int64
dtypes: float64(8), int64(11), object(2)
memory usage: 3.5+ MB
```

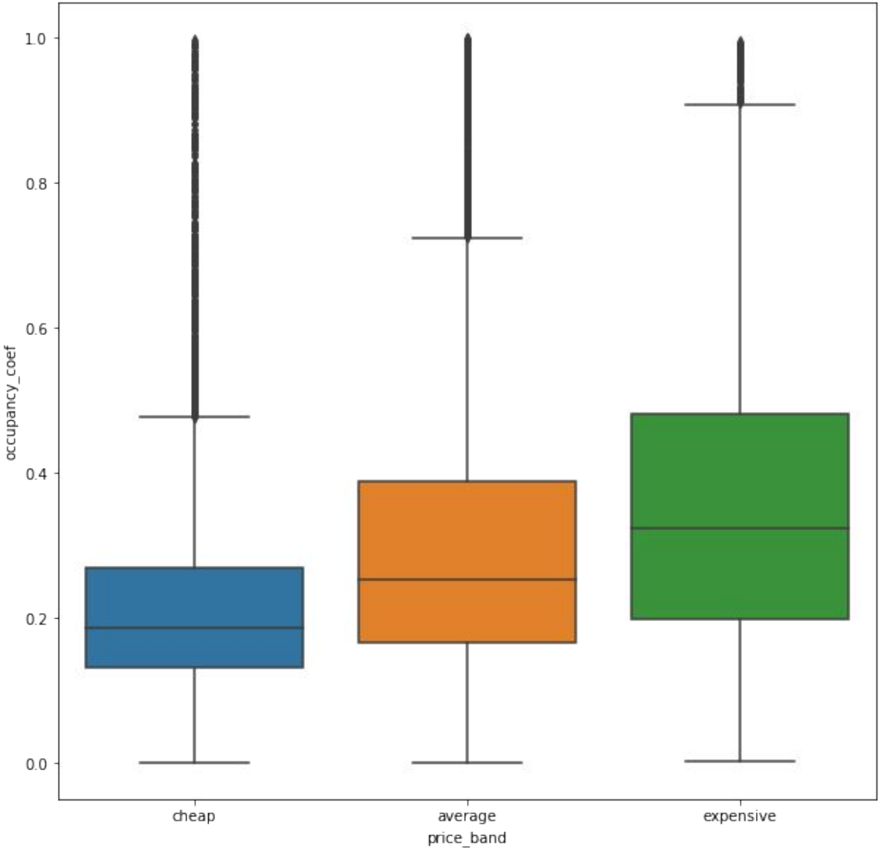
a. Data import
b. Checking data types
c. Resolving missing values
d. Removing outliers

2. Exploratory Data Analysis



a. How does location have an impact on price?

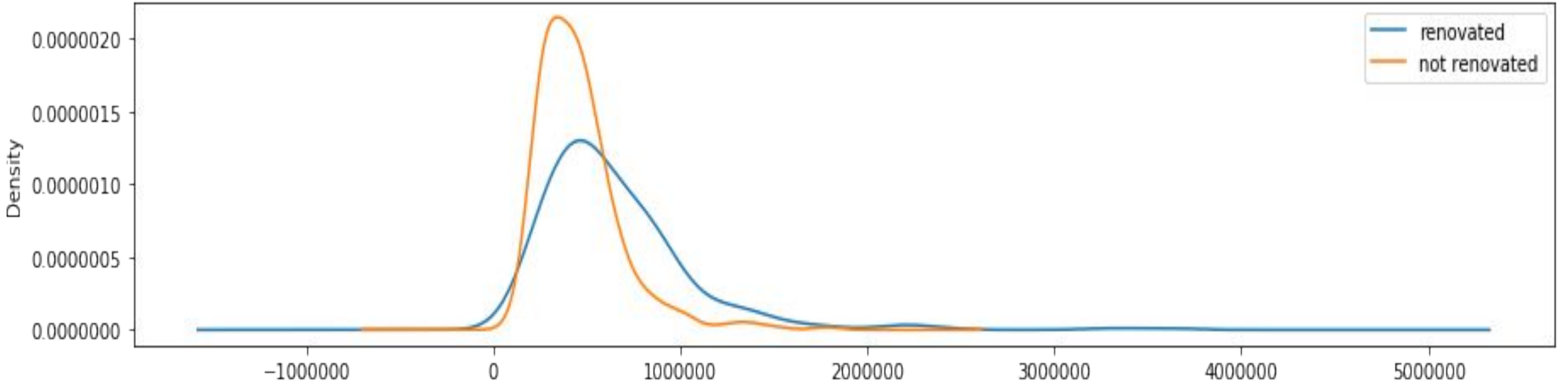
2. Exploratory Data Analysis

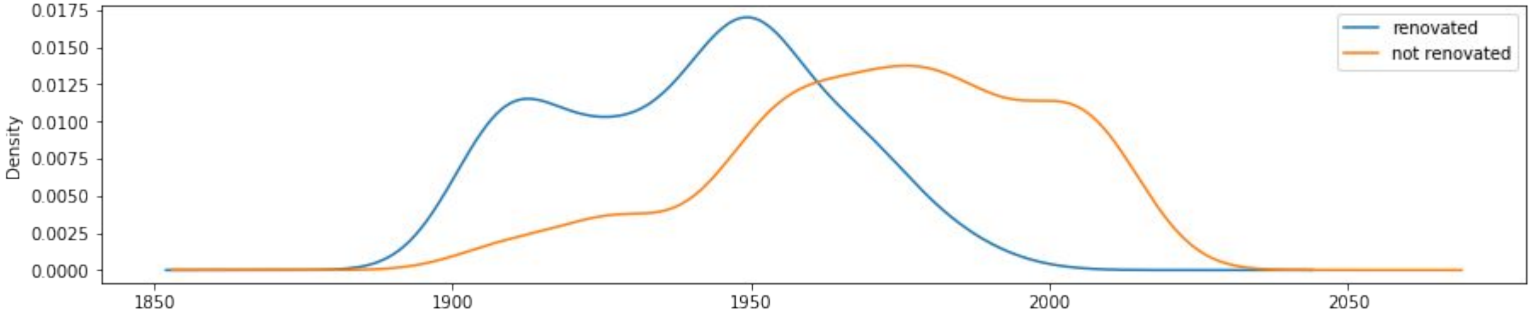


b. How does living density have an impact on price?

2. Exploratory Data Analysis

c. How does renovation have an impact on price?





3. Modeling

	correlation
price	1.000000
sqft_living	0.701554
grade	0.668262
sqft_above	0.605510
sqft_living15	0.585597
bathrooms	0.524823
view	0.395640
sqft_basement	0.319199
bedrooms	0.315193
lat	0.308032

a. Adding predictors to the model

3. Modeling

OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.862
Model:	OLS	**Adj. R-squared (uncentered):**	0.862
Method:	Least Squares	**F-statistic:**	2.248e+04
Date:	Tue, 22 Oct 2019	**Prob (F-statistic):**	0.00
Time:	13:46:08	**Log-Likelihood:**	-2.9744e+05
No. Observations:	21529	**AIC:**	5.949e+05
Df Residuals:	21523	**BIC:**	5.949e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
sqft_living	285.7753	2.338	122.205	0.000	281.192	290.359
view	7.198e+04	2433.015	29.586	0.000	6.72e+04	7.68e+04
bedrooms	-5.067e+04	2281.429	-22.208	0.000	-5.51e+04	-4.62e+04
lat	1935.5536	133.999	14.445	0.000	1672.906	2198.201
waterfront	5.533e+05	2.19e+04	25.304	0.000	5.1e+05	5.96e+05
yr_renovated	61.1899	4.552	13.443	0.000	52.268	70.112

Omnibus:	12922.614	**Durbin-Watson:**	1.981
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	418397.294
Skew:	2.347	**Prob(JB):**	0.00
Kurtosis:	24.081	**Cond. No.**	3.02e+04

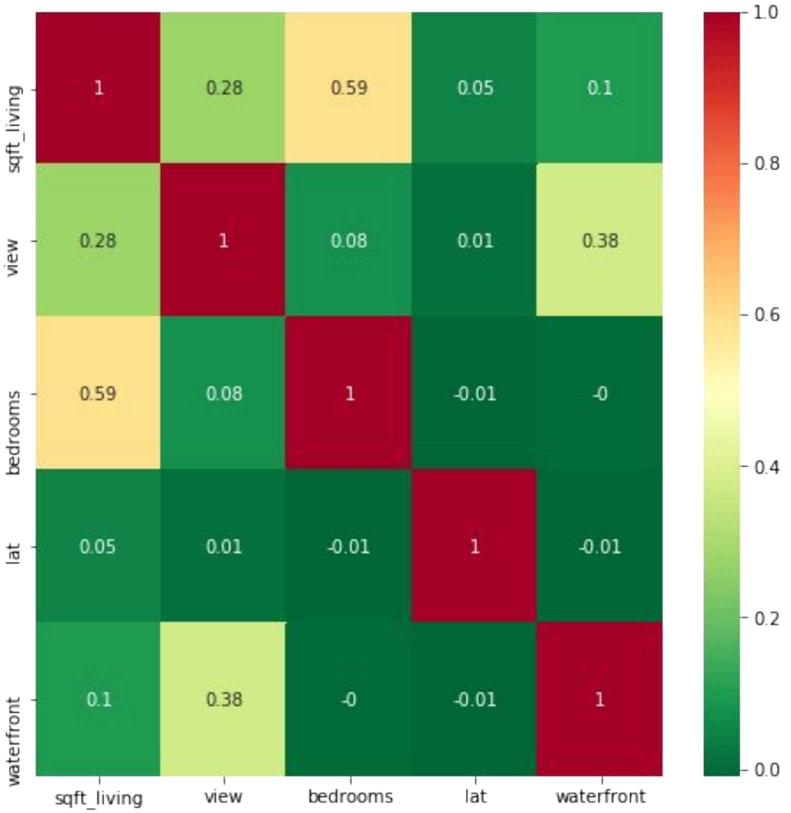
a. Evaluating the coefficients

3. Modeling

	correlation
sqft_living	1.000000
sqft_above	0.945416
sqft_living_per_bed	0.786544
grade	0.782506
bathrooms	0.777182
sqft_living15	0.772939

c. Checking for collinearity between predictors

4. Possible extensions



a. Reducing collinearity using feature engineering

4. Possible extensions

b. Using feature scaling to scale model coefficients

c. Splitting the data into training and test sets

d. Checking for normality of predictors