FAIR HOME PRICE PREDICTOR

A multiple linear regression analysis on the King County home sales dataset

“To us, our house was not unsentient matter -- it had a heart, and a soul, and eyes to see us with... we lived in its grace and in the peace of its benediction.”

TWAIN

AGENDA

1. Overview
2. The data
3. EDA and data cleaning
4. The solution: Models 1 - 9
5. Looking at a few coefficients
6. The home price predictor
7. Conclusion

OVERVIEW AND BUSINESS UNDERSTANDING

1. Not always easy to determine “fair price” of a home

   The housing market has been on a wild ride the last few years

2. Assist homebuyers moving to King County, WA

   Could also be used by KC government or local RE agencies

3. Input home sale data and receive a fair price estimate

   Using multiple linear regression machine learning model

4. Models judged on R-squared value and root mean squared error

   Recommendations, inferences, limitations

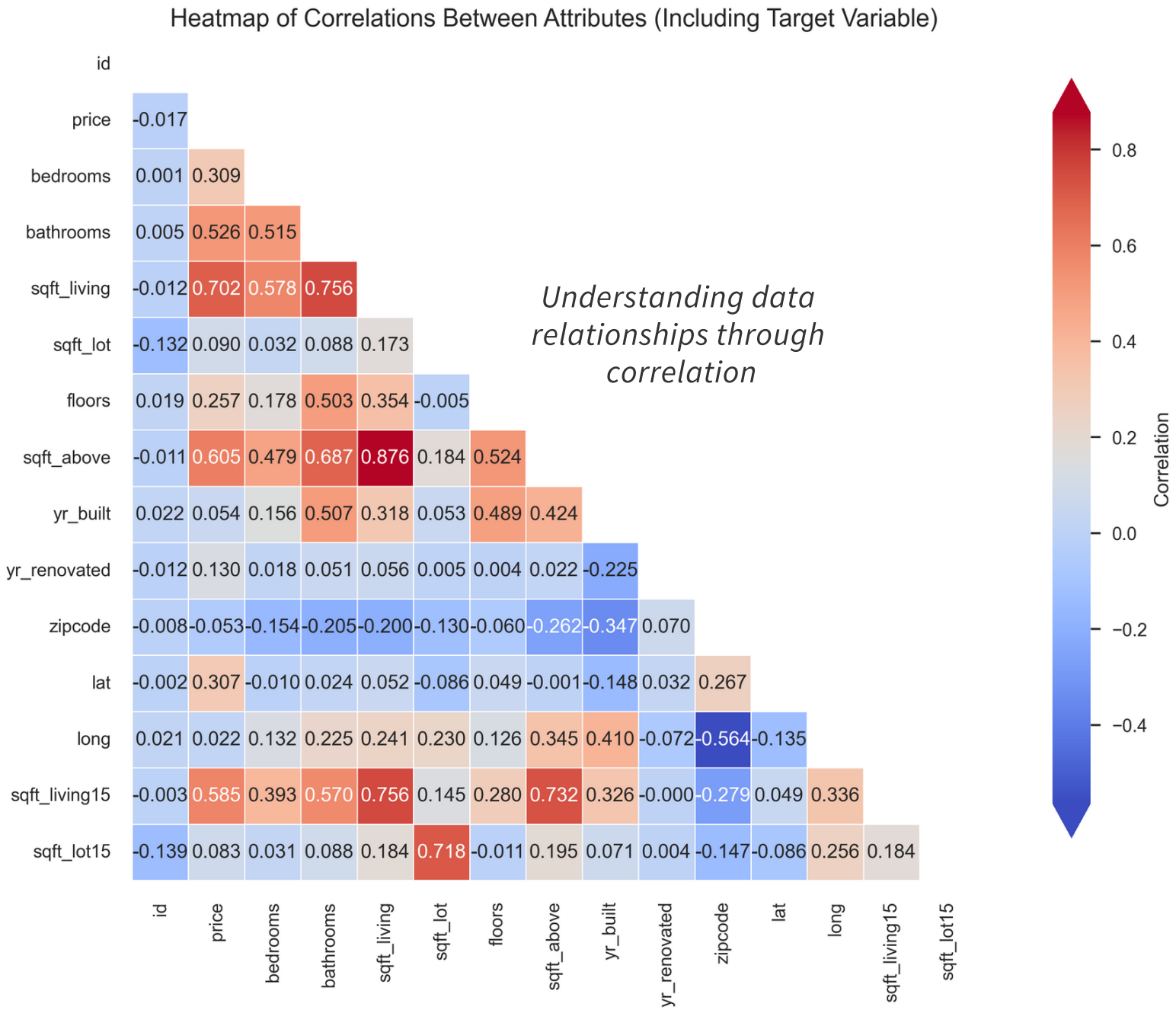
DATA UNDERSTANDING

> 21k home sales in King County, WA

From May 2014 - May 2015

21 unique variables

Price, date, square footage, etc.



EXPLORATORY DATA ANALYSIS

- Strong degree of outliers - bedrooms, bathrooms, square footage
- Visualization of Numerical Features with Correlation heatmap
- NaN values; '?' values

DATA CLEANING

- Replace missing values
- Transform datatypes to more workable formats
- Transform 'waterfront' into binary categorical

BASELINE MODEL AND MODEL ITERATIONS 1 - 4

Baseline model: Square feet of living space	R^2 = 0.48 RMSE = $266,730
Model 1: Add all numerical features	R^2 = 0.59 RMSE = $238,655
Model 2: Transform zero-inflated to binary form	R^2 = 0.59 RMSE = $238,411
Model 3: Add ordinal categorical variables	R^2 = 0.64 RMSE = $224,209
Model 4: Add 'month' dummy variables	R^2 = 0.64 RMSE = $223,916

X

MODEL ITERATIONS 5 - 9

Model 5: Add zip code dummy variables	R^2 = 0.79 RMSE = $168,489	
Model 6: 'total_sqft' & 'total_rooms'	R^2 = 0.79 RMSE = $169,175	X
Model 7: Log transform right-skewed features and sale price	R^2 = 0.888 RMSE = $118,749	
Model 8: Polynomial features to 2nd degree	R^2 = 0.891 RMSE = $115,689	
Model 9: Recursive feature elimination	R^2 = 0.891 RMSE = $115,719	

CONCLUSIONS, RECOMMENDATIONS, AND NEXT STEPS

Conclusions

- Certain zip codes and waterfront correlated with high sale price
- Large outliers beyond $2M sale mark

	Feature	Scaled Coefficient
34	x0_98039	133.380325
16	x0_98004	115.342462
5	waterfront	56.208412

Recommendations

- Use model to set a benchmark, but don't expect perfection
- Use model to look for bargains, and determine “how much” house you can afford

Next steps

- Gather more years of data, and more features
- What if outliers were taken out? What about variables containing distance to points of interest?

THOUGHTS? QUESTIONS?

Email: rposternak@yahoo.com
GitHub: github.com/rjpost20
LinkedIn: linkedin.com/in/ryanposternak