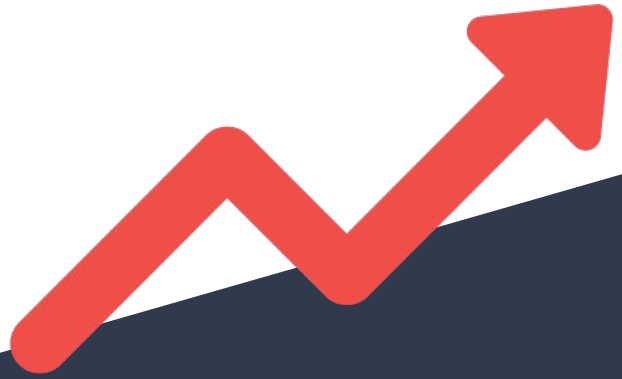
Predicting House Prices in King County, USA



MACS 33002 Project 1
Yonjoo Kim, Nicole Yip, Leon Zhang

Overview

Task: to train regression models and compare their performances in predicting house prices in King County, USA

Dataset: the dataset contains house sale prices for King County. It includes homes sold between May 2014 and May 2015. There are 21,613 instances in this dataset and 21 attributes

Models/Method: we will compare 3 different regression models: DT, RF, and linear regression. This is a supervised regression task as we have labeled data.

Dataset (summary)

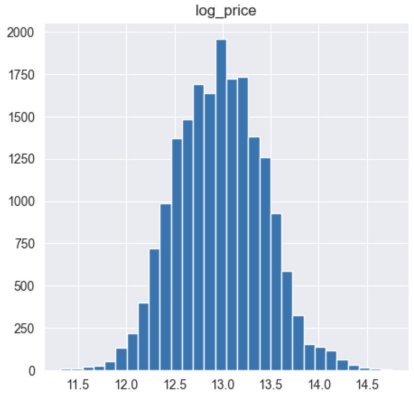
Number of Feature Variables	17
Type of Feature Variables	Categorical Numerical
Number of Instances	21,613
Target Variable	Log_price (transformed)
Number of missing values	None

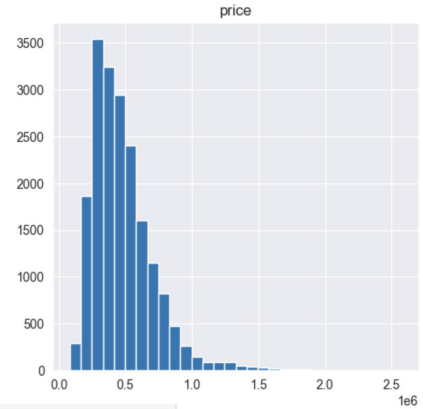
```
In [4]: df['date'] = pd.to_datetime(df['date'])
        df.head()
```

Out[4]:

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	yr_renovated
0	7129300520	2014-10-13	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180	0	1955	0
1	6414100192	2014-12-09	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170	400	1951	1991
2	5631500400	2015-02-25	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770	0	1933	0
3	2487200875	2014-12-09	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050	910	1965	0
4	1954400510	2015-02-18	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680	0	1987	0

5 rows × 21 columns





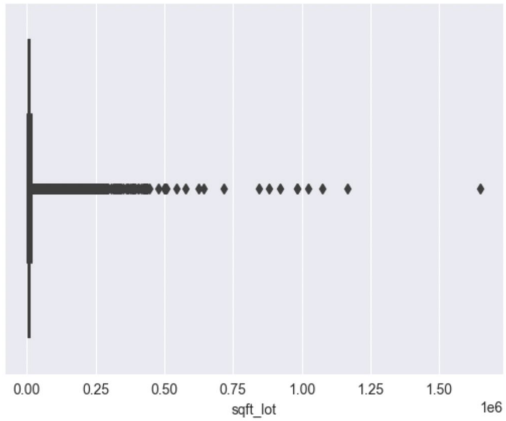
zipcode	lat	long	sqft_living15	sqft_lot15
98178	47.5112	-122.257	1340	5650
98125	47.7210	-122.319	1690	7639
98028	47.7379	-122.233	2720	8062
98136	47.5208	-122.393	1360	5000
98074	47.6168	-122.045	1800	7503

Data Cleaning & Pre-processing

sqft_lot

In [20]:
```
# draw box plot to check outliers
sns.set_style("darkgrid",{'axes.grid' : True})
sns.boxplot(x=df['sqft_lot'], orient='h')
```

Out[20]: <AxesSubplot: xlabel='sqft_lot'>



In [21]:
```
lower_thresh, upper_thresh = summarise_quartiles(df, 'sqft_lot')
```

```
Q1: 5000.0 Q3: 10300.0 min: -2950.0 max: 18250.0
Number of outliers under min_thresh: 0
Number of outliers above max_thresh: 2148
```

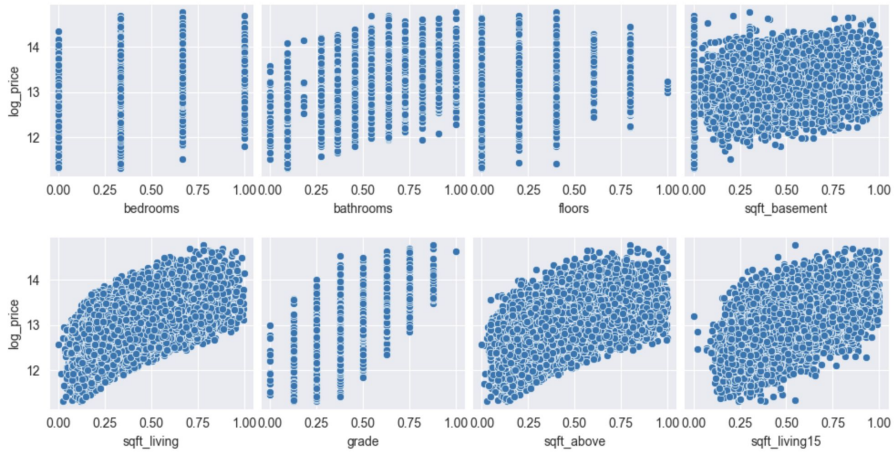
Outlier removal Standard: if in the thousands (21k original instances), it may be natural variation and therefore we will not remove. Adjusted for analysis of each variable's boxplot

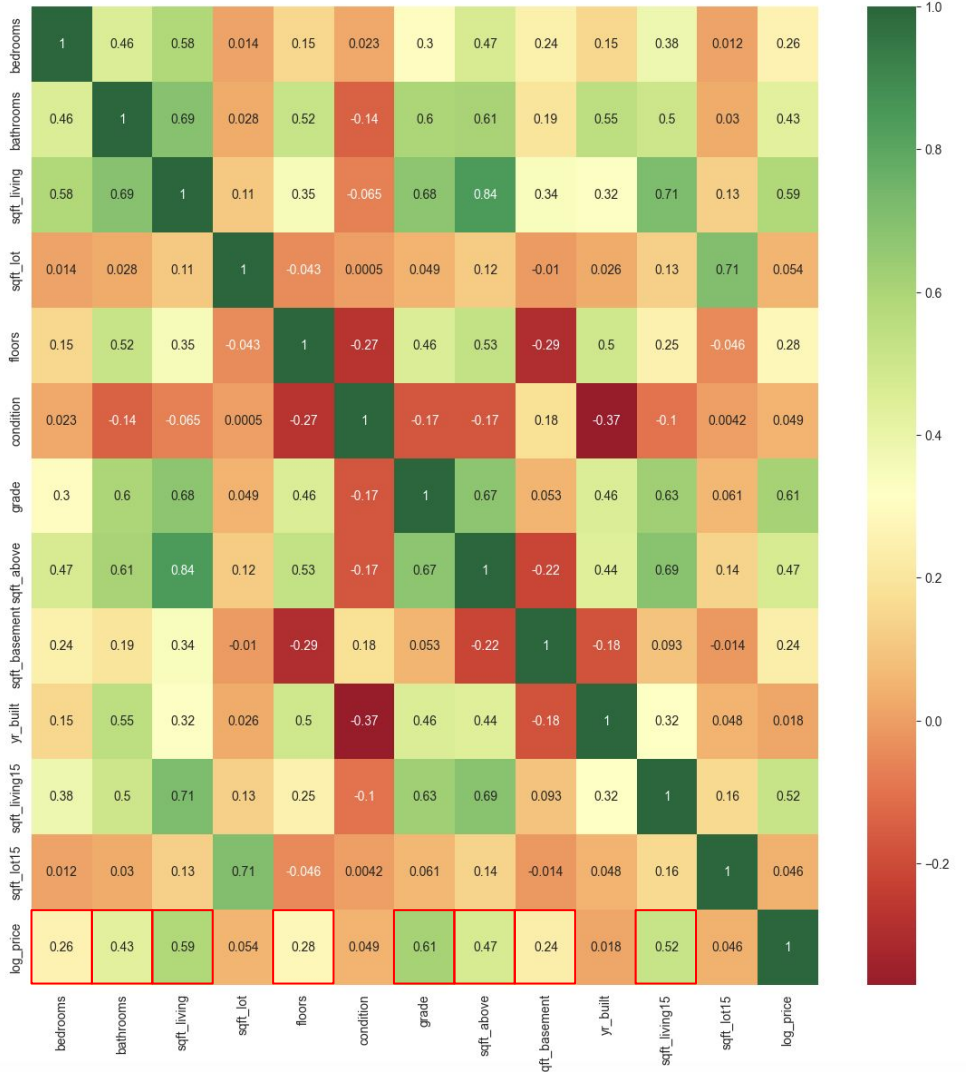
Log normalization on price, MinMaxScaler for all feature variables (numerical)

Training / Testing split 80% training, 20% testing

EDA

- Heat map
- Scatter plots
- Chose features with > 0.2 correlation





PCA

	Variance explained
0	0.299
1	0.122
2	0.110
3	0.079
4	0.067
5	0.057
6	0.048
7	0.041
8	0.037
9	0.030
10	0.024
11	0.022
12	0.016
13	0.016
14	0.015
15	0.014
16	0.005
17	0.000

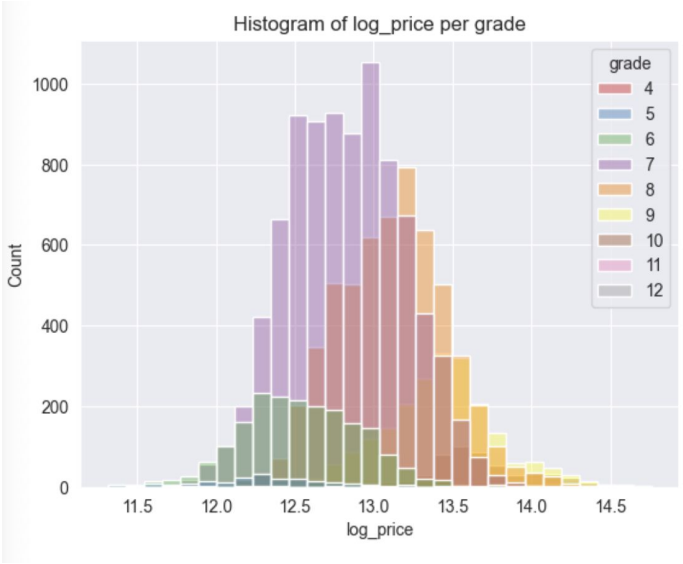
```
Total variance explained in top 3 principal components: 61.01%
Total variance explained in top 7 principal components: 82.20%
Total variance explained in top 10 principal components: 91.27%
```

	Model	Model Parameters	Accuracy/R^2	Test MSE	Train MSE	Test MAE	Train MAE
9	Random Forest	PCA	0.791	0.044	0.006	0.153	0.055
8	Decision Tree	PCA	0.692	0.065	0.041	0.186	0.150
10	Linear Regression: OLS	PCA	0.683	0.067	0.064	0.201	0.196
12	Linear Regression: Ridge	PCA, alpha=1.0	0.683	0.067	0.064	0.201	0.196
11	Linear Regression: Lasso	PCA, alpha=0.01	0.677	0.068	0.065	0.202	0.197
1	Random Forest	max_depth=11, min_samples_split=50	0.501	0.105	0.085	0.257	0.235
5	Linear Regression: OLS	Polynomial Degree 4	0.496	0.106	0.096	0.259	0.249
7	Linear Regression: Ridge	Polynomial Degree 4, alpha=1.0	0.491	0.107	0.102	0.262	0.257
0	Decision Tree	max_depth=9, min_samples_split=200	0.475	0.110	0.099	0.264	0.253
2	Linear Regression: OLS		0.444	0.117	0.113	0.275	0.271
4	Linear Regression: Ridge	alpha=1.0	0.444	0.117	0.113	0.275	0.271
3	Linear Regression: Lasso	alpha=0.01	0.402	0.126	0.123	0.287	0.284
6	Linear Regression: Lasso	Polynomial Degree 4, alpha=0.01	0.402	0.126	0.123	0.287	0.284

DT

DT models are well-suited for capturing linear relationships between the features and target variable.

Parameters	PCA
R-Square (accuracy)	0.692
MAE (training)	0.150
MAE (testing)	0.186
MSE (training)	0.041
MSE (testing)	0.065



RF

RF is used for predicting continuous target variables and is utilized because of its ability to make accurate predictions by combining the results from multiple DTs and reducing overfitting through random subset selection

Parameters	PCA
R-Square (accuracy)	0.791
MSE (Training)	0.006
MSE (Testing)	0.044
MAE (Training)	0.055
MAE (testing)	0.153

Linear Regression

LR is a simple and commonly used ML algorithm for predicting continuous target variables. It is used when there is a linear relationship between the independent and dependent variables; often provides a quick and straightforward solution.

Final best model for linear (Ridge)

Parameters	PCA, alpha = 1.0
R-Square (accuracy)	0.683
MAE (training)	0.196
MAE (testing)	0.201
MSE (training)	0.064
MSE (testing)	0.067

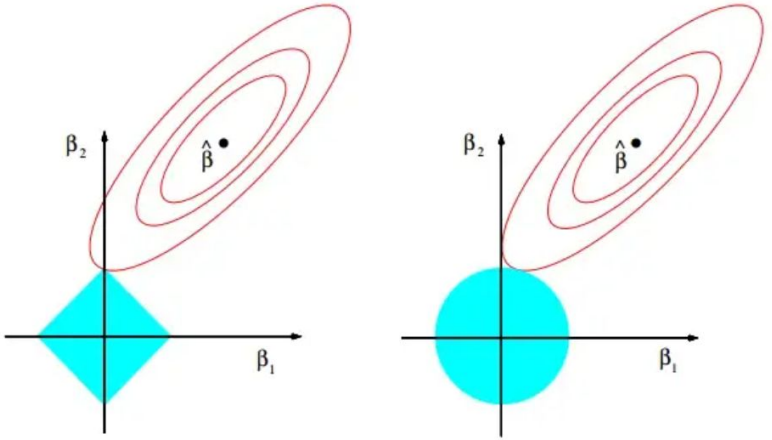


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Hastie et al. "The Elements of Statistical Learning" (2009)

Model Comparison

	Model	Model Parameters	Accuracy/R^2	Test MSE	Train MSE	Test MAE	Train MAE
9	Random Forest	PCA	0.791	0.044	0.006	0.153	0.055
8	Decision Tree	PCA	0.692	0.065	0.041	0.186	0.150
12	Linear Regression: Ridge	PCA, alpha=1.0	0.683	0.067	0.064	0.201	0.196

Error Analysis

Improvements for future analysis

- Wider and richer data (e.g. crime rates, distance to nearest school, how impatient the homeowner is to sell their house, etc.)
- Use more tailored Machine Learning models
- Optimise parameters (e.g. grid search)

Some solutions for improving complexity and the model in general