Housing Prices in King County, Washington

—

By: Andrew, Christopher, Nathaniel, Toufik

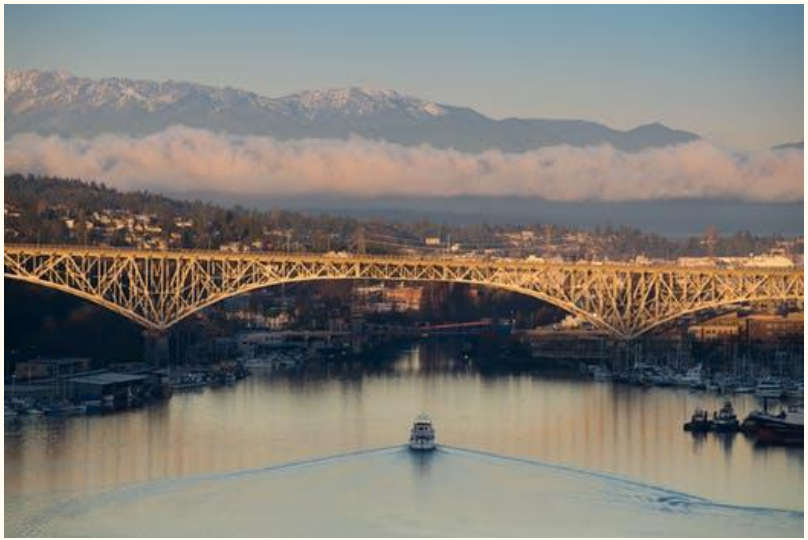
Data Source

1. Data set was obtained from Kaggle.com under the CC0: Public Domain License

2. Dataset was originally posted for a Kaggle competition

3. Includes sale prices and descriptive fields for homes sold between May 2014 and May 2015

4. Covers King County USA, the county which includes Seattle



Data Set

Available fields for the analysis

- ID
- Date (yyyy/mm/dd)
- Price ($US)
- Bedrooms
- Bathrooms
- Square Ft. of Living (Total)
- Square Ft. Lot Size
- Floors
- Waterfront (Y/N)
- Has Been Viewed (1-4)
- Condition
- Grade
- Square Ft. (Upstairs)
- Square Ft. (Basement)
- Year Built
- Year Renovated
- ZIP Code
- Latitude
- Longitude
- Square Ft. of Living in 2015
- Square Ft. of Lot in 2015

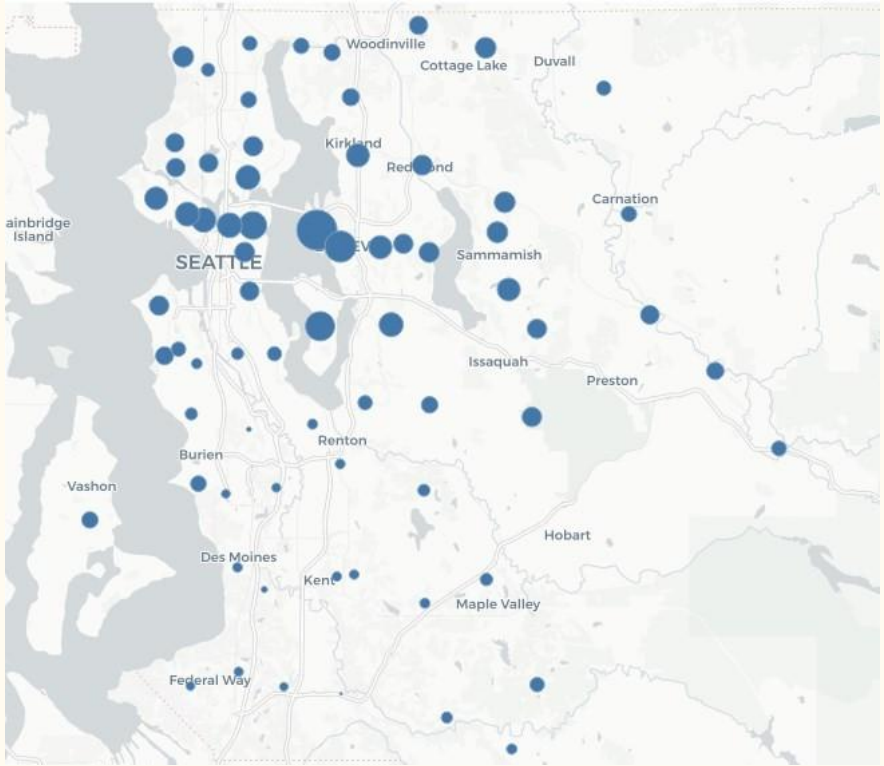
Background Research in Predicting Housing Prices

- Economic perspective (housing booms and busts)
  - Recovery phase, expansion phase, hyper supply phase and recession phase
- Current Housing Market Economy
  - Stable Economy (mortgage rates remain low)
  - Price appreciation, increase in ‘millennial’ families, high rent for building owners
  - More Canadians
  - One top forecast market is Seattle (11% price appreciation)
- Multivariate regression model
  - Environmental info versus characteristics info
  - Hedonic Pricing model/regression

SMART Question: Can we develop a model that accurately (at the 5% level) predicts housing prices in King’s County, Washington?

Exploratory Data Analysis

```
## Classes 'tbl_df', 'tbl' and 'data.frame': 21613 obs. of  21 variables:
##  $ id           : chr  "7129300520"
##  $ date         : POSIXct, format: "2014-10-13"
##  $ price        : num  221900
##  $ bedrooms     : int  3
##  $ bathrooms    : num  1
##  $ sqft_living  : int  1180
##  $ sqft_lot     : int  5650
##  $ floors       : num
##  $ waterfront   : int  0
##  $ view         : int  0
##  $ condition    : int  3
##  $ grade        : int  7
##  $ sqft_above   : int  1180
##  $ sqft_basement: int  0
##  $ yr_built     : int  1955
##  $ yr_renovated : int  0
##  $ zipcode      : int  98178
##  $ lat          : num  47.5
##  $ long         : num  -122
##  $ sqft_living15: int  1340
##  $ sqft_lot15   : int  5650
```



Mean Sale Price Map

Exploratory Data Analysis

Variables	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	Month
price	1	0.31	0.53	0.7	0.09	0.26	0.27	0.4	0.04	0.67	0.61	0.32	-0.05	-0.11	-0.01
bedrooms	0.31	1	0.52	0.58	0.03	0.18	-0.01	0.08	0.03	0.36	0.48	0.3	-0.15	-0.17	0
bathrooms	0.53	0.52	1	0.75	0.09	0.5	0.06	0.19	-0.12	0.66	0.69	0.28	-0.51	-0.54	0.01
sqft_living	0.7	0.58	0.75	1	0.17	0.35	0.1	0.28	-0.06	0.76	0.88	0.44	-0.32	-0.34	0.01
sqft_lot	0.09	0.03	0.09	0.17	1	-0.01	0.02	0.07	-0.01	0.11	0.18	0.02	-0.05	-0.05	0
floors	0.26	0.18	0.5	0.35	-0.01	1	0.02	0.03	-0.26	0.46	0.52	-0.25	-0.49	-0.51	0.01
waterfront	0.27	-0.01	0.06	0.1	0.02	0.02	1	0.4	0.02	0.08	0.07	0.08	0.03	0	0.01
view	0.4	0.08	0.19	0.28	0.07	0.03	0.4	1	0.05	0.25	0.17	0.28	0.05	0.02	-0.01
condition	0.04	0.03	-0.12	-0.06	-0.01	-0.26	0.02	0.05	1	-0.14	-0.16	0.17	0.36	0.4	0.02
grade	0.67	0.36	0.66	0.76	0.11	0.46	0.08	0.25	-0.14	1	0.76	0.17	-0.45	-0.46	0.01
sqft_above	0.61	0.48	0.69	0.88	0.18	0.52	0.07	0.17	-0.16	0.76	1	-0.05	-0.42	-0.44	0.01
sqft_basement	0.32	0.3	0.28	0.44	0.02	-0.25	0.08	0.28	0.17	0.17	-0.05	1	0.13	0.1	0.01
yr_built	-0.05	-0.15	-0.51	-0.32	-0.05	-0.49	0.03	0.05	0.36	-0.45	-0.42	0.13	1	0.91	-0.01
yr_renovated	-0.11	-0.17	-0.54	-0.34	-0.05	-0.51	0	0.02	0.4	-0.46	-0.44	0.1	0.91	1	-0.01
Month	-0.01	0	0.01	0.01	0	0.01	0.01	-0.01	0.02	0.01	0.01	0.01	-0.01	-0.01	1

Procedure

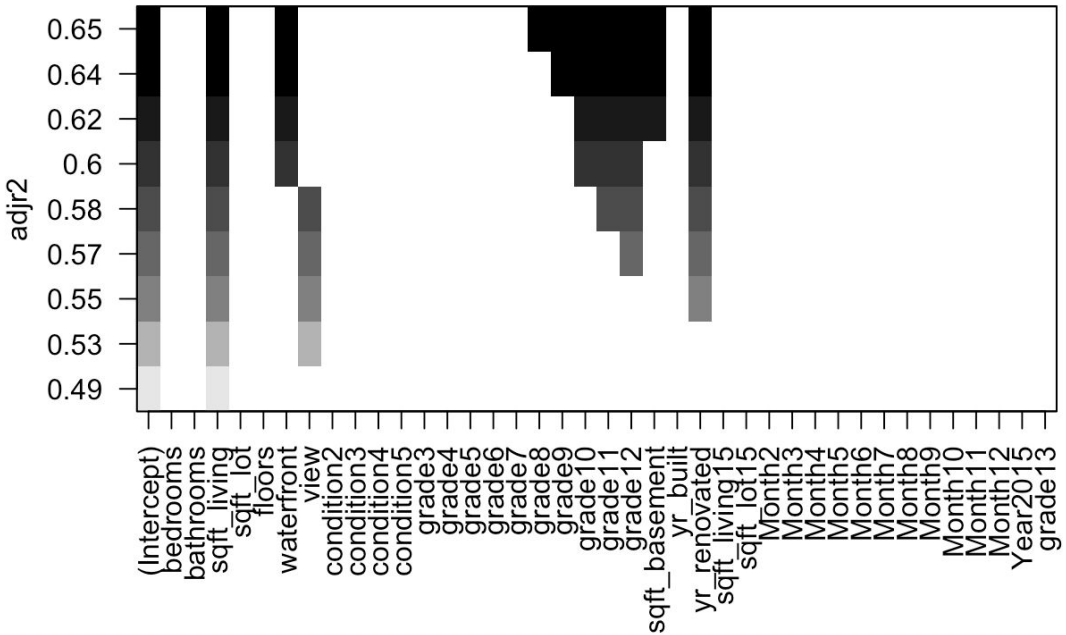
1. Random sample the dataset into a training set and a test set
2. Using Regsubsets method within the training set, select which factors have the best predictive power for home price.
3. Take the coefficients of the best model and assess its predictive power within the test set using a t-test.

Best Fit Model

Sequential Replacement Selection Method:

Factors - sqft_living, waterfront,grade,sqft_basement,and yr_renovated

Adjusted R squared at **.65**



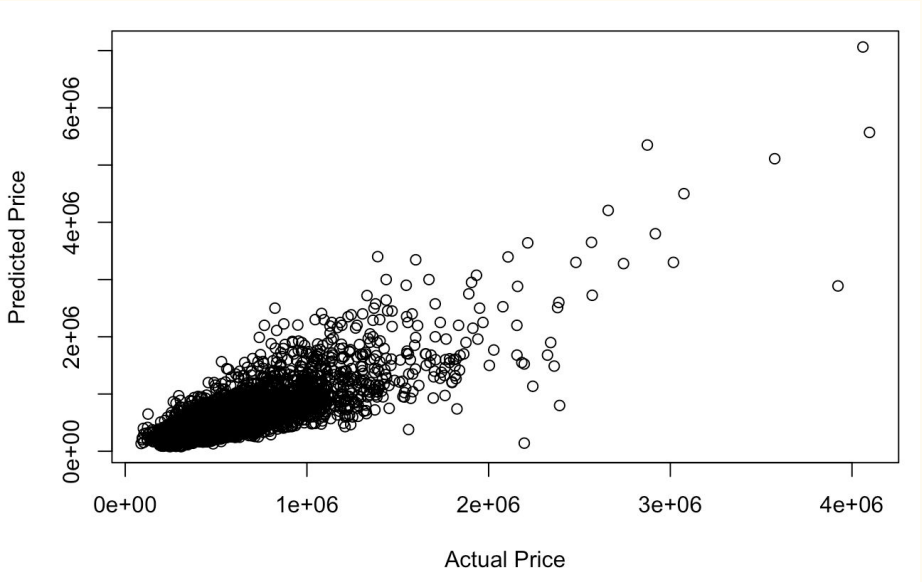
Best Fit Model - King County Sale Prices

Model Results

Final Model:

Price ~ sqft_living + sqft_living^2 + waterfront + grade + sqft_basement + sqft_basement^2 + yr_renovated

	Estimate	Std. Error	t value	Pr(>|t|)
poly(sqft_living, 2)1	15417850	400467	38.5	9.881e-309
poly(sqft_living, 2)2	5574465	302081	18.45	3.588e-75
waterfront	704314	20184	34.89	3.702e-256
grade3	-2046176	144100	-14.2	1.864e-45
grade4	-2020904	92185	-21.92	8.189e-105
grade5	-2055393	77418	-26.55	1.14e-151
grade6	-1996664	75794	-26.34	2.078e-149
grade7	-1891599	75681	-24.99	5.236e-135
grade8	-1763632	75632	-23.32	4.382e-118
grade9	-1610515	75323	-21.38	7.064e-100
grade10	-1426587	74761	-19.08	3.473e-80
grade11	-1209607	74172	-16.31	2.93e-59
grade12	-845735	75389	-11.22	4.379e-29
poly(sqft_basement, 2)1	2245974	260932	8.608	8.232e-18
poly(sqft_basement, 2)2	-1502258	264888	-5.671	1.444e-08
yr_renovated	2517	73.67	34.17	4.721e-246
(Intercept)	2220710	75405	29.45	3.66e-185



Correlation between predicted price and actual price = **0.82;** T-test statistics: (t=-1.503, df=7204, **p-value=.133**)

Conclusions

1. Housing sale prices can be predicted using intrinsic data points of the property
2. ZIP code needs to be removed from any generalized linear regression model
3. A second order model helps to reduce some non-linearity within the model