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

John Kline

11/8/19

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

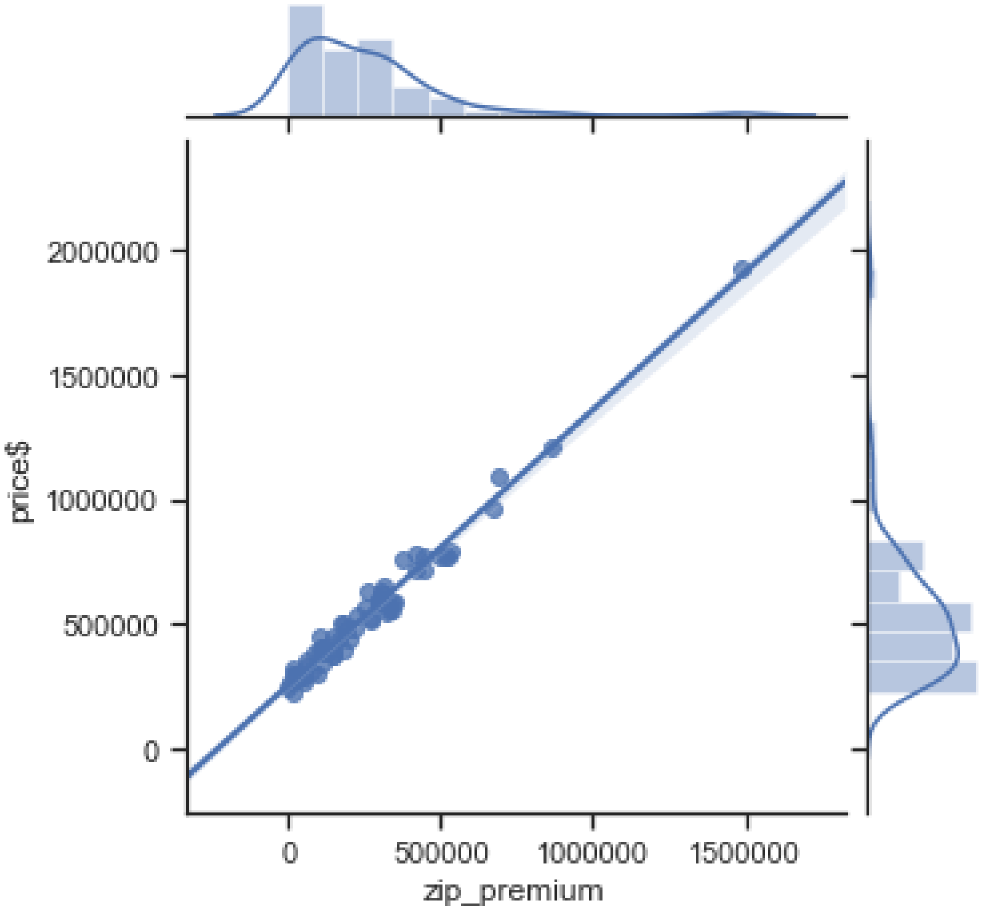
Multivariate linear regression in python

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

52%

of price variance

With Zipcodes

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)**

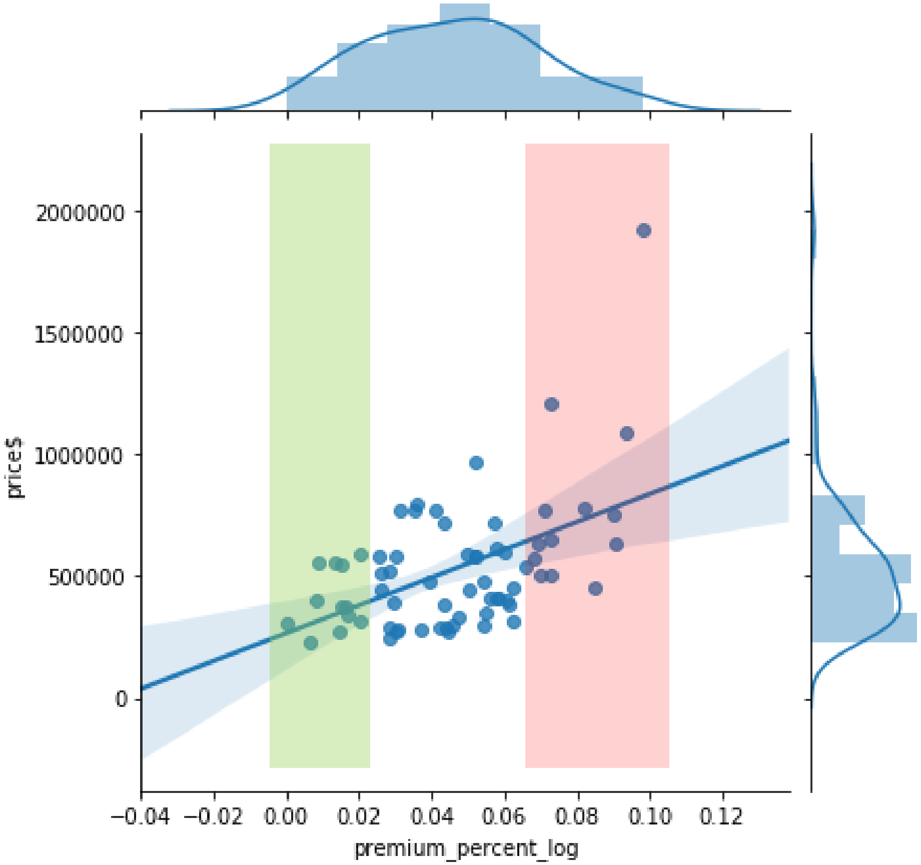
Client Prioritization Areas

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

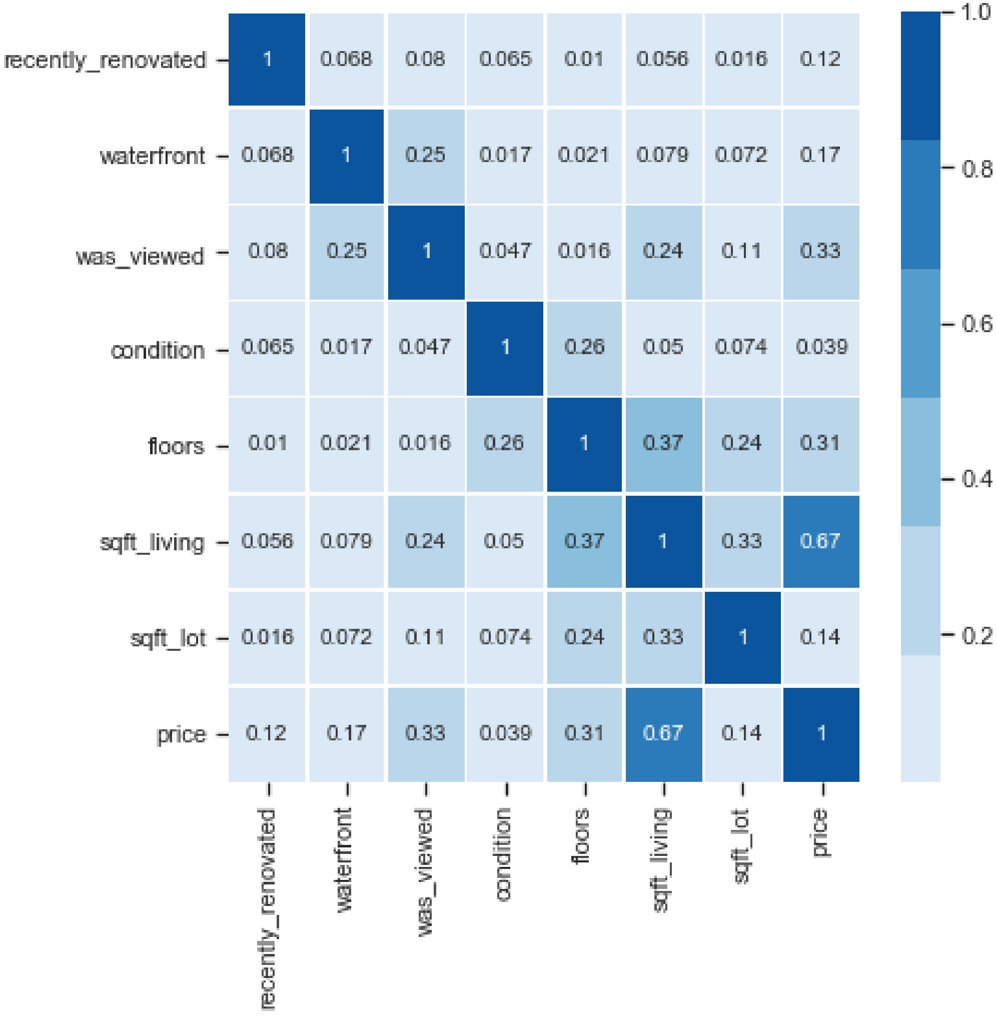
High-Priority	Low-Priority	Medium-Priority
98103	98004	All Others
98106	98005	
98107	98006	
98108	98024	
98117	98027	
98118	98039	
98122	98040	
98126	98053	
98133	98065	
98146	98070	
98148	98072	
98168	98074	
	98075	
	98077	

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

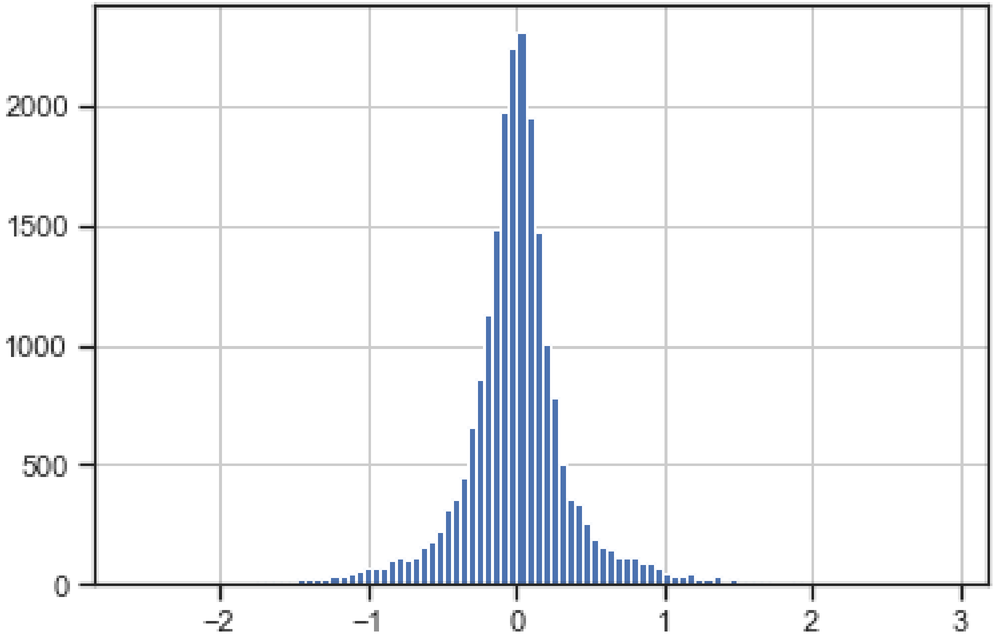
Correlation Heatmap



	recently_renovated	waterfront	was_viewed	condition	floors	sqft_living	sqft_lot	price
recently_renovated	1	0.068	0.08	0.065	0.01	0.056	0.016	0.12
waterfront	0.068	1	0.25	0.017	0.021	0.079	0.072	0.17
was_viewed	0.08	0.25	1	0.047	0.016	0.24	0.11	0.33
condition	0.065	0.017	0.047	1	0.26	0.05	0.074	0.039
floors	0.01	0.021	0.016	0.26	1	0.37	0.24	0.31
sqft_living	0.056	0.079	0.24	0.05	0.37	1	0.33	0.67
sqft_lot	0.016	0.072	0.11	0.074	0.24	0.33	1	0.14
price	0.12	0.17	0.33	0.039	0.31	0.67	0.14	1

Tests for OLS Validity - Main Regression

Regression residual graph does not demonstrate any marked heteroskedasticity…



Histogram of Regression Residuals
(Home Predictors & Zip Codes)

Count of Observations

Difference Between Observed ln(Price) and Predicted ln(Price)

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

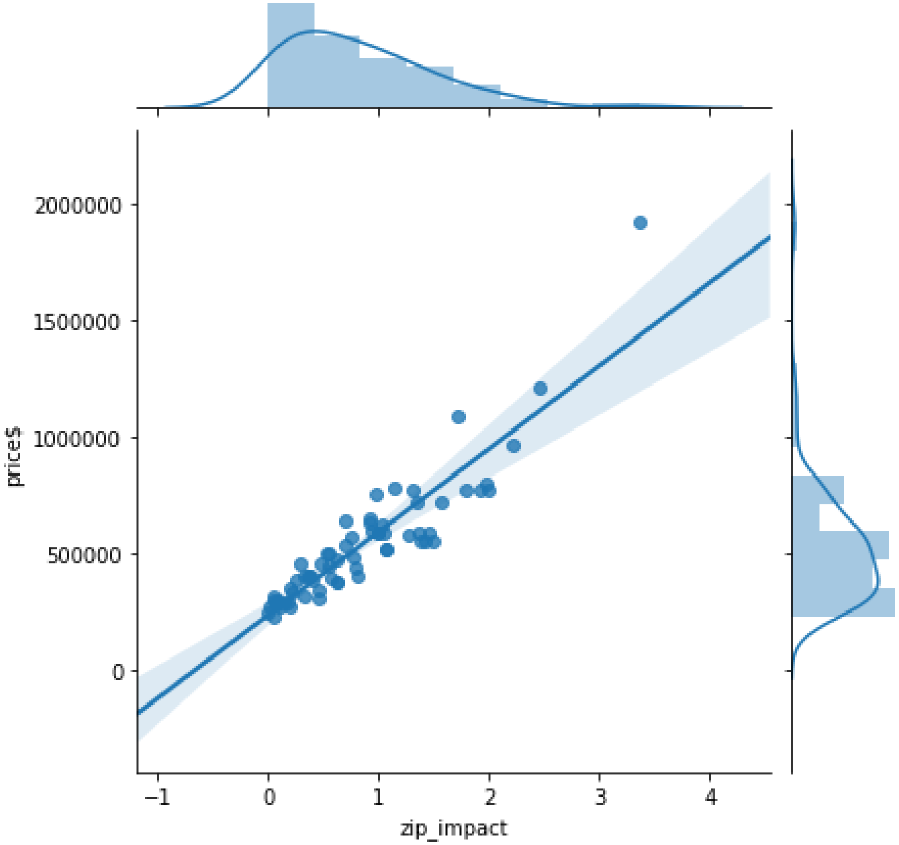
Log-Transformed Outcome Detail

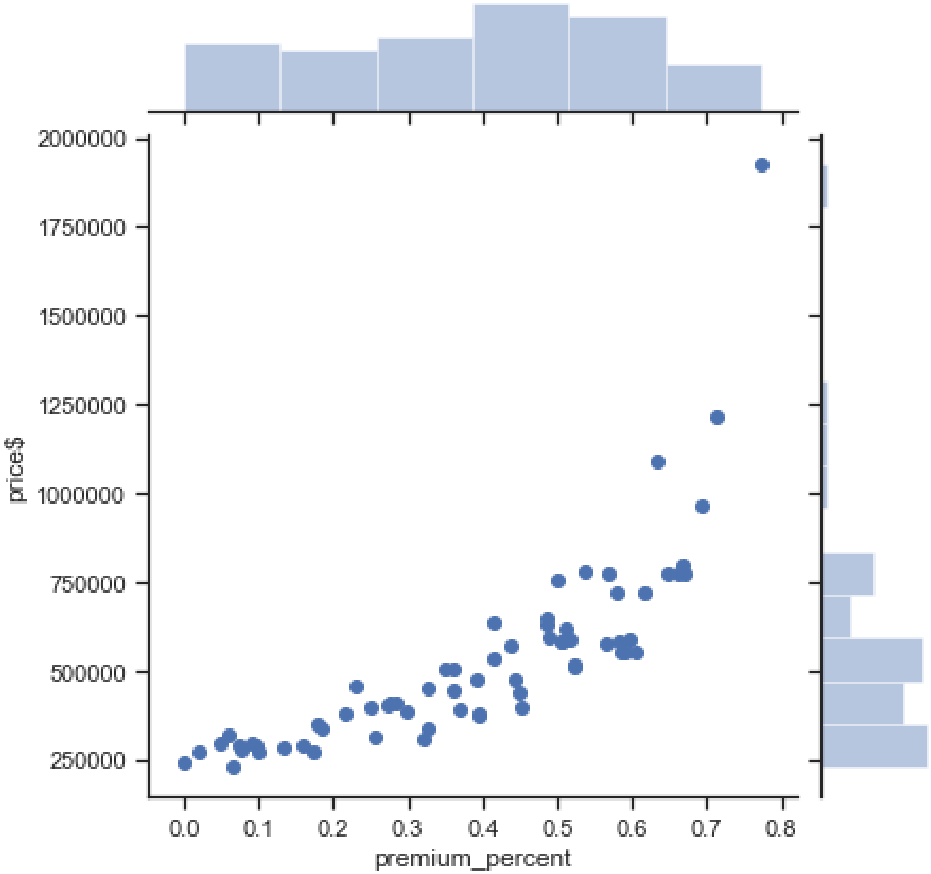
Econometric approach: exp(coeff)-1

- 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


price$
zip_impact


price$
premium_percent