

The background features a complex network of thin grey lines connecting various points, creating a web-like structure. Scattered throughout are numerous triangles of different sizes and orientations, some with solid black dots at their vertices. The overall aesthetic is modern and geometric.

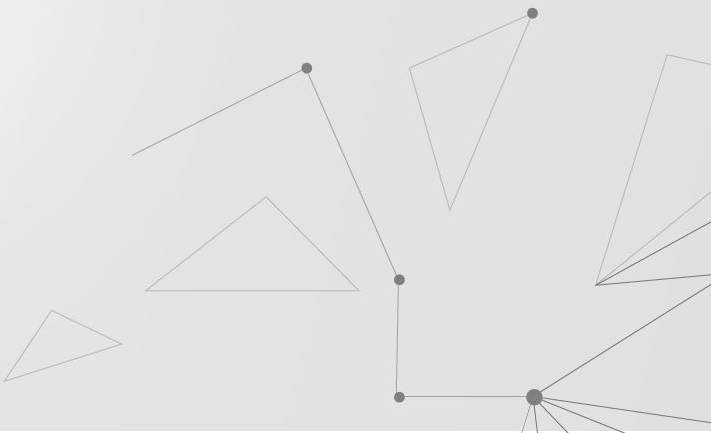
Kings County House Sales

Bonny Nichol
February 2020



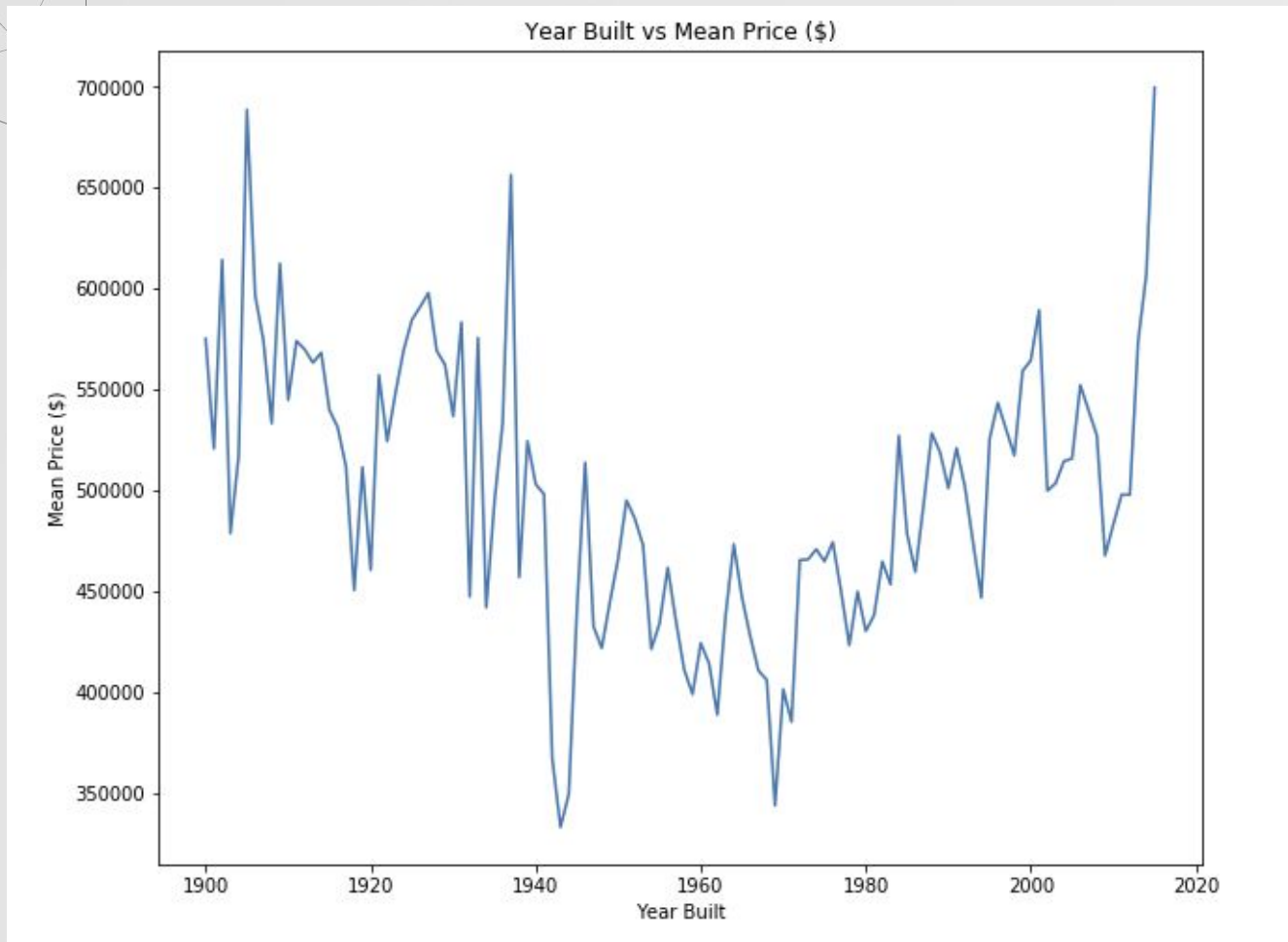
Business Proposal

Predicting the sales prices of houses in Kings
County, WA with pre-existing Data

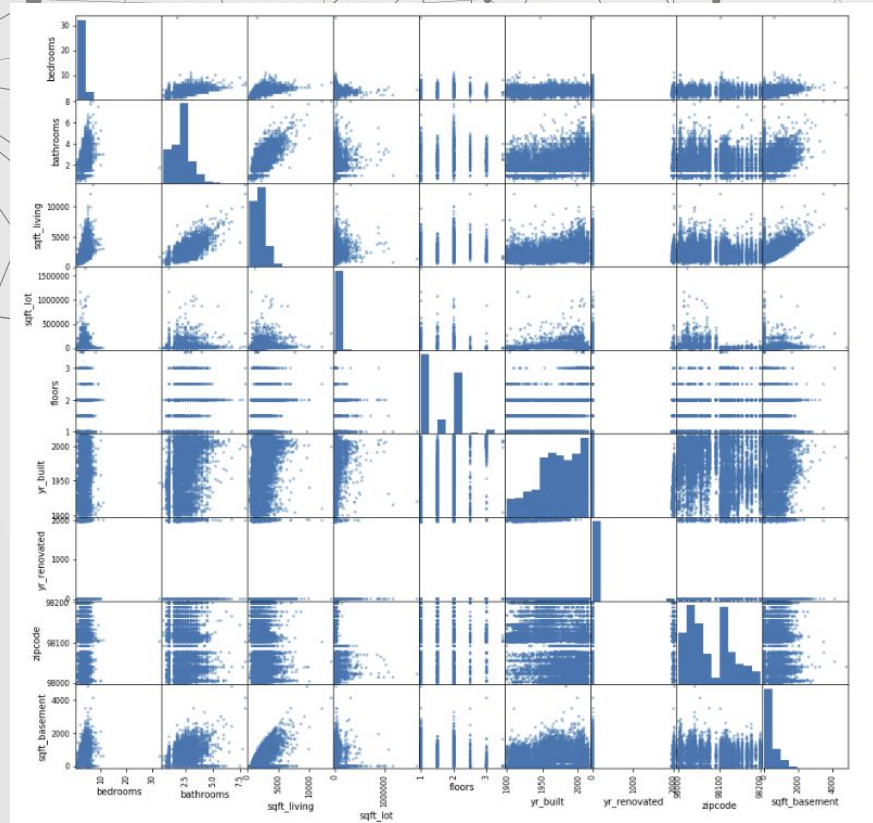


Given Features

id id number for the data entry
date date the house was sold
price price of the house
bedrooms number of bedrooms in the house
bathrooms number of bathrooms in the house
sqft_living square footage of the living room
sqft_lot square footage of the lot size
floors number of floors in the house
waterfront if the house is located at the waterfront (1 = yes, 0 = no)
view does the house have a view
condition condition of the house
grade grade level of the house
sqft_above square footage of the house not including the basement
sqft_basement square footage of only the basement
yr_built year the house was built
yr_renovated year the house was renovated
zipcode zipcode where the house is located
lat latitude where the house is located
long longitude where the house is located
sqft_living15 square footage of the living area in 2015 (this implies there were renovations)
sqft_lot15 square footage of the lot size in 2015

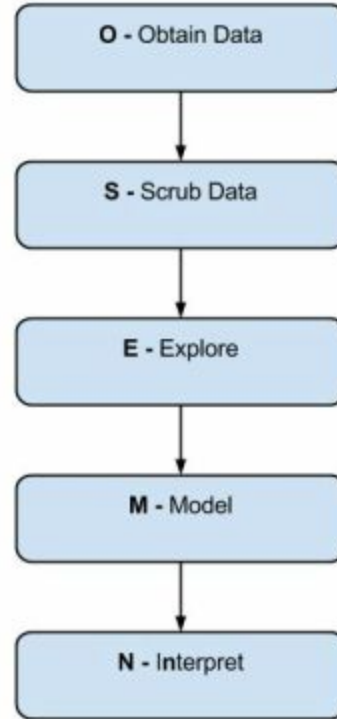


Methodology



OSEMN: Data Science Method

The OSEMN Workflow



No
significant
linear
correlation

Dep. Variable:	price	R-squared:	0.492
Model:	OLS	Adj. R-squared:	0.491
Method:	Least Squares	F-statistic:	740.2
Date:	Fri, 28 Feb 2020	Prob (F-statistic):	0.00
Time:	14:19:55	Log-Likelihood:	-5063.4
No. Observations:	14559	AIC:	1.017e+04
Df Residuals:	14539	BIC:	1.032e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	18.6592	0.258	72.202	0.000	18.153	19.166
sqft_living	0.0002	1.53e-05	13.517	0.000	0.000	0.000
sqft_lot	-8.731e-06	5.07e-07	-17.209	0.000	-9.73e-06	-7.74e-06
sqft_above	0.0003	1.6e-05	17.473	0.000	0.000	0.000
yr_built	-0.0035	0.000	-28.037	0.000	-0.004	-0.003
basement_1	0.1826	0.011	16.148	0.000	0.160	0.205
view_1	0.1847	0.024	7.742	0.000	0.138	0.231
view_2	0.1652	0.015	11.045	0.000	0.136	0.194
view_3	0.2370	0.022	10.890	0.000	0.194	0.280
view_4	0.4909	0.030	16.562	0.000	0.433	0.549
cond_2	0.0009	0.083	0.010	0.992	-0.163	0.164
cond_3	0.2905	0.077	3.766	0.000	0.139	0.442
cond_4	0.3121	0.077	4.045	0.000	0.161	0.463
cond_5	0.3759	0.078	4.841	0.000	0.224	0.528
bed_2	0.0635	0.030	2.093	0.036	0.004	0.123
bed_3	-0.0254	0.030	-0.842	0.400	-0.084	0.034
bed_4	-0.0894	0.031	-2.899	0.004	-0.150	-0.029
bath_1	0.1147	0.048	2.394	0.017	0.021	0.209
bath_2	0.2028	0.048	4.185	0.000	0.108	0.298
bath_3	0.2871	0.050	5.741	0.000	0.189	0.385

Omnibus:	28.875	Durbin-Watson:	2.026
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.005
Skew:	-0.107	Prob(JB):	5.03e-07
Kurtosis:	2.953	Cond. No.	9.88e+05



Conclusions



- Analysis of Kings County Housing Sales dataset.
- OSEMiN Data Science Framework
- The project used a simple linear regression model, OLS regression model and Stepwise Selection to choose potential features that could contribute to the sales price.
- After using these selected features in another linear regression model we have seen that the model accuracy actually worsened. (From 46.2% accuracy to 35.8%). This shows the first model was not a perfect fit and therefore can be improved upon to improve feature selection.
- Future explorations of this project will be focusing on categorical data and its possible impact on the models as well as exploring other models and improving accuracy.