Group 3 Project 2

Jance Ng

Jin Min Wood

Adrian Kong

PROBLEM STATEMENT

As a property agency, it is very important to us that we are able to estimate what is the estimated worth of the property. By predicting the worth of a property accurately, we have higher confidence in selling the property and also receiving our well deserved commission. There are many manipulating variables that are affecting the price of a property. We will make use of regression model to estimate the SalePrice of a property and measure how well is our prediction and the actual SalePrice.

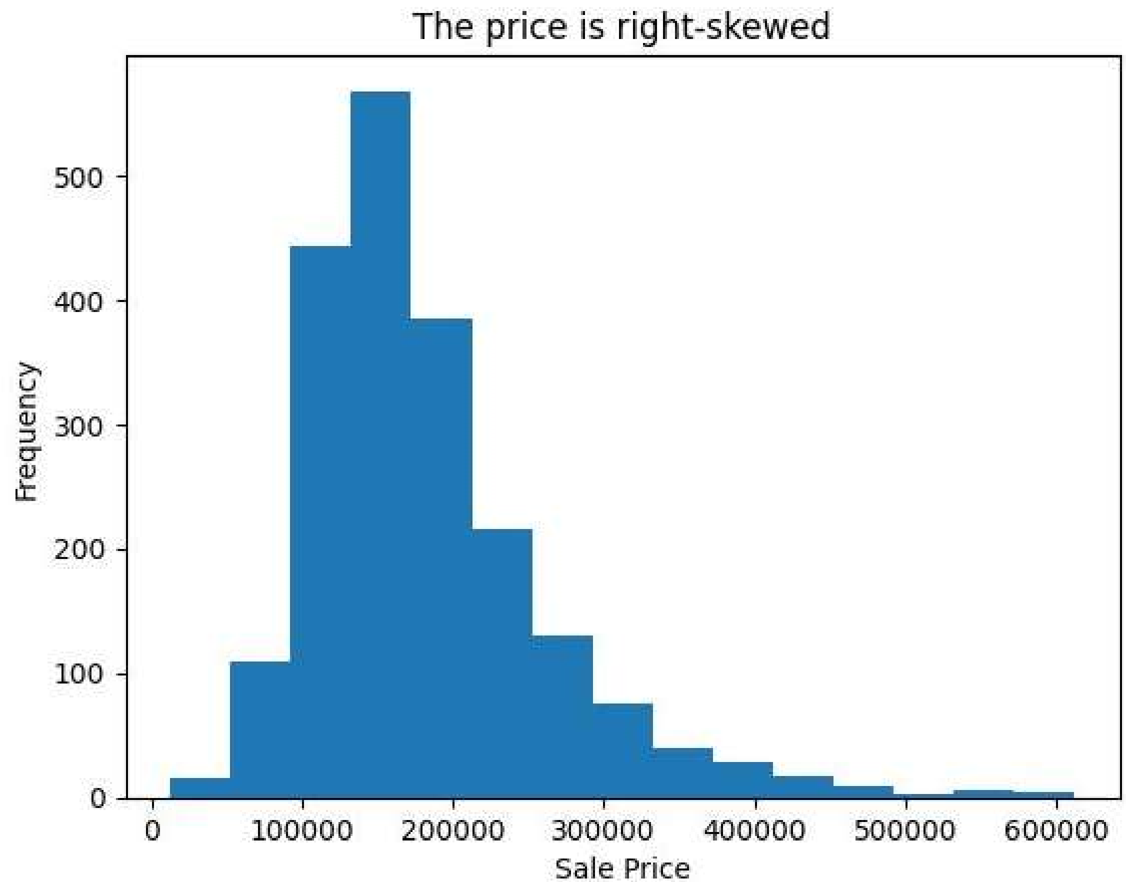
DATASET

Number of features: 81; examples below:

Area of the House	**Data Column (Feature)**	**Data Description**
Bathroom	bsmt_full_bath	Basement full bathrooms
	bsmt_half_bath	Basement half bathrooms
	full_bath	Full bathrooms above grade
	half_bath	Half baths above grade
Kitchen	kitchen_abvgr	Kitchens above grade
	kitchen_qual	Kitchen quality
Fireplace	fireplaces	Number of fireplaces
	fireplace_qu	Fireplace quality

For more information, please refer to: http://jse.amstat.org/v19n3/decock/DataDocumentation.txt)

EXPLORATORY VISUALIZATION



Convert Sale Price to logarithmic scale: **log(saleprice)**

DATA CLEANING - Missing Values

Feature	**Remarks**
Alley	It can be deduced that there is no alley access.
Bsmt Qual	The missing value count is the same as Bsmt Cond, it can be deduced that there is no basement for these units.
Bsmt Cond	Same as Bsmt Qual.
Garage Type	We can assume these houses have no garage.
Pool QC	We have no other reference, hence we will assume there is no pool.

DATA CLEANING - Missing Values - Exception 1

Feature	Remarks
Lot Frontage	We can assume that the unit is an apartment or condominium so it does not have lot frontage. However, looking through the MSSubclass, there is no apartment or condominium. Therefore, we will need to replace the missing values with the mean of the Lot Frontage.

Mean Value

	Lot Frontage
Lot Config	
Corner	83.245552
CulDSac	55.228571
FR2	60.836735
FR3	87.000000
Inside	66.952780

DATA CLEANING - Missing Values - Exception 1

Feature	Remarks
Garage Yr Blt	There is one house with value in under Garage Type but has missing garage year built. We replace missing value under numerical column with the mean value and missing value under categorical column with mode value

	Garage Type	Garage Yr Blt	Garage Finish	Garage Cars	Garage Area	Garage Cond
1712	Detchd	NaN	NaN	NaN	NaN	NaN

Garage Type	Detchd
Garage Yr Blt	1923.0
Garage Finish	Unf
Garage Cars	2.0
Garage Area	473.671707
Garage Cond	TA

PREPROCESSING

1. Split dataset into numerical and categorical features.
2. One-hot encode all categorical features.
3. Merge all numerical and the one-hot encoded categorical features
4. Split the dataset into 80% train and 20% test

Data	**Number of rows**	**Number of columns**
X_train	1640	274
X_test	411	274
y_train	1640	1
y_test	411	1

BASELINE MODEL

Model Used	Linear Regression
Features Used	Lot Area & Overall Quality
R2 score	0.7142968805431572

MODELING PART 1

Steps:

1. Standardize all features
2. Fit standardized features to Linear Regression model.

Model Used	Linear Regression
Features Used	Standardized all features
R2 score	-4.6009062687285534e+21

MODELING PART 1 (cont.)

Steps:

1. Find optimal alpha value
2. Fit standardized features to Ridge model with optimal alpha value.

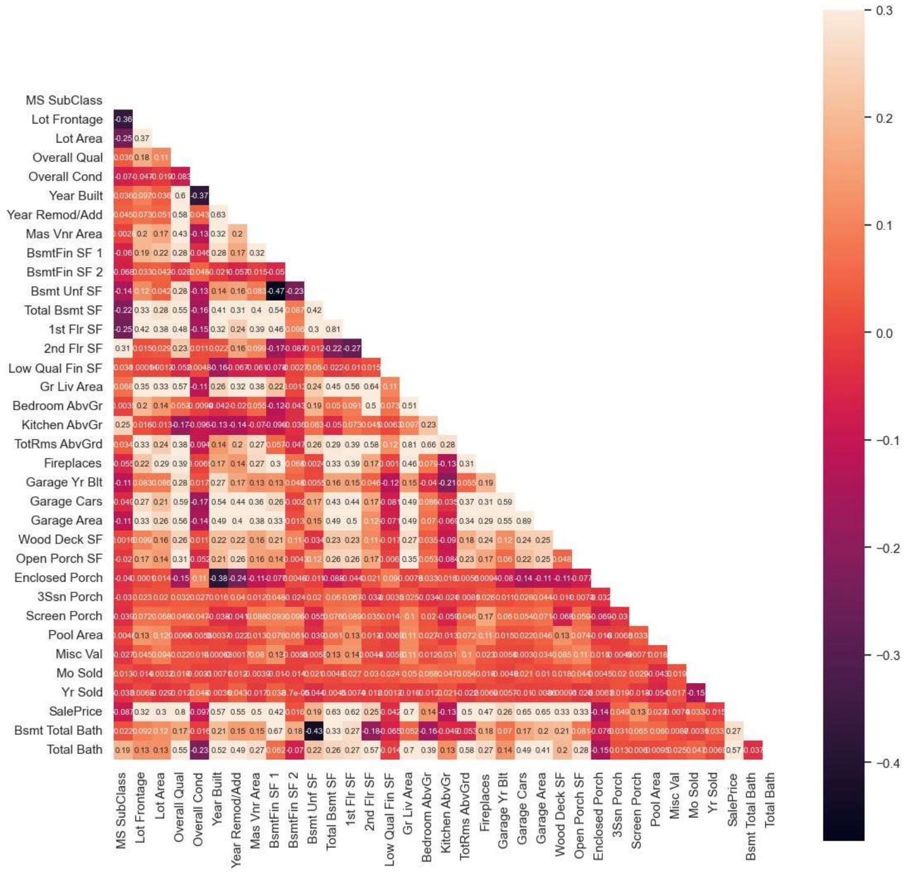
Model Used	Ridge
Features Used	Standardized all features
R2 score	0.8397908717781926
RMSE	199325.9060718847

MODELING PART 1 (cont.)

Steps:
1. Find optimal alpha value
2. Fit standardized features to Lasso model with optimal alpha value.

Model Used	Lasso
Features Used	Standardized all features
R2 score	0.8437015234296557
RMSE	199147.08628262393

DATA PROCESSING PART 2



➔ Heatmap of correlation of all numerical features.

FEATURE	**VIF**
Overall Qual	60.667728
Year Built	9058.56774
Year Remod/Add	8910.26204
Mas Vnr Area	1.85017
Total Bsmt SF	22.201645
1st Flr SF	34.998552
Gr Liv Area	59.79677
TotRms AbvGrd	55.854408
Garage Cars	36.763833
Garage Area	32.589563
SalePrice	28.860521
Total Bath	22.967767

- These are the features with correlation to SalePrice higher than **0.5**
- Year Built and Year Remod/Add has high Variance Inflation Factor

Final features selected in Part 2 are:

- Overall Qual
- Mas Vnr Area
- Total Bsmt SF
- 1st Flr SF
- Gr Liv Area
- TotRms AbvGrd
- Garage Cars
- Garage Area
- Total Bath
- Age

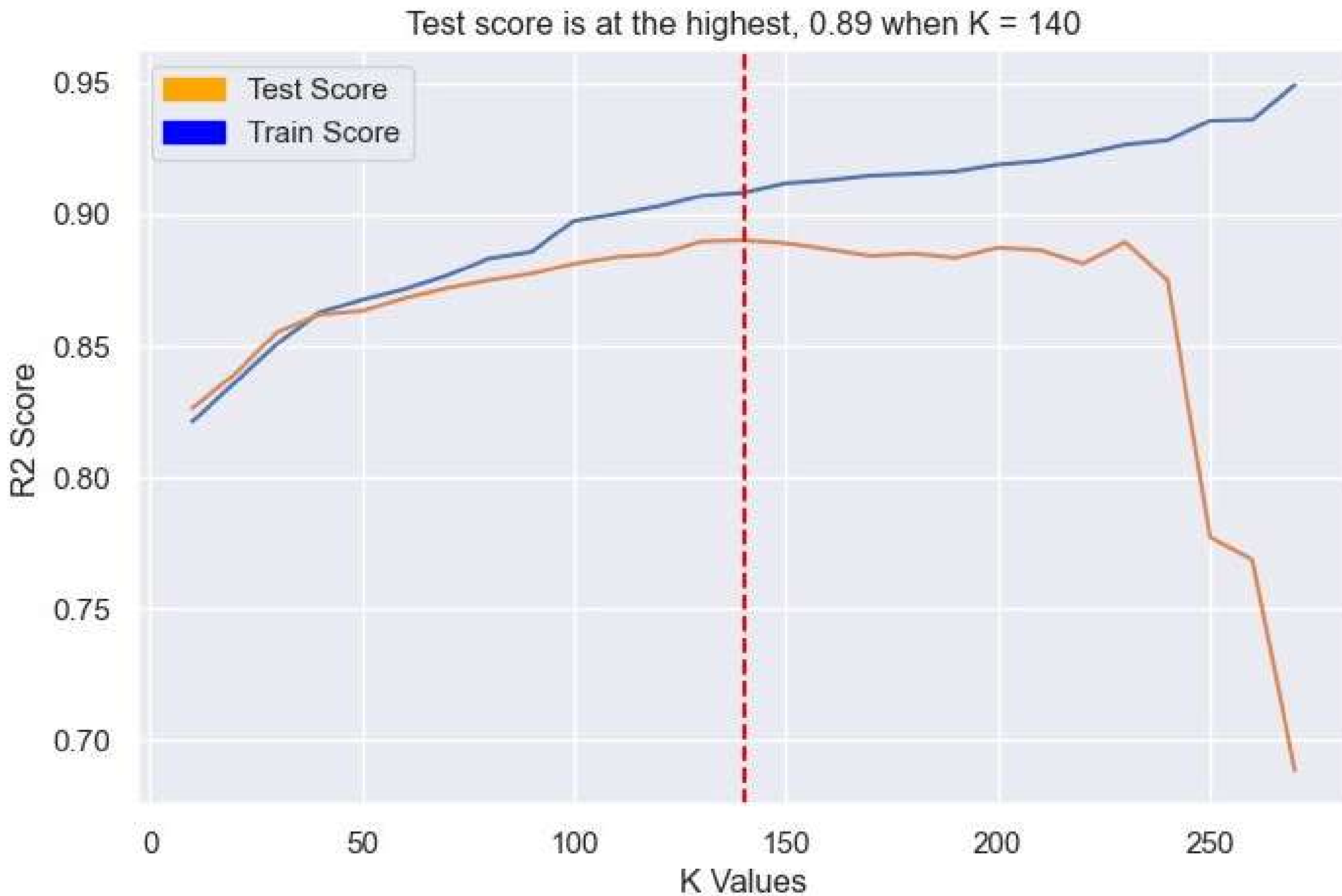
MODELING PART 2

Steps:

1. Standardize all features selected in Data Processing Part 2
2. Fit standardized features to Linear Regression model.

Model Used	Linear Regression
Features Used	Standardized features selected in Part 2
R2 score	0.7702069554872076
RMSE	129467.33485416573

MODELING PART 3 (With SelectKBest)



MODELING PART 3

Steps:

1. Make use of SelectKBest function to select optimum number of features, K.
2. Fit and transform the training data with K number features.
3. Fit transformed data to Linear Regression model.

Model Used	Linear Regression
Features Used	140 features selected with SelectKBest function
R2 score	0.890137859811365
RMSE	23914.714072188213

CONCLUSION

Model	In Section	R2 Score	RMSE
Linear Regression	Baseline Model	0.7142968805431572	Not Calculated
Linear Regression	Modeling Part 1	-4.6009062687285534e+21	Not Calculated
Ridge	Modeling Part 1	0.8397908717781926	199,325.9060718847
Lasso	Modeling Part 1	0.8437015234296557	199,147.08628262393
Linear Regression	Modeling Part 2	0.7702069554872076	129,467.33485416573
Linear Regression	Modeling Part 3	0.890137859811365	23,914.714072188213

The method and model in Modeling Part 3 gave us the best R2 score and RMSE. This model will be used to predict house sale price.