House-Specific vs. Neighborhood Effects on Home Sale Price in King County, Washington State

John Kline

10/28/19

Framework & Data Sources

GOALS

What portion of house price is determined by neighborhood desirability vs. home-specific characteristics?

***Business Problem**: A real estate investor believes that rental rates for homes are driven by house-specific characteristics, while purchase price is determined by a combination of neighborhood and house-specific characteristics. This investor wants to identify regions where it will be a more capital-efficient way to generate rental cash flow i.e. a high ratio of house characteristics driving sales prices vs. neighborhood characteristics.*

METHODS

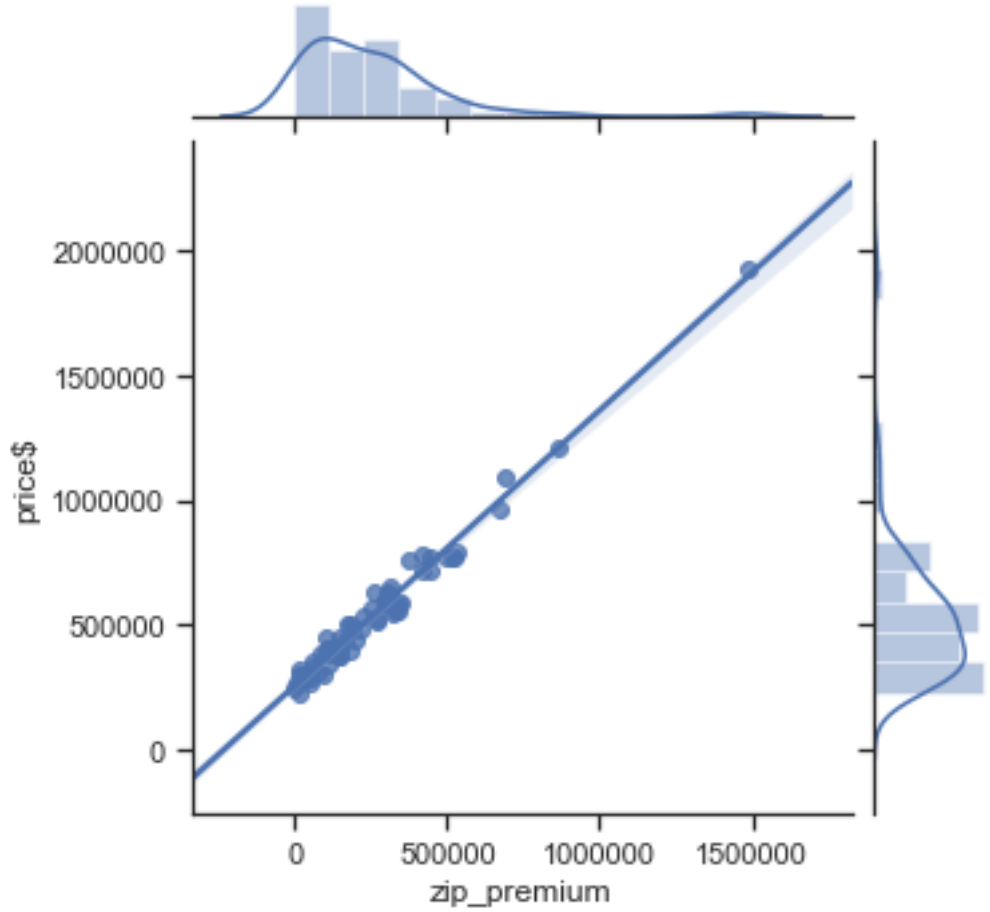
Multivariate linear regression in python

DATA SOURCES

- *2014 (May) – 2015 (May) housing price data from Kaggle 2016 dataset (*https://www.kaggle.com/harlfoxem/housesalesprediction)

Zipcode Price Premium

Avg. Home Price vs. Zipcode Price Premium



Note: *Each point on the graph represents a unique zip code in King County*

Findings Summary

- **Location, Location, Location** – a home's neighborhood (i.e. zipcode) is a crucial predictor of average home sale price
- **Zipcode adds predictive power even controlling for other factors**

Without Zipcodes	*With Zipcodes*
52% *of price variance*	**85%** *of price variance*

Methods Overview

1

Data Detail:

- **~21,500 home sales in King County, from 2014-2015**

2

Home Factors Regression:

- **Identified top usable home-specific predictors of price**
  - Living space square footage
  - Lot size
  - Home condition rating
  - Waterfront proximity
  - Renovation status
  - Number of floors
  - Whether house was viewed
- **Regressed predictors against the natural logarithm of price**

2

Zip Code Regression

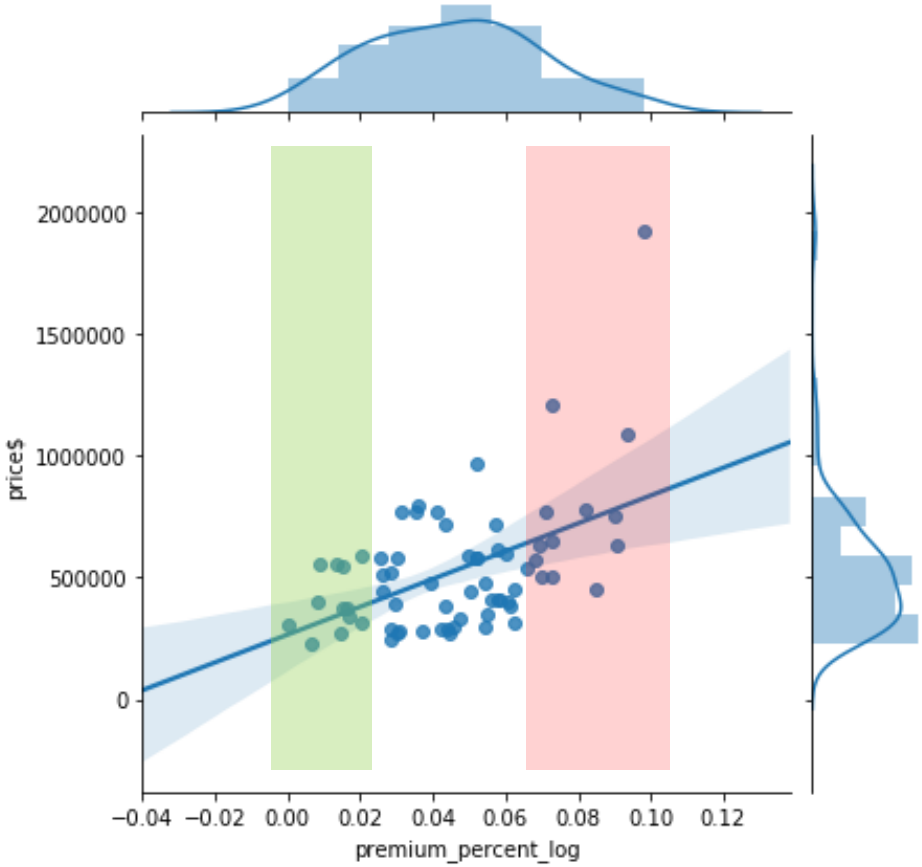
- **Controlling for the top home-specific price predictors listed above, regressed each zip code against the natural log of home sale price**
- **Grouped results by zip code to produce average home sale by zip code and average zip code impact on price (zip code premium)**

Client Prioritization Areas

Findings

- **Our client should prioritize the zip codes in green and avoid the zip codes in red:** these should have the highest and lowest respective ratios of potential rental cash flow to home sale price
- **All premiums aren't equal:** the average percent of home value due to zip code premium varies greatly (0-10%) and tends to be higher for more expensive homes.

Avg Home Price vs. Proportion of Home Price Explained by Zipcode Premium



Client Recommendation Summary

Recommendations

1. **We have identified 12 zip codes that should generate the highest rental cashflow per dollar spent on purchase price** – these have the lowest proportion of home sale value driven by neighborhood characteristics

2. **14 zip codes should be avoided as they provide the lowest rental cashflow per dollar spent on purchase price** – they exhibit disproportionately high percentages of the home sale value driven by neighborhood characteristics

3. **The middle 39 zip codes may have some deals available, but should be investigated after the top-priority regions** – these regions have a moderate amount of sale price determined by neighborhood characteristics

Next Analytic Steps

1. **Comparison analysis of different ways of grouping houses by geography – do results persist? Are there more accurate groupings?**
   - Zip code vs. long/lat binning
   - Mapping socially understood “neighborhoods”, e.g. “Pac Heights” vs the “Tenderloin” in SF are not captured in zip codes
2. **What elements of the zip code premium can be explained by other quantitative factors?**
   - Distance to work areas / transit options
   - School quality
   - Unexamined house quality factors

Discussion & Questions

Thanks!

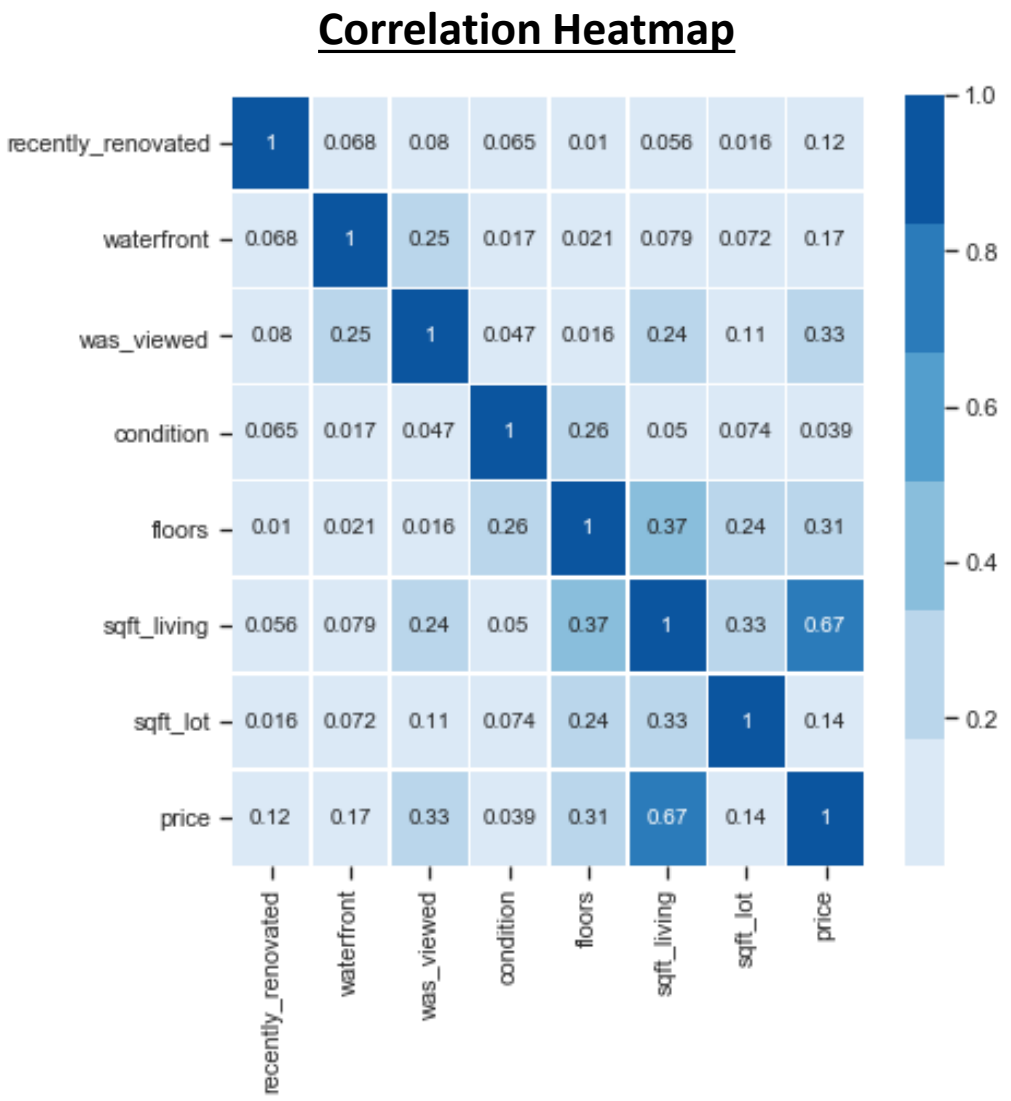
Appendix

Zipcodes by Priority Level

High-Priority	Moderate-Priority		Low-Priority
98022	98006	98065	98004
98023	98007	98070	98005
98030	98008	98072	98033
98031	98010	98074	98039
98032	98011	98075	98040
98038	98014	98077	98102
98042	98019	98106	98103
98055	98024	98108	98105
98058	98027	98118	98107
98092	98028	98125	98112
98148	98029	98126	98115
98168	98034	98133	98116
98178	98045	98136	98117
98188	98052	98144	98119
98198	98053	98146	98122
	98056	98155	98199
	98059	98166	
		98177	

Tests for OLS Validity - Main Regression

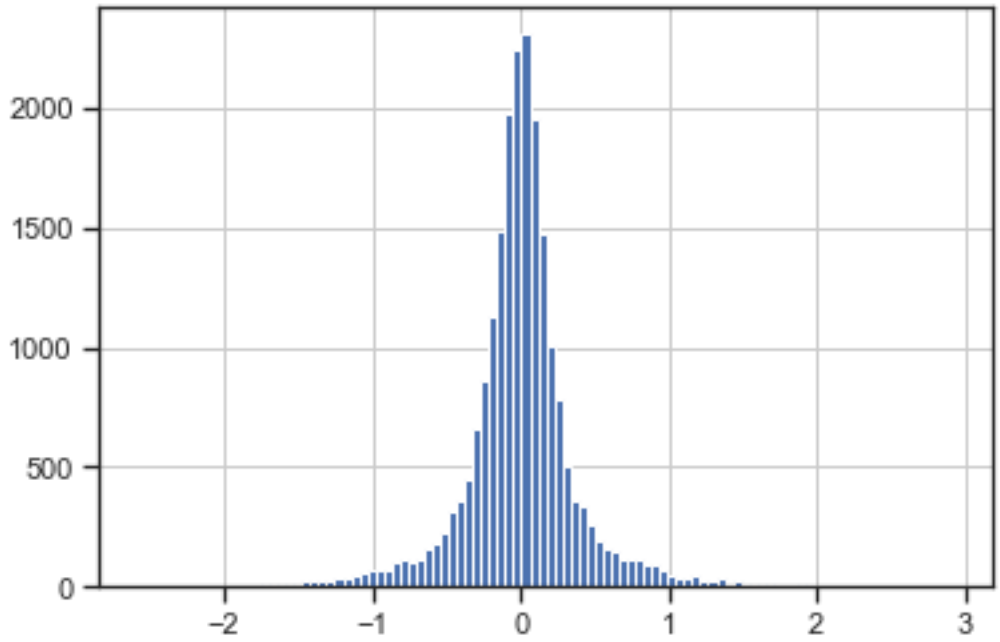
- Correlation among house-specific independent variables chosen for the analysis remained low – with a max of 0.37 in magnitude
- Price is the dependent variable – sqft_living is the independent variable with the highest correlation with price



Tests for OLS Validity - Main Regression

Regression residual graph does not demonstrate any marked heteroskedasticity…

Histogram of Regression Residuals
(Home Predictors & Zip Codes)

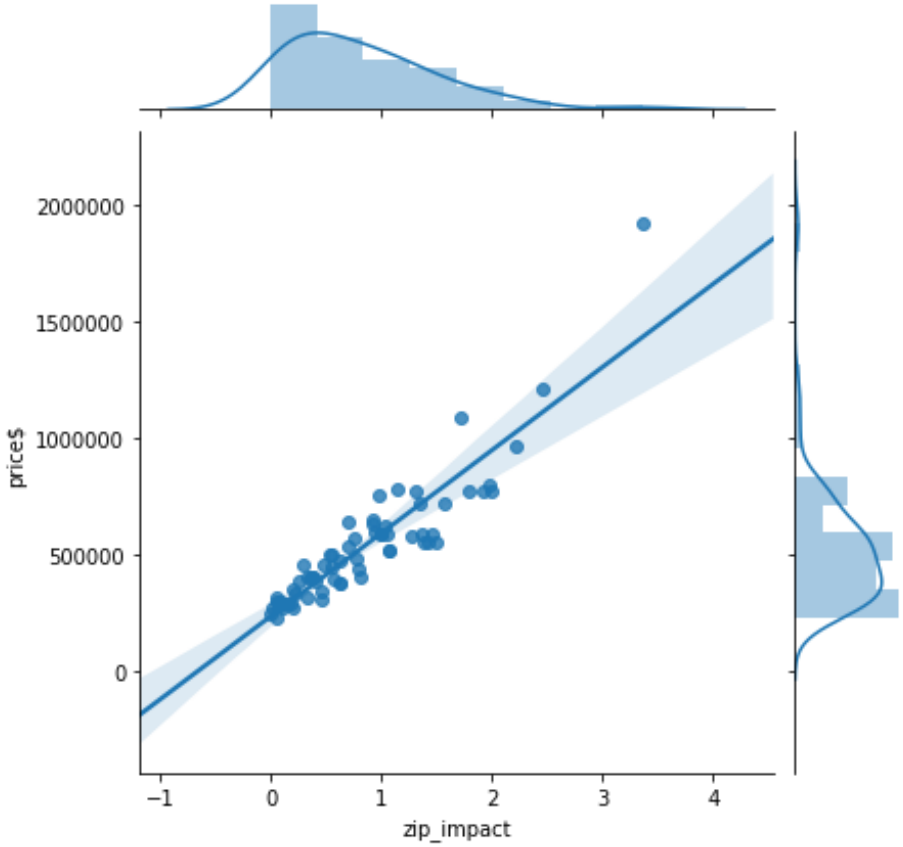


- **Adj. R-Squared:** 0.85
- **F-Stat. Prob.:** 0.00
- **Skew**: -0.135
- **Kurtosis**: 5.06 (slightly leptokurtic / fat-tailed)

Log-Transformed Outcome Detail

Econometric approach: exp(coeff)-1

- 1.00 corresponds to a +100% impact of the zip code dummy var.
- 3.4 corresponds to a +340% impact of the zipcode dummy var



Naïve / High-Impact: exp(total price-coeff)

- Cherry-picks last 0-1.4 ln(dollars), the most high-impact part of the contribution, and attributes it entirely to the zip code
- Seems to correspond to the econometric values