FLIP ROBO

NAME OF THE PROJECT

HOUSING: PRICE PREDICTION

Submitted by:

Yogesh.C.Mudliar

ACKNOWLEDGMENT

It is a genuine pleasure to express my deep sense of thanks and gratitude to my guide, Ms.Khushboo Garg, for allowing me to work on this project. It was a great way to expose myself to the actual research environment.

I thank FLIPROBO for permitting me to work with them.

I take this opportunity to say heartfelt thanks to Dr. Deepika Sharma, VP-learning And development DataTraind for her overall dedication, devotion, and support towards me. I convey my sincere regards to all the DataTraind team thanks for supporting me during academic years of my post-graduation course in data science.

I express my profound sense of gratitude to my mentor Ms.Khushboo Garg, FLIPROBO for her guidance at every step of my research work.

Apart from the project, I learned a lot from her, she gave me valuable thought- "To think"; that I will benefit from, for a long time to come. I am indebted to her more than she knows.

INTRODUCTION

- ## Business Problem Framing

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.

- ## Conceptual Background of the Domain Problem

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.

- ## Review of Literature

The aim to work on this project is to find the variables are important to predict the price of variable and how do these variables describe the price of the house. I have done EDA (Exploratory data analysis) to learn about dataset and the problems. I used diferent types of library to find insights of the dataset like pandas, numphy, seaborn, and sklearn

- ## Motivation for the Problem Undertaken

Houses are one of the necessary needs 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.

Analytical Problem Framing

- ## Mathematical/ Analytical Modelling of the Problem

I done Extensive EDA has to be performed to gain relationships of important variable and price. I used linear regression (sklearn.linear_model.LinearRegression) analysis because it used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

- ## Data Sources and their formats

We have different types of important data which is used for predictions :

-MSZoning: Identifies the general zoning classification of the sale.

-Neighborhood: Physical locations within Ames city limits

- SaleType: Type of sale

- SaleCondition: Condition of sale

- BldgType: Type of dwelling

- HouseStyle: Style of dwelling

- Street: Type of road access to property

- OverallQual: Rates the overall material and finish of the house

- OverallCond: Rates the overall condition of the house

- ExterQual: Evaluates the quality of the material on the exterior

• Data Pre-processing Done

After collecting a data, the next step is to get it ready for analysis. This means cleaning, or ‘scrubbing’ it, and is crucial in making sure that you’re working with high quality data.

- **Removing unwanted data points**—extracting irrelevant observations that have no bearing on intended analysis.
- **Bringing structure to data**—fixing layout issues, which will help to map and manipulate this data more easily.
- **Filling in major gaps**—this data contains null values and I notice that important data are missing. Once we identified gaps, we can go about filling them.

During cleaning the data we have to do EDA(Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations). This helps identify initial trends and characteristics, and can even refine our hypothesis.

• Data Inputs- Logic- Output Relationships

There are some inputs which is important to find our outputs like the price of houses with the available independent variables.

-MSZoning: Identifies the general zoning classification of the sale.

-Neighborhood: Physical locations within Ames city limits

- SaleType: Type of sale

- SaleCondition: Condition of sale

- BldgType: Type of dwelling

- HouseStyle: Style of dwelling

- Street: Type of road access to property

- OverallQual: Rates the overall material and finish of the house

- OverallCond: Rates the overall condition of the house

- ExterQual: Evaluates the quality of the material on the exterior

- Hardware and Software Requirements and Tools Used

Software requirements:

numpy : library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

pandas : software library written for the Python programming language for data manipulation and analysis

sklearn: Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines

seaborn: Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.

matplotlib.pyplot : is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

sklearn.linear_model _ LinearRegression: the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable.

sklearn.metrics _ mean_squared_error,mean_absolute_error

sklearn.model_selection _ train_test_split: to get training and test sets

sklearn.impute import SimpleImputer: scikit-learn class which is helpful in handling the missing data in the predictive model dataset.

Sklearn_ preprocessing: provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.

sklearn.preprocessing import LabelEncoder: to normalize labels. It can also be used to transform non-numerical labels (as long as they are hashable and comparable) to numerical labels.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

After collecting a data, the next step is to get it ready for analysis. This means cleaning, or ‘scrubbing’ it, and is crucial in making sure that you’re working with high quality data.

- Removing unwanted data points—extracting irrelevant observations that have no bearing on intended analysis.
- Bringing structure to data—fixing layout issues, which will help to map and manipulate this data more easily.
- Filling in major gaps—This data contains null values and I notice that important data are missing. Once we identified gaps, we can go about filling them.

During cleaning the data we have to do EDA(Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations). This helps identify initial trends and characteristics, and can even refine our hypothesis.

I done Predictive analysis (Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyse current and historical facts to make predictions about future or otherwise unknown events)

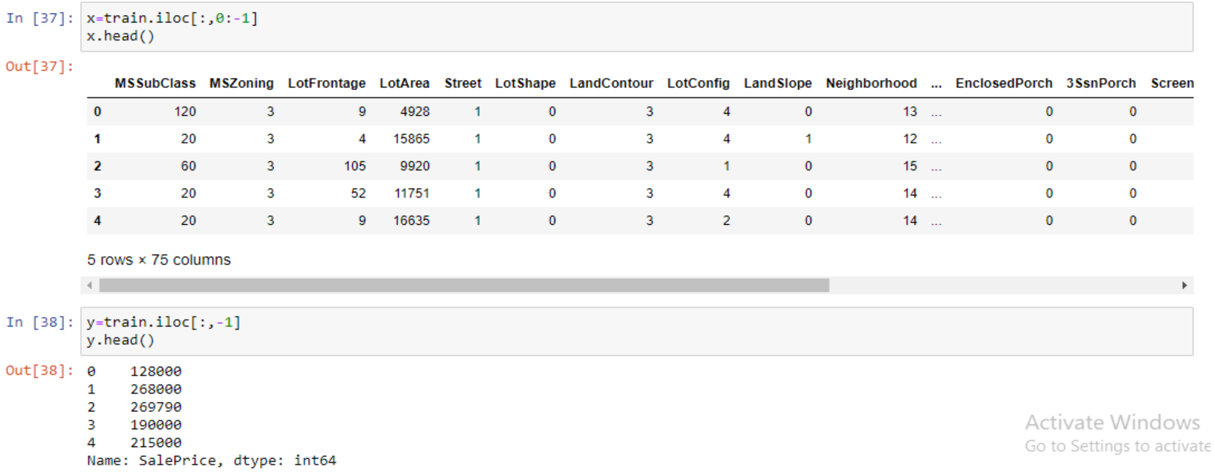
predictions about Which variables are important to predict the price of variable, and How do these variables describe the price of the house.

- # Testing of Identified Approaches (Algorithms)

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e. it finds the linear relationship between the dependent and independent variable.

- ## Run and Evaluate selected models

-First I lock dependent and independent variables for test and train



```
In [37]: x=train.iloc[:,0:-1]
         x.head()
Out[37]:
```

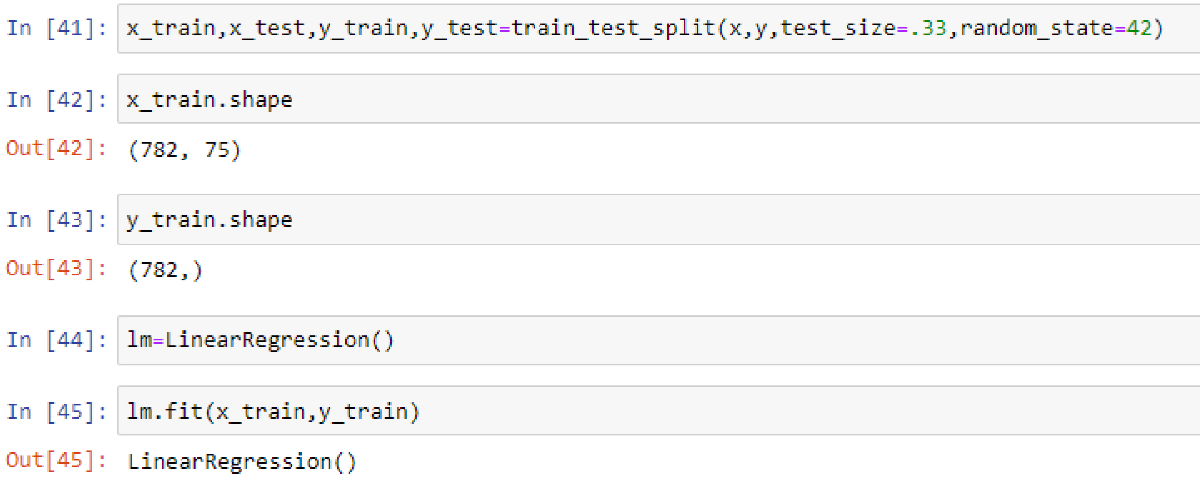
	MSSubClass	MSZoning	LotFrontage	LotArea	Street	LotShape	LandContour	LotConfig	LandSlope	Neighborhood	...	EnclosedPorch	3SsnPorch	Screen
0	120	3	9	4928	1	0	3	4	0	13	...	0	0	
1	20	3	4	15865	1	0	3	4	1	12	...	0	0	
2	60	3	105	9920	1	0	3	1	0	15	...	0	0	
3	20	3	52	11751	1	0	3	4	0	14	...	0	0	
4	20	3	9	16635	1	0	3	2	0	14	...	0	0	

5 rows × 75 columns

```
In [38]: y=train.iloc[:,-1]
         y.head()
Out[38]: 0    128000
         1    268000
         2    269790
         3    190000
         4    215000
         Name: SalePrice, dtype: int64
```

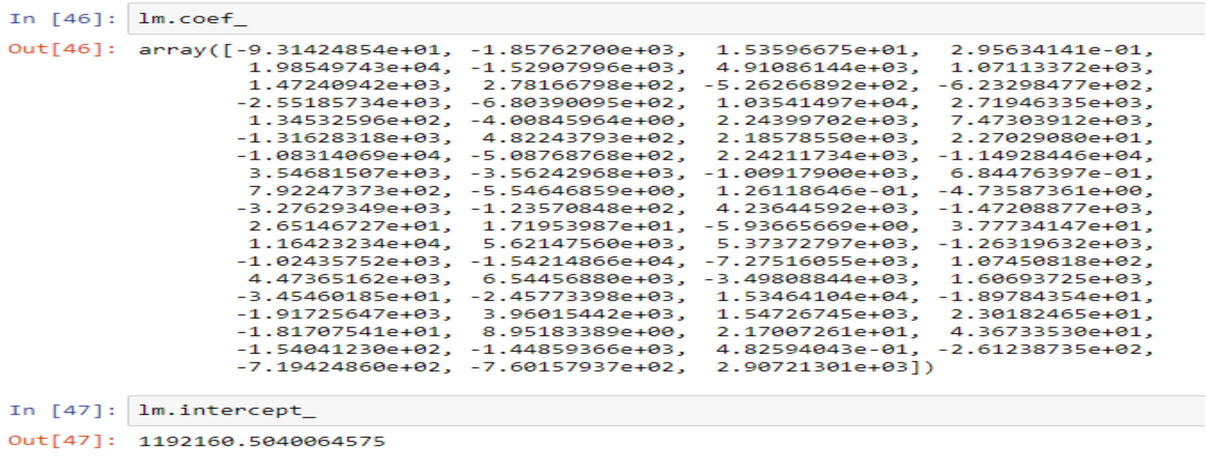
Activate Windows
Go to Settings to activate

-After that put all variables in training and testing



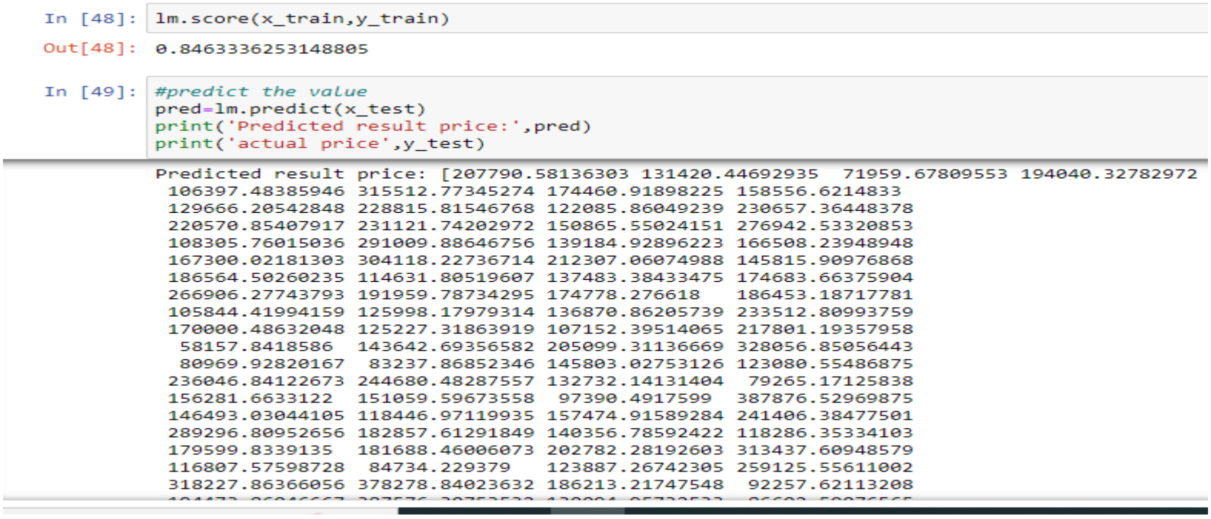
```
In [41]: x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.33,random_state=42)

In [42]: x_train.shape

Out[42]: (782, 75)

In [43]: y_train.shape

Out[43]: (782,)

In [44]: lm=LinearRegression()

In [45]: lm.fit(x_train,y_train)

Out[45]: LinearRegression()
```

-Finding the Coefficient and intercept



```
In [46]: lm.coef_
Out[46]: array([-9.31424854e+01, -1.85762700e+03,  1.53596675e+01,  2.95634141e-01,
                 1.98549743e+04, -1.52907996e+03,  4.91086144e+03,  1.07113372e+03,
                 1.47240942e+03,  2.78166798e+02, -5.26266892e+02, -6.23298477e+02,
                -2.55185734e+03, -6.80390095e+02,  1.03541497e+04,  2.71946335e+03,
                 1.34532596e+02, -4.00845964e+00,  2.24399702e+03,  7.47303912e+03,
                -1.31628318e+03,  4.82243793e+02,  2.18578550e+03,  2.27029080e+01,
                -1.08314069e+04, -5.08768768e+02,  2.24211734e+03, -1.14928446e+04,
                 3.54681507e+03, -3.56242968e+03, -1.00917900e+03,  6.84476397e-01,
                 7.92247373e+02, -5.54646859e+00,  1.26118646e-01, -4.73587361e+00,
                -3.27629349e+03, -1.23570848e+02,  4.23644592e+03, -1.47208877e+03,
                 2.65146727e+01,  1.71953987e+01, -5.93665669e+00,  3.77734147e+01,
                 1.16423234e+04,  5.62147560e+03,  5.37372797e+03, -1.26319632e+03,
                -1.02435752e+03, -1.54214866e+04, -7.27516055e+03,  1.07450818e+02,
                 4.47365162e+03,  6.54456880e+03, -3.49808844e+03,  1.60693725e+03,
                -3.45460185e+01, -2.45773398e+03,  1.53464104e+04, -1.89784354e+01,
                -1.91725647e+03,  3.96015442e+03,  1.54726745e+03,  2.30182465e+01,
                -1.81707541e+01,  8.95183389e+00,  2.17007261e+01,  4.36733530e+01,
                -1.54041230e+02, -1.44859366e+03,  4.82594043e-01, -2.61238735e+02,
                -7.19424860e+02, -7.60157937e+02,  2.90721301e+03])

In [47]: lm.intercept_
Out[47]: 1192160.5040064575
```

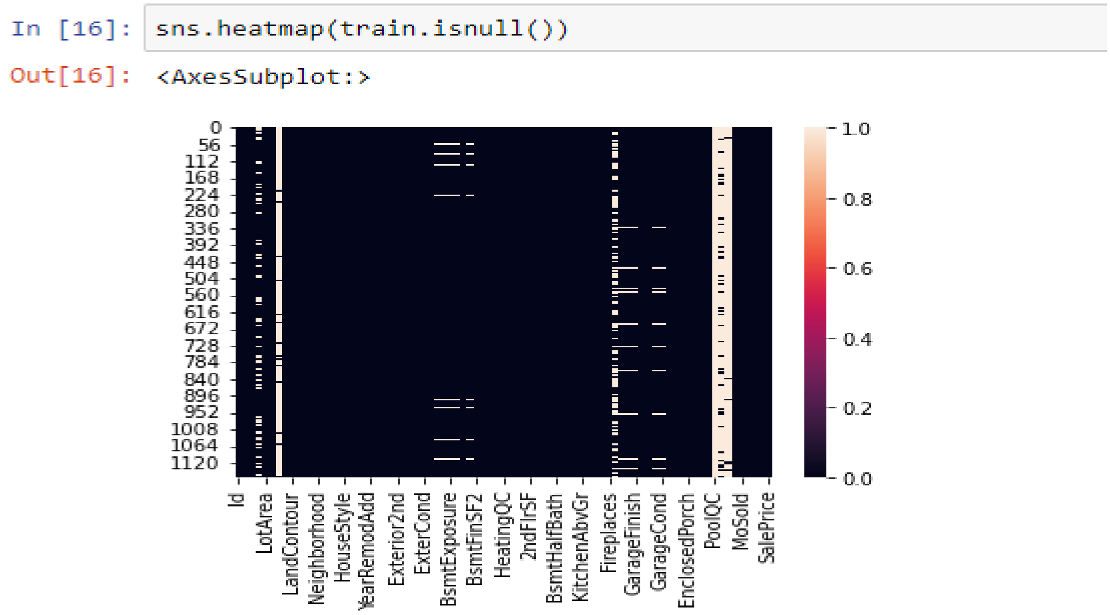
-Finding the prediction and scores of accuracy



```
In [48]: lm.score(x_train,y_train)
Out[48]: 0.8463336253148805

In [49]: #predict the value
         pred=lm.predict(x_test)
         print('Predicted result price:',pred)
         print('actual price',y_test)

         Predicted result price: [207790.58136303 131420.44692935  71959.67809553 194040.32782972
          106397.48385946 315512.77345274 174460.91898225 158556.6214833
          129666.20542848 228815.81546768 122085.86049239 230657.36448378
          220570.85407917 231121.74202972 150865.55024151 276942.53320853
          108305.76015036 291009.88646756 139184.92896223 166508.23948948
          167300.02181303 304118.22736714 212307.06074988 145815.90976868
          186564.50260235 114631.80519607 137483.38433475 174683.66375904
          266906.27743793 191959.78734295 174778.276618   186453.18717781
          105844.41994159 125998.17979314 136870.86205739 233512.80993759
          170000.48632048 125227.31863919 107152.39514065 217801.19357958
           58157.8418586  143642.69356582 205099.31136669 328056.85056443
           80969.92820167  83237.86852346 145803.02753126 123080.55486875
          236046.84122673 244680.48287557 132732.14131404  79265.17125838
          156281.6633122  151059.59673558  97390.4917599  387876.52969875
          146493.03044105 118446.97119935 157474.91589284 241406.38477501
          289296.80952656 182857.61291849 140356.78592422 118286.35334103
          179599.8339135  181688.46006073 202782.28192603 313437.60948579
          116807.57598728  84734.229379   123887.26742305 259125.55611002
          318227.86366056 378278.84023632 186213.21747548  92257.62113208
```

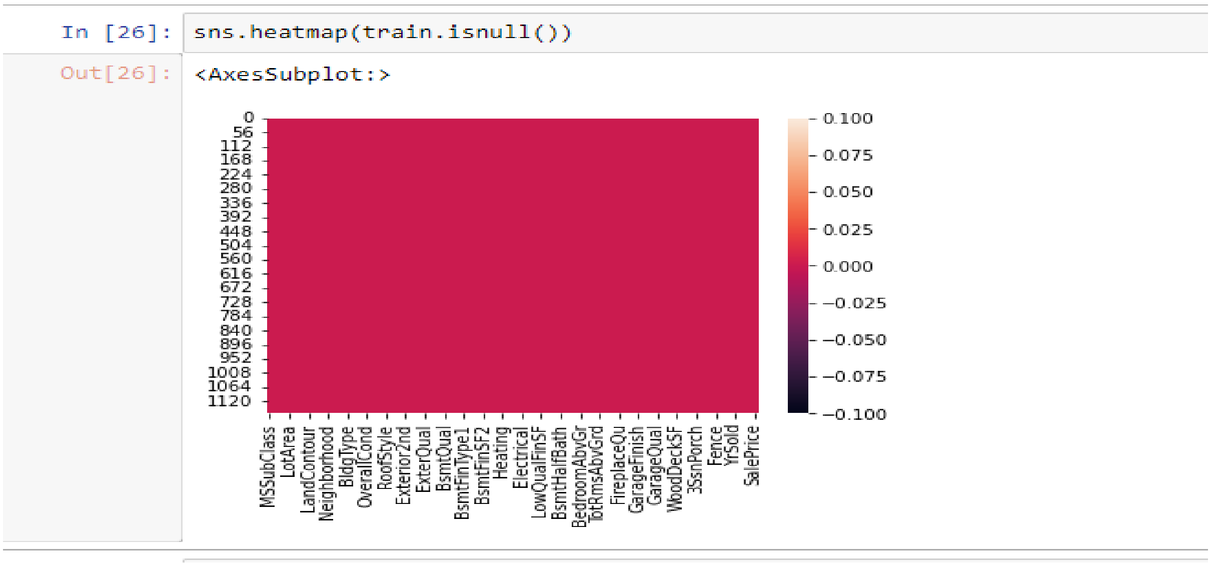
- Visualizations

-We can observe below given heat map before cleaning dataset



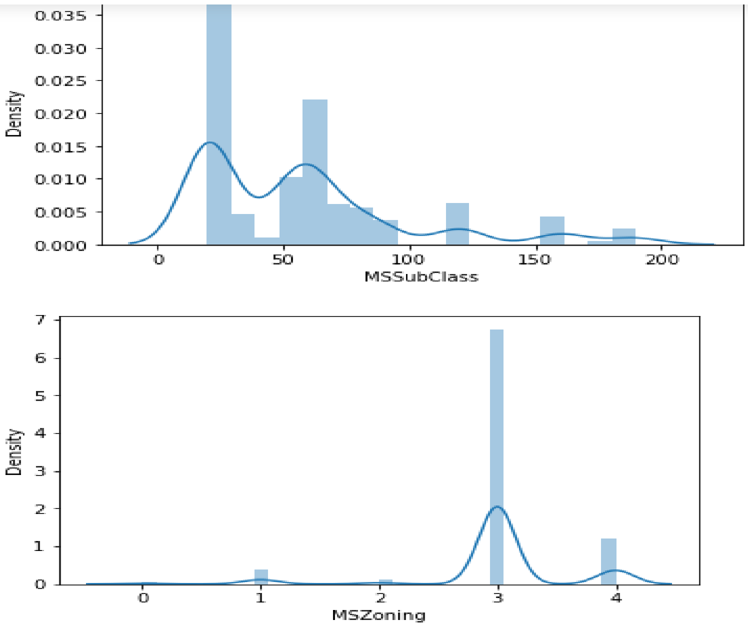
```
In [16]: sns.heatmap(train.isnull())
Out[16]: <AxesSubplot:>
```

-After cleaning ad removing null values the data set I found below mentioned heat map

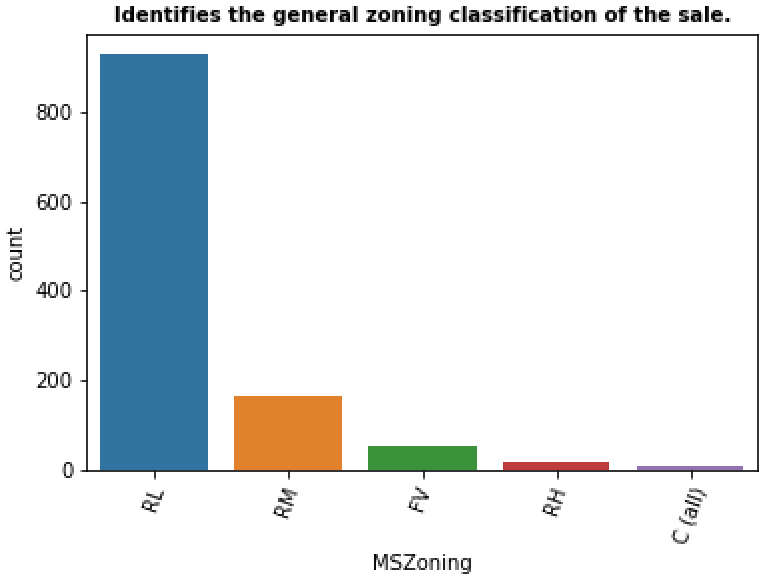


```
In [26]: sns.heatmap(train.isnull())
Out[26]: <AxesSubplot:>
```

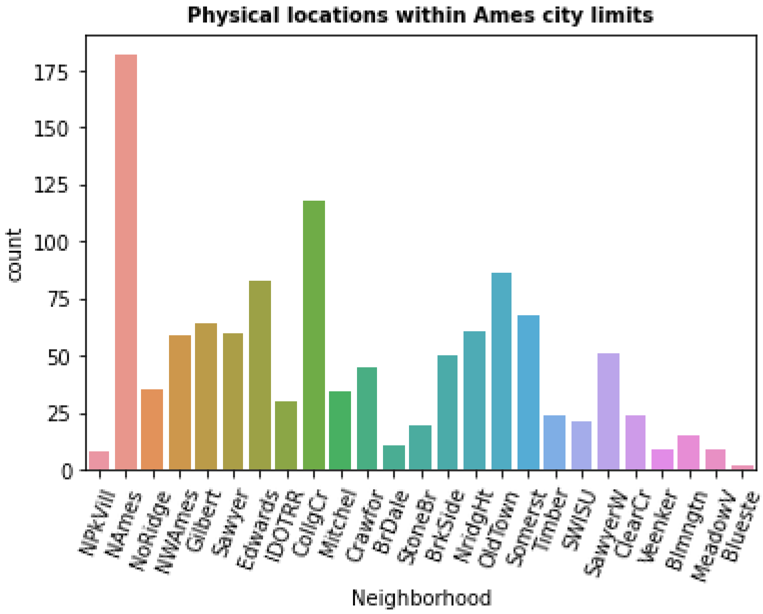
-I also plot distplot to find out the skewness of data



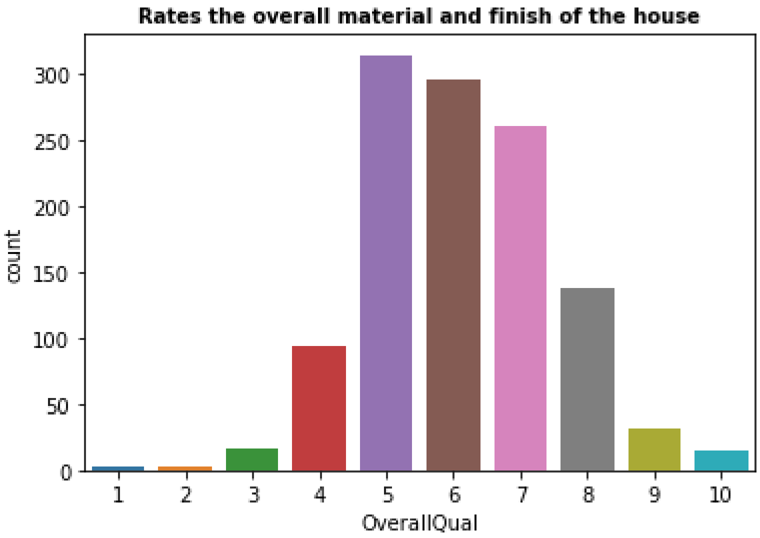
- Identifies the general zoning classification of the sale



- Physical locations within Ames city limits



- Rates the overall material and finish of the house

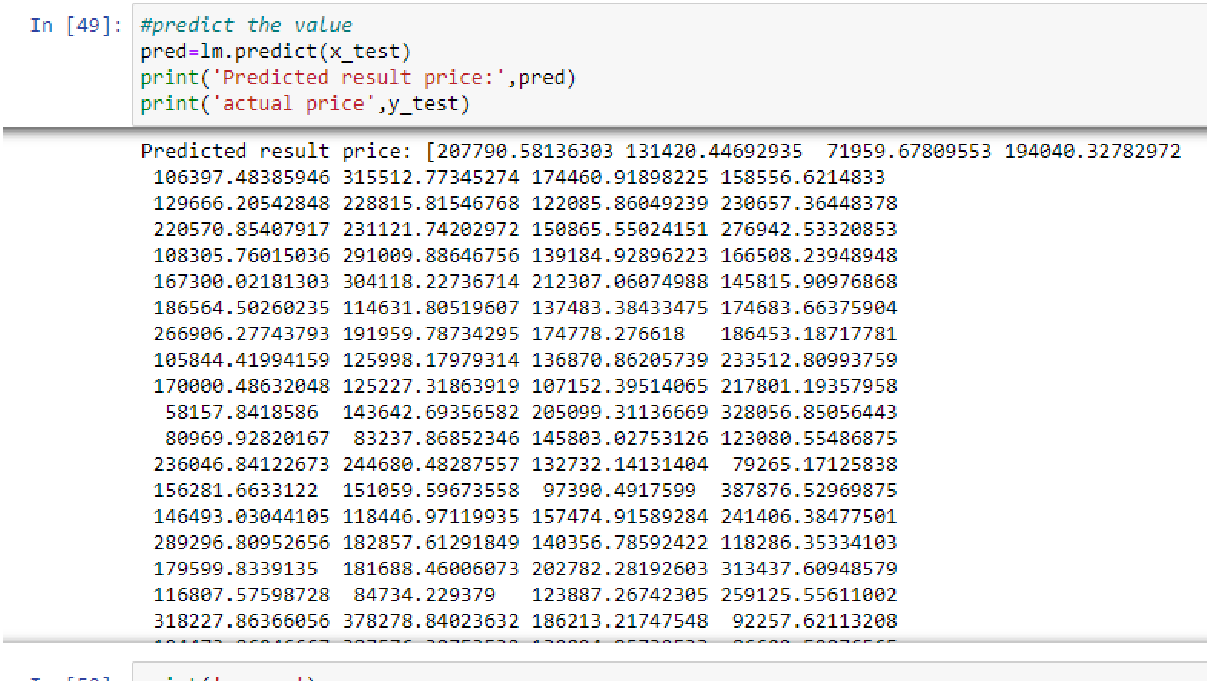


- Interpretation of the Results

-I find out the price of houses with the available independent variables

-I find out which variables are important to predict the price of variable by ploting countplot.

-these variables describe the price of the house



```
In [49]: #predict the value
         pred=lm.predict(x_test)
         print('Predicted result price:',pred)
         print('actual price',y_test)
```

```
Predicted result price: [207790.58136303 131420.44692935  71959.67809553 194040.32782972
 106397.48385946 315512.77345274 174460.91898225 158556.6214833
 129666.20542848 228815.81546768 122085.86049239 230657.36448378
 220570.85407917 231121.74202972 150865.55024151 276942.53320853
 108305.76015036 291009.88646756 139184.92896223 166508.23948948
 167300.02181303 304118.22736714 212307.06074988 145815.90976868
 186564.50260235 114631.80519607 137483.38433475 174683.66375904
 266906.27743793 191959.78734295 174778.276618   186453.18717781
 105844.41994159 125998.17979314 136870.86205739 233512.80993759
 170000.48632048 125227.31863919 107152.39514065 217801.19357958
  58157.8418586  143642.69356582 205099.31136669 328056.85056443
  80969.92820167  83237.86852346 145803.02753126 123080.55486875
 236046.84122673 244680.48287557 132732.14131404  79265.17125838
 156281.6633122  151059.59673558  97390.4917599  387876.52969875
 146493.03044105 118446.97119935 157474.91589284 241406.38477501
 289296.80952656 182857.61291849 140356.78592422 118286.35334103
 179599.8339135  181688.46006073 202782.28192603 313437.60948579
 116807.57598728  84734.229379   123887.26742305 259125.55611002
 318227.86366056 378278.84023632 186213.21747548  92257.62113208
```

CONCLUSION

Key Findings and Conclusions of the Study

-I find out the price of houses with the available independent variables

-I find out which variables are important to predict the price of variable by ploting countplot

- I fined which variables are important to predict the price of variable

- I fined how these variables describe the price of the house

Learning Outcomes of the Study in respect of Data Science

1. First I define the question and done EDA(Exploratory data analysis)
2. I collected all necessary dataset and drop all unnecessary data .
3. Cleaning the data
4. Analyzing the data

   -I done comparison between Sale Price & Lot Area

   - Identifies the general zoning classification of the sale.

   - Identified Rates the overall material and finish of the house
5. Sharing your results

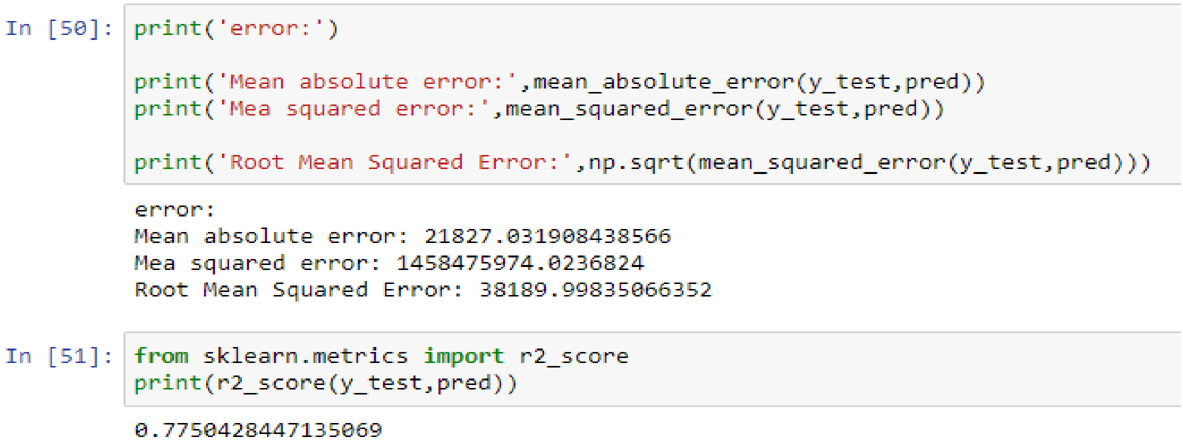
   -I find out the price of houses with the available independent variables

   -I find out which variables are important to predict the price of variable by ploting countplot

   -these variables describe the price of the house

- ## Limitations of this work and Scope for Future Work

There are still errors present in findings as mention below



```
In [50]: print('error:')

         print('Mean absolute error:',mean_absolute_error(y_test,pred))
         print('Mea squared error:',mean_squared_error(y_test,pred))

         print('Root Mean Squared Error:',np.sqrt(mean_squared_error(y_test,pred)))

         error:
         Mean absolute error: 21827.031908438566
         Mea squared error: 1458475974.0236824
         Root Mean Squared Error: 38189.99835066352

In [51]: from sklearn.metrics import r2_score
         print(r2_score(y_test,pred))

         0.7750428447135069
```

We can observe the **R2 score;** it is a very important metric that is used to evaluate the performance of a regression-based machine learning model. It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset.

We can improve the results by removing more outliers and regularization (making the resulting more regular).