FLIP ROBO

HOUSING: PRICE PREDICTION

Submitted by:
CHESTHA SHARMA

ACKNOWLEDGMENT

The project entitled “Housing: Price Prediction” is done by me during my internship with Flip Robo Technologies. I am grateful to Data Trained and Flip Robo Technologies for their guidance during this project

INTRODUCTION

- Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain.

- Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

- We are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market

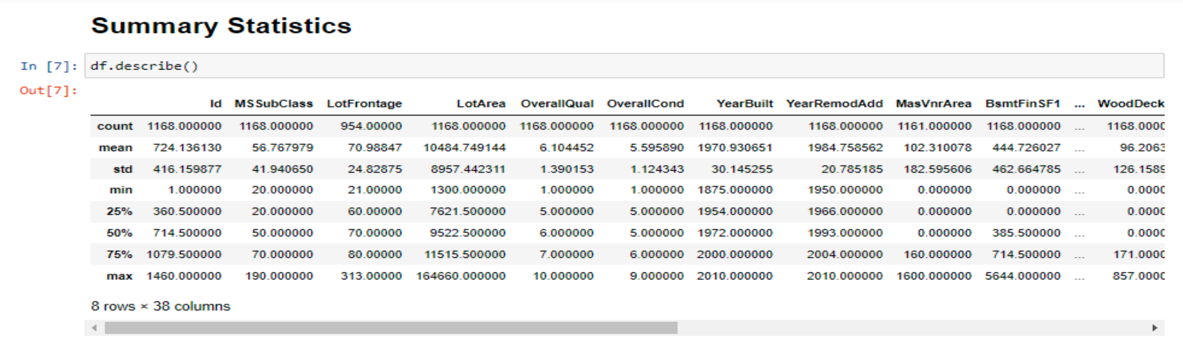
- A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

- The company is looking at prospective properties to buy houses to enter the market. We have to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:
- Which variables are important to predict the price of variable?
- How do these variables describe the price of the house?

Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

In this project we have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them.



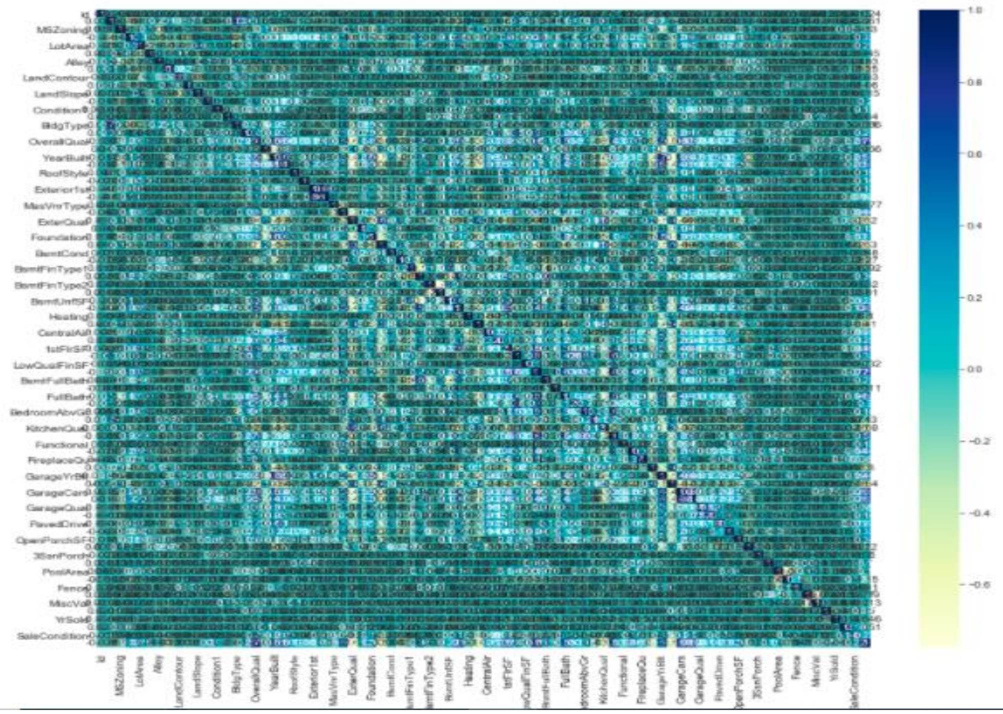
Summary Statistics

```
In [7]: df.describe()
Out[7]:
```

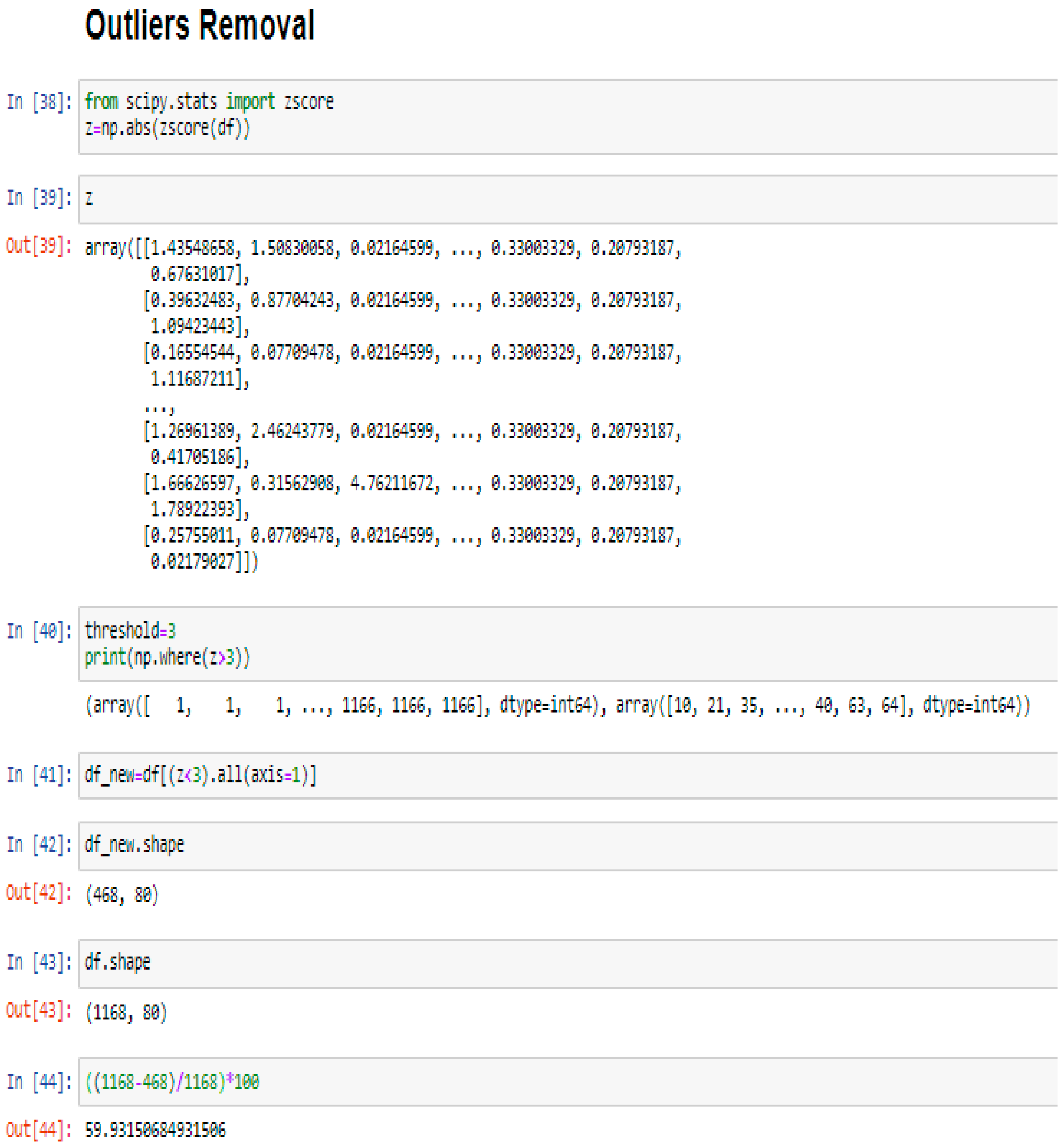
	Id	MSSubClass	LotFrontage	LotArea	OverallQual	OverallCond	YearBuilt	YearRemodAdd	MasVnrArea	BsmtFinSF1	...	WoodDeck
count	1168.000000	1168.000000	954.00000	1168.000000	1168.000000	1168.000000	1168.000000	1168.000000	1161.000000	1168.000000	...	1168.0000
mean	724.136130	56.767979	70.98847	10484.749144	6.104452	5.595890	1970.930651	1984.758562	102.310078	444.726027	...	96.2063
std	416.159877	41.940650	24.82875	8957.442311	1.390153	1.124343	30.145255	20.785185	182.595606	462.664785	...	126.1589
min	1.000000	20.000000	21.00000	1300.000000	1.000000	1.000000	1875.000000	1950.000000	0.000000	0.000000	...	0.0000
25%	360.500000	20.000000	60.00000	7621.500000	5.000000	5.000000	1954.000000	1966.000000	0.000000	0.000000	...	0.0000
50%	714.500000	50.000000	70.00000	9522.500000	6.000000	5.000000	1972.000000	1993.000000	0.000000	385.500000	...	0.0000
75%	1079.500000	70.000000	80.00000	11515.500000	7.000000	6.000000	2000.000000	2004.000000	160.000000	714.500000	...	171.0000
max	1460.000000	190.000000	313.00000	164660.000000	10.000000	9.000000	2010.000000	2010.000000	1600.000000	5644.000000	...	857.0000

8 rows × 38 columns

```
In [32]: plt.figure(figsize=(15,13))
         sns.heatmap(dfcorr,cmap='YlGnBu',annot=True)
Out[32]: <matplotlib.axes._subplots.AxesSubplot at 0x276019ebb48>
```



Removing the Outliers using Z-score



Outliers Removal

```
In [38]: from scipy.stats import zscore
         z=np.abs(zscore(df))
```

```
In [39]: z
```

```
Out[39]: array([[1.43548658, 1.50830058, 0.02164599, ..., 0.33003329, 0.20793187,
                 0.67631017],
                [0.39632483, 0.87704243, 0.02164599, ..., 0.33003329, 0.20793187,
                 1.09423443],
                [0.16554544, 0.07709478, 0.02164599, ..., 0.33003329, 0.20793187,
                 1.11687211],
                ...,
                [1.26961389, 2.46243779, 0.02164599, ..., 0.33003329, 0.20793187,
                 0.41705186],
                [1.66626597, 0.31562908, 4.76211672, ..., 0.33003329, 0.20793187,
                 1.78922393],
                [0.25755011, 0.07709478, 0.02164599, ..., 0.33003329, 0.20793187,
                 0.02179027]])
```

```
In [40]: threshold=3
         print(np.where(z>3))

         (array([   1,    1,    1, ..., 1166, 1166, 1166], dtype=int64), array([10, 21, 35, ..., 40, 63, 64], dtype=int64))
```

```
In [41]: df_new=df[(z<3).all(axis=1)]
```

```
In [42]: df_new.shape
```

```
Out[42]: (468, 80)
```

```
In [43]: df.shape
```

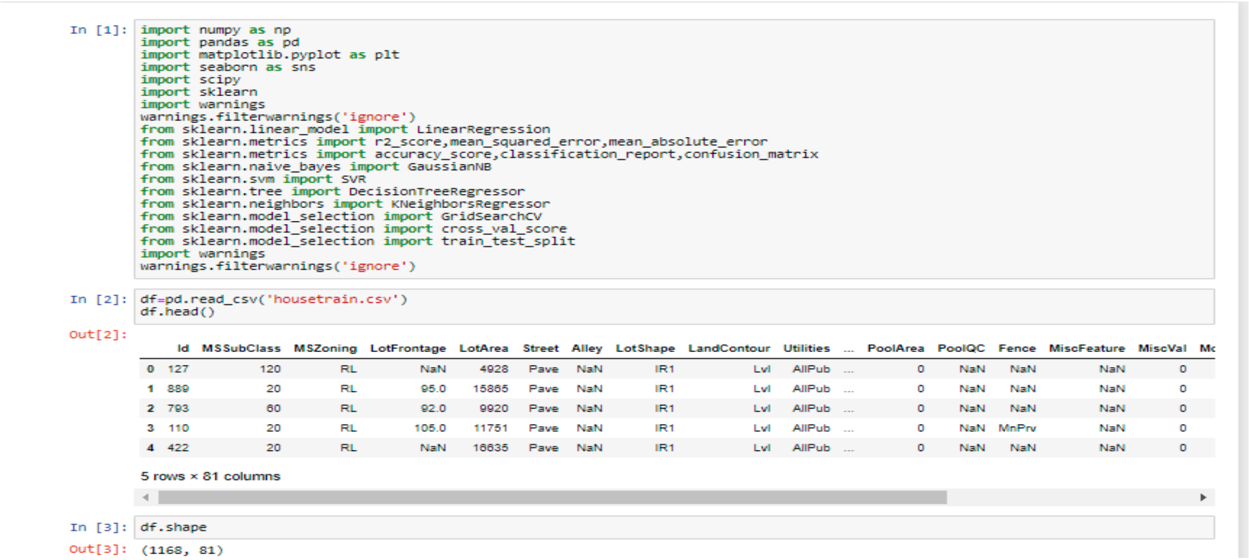
```
Out[43]: (1168, 80)
```

```
In [44]: ((1168-468)/1168)*100
```

```
Out[44]: 59.93150684931506
```

- <u>Data Sources and their formats</u>

The sample data is provided to us from our client database. It is provided in csv format and hence we import it using pandas. Then we further checked more about data using info, checked data types using dtypes, shapes using .shape, columns using .columns, null values using .isnull.sum, and further visualize it through heatmap as follows:

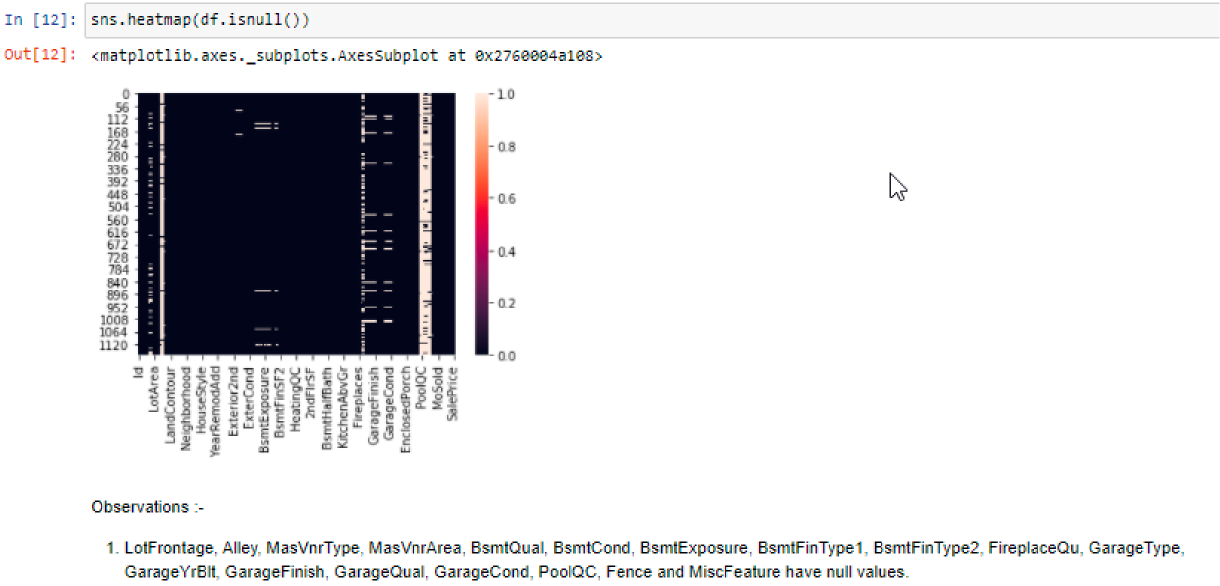


```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import scipy
        import sklearn
        import warnings
        warnings.filterwarnings('ignore')
        from sklearn.linear_model import LinearRegression
        from sklearn.metrics import r2_score,mean_squared_error,mean_absolute_error
        from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
        from sklearn.naive_bayes import GaussianNB
        from sklearn.svm import SVR
        from sklearn.tree import DecisionTreeRegressor
        from sklearn.neighbors import KNeighborsRegressor
        from sklearn.model_selection import GridSearchCV
        from sklearn.model_selection import cross_val_score
        from sklearn.model_selection import train_test_split
        import warnings
        warnings.filterwarnings('ignore')

In [2]: df=pd.read_csv('housetrain.csv')
        df.head()

Out[2]:
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	Mc
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0	
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0	

5 rows × 81 columns

```
In [3]: df.shape
Out[3]: (1168, 81)
```



```
In [12]: sns.heatmap(df.isnull())
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x2760004a108>
```

Observations :-

1. LotFrontage, Alley, MasVnrType, MasVnrArea, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond, PoolQC, Fence and MiscFeature have null values.

• Data Preprocessing Done

First we will determine whether there are any null values and since there were null values as well as NaN vales present in the dataset we proceeded further by imputing them using Simple Imputer with mean and most frequent as strategies respectively. Next we did Label encoding using label encoder. Then we performed some data visualization in which we observed certain attributes were having skewness and outliers that were plotted using distplot and boxplot. Outliers were removed with the help of Zscore in which 685 rows were removed.

• Data Inputs- Logic- Output Relationships

The data consists of 80 inputs and one output-“SalePrice”. MSSubClass,OverallCond,KitchenAbvGr,EnclosedPorch and Yr Sold are the least/negatively correlated column with target('SalePrice') variable. OverallQual is highly correlated column with target variable followed by GrLivArea and other attributes.

• Hardware and Software Requirements and Tools Used

In this project we have used HP Pavilion PC with 64-bit operating system and have Windows 10 pro. We have used python to develop this project in which we have used various libraries such as numpy, pandas, matplotlib, seaborn for handling data or arrays and their visualization. For statistical purpose we have used zscore from scipy.stats to remove outliers. Lastly, to develop the model we have used various libraries and metrics from sklearn such as train_test_split, Linear Regression, Lasso, Ridge, Elastic Net, SVR, Decision Tree Regressor, KNeighbors Regressor, Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor, mean_squared_error, mean_absolute_error and r2_score.

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)

  We have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them. We have used distplot to find the distribution of all attributes.

- ## Testing of Identified Approaches (Algorithms)

  We have used following algorithms such as: LinearRegression, Lasso, Ridge, ElasticNet, SVR, DecisionTreeRegressor, KNeighborsRegressor, RandomForestRegressor, AdaBoostRegressor and GradientBoostingRegressor.

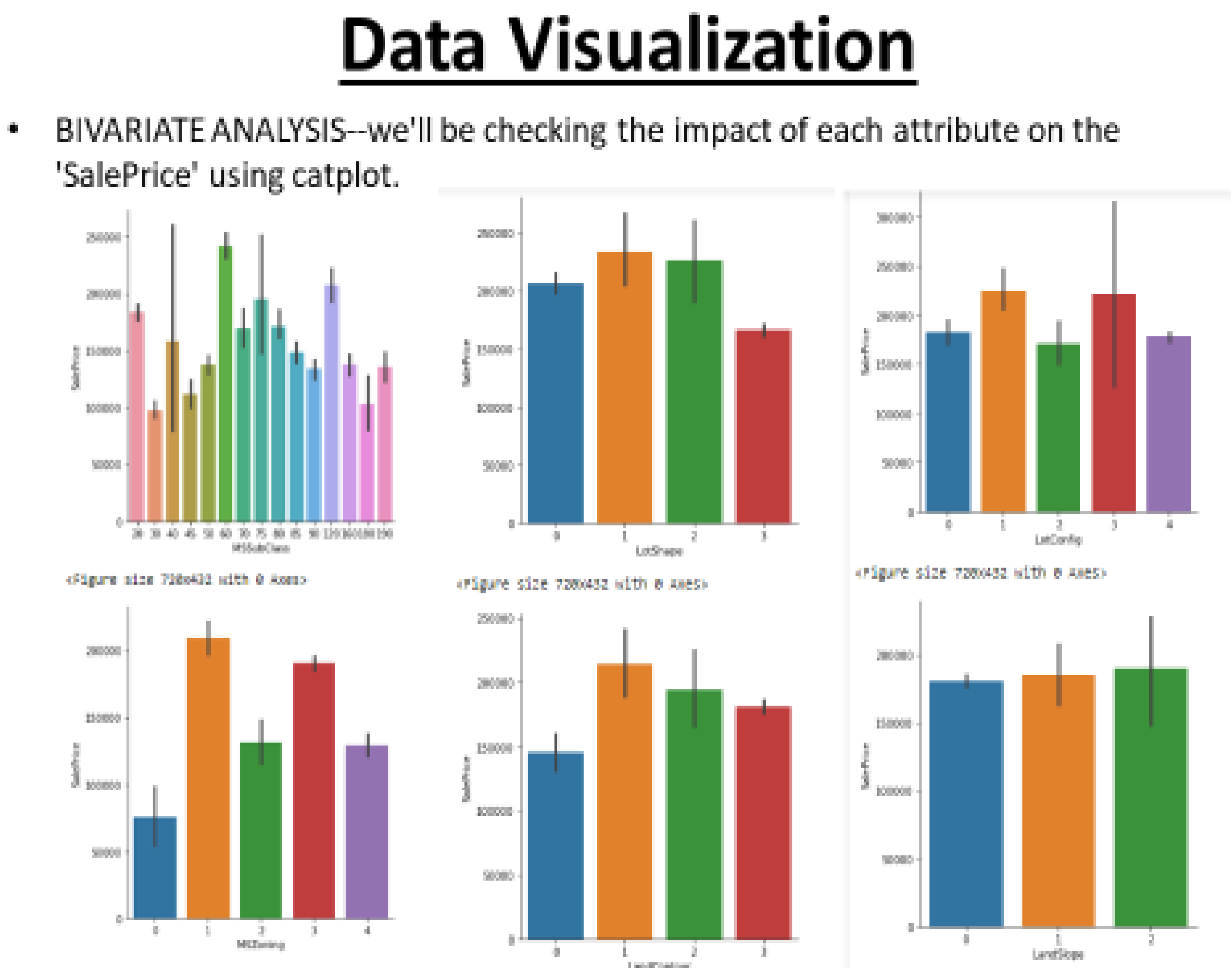
- ## Run and Evaluate selected models

  We have formed a loop where all the algorithms will be used one by one and their corresponding Score, Mean Absolute Error, Mean Squared Error, RMSE and r2_score will be evaluated.
  • I chose GradientBoostingRegressor as our best model since it's giving us best score and it's performing well. It's r2_score is also satisfactory and it shows that our model is neither underfitting/overfitting. Then we performed hyperparamter tuning using GridSearchCV on GradientBoostingRegressor from which got 'learning_rate': 0.1, 'n_estimators': 500 as best parameters. We got score : 0.999517991577412 after performing hyperparameter tuning and earlier it was 0.9846658425719441. Its r2_score is also satisfactory.

  Hence we saved GradientBoostingRegressor as our final model using joblib.

- <u>Key Metrics for success in solving problem under consideration</u>

  Key metrics used for finalising the model was Score and r2_score. Since in case of GradientBoostingRegressor it's giving us good score among all other models and it's performing well. It's r2_score is also satisfactory and it shows that our model is neither underfitting/overfitting .



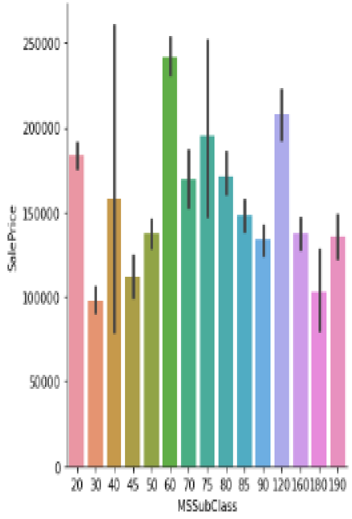
Data Visualization

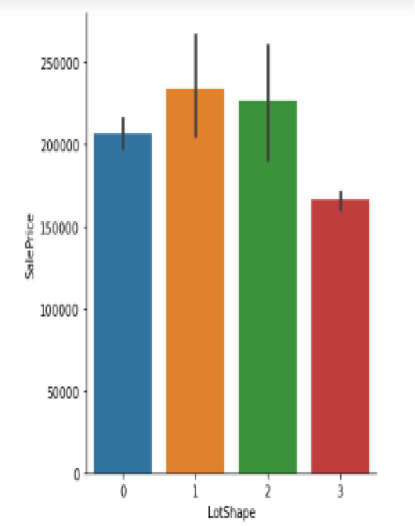
- BIVARIATE ANALYSIS--we'll be checking the impact of each attribute on the 'SalePrice' using catplot.

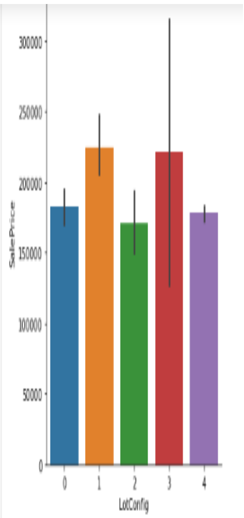
```
<Figure size 720x432 with 0 Axes>
```

```
<Figure size 720x432 with 0 Axes>
```

```
<Figure size 720x432 with 0 Axes>
```


SalePrice
0
50000
100000
150000
200000
250000
20 30 40 45 50 60 70 75 80 85 90 120 160 180 190
MSSubClass

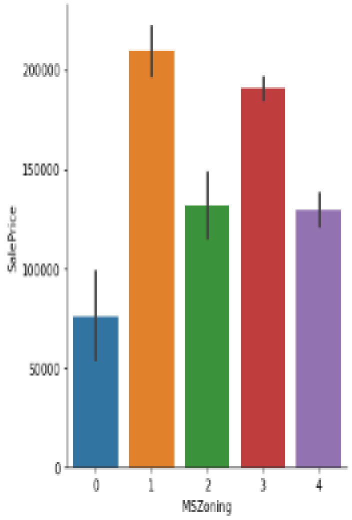

SalePrice
0
50000
100000
150000
200000
250000
0 1 2 3
LotShape

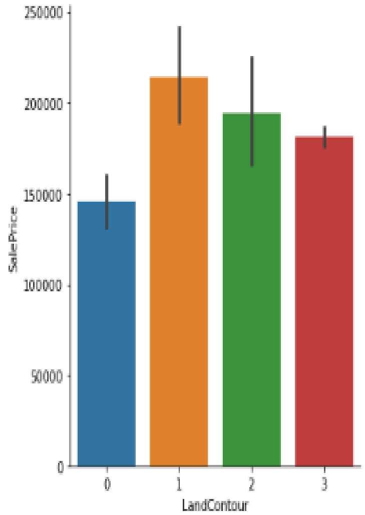

SalePrice
0
50000
100000
150000
200000
250000
300000
0 1 2 3 4
LotConfig

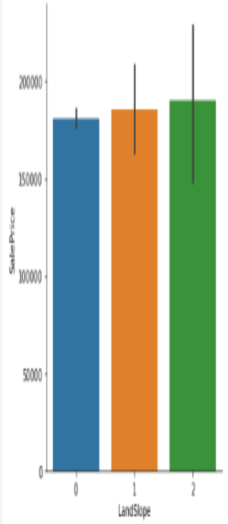
```
<Figure size 720x432 with 0 Axes>
```

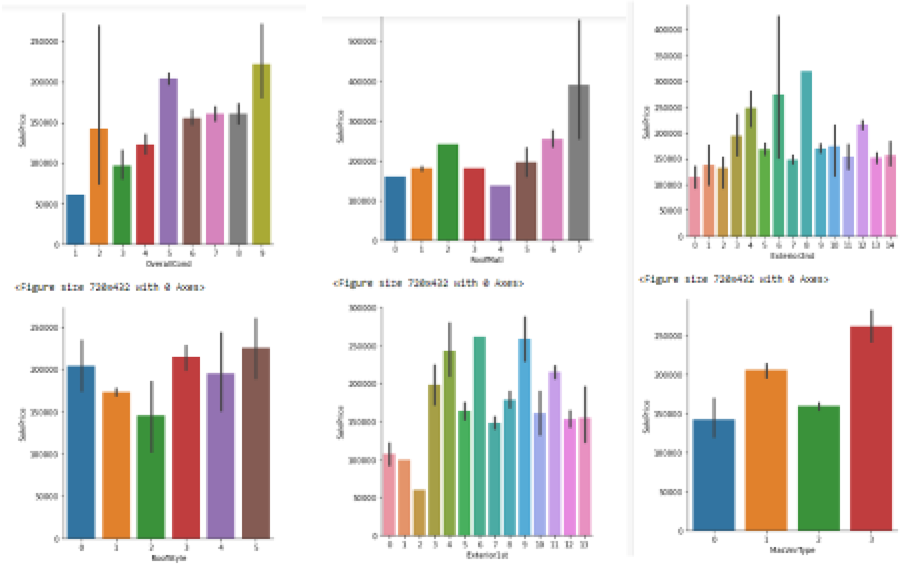
```
<Figure size 720x432 with 0 Axes>
```

```
<Figure size 720x432 with 0 Axes>
```

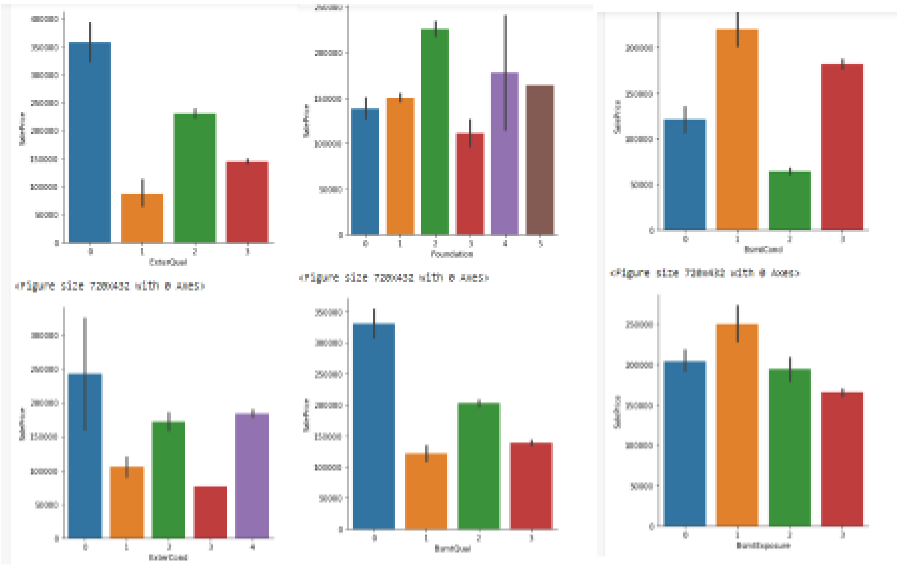

SalePrice
0
50000
100000
150000
200000
0 1 2 3 4
MSZoning


SalePrice
0
50000
100000
150000
200000
250000
0 1 2 3
LandContour

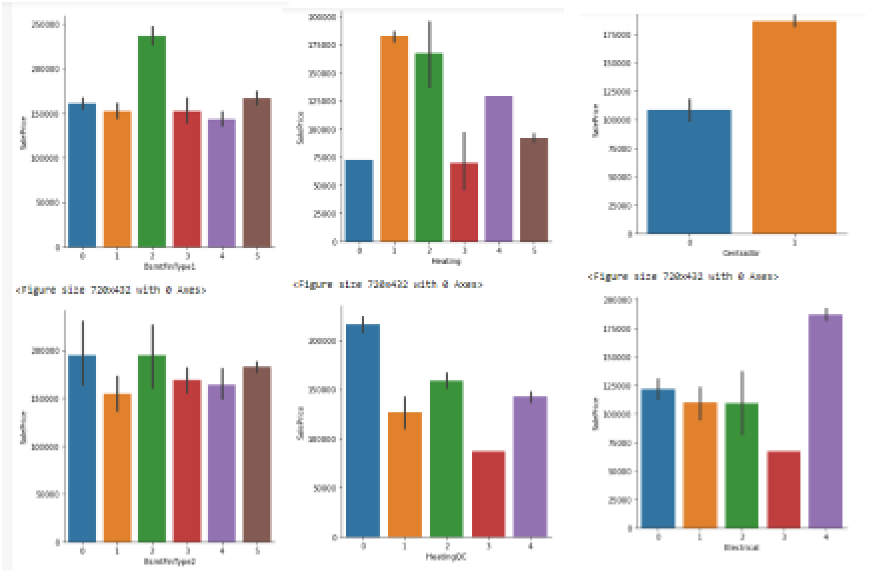

SalePrice
0
50000
100000
150000
200000
0 1 2
LandSlope


SalePrice
OverallCond
RoofMatl
Exterior2nd
RoofStyle
Exterior1st
MasVnrType

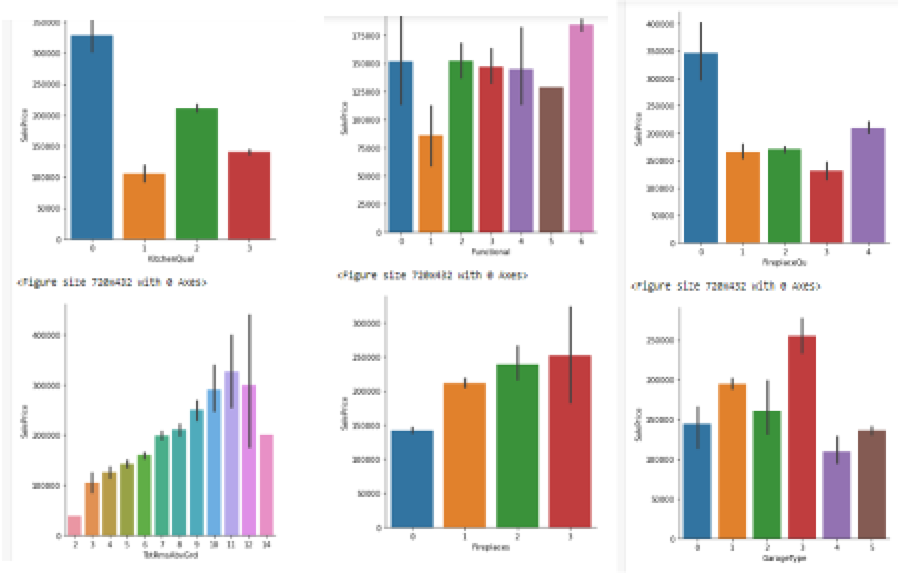
```
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
```


SalePrice
ExterQual
Foundation
BsmtCond
ExterCond
BsmtQual
BsmtExposure

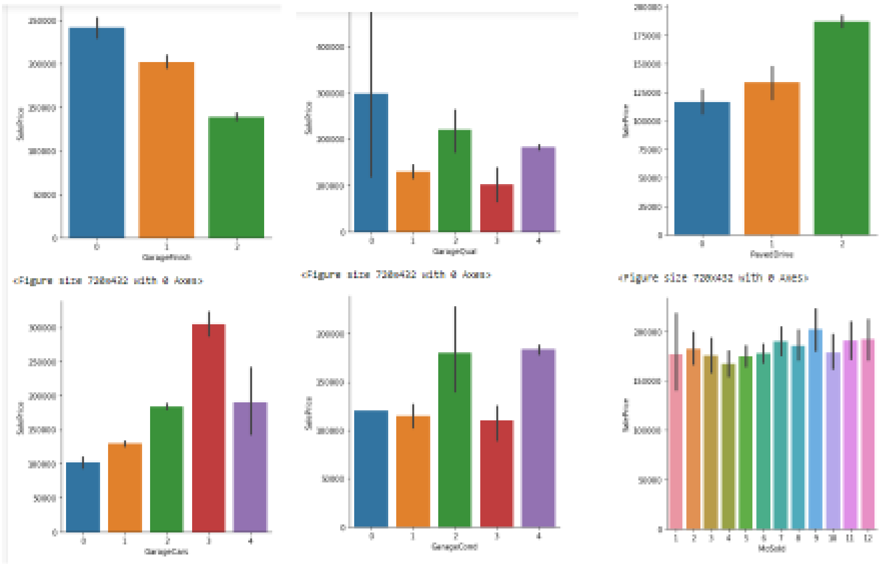
```
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
```


SalePrice
BsmtFinType1
Heating
CentralAir
BsmtFinType2
HeatingQC
Electrical

```
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
```

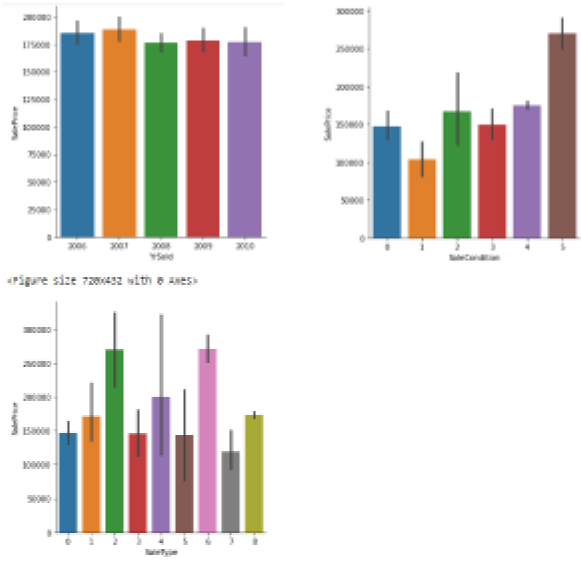

SalePrice
KitchenQual
Functional
FireplaceQu
TotRmsAbvGrd
Fireplaces
GarageType

```
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
```



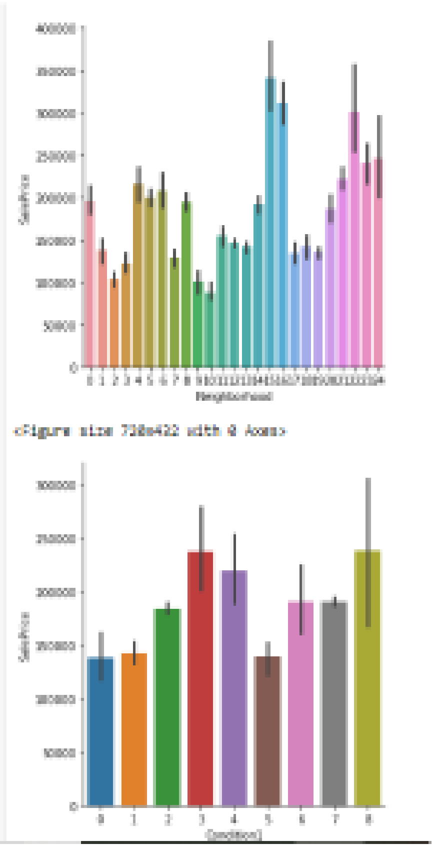
```
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
<Figure size 720x432 with 0 Axes>
```

SalePrice
250000
200000
150000
100000
50000
0
0
1
2
GarageFinish
SalePrice
400000
300000
200000
100000
0
0
1
2
3
4
GarageQual
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
2
PavedDrive
SalePrice
300000
250000
200000
150000
100000
50000
0
0
1
2
3
4
GarageCars
SalePrice
200000
150000
100000
50000
0
0
1
2
3
4
GarageCond
SalePrice
200000
150000
100000
50000
0
1
2
3
4
5
6
7
8
9
10
11
12
MoSold

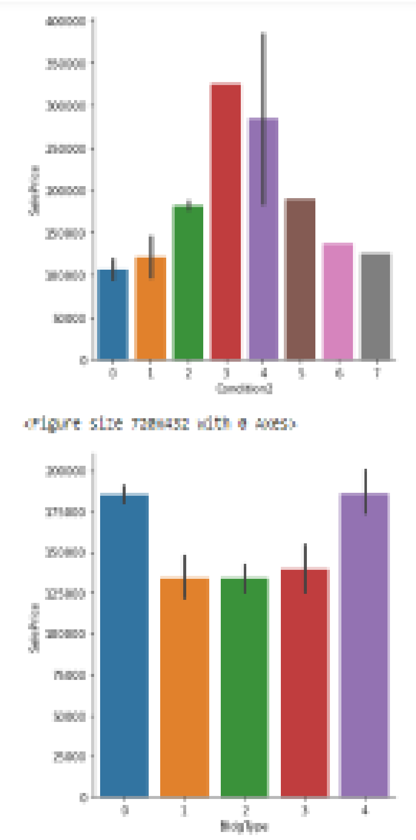


```
<Figure size 720x432 with 0 Axes>
```

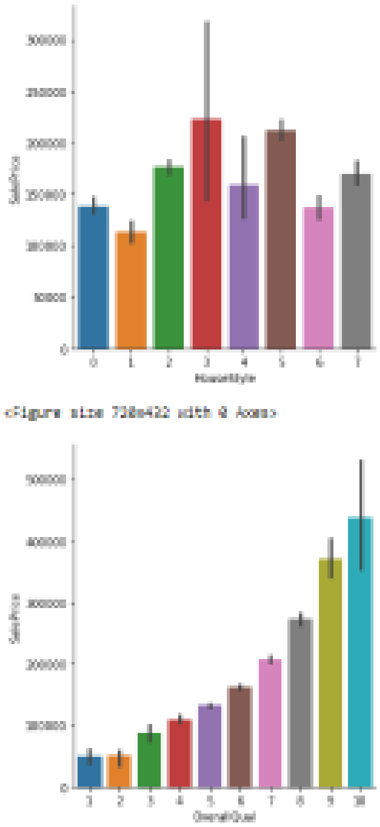
SalePrice
200000
175000
150000
125000
100000
75000
50000
25000
0
2006
2007
2008
2009
2010
YrSold
SalePrice
300000
250000
200000
150000
100000
50000
0
0
1
2
3
4
5
SaleCondition
SalePrice
300000
250000
200000
150000
100000
50000
0
0
1
2
3
4
5
6
7
8
SaleType


SalePrice
0
50000
100000
150000
200000
250000
300000
350000
400000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Neighborhood
SalePrice
0
50000
100000
150000
200000
250000
300000
0 1 2 3 4 5 6 7 8
Condition1

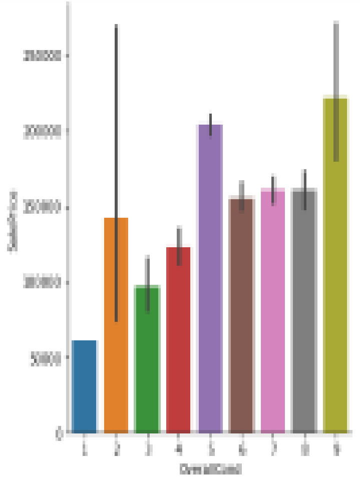
<Figure size 720x432 with 0 Axes>


SalePrice
0
50000
100000
150000
200000
250000
300000
350000
400000
0 1 2 3 4 5 6 7
Condition2
SalePrice
0
25000
50000
75000
100000
125000
150000
175000
200000
0 1 2 3 4
BldgType

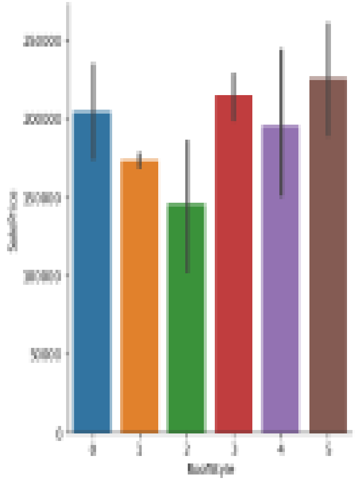
<Figure size 720x432 with 0 Axes>

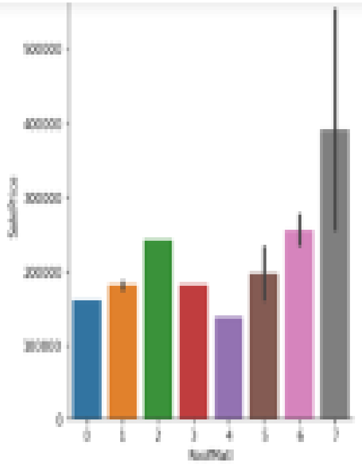

SalePrice
0
50000
100000
150000
200000
250000
300000
0 1 2 3 4 5 6 7
HouseStyle
SalePrice
0
100000
200000
300000
400000
500000
1 2 3 4 5 6 7 8 9 10
OverallQual

<Figure size 720x432 with 0 Axes>

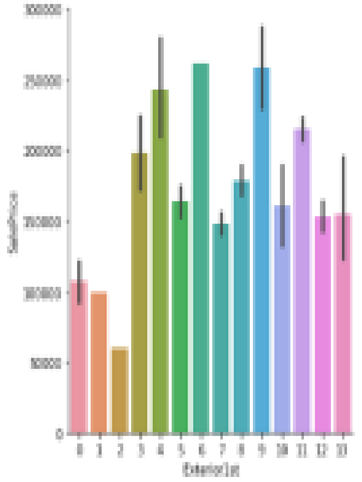

SalePrice
0
50000
100000
150000
200000
250000
1
2
3
4
5
6
7
8
9
OverallCond

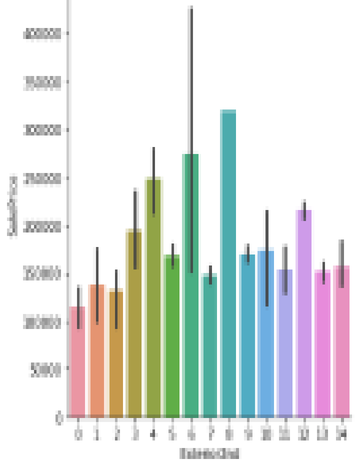
```
<Figure size 720x432 with 0 Axes>
```


SalePrice
0
50000
100000
150000
200000
250000
0
1
2
3
4
5
RoofStyle

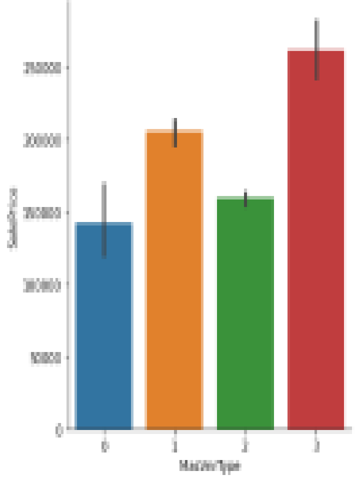

SalePrice
0
100000
200000
300000
400000
500000
0
1
2
3
4
5
6
7
RoofMatl

```
<Figure size 720x432 with 0 Axes>
```


SalePrice
0
50000
100000
150000
200000
250000
300000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Exterior1st


SalePrice
0
50000
100000
150000
200000
250000
300000
350000
400000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Exterior2nd

```
<Figure size 720x432 with 0 Axes>
```


SalePrice
0
50000
100000
150000
200000
250000
0
1
2
3
MasVnrType

Interpretation of the Results

- Least SalePrice is for 30:1-STORY 1945 & OLDER and maximum for 60:2-STORY 1946 & NEWER
- In MSZoing maximum is for category 1 i.e, Floating Village Residential
- Lotshape 1 and 2 have almost similar price and 3 has least.
- Landconotur corresponding to 1 i.e, HLS Hillside - Significant slope from side to side has maximum price.
- Lotconfig corresponding to 1 and 3 have similar price.
- Neighborhoot with (15)NPkVill Northpark Villa has maximum sales price and (10)IDOTRR Iowa DOT and Rail Road has least.
- Normal condition houses have highest saleprice
- 1Fam Single-family Detached and TwnhsI Townhouse Inside Unit have maximum saleprice.
- In HouseStyle category 3: 2Story Two story has max sale price.
- In OverallQual: SalePrice increase as Ratings increase.
- Similary for OverallCond 5 and 9 have max sale price
- In RoofStyle 5:Shed has maximum.
- In Exterior1st 6:HardBoard and 9:Other have Saleprice
- In Exterior2nd 8:MetalSd Metal Siding
- In MasVnrType, 3:stone has max saleprice and 0:BrkCmn Brick Common has least
- In ExterQual 0:Excellent has maximum price. Similary for ExterCond
- In Foundation 2:PConc Poured Contrete has max price
- In BsmtQual 0: Ex Excellent (100+ inches), In BsmtCond 1: Gd Good, In BsmtExposure 1: Av Average Exposure (split levels or foyers typically score average or above) have max sale prices
- In BsmtFinType1: Rating of basement finished area - 2:GLQ Good Living Quarters has max price
- In HeatingQC: Heating quality and condition 0:Ex Excellent has max price.
- Houses with CentralAir has higher saleprice
- In FireplaceQu: Fireplace quality 0:Ex Excellent - Exceptional Masonry Fireplace has max saleprice

- GarageType 3:BuiltIn Built-In (Garage part of house - typically has room above garage) has max saleprice
- Finished Garage has more price
- Paved Driveway has more price
- In 2007 maximum houses are sold followed by 2006
- In saletype category 2 and 6 have max sale price
- Normal sale condition has max price.

CONCLUSION

- Key Findings and Conclusions of the Study
- Lotshape 1 and 2 have almost similar price and 3 has least.
- Landconotur corresponding to 1 i.e, HLS Hillside - Significant slope from side to side has maximum price.
- Neighborhoot with (15)NPkVill Northpark Villa has maximum sales price and (10)IDOTRR Iowa DOT and Rail Road has least.
- Normal condition houses have highest saleprice
- 1Fam Single-family Detached and TwnhsI Townhouse Inside Unit have maximum saleprice.
- In HouseStyle category 3: 2Story Two story has max sale price.
- In RoofStyle 5:Shed has maximum.
- In Exterior1st 6:HardBoard and 9:Other have Saleprice
- In MasVnrType, 3:stone has max saleprice and 0:BrkCmn Brick Common has least
- Houses with CentralAir has higher saleprice
- GarageType 3:BuiltIn Built-In (Garage part of house - typically has room above garage) has max saleprice
- In 2007 maximum houses are sold followed by 2006
- In LotArea, initially the price keep on increasing as LotArea increases but after 70000 it becomes constant till 160000 and then drops.
- In MasVnrArea, at 1200 saleprice is maximum and then it drops drastically.

- For 1stFlrSF:first floor square feet till 2500 the price is increasing uniformly but after that it decreases and drops after 3000
- For 2ndFlrSF:Second floor square feet the price is increasing as the area increases.

Learning Outcomes of the Study in respect of Data Science

With the help of visualization tools such as matplotlib and seaborn we have visualized the impact of each attributes on our target variable. For cleaning the data and plotting outliers we have used distplot and boxplot and for removing outliers we have used zscore which is a statistical tool. At last we got GradientBoostingRegressor as our best model.

Limitations of this work and Scope for Future Work

The model is working well and we have performed hyperparameter tuning and we have concluded our project by choosing GradientBoostingRegressor as our best model.