Module 1 Project

Home Price Factors in King County

Matthew Onstott

Table of Contents

Project Overview

Goal

- Determine if and how features affect prices
- Predict future housing prices

Dataset

- House sales in King County, Washington, from May 2014 to May 2015
- 21,597 records with date, price and feature information

Method

- Multivariate Linear Regression

Features

- **Count**
  - Bedrooms
  - Bathrooms
  - Floors
  - Views
- **Square Footage**
  - Home
  - Lot
  - Nearest 15 Neighbors' Homes
  - Nearest 15 neighbors' Lots
- **Binary**
  - Basement
  - Waterfront
- **Scale**
  - Overall Condition
  - King County Grade
- **Date**
  - Year Built
  - Year Renovated
- **Location**
  - ZIP Code
  - Latitude
  - Longitude

Process

Methodology

OSEMN Framework

- **Obtain**
  - Gather data from sources
- **Scrub**
  - Clean and pre-process
- **Explore**
  - Find patterns and trends
- **Model**
  - Predict and forecast
- **Interpret**
  - Review results and next steps

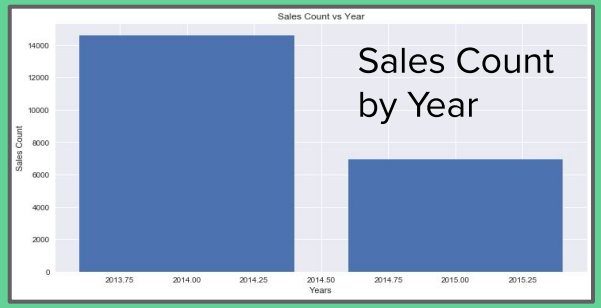
Questions

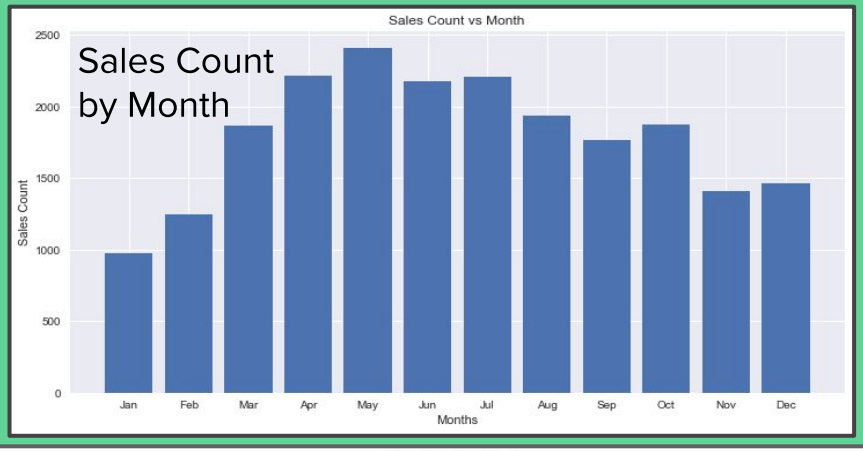
1. How does time affect price?
2. How does location affect price?
3. How does construction date affect price?
4. What factors best predict price?

Time Dependency

Recommendation:

- Buy in winter
- Sell in summer





Year

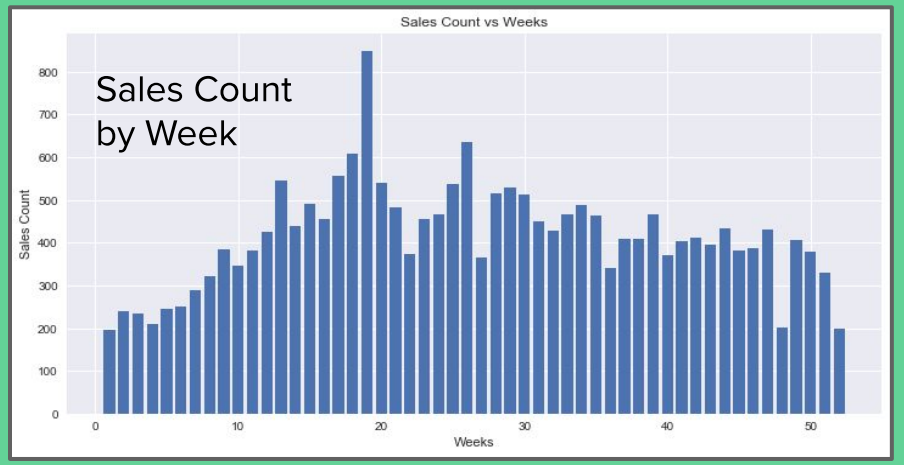
- There is not enough data to comment reliably on year-long patterns

Month

- Winter months have the fewest sales (avg. **1,250**)
- Late spring and early summer months have the most sales (APR thru JUL above **2,000**)

Week

- Certain weeks have very large or small sales (start of May above **800** vs start of January at **200)**



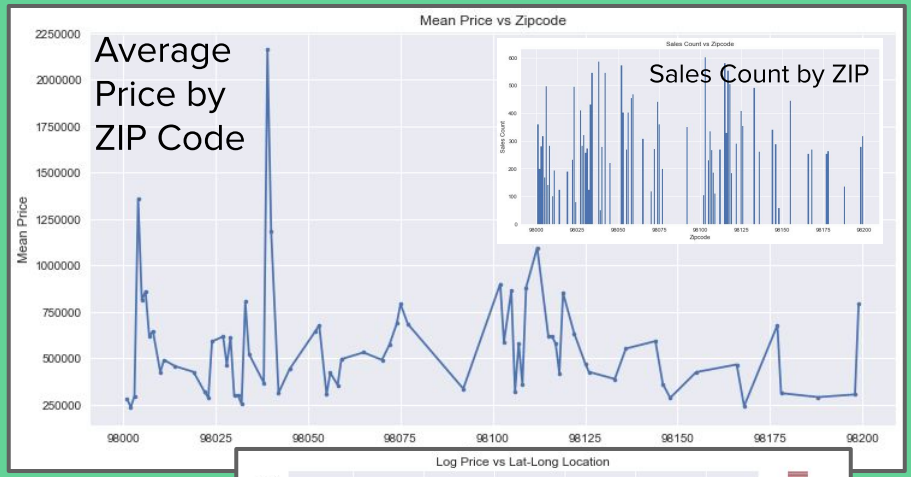
Location Dependency

Recommendation:

- Invest above lat 47.5

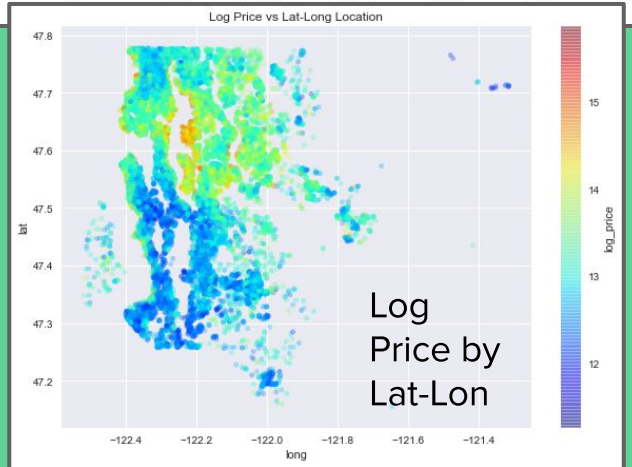
ZIP Code

- sales vary greatly by ZIP Code
- The largest has an average sales price is nearly **$2.25 million**
- For comparison, the second highest ZIP Code has a mean sale price of just over **$1.25 million**
- Beyond this, price are well below **$1 million**



Latitude

- There is a clear separation of high and low priced homes at **47.5** (expensive homes at higher latitudes)



Longitude

- The most expensive homes are at central latitude values (**-122.2**) while cheaper homes are found at the edges (**±0.2**)

Construction Dependency

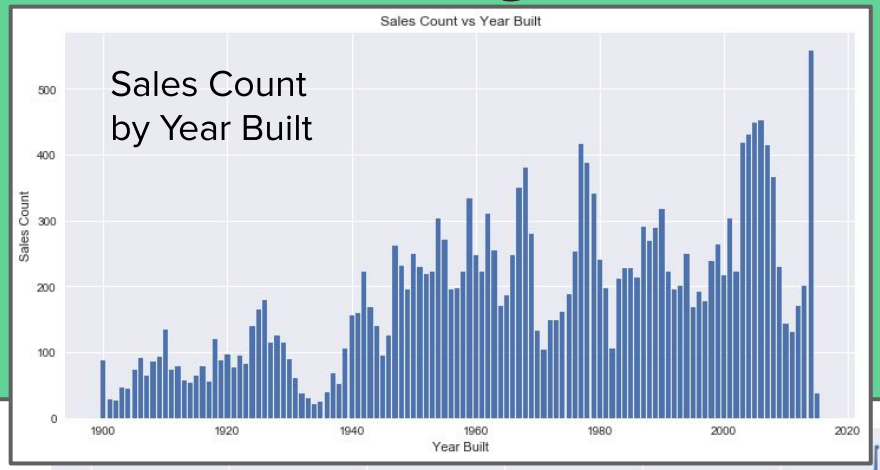
Recommendation:
- Stick to new units

Year Built

- There is a very strong relationship with price
- The later a house is built, the more expensive it is to purchase (2015 above **$500k**, 1971 at **$100k**)

```
Number of Years Built in Dataset: 116

Year Built Statistics:
count    21534
mean      1971
std         29
min       1900
25%       1951
50%       1975
75%       1997
max       2015
```



Iterative Modeling

Lessons Learned

- Reducing the number of predictors increases quality of fit for a small cost in explained variance
- Log transformation of price satisfies normality (skew, kurtosis)
- Overfitting is not an issue in the final model (Accuracy, RMSE)

#	Model	R^2	Adjusted R^2	F Statistic	Predictors
1	Baseline	0.699	0.698	1426	33
2	Only Significant Predictors	0.699	0.698	2137	22
3	High t-Statistics	0.621	0.621	4758	7
4	Log Price	0.718	0.718	7385	7
5	Final Model Training Set	0.720	0.719	5952	7

Final Model	Accuracy	RMSE	% of Data
Training Set	72.0%	0.279	80%
Testing Set	71.2%	0.281	20%

Price Factors

Influencers

- The final model has 7 coefficients built from 5 predictors that impact housing sales price
- **Grade**
  - Expensive homes are given higher ratings by King County
- **Latitude**
  - Location is critical to determining the final price
- **View**
  - The number of views for a house is related to its perceived value
- **Waterfront**
  - Homes on the water are highly coveted
- **Year Built**
  - Newer homes are more costly and must be sold at higher prices

Final Model

$$[log\_price] = 11.45 + 0.62[lat75] + 2.66[grade\_s] + 0.44[lat100] + 0.30[lat50] + 0.35[waterfront] - 0.06[yrbuilt\_75] + 0.43[view\_s]$$

Definitions

- **log_price** = log(**price**)
- **lat75** = 1 for **lat** > 47.57 and **lat** <= 47.68, 0 otherwise
- **grade_s** = (**grade**(i) - min(**grade**)) / (max(**grade**) - min(**grade**)) where min(**grade**) = 3 and max(**grade**) = 13
- **lat100** = 1 for **lat** >= 47.68, 0 otherwise
- **lat50** = 1 for **lat** > 47.47 and **lat** <= 47.57, 0 otherwise
- **waterfront** = 1 for houses with **waterfront** view, 0 otherwise
- **yrbuilt_75** = 1 for **yr_built** > 1975 and **yr_built** <= 1997, 0 otherwise
- **view_s** = (**view**(i) - min(**view**)) / (max(**view**) - min(**view**)) where min(**view**) = 0 and max(**view**) = 4

Future Work

- Inspect price models by the number of views
- Inspect sales of low valued houses in expensive neighborhoods
- Analyze how remodeling affects the resale value of flipped houses
- Include other factors that may affect housing prices
    - demographics, education, crime, jobs, entertainment