Assignment 1 - Regression

Student Details:

- Name : Anjishnu Mukherjee
- Registration Number : B05-511017020
- Class Roll Number : CS Gy-70
- Exam Roll Number : 510517086
- Email : 511017020.anjishnu@students.iiests.ac.in

Project Setup

Mount Google Drive

```
Mounted at /content/drive/
```

Load files and libraries, set seed

Source of Data : Kaggle House Prices Dataset

```
Environment Information :
-----------------------------------
OS: Linux-4.19.112+-x86_64-with-Ubuntu-18.04-bionic
Python version: 3.6.9 (default, Jul 17 2020, 12:50:27)  [GCC 8.4.0]
Numpy version: 1.18.5
Pandas version: 1.0.5
Matplotlib version: 3.2.2
Seaborn version: 0.10.1
Scikitlearn version: 0.22.2.post1
-----------------------------------
```

Read dataset into Pandas dataframe

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Conditic
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Nc
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Nc
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Nc
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Nc
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Nc

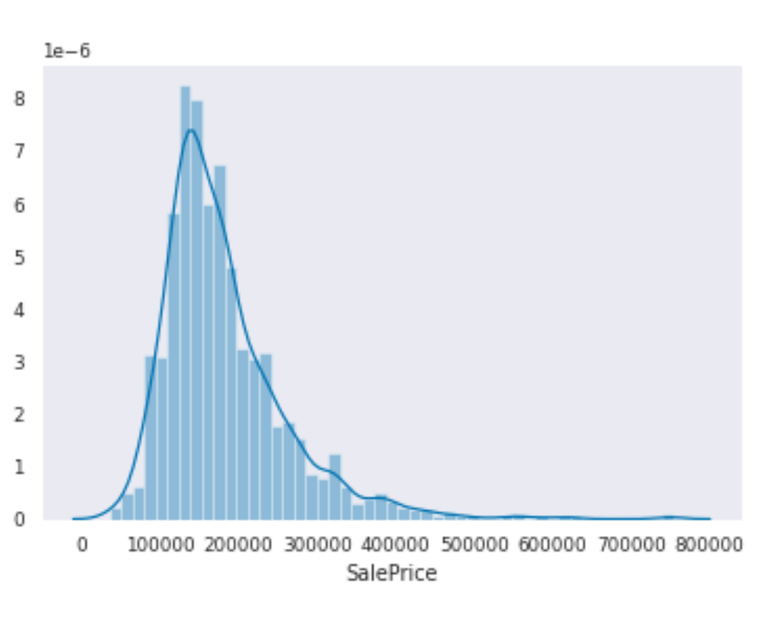
5 rows × 81 columns

Analyze the columns of the dataset.

SalePrice distribution isn't a perfect Gaussian curve.

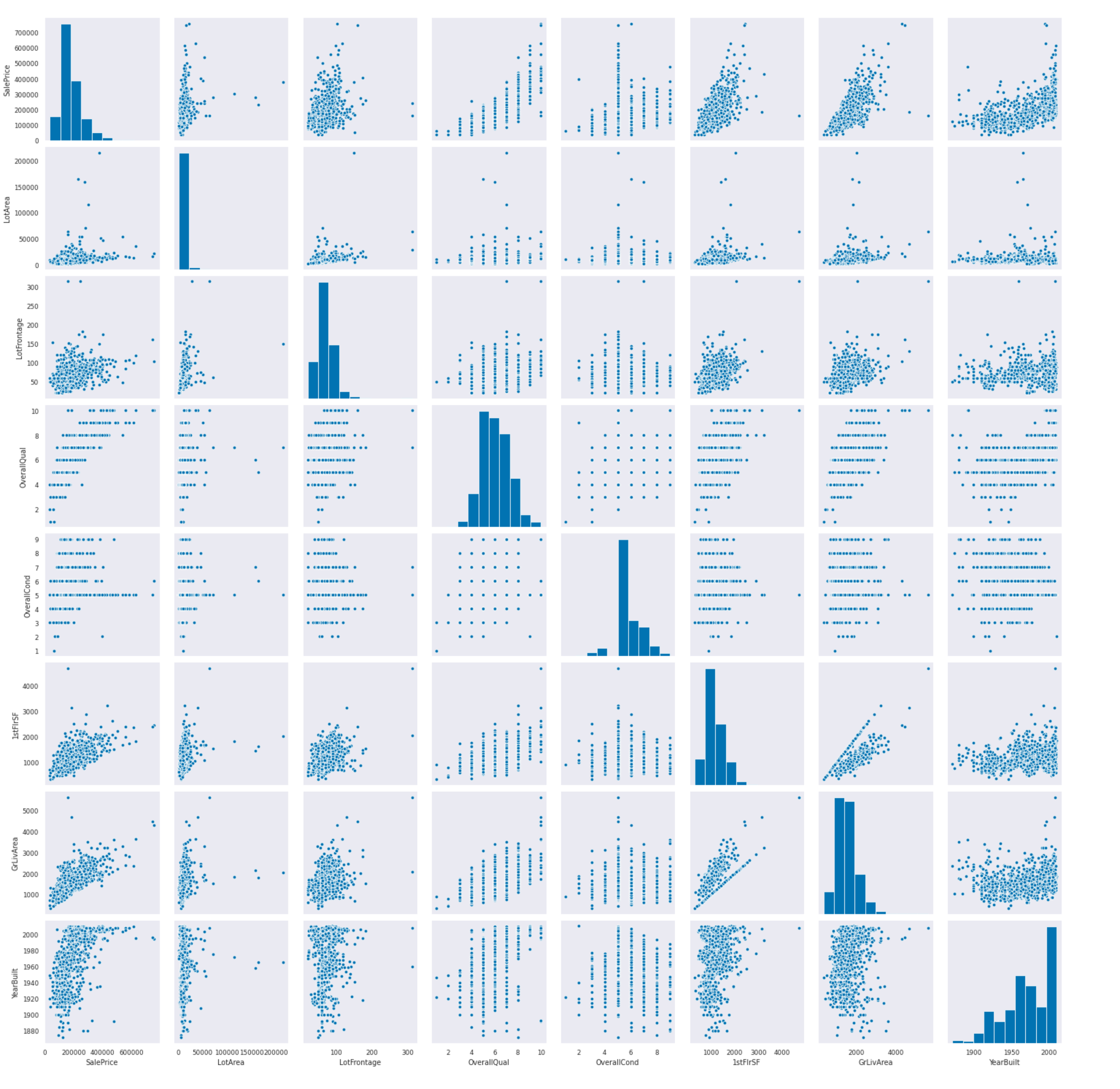
```
count      1460.000000
mean     180921.195890
std       79442.502883
min       34900.000000
25%      129975.000000
50%      163000.000000
75%      214000.000000
max      755000.000000
Name: SalePrice, dtype: float64
```

The mean SalePrice is nearly 181,000 units which seems to be a reasonable number given the context of this dataset.



Clearly SalePrice follows a slightly shifted normal distribution.

Let's now compare the pairwise distributions of all the relevant features of analysis.



Remove the rows with a missing value.

Additionally I retain only those columns which will be used for this analysis and drop all rows which have missing values in those columns.

```
Data shape :

(1460, 81)

Data with only required columns :

(1460, 10)

Data after dropping all rows with missing values :

(1201, 10)

Number of missing values in data :

0
```

For all the cells in the final dataframe, there are no missing values.

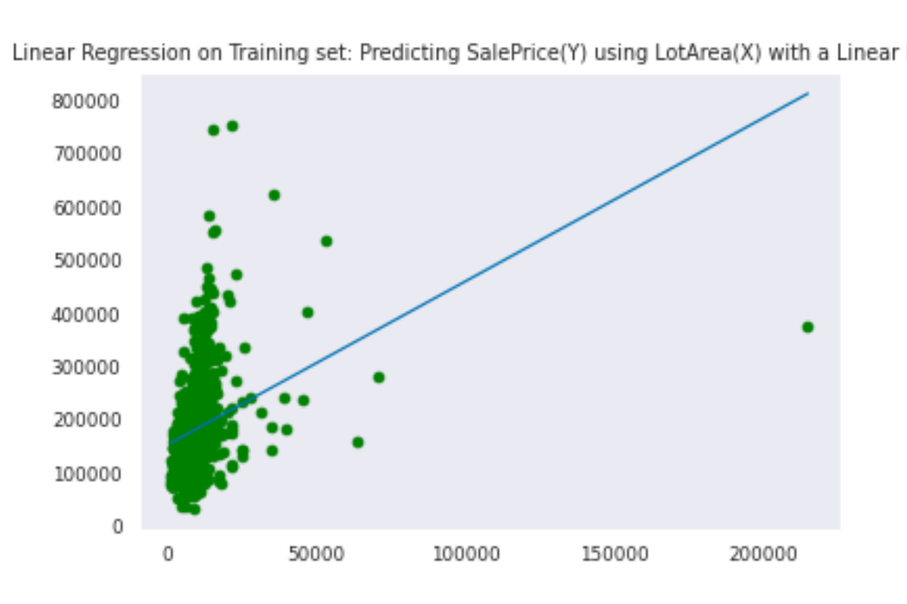
Divide the training.csv into two sets of ratio 80:20 entitled to train and test set respectively.

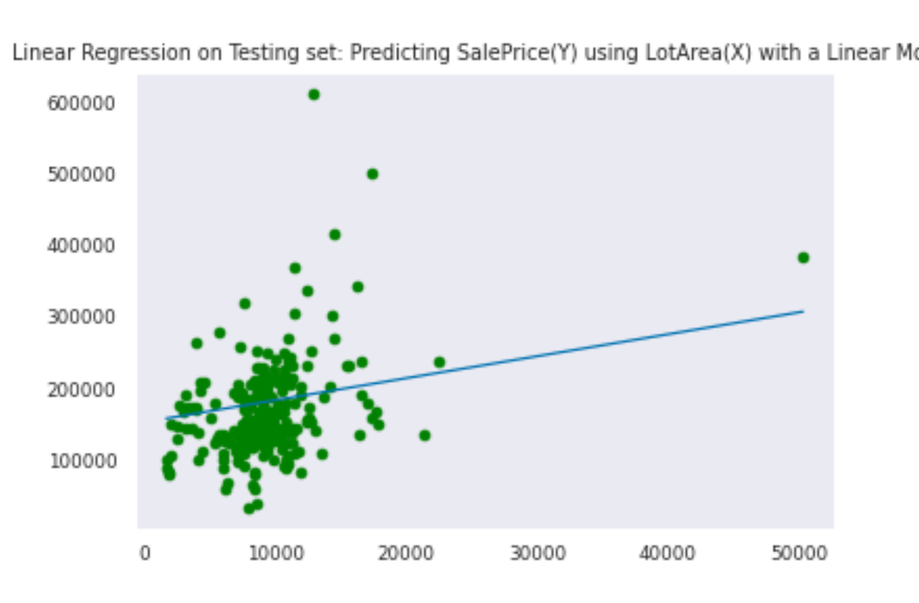
```
Training set shape:

(960, 10)

Testing set shape:

(241, 10)
```

Use the linear regression method to estimate the slope and intercept for predicting 'SalePrice' based on 'LotArea'

- Model 0 : LotArea

```
-----------------------------------
Slope : [3.06669265]
Intercept:  153128.20137698896
Training R2 Score 0.09270179567453762
Test R2 Score 0.06718954346588057
Training MSE Score: 6801580464.01976
Test MSE Score: 4228531500.5697026
-----------------------------------
```





Use the multiple regression method to estimate the value of the weights/coefficients for predicting 'SalePrice' based on the following features:

- Model 1: LotFrontage, LotArea
- Model 2: LotFrontage, LotArea, OverallQual, OverallCond
- Model 3: LotFrontage, LotArea, OverallQual, OverallCond, 1stFlrSF, GrLivArea

Model 1: LotFrontage, LotArea

```
Coefficients : [   1.90578562 947.71569695]
Intercept:  98252.52075249185
Training R2 Score 0.15548272542793504
Test R2 Score 0.10330767116226003
Training MSE Score: 6330941876.521212
Test MSE Score: 4064804089.8873663
-----------------------------------
```

Model 2: LotFrontage, LotArea, OverallQual, OverallCond

```
Coefficients : [ 1.24181274e+00  3.61230895e+02  4.50837864e+04 -6.39899391e+02]
Intercept:  -127974.7555503003
Training R2 Score 0.6855835863054668
Test R2 Score 0.6307821250059973
Training MSE Score: 2357029394.257201
Test MSE Score: 1673704881.8968096
-----------------------------------
```

Model 3: LotFrontage, LotArea, OverallQual, OverallCond, 1stFlrSF, GrLivArea

```
Coefficients : [7.17308596e-01 1.23781110e+01 3.32884093e+04 7.11055322e+02
 3.35584568e+01 3.94411708e+01]
Intercept:  -132673.55210699735
Training R2 Score 0.7460817293514201
Test R2 Score 0.7127218913283009
Training MSE Score: 1903503766.3113735
Test MSE Score: 1302262987.5482461
-----------------------------------
```

Use the multiple regression method to estimate the value of the weights/coefficients for predicting 'SalePrice' based on the following set of mixed (numerical and categorical) features:

- Model 4: LotArea, Street
- Model 5: LotArea, OverallCond, Street, Neighborhood
- Model 6: LotArea, OverallCond, Street, 1stFlrSF, Neighborhood, Year

Check the range of years in the training data.

```
Most recent year :
2010
Least recent year :
1872
```

Testing dataset doesn't have the *ClearCr* neighborhood or *Gravel* streets, so add a column of zeros for them manually

```
Training set :
['Blmngtn', 'Blueste', 'BrDale', 'BrkSide', 'ClearCr', 'CollgCr', 'Crawfor', 'Edwards', 'Gilbert', 'IDOTRR', 'MeadowV
', 'Mitchel', 'NAmes', 'NPkVill', 'NWAmes', 'NoRidge', 'NridgHt', 'OldTown', 'SWISU', 'Sawyer', 'SawyerW', 'Somerst',
'StoneBr', 'Timber', 'Veenker']
Testing set :
['Blmngtn', 'Blueste', 'BrDale', 'BrkSide', 'CollgCr', 'Crawfor', 'Edwards', 'Gilbert', 'IDOTRR', 'MeadowV', 'Mitchel
', 'NAmes', 'NPkVill', 'NWAmes', 'NoRidge', 'NridgHt', 'OldTown', 'SWISU', 'Sawyer', 'SawyerW', 'Somerst', 'StoneBr',
'Timber', 'Veenker']
```

```
Training set :
['Pave' 'Grvl']
Testing set :
['Pave']
```

```
Testing set shape:

(960, 10)

Testing set shape:

(241, 12)
```

Create dummy variables for each categorical feature

```
Testing set shape:

(960, 37)

Testing set shape:

(241, 37)
```

Model 4: LotArea, Street

```
Coefficients : [ 3.10841101e+00 -4.69160661e+04  4.69160661e+04]
Intercept:  106278.49976814234
Training R2 Score 0.09876983231048453
Test R2 Score 0.06530285163673055
Training MSE Score: 6756091297.127055
Test MSE Score: 4237084080.3307295
-----------------------------------
```

Model 5: LotArea, OverallCond, Street, Neighborhood

```
Coefficients : [ 2.03702836e+00 -1.63459937e+04  1.63459937e+04  7.45278117e+03
  2.13116799e+04 -4.83540948e+04 -6.39468239e+04 -5.98834775e+04
 -6.70195991e+03  1.50231836e+04  6.19760751e+03 -5.59521593e+04
  7.56727756e+03 -7.89436935e+04 -6.90762057e+04 -3.69137601e+04
 -4.22548398e+04 -2.58279399e+04 -2.73427218e+03  1.48192048e+05
  1.31809577e+05 -6.08112809e+04 -3.48747606e+04 -5.25354616e+04
  6.65346685e+03  5.68583079e+04  1.50813098e+05  4.47176790e+04
  4.96668034e+04]
Intercept:  108643.27719470818
Training R2 Score 0.6040465602247971
Test R2 Score 0.616775885881661
Training MSE Score: 2968273460.475608
Test MSE Score: 1737196690.899276
-----------------------------------
```

Model 6: LotArea, OverallCond, Street, 1stFlrSF, Neighborhood, Year

```
Coefficients : [ 1.17573136e+00 -1.76788941e+04  1.76788941e+04  6.12520486e+02
  1.10509350e+04  7.80255370e+01 -2.29032536e+04 -1.62884836e+04
 -2.70240525e+04 -2.21325300e+04 -5.01082596e+02 -1.86051072e+03
  2.01715552e+04 -4.41029802e+04  6.65234229e+03 -2.77880153e+04
 -4.48760776e+04 -4.52879121e+04 -3.56778091e+04 -1.37721523e+04
 -1.92188824e+04  1.12158996e+05  8.16911759e+04 -2.03724941e+04
 -5.03280613e+03 -4.63537441e+04 -6.69531638e+03  3.22316570e+04
  1.04135144e+05  2.04784180e+04  2.23688141e+04]
Intercept:  -1201787.3598678526
Training R2 Score 0.7100668171505187
Test R2 Score 0.725195006021002
Training MSE Score: 2173490328.6909986
Test MSE Score: 1245720998.7456412
-----------------------------------
```

List of all the models and features each one uses.

Model	Features used
Model 0	LotArea
Model 1	LotFrontage, LotArea
Model 2	LotFrontage, LotArea, OverallQual, OverallCond
Model 3	LotFrontage, LotArea, OverallQual, OverallCond, 1stFlrSF, GrLivArea
Model 4	LotArea, Street
Model 5	LotArea, OverallCond, Street, Neighborhood
Model 6	LotArea, OverallCond, Street, 1stFlrSF, Neighborhood, Year

Calculate and compare the Mean squared Error, R2 score for each of the model for test and training set for the above models.

Model	MSE (training)	MSE(testing)	R2 (training)	R2(testing)
Model 0	6801580464.01976	4228531500.5697026	0.09270179567453762	0.06718954346588057
Model 1	6330941876.521212	4064804089.8873663	0.15548272542793504	0.10330767116226003
Model 2	2357029394.257201	1673704881.8968096	0.6855835863054668	0.6307821250059973
Model 3	1903503766.3113735	1302262987.5482461	0.7460817293514201	0.7127218913283009
Model 4	6756091297.127055	4237084080.3307295	0.09876983231048453	0.06530285163673055
Model 5	2968273460.475608	1737196690.899276	0.6040465602247971	0.616775885881661
Model 6	2173490328.6909986	1245720998.7456412	0.7100668171505187	0.725195006021002

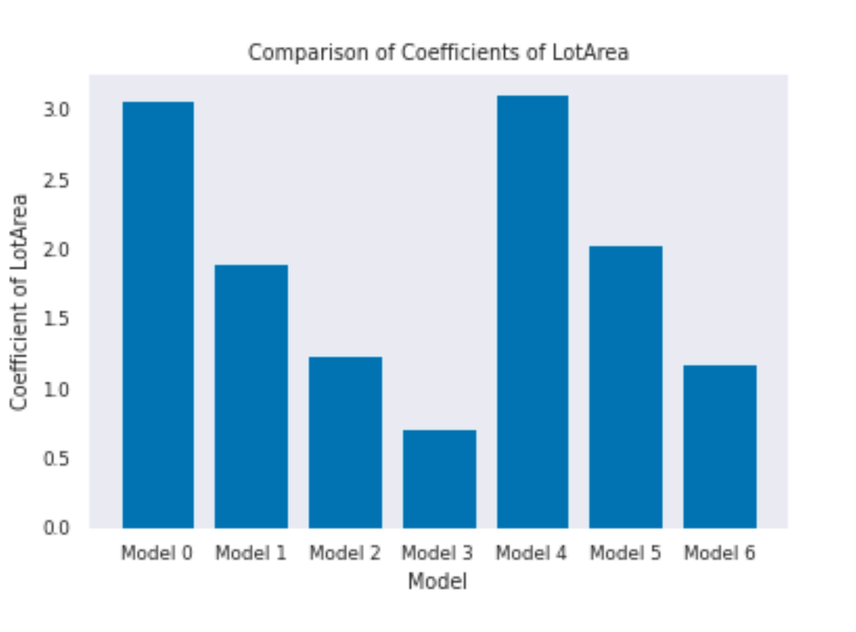
```
R2_training : Model_3 > Model_6 > Model_2 > Model_5 > Model_1 > Model_4 > Model_0
R2_testing : Model_6 > Model_3 > Model_2 > Model_5 > Model_1 > Model_0 > Model_4
```

Usually, the larger the R2, the better the regression model fits your observations.

Thus Model_3 and Model_6 with R2 scores of nearly 75% appear to be the best out of these 7 models.

Compare the feature “LotArea” weights/coefficients for all the seven trained models.

Model	Coefficients of LotArea
Model 0	3.06669265
Model 1	1.9057856
Model 2	1.24181274e+00
Model 3	7.17308596e-01
Model 4	3.10841101e+00
Model 5	2.03702836e+00
Model 6	1.17573136e+00



We can clearly see that the models which had the best R2 scores (Model_3 and Model_6) assign very small coefficients to LotArea, indicating that it is not really the most important feature for SalePrice prediction due to being highly correlated with other features in the feature set .

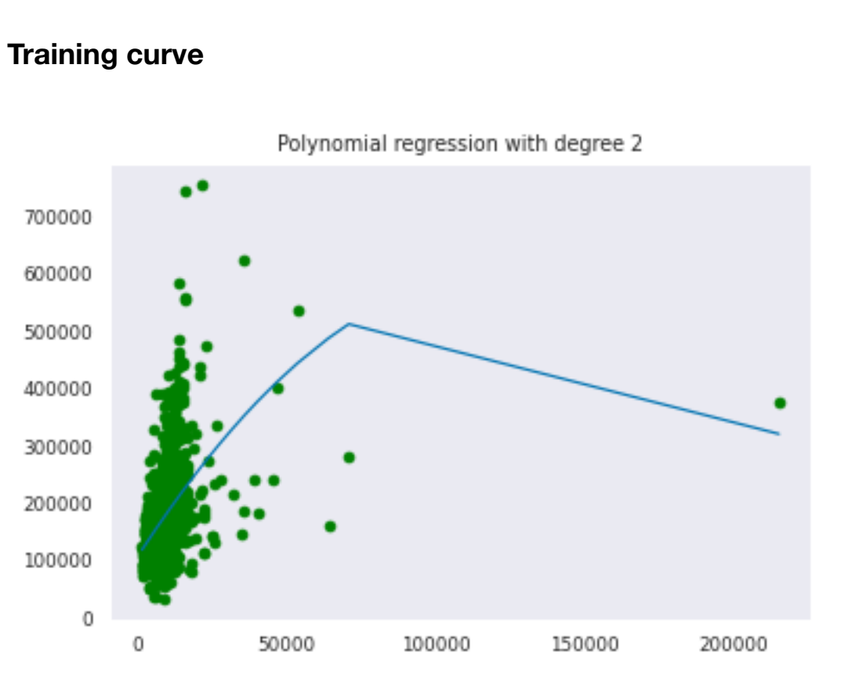
Use the polynomial regression of degree (2 and 3), to estimate the value of the weights/coefficients for predicting 'SalePrice' based on 'LotArea'.

- Also, print the graph on the training and test set.

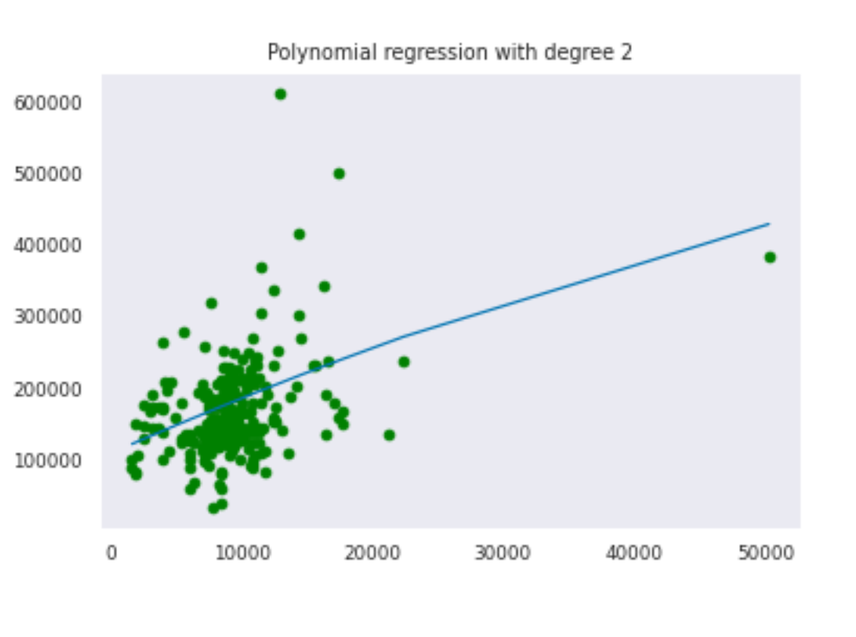
Polynomial Regression of degree 2

```
Training R2 Score 0.1779902407230977
Test R2 Score 0.11670371531453783
Training MSE Score: 6162213804.983787
Test MSE Score: 4004078361.22069
```

Training curve



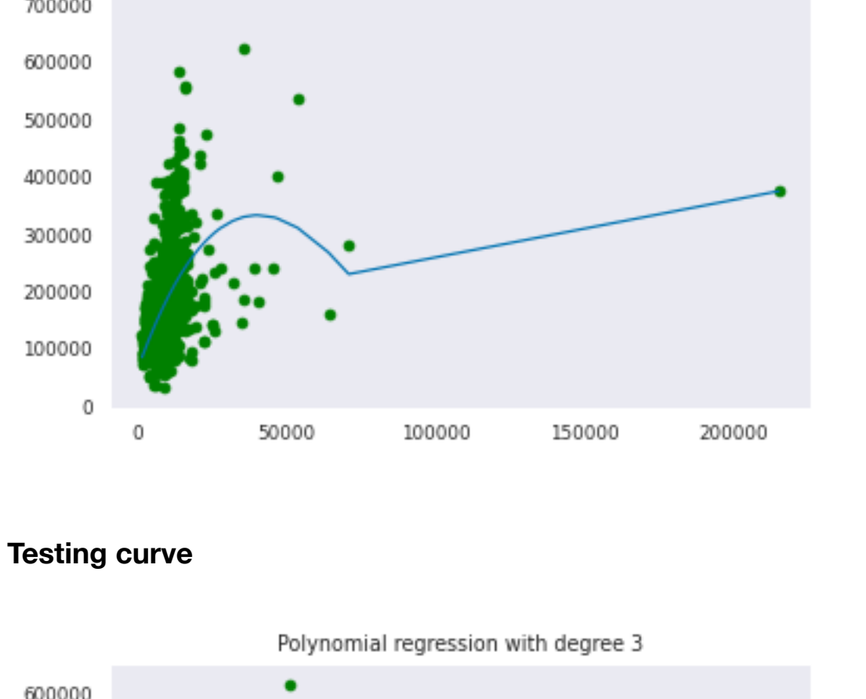
Testing curve



Polynomial Regression of degree 3

```
Training R2 Score 0.2156904294515526
Test R2 Score 0.07836163416724373
Training MSE Score: 5879593530.94064
Test MSE Score: 4177887195.365982
```

Training curve



Testing curve