

# Variation in Housing Price in King County by Zip Code, Washington State

John Kline

10/28/19

# Framework & Data Sources

---

## GOALS

*Evaluate the impact of zip code on the expected sale price of residential housing units – controlling for indicators of house quality*

## 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>)*

# Findings - Summary

***Housing prices are highly stratified by zip code – using zip code as a factor in the analysis improves the percent of housing price variance explained from 55% to 86%***

**Without Zips**

**55%**  
*of price variance*

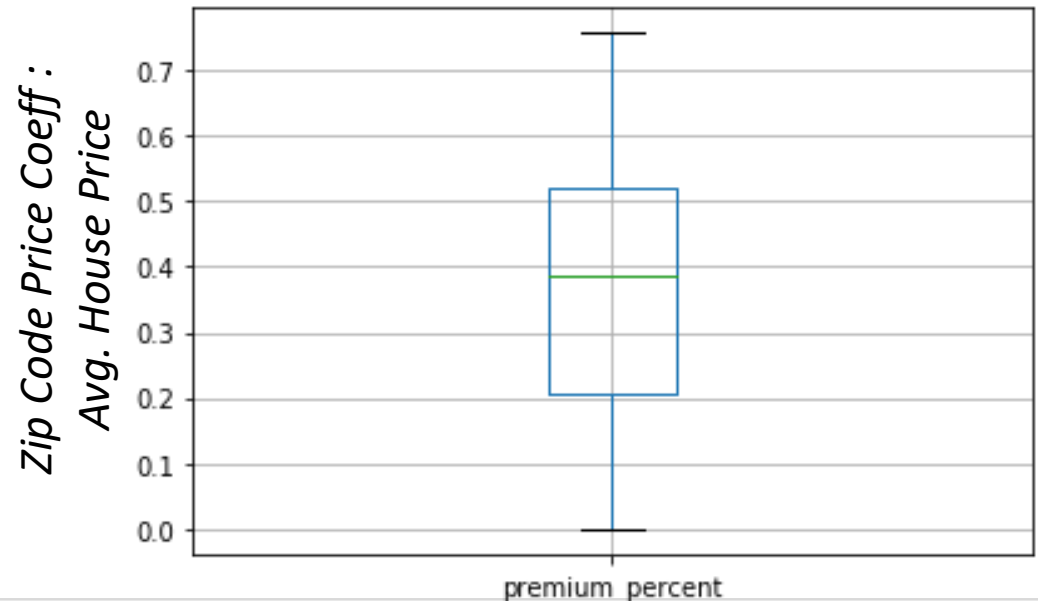


**With Zips**

**86%**  
*of price variance*

***Zip code explained a high percentage (37%) of the total house price on their own: this factor explained values independently from such core features as the house square footage, relative size to near neighbors, view, rated condition, bedrooms, floors, lot size, whether it was viewed, whether it was on the water, or renovation status***

***Boxplot Distribution of Zip Codes by Ratio of Zip Code Premium to Avg. House Price***



**John Kline**

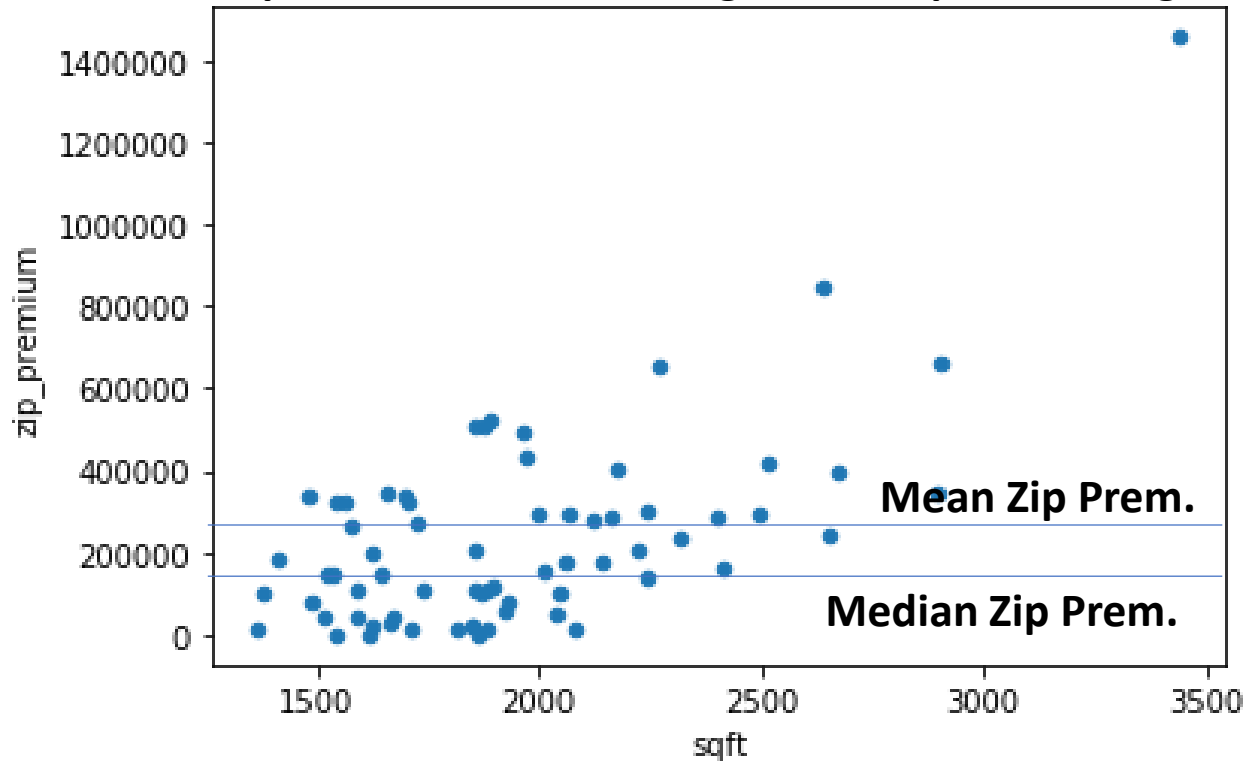
Non-Technical Presentation

Module 1 - Flatiron School - Data Science Program

# Findings - Detailed

*Zip Code Premium, or the portion of housing price explained by the zip code of the house instead of any other observed factor, is a powerful for identifying house value between neighborhoods that might otherwise seem very similar on the basis of house size (the second-most important explanatory price variable).*

**Zip Code Premium vs. Avg. Home Square Footage**



*Zip Code Premium is not distributed linearly – the most exclusive zip codes demand disproportionately higher premiums than the average.*

- The mean zip premium is significantly higher than the median
- The highest zip premium is almost an entire interquartile range above the next closest value

# Recommendations & Comments

---

- This analysis ratifies a cliché realtor saying, “Location, location, location”, but there are not a lot of consumable, rigorous analytic support tools on online realty websites like Zillow –grouped geographic analyses like may be able to help homeowners make better, faster decisions on the tradeoff of housing price, premiums paid, and actual important factors (e.g. school, distance to work, ec.)
- The most highly desirable zip codes were disproportionately expensive – however this analysis was limited in being able to determine the extent to which these reflect unexamined, but quantitative factors like school quality, or social constructions like neighborhood prestige

# Next Analytic Steps

---

- 1. Comparison analysis of different ways of grouping houses by geography – do results persist?**
  - 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

# Discussion & Questions

---

Thanks!

# Appendix

---



# Methods Discussion

---

## Model

- *Sklearn package: .ols() method*

## Key Transformations

- *The dependent variable in this analysis was a natural logarithmic transformation of the sale price in \$USD*
- *Independent variables with continuous values were sometimes log transformed*
  - *Sqft\_living*
  - *Sqft\_lot*
  - *Joneses\_living*

# Tests for OLS Validity

- Correlation among variables chosen for the analysis remained low – with a max of 0.59 in magnitude

