King County Housing Data

Flatiron Project 1

Akshay Ghalsasi

The data and the questions for analysis

- King County Housing prices May 2014- May 2015

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
0	7129300520	10/13/2014	221900.0	3	1.00	1180	5650	1.0	NaN	0.0	3	7	1180	0.0	1955	0.0	98178	47.5112	-122.257	1340	5650
1	6414100192	12/9/2014	538000.0	3	2.25	2570	7242	2.0	0.0	0.0	3	7	2170	400.0	1951	1991.0	98125	47.7210	-122.319	1690	7639
2	5631500400	2/25/2015	180000.0	2	1.00	770	10000	1.0	0.0	0.0	3	6	770	0.0	1933	NaN	98028	47.7379	-122.233	2720	8062
3	2487200875	12/9/2014	604000.0	4	3.00	1960	5000	1.0	0.0	0.0	5	7	1050	910.0	1965	0.0	98136	47.5208	-122.393	1360	5000
4	1954400510	2/18/2015	510000.0	3	2.00	1680	8080	1.0	0.0	0.0	3	8	1680	0.0	1987	0.0	98074	47.6168	-122.045	1800	7503

- Original data has 20 relevant columns
- Questions we try to answer
  - How well can we predict the price of the house using linear models?
  - Can we make recommendations to the seller of the house regarding when to sell it?
  - Can we the give the buyer of the house a list of houses for his needs?

Business Cases for questions

- Q1 - How well can we predict the price of the house using linear models?

Ans: Useful for both buyers and sellers to get a fair price. Good predictions of prices can be monetized

- Q2 - Can we make recommendations to the seller of the house regarding when to sell it?

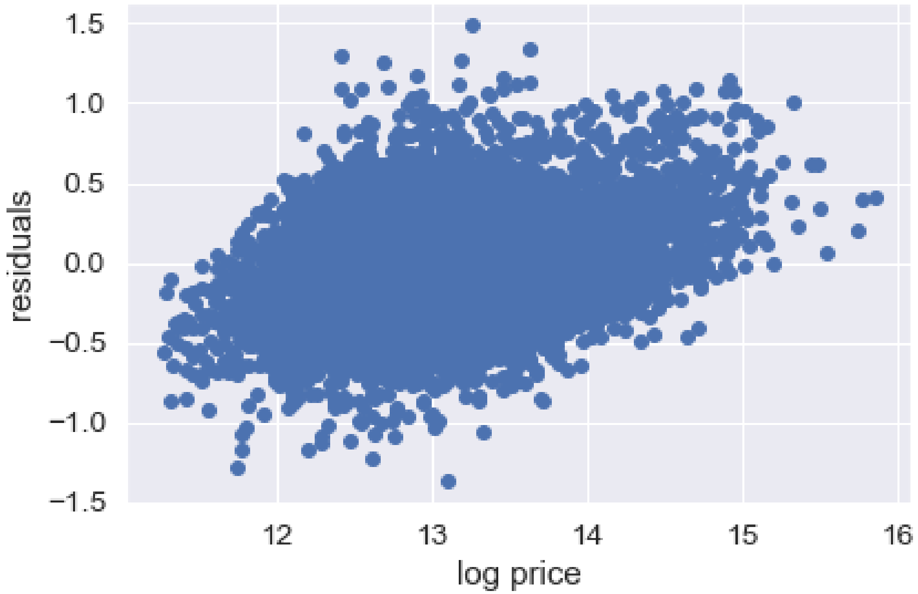
Ans: Useful for sellers. A sophisticated analysis will give confidence levels on when best to sell houses

Q1 - The Benchmark

- Need a benchmark to know if our model is doing good

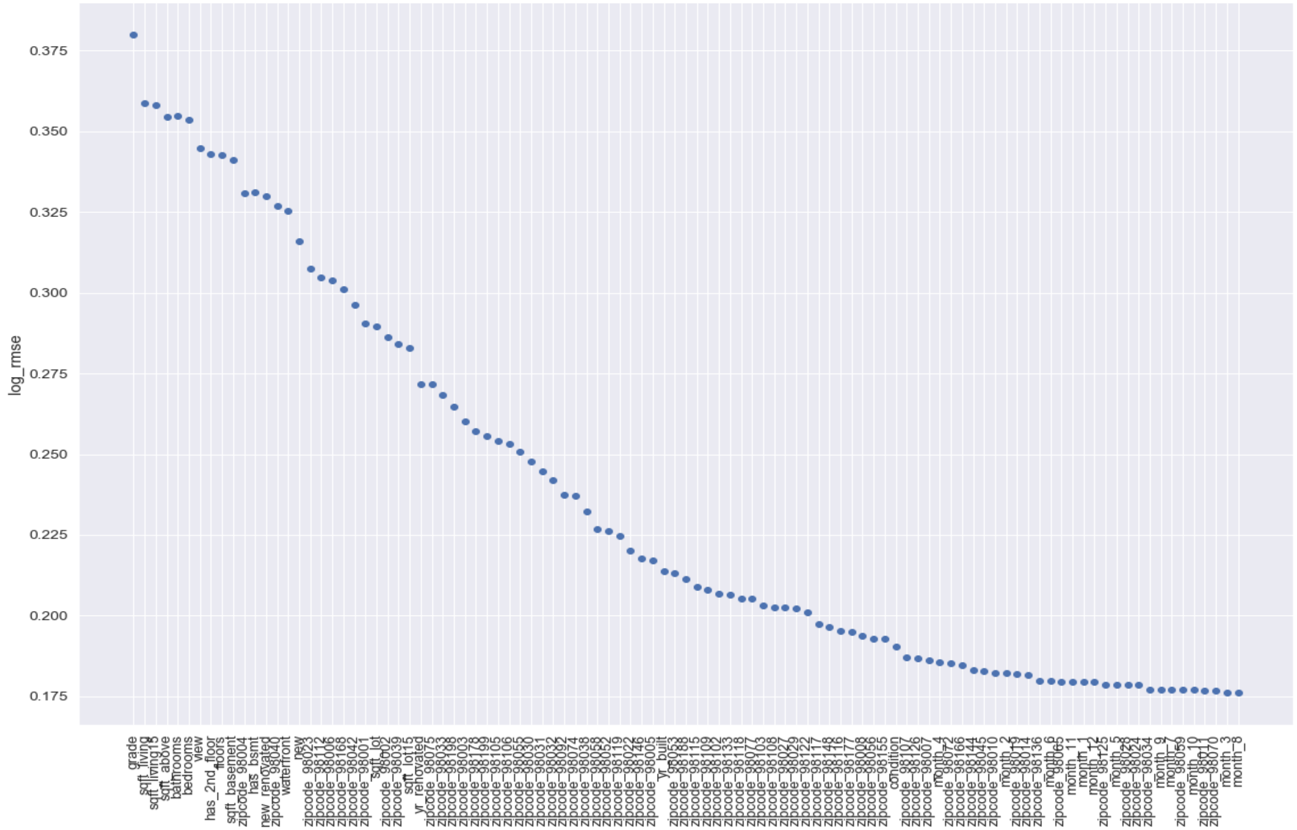
$$Price = Avg\left(\frac{price}{sqft}\right) * sqft$$

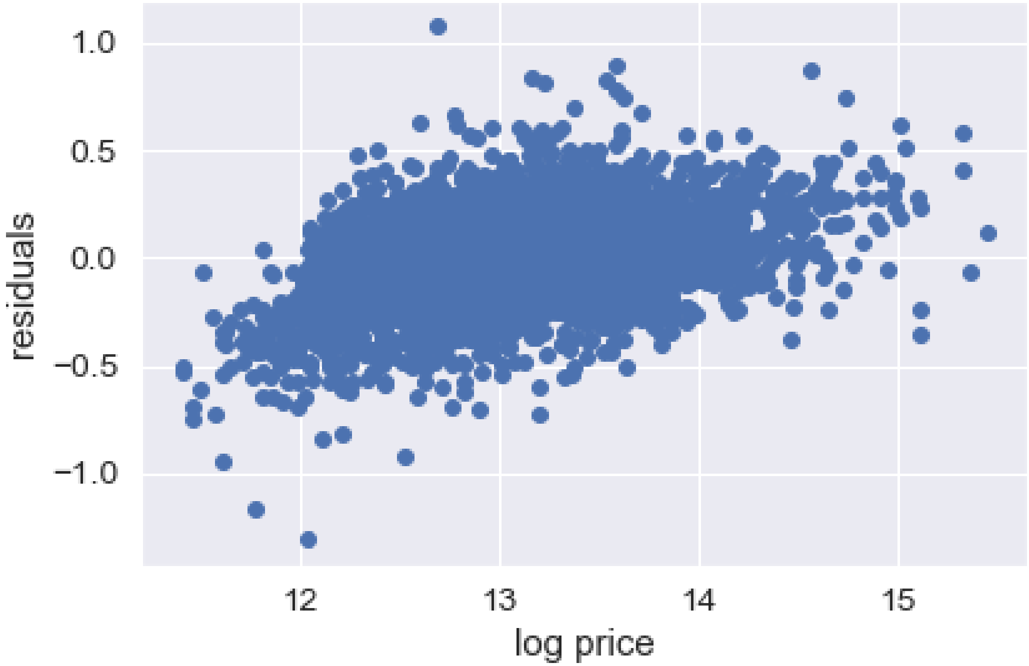
- Gives log RMSE of 0.25, need to beat this



Q1 - Final Model

- Final model (Ridge Regression)contains 101 features predicting price (see blog)
- We get a log RMSE of 0.18, much better than benchmark model



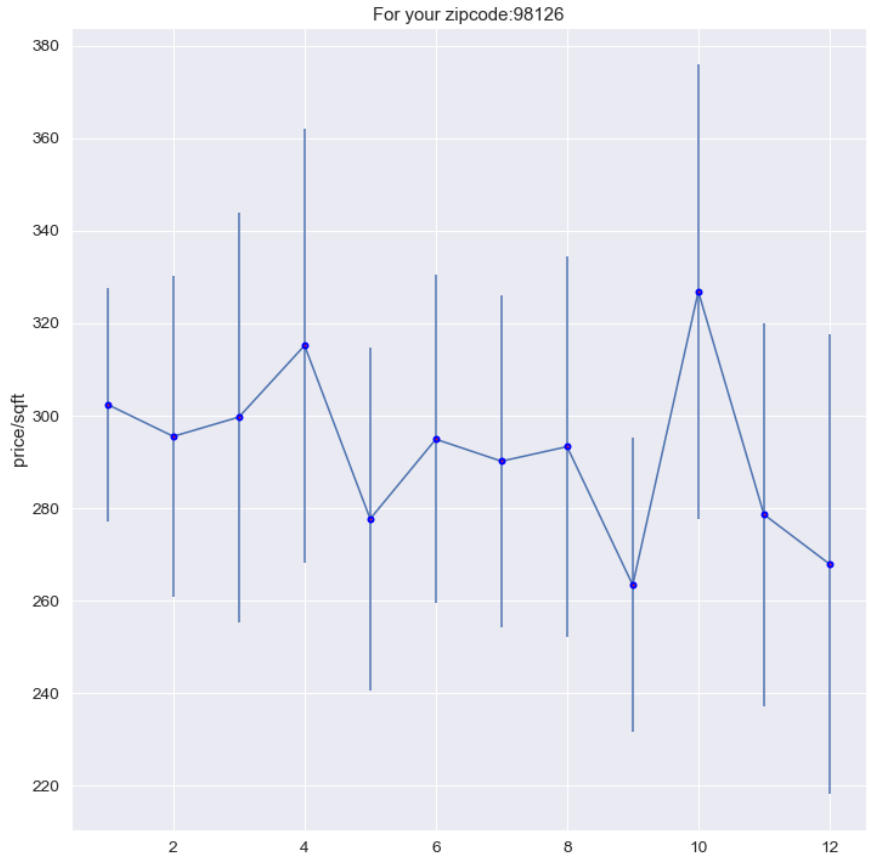


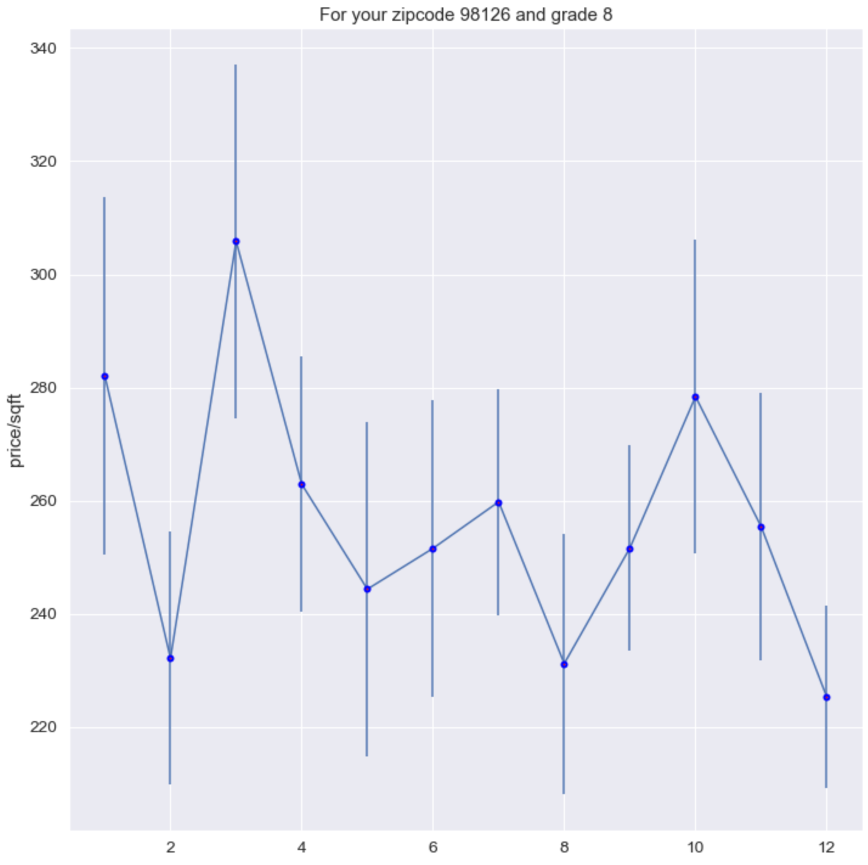
Q1 Conclusions and Future Work

- Final score of 0.176 is much better than our benchmark naive predictions
- The residuals look better but still a dependence on target variable. Room for improvement
- Need better feature engineering.
- Try alternatives to ridge regression

Q2 – Zipcode and Grade of House

- Use zipcode and grade to group data. Use mean to give predictions and std dev to visualize





Q2 – Conclusion and Future work

- Trying to predict best/worst times over all zipcodes and grades we get on average the best time to sell in April, worst is January
- For future work, can conduct hypothesis testing to give confidence levels on best time to sell house