King County Housing Market: Regression Analysis

Adina Steinman and George Paskalev

Business Problem

- Which areas of King County should a housing development company build properties?

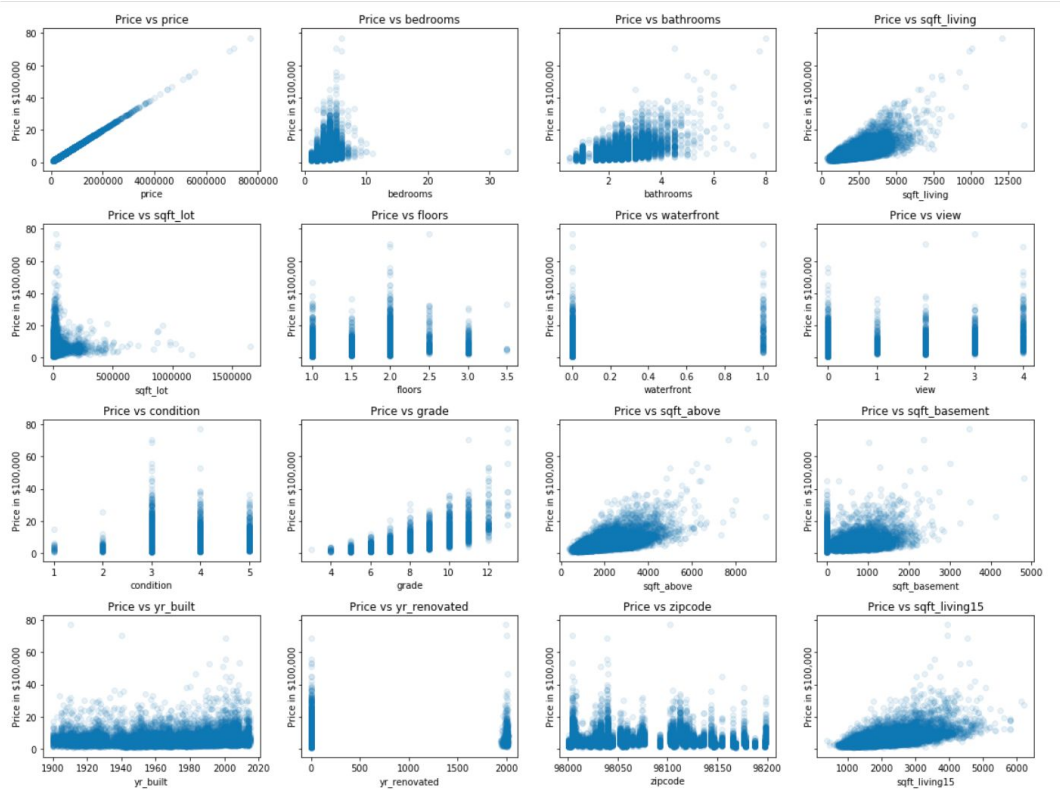
- What features should these properties have in order for them to sell at high prices and yield the company high profits?

Data Used

- Used housing data from King County, Washington
  - Initial dataset included 21 features, with 21,597 rows of data
- Dataset included various features, most relevant to our analysis were:
  - House Price
  - Bedrooms
  - Square footage of interior living space
  - Floors
  - House Condition
  - Zip Code

EDA: Feature Analysis

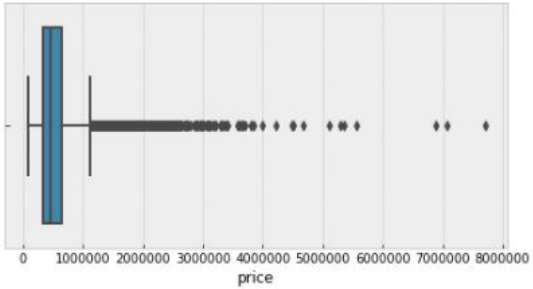
- Scatter plots helped identify categorical vs. continuous variables
- Initial look at whether any variables follow a normal distribution

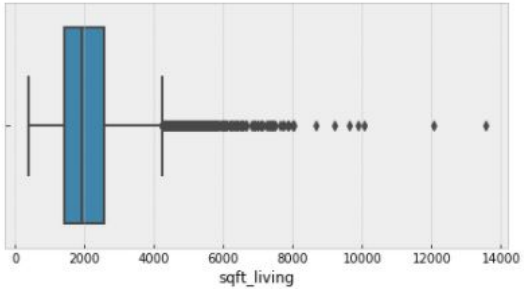


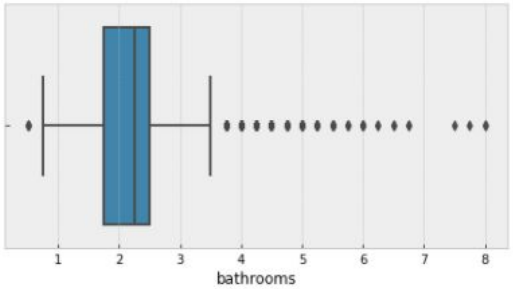
Variable Definitions

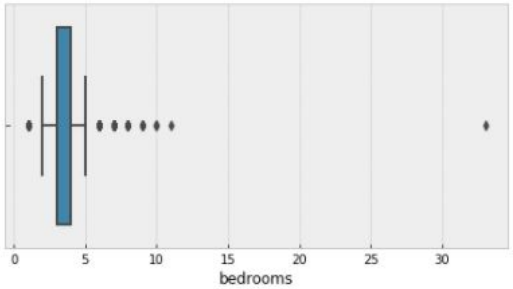
- **Sqft_living:** square footage of interior living space
- **Sqft_lot:** square footage of land space
- **Waterfront:** whether or not unit overlooks waterfront
- **View:** Index from 0-4 of how good home view is
- **Grade:** Index from 1-13 which measures the quality of construction and design
- **Condition:** Index 1-5 on condition of the unit
- **Sqft_above:** square footage of living space above ground level
- **Sqft_basement:** square footage of living space below ground level
- **Sqft_living15:** square footage of interior living space for the nearest 15 neighbors

EDA: Outliers









- Boxplots helped identify outliers for price, bathrooms, sqft_living and bedrooms
- Cleaned our data further by removing homes with values of price greater than $700,000, bedrooms greater than 13, and square foot living greater than 11,000.

Baseline Model

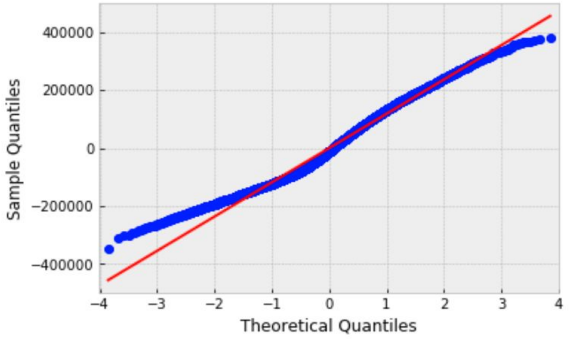
OLS Regression Results

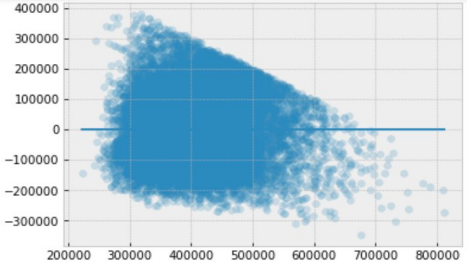
Dep. Variable:	price	R-squared:	0.269
Model:	OLS	Adj. R-squared:	0.268
Method:	Least Squares	F-statistic:	569.0
Date:	Sat, 28 Nov 2020	Prob (F-statistic):	0.00
Time:	10:00:48	Log-Likelihood:	-2.2299e+05
No. Observations:	17024	AIC:	4.460e+05
Df Residuals:	17012	BIC:	4.461e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.231e+07	1.73e+06	-18.634	0.000	-3.57e+07	-2.89e+07
bedrooms	-1.722e+04	1365.837	-12.609	0.000	-1.99e+04	-1.45e+04
bathrooms	1.577e+04	2096.954	7.520	0.000	1.17e+04	1.99e+04
sqft_living	101.8386	2.177	46.785	0.000	97.572	106.105
sqft_lot	0.0728	0.028	2.577	0.010	0.017	0.128
floors	2.626e+04	2040.762	12.868	0.000	2.23e+04	3.03e+04
waterfront	2.558e+04	2.42e+04	1.059	0.290	-2.18e+04	7.29e+04
view	1.92e+04	1867.791	10.280	0.000	1.55e+04	2.29e+04
condition	2.09e+04	1484.173	14.083	0.000	1.8e+04	2.38e+04
yr_renovated	7.0521	3.760	1.876	0.061	-0.318	14.422
zipcode	330.8279	17.670	18.723	0.000	296.193	365.462
renovated	1.616e+04	7362.390	2.195	0.028	1725.824	3.06e+04

Omnibus:	974.436	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	583.272
Skew:	0.316	Prob(JB):	2.21e-127
Kurtosis:	2.351	Cond. No.	1.90e+08

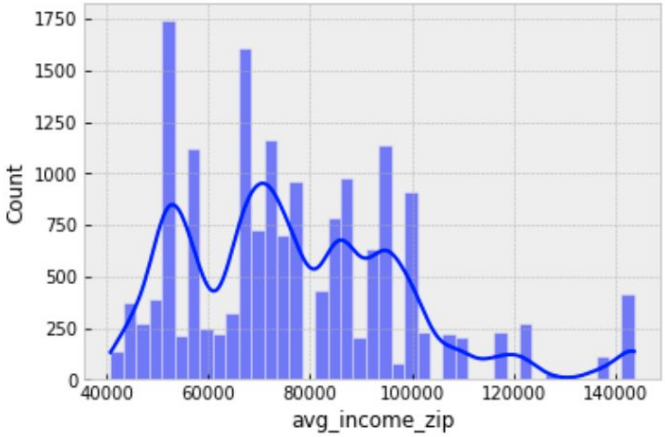
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.9e+08. This might indicate that there are
strong multicollinearity or other numerical problems.

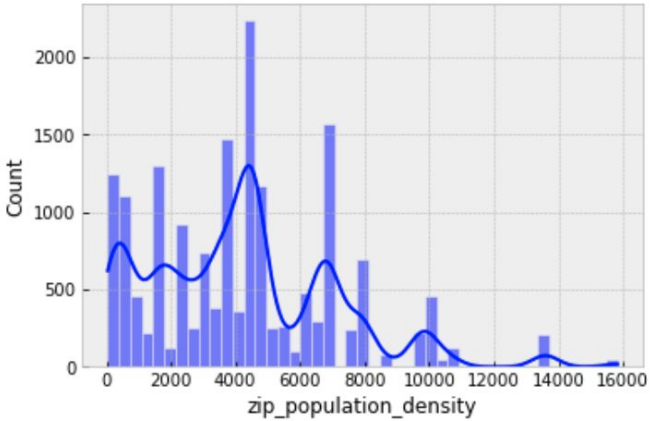




The P-value is: 4.1e-22
The F-statistic is 3.4e-22

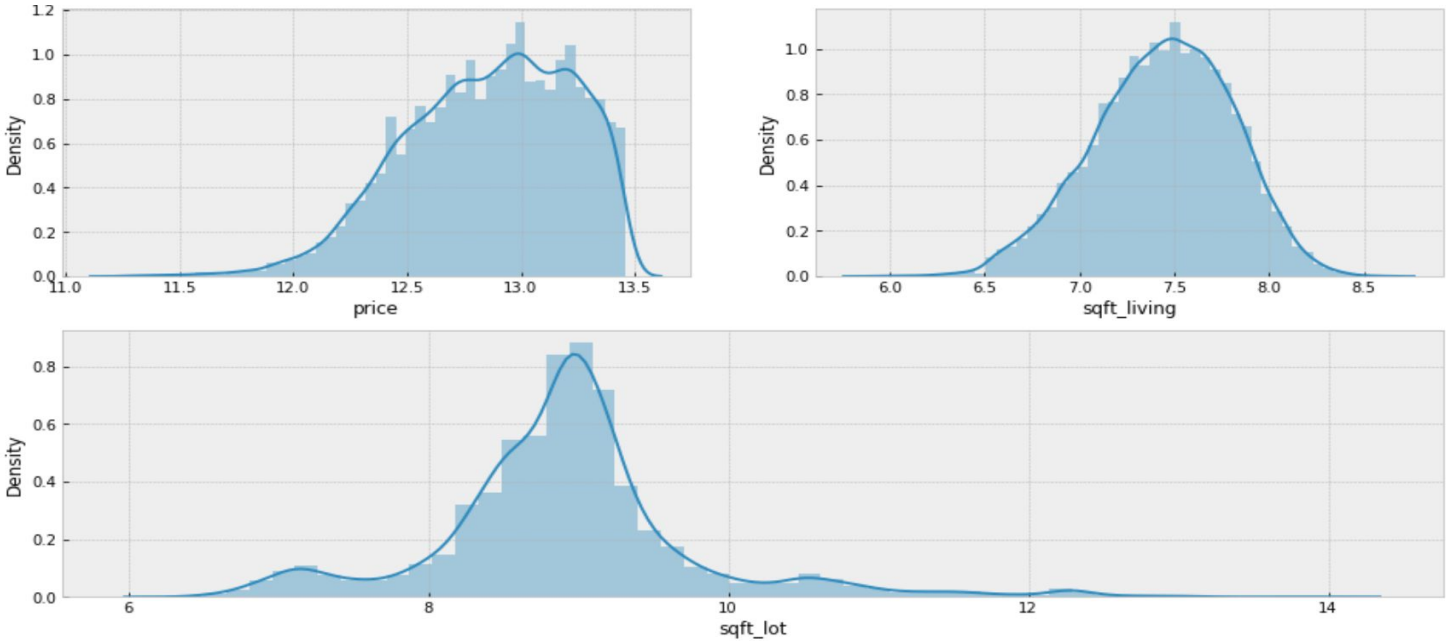
Model 2: Feature Engineering





- Avg_income_zip: The first new feature we are adding is the median household income for each zip code. Household finance experts assert that buyers can afford to pay up to three times their annual incomes for a home.
- Zip_population_density: The next feature is identifying the population density for each zip code. We assume that zip codes with higher population density are more desirable and prices are higher than in zip codes with lower density.

Model 3: Log Transformations

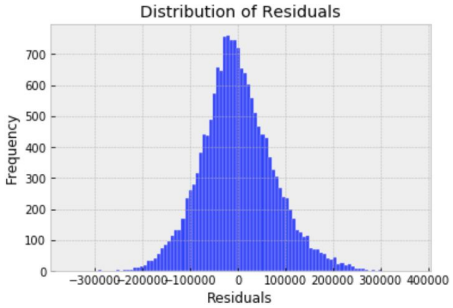


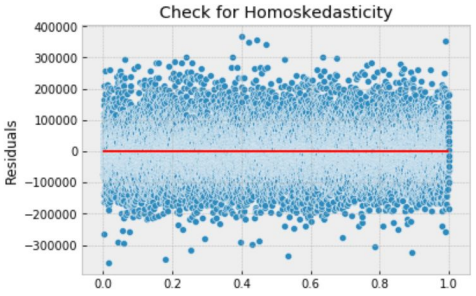
- Log transformations of price, sqft_livilng, and sqft_lot are shown above. Once these variables are logged, their distributions appear significantly more normal.

Model 4: Drop correlated variables, engineer and dummy city variables

```
In [43]: df6.city.value_counts()

Out[43]: Seattle          7189
         Renton           1487
         Kent             1184
         Auburn            896
         Federal Way       762
         Kirkland          707
         Bellevue          615
         Redmond           593
         Maple Valley      569
         Issaquah          552
         Sammamish         415
         Woodinville       333
         Kenmore           270
         Snoqualmie        269
         Enumclaw          228
         North Bend        200
         Duvall            186
         Bothell           182
         Carnation         109
         Vashon            102
         Black Diamond      88
         Fall City          58
         Mercer Island      30
         Name: city, dtype: int64
```





- Engineering the “city” dummy variables and removing features highly correlated with one another and features that are lowly correlated with our target variable, price, as well as removing low p-values, improved the linear assumptions of our model
- Our residuals follow a normal distribution and are homoscedastic

Final Model: Interactions & Final Results

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.699
Model:	OLS	**Adj. R-squared:**	0.699
Method:	Least Squares	**F-statistic:**	1411.
Date:	Sat, 28 Nov 2020	**Prob (F-statistic):**	0.00
Time:	10:02:39	**Log-Likelihood:**	-2.1544e+05
No. Observations:	17024	**AIC:**	4.309e+05
Df Residuals:	16995	**BIC:**	4.312e+05
Df Model:	28		
Covariance Type:	nonrobust		

- Added interaction terms: sqft_living*floors and sqft_living*bathrooms as they are significant in our model
- Final model has an R-squared value of 0.699, which is a large improvement from our baseline model value of 0.269
- All our remaining features have a p-value less than 0.05

Final Model: Coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.559e+05	4.02e+04	-13.830	0.000	-6.35e+05	-4.77e+05
bedrooms	-1.173e+04	900.580	-13.023	0.000	-1.35e+04	-9963.216
bathrooms	-9.924e+04	2.03e+04	-4.884	0.000	-1.39e+05	-5.94e+04
sqft_living	7.924e+04	5593.345	14.166	0.000	6.83e+04	9.02e+04
floors	-5.438e+05	3.08e+04	-17.668	0.000	-6.04e+05	-4.83e+05
view	2.426e+04	1166.934	20.788	0.000	2.2e+04	2.65e+04
condition	1.693e+04	961.966	17.604	0.000	1.5e+04	1.88e+04
avg_income_zip	3.0266	0.048	62.759	0.000	2.932	3.121
zip_population_density	24.5037	0.294	83.462	0.000	23.928	25.079
city_Auburn	-6.667e+04	2940.767	-22.672	0.000	-7.24e+04	-6.09e+04
city_Bellevue	4.732e+04	3516.923	13.454	0.000	4.04e+04	5.42e+04
city_Bothell	3.682e+04	5757.913	6.394	0.000	2.55e+04	4.81e+04
city_Fall_City	4.52e+04	1.02e+04	4.442	0.000	2.53e+04	6.51e+04
city_Federal_Way	-1.082e+05	2971.929	-36.392	0.000	-1.14e+05	-1.02e+05
city_Issaquah	5.118e+04	3850.319	13.291	0.000	4.36e+04	5.87e+04
city_Kenmore	-1.291e+04	4803.243	-2.688	0.007	-2.23e+04	-3495.800
city_Kent	-9.621e+04	2544.701	-37.807	0.000	-1.01e+05	-9.12e+04
city_Kirkland	2.176e+04	3160.847	6.883	0.000	1.56e+04	2.8e+04
city_Maple_Valley	-7.34e+04	3827.903	-19.175	0.000	-8.09e+04	-6.59e+04
city_Mercer_Island	6.572e+04	1.42e+04	4.639	0.000	3.79e+04	9.35e+04
city_North_Bend	2.198e+04	5733.272	3.833	0.000	1.07e+04	3.32e+04
city_Redmond	4.526e+04	3812.316	11.871	0.000	3.78e+04	5.27e+04
city_Renton	-5.483e+04	2364.156	-23.191	0.000	-5.95e+04	-5.02e+04
city_Sammamish	-6.644e+04	5279.312	-12.586	0.000	-7.68e+04	-5.61e+04
city_Snoqualmie	-4.975e+04	5512.356	-9.025	0.000	-6.06e+04	-3.89e+04
city_Vashon	9.083e+04	7798.541	11.647	0.000	7.55e+04	1.06e+05
city_Woodinville	-1.566e+04	4964.593	-3.155	0.002	-2.54e+04	-5931.546
sqft_living:floors	7.38e+04	4116.014	17.929	0.000	6.57e+04	8.19e+04
sqft_living:bathrooms	1.476e+04	2699.614	5.468	0.000	9468.780	2.01e+04

Omnibus:	474.537	Durbin-Watson:	1.969
Prob(Omnibus):	0.000	Jarque-Bera (JB):	774.917
Skew:	0.261	Prob(JB):	5.36e-169
Kurtosis:	3.905	Cond. No.	6.30e+06

<u>Sample Interpretations:</u>

- A one unit increase of bedrooms would **decrease** price by $11,730
- A one percent increase of sqft_living would **increase** price by $792
- The impact of whether a house is located in a certain city has a specific impact on price; as an example, if the home is located in Bellevue, in relation to Seattle (the dropped dummy term), the **increase** on price will be $47,320
- The effect of sqft_living on price is different for different values of floors; this interaction has a positive effect on price, and as floors increase by one unit, the impact of sqft_living on price will **increase** by $73,800

Conclusion & Recommendations

- Our model is able to predict housing price with approximately 70% accuracy
- Our model satisfies the linear assumptions of a regression model: residuals are normally distributed, features are not correlated with one another and the residuals follow homoscedasticity.
- Business Recommendations:
    - The following areas are positively related to price: Bellevue, Bothell, Fall City, Issaquah, Kirkland, Mercer Island, North Bend, Redmond and Vashon. We recommend the housing development company builds in one of these areas.
    - Certain features of homes will also positively impact their price: a nice view of the property, good condition of the homes, and a high square footage of the home's living space, especially for homes with a larger number of floors and bathrooms.

Thank you!

- Github Repo: https://github.com/adinas94/Phase-2-Project
- Contact information:
  - Adina Steinman: adinasteinman@gmail.com
  - George Paskalev: georgeppaskalev@gmail.com