Regression Models Housing Prices King County, USA

Tal Gottfeld and Daniel Rodin

,Project #3

Naya College, August ‘21

Content

- Goal and Objectives
- The Problem
- The Data
- The Features
- The Model
  - Conclusions

Goal and Objectives

This project is designed to enable an initial trial with building a robust and flexible regression pipeline.

The Problem

- Designing a regression model to predict house prices in King County, USA

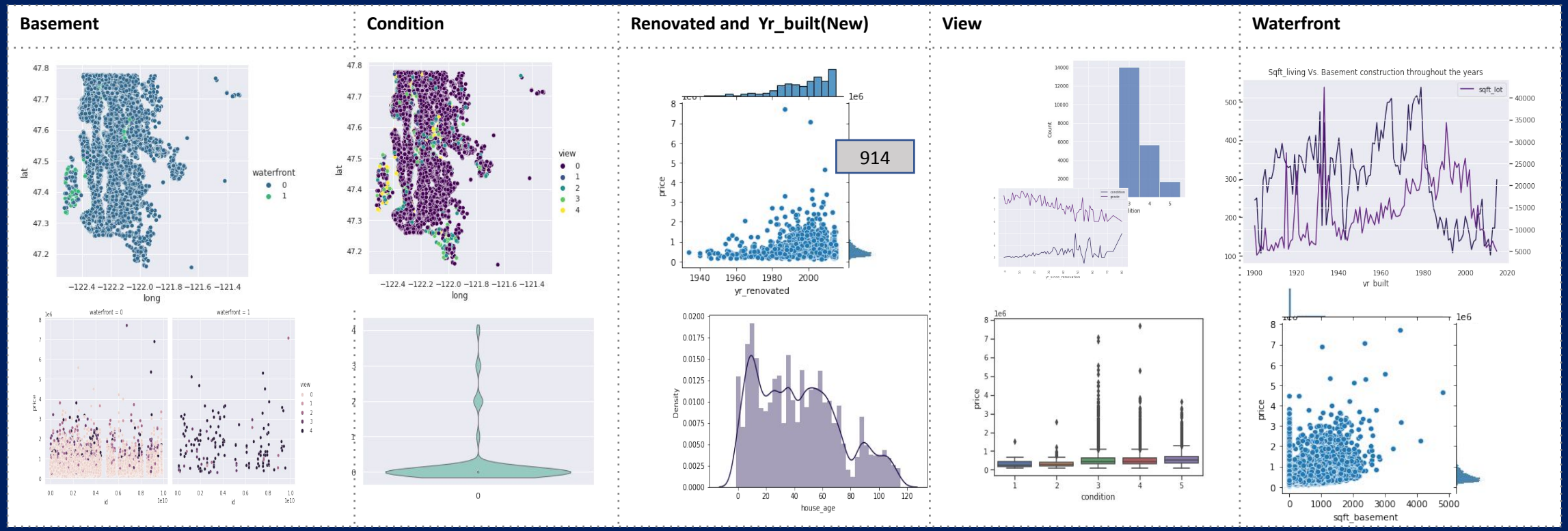
The Data

- Data source: Kaggle website
  https://www.kaggle.com/harlfoxem/housesalesprediction
- House price in King County (including Seattle), May 2014 to May 2015.

Features – What doesn't work

Method: Review every feature, analyze its behaviors, population, deviation and correlation

- 21 columns, 21613 rows
- After analysis and corr. review, we remained with only 7 features

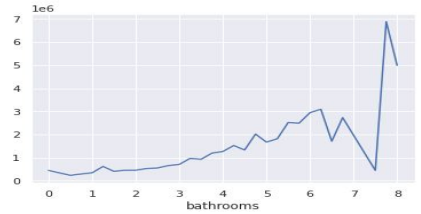


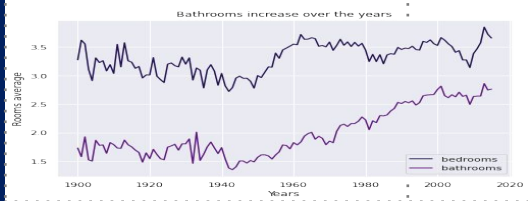
Additional colunms we removed: 'id', 'date,'**sqft_lot**', '**floors**', 'yr_built', '**renovated**', 'zip code', 'lat', 'long', 'sqft_living15', 'sqft_lot15', 'year_sold', 'yr_since_renovation', '**house_age**'],

Features – Our winning team

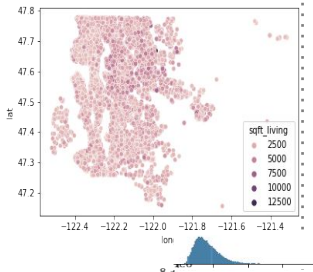
Method: Review every feature, analyze its behaviors, population, deviation and correlation

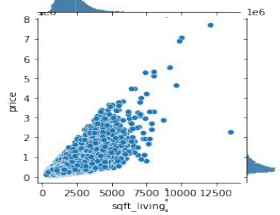
Grade



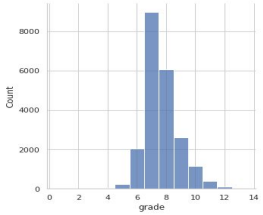


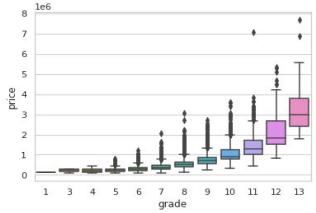
Sqft_living



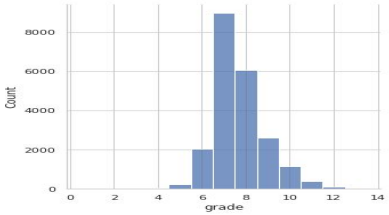


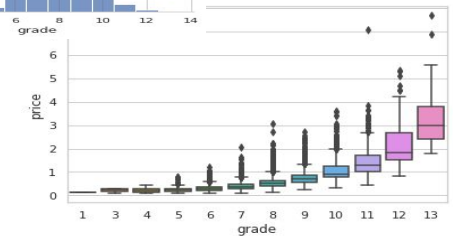
Bedrooms ans bathrooms



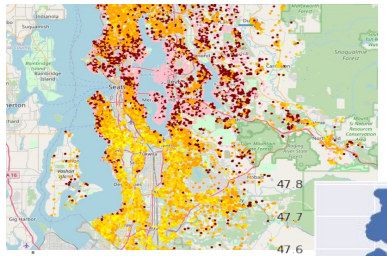


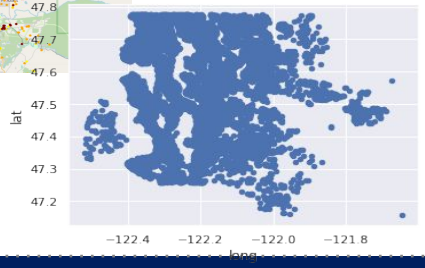
Zip_Value

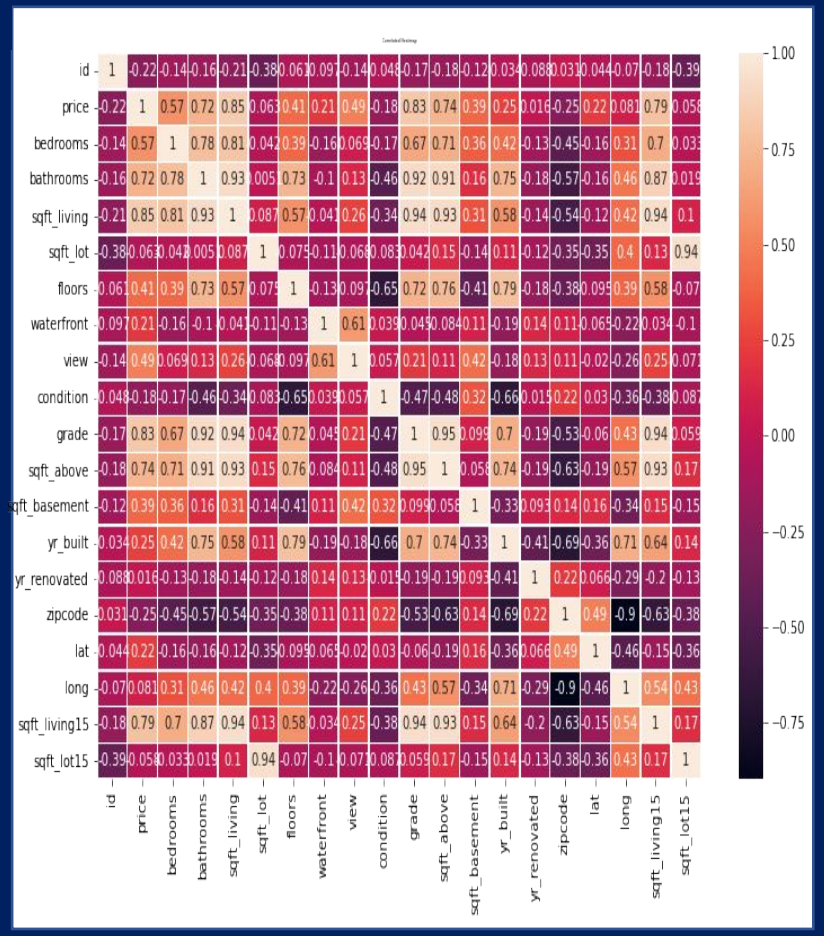


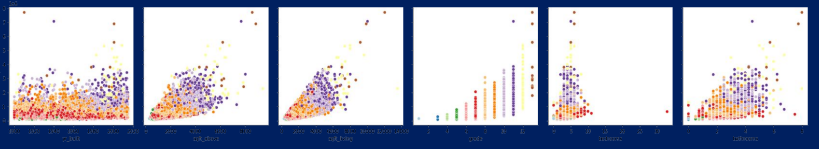


Grade



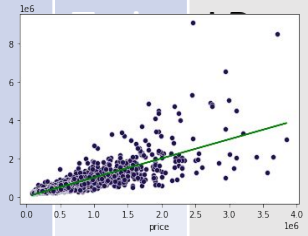
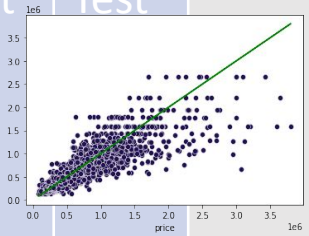
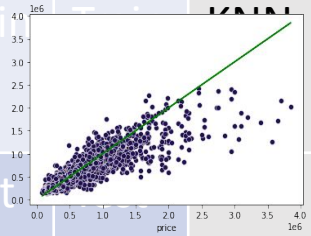






Models

Method: Columns removal, Outliers, scale of Max AbsScaler for X, Log our Y

Num of Tweak	RMSE	RMSLE	Scaler	LR	
Features 1. Sqft_above 2. Bedrooms 3. Yr_build	16%	0.25 0.25	Max AbsScaler Log for y		
Hyper param 1. max_depth from 20 to 30 & min_ samples leaf from to 500 2. leaf_depth - 40, max_leaf_nodes – 80 3. Transform Criterion from mse to mae with depth of 40	18%	0.2 0.22	Log to y_train Log to y_test	Train Test 	DT
Hyperparameters & Metric 1. We tired all hyper param: n_neighbors, depth, min and max leaf, 2. Applied all metrics including **User defined** – best yet! – 2%	2%	0.21 0.22	Log to y_train Log to y_test		KNN

Conclusion - Key notes and challenges

- EDA was a significant milestone in understanding our features and building the models
- Scaling is impactful in designing and optimizing our models:
  - MaxScaler for our X
  - Log for our y
  - Appling RMSLE on the results (data with large variance impacting RMSE)

Thank You

:Further reading: Colabs notebook and dataset

https://drive.google.com/drive/folders/1_FIODD5tLLDWMKyF8hnOaT8m4ap14g9w?usp=sharing

Legend

The below are descriptions of all the columns of our dataset

ID –sold house each for unique id

date - Date of the home sale

price - Price of each home sold

bedrooms - Number of bedrooms

bathrooms - Number of bathrooms, where .5 accounts for a room with a toilet but no shower

sqft_living - Square footage of the apartments interior living space

sqft_lot - Square footage of the land space

floors - Number of floors

waterfront - A dummy variable for whether the apartment was overlooking the waterfront or not

view - An index from 0 to 4 of how good the view of the property was

Condition - An index from 1 to 5 on the condition of the apartment

grade - An index from 1 to 13, where 1-3 falls short of building construction and design,

sqft_above - The square footage of the interior housing space that is above ground level

sqft_basement - The square footage of the interior housing space that is below ground level

yr_built - The year the house was initially built

yr_renovated - The year of the house's last renovation

zipcode - What zipcode area the house is in

lat - Lattitude

long - Longitude

sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors

sqft_lot15- The square footage of the land lots of the nearest 15 neighbors

Max AbsScaler

	price	bedrooms	bathrooms	sqft_living	grade	sqft_above	zip_value
0	221900.0	3	1.00	1180	7	1180	311225.161538
1	538000.0	3	2.25	2570	7	2170	470190.848039
2	180000.0	2	1.00	770	6	770	463304.432624
3	604000.0	4	3.00	1960	7	1050	552920.118321
4	510000.0	3	2.00	1680	8	1680	686357.152273

	bedrooms	bathrooms	sqft_living	grade	sqft_above	yr_built	zip_value
0	0.444444	0.454545	0.405483	0.692308	0.469115	0.991563	0.175352
1	0.444444	0.454545	0.389610	0.692308	0.450751	0.994541	0.324636
2	0.333333	0.545455	0.340548	0.538462	0.297162	0.984119	0.245948
3	0.444444	0.454545	0.476190	0.692308	0.287145	0.978164	0.345532
4	0.333333	0.272727	0.369408	0.615385	0.277129	0.956328	0.569492