King County House Pricing Predictive Analysis

By Supermodels RealState GmbH
Henrike Sahnwaldt, Enrique Hernani Ros & Rand Abu Ajamia

Our Innovative Solution

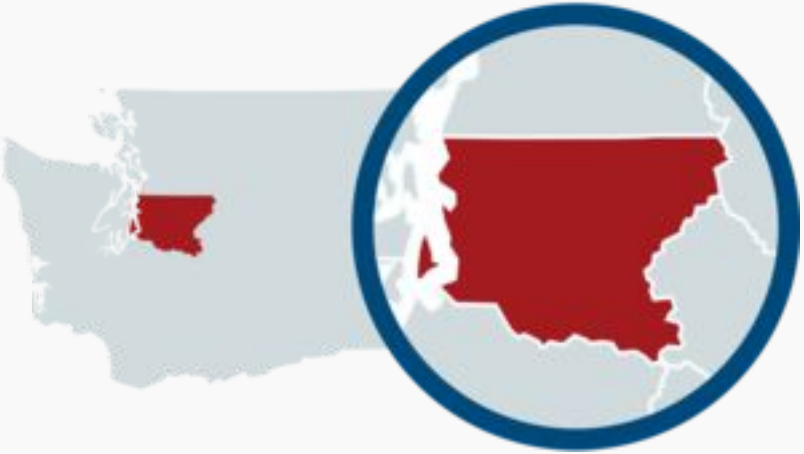
Objective:
predictive machine learning model

Our solution:

- Predicts price of properties with **90%** accuracy
- Engineered with unique “hotspot” feature to include distance from any location
- Tailored to incorporate new data

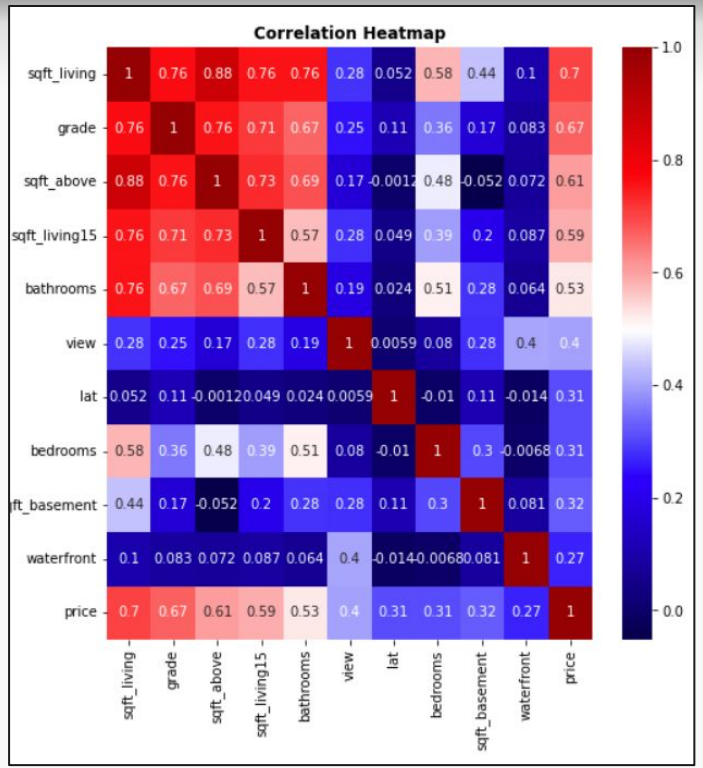
Data Set Description

Source	Kaggle
Time Period	May 2014 - May 2015
Location	King County, Washington, USA
Amount of Data	21,597 Records 20 Features
Data Quality	High
Clean Data	No null values, typos, incorrect data types or duplicate records



What is driving the price according to the data?

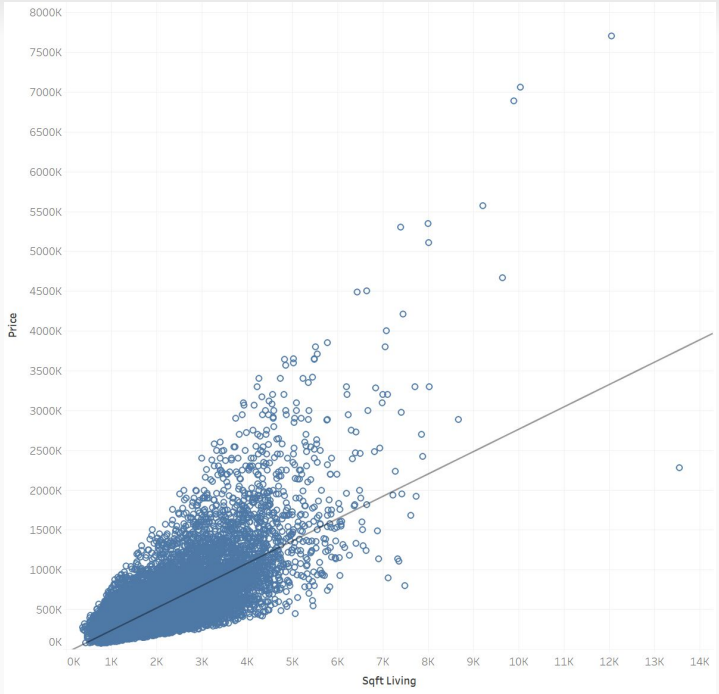
Correlation Heatmap of the 10 most relevant estimators in terms of correlation coefficient



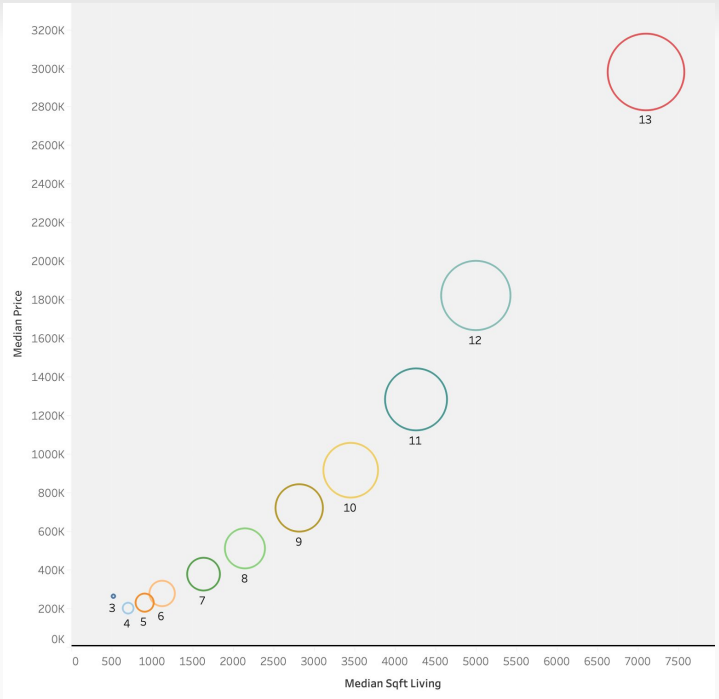
High correlation with living area and building condition	Further explore these estimators
Low correlation with the location data	How to treat the location data to improve our model?
House area estimators are highly intercorrelated	Redundant information. Drop all of them except sqft_living

Getting to know our main variables

Price vs ***living area*** (sqft)



Grade (1-13) by price & ***living area*** (sqft)

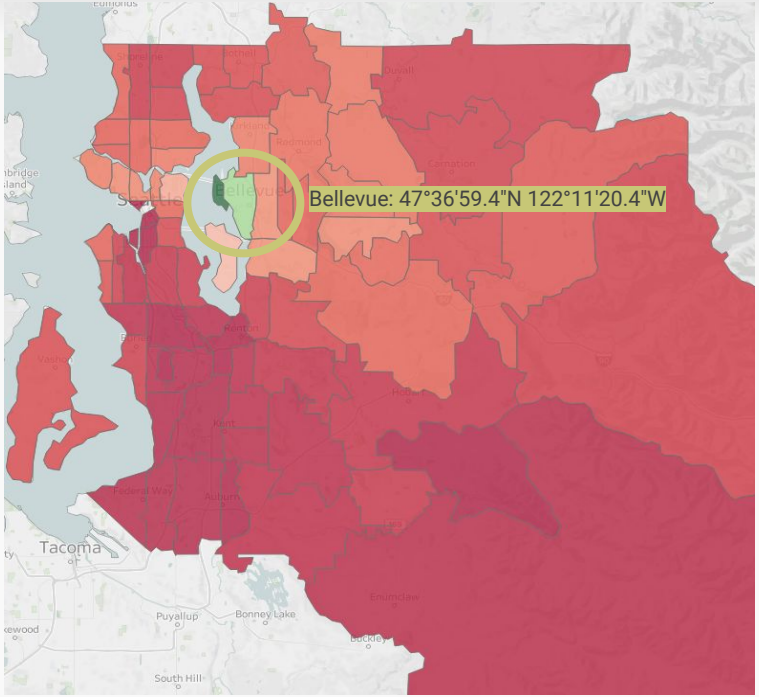


Location Feature Engineering

R2 ▲ +6%

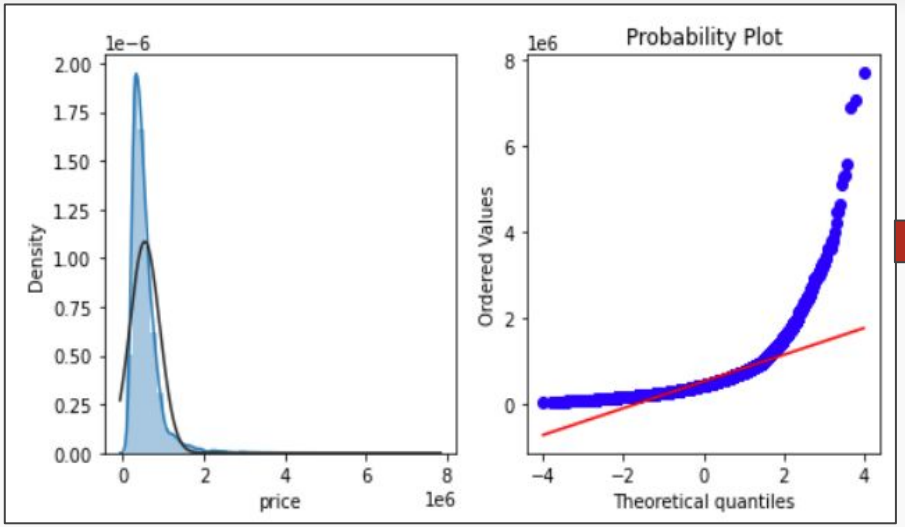
- Problem: no correlation between price & ZIP 😢
- Solution: engineer a new feature 😍
- Compute distance to price “hot spot” Bellevue, using ***Haversine*** formula

Feature	Correlation w. price
ZIP, longitude	~ 0.05
latitude	0.31
distance to Bellevue	**0.42**

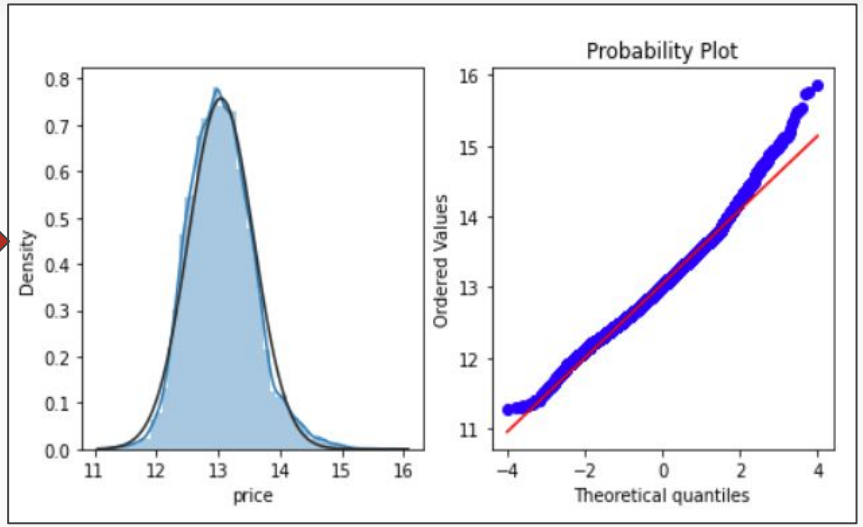


Normalization Transform

R2 ▲ +4%



LN



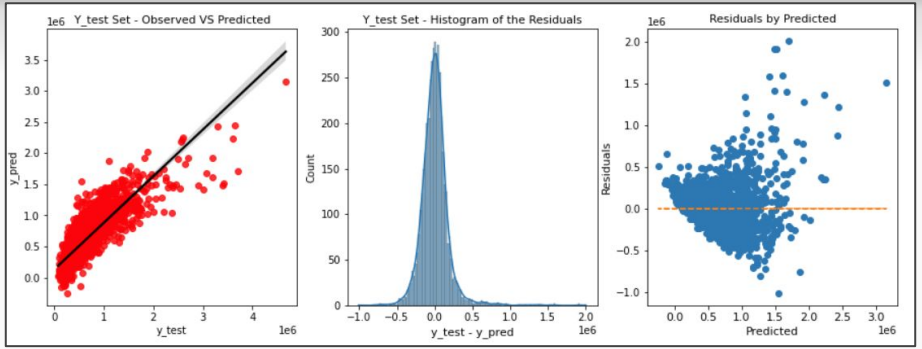
Original Price Data

Transformed Price Data (Logarithmic)

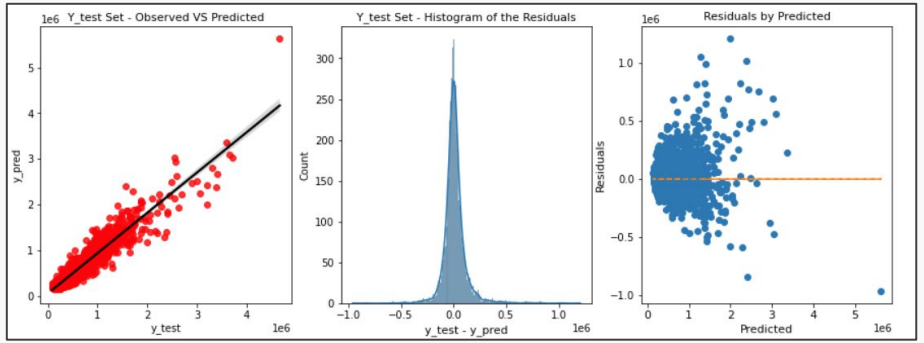
Improving the model

Regression Method	N of Predictors	Transf.?	Train Set		Test Set	
			RMSE ($)	Adjusted R2	RMSE ($)	Adjusted R2
Linear (Baseline)	20	No	204,389	0.70	187,189	**0.71**
Linear Transformed	12	Yes	168,308	0.79	156,397	0.81
Ridge Regression (α = 1)	12	Yes	168,316	0.79	156,348	0.81
Polynomial (n=2)	12	Yes	135,866	0.87	129,882	0.86
Extreme Gradient Boosting (Final)	12	Yes	98,718	0.93	108,025	**0.90**

Baseline Model



Extreme Gradient Boosting



Outro

Achievement

- Predictive model with 90% accuracy has been developed
- With the right transformations, location data has significant impact on price prediction

Next Steps

- Further feature engineering: include distance to next metro station, supermarket, school…
- Develop an interactive tool for house price prediction

Thank you

BackUp Slides

Prepare the data for modelling

- Data Cleaning - the quality of the data was very high. No cleaning was required
- Data Processing for Modelling - the following data transformations were carried out in order to improve the data for modelling purposes
  - Dropping features highly correlated between them or with very low correlation with the sale price
  - Treatment of the location data
  - Better capture the information regarding building age / renovation year
  - Outliers treatment
  - Normalization of the distribution of important features
  - Standard Scaling of Numerical Features

Features Selection

Features Dropped

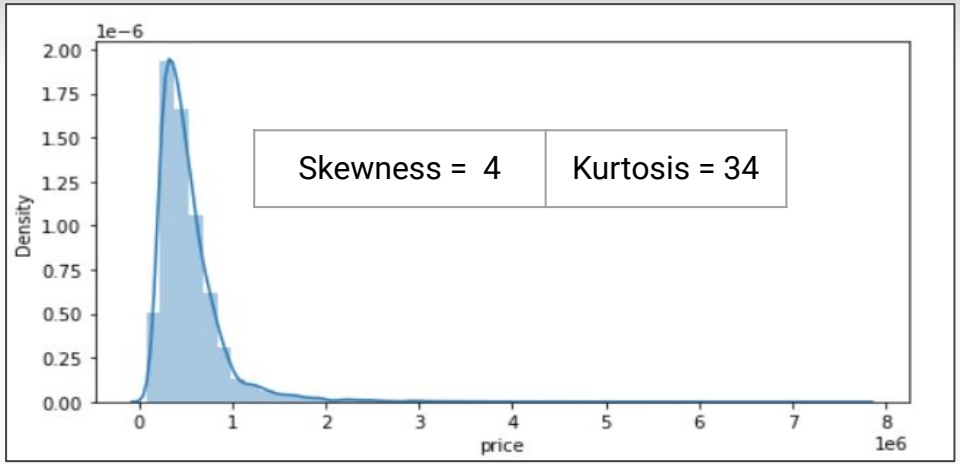
Feature	Reason to drop it
Condition	Building Condition information already contained in Grade
Floor	Hypothesis testing shows a p-value >0.05
yr_renovated	
sqft_lot	
sqft_above	
sqft_basement	
sqft_living15	
sqft_lot15	

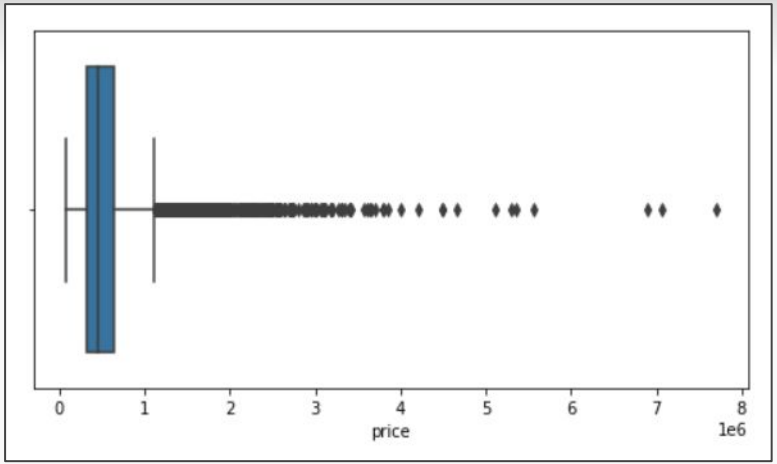
What about Outliers?

- *'bedrooms'* has one outlier value: 33 bedrooms in a house with 2 bathrooms
- Conclusion: obviously inconsistent data, delete row

- Target variable *'price'* has large outliers (houses between 2 - 7.7 million $)
- But they correspond to outliers in independent features (sqft_living, rooms)
- Outliers capture valuable information and variability of our data
- Conclusion: Keep outliers to ensure realistic predictions

Getting to know our target - Sale Price





The sale price deviates from a normal distribution with:

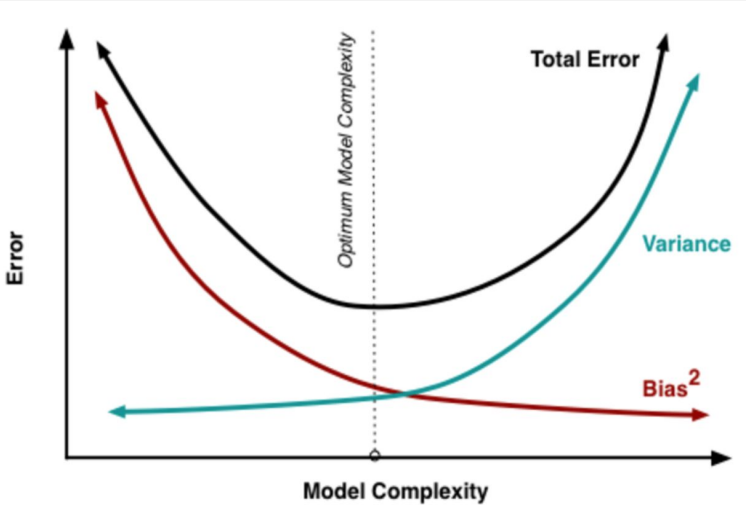
- Positive Skewness - large right tail
- Very High Kurtosis - very high number of extreme values. Are they really outliers?

Standard Deviation	367K $
Mean	540K $
Max	7.7M $

Regularization (Ridge/Lasso Regression)

R2 —

- In every regression there is a trade-off betw
- Regularization algorithms (for example Ridg reduce the magnitude of the fitting coefficie complexity of the model and the variance
- No improvement in the adjusted R2 coefficie or Lasso Regressions ➔ Our model is not o

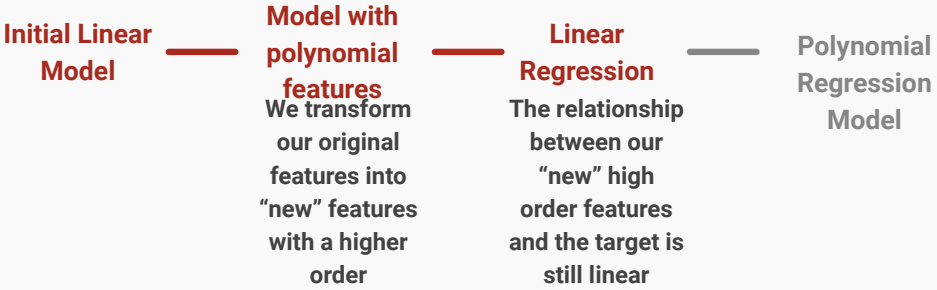


Source: researchgate.net

R2 ▲ +5.8%

Polynomial Regression

- The target is to increase the complexity of the model by increasing the order of the fitting coefficients
- Risk of overfitting if the order is increased too much

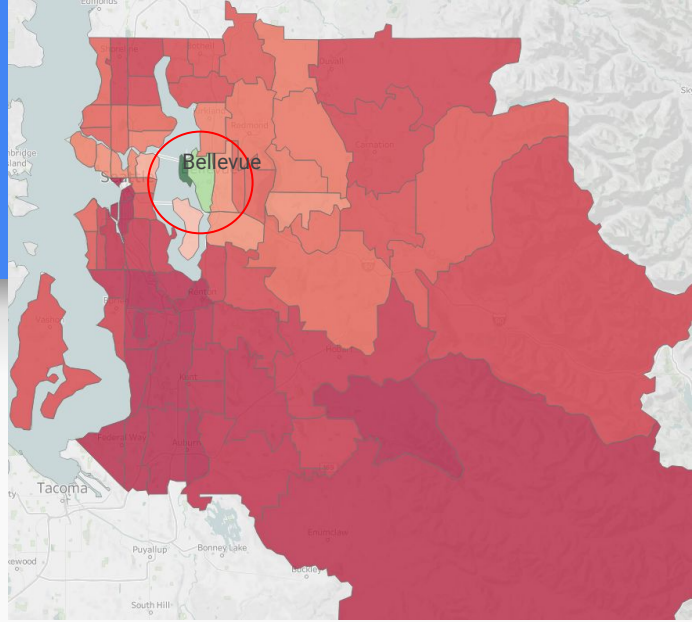


R2 ▲ +4.3%

Extreme Gradient Boosting

- Machine learning technique for regression and classification problems based on ensembling simpler models, normally decision trees
- Implemented in Python with XGBoost library

Location Feature Engineering



- Problem: no correlation between price & ZIP :(
- Solution: engineer a new feature! :)
- Compute distance to price “hot spot”
  using ***Haversine*** formula:

house GPS data -> Bellevue GPS data -> distance in km

	lat	long
0	47.5112	-122.257
1	47.7210	-122.319
2	47.7379	-122.233
3	47.5208	-122.393
4	47.6168	-122.045
...	...	...

0.31 0.02

47.616492, -122.188985

$a = \sin^2(\Delta latDifference/2) + \cos(lat1).\cos(lt2).\sin^2(\Delta lonDifference/2)$

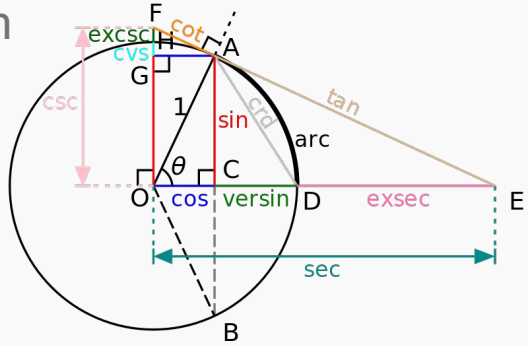
$c = 2.\text{atan2}(\sqrt{a}, \sqrt{(1-a)})$

$d = R.c$

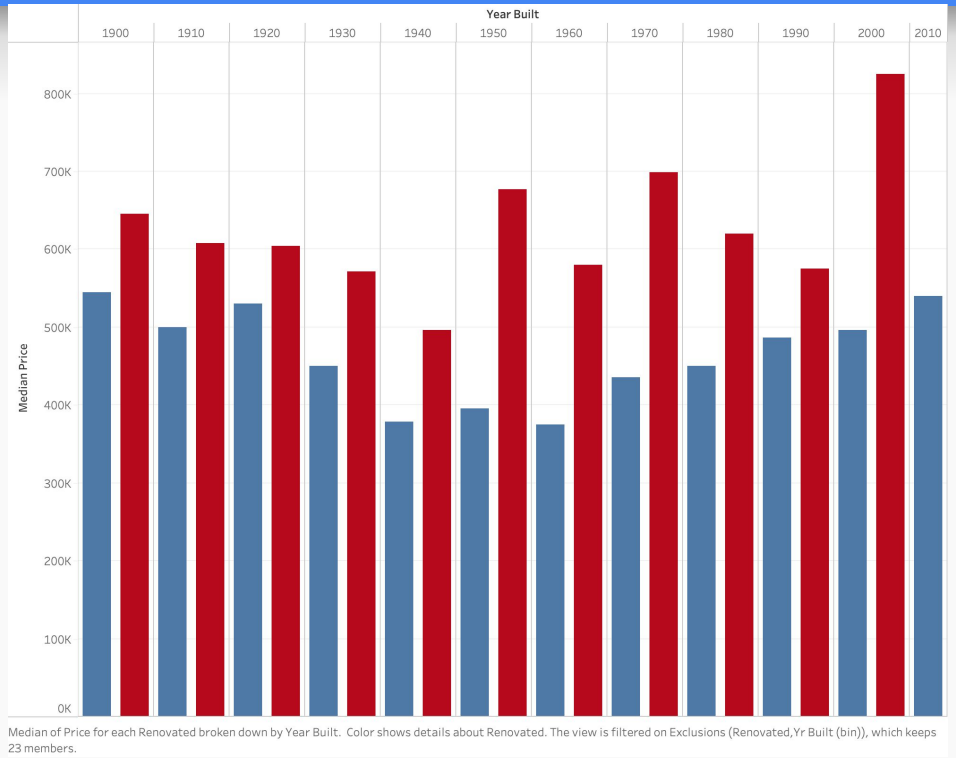
<- correlation w. price increased ->

	dist_to_seattle	dist_to_bellevue
0	12.434278	12.771789
1	12.477217	15.159928
2	16.247460	13.896328
3	10.731122	18.641163
4	21.850148	10.792469
...	...	...

0.29 0.42



Year Built VS Median Price



Median of Price for each Renovated broken down by Year Built. Color shows details about Renovated. The view is filtered on Exclusions (Renovated,Yr Built (bin)), which keeps 23 members.

Regression Method	N of Predictors	Transf.?	Train Set		Test Set	
			RMSE ($)	Adjusted R2	RMSE ($)	Adjusted R2
Linear (Baseline)	18	No	204,389	0.70	187,189	**0.71**
Linear	11	No	205,435	0.69	187,945	0.71
Linear Transformed	12	Yes	168,042	0.80	155,021	0.80
Ridge Regression (α = 1)	12	Yes	168,056	0.80	154,984	0.80
Lasso Regression (α = 0.1)	12	Yes	255,210	0.53	233,032	0.55
Polynomial (n=2)	12	Yes	131,009	0.88	128,919	0.86
Extreme Gradient Boosting (Final)	12	Yes	93,118	0.94	108,315	**0.90**