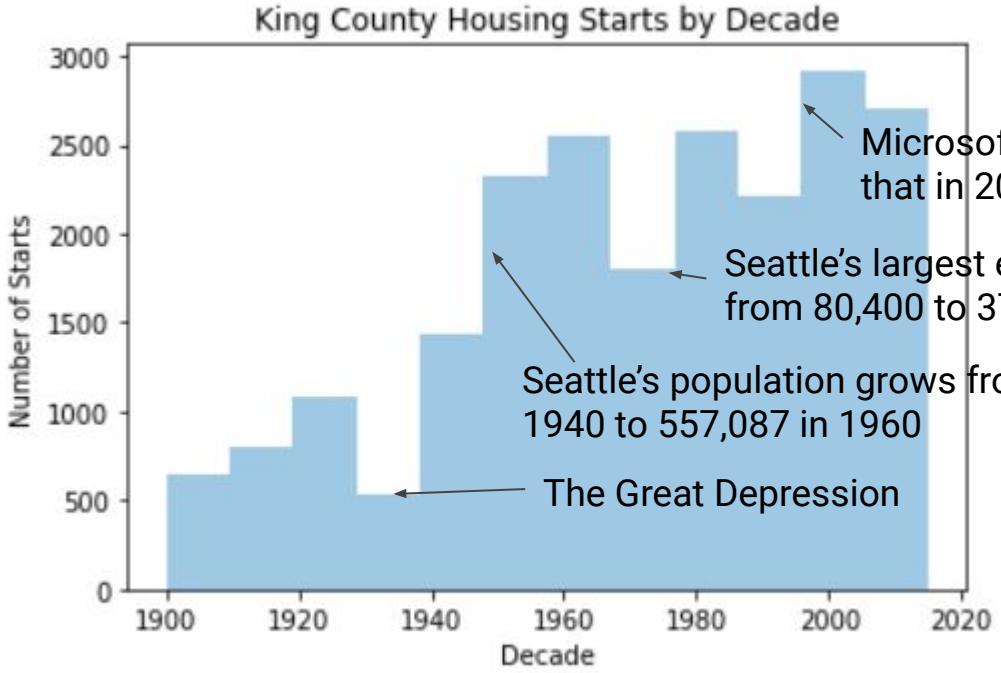
King County Housing Data

May 10, 2019

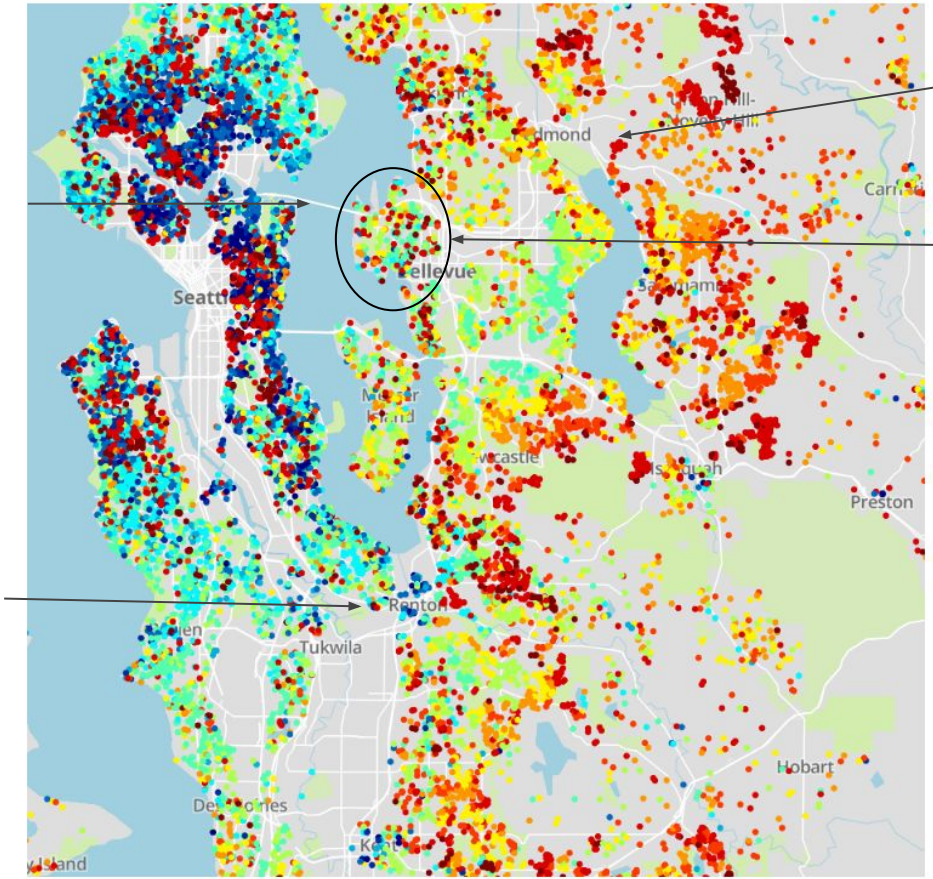
Linh Pham
Helen Levy-Myers

Housing Data is Important to Communities



Final bar represents data until early 2015, not a complete decade

Housing Starts by Decade in King County



Causeway opens 1969.

Large Boeing manufacturing plant in Renton

Microsoft headquarters are in Redmond

Two most expensive zip codes here.

Project objective:

To create a linear model that can predict the home sale price and make suggestions about how to improve home value for future sales.

The Data Set

Data Source

- 21,597 Home Sales in King County, Washington, which includes Seattle.
- All sales between May 2014 and May 2015

Organizing The Data

Four Categories

All four categories included Price and ID.

- Size of Home
  - Bedrooms
  - Bathrooms
  - Floors
  - Living space (sq. ft.)
  - Lot size (sq. ft.)
  - Basement (sq. ft.)
  - Above Grade (sq. ft)

- Neighborhood
  - Zip code
  - Latitude
  - Longitude
  - Living space of nearest 15 neighbors (sq. ft.)
  - Lot size of nearest 15 neighbors (sq. ft.)

- Building Status
  - Condition
  - Grade
  - Year Built
  - Year Renovated

- Other
  - Waterfront
  - Number of beautiful views
  - Date Sold

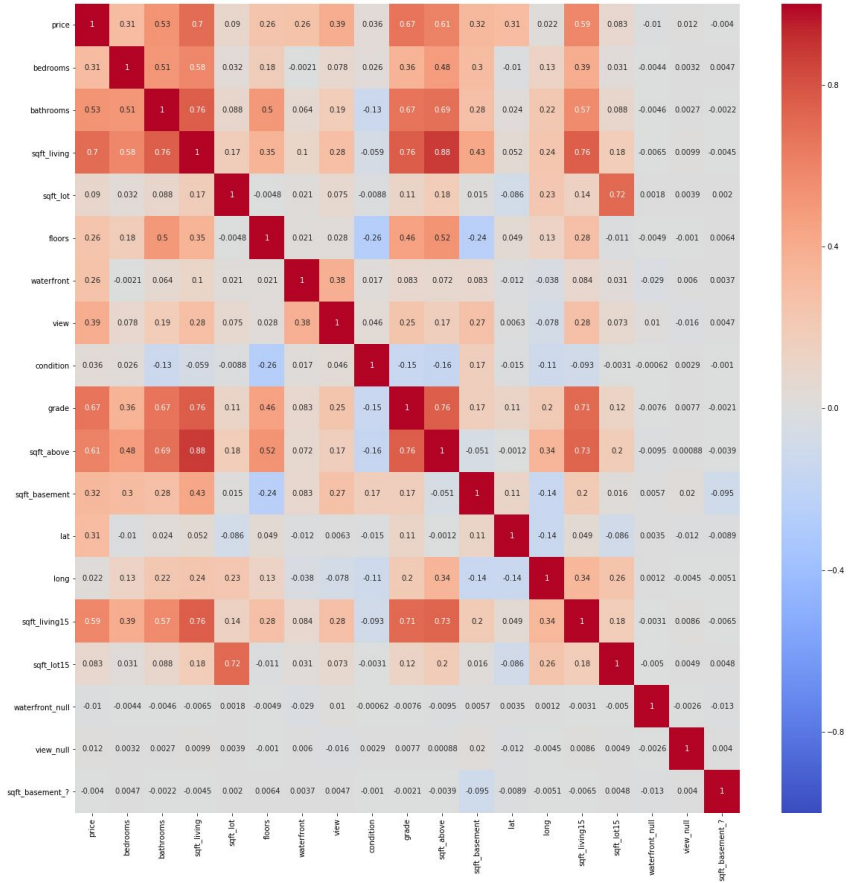
"This is not what I meant when I said 'we need better data cleansing!'"

Data Cleaning

- Created new ‘Null’ value columns for Waterfront, Year Renovated and Views
- Created new ‘Question Mark’ column for Basement
- Binned Year Built and Year Renovated by Decade
- Converted Zip Codes to categorical data
- Looked at three new variables

New Variables

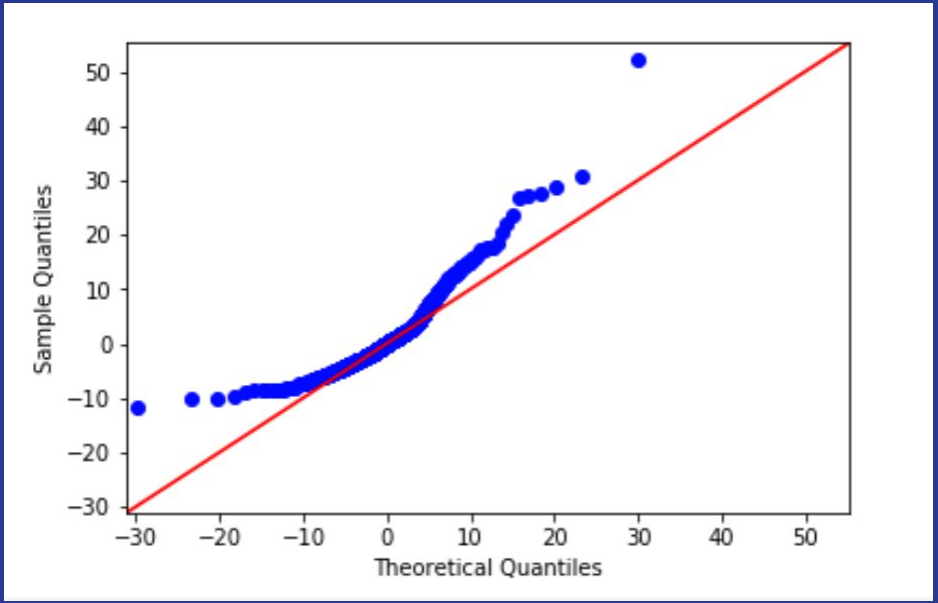
- Living Space Ratio - Living Space/Living Space 15
- Lot Size Ratio - Lot Size/Lot Size 15
- Month variable from Date



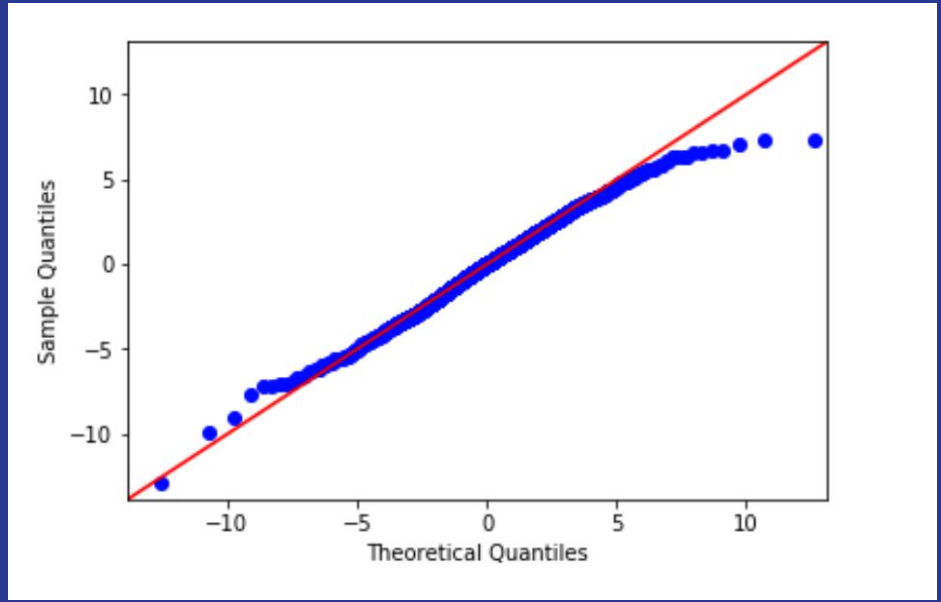
- Correlation heat map to easily discover which variables have high collinearity.
- Colored values show high multicollinearity between the variables.

Residual Plots

Model Before Transformation
Price is Regressed on the Variables



Model After Transformation
Price is Transformed Using Logarithm



The Models We Examined

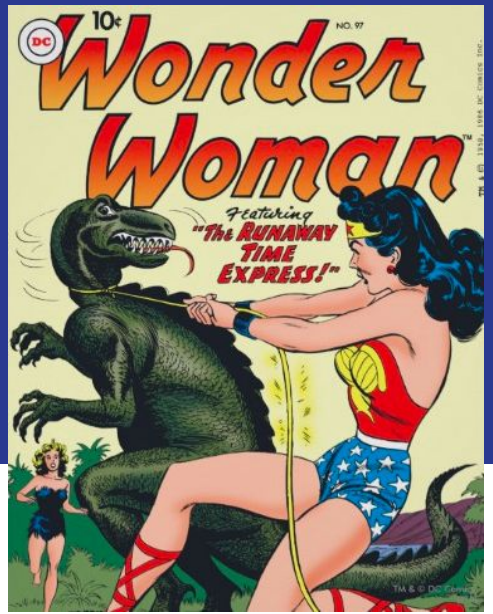
	Our model	Guess 1	Guess 2
Under predict (%)	-13	-35	28
Over predict (%)	15	48	67

- Our Model - includes 83 variables and tries to predict the log of price
- Guess 1 - Mean of the Log Price
- Guess 2 - Mean of the Price

Variables that Contribute to Home Sales Price

Variables	Increase in Percentage	Increase in Dollars Compared to Avergage Price
Zipcode 98004	44.93	$242,754.99
Living Space Ratio (sqft)	-7.61	-$41,116.53
Renovated 2006-2015	17.16	$92,714.79
Bathrooms	3.18	$17,181.41
Grade	10.40	$56,190.78
View	6.87	$37,118.34

The Team



Linh Pham

Helen Levy-Myers