NAME OF THE PROJECT

HOUSING PRICE PREDICTION MODEL

Submitted by: **AKSHAY RUNTHALA**

ACKNOWLEDGMENT

I took the help of various sources (mentioned below) to complete my project:-

- **SCIKIT-LEARN DOCUMENTATION(https://scikit-learn.org/stable/)**
- **NOTES PREPARED BY MYSELF RESOURCES PROVIDED BY MY TRAINING INSTITUTE(DATA TRAINED)**
- **PANDAS DOCUMENTATION(https://pandas.pydata.org/docs/)**
- **MATPLOTLIB DOCUMENTATION(https://matplotlib.org/)**
- **SEABORN DOCUMENTATION(https://seaborn.pydata.org/)
  POWER BI DOCUMENTATION**
- **TABLEAU RESOURCES**

INTRODUCTION

- **Business Problem Framing**
  Describe the business problem and how this problem can be related to the real world.

MY ANSWER-
The business problem I came across while preparing this project was that if any business/individual chooses to purchase a house residential,commercial or business purpose, he needs to take advice or domain knowledge about a house from a real estate appraiser regarding-

- What value can the house be purchased in the real estate market?
- What are the features which affect the value of the house most?
- What are the least important factors affecting the price of a house?
- If the business is buying house for selling at a later date--
  - What price should be set for the house to be sold?
  - How much does the inflation factor affect the sale price?

A **real estate appraiser** provides an objective and unbiased estimate or appraisal of the value of a property.
A real estate appraiser is a human being and though he is supposed to make an unbiased estimate for the value of the house or property,**it takes time for him to decide the best price for the house given the availability of data.A real estate appraiser's estimate varies person by person.**
I can say the accuracy of estimate and time taken to estimate the price/value of a property are the main limitations faced in today's world nowadays.The values predicted by the real estate appraiser in the end are an estimate and actual price of the house varies from

estimate.Many businesses/individuals rely on the estimates/decision of the real estate appraiser/agents to buy/or sell a particular property. Buying/selling a house/property or doing real estate businesses is a huge financial matter.If the appraiser's estimate regarding a house is even slightly wrong or varies,then it may cause huge financial loss to the businesses dealing in real estate. The individual choosing to purchase a house may suffer huge financial losses if he purchased a house at a higher cost than the reasonable price in the industry.
There is risk of financial loss in both selling and buying a house if the appraiser's estimate is incorrect.
There is a need for a pattern to estimate the houses' values as accurately as possible. Human beings' abilities are limited due to time factor as well as accuracy in estimation of values.
Machine learning tasks can easily overcome these limitations of human beings because they are exponentially faster than human appraisers to predict the value of a house based on the given data (house features).Most importantly,the machine learning models can predict the values of houses more accurately than humans.

Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

MY ANSWER-

- Historical cost of the house
- Quality of materials used in construction of the house.
- Data available regarding cost of materials used in the houses' construction (based on the country or region).

Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

MY ANSWER:--

I searched the websites for understanding the meaning of the terms and concepts used in the real estate industry.I needed to understand these concepts so that I can perform my machine learning tasks better.

- ## Motivation for the Problem Undertaken

Describe your objective behind making this project, this domain and what is the motivation behind.

MY ANSWER:--

My objective for making this project is to-

- Make a machine learning model that can most accurately predict the **SALE PRICES** of every house regarding which data is available to the company.
- To make this model as efficient as possible to predict the future houses based on data available to the company.
- Find out the **important variables** affecting the price of a single house.
- Develop my skills for data science further by undertaking this project. This is a great project which directly relates to the real world scenario. I was motivated by the various ideas that came when I first looked into this project.
- Test and upgrade my various python coding skills while working on my project.

- Test my tableau software and business intelligence skills.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

I calculated the skewness of the whole dataset to know whether the data is normal or not. I also used skewness to impute missing values in numerical columns of the dataset. If my data is not normal,It is better to impute missing numerical values with the median value and if it is normal,then it is advisable to impute missing values with mean.

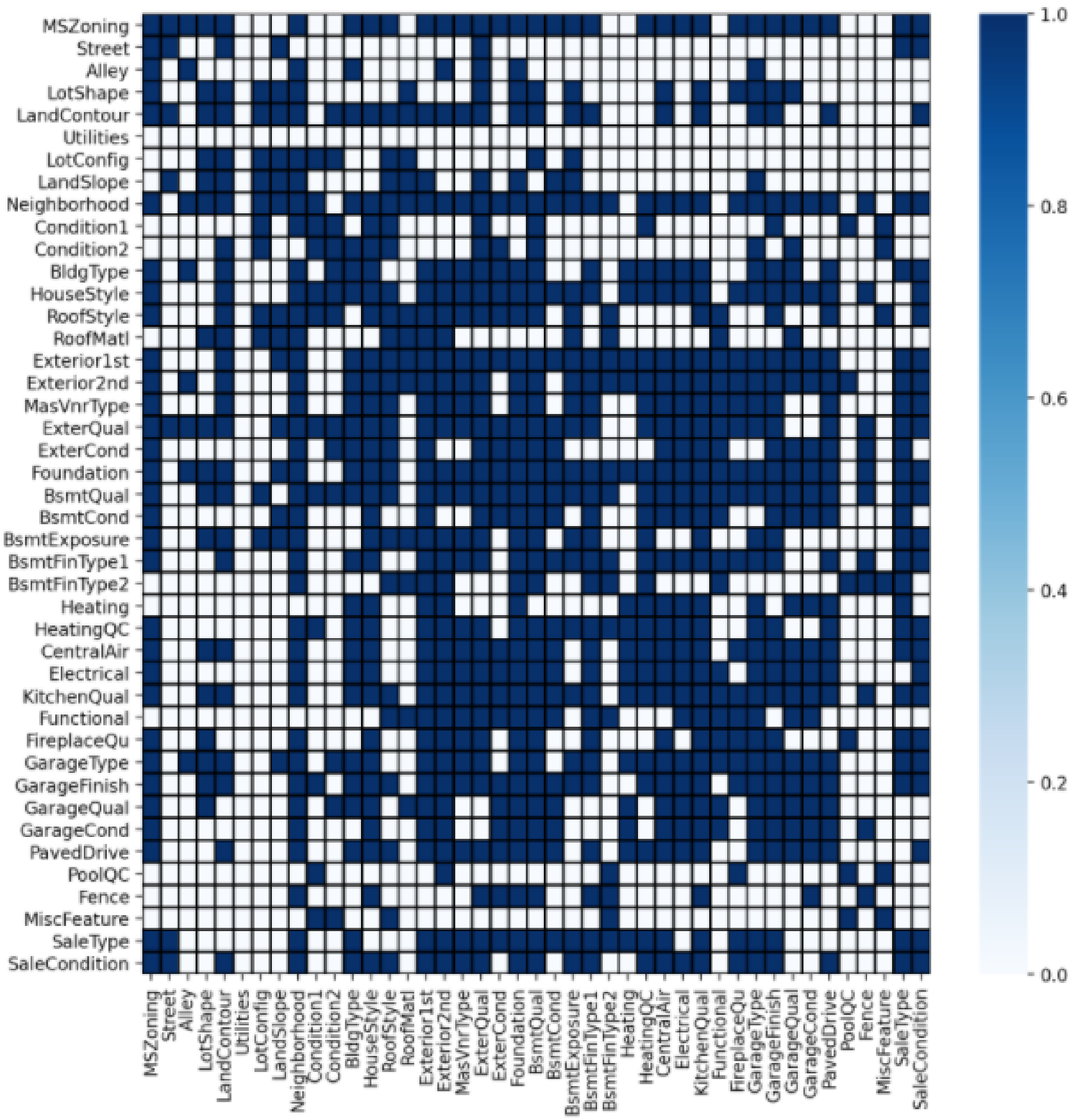
I imputed my categorical columns' missing values with the most frequent value.

SKEWNESS CHECK AND OBSERVATIONS:-

I Plotted Skewness For All Data Distributions Against Columns Of The Data. Plotted Green Horizontal Line For A Skewness Of Zero(Normal Distribution Has A Skewness Of 0). Plotted Red Lines Around The Green Line Denoting A Range Of (- 0.5,0.5).If Data Points Fall Within The Above Mentioned Range,Then They Are Approximately Normally Distributed . The rule for skewness seems to be: If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. If the skewness is between -1 and -0.5 (negatively skewed) or between 0.5 and 1 (positively skewed), the data are moderately skewed. Skewness of the normal distribution is zero.Most of the dataset columns seem to have a skewness of more than 0.5 (except for the label) and even more than 1,so they all are highly skewed.They all have a right-skewed data or a positively-skewed data curve since the right tail is longer and mass of the distribution is concentrated on the left of the figure.**I did not consider skewness for the categorical columns as well as the target label - SalePrice in my dataset.** Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y. A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables x and y.

I calculated the correlation between my target label (SALE PRICE) and other columns of the data to know which variables affect my Sale price the most.

I used chi square tests from scipy library to find relationships between the categorical variables of the data. Below is the heatmap for chi-square test results.

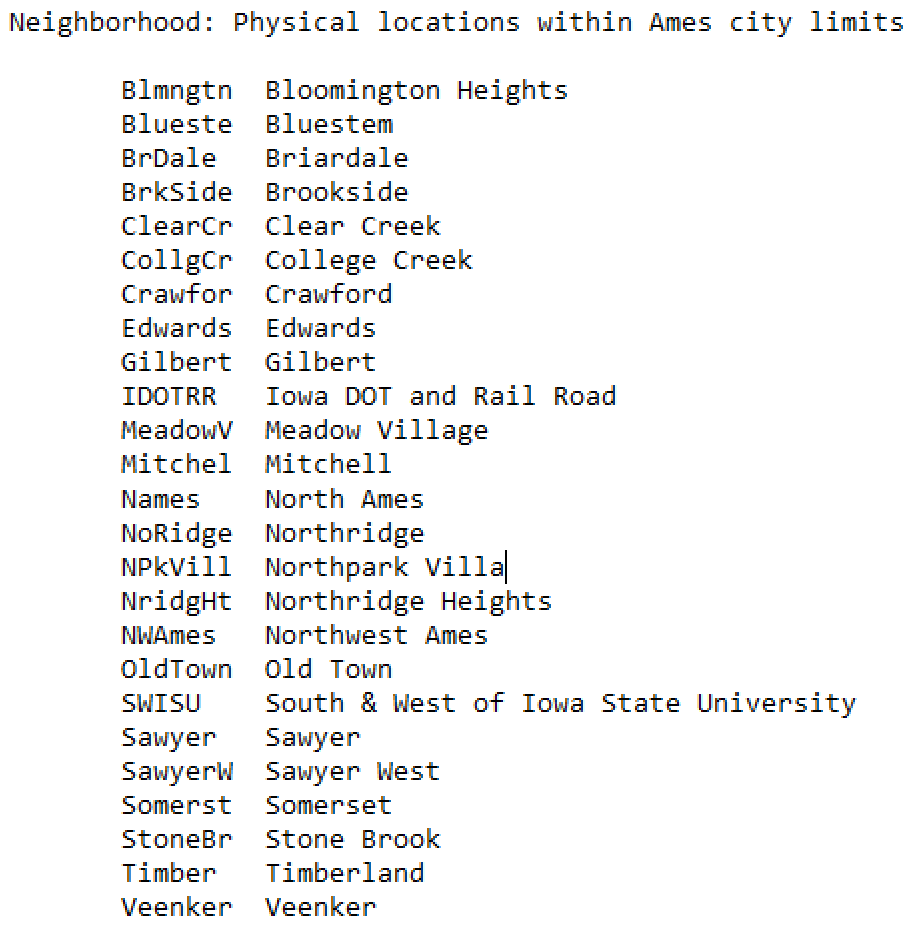


As shown in the above heatmap figure,the 'blue' colour shows the relationship status to be 'yes'or (1),whereas the 'white' colour shows the relationship status to be 'no'or (0) between the variables of the dataset.

- ## Data Sources and their formats

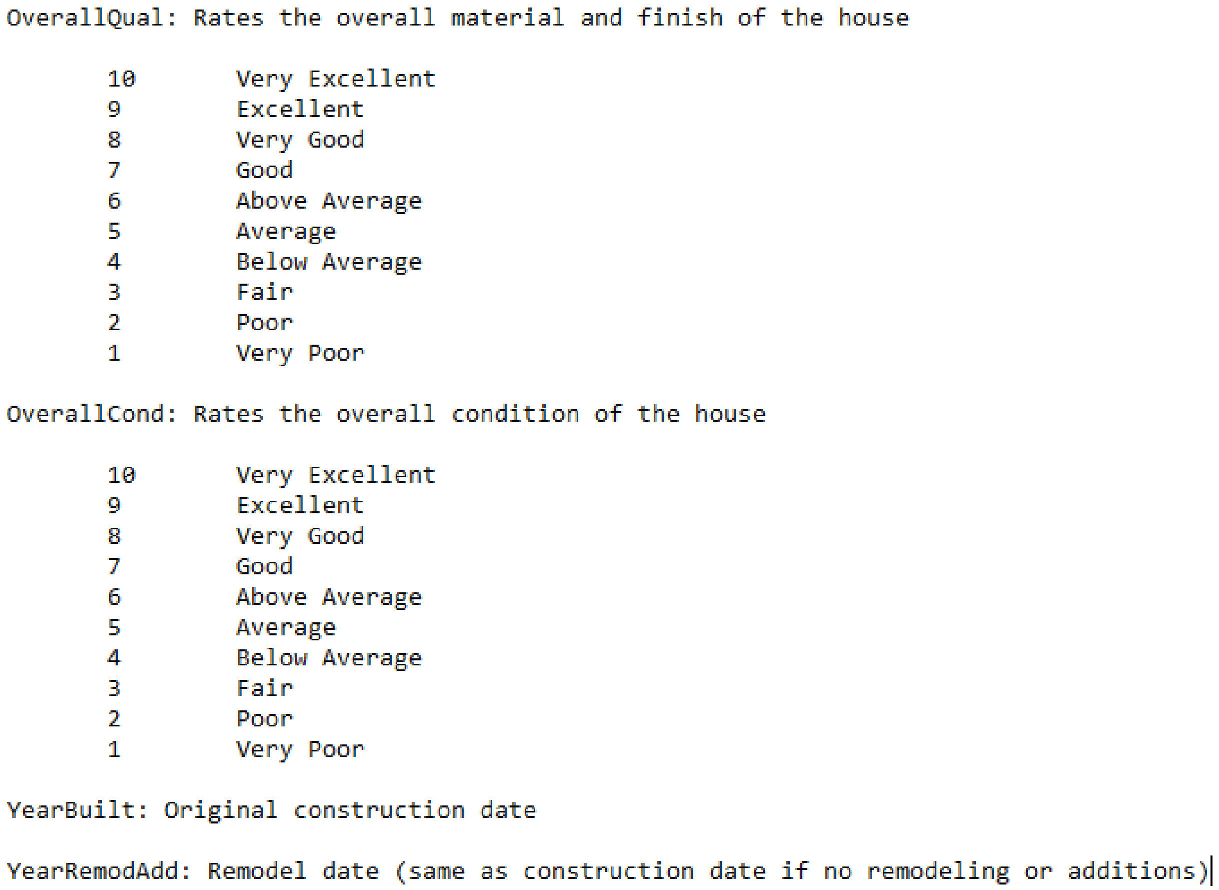
What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

Following data description I found necessary-

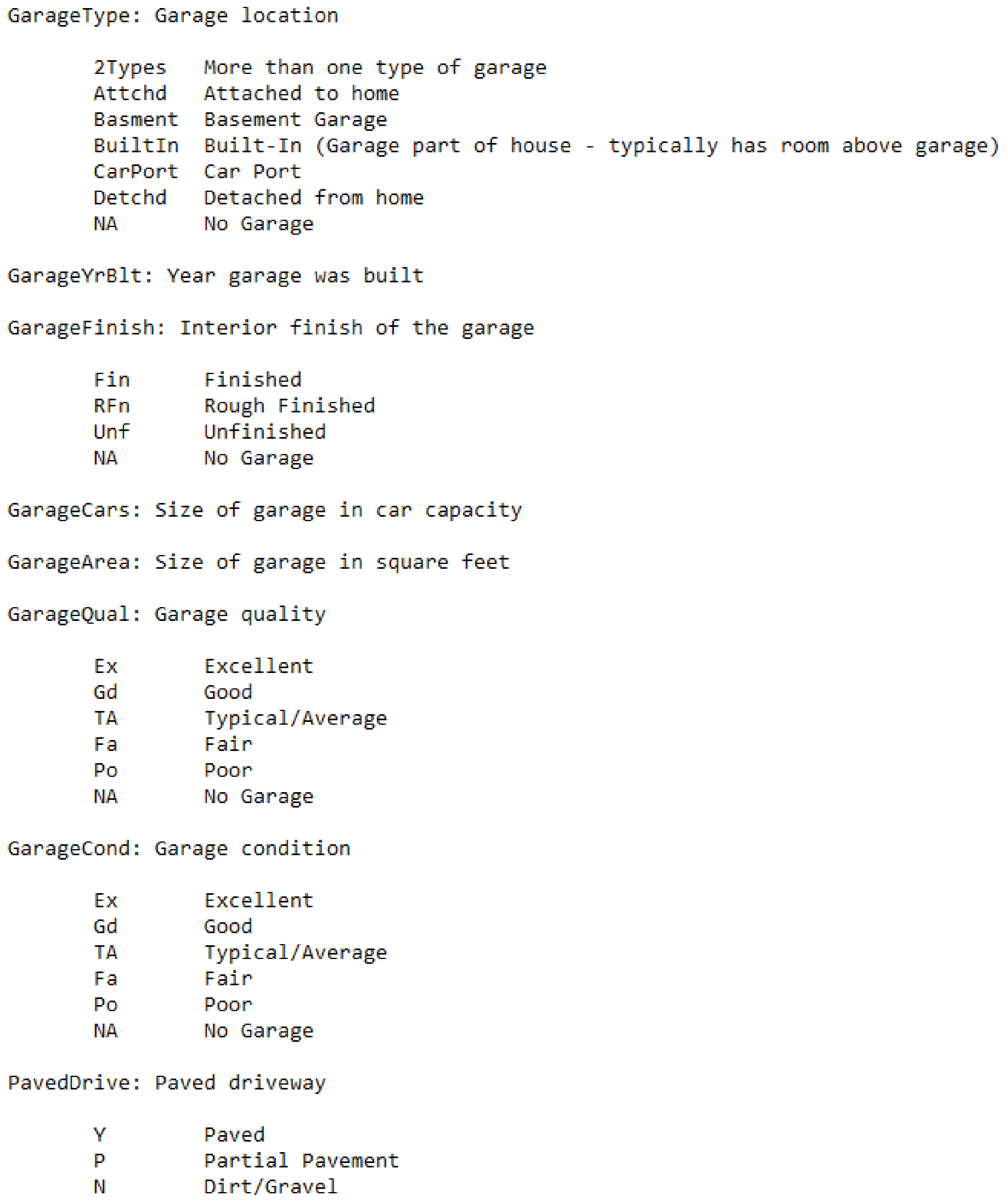


Neighborhood: Physical locations within Ames city limits

Blmngtn	Bloomington Heights
Blueste	Bluestem
BrDale	Briardale
BrkSide	Brookside
ClearCr	Clear Creek
CollgCr	College Creek
Crawfor	Crawford
Edwards	Edwards
Gilbert	Gilbert
IDOTRR	Iowa DOT and Rail Road
MeadowV	Meadow Village
Mitchel	Mitchell
Names	North Ames
NoRidge	Northridge
NPkVill	Northpark Villa
NridgHt	Northridge Heights
NWAmes	Northwest Ames
OldTown	Old Town
SWISU	South & West of Iowa State University
Sawyer	Sawyer
SawyerW	Sawyer West
Somerst	Somerset
StoneBr	Stone Brook
Timber	Timberland
Veenker	Veenker



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

       10      Very Excellent
       9       Excellent
       8       Very Good
       7       Good
       6       Above Average
       5       Average
       4       Below Average
       3       Fair
       2       Poor
       1       Very Poor

OverallCond: Rates the overall condition of the house

       10      Very Excellent
       9       Excellent
       8       Very Good
       7       Good
       6       Above Average
       5       Average
       4       Below Average
       3       Fair
       2       Poor
       1       Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
```



```
GarageType: Garage location

       2Types   More than one type of garage
       Attchd   Attached to home
       Basment  Basement Garage
       BuiltIn  Built-In (Garage part of house - typically has room above garage)
       CarPort  Car Port
       Detchd   Detached from home
       NA       No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

       Fin      Finished
       RFn      Rough Finished
       Unf      Unfinished
       NA       No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

       Ex       Excellent
       Gd       Good
       TA       Typical/Average
       Fa       Fair
       Po       Poor
       NA       No Garage

GarageCond: Garage condition

       Ex       Excellent
       Gd       Good
       TA       Typical/Average
       Fa       Fair
       Po       Poor
       NA       No Garage

PavedDrive: Paved driveway

       Y        Paved
       P        Partial Pavement
       N        Dirt/Gravel
```

Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

My Answer--

- There were many missing values in the dataset. I used the Simple imputer method from sklearn library to impute the missing values in the data.I calculated skewness for the data and based on skewness data,I imputed the missing values.
- I used One Hot encoding method to convert the ‘object’ data type columns in the data to float or integer.
- I used the Elliptic Envelope to detect the outliers and removed them(the removal was limited to 5% of the dataset).
- I used the log1p method from numpy library to adjust my skewness of the dataset.The log1p() function **computes the value of log(1+x) accurately** even for a tiny argument x.
- I used the Standard Scalar from sklearn.preprocessing to scale the data because scaling was necessary for better regression tasks.

I applied the same preprocessing methods to the test.csv which I imported in my jupyter notebook so that the model correctly predicts the **sale price.**

• Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

My Answer--

The relationship between data input and output is very important in machine learning model preparation. If the data we import is incorrect, then the output will also be incorrect. Every machine learning task is dependent on the data we import to prepare the machine learning model.

I used the correlation matrix to find the relationship between inputs and output. Correlation greater than 0.5 is considered a strong positive relationship.

There is strong positive relationship between below mentioned inputs/columns and the output (SalePrice)--

1. Overall Quality
2. The year house was built
3. The year house was sold
4. Total square feet of basement area
5. First Floor square feet of house
6. Above grade (ground) living area square feet
7. Full bathrooms above grade
8. Total rooms above grade (does not include bathrooms)
9. Size of garage in car capacity
10. Size of garage in square feet

- State the set of assumptions (if any) related to the problem under consideration

My Answer--

1. I chose a random state for all my regression models as well as a random state for splitting my data into training and test data to be **14.**
2. I used an elliptic envelope method to detect the outliers in my dataset.

- Hardware and Software Requirements and Tools Used

MY ANSWER-

Libraries and packages used :-

- Pandas-used for Exploratory data analysis

  I used-

  - □ .get_dummies method to perform one hot encoding on my categorical or object data type columns of the data.
  - □ .describe() method to understand the data
  - □ .shape to see the shape of my rows and columns of the data.
  - □ Pivot table method from pandas library to analyse my data better and derive meaningful insights for my exploratory data analysis.
  - □ .info() method to understand the data type of columns in the data.
- Matplotlib- used for Exploratory data analysis
- Seaborn- used for Exploratory data analysis
  - □ Used bar plot,scatter plot to visualize my data for EXPLORATORY DATA ANALYSIS.

  - □ Used **heatmap** to show **correlation/relationship** between the columns/variables of the data.
  - □ Used boxplot to detect outliers in my data.
- Numpy- used for Exploratory data analysis
    - □ Used **log1p** method to reduce excess skewness of the data.
- Sklearn-used for machine learning
  - □ Imported various regression models like KNN,DECISION TREE,RANDOM FORESTS,LIGHT GRADIENT BOOSTING REGRESSOR,XGBOOST
  - □ Performed standard scaling using standard scaler from preprocessing.
  - □
- Scipy-used for statistics
  - □ For computing Chi-square tests.**CHI-SQUARE TEST IS USED TO DETERMINE THE RELATIONSHIP BETWEEN THE CATEGORICAL VARIABLES OF THE DATASET.**
- Light gbm regression –used for machine learning
- Xgboost regression -used for machine learning
- Tableau desktop software for model dashboard

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Describe the approaches you followed, both statistical and analytical, for solving this problem.

  MY ANSWER-

  - □ I calculated the skewness method and plotted the skewness data to determine the normality of the data rather than kdeplot and histogram because I found it more effective.
  - □ I used the ***ELLIPTIC ENVELOPE*** method to detect my outliers upto a predefined limit(I chose around 5% of my data as outliers).

- Testing of Identified Approaches (Algorithms)

  Listing down all the algorithms used for the training and testing.

- Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

MY ANSWER-

Listing down all the algorithms used for the training and testing.

- □ RANDOM FOREST REGRESSOR
- □ GRADIENT BOOSTING REGRESSOR
- □ ADABOOST-REGRESSOR
- □ XGBOOST REGRESSOR
- □ LIGHT GBM REGRESSOR
- □ ELASTIC NET REGRESSION
- □ K-NEAREST NEIGHBORS REGRESSOR
- □ DECISION TREE REGRESSOR
- □ SUPPORT VECTOR MACHINE REGRESSION

I used various metrics to evaluate my model performance and based on them, I selected my models for hyper parameter tuning.

ElasticNet()

```
Root_mean_squared_error:  34029.594873146656

mean_absolute_error:  18626.718551395876

mean_squared_error:  1158013327.2304893

r2:  0.8343928210093544



cross validation scores below:--   ElasticNet()

root_mean_squared_error:  -44511.54774780596

mean_absolute_error_cross_val_score:  -28426.545544544882

mean_squared_error_cross_val_score:  -1997358162.2334046



DecisionTreeRegressor(random_state=14)

Root_mean_squared_error:  37760.2392815826

mean_absolute_error:  26286.359050445102

mean_squared_error:  1425835670.6023738

r2:  0.7960916186712454



cross validation scores below:--   DecisionTreeRegressor(random_state=14)

root_mean_squared_error:  -42358.4080056002

mean_absolute_error_cross_val_score:  -27296.333837301587

mean_squared_error_cross_val_score:  -1847648959.9488056
```

```
RandomForestRegressor(random_state=14)
Root_mean_squared_error:  29549.576782142663
mean_absolute_error:  17574.200356083085
mean_squared_error:  873177488.0037448
r2:  0.8751271188801641


cross validation scores below:--   RandomForestRegressor(random_state=14)
root_mean_squared_error:  -30600.75795632102
mean_absolute_error_cross_val_score:  -18117.537587857147
mean_squared_error_cross_val_score:  -971622160.8938065


AdaBoostRegressor(random_state=14)
Root_mean_squared_error:  34780.46796672726
mean_absolute_error:  23115.97913573362
mean_squared_error:  1209680951.984541
r2:  0.8270038476880114


cross validation scores below:--   AdaBoostRegressor(random_state=14)
root_mean_squared_error:  -35574.10379973371
mean_absolute_error_cross_val_score:  -24145.16320502114
mean_squared_error_cross_val_score:  -1283316722.4231887
```

```
GradientBoostingRegressor(random_state=14)

Root_mean_squared_error:  28272.476490941488

mean_absolute_error:  15990.687443093462

mean_squared_error:  799332926.9308392

r2:  0.8856876099863707



cross validation scores below:--   GradientBoostingRegressor(random_state=14)

root_mean_squared_error:  -28520.808472163342

mean_absolute_error_cross_val_score:  -16752.81256483122

mean_squared_error_cross_val_score:  -875154641.6405681



SVR()

Root_mean_squared_error:  85955.69485358296

mean_absolute_error:  57195.91035129185

mean_squared_error:  7388381477.762269

r2:  -0.05661047681138576



cross validation scores below:--   SVR()

root_mean_squared_error:  -80293.21542089937

mean_absolute_error_cross_val_score:  -54983.85554678177

mean_squared_error_cross_val_score:  -6508283790.638884
```

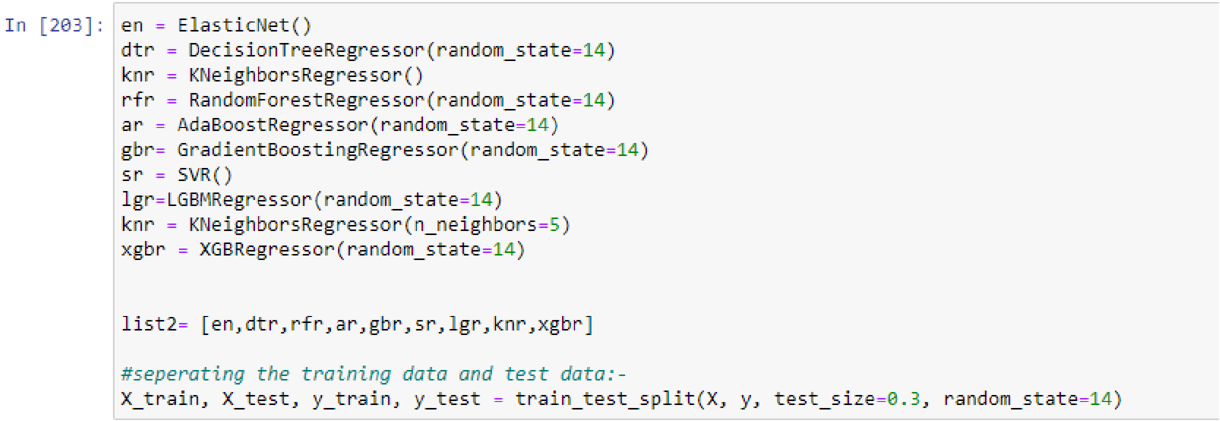
```
LGBMRegressor(random_state=14)

Root_mean_squared_error:  29077.787117232743

mean_absolute_error:  16983.20731206699

mean_squared_error:  845517703.6351064

r2:  0.8790827373113742



cross validation scores below:--   LGBMRegressor(random_state=14)

root_mean_squared_error:  -29191.21648098475

mean_absolute_error_cross_val_score:  -17595.12283058051

mean_squared_error_cross_val_score:  -880897956.9227698



KNeighborsRegressor()

Root_mean_squared_error:  47305.98881546126

mean_absolute_error:  27638.755489614243

mean_squared_error:  2237856577.808546

r2:  0.6799647239614465



cross validation scores below:--   KNeighborsRegressor()

root_mean_squared_error:  -52406.672950534245

mean_absolute_error_cross_val_score:  -34497.00751666667
```

```
mean_squared_error_cross_val_score:  -2771421712.784726


XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=0.300000012, max_delta_step=0, max_depth=6,
             min_child_weight=1, missing=nan, monotone_constraints='()',
             n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=14,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)
Root_mean_squared_error:  29352.619063833678
mean_absolute_error:  17789.52981268546
mean_squared_error:  861576245.9065322
r2:  0.8767862094375257


cross validation scores below:--   XGBRegressor(base_score=0.5, booster='gbtree',
colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=0.300000012, max_delta_step=0, max_depth=6,
             min_child_weight=1, missing=nan, monotone_constraints='()',
             n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=14,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)
```

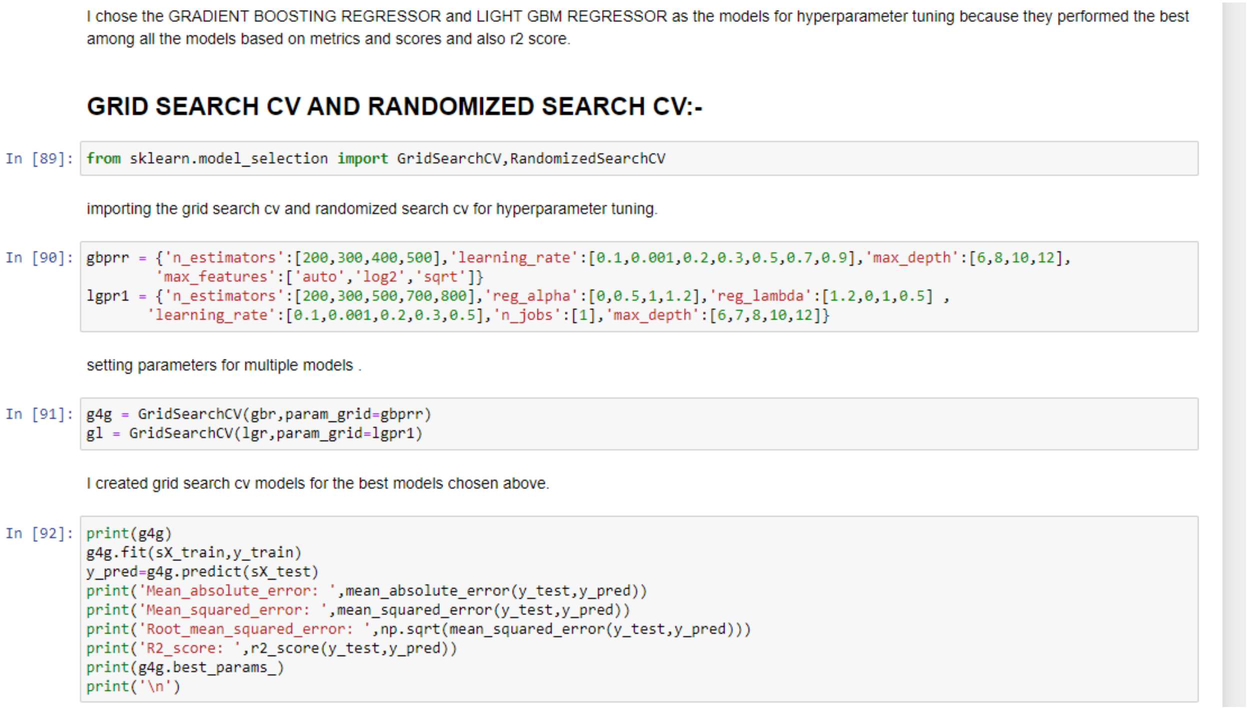
root_mean_squared_error: -31434.881271289632

mean_absolute_error_cross_val_score: -18959.158092649428

mean_squared_error_cross_val_score: -1026883932.8239281



```
In [203]: en = ElasticNet()
          dtr = DecisionTreeRegressor(random_state=14)
          knr = KNeighborsRegressor()
          rfr = RandomForestRegressor(random_state=14)
          ar = AdaBoostRegressor(random_state=14)
          gbr= GradientBoostingRegressor(random_state=14)
          sr = SVR()
          lgr=LGBMRegressor(random_state=14)
          knr = KNeighborsRegressor(n_neighbors=5)
          xgbr = XGBRegressor(random_state=14)


          list2= [en,dtr,rfr,ar,gbr,sr,lgr,knr,xgbr]

          #seperating the training data and test data:-
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=14)
```



I chose the GRADIENT BOOSTING REGRESSOR and LIGHT GBM REGRESSOR as the models for hyperparameter tuning because they performed the best among all the models based on metrics and scores and also r2 score.

GRID SEARCH CV AND RANDOMIZED SEARCH CV:-

```
In [89]: from sklearn.model_selection import GridSearchCV,RandomizedSearchCV
```

importing the grid search cv and randomized search cv for hyperparameter tuning.

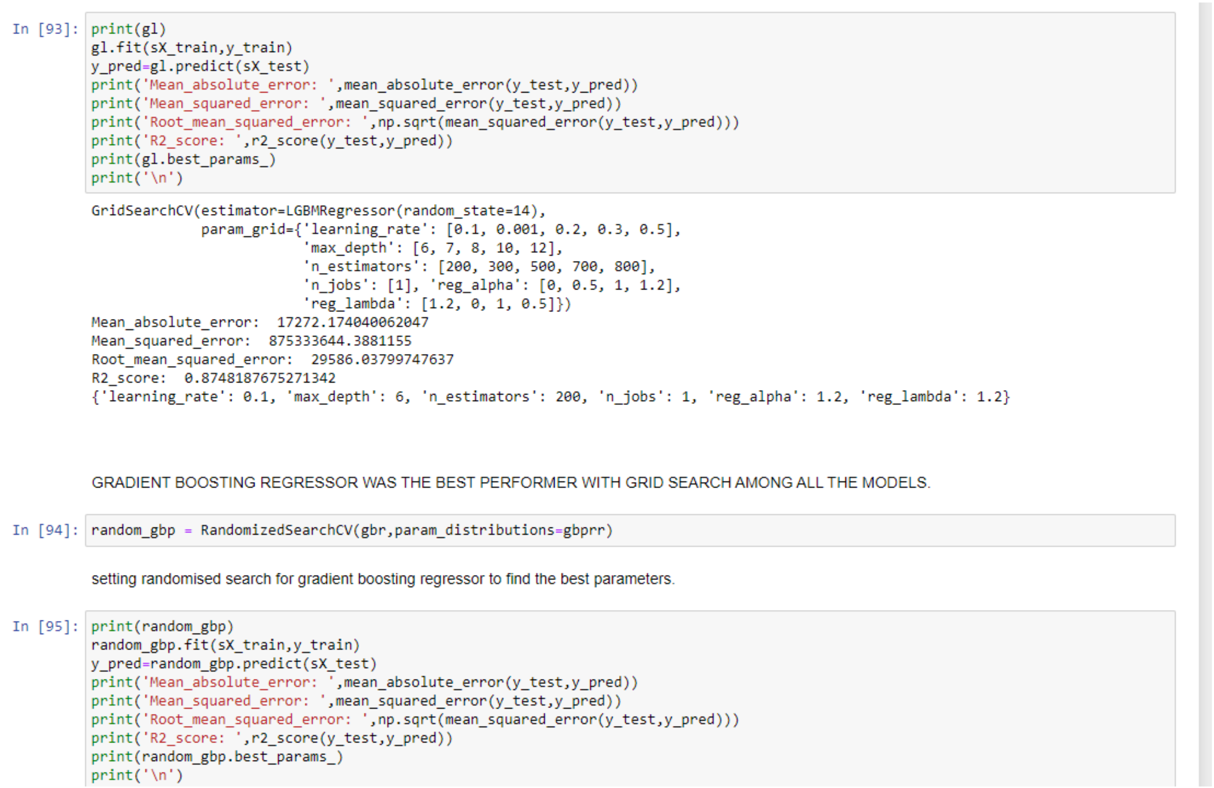
```
In [90]: gbprr = {'n_estimators':[200,300,400,500],'learning_rate':[0.1,0.001,0.2,0.3,0.5,0.7,0.9],'max_depth':[6,8,10,12],
                  'max_features':['auto','log2','sqrt']}
         lgpr1 = {'n_estimators':[200,300,500,700,800],'reg_alpha':[0,0.5,1,1.2],'reg_lambda':[1.2,0,1,0.5] ,
                 'learning_rate':[0.1,0.001,0.2,0.3,0.5],'n_jobs':[1],'max_depth':[6,7,8,10,12]}
```

setting parameters for multiple models .

```
In [91]: g4g = GridSearchCV(gbr,param_grid=gbprr)
         gl = GridSearchCV(lgr,param_grid=lgpr1)
```

I created grid search cv models for the best models chosen above.

```
In [92]: print(g4g)
         g4g.fit(sX_train,y_train)
         y_pred=g4g.predict(sX_test)
         print('Mean_absolute_error: ',mean_absolute_error(y_test,y_pred))
         print('Mean_squared_error: ',mean_squared_error(y_test,y_pred))
         print('Root_mean_squared_error: ',np.sqrt(mean_squared_error(y_test,y_pred)))
         print('R2_score: ',r2_score(y_test,y_pred))
         print(g4g.best_params_)
         print('\n')
```



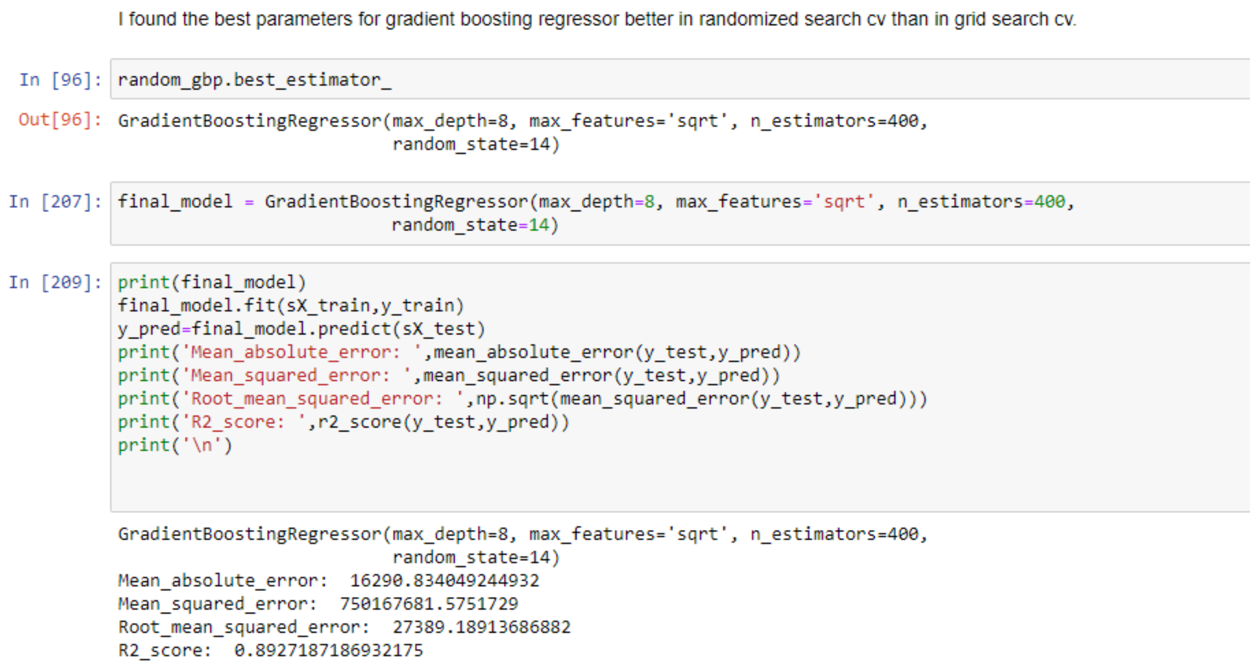
```
In [93]: print(gl)
         gl.fit(sX_train,y_train)
         y_pred=gl.predict(sX_test)
         print('Mean_absolute_error: ',mean_absolute_error(y_test,y_pred))
         print('Mean_squared_error: ',mean_squared_error(y_test,y_pred))
         print('Root_mean_squared_error: ',np.sqrt(mean_squared_error(y_test,y_pred)))
         print('R2_score: ',r2_score(y_test,y_pred))
         print(gl.best_params_)
         print('\n')

         GridSearchCV(estimator=LGBMRegressor(random_state=14),
                      param_grid={'learning_rate': [0.1, 0.001, 0.2, 0.3, 0.5],
                                  'max_depth': [6, 7, 8, 10, 12],
                                  'n_estimators': [200, 300, 500, 700, 800],
                                  'n_jobs': [1], 'reg_alpha': [0, 0.5, 1, 1.2],
                                  'reg_lambda': [1.2, 0, 1, 0.5]})
         Mean_absolute_error:  17272.174040062047
         Mean_squared_error:  875333644.3881155
         Root_mean_squared_error:  29586.03799747637
         R2_score:  0.8748187675271342
         {'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200, 'n_jobs': 1, 'reg_alpha': 1.2, 'reg_lambda': 1.2}
```

GRADIENT BOOSTING REGRESSOR WAS THE BEST PERFORMER WITH GRID SEARCH AMONG ALL THE MODELS.

```
In [94]: random_gbp = RandomizedSearchCV(gbr,param_distributions=gbprr)
```

setting randomised search for gradient boosting regressor to find the best parameters.

```
In [95]: print(random_gbp)
         random_gbp.fit(sX_train,y_train)
         y_pred=random_gbp.predict(sX_test)
         print('Mean_absolute_error: ',mean_absolute_error(y_test,y_pred))
         print('Mean_squared_error: ',mean_squared_error(y_test,y_pred))
         print('Root_mean_squared_error: ',np.sqrt(mean_squared_error(y_test,y_pred)))
         print('R2_score: ',r2_score(y_test,y_pred))
         print(random_gbp.best_params_)
         print('\n')
```



I found the best parameters for gradient boosting regressor better in randomized search cv than in grid search cv.

```
In [96]: random_gbp.best_estimator_

Out[96]: GradientBoostingRegressor(max_depth=8, max_features='sqrt', n_estimators=400,
                                   random_state=14)

In [207]: final_model = GradientBoostingRegressor(max_depth=8, max_features='sqrt', n_estimators=400,
                                   random_state=14)

In [209]: print(final_model)
          final_model.fit(sX_train,y_train)
          y_pred=final_model.predict(sX_test)
          print('Mean_absolute_error: ',mean_absolute_error(y_test,y_pred))
          print('Mean_squared_error: ',mean_squared_error(y_test,y_pred))
          print('Root_mean_squared_error: ',np.sqrt(mean_squared_error(y_test,y_pred)))
          print('R2_score: ',r2_score(y_test,y_pred))
          print('\n')

          GradientBoostingRegressor(max_depth=8, max_features='sqrt', n_estimators=400,
                                    random_state=14)
          Mean_absolute_error:  16290.834049244932
          Mean_squared_error:  750167681.5751729
          Root_mean_squared_error:  27389.18913686882
          R2_score:  0.8927187186932175
```

- Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

My answer-

The metrics used were--

- □ R2 score-- the closer it is to 1,the better the model performance is.
- □ Mean absolute error
- □ Mean squared error
- □ Root mean squared error(RMSE)

All the errors should be as minimal as possible.the lesser they are,the better the model performance is.

- ## Visualizations

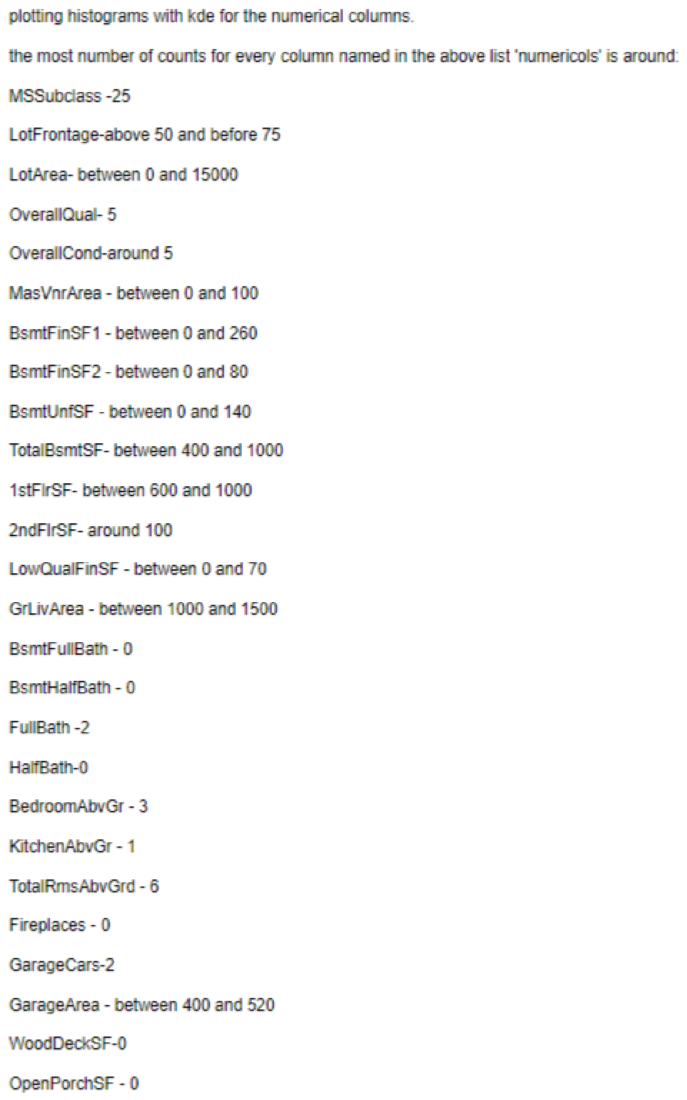
  Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

  If different platforms were used, mention that as well.

My answer-

1. Histogram

   I plotted it to know about the skewness and distribution of the numerical data columns.



plotting histograms with kde for the numerical columns.

the most number of counts for every column named in the above list 'numericols' is around:

MSSubclass -25

LotFrontage-above 50 and before 75

LotArea- between 0 and 15000

OverallQual- 5

OverallCond-around 5

MasVnrArea - between 0 and 100

BsmtFinSF1 - between 0 and 260

BsmtFinSF2 - between 0 and 80

BsmtUnfSF - between 0 and 140

TotalBsmtSF- between 400 and 1000

1stFlrSF- between 600 and 1000

2ndFlrSF- around 100

LowQualFinSF - between 0 and 70

GrLivArea - between 1000 and 1500

BsmtFullBath - 0

BsmtHalfBath - 0

FullBath -2

HalfBath-0

BedroomAbvGr - 3

KitchenAbvGr - 1

TotalRmsAbvGrd - 6

Fireplaces - 0

GarageCars-2

GarageArea - between 400 and 520

WoodDeckSF-0

OpenPorchSF - 0

2. Scatter plot

   To know about the relationship between the numerical variables and the target label --SALE PRICE.

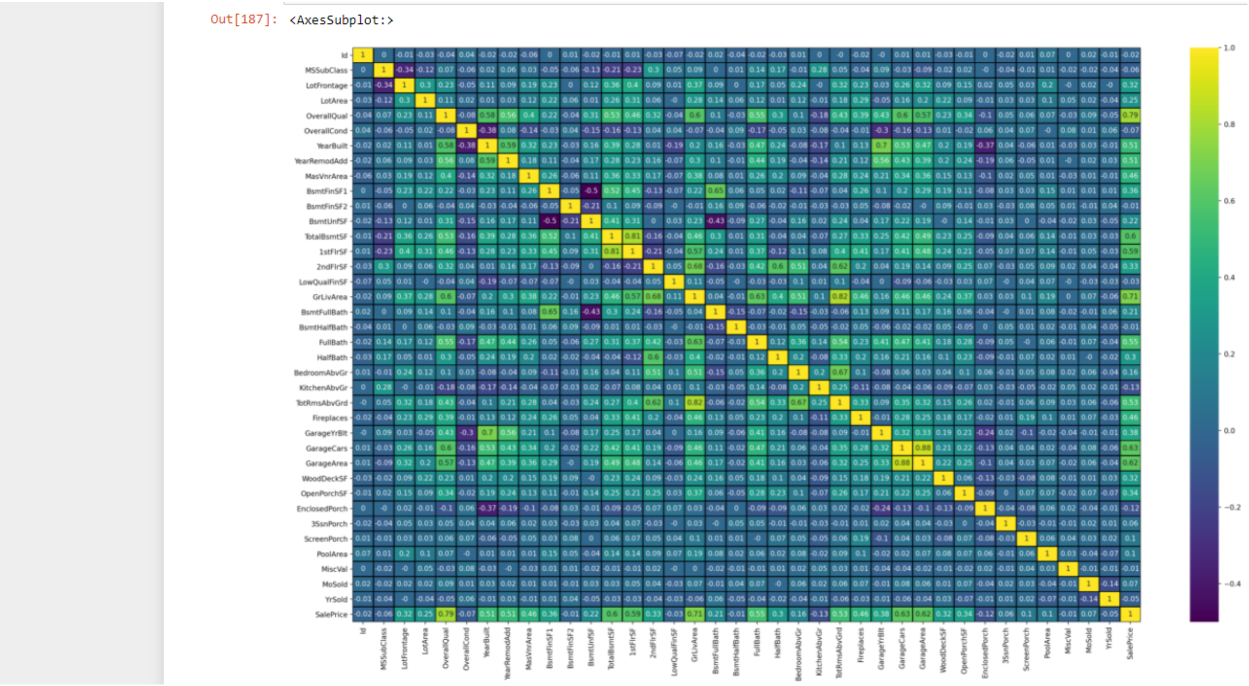
3. Bar plot

   To show comparison of different data based on categorical columns.

4. Heatmap

☐ To show correlation between the variables and target label of the data.

☐ To show the chi square test results.



```
Out[187]: <AxesSubplot:>
```

- ## Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

I found that while calculating skewness of the dataset,***many*** of the columns /variables were **not normally distributed**. When comparing all the models over r2 score,mean absolute error,mean squared error and root mean squared error---light gbm regression and gradient boosting regressor were consistent. After doing grid search cv, I found that gradient boosting regressor performed better than light gbm regressor.I used randomized search cv to find better scores for my gradient boosting regressor,so I saved that model with the help of the joblib library.

CONCLUSION

- Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

I found that most of the data were highly skewed and there were many missing values in both the test as well as train data.

- Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how you overcame that.

My answer-

- I learned to use tableau desktop software to make my model dashboard.It was a wonderful experience to apply my skills firsthand.I found out that tableau to be a great tool for data visualization and analysis.
- I learned how to organize my data for the machine learning process.

- I applied one of my favourite methods for data analysis which I often see in Excel--**pivot tables.**
- I found out that skewness can be a great technique to know a lot about the data like--
  - □ whether it is normal or not,
  - □ how it affects outliers.
  - □ How it can be used to handle missing values.
- I found the seaborn heatmap plotting correlation to be better than scatterplots at finding out the relationships between the variables of the dataset.