4th December 2019

Rush Hour 4 Consultancy

Property Consultants

Presenters: Finn & Thomas
Location : Kings County

Your Questions Answered

- What are the top 5 house features that are the most important in determining house prices?
- How much in average is the price difference between waterfront and non-waterfront houses?
- Is timing important when selling / buying a house? Is there a relationship between the number of house transactions and house prices?

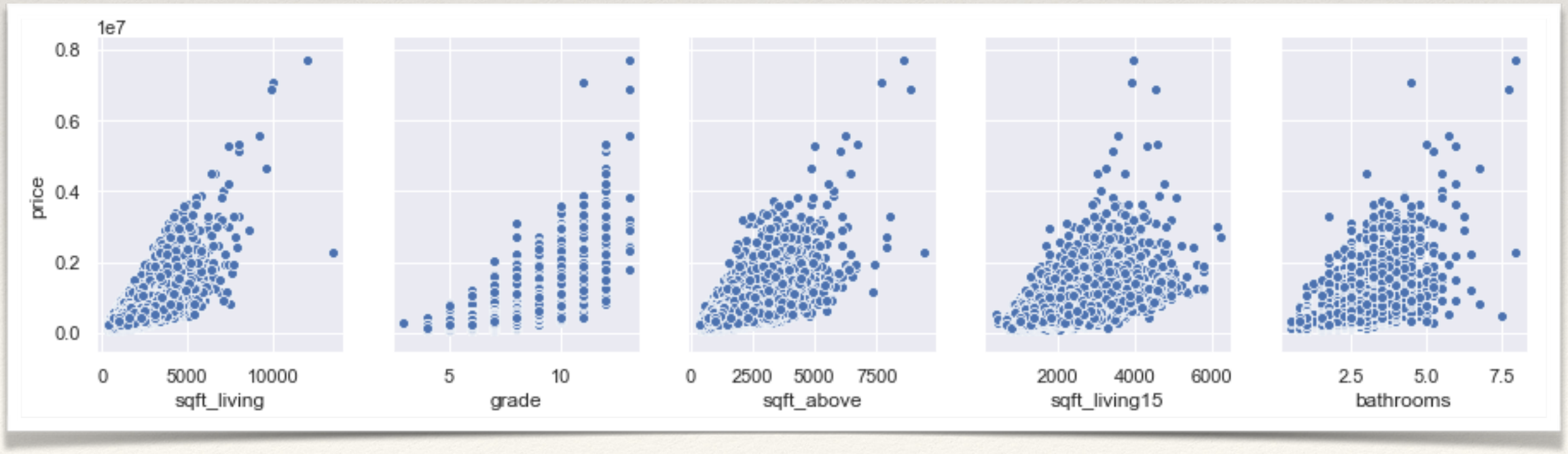
Top 5 House Features

How did we pick the top 5 features?

- Based on the available data, we used a correlation matrix and picked the features with the strongest relationship with price

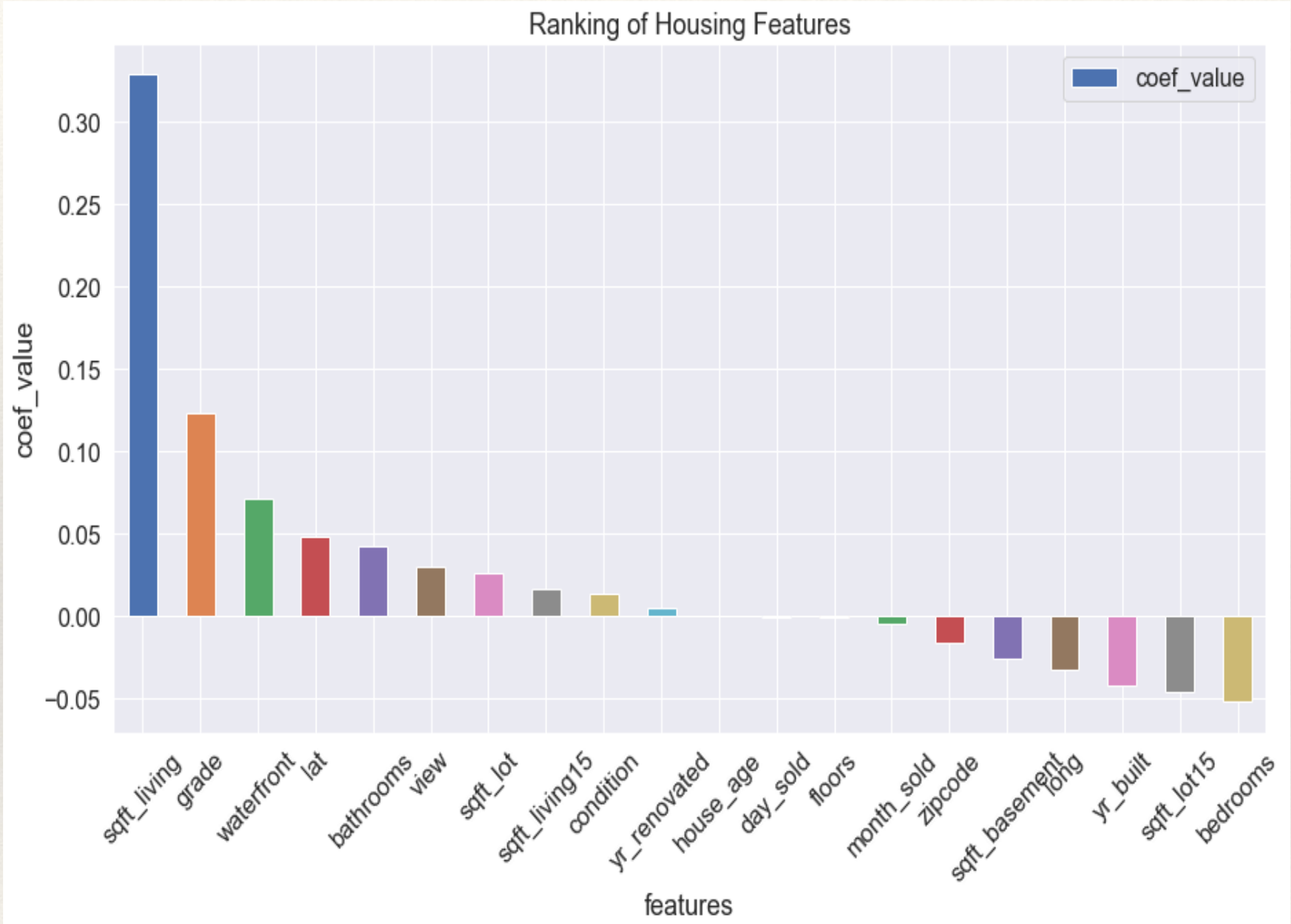
- As a result, we obtained the following features:
  - **Sqft living [0.7]** - a combination of sqft basement and soft above
  - **Grade [0.67]** - measures the quality of the house (architectural design, workmanship, materials), ranking from 1 to 13
  - **Sqft above [0.61]** - sqft of house excluding the basement
  - **Sqft living 15 [0.59]** - sqft living of your 15 nearest neighbours
  - **Bathrooms [0.53]** - number of bathrooms

- This is evident purely just by looking at scatter plots of each of the above house features against price as depicted below



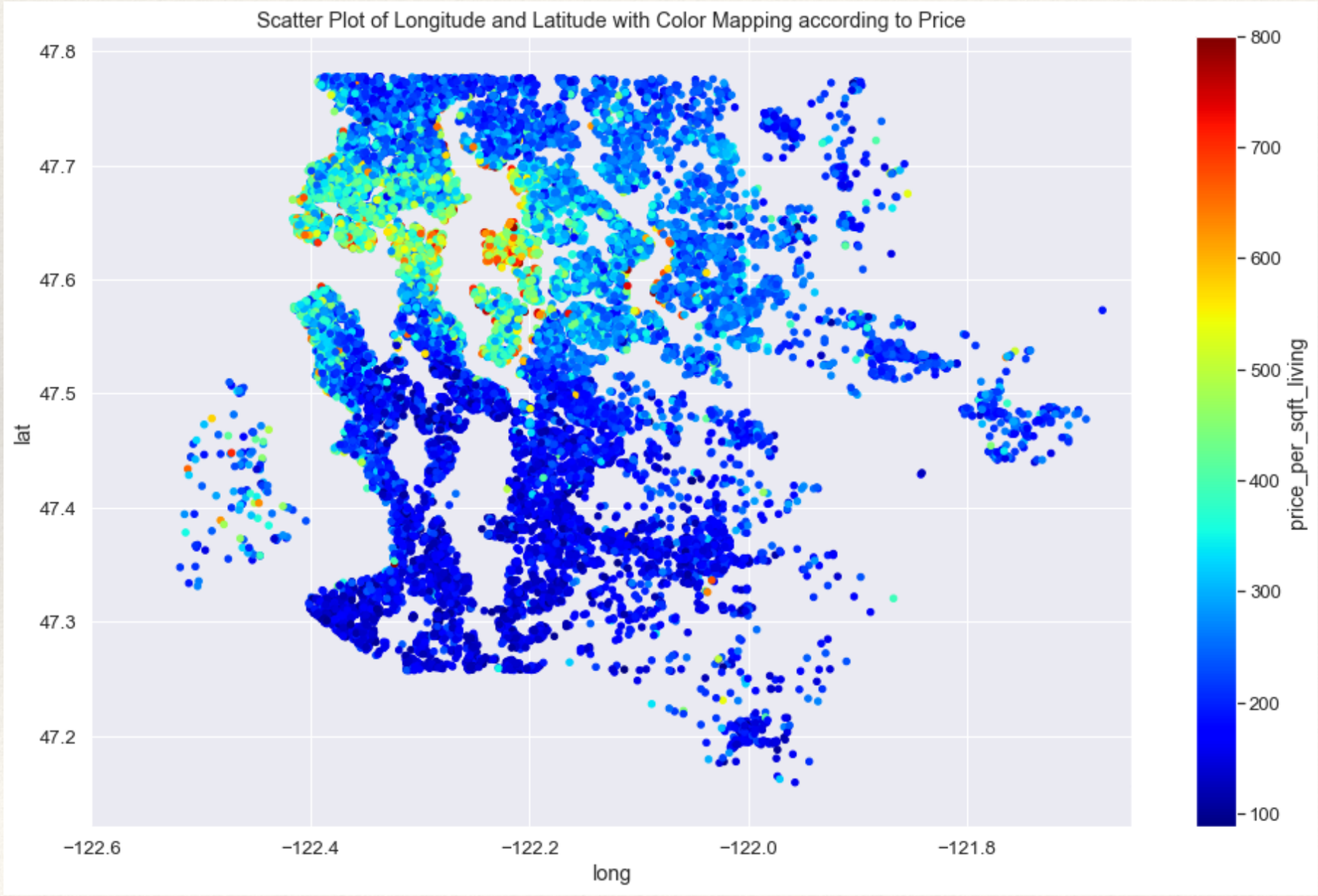
Other Remaining Features

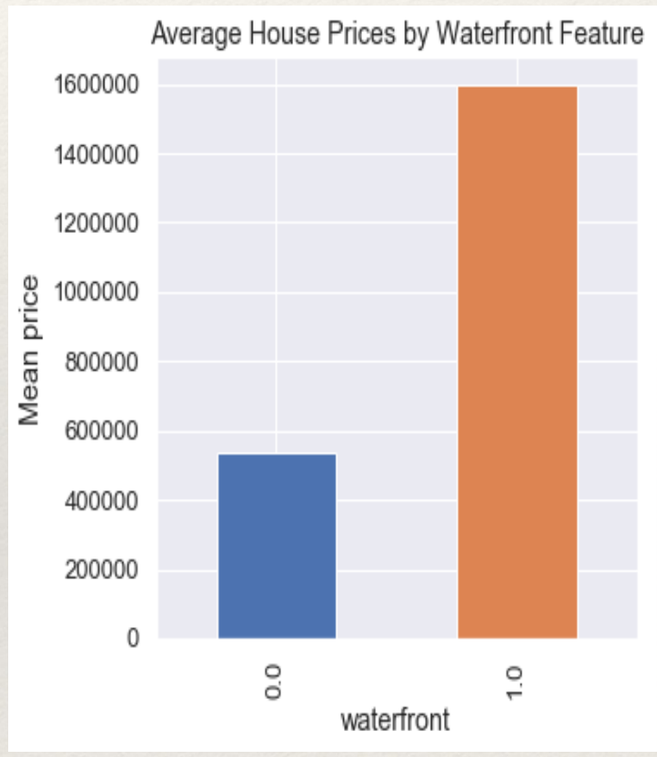
- Cautious in interpreting as one might think the features to the right are less important which is not true!
- Features on the right means they have negative relationship to price
- Whilst, features in the middle; house_age, day_sold and floors suggests no significant relationship to price
- Another interesting point here is that the first 5 features do not tie in with out initial top 5 features in the previous slide, largely due to various effects of interaction between features in a more complex model
- Nonetheless we can at least confirm that our earlier top 5 features are at within the top half of features!



Waterfront and Non-Waterfront Houses

- Initial insight suggests that houses near to beds of water tend to attract higher prices
- On average, waterfront properties yields in an average of $1m than non-waterfront properties
- That said, these values tend to differ for each of the top 5 features identified earlier which will be explored in the next slide



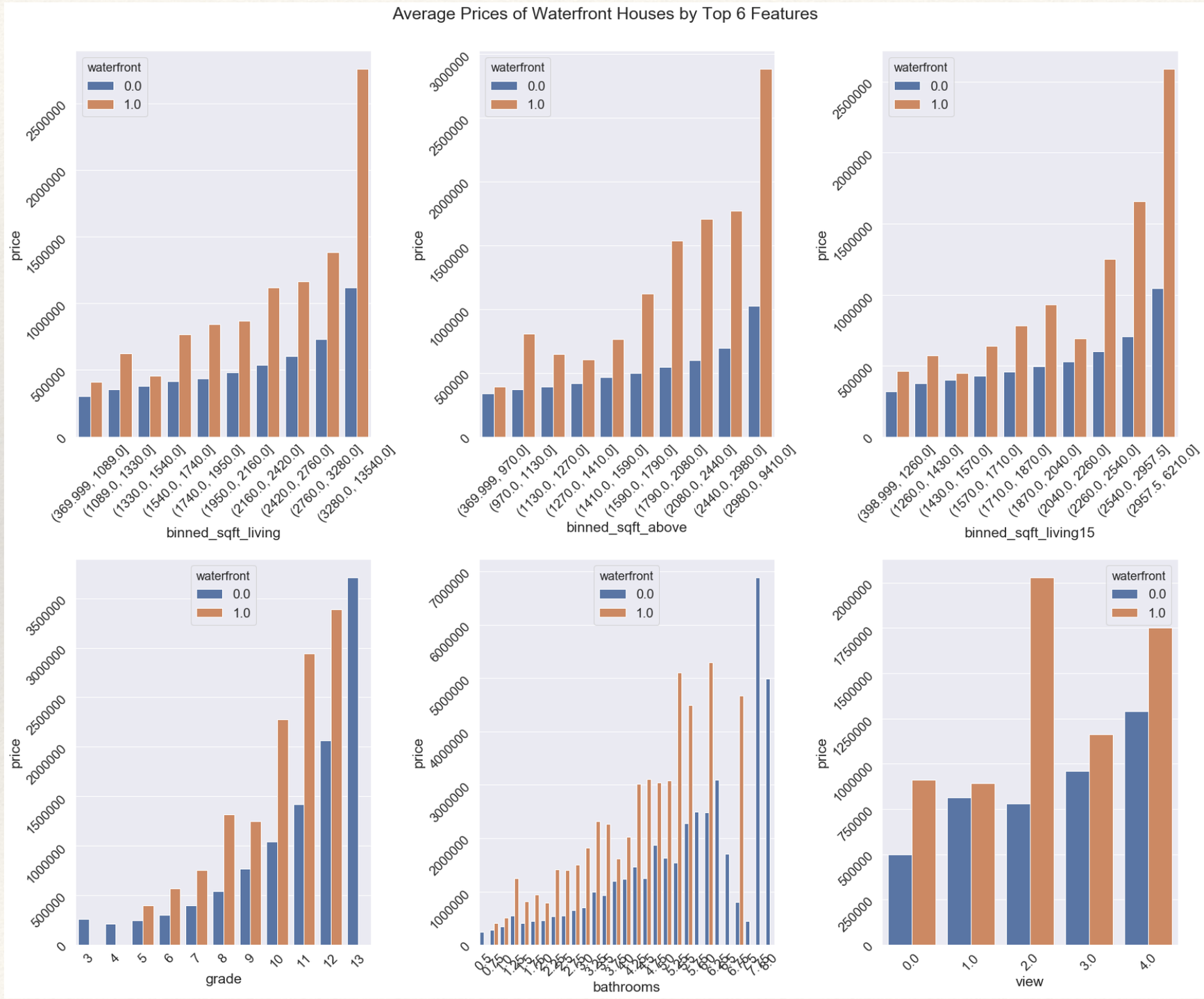


*** Would caveat that the plot on the left is not a true representative of the Kings County region but only serves as an approximation noting that there are areas without any transactions during the timeframe of the available data.*

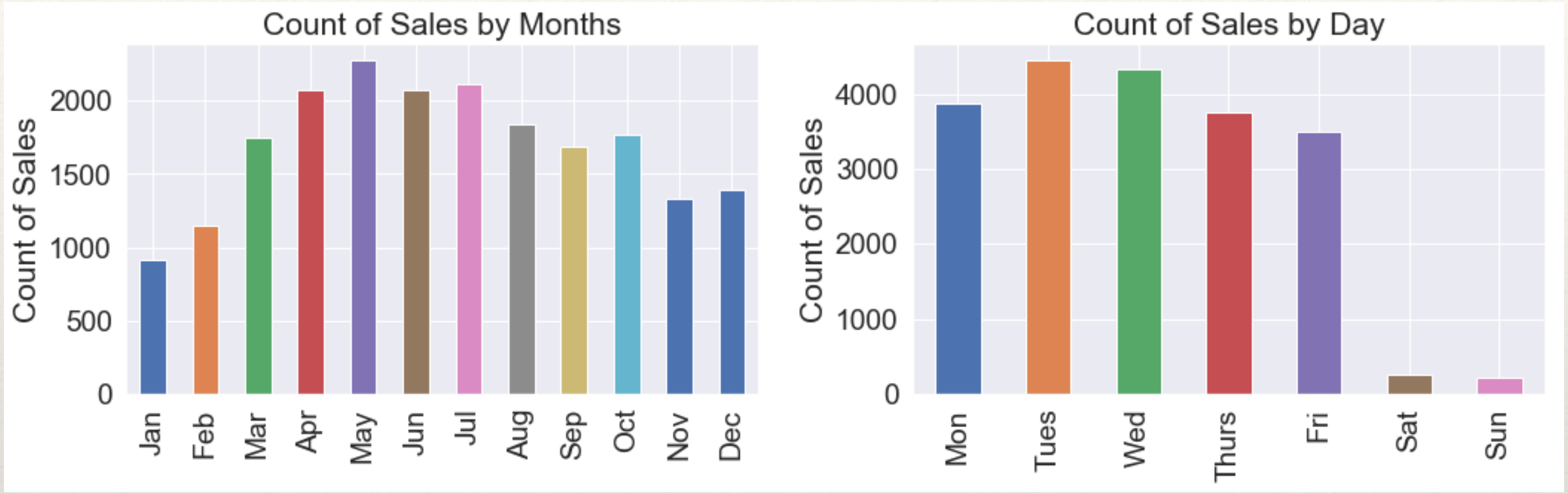
Average Prices of Waterfront Houses

- There seems to be a general trend across all of the top 5 features; waterfront properties attract higher prices across each unique elements in each feature

- Would however point out that, waterfront properties only represent a minor proportion of the total dataset hence would warn against putting too much weight / emphasis on this whenever deciding on a house purchase / sale.



Timing of Transactions

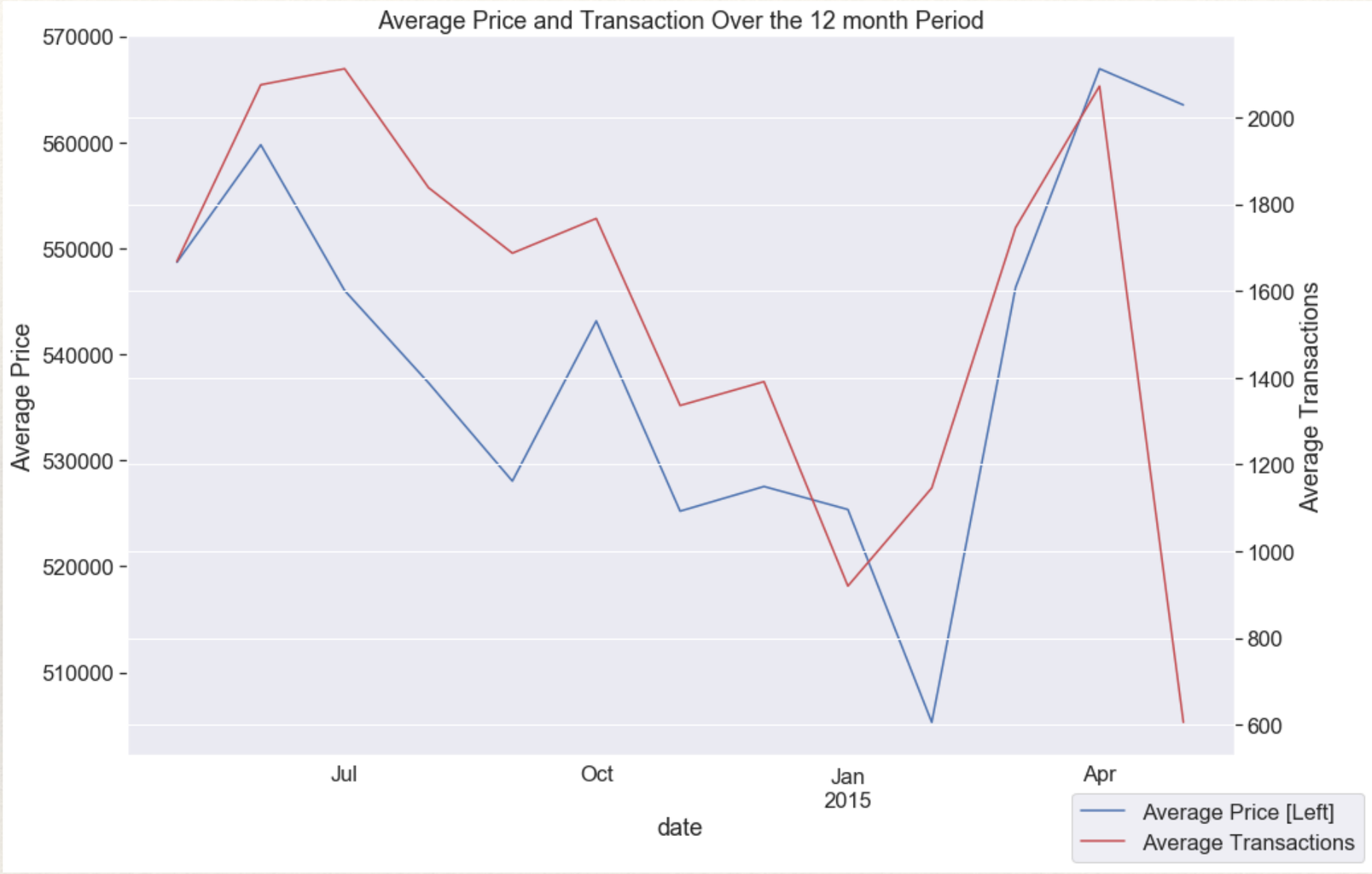


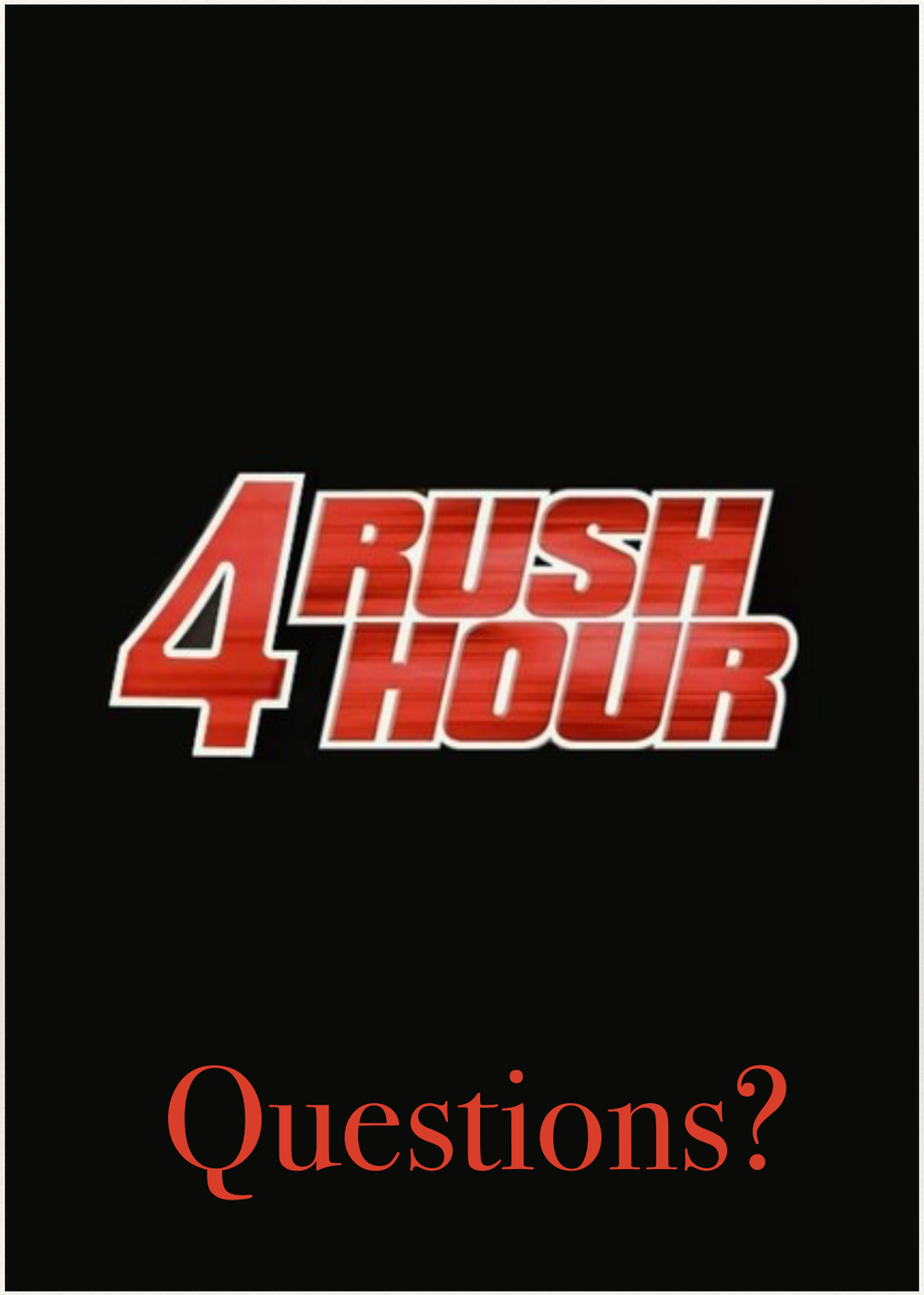
- Month of 'May' tends to have higher activity possibly due to various factors:
  - employee bonuses are typically paid in March / April, leading to an uptick in house transactions in May
  - warmer months meaning more viewings and a better conversion rate
  - lower purchases in January possibly due to higher Christmas season spending

- Referring to the 'day' count chart, weekend attracts the least number of transactions possibly due to:
  - not many property agents working during weekends
  - large fund transfers across the weekend not supported by majority of banks

Relationship of average price and transactions

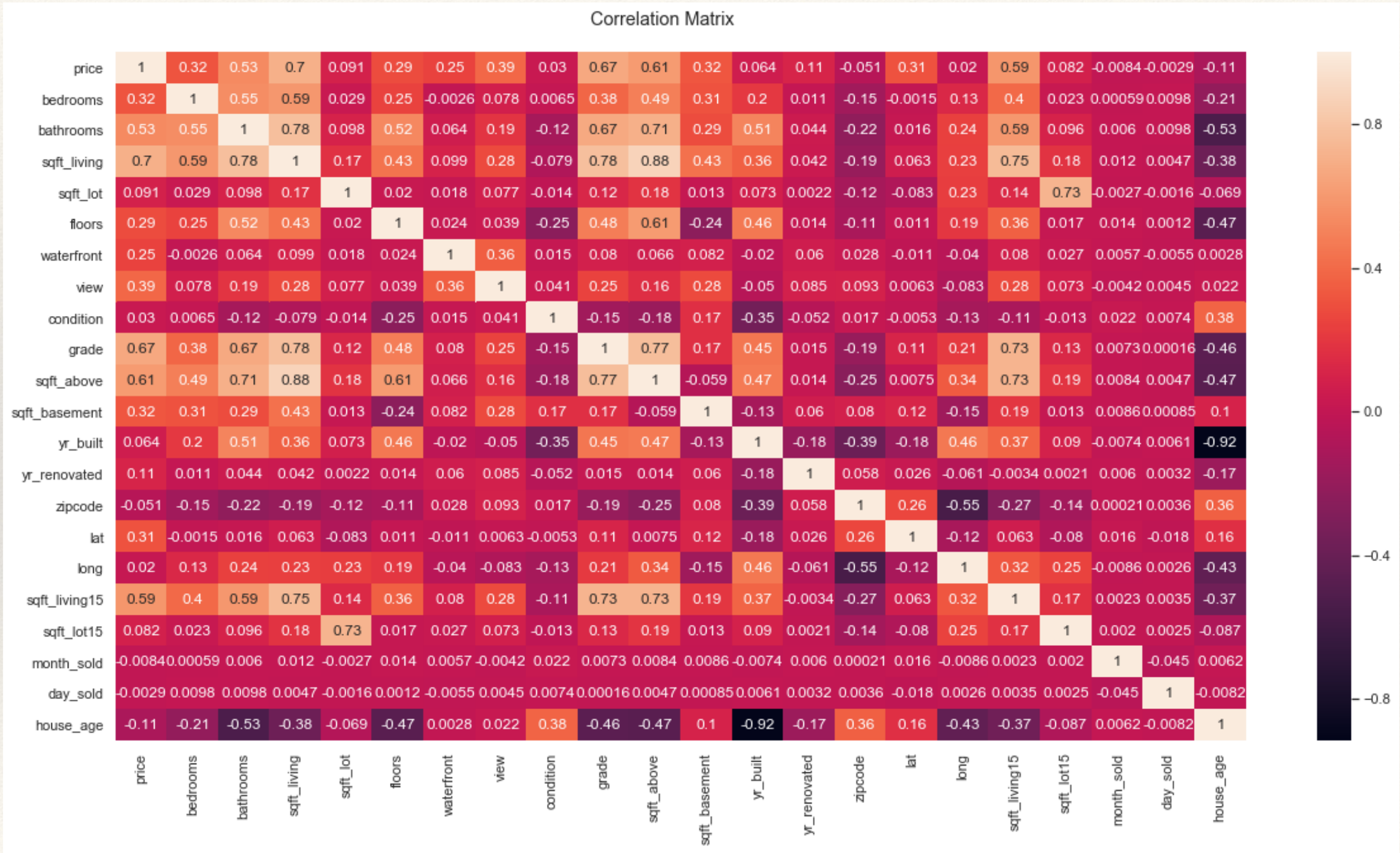
- Two separate line plots over the period of May-14 to May -15
- Blue represents the average price. Red represents the average number of transaction.
- Up to Apr-15, there is a clear trend between both lines suggesting as the avg. transaction falls, the average price will fall too. Feels like a simple economics 101, lower demand, lower prices.
- One shortfall is that the graph does not necessarily tell if there is a time lag between a drop in number of transactions and the prices. Could potentially be a good indicator of house prices




4 HOUR
Questions?

Appendix

Correlation Matrix



Top 5 Features OLS Breakdown

Simple Linear Regression:

	feature	Pearson_r	Pearson_r2	R2	P_value	Coef_value	Coef_interval	F_value	T_value	Jacque-Bera	Resid. Skew	Resid. Kurtosis
0	sqft_living	0.70	0.50	0.50	0.00	283.07	[279.14, 286.99]	19,965.33	141.30	505,678.25	2.83	26.75
1	grade	0.67	0.45	0.45	0.00	210,735.70	[207531.15, 213940.25]	16,614.63	128.90	1,918,106.94	4.08	49.85
2	sqft_above	0.61	0.37	0.37	0.00	269.46	[264.6, 274.32]	11,817.67	108.71	686,106.48	3.29	30.67
3	sqft_living15	0.59	0.35	0.35	0.00	317.41	[311.41, 323.41]	10,752.81	103.70	1,810,094.57	4.24	48.41
4	bathrooms	0.53	0.28	0.28	0.00	257,175.01	[251573.63, 262776.4]	8,098.69	89.99	819,912.33	3.43	33.32
5	view	0.39	0.15	0.15	0.00	189,848.61	[183739.73, 195957.5]	3,710.55	60.91	987,086.37	3.64	36.33

Simple Linear Regression with minmax scaling:

	feature	Pearson_r	Pearson_r2	R2	P_value	Coef_value	Coef_interval	F_value	T_value	Jacque-Bera	Resid. Skew	Resid. Kurtosis
0	sqft_living_minmax	0.70	0.50	0.50	0.00	0.49	[0.48, 0.5]	19,965.33	141.30	505,678.25	2.83	26.75
1	grade_minmax	0.67	0.45	0.45	0.00	0.28	[0.27, 0.28]	16,614.63	128.90	1,918,106.94	4.08	49.85
2	sqft_above_minmax	0.61	0.37	0.37	0.00	0.32	[0.31, 0.33]	11,817.67	108.71	686,106.48	3.29	30.67
3	sqft_living15_minmax	0.59	0.35	0.35	0.00	0.24	[0.24, 0.25]	10,752.81	103.70	1,810,094.57	4.24	48.41
4	bathrooms_minmax	0.53	0.28	0.28	0.00	0.25	[0.25, 0.26]	8,098.69	89.99	819,912.33	3.43	33.32
5	view_minmax	0.39	0.15	0.15	0.00	0.10	[0.1, 0.1]	3,710.55	60.91	987,086.37	3.64	36.33

Simple Linear Regression with log transformation:

	feature	Pearson_r	Pearson_r2	R2	P_value	Coef_value	Coef_interval	F_value	T_value	Jacque-Bera	Resid. Skew	Resid. Kurtosis
0	grade_log	0.70	0.48	0.48	0.00	2.46	[2.42, 2.49]	19,099.57	138.20	173.68	0.20	3.20
1	sqft_living_log	0.68	0.46	0.46	0.00	0.85	[0.83, 0.86]	17,385.32	131.85	144.23	0.18	2.78
2	sqft_living15_log	0.61	0.38	0.38	0.00	1.00	[0.98, 1.01]	12,328.54	111.03	518.76	0.32	3.46
3	sqft_above_log	0.59	0.35	0.35	0.00	0.73	[0.72, 0.74]	10,855.30	104.19	170.29	0.21	2.87
4	bathrooms_log	0.53	0.28	0.28	0.00	0.71	[0.69, 0.72]	7,766.04	88.13	284.13	0.29	3.06

- Looks like there were no improvements in the model apart from smaller coef_values and narrower confidence intervals

- Comparing both log-transform and non-log transform data, we see improvements in:
  - R2 values (except sqft_living) though it is fairly marginal
  - Narrower confidence interval
  - More uniform residuals as seen by the JB, skew and kurtosis measurements

Simple Linear Regression Error Graphs

Residuals for the log-transformed model is now more uniformed across the spectrum of the log feature

Multiple Linear Regression Results

Non-scaled

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.701
Model:	OLS	**Adj. R-squared:**	0.700
Method:	Least Squares	**F-statistic:**	2381.
Date:	Tue, 03 Dec 2019	**Prob (F-statistic):**	0.00
Time:	21:41:23	**Log-Likelihood:**	-2.7777e+05
No. Observations:	20359	**AIC:**	5.556e+05
Df Residuals:	20338	**BIC:**	5.558e+05
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.096e+07	3.08e+06	3.564	0.000	4.93e+06	1.7e+07
bedrooms	-3.911e+04	2063.061	-18.956	0.000	-4.32e+04	-3.51e+04
bathrooms	4.389e+04	3445.489	12.737	0.000	3.71e+04	5.06e+04
sqft_living	190.2965	3.906	48.714	0.000	182.640	197.953
sqft_lot	0.1217	0.051	2.405	0.016	0.023	0.221
floors	-2784.3607	4049.915	-0.688	0.492	-1.07e+04	5153.800
waterfront	5.445e+05	1.74e+04	31.258	0.000	5.1e+05	5.79e+05
view	5.771e+04	2178.903	26.485	0.000	5.34e+04	6.2e+04
condition	2.679e+04	2420.030	11.071	0.000	2.2e+04	3.15e+04
grade	9.45e+04	2259.628	41.821	0.000	9.01e+04	9.89e+04
sqft_basement	-39.8629	4.610	-8.647	0.000	-48.899	-30.827
yr_built	-2790.5254	236.877	-11.780	0.000	-3254.824	-2326.226
yr_renovated	20.1908	6.844	2.950	0.003	6.777	33.605
zipcode	-609.3829	34.099	-17.871	0.000	-676.219	-542.546
lat	5.972e+05	1.11e+04	53.924	0.000	5.76e+05	6.19e+05
long	-2.065e+05	1.37e+04	-15.127	0.000	-2.33e+05	-1.8e+05
sqft_living15	21.6798	3.559	6.092	0.000	14.704	28.656
sqft_lot15	-0.3970	0.077	-5.186	0.000	-0.547	-0.247
month_sold	-2817.9972	460.450	-6.120	0.000	-3720.516	-1915.479
day_sold	-161.3426	979.504	-0.165	0.869	-2081.249	1758.564
house_age	6.3746	240.301	0.027	0.979	-464.634	477.383

Omnibus:	17137.604	**Durbin-Watson:**	1.995
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1649740.336
Skew:	3.522	**Prob(JB):**	0.00
Kurtosis:	46.534	**Cond. No.**	2.17e+08

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.17e+08. This might indicate that there are
strong multicollinearity or other numerical problems.

Min-Max Scaled

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.701
Model:	OLS	**Adj. R-squared:**	0.700
Method:	Least Squares	**F-statistic:**	2381.
Date:	Tue, 03 Dec 2019	**Prob (F-statistic):**	0.00
Time:	21:41:23	**Log-Likelihood:**	44847.
No. Observations:	20359	**AIC:**	-8.965e+04
Df Residuals:	20338	**BIC:**	-8.949e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0399	0.004	-10.292	0.000	-0.048	-0.032
bedrooms	-0.0513	0.003	-18.956	0.000	-0.057	-0.046
bathrooms	0.0432	0.003	12.737	0.000	0.037	0.050
sqft_living	0.3288	0.007	48.714	0.000	0.316	0.342
sqft_lot	0.0264	0.011	2.405	0.016	0.005	0.048
floors	-0.0009	0.001	-0.688	0.492	-0.004	0.002
waterfront	0.0714	0.002	31.258	0.000	0.067	0.076
view	0.0303	0.001	26.485	0.000	0.028	0.033
condition	0.0141	0.001	11.071	0.000	0.012	0.017
grade	0.1240	0.003	41.821	0.000	0.118	0.130
sqft_basement	-0.0252	0.003	-8.647	0.000	-0.031	-0.019
yr_built	-0.0421	0.004	-11.780	0.000	-0.049	-0.035
yr_renovated	0.0053	0.002	2.950	0.003	0.002	0.009
zipcode	-0.0158	0.001	-17.871	0.000	-0.018	-0.014
lat	0.0487	0.001	53.924	0.000	0.047	0.050
long	-0.0326	0.002	-15.127	0.000	-0.037	-0.028
sqft_living15	0.0165	0.003	6.092	0.000	0.011	0.022
sqft_lot15	-0.0453	0.009	-5.186	0.000	-0.062	-0.028
month_sold	-0.0041	0.001	-6.120	0.000	-0.005	-0.003
day_sold	-0.0001	0.001	-0.165	0.869	-0.002	0.001
house_age	9.618e-05	0.004	0.027	0.979	-0.007	0.007

Omnibus:	17137.604	**Durbin-Watson:**	1.995
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1649740.336
Skew:	3.522	**Prob(JB):**	0.00
Kurtosis:	46.534	**Cond. No.**	127.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Issues encountered during modelling

- The multiple linear regression model is probably not the best model for the task given the some of it assumptions have not been satisfied:
  - still remains significant multicollinearity
  - non-normal error distribution
  - non-linear relationship for some features

- Other potential issues:
  - possibility of over fitting as we've not done any train-test split / cross validation here
  - interpreting the coef values can be quite misleading as we've done some feature scaling here
  - dropping more features with higher correlation (set initially at 0.8) may reduce R2, hence need to find balance here

Ranking Feature Importance by P-Value