Cracking the Code on Homes

Nick Catalano, Alaska Lam, Zach Paul

Problem Statement:

We are aiming to better understand the housing market in the Kings County area, specifically for buying/selling houses.

In this presentation, we will investigate how several factors significantly impact sale prices.

Business Value:

By creating a model that can accurately predict sale prices for houses:

Sellers will be able to more competitively price their property armed with this knowledge

Realty companies or individuals who flip houses will be better able to increase their ROI

Methodology:

Analyze past house-sale data to make recommendations to realtors, sellers and/or individuals who flip houses, on how to sell houses at the highest price point possible for their situation

Some of the topics we will explore:

Waterfront properties

Distance to a coast of water

Zipcodes

Model 1 - R accuracy value of .933

Model 1: Prices less than $1M

In [43]:
```
model_1 = linear_regression_model(df[df['price'] < 1000000],
                                  ['price', 'age', 'sqft_living', 'waterfront', 'd_coast','renovated','zipcode_grade'])
```

OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.933
Model:	OLS	**Adj. R-squared (uncentered):**	0.933
Method:	Least Squares	**F-statistic:**	4.682e+04
Date:	Wed, 30 Sep 2020	**Prob (F-statistic):**	0.00
Time:	18:20:29	**Log-Likelihood:**	-2.6543e+05
No. Observations:	20107	**AIC:**	5.309e+05
Df Residuals:	20101	**BIC:**	5.309e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	1645.6295	30.737	53.539	0.000	1585.383	1705.876
sqft_living	157.8281	1.197	131.860	0.000	155.482	160.174
renovated	1.199e+05	5459.240	21.958	0.000	1.09e+05	1.31e+05
d_coast_bins	-6.648e+04	3309.393	-20.087	0.000	-7.3e+04	-6e+04
grade_bins_1	1.101e+05	2957.332	37.223	0.000	1.04e+05	1.16e+05
grade_bins_2	2.766e+05	3473.197	79.648	0.000	2.7e+05	2.83e+05

Omnibus:	677.086	**Durbin-Watson:**	1.916
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	787.970
Skew:	0.421	**Prob(JB):**	7.84e-172
Kurtosis:	3.480	**Cond. No.**	1.24e+04

Model 2 - R accuracy value of .924

Model 2: Prices greater than or equal to $1M

```
In [35]: model_2 = linear_regression_model(df[df['price'] >= 1000000],
                                            ['price', 'age', 'sqft_living', 'waterfront', 'd_coast','renovated','zipcode_grade'])
```

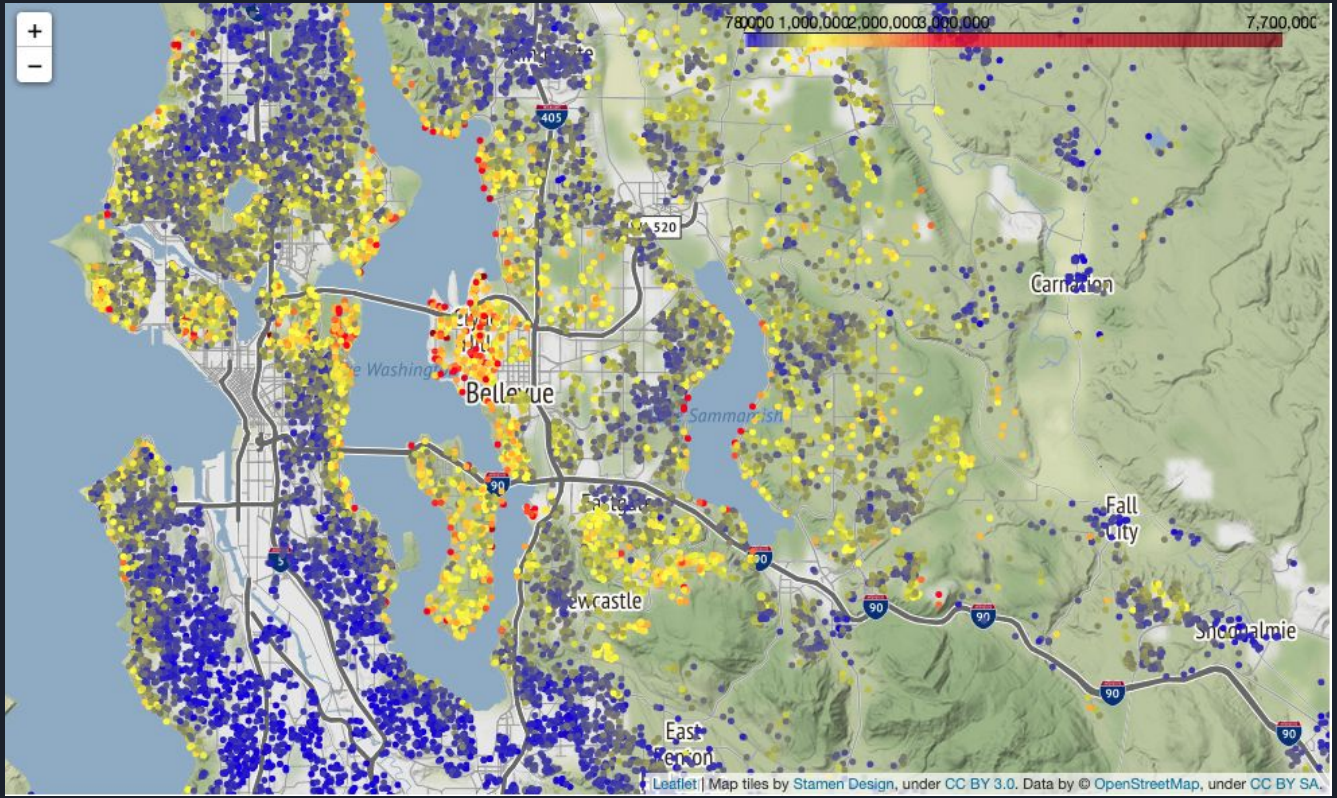
OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.924
Model:	OLS	**Adj. R-squared (uncentered):**	0.923
Method:	Least Squares	**F-statistic:**	2560.
Date:	Wed, 30 Sep 2020	**Prob (F-statistic):**	0.00
Time:	17:58:57	**Log-Likelihood:**	-21531.
No. Observations:	1490	**AIC:**	4.308e+04
Df Residuals:	1483	**BIC:**	4.311e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
age	2833.8484	398.061	7.119	0.000	2053.027	3614.670
sqft_living	316.8489	9.922	31.935	0.000	297.387	336.311
waterfront	6.305e+05	5.11e+04	12.328	0.000	5.3e+05	7.31e+05
renovated	2.257e+05	4.1e+04	5.507	0.000	1.45e+05	3.06e+05
d_coast_bins	-2.27e+05	2.77e+04	-8.202	0.000	-2.81e+05	-1.73e+05
grade_bins_1	1.786e+05	5.29e+04	3.378	0.001	7.49e+04	2.82e+05
grade_bins_2	3.702e+05	5.13e+04	7.223	0.000	2.7e+05	4.71e+05

Omnibus:	527.675	**Durbin-Watson:**	1.992
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3717.312
Skew:	1.470	**Prob(JB):**	0.00
Kurtosis:	10.158	**Cond. No.**	2.44e+04

Recommendation 1:



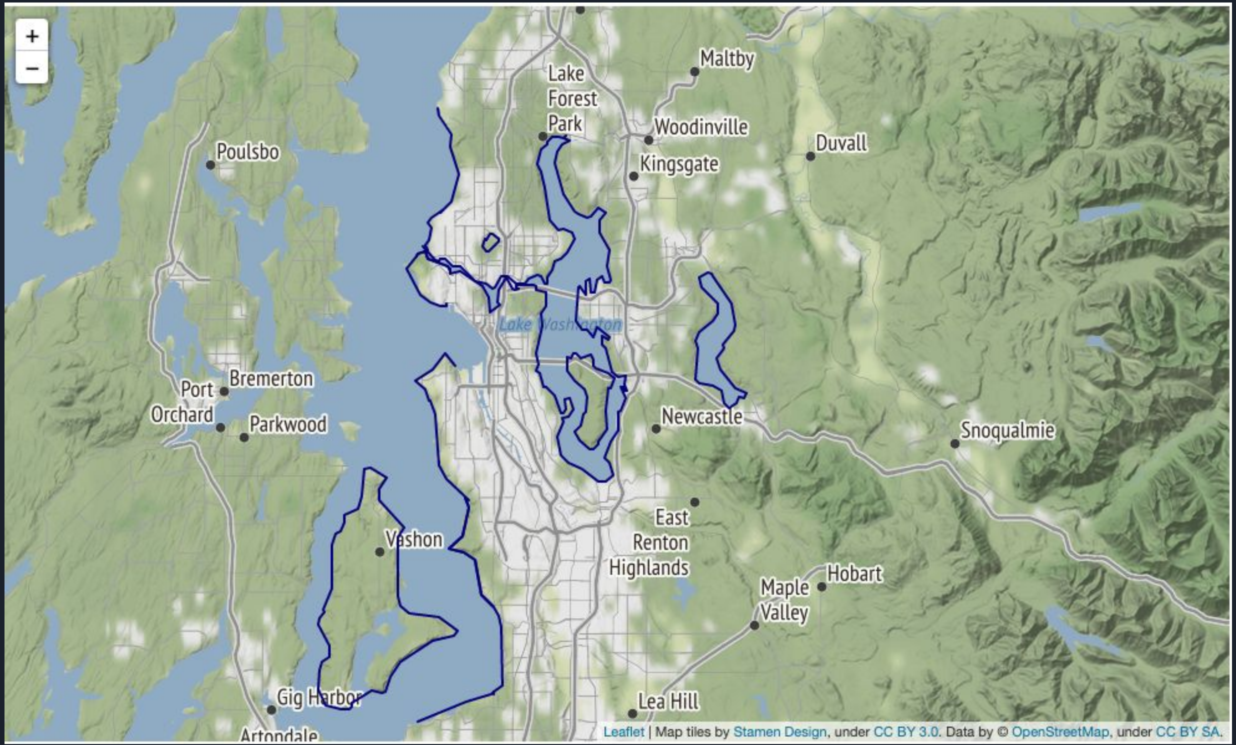
<u>“Waterfront Property”</u>

Waterfront classification increases price

What is “waterfront”?

According to dataset, “has a view to a waterfront”

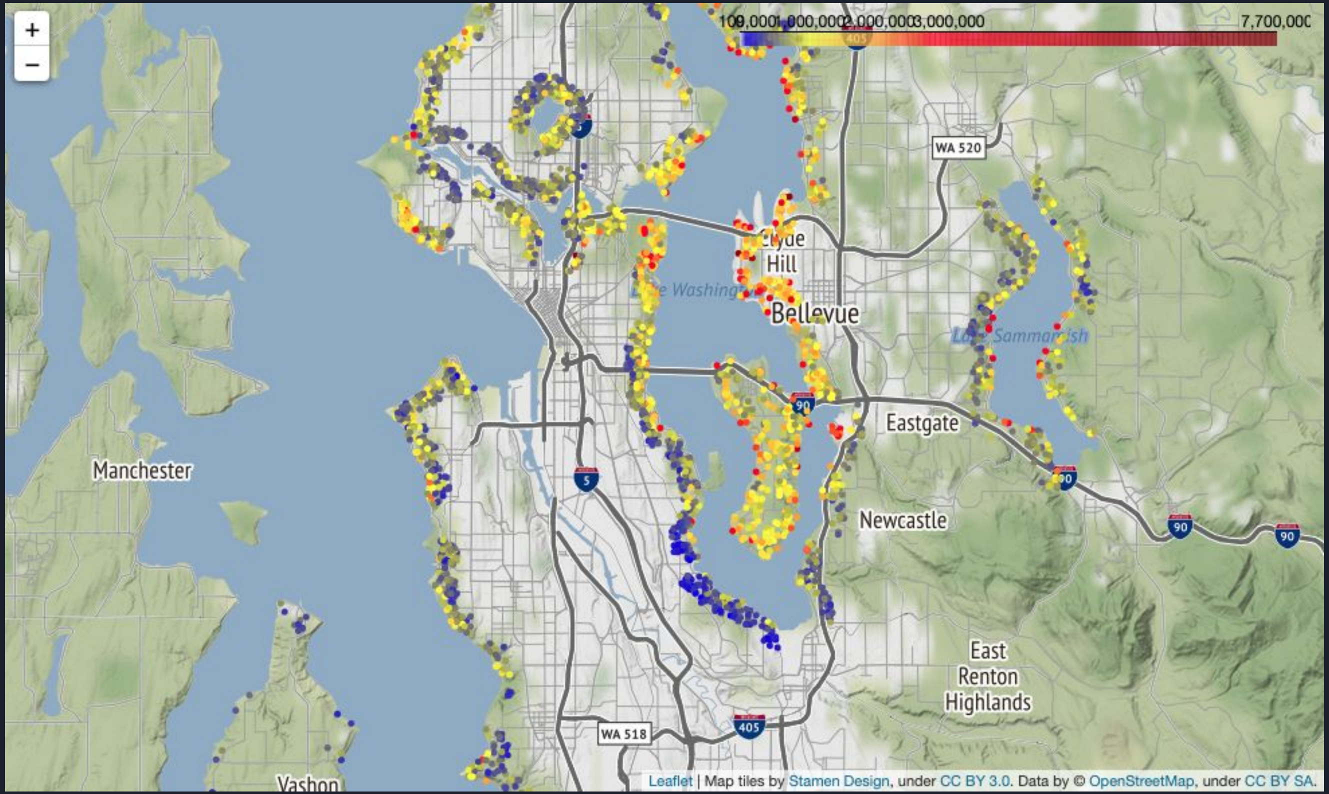
What about houses without direct view, but still near a coast of water?



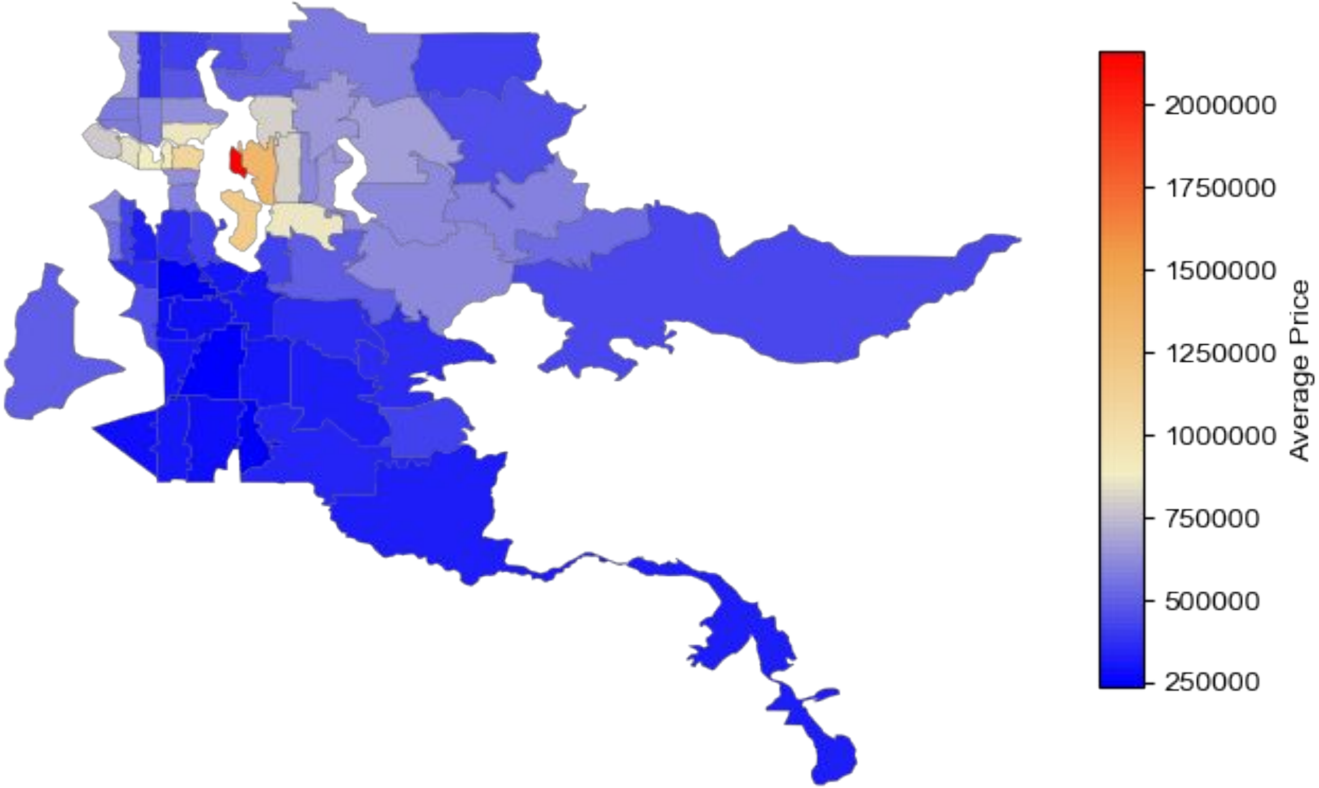
Recommendation 2:

<u>Distance from Coastline</u>

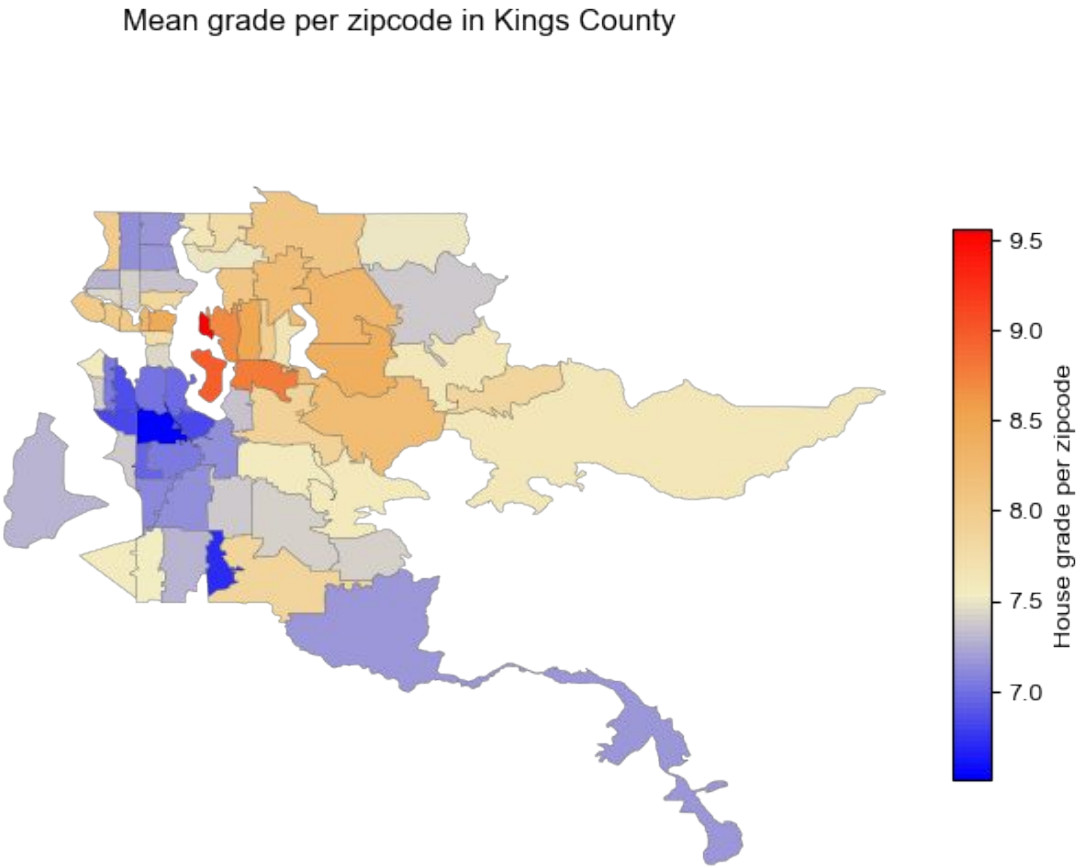
Houses less than .25 miles from shore have mostly higher prices than other houses



Average Sale Price per ZIP Code in Kings County



Recommendation 3 :



Grade of house,
given by King County grading system

Regions with highest grades correlate with regions with highest sale prices

Future Considerations

Analyzing luxury homes [$1 million+ sale price] separately from non-luxury

Creating separate prediction models for different neighborhoods or regions, and/or change coastal points

Analyzing burglary patterns vs sale price

What about impact of specific construction or realty company?

Thank you!