King County Housing Consultants (Market overview & price prediction)

Ravi Dahiya & Marc Inizan

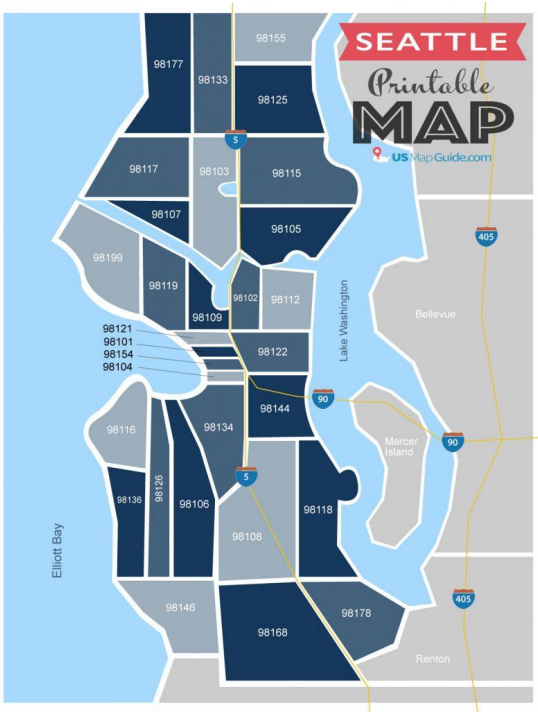
Objectives

* Analyze the geographical data to provide in-depth information about the general market scenario.

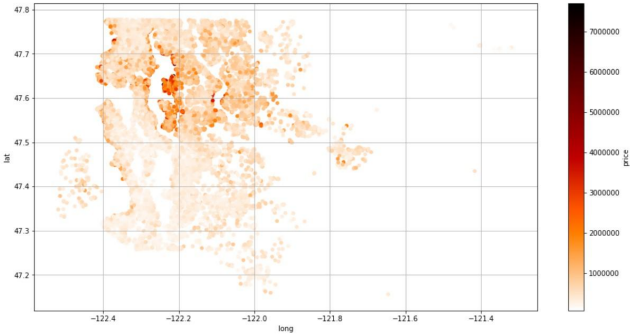
* Find out the features that have the highest impact on the price of the house.

* Predict the price of a house based on its most important features.

The market

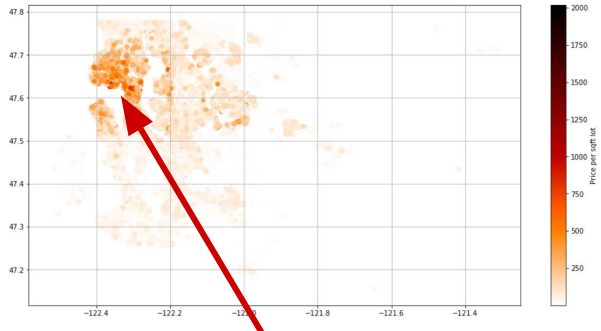


Price as per location



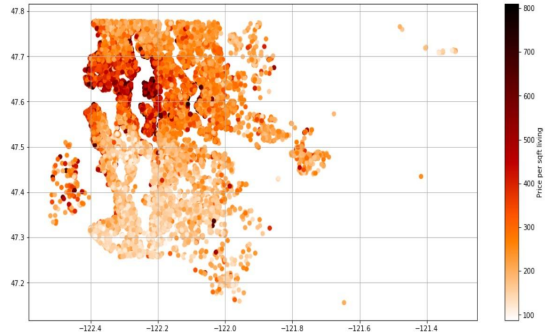
As we can see that the prices are highest around (-47.7, -122.2)

Price as per sqft lot as per location



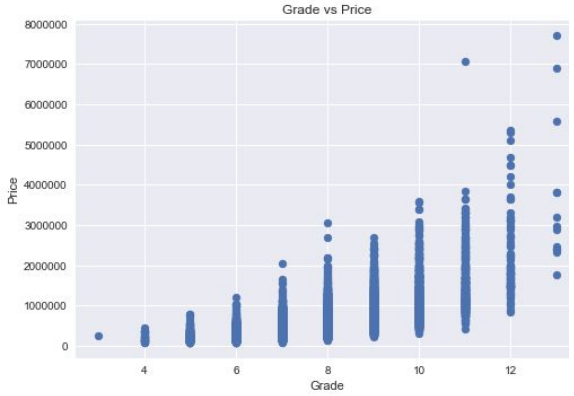
Areas in the darker shades will attract higher prices per sqft of land

Price per sqft living as per location



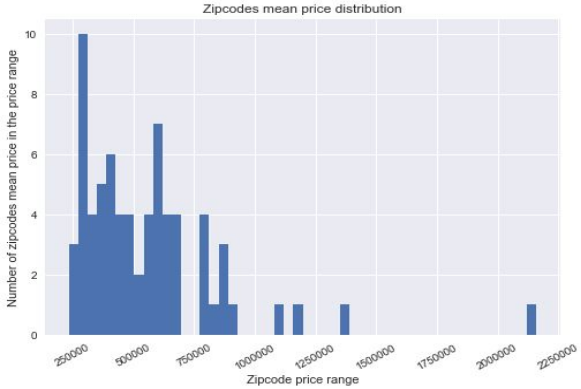
As we can see that even though the houses with the highest prices are scattered over a large area, the per sqft lot price is highest in the North West of Seattle's King County

Price indicators



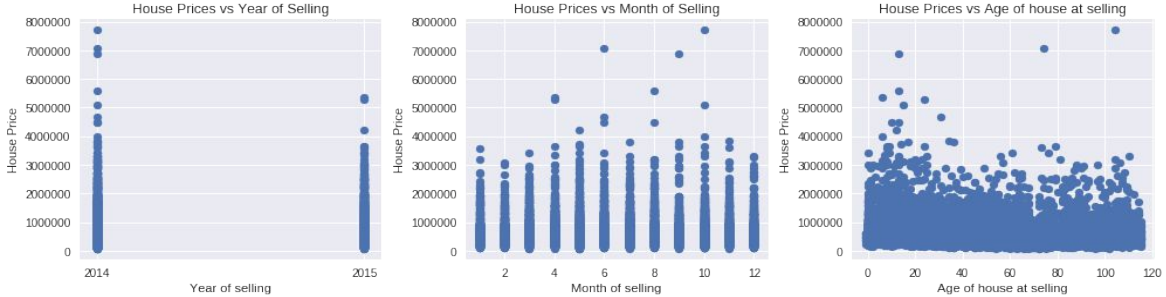
Higher the grade - higher the price of the house

Higher the living area - higher the price



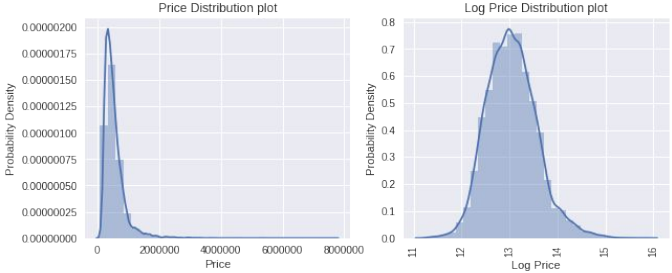
Uneven mean prices by zip codes

Data Cleaning & Data Transformation

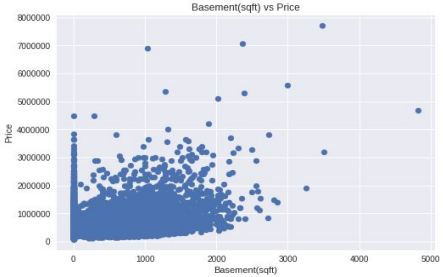


Merging the Columns

Binning



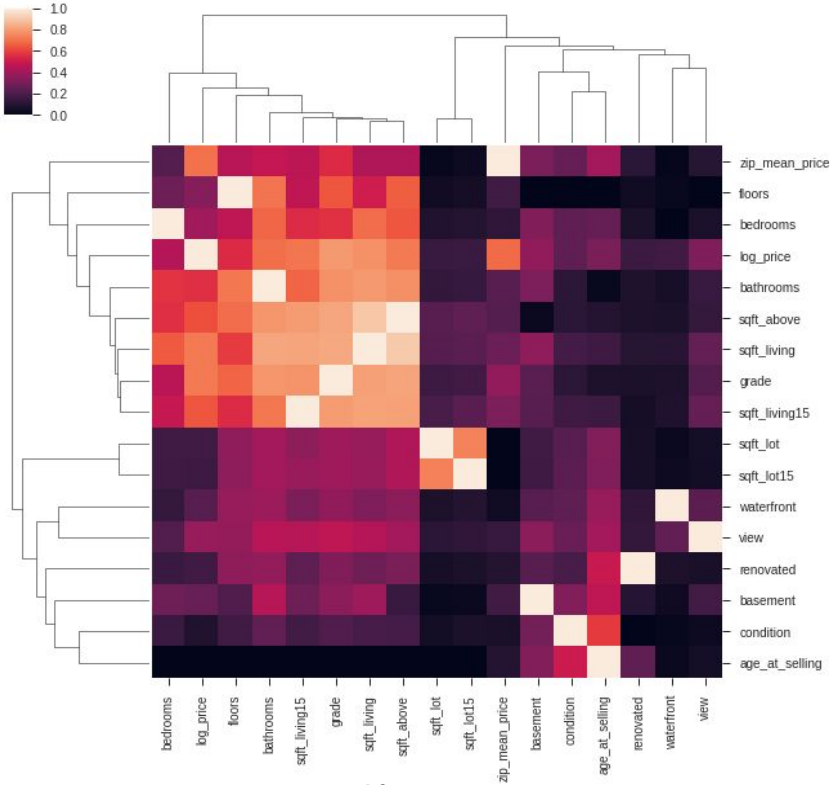
Log Transformation



Yes/No

Zipcode Bins

Feature Selection



Cluster Map

Top 5 features correlated with Log Price:
- -Grade
- -Square footage of the living area
- -Zipcode
- -Square footage of the living area of the 15 closest houses
- -Square footage of house apart from basement

Those last 2 features are highly correlated with the square footage of the living area.

Prediction Model

Dep. Variable:	log_price	**R-squared:**	0.776
Model:	OLS	**Adj. R-squared:**	0.776
Method:	Least Squares	**F-statistic:**	2.491e+04
Date:	Tue, 03 Dec 2019	**Prob (F-statistic):**	0.00
Time:	12:55:04	**Log-Likelihood:**	-640.13
No. Observations:	21594	**AIC:**	1288.
Df Residuals:	21590	**BIC:**	1320.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.8141	0.005	2415.595	0.000	11.804	11.824
grade	0.1061	0.002	44.570	0.000	0.101	0.111
sqft_living	0.0002	2.79e-06	83.541	0.000	0.000	0.000
zip_mean_price	0.3027	0.002	145.806	0.000	0.299	0.307

Omnibus:	707.520	**Durbin-Watson:**	1.975
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1588.247
Skew:	0.190	**Prob(JB):**	0.00
Kurtosis:	4.273	**Cond. No.**	6.69e+03

With 3 featured: R-squared value of 0.776, which means that 77.6 percent variance in the dependent variable can be explained by independent variables.

Model in the working

```
1 predict(1200, 98117, 7)
```

```
404865.0
```

A 1200 square feet house located in the 98117 zipcode with a grade of 7 would be worth 405K$

Recommendations

1. Should include features such as custom design, high quality cabinet work, wood trim ,marble, bigger entries

2. As the size of lot has small impact on the price, focus should be on high living area. Also, basement doesn't add much value to the house.

3. Houses in zipcodes with high average prices will be priced higher

Questions & Answers