# SURPRISE HOUSING REGRESSION CASE STUDY

## 1. ABSTRACT

**The Surprise Housing case study aims to predict house prices based on various factors like location, size, number of rooms, and other attributes. By utilizing machine learning and data science techniques, the study develops predictive models to estimate house prices, providing insights into the factors influencing real estate markets. Multiple regression models are experimented with, and the most effective model is selected based on performance metrics.**

**Surprise Housing** is US based housing Company. It uses data analytics to purchase houses at a price below their actual values and flip them on at a higher price. Company has now decided to enter Australian market. It is looking at prospective properties to buy to enter the market.

### Business Goal

**Surprise Housing** requires a model that can 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 management to understand the pricing dynamics of a new market.

### Aim of the case study

1. To understand the factors affecting the price of houses in Australian market. The company wants to know:
    - Which variables are significant in predicting the price of a house, and
    - How well those variables describe the price of a house.
2. To create regression model using regularisation in order to predict the actual value of the prospective properties:
    - determine the optimal value of lambda for ridge and lasso regression.

### Following are steps used for creation and evaluation of model for case study

- Step 1: Reading and Understanding the Data
- Step 2: Data Analysis
- Step 3: Data Preparation
- Step 4: Splitting the Data into Training and Testing Sets
- Step 5: Model Building and Evaluation
- Step 6: Conclusion

- Step 7: Summary

**OBJECTIVE**

The primary objective of this case study is to predict housing prices using machine learning algorithms, by exploring and analyzing a dataset of historical housing prices. Additionally, the study seeks to identify the key factors that most influence the pricing of houses.

## 3. INTRODUCTION

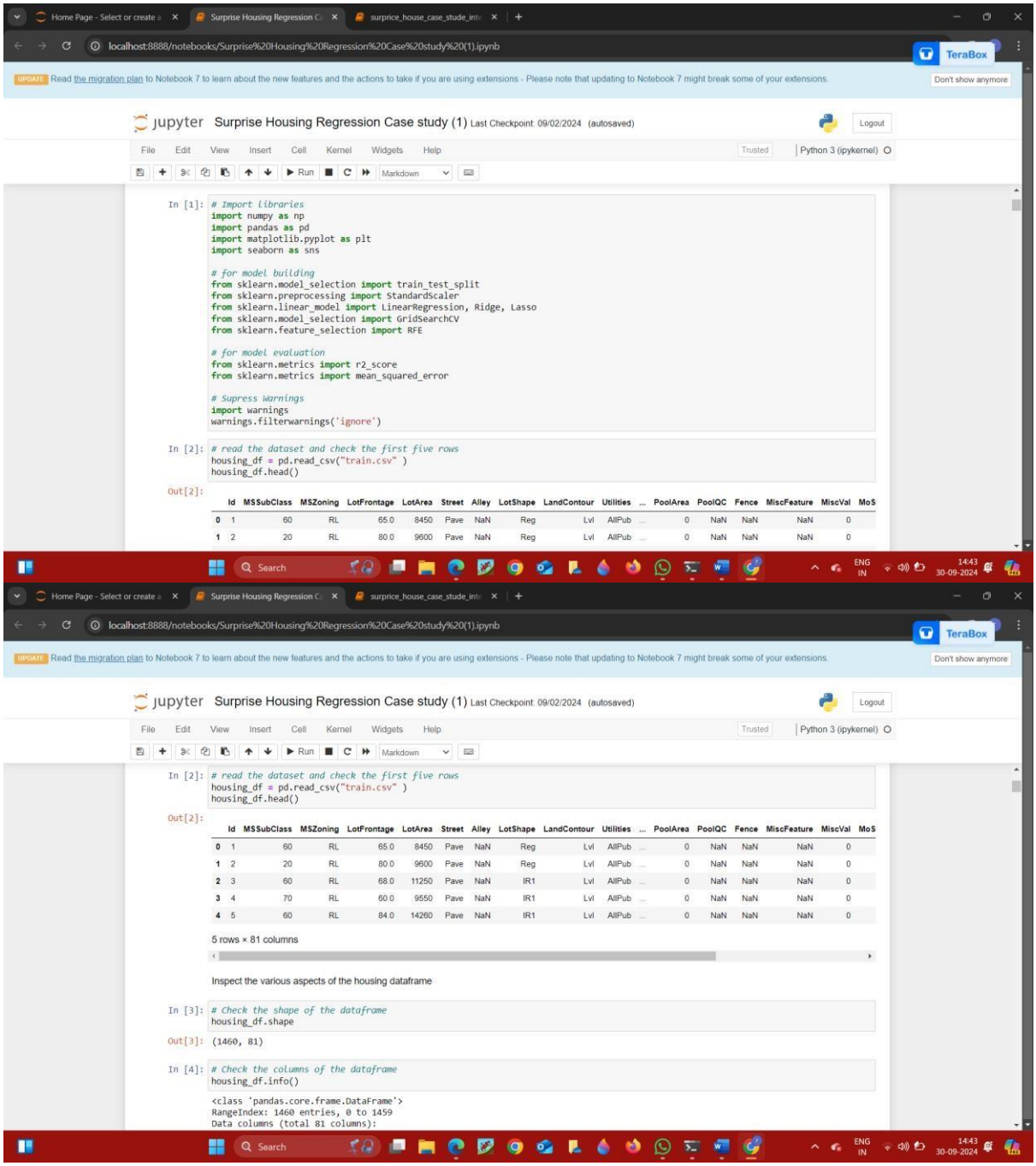
Housing prices are affected by various factors including geographical location, house size, age of the house, number of bedrooms, and market conditions. Accurate prediction of these prices is crucial for buyers, sellers, and investors. In this case study, we apply data science and machine learning techniques to build a model capable of predicting house prices. The dataset comprises multiple features, which will be used to train different models, evaluate their performance, and choose the best one.

## 4. METHODOLOGY

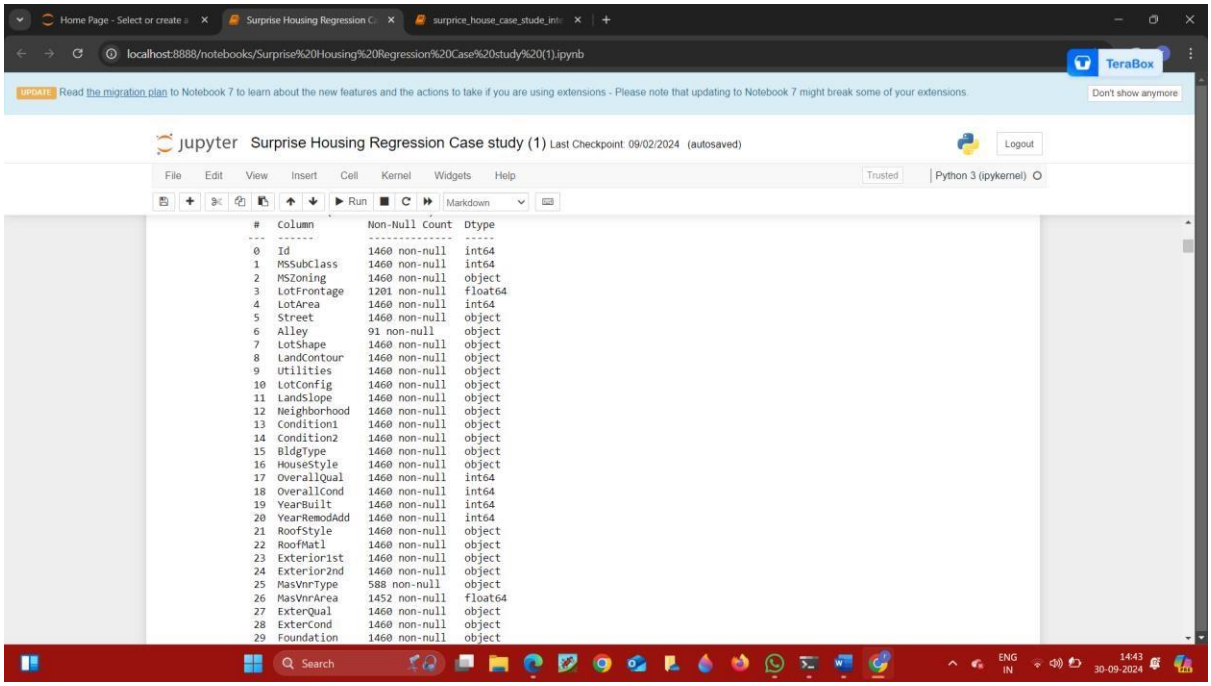
- **Data Collection and Understanding**: The dataset used for the study contains historical house prices and relevant features (e.g., area, number of rooms, neighborhood, etc.). Initial exploration helps understand the distribution, missing values, and outliers in the data.
- **Data Preprocessing**:
    - Handling missing values and outliers.
    - Encoding categorical variables (such as location). o Feature scaling (normalizing or standardizing numeric features).
    - Splitting data into training and testing sets.
- **Exploratory Data Analysis (EDA)**: Visualize the relationships between different features and the target variable (house price). Use correlation matrices and plots (scatter plots, histograms, etc.) to uncover patterns in the data.
- **Feature Engineering**:
    - Creation of new features if necessary (e.g., price per square foot).
    - Selecting the most relevant features using statistical tests, domain knowledge, and correlation analysis.
- **Model Building**:
    - Train multiple models such as:
        1. Linear Regression

2. Ridge and Lasso Regression
3. Decision Trees
4. Random Forest
5. Gradient Boosting o Tune hyperparameters using grid search or cross-validation techniques.

- **Model Evaluation**: Evaluate the models using performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared value.
- **Model Deployment**: The best-performing model is finalized and can be deployed for real-world use or further refined.


Home Page - Select or create a
Surprise Housing Regression C
surprice_house_case_stude_inte
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown
In [1]:
# Import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# for model building
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression, Ridge, Lasso
from sklearn.model_selection import GridSearchCV
from sklearn.feature_selection import RFE
# for model evaluation
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
# Supress Warnings
import warnings
warnings.filterwarnings('ignore')
In [2]:
# read the dataset and check the first five rows
housing_df = pd.read_csv("train.csv" )
housing_df.head()
Out[2]:
Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour Utilities ... PoolArea PoolQC Fence MiscFeature MiscVal MoS
0 1 60 RL 65.0 8450 Pave NaN Reg Lvl AllPub ... 0 NaN NaN NaN 0
1 2 20 RL 80.0 9600 Pave NaN Reg Lvl AllPub ... 0 NaN NaN NaN 0
Search
ENG IN
14:43 30-09-2024

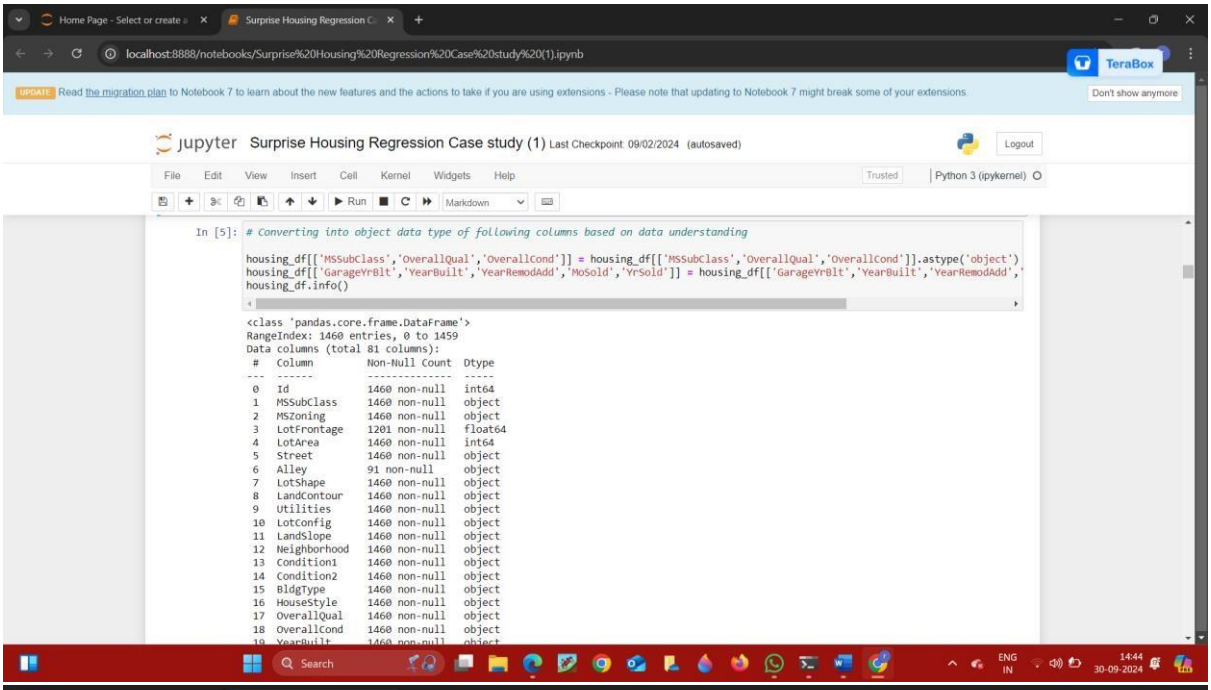

Home Page - Select or create a
Surprise Housing Regression C
surprice_house_case_stude_inte
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown
In [2]:
# read the dataset and check the first five rows
housing_df = pd.read_csv("train.csv" )
housing_df.head()
Out[2]:
Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour Utilities ... PoolArea PoolQC Fence MiscFeature MiscVal MoS
0 1 60 RL 65.0 8450 Pave NaN Reg Lvl AllPub ... 0 NaN NaN NaN 0
1 2 20 RL 80.0 9600 Pave NaN Reg Lvl AllPub ... 0 NaN NaN NaN 0
2 3 60 RL 68.0 11250 Pave NaN IR1 Lvl AllPub ... 0 NaN NaN NaN 0
3 4 70 RL 60.0 9550 Pave NaN IR1 Lvl AllPub ... 0 NaN NaN NaN 0
4 5 60 RL 84.0 14260 Pave NaN IR1 Lvl AllPub ... 0 NaN NaN NaN 0
5 rows × 81 columns
Inspect the various aspects of the housing dataframe
In [3]:
# Check the shape of the dataframe
housing_df.shape
Out[3]:
(1460, 81)
In [4]:
# Check the columns of the dataframe
housing_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
Search
ENG IN
14:43 30-09-2024

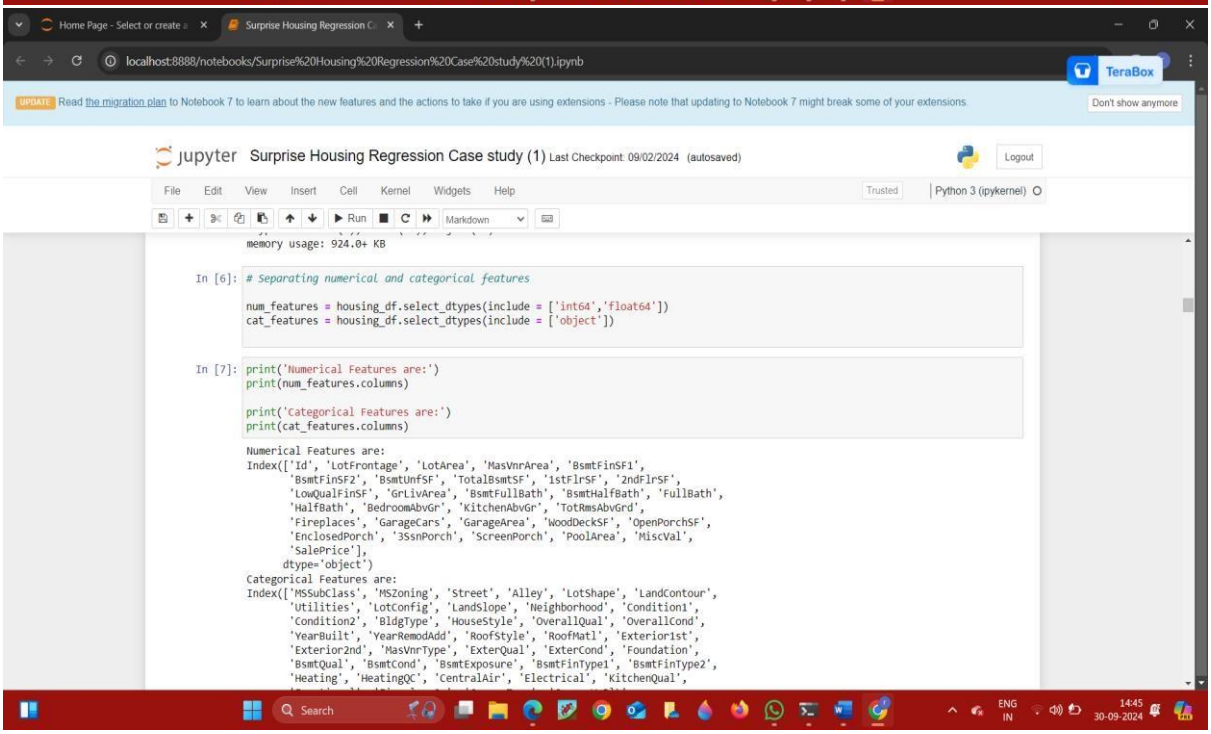



**Step 2: Data Analysis**

Here following things will be done:

- Classifying Numerical and Categorical variables
- Analysing Numerical Variables
- Analysing Categorical Variables

- 
Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown
In [5]:
# Converting into object data type of following columns based on data understanding
housing_df[['MSSubClass','OverallQual','OverallCond']] = housing_df[['MSSubClass','OverallQual','OverallCond']].astype('object')
housing_df[['GarageYrBlt','YearBuilt','YearRemodAdd','MoSold','YrSold']] = housing_df[['GarageYrBlt','YearBuilt','YearRemodAdd','
housing_df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
# Column Non-Null Count Dtype
0 Id 1460 non-null int64
1 MSSubClass 1460 non-null object
2 MSZoning 1460 non-null object
3 LotFrontage 1201 non-null float64
4 LotArea 1460 non-null int64
5 Street 1460 non-null object
6 Alley 91 non-null object
7 LotShape 1460 non-null object
8 LandContour 1460 non-null object
9 Utilities 1460 non-null object
10 LotConfig 1460 non-null object
11 LandSlope 1460 non-null object
12 Neighborhood 1460 non-null object
13 Condition1 1460 non-null object
14 Condition2 1460 non-null object
15 BldgType 1460 non-null object
16 HouseStyle 1460 non-null object
17 OverallQual 1460 non-null object
18 OverallCond 1460 non-null object
Search
ENG
IN
14:44
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown
memory usage: 924.0+ KB
In [6]:
# Separating numerical and categorical features
num_features = housing_df.select_dtypes(include = ['int64','float64'])
cat_features = housing_df.select_dtypes(include = ['object'])
In [7]:
print('Numerical Features are:')
print(num_features.columns)
print('Categorical Features are:')
print(cat_features.columns)
Numerical Features are:
Index(['Id', 'LotFrontage', 'LotArea', 'MasVnrArea', 'BsmtFinSF1',
'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF',
'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath',
'HalfBath', 'BedroomAbvGr', 'KitchenAbvGr', 'TotRmsAbvGrd',
'Fireplaces', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF',
'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MiscVal',
'SalePrice'],
dtype='object')
Categorical Features are:
Index(['MSSubClass', 'MSZoning', 'Street', 'Alley', 'LotShape', 'LandContour',
'Utilities', 'LotConfig', 'LandSlope', 'Neighborhood', 'Condition1',
'Condition2', 'BldgType', 'HouseStyle', 'OverallQual', 'OverallCond',
'YearBuilt', 'YearRemodAdd', 'RoofStyle', 'RoofMatl', 'Exterior1st',
'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation',
'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2',
'Heating', 'HeatingQC', 'CentralAir', 'Electrical', 'KitchenQual',
Search
ENG
IN
14:45
30-09-2024

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help | Trusted | Python 3 (ipykernel)

Run | Markdown

## Analysing Numerical features

Here we will analyse the numerical data, visualize it and clean it as well.

In [8]:
```
# Checking the statistical summary of numerical variables
num_features.describe()
```

Out[8]:

| | Id | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtFinSF2 | BsmtUnfSF | TotalBsmtSF | 1stFlrSF | 2ndFlrSF | ... | GarageCars |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 1460.000000 | 1201.000000 | 1460.000000 | 1452.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | ... | 1460.000000 |
| mean | 730.500000 | 70.049958 | 10516.828082 | 103.685262 | 443.639726 | 46.549315 | 567.240411 | 1057.429452 | 1162.626712 | 346.992466 | ... | 1.767123 |
| std | 421.610009 | 24.284752 | 9981.264932 | 181.066207 | 456.098091 | 161.319273 | 441.866955 | 438.705324 | 386.587738 | 436.528436 | ... | 0.747315 |
| min | 1.000000 | 21.000000 | 1300.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 334.000000 | 0.000000 | ... | 0.000000 |
| 25% | 365.750000 | 59.000000 | 7553.500000 | 0.000000 | 0.000000 | 0.000000 | 223.000000 | 795.750000 | 882.000000 | 0.000000 | ... | 1.000000 |
| 50% | 730.500000 | 69.000000 | 9478.500000 | 0.000000 | 383.500000 | 0.000000 | 477.500000 | 991.500000 | 1087.000000 | 0.000000 | ... | 2.000000 |
| 75% | 1095.250000 | 80.000000 | 11601.500000 | 166.000000 | 712.250000 | 0.000000 | 808.000000 | 1298.250000 | 1391.250000 | 728.000000 | ... | 2.000000 |
| max | 1460.000000 | 313.000000 | 215245.000000 | 1600.000000 | 5644.000000 | 1474.000000 | 2336.000000 | 6110.000000 | 4692.000000 | 2065.000000 | ... | 4.000000 |

8 rows × 30 columns

In [9]:
```
# here field ID is just index and will not play any role in analysis so dropping it off
num_features.drop(['Id'], axis = 1 , inplace = True)
num_features.columns
```

Out[9]:
```
Index(['LotFrontage', 'LotArea', 'MasVnrArea', 'BsmtFinSF1', 'BsmtFinSF2',
       'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF',
```



Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore

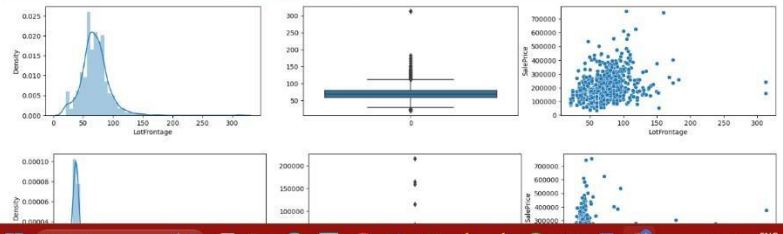
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help | Trusted | Python 3 (ipykernel)

Run | Markdown

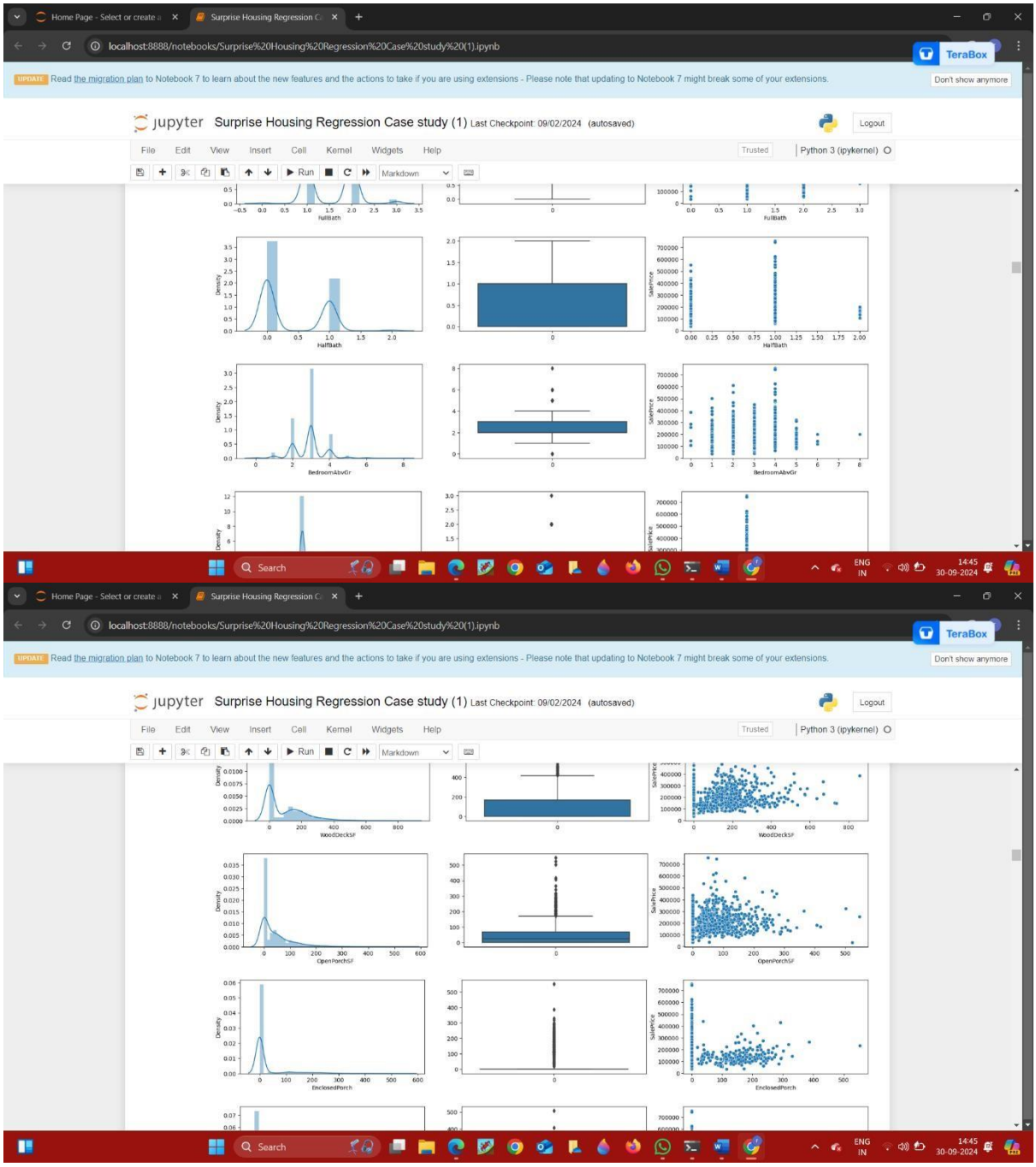
In [10]:
```
# Lets visualize the numerical features by doing univariate and bi-variate analysis

for col in (num_features.columns):
    plt.figure(figsize=(20, 3))
    plt.title(col, fontdict={'fontsize': 18})

    plt.subplot(1,3,1)
    sns.distplot(housing_df[col])

    plt.subplot(1,3,2)
    sns.boxplot(housing_df[col])

    plt.subplot(1,3,3)
    sns.scatterplot(x=housing_df[col],y="SalePrice", data=housing_df)
```







Search | ENG IN | 14:45 30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown
FullBath
HalfBath
BedroomAbvGr
Density
SalePrice
Search
ENG
IN
14:45
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown
WoodDeckSF
OpenPorchSF
EnclosedPorch
Density
SalePrice
Search
ENG
IN
14:45
30-09-2024

Home Page - Select or create a | Surprise Housing Regression Ca

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

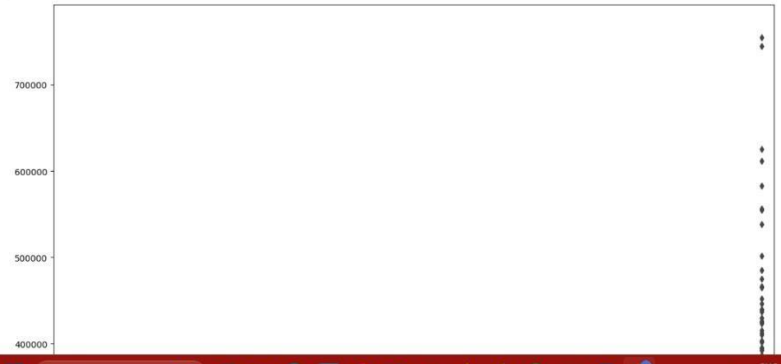
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

In [11]:
```
# Lets check what the box plot says:
plt.figure(figsize=[15,15])
plt.xticks(rotation=90)
sns.boxplot(data=num_features);
```




Search ENG IN 14:45 30-09-2024

Home Page - Select or create a | Surprise Housing Regression Ca

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

```
count_outlier = []
per_outlier   = []
for col in (num_features.columns):
    IQR = num_features[col].quantile(.75) - num_features[col].quantile(.25)
    max_val = num_features[col].quantile(.75) + IQR*1.5
    min_val = num_features[col].quantile(.25) - IQR*1.5
    count = ( ((num_features[col] > max_val).sum()) + ((num_features[col] < min_val).sum()) )
    count_outlier.append(count)
    per = ( round((count / housing_df.shape[0]) * 100,2) )
    per_outlier.append(per)
    data_outlier = {'Feature': num_features.columns , 'Outlier Count': count_outlier, 'Outlier Per': per_outlier}
(pd.DataFrame(data_outlier).sort_values(by='Outlier Count', ascending = False))
```

Out[12]:

| | Feature | Outlier Count | Outlier Per |
|---|---|---|---|
| 23 | EnclosedPorch | 208 | 14.25 |
| 4 | BsmtFinSF2 | 167 | 11.44 |
| 25 | ScreenPorch | 116 | 7.95 |
| 2 | MasVnrArea | 96 | 6.58 |
| 0 | LotFrontage | 88 | 6.03 |
| 12 | BsmtHalfBath | 82 | 5.62 |
| 22 | OpenPorchSF | 77 | 5.27 |
| 1 | LotArea | 69 | 4.73 |
| 16 | KitchenAbvGr | 68 | 4.66 |
| 28 | SalePrice | 61 | 4.18 |
| 6 | TotalBsmtSF | 61 | 4.18 |

Search ENG IN 14:45 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

In [13]:
```
# Outlier treatment : Capping to max and min value
for col in (num_features.columns):
    IQR = num_features[col].quantile(.75) - num_features[col].quantile(.25)
    max_val = num_features[col].quantile(.75) + IQR*1.5
    min_val = num_features[col].quantile(.25) - IQR*1.5
   # count = ((num_features[col] > max_val).sum()) + ((num_features[col] < min_val).sum())
    num_features[col][num_features[col] > max_val] = max_val
    num_features[col][num_features[col] < min_val] = min_val
```

In [14]:
```
num_features.describe()
```

Out[14]:

| | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtFinSF2 | BsmtUnfSF | TotalBsmtSF | 1stFlrSF | 2ndFlrSF | LowQualFinSF | ... | GarageCars |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 1201.000000 | 1460.000000 | 1452.000000 | 1460.000000 | 1460.0 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.0 | ... | 1460.000000 |
| mean | 69.134888 | 9647.388014 | 90.192149 | 439.997517 | 0.0 | 563.777740 | 1050.254795 | 1157.018151 | 346.789041 | 0.0 | ... | 1.765411 |
| std | 19.662022 | 3594.356399 | 134.925253 | 433.219435 | 0.0 | 431.710214 | 397.937878 | 362.583002 | 435.791621 | 0.0 | ... | 0.742753 |
| min | 27.500000 | 1481.500000 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 42.000000 | 334.000000 | 0.000000 | 0.0 | ... | 0.000000 |
| 25% | 59.000000 | 7553.500000 | 0.000000 | 0.000000 | 0.0 | 223.000000 | 795.750000 | 882.000000 | 0.000000 | 0.0 | ... | 1.000000 |
| 50% | 69.000000 | 9478.500000 | 0.000000 | 383.500000 | 0.0 | 477.500000 | 991.500000 | 1087.000000 | 0.000000 | 0.0 | ... | 2.000000 |
| 75% | 80.000000 | 11601.500000 | 166.000000 | 712.250000 | 0.0 | 808.000000 | 1298.250000 | 1391.250000 | 728.000000 | 0.0 | ... | 2.000000 |
| max | 111.500000 | 17673.500000 | 415.000000 | 1780.625000 | 0.0 | 1685.500000 | 2052.000000 | 2155.125000 | 1820.000000 | 0.0 | ... | 3.500000 |

8 rows × 29 columns

In [15]:
```
#Checking the box plot to see the outlier treatment
```

Search ENG IN 14:45 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

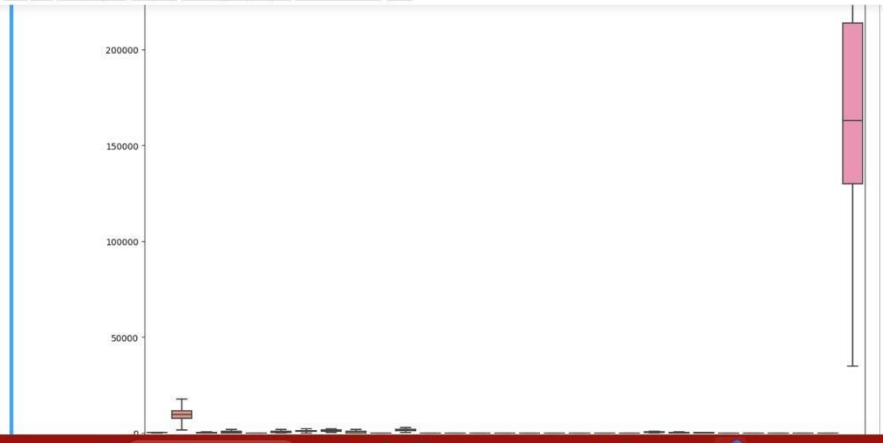
TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

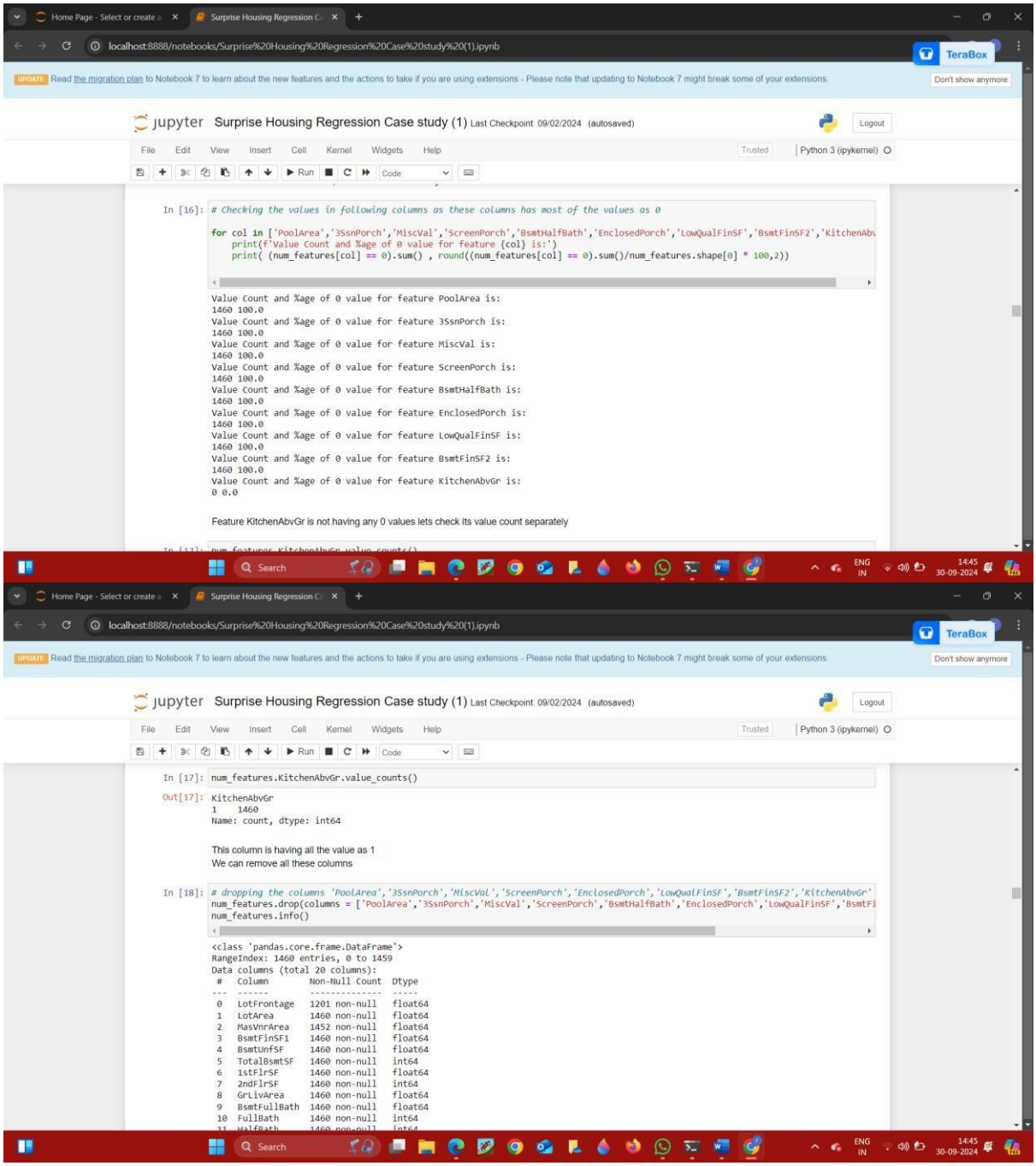
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code




Search ENG IN 14:45 30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [16]:
# Checking the values in following columns as these columns has most of the values as 0
for col in ['PoolArea','3SsnPorch','MiscVal','ScreenPorch','BsmtHalfBath','EnclosedPorch','LowQualFinSF','BsmtFinSF2','KitchenAbv
print(f'Value Count and %age of 0 value for feature {col} is:')
print( (num_features[col] == 0).sum() , round((num_features[col] == 0).sum()/num_features.shape[0] * 100,2))
Value Count and %age of 0 value for feature PoolArea is:
1460 100.0
Value Count and %age of 0 value for feature 3SsnPorch is:
1460 100.0
Value Count and %age of 0 value for feature MiscVal is:
1460 100.0
Value Count and %age of 0 value for feature ScreenPorch is:
1460 100.0
Value Count and %age of 0 value for feature BsmtHalfBath is:
1460 100.0
Value Count and %age of 0 value for feature EnclosedPorch is:
1460 100.0
Value Count and %age of 0 value for feature LowQualFinSF is:
1460 100.0
Value Count and %age of 0 value for feature BsmtFinSF2 is:
1460 100.0
Value Count and %age of 0 value for feature KitchenAbvGr is:
0 0.0
Feature KitchenAbvGr is not having any 0 values lets check its value count separately
Search
ENG
IN
14:45
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [17]:
num_features.KitchenAbvGr.value_counts()
Out[17]:
KitchenAbvGr
1 1460
Name: count, dtype: int64
This column is having all the value as 1
We can remove all these columns
In [18]:
# dropping the columns 'PoolArea','3SsnPorch','MiscVal','ScreenPorch','EnclosedPorch','LowQualFinSF','BsmtFinSF2','KitchenAbvGr'
num_features.drop(columns = ['PoolArea','3SsnPorch','MiscVal','ScreenPorch','BsmtHalfBath','EnclosedPorch','LowQualFinSF','BsmtFi
num_features.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 20 columns):
# Column Non-Null Count Dtype
--- ------ -------------- -----
0 LotFrontage 1201 non-null float64
1 LotArea 1460 non-null float64
2 MasVnrArea 1452 non-null float64
3 BsmtFinSF1 1460 non-null float64
4 BsmtUnfSF 1460 non-null float64
5 TotalBsmtSF 1460 non-null int64
6 1stFlrSF 1460 non-null float64
7 2ndFlrSF 1460 non-null int64
8 GrLivArea 1460 non-null float64
9 BsmtFullBath 1460 non-null float64
10 FullBath 1460 non-null int64
Search
ENG
IN
14:45
30-09-2024

Home Page - Select or create a | Surprise Housing Regression C

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [19]:
```
# creating heatmap

plt.figure(figsize=(25, 20))
sns.heatmap(num_features.corr(), annot=True, cmap="YlGnBu")
plt.show()
```

Search ENG IN 14:45 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

Here we can see some good correlations like:

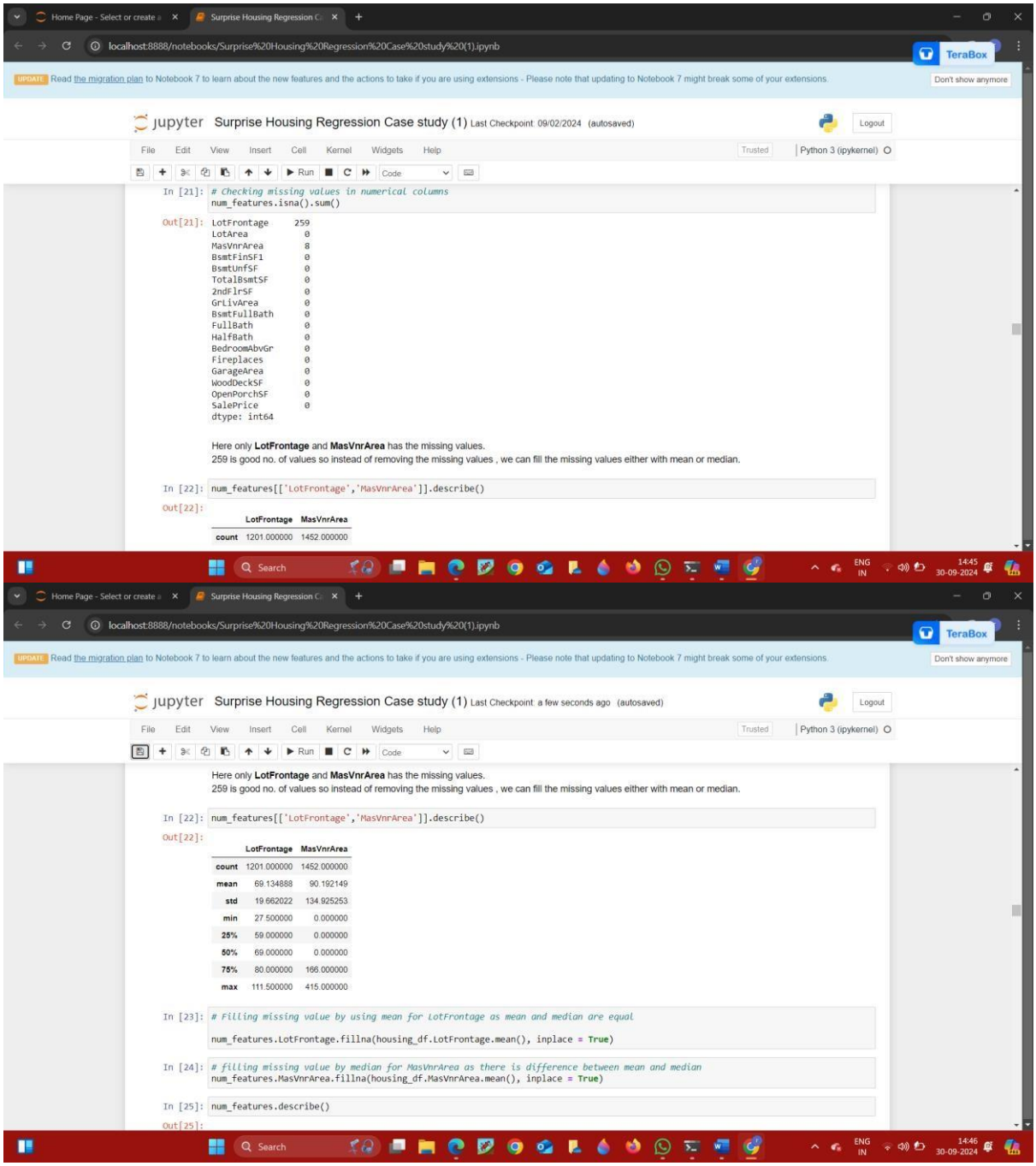
- GarageCars and GarageArea (0.89)
- TotRmsAbvGrd and GrLivArea (0.84)
- TotalBsmtSF and 1stFlrSF (0.81)
- Features like 'TotalBsmtSF', '1stFlrSF', 'GrLivArea' are showing positive correlation with SalePrice.

So we can remove GarageCars, TotRmsAbvGrd and 1stFlrSF

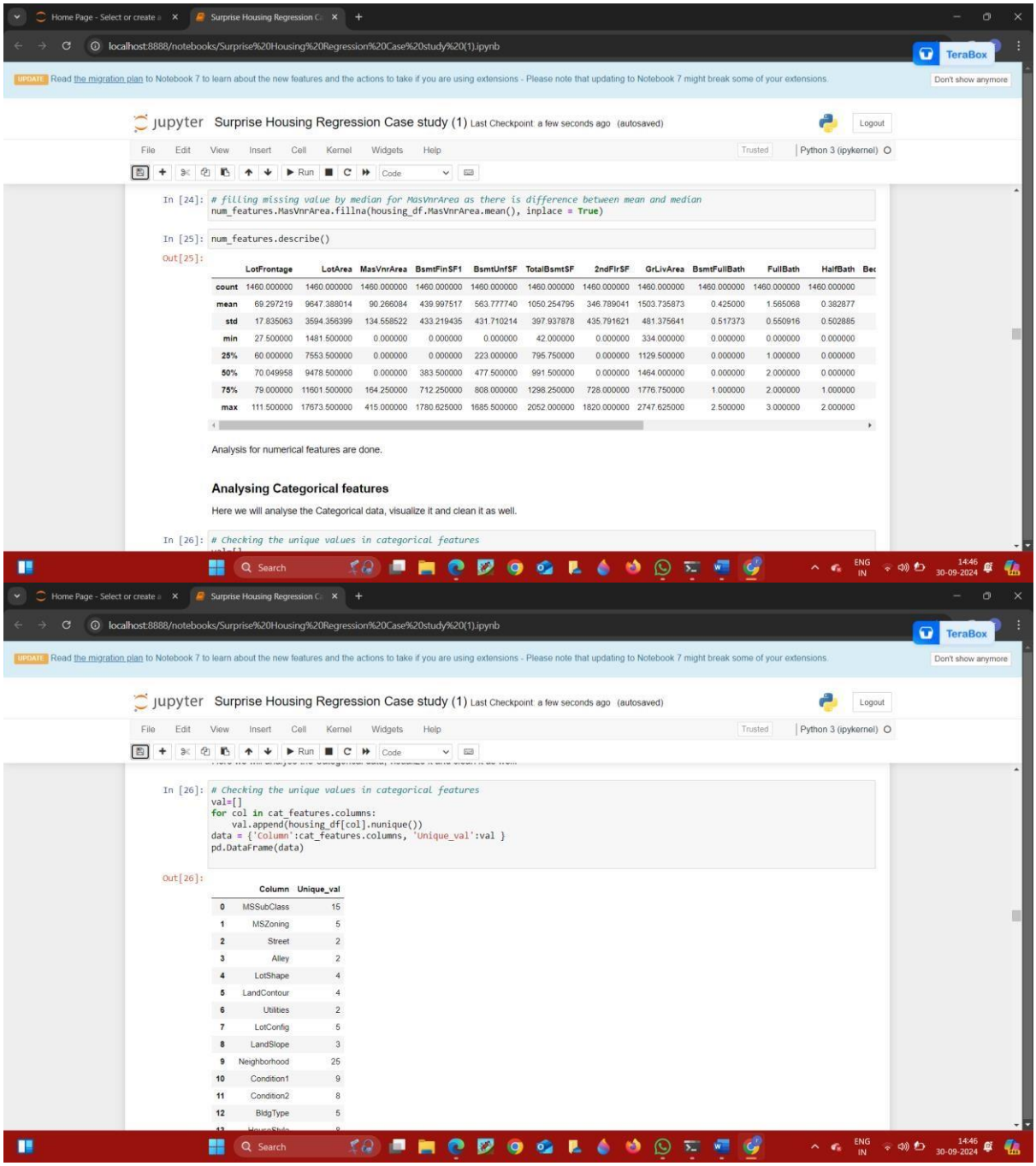
In [20]:
```
# dropping the columns GarageCars, TotRmsAbvGrd and 1stFlrSF
num_features.drop(columns=['GarageCars','TotRmsAbvGrd','1stFlrSF'], inplace = True)
num_features.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 17 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   LotFrontage   1201 non-null   float64
 1   LotArea       1460 non-null   float64
 2   MasVnrArea    1452 non-null   float64
 3   BsmtFinSF1    1460 non-null   float64
 4   BsmtUnfSF     1460 non-null   float64
 5   TotalBsmtSF   1460 non-null   int64
 6   2ndFlrSF      1460 non-null   int64
 7   GrLivArea     1460 non-null   float64
 8   BsmtFullBath  1460 non-null   float64
 9   FullBath      1460 non-null   int64
```

Search ENG IN 14:45 30-09-2024


Home Page - Select or create a
Surprise Housing Regression C
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 09/02/2024 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [21]: # Checking missing values in numerical columns
num_features.isna().sum()
Out[21]:
LotFrontage 259
LotArea 0
MasVnrArea 8
BsmtFinSF1 0
BsmtUnfSF 0
TotalBsmtSF 0
2ndFlrSF 0
GrLivArea 0
BsmtFullBath 0
FullBath 0
HalfBath 0
BedroomAbvGr 0
Fireplaces 0
GarageArea 0
WoodDeckSF 0
OpenPorchSF 0
SalePrice 0
dtype: int64
Here only LotFrontage and MasVnrArea has the missing values.
259 is good no. of values so instead of removing the missing values , we can fill the missing values either with mean or median.
In [22]: num_features[['LotFrontage','MasVnrArea']].describe()
Out[22]:
LotFrontage MasVnrArea
count 1201.000000 1452.000000
Search
ENG IN
14:45 30-09-2024


Home Page - Select or create a
Surprise Housing Regression C
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Here only LotFrontage and MasVnrArea has the missing values.
259 is good no. of values so instead of removing the missing values , we can fill the missing values either with mean or median.
In [22]: num_features[['LotFrontage','MasVnrArea']].describe()
Out[22]:
LotFrontage MasVnrArea
count 1201.000000 1452.000000
mean 69.134888 90.192149
std 19.662022 134.925253
min 27.500000 0.000000
25% 59.000000 0.000000
50% 69.000000 0.000000
75% 80.000000 166.000000
max 111.500000 415.000000
In [23]: # Filling missing value by using mean for LotFrontage as mean and median are equal
num_features.LotFrontage.fillna(housing_df.LotFrontage.mean(), inplace = True)
In [24]: # filling missing value by median for MasVnrArea as there is difference between mean and median
num_features.MasVnrArea.fillna(housing_df.MasVnrArea.mean(), inplace = True)
In [25]: num_features.describe()
Out[25]:
Search
ENG IN
14:46 30-09-2024



Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [24]:
```
# filling missing value by median for MasVnrArea as there is difference between mean and median
num_features.MasVnrArea.fillna(housing_df.MasVnrArea.mean(), inplace = True)
```

In [25]:
```
num_features.describe()
```

Out[25]:

| | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtUnfSF | TotalBsmtSF | 2ndFlrSF | GrLivArea | BsmtFullBath | FullBath | HalfBath | Bed |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | |
| mean | 69.297219 | 9647.388014 | 90.266084 | 439.997517 | 563.777740 | 1050.254795 | 346.789041 | 1503.735873 | 0.425000 | 1.565068 | 0.382877 | |
| std | 17.835063 | 3594.356399 | 134.558522 | 433.219435 | 431.710214 | 397.937878 | 435.791621 | 481.375641 | 0.517373 | 0.550916 | 0.502885 | |
| min | 27.500000 | 1481.500000 | 0.000000 | 0.000000 | 0.000000 | 42.000000 | 0.000000 | 334.000000 | 0.000000 | 0.000000 | 0.000000 | |
| 25% | 60.000000 | 7553.500000 | 0.000000 | 0.000000 | 223.000000 | 795.750000 | 0.000000 | 1129.500000 | 0.000000 | 1.000000 | 0.000000 | |
| 50% | 70.049958 | 9478.500000 | 0.000000 | 383.500000 | 477.500000 | 991.500000 | 0.000000 | 1464.000000 | 0.000000 | 2.000000 | 0.000000 | |
| 75% | 79.000000 | 11601.500000 | 164.250000 | 712.250000 | 808.000000 | 1298.250000 | 728.000000 | 1776.750000 | 1.000000 | 2.000000 | 1.000000 | |
| max | 111.500000 | 17673.500000 | 415.000000 | 1780.625000 | 1685.500000 | 2052.000000 | 1820.000000 | 2747.625000 | 2.500000 | 3.000000 | 2.000000 | |

Analysis for numerical features are done.

### Analysing Categorical features

Here we will analyse the Categorical data, visualize it and clean it as well.

In [26]:
```
# Checking the unique values in categorical features
```

Search ENG IN 14:46 30-09-2024



Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

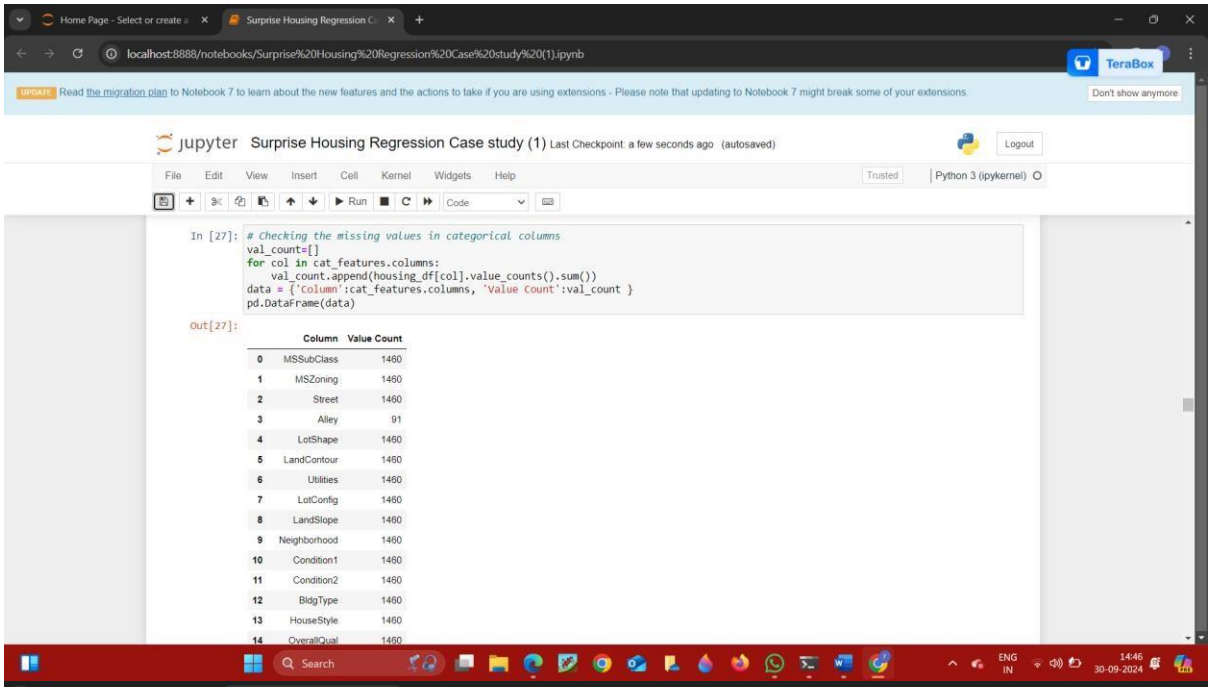
Run Code

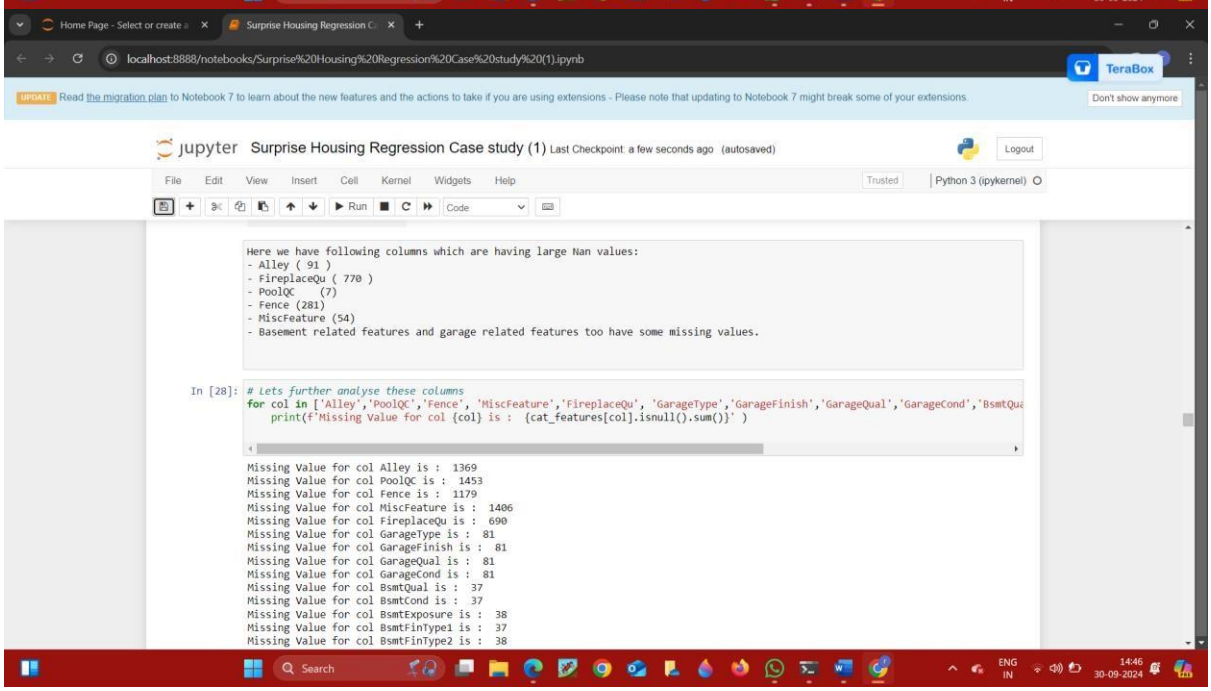
In [26]:
```
# Checking the unique values in categorical features
val=[]
for col in cat_features.columns:
    val.append(housing_df[col].nunique())
data = {'Column':cat_features.columns, 'Unique_val':val }
pd.DataFrame(data)
```

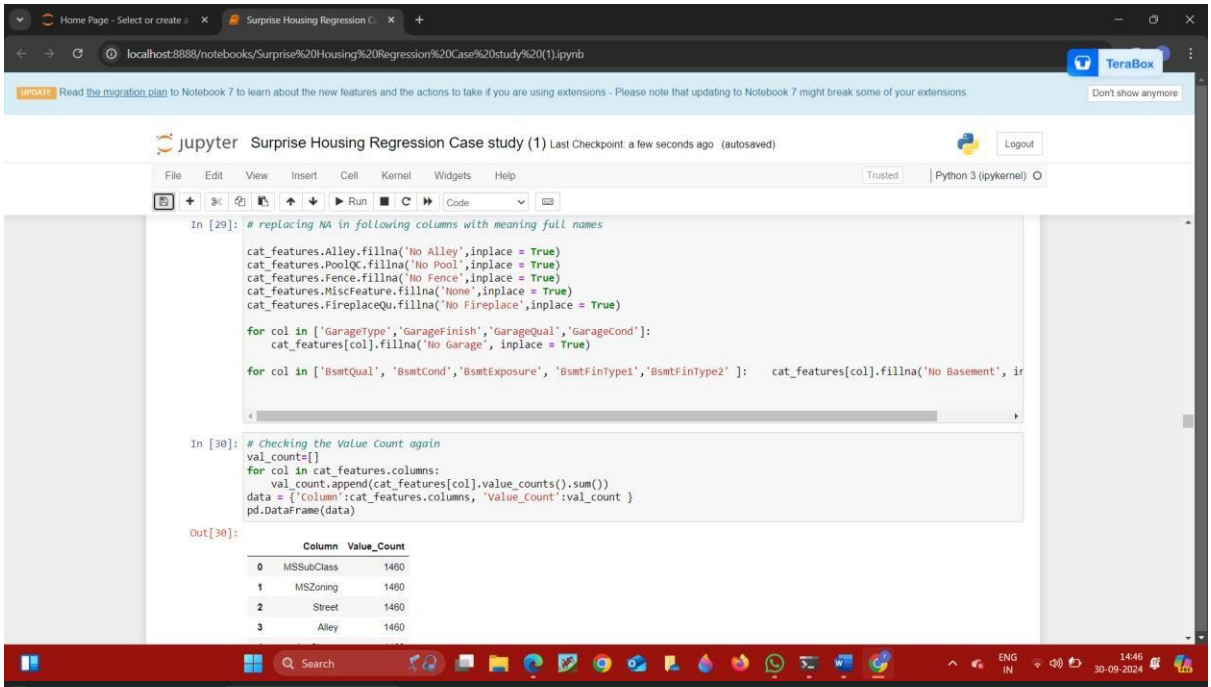
Out[26]:

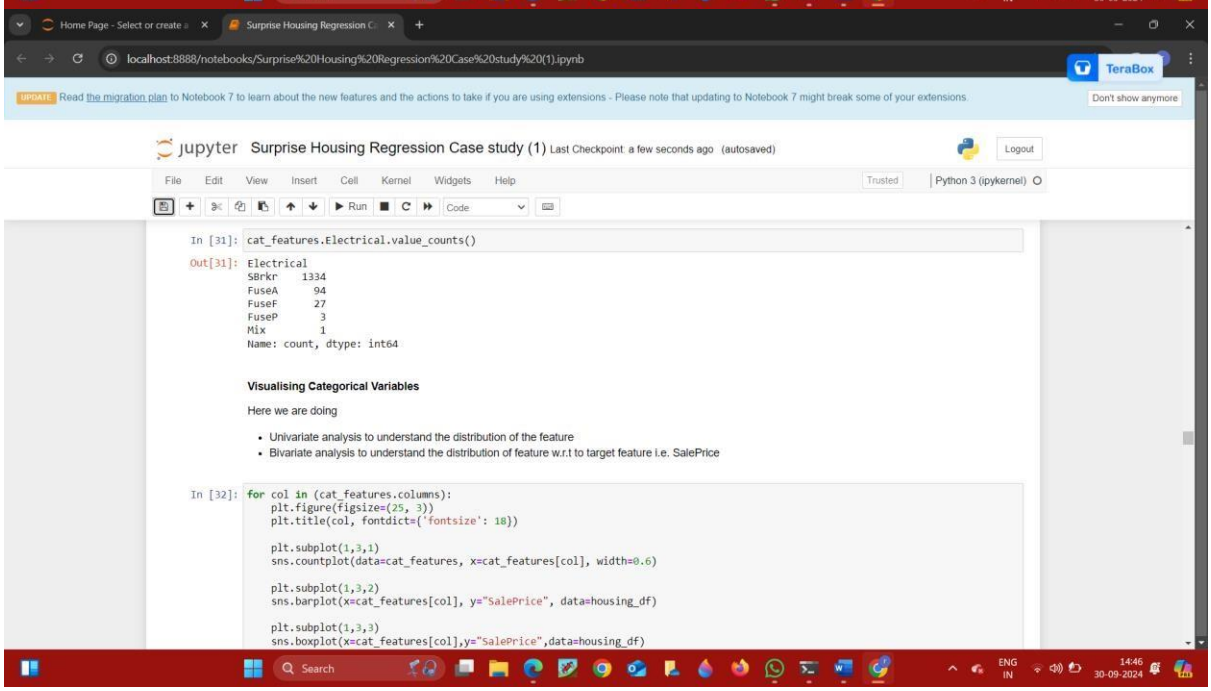
| | Column | Unique_val |
|---|---|---|
| 0 | MSSubClass | 15 |
| 1 | MSZoning | 5 |
| 2 | Street | 2 |
| 3 | Alley | 2 |
| 4 | LotShape | 4 |
| 5 | LandContour | 4 |
| 6 | Utilities | 2 |
| 7 | LotConfig | 5 |
| 8 | LandSlope | 3 |
| 9 | Neighborhood | 25 |
| 10 | Condition1 | 9 |
| 11 | Condition2 | 8 |
| 12 | BldgType | 5 |

Search ENG IN 14:46 30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [27]:
# Checking the missing values in categorical columns
val_count=[]
for col in cat_features.columns:
val_count.append(housing_df[col].value_counts().sum())
data = {'Column':cat_features.columns, 'Value Count':val_count }
pd.DataFrame(data)
Out[27]:
Column Value Count
0 MSSubClass 1460
1 MSZoning 1460
2 Street 1460
3 Alley 91
4 LotShape 1460
5 LandContour 1460
6 Utilities 1460
7 LotConfig 1460
8 LandSlope 1460
9 Neighborhood 1460
10 Condition1 1460
11 Condition2 1460
12 BldgType 1460
13 HouseStyle 1460
14 OverallQual 1460
Search
ENG
IN
14:46
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Here we have following columns which are having large Nan values:
- Alley ( 91 )
- FireplaceQu ( 770 )
- PoolQC (7)
- Fence (281)
- MiscFeature (54)
- Basement related features and garage related features too have some missing values.
In [28]:
# Lets further analyse these columns
for col in ['Alley','PoolQC','Fence', 'MiscFeature','FireplaceQu', 'GarageType','GarageFinish','GarageQual','GarageCond','BsmtQua
print(f'Missing Value for col {col} is : {cat_features[col].isnull().sum()}' )
Missing Value for col Alley is : 1369
Missing Value for col PoolQC is : 1453
Missing Value for col Fence is : 1179
Missing Value for col MiscFeature is : 1406
Missing Value for col FireplaceQu is : 690
Missing Value for col GarageType is : 81
Missing Value for col GarageFinish is : 81
Missing Value for col GarageQual is : 81
Missing Value for col GarageCond is : 81
Missing Value for col BsmtQual is : 37
Missing Value for col BsmtCond is : 37
Missing Value for col BsmtExposure is : 38
Missing Value for col BsmtFinType1 is : 37
Missing Value for col BsmtFinType2 is : 38
Search
ENG
IN
14:46
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [29]:
# replacing NA in following columns with meaning full names
cat_features.Alley.fillna('No Alley',inplace = True)
cat_features.PoolQC.fillna('No Pool',inplace = True)
cat_features.Fence.fillna('No Fence',inplace = True)
cat_features.MiscFeature.fillna('None',inplace = True)
cat_features.FireplaceQu.fillna('No Fireplace',inplace = True)
for col in ['GarageType','GarageFinish','GarageQual','GarageCond']:
cat_features[col].fillna('No Garage', inplace = True)
for col in ['BsmtQual', 'BsmtCond','BsmtExposure', 'BsmtFinType1','BsmtFinType2' ]: cat_features[col].fillna('No Basement', in
In [30]:
# Checking the Value Count again
val_count=[]
for col in cat_features.columns:
val_count.append(cat_features[col].value_counts().sum())
data = {'Column':cat_features.columns, 'Value_Count':val_count }
pd.DataFrame(data)
Out[30]:
Column Value_Count
0 MSSubClass 1460
1 MSZoning 1460
2 Street 1460
3 Alley 1460
Search
ENG
IN
14:46
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [31]:
cat_features.Electrical.value_counts()
Out[31]:
Electrical
SBrkr 1334
FuseA 94
FuseF 27
FuseP 3
Mix 1
Name: count, dtype: int64
Visualising Categorical Variables
Here we are doing
Univariate analysis to understand the distribution of the feature
Bivariate analysis to understand the distribution of feature w.r.t to target feature i.e. SalePrice
In [32]:
for col in (cat_features.columns):
plt.figure(figsize=(25, 3))
plt.title(col, fontdict={'fontsize': 18})
plt.subplot(1,3,1)
sns.countplot(data=cat_features, x=cat_features[col], width=0.6)
plt.subplot(1,3,2)
sns.barplot(x=cat_features[col], y="SalePrice", data=housing_df)
plt.subplot(1,3,3)
sns.boxplot(x=cat_features[col],y="SalePrice",data=housing_df)
Search
ENG
IN
14:46
30-09-2024

Home Page - Select or create a | Surprise Housing Regression C

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

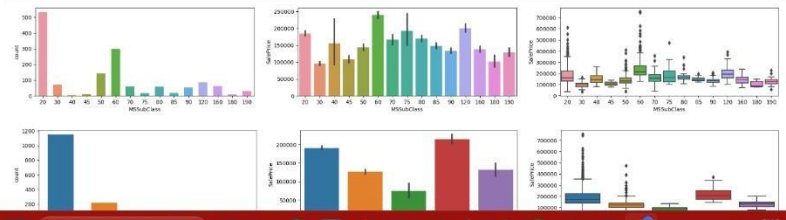
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [32]:
```
for col in (cat_features.columns):
    plt.figure(figsize=(25, 3))
    plt.title(col, fontdict={'fontsize': 18})

    plt.subplot(1,3,1)
    sns.countplot(data=cat_features, x=cat_features[col], width=0.6)

    plt.subplot(1,3,2)
    sns.barplot(x=cat_features[col], y="SalePrice", data=housing_df)

    plt.subplot(1,3,3)
    sns.boxplot(x=cat_features[col],y="SalePrice",data=housing_df)

    plt.show()
```



Search ENG IN 14:46 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

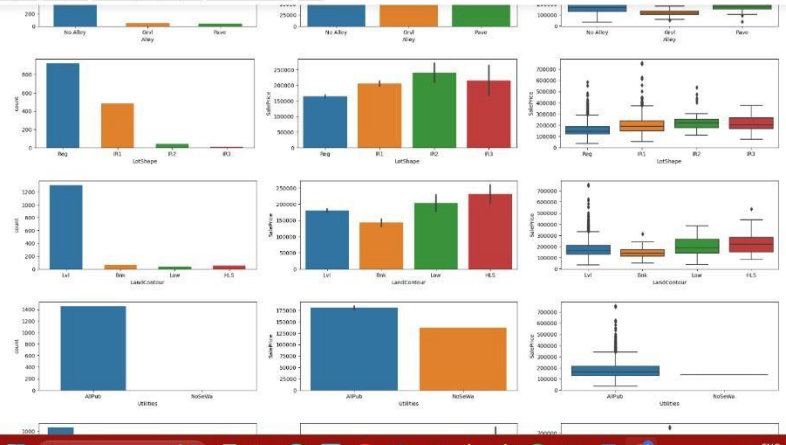
TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

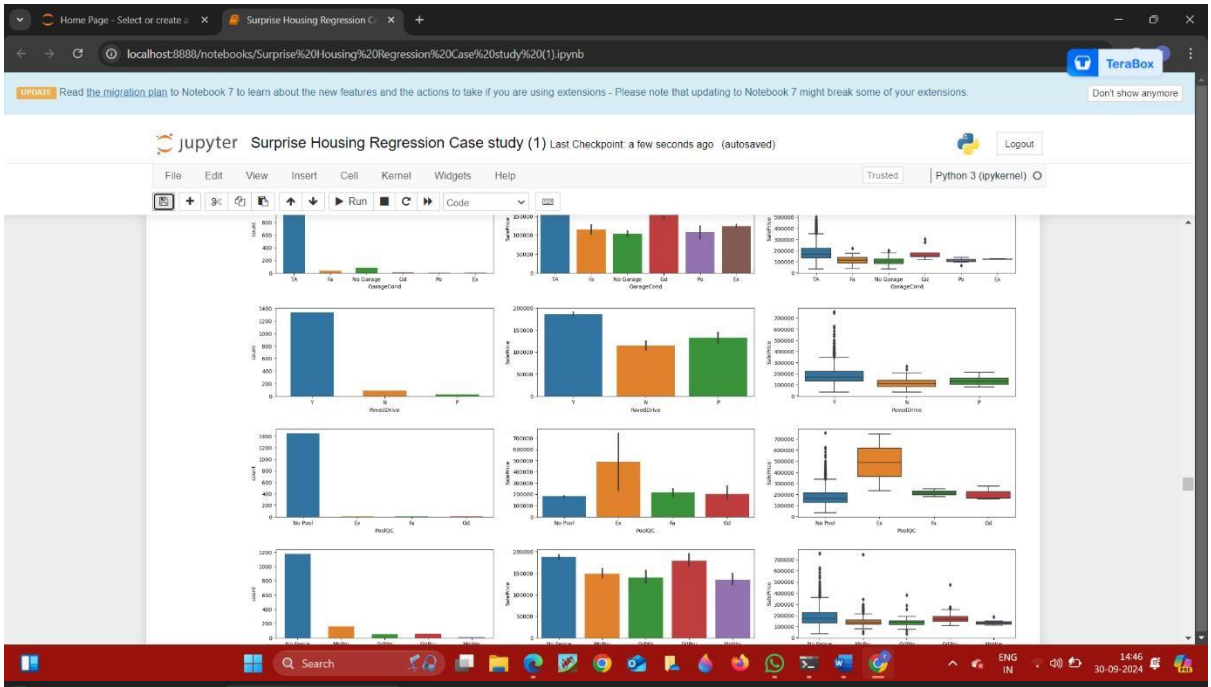
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code



Search ENG IN 14:46 30-09-2024



Home Page - Select or create a | Surprise Housing Regression Ca

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved)

Logout

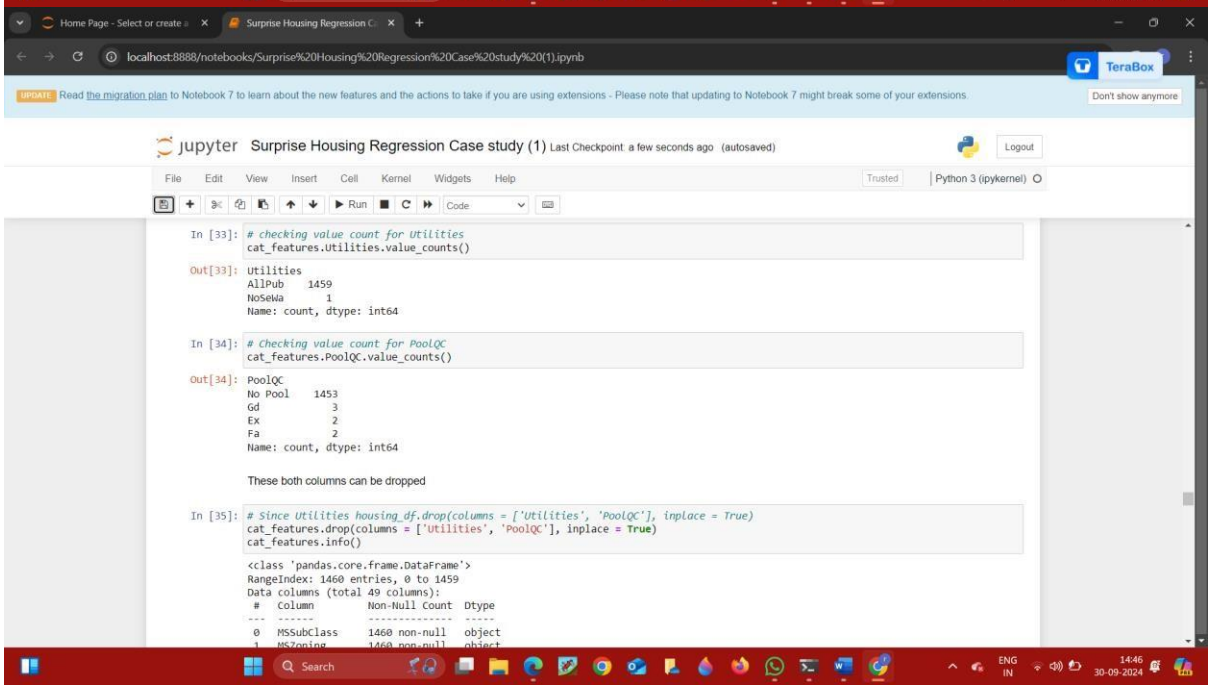
File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code



Search ENG IN 14:46 30-09-2024



Home Page - Select or create a | Surprise Housing Regression Ca

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3 (ipykernel)

Run Code

In [33]:
```
# checking value count for Utilities
cat_features.Utilities.value_counts()
```

Out[33]:
```
Utilities
AllPub    1459
NoSeWa       1
Name: count, dtype: int64
```

In [34]:
```
# Checking value count for PoolQC
cat_features.PoolQC.value_counts()
```

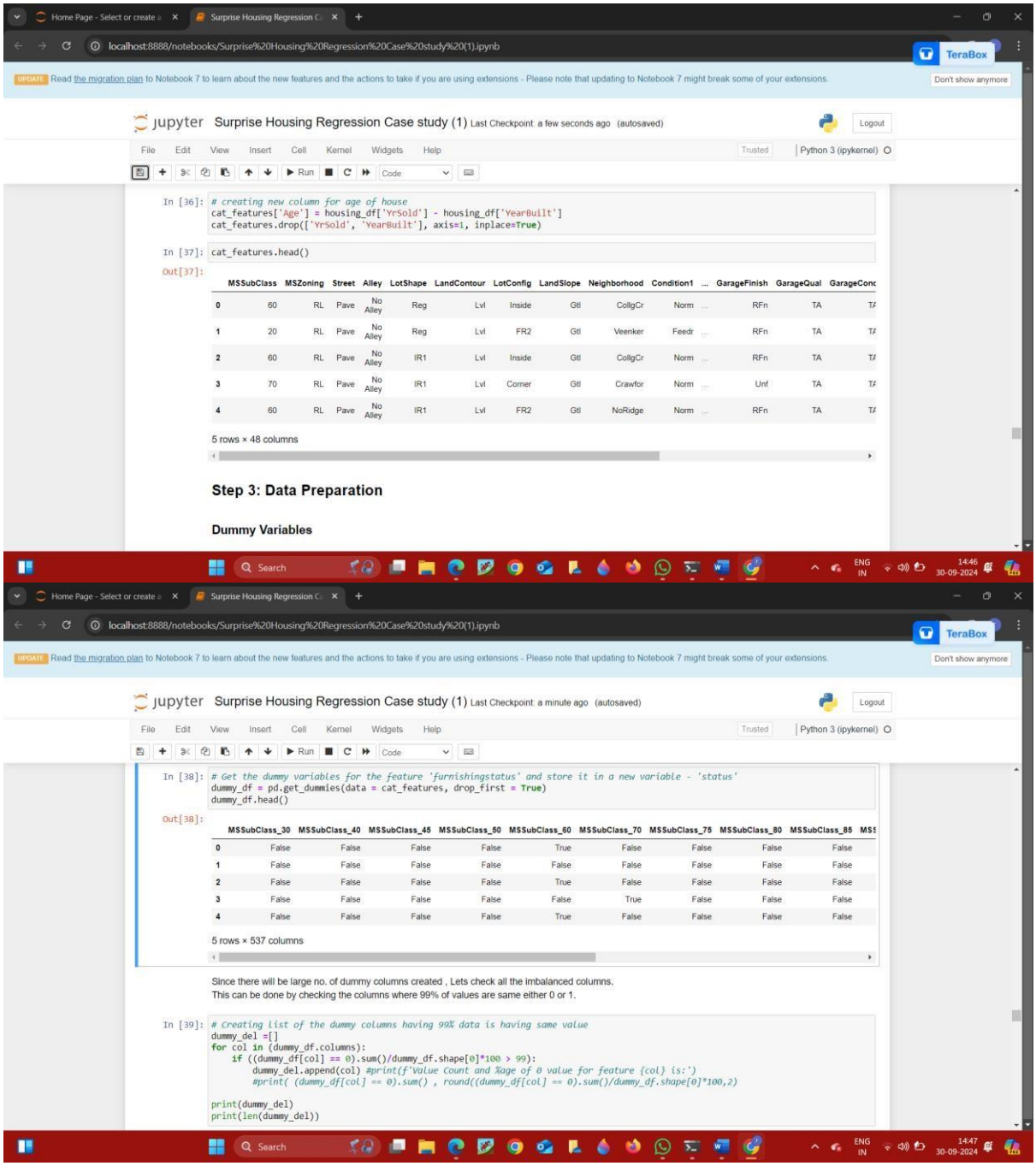
Out[34]:
```
PoolQC
No Pool    1453
Gd            3
Ex            2
Fa            2
Name: count, dtype: int64
```

These both columns can be dropped

In [35]:
```
# Since Utilities housing_df.drop(columns = ['Utilities', 'PoolQC'], inplace = True)
cat_features.drop(columns = ['Utilities', 'PoolQC'], inplace = True)
cat_features.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 49 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   MSSubClass     1460 non-null   object
 1   MSZoning       1460 non-null   object
```

Search ENG IN 14:46 30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [36]:
# creating new column for age of house
cat_features['Age'] = housing_df['YrSold'] - housing_df['YearBuilt']
cat_features.drop(['YrSold', 'YearBuilt'], axis=1, inplace=True)
In [37]:
cat_features.head()
Out[37]:
MSSubClass MSZoning Street Alley LotShape LandContour LotConfig LandSlope Neighborhood Condition1 ... GarageFinish GarageQual GarageCond
0 60 RL Pave No Alley Reg Lvl Inside Gtl CollgCr Norm ... RFn TA TA
1 20 RL Pave No Alley Reg Lvl FR2 Gtl Veenker Feedr ... RFn TA TA
2 60 RL Pave No Alley IR1 Lvl Inside Gtl CollgCr Norm ... RFn TA TA
3 70 RL Pave No Alley IR1 Lvl Corner Gtl Crawfor Norm ... Unf TA TA
4 60 RL Pave No Alley IR1 Lvl FR2 Gtl NoRidge Norm ... RFn TA TA
5 rows × 48 columns
Step 3: Data Preparation
Dummy Variables
Search
ENG IN
14:46 30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a minute ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [38]:
# Get the dummy variables for the feature 'furnishingstatus' and store it in a new variable - 'status'
dummy_df = pd.get_dummies(data = cat_features, drop_first = True)
dummy_df.head()
Out[38]:
MSSubClass_30 MSSubClass_40 MSSubClass_45 MSSubClass_50 MSSubClass_60 MSSubClass_70 MSSubClass_75 MSSubClass_80 MSSubClass_85 MS
0 False False False False True False False False False
1 False False False False False False False False False
2 False False False False True False False False False
3 False False False False False True False False False
4 False False False False True False False False False
5 rows × 537 columns
Since there will be large no. of dummy columns created , Lets check all the imbalanced columns.
This can be done by checking the columns where 99% of values are same either 0 or 1.
In [39]:
# Creating list of the dummy columns having 99% data is having same value
dummy_del =[]
for col in (dummy_df.columns):
    if ((dummy_df[col] == 0).sum()/dummy_df.shape[0]*100 > 99):
        dummy_del.append(col) #print(f'Value Count and %age of 0 value for feature {col} is:')
        #print( (dummy_df[col] == 0).sum() , round((dummy_df[col] == 0).sum()/dummy_df.shape[0]*100,2)

print(dummy_del)
print(len(dummy_del))
Search
ENG IN
14:47 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a minute ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [40]:
```
# dropping all the columns which are imbalanced in nature
dummy_df.drop(columns = dummy_del, inplace = True)
print(dummy_df.shape)
```

```
(1460, 295)
```

In [41]:
```
num_features.columns
```

Out[41]:
```
Index(['LotFrontage', 'LotArea', 'MasVnrArea', 'BsmtFinSF1', 'BsmtUnfSF',
       'TotalBsmtSF', '2ndFlrSF', 'GrLivArea', 'BsmtFullBath', 'FullBath',
       'HalfBath', 'BedroomAbvGr', 'Fireplaces', 'GarageArea', 'WoodDeckSF',
       'OpenPorchSF', 'SalePrice'],
      dtype='object')
```

In [42]:
```
# Concating the numrical features and dummy features created out of categorical features
df = pd.concat([num_features,dummy_df], axis = 1)
df.head()
```

Out[42]:

| | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtUnfSF | TotalBsmtSF | 2ndFlrSF | GrLivArea | BsmtFullBath | FullBath | ... | Age_46 | Age_47 | Age_48 | Age_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 65.0 | 8450.0 | 196.0 | 706.0 | 150.0 | 856 | 854 | 1710.0 | 1.0 | 2 | ... | False | False | False | Fal |
| 1 | 80.0 | 9600.0 | 0.0 | 978.0 | 284.0 | 1262 | 0 | 1262.0 | 0.0 | 2 | ... | False | False | False | Fal |
| 2 | 68.0 | 11250.0 | 162.0 | 486.0 | 434.0 | 920 | 866 | 1786.0 | 1.0 | 2 | ... | False | False | False | Fal |
| 3 | 60.0 | 9550.0 | 0.0 | 216.0 | 540.0 | 756 | 756 | 1717.0 | 1.0 | 1 | ... | False | False | False | Fal |
| 4 | 84.0 | 14260.0 | 350.0 | 655.0 | 490.0 | 1145 | 1053 | 2198.0 | 1.0 | 2 | ... | False | False | False | Fal |

5 rows × 312 columns

Search ENG IN 14:47 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a minute ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

## Step 4: Splitting the Data into Training and Testing Sets

As you know, the first basic step for regression is performing a train-test split.

In [43]:
```
# creating independent data set by dropping dependent variable SalePrice
X = df.drop(['SalePrice'], axis = 1)
X.head()
```

Out[43]:

| | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtUnfSF | TotalBsmtSF | 2ndFlrSF | GrLivArea | BsmtFullBath | FullBath | ... | Age_46 | Age_47 | Age_48 | Age_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 65.0 | 8450.0 | 196.0 | 706.0 | 150.0 | 856 | 854 | 1710.0 | 1.0 | 2 | ... | False | False | False | Fal |
| 1 | 80.0 | 9600.0 | 0.0 | 978.0 | 284.0 | 1262 | 0 | 1262.0 | 0.0 | 2 | ... | False | False | False | Fal |
| 2 | 68.0 | 11250.0 | 162.0 | 486.0 | 434.0 | 920 | 866 | 1786.0 | 1.0 | 2 | ... | False | False | False | Fal |
| 3 | 60.0 | 9550.0 | 0.0 | 216.0 | 540.0 | 756 | 756 | 1717.0 | 1.0 | 1 | ... | False | False | False | Fal |
| 4 | 84.0 | 14260.0 | 350.0 | 655.0 | 490.0 | 1145 | 1053 | 2198.0 | 1.0 | 2 | ... | False | False | False | Fal |

5 rows × 311 columns

In [44]:
```
# we have already seen the distribution of sales price before. Lets check it again
plt.title('Distribution of SalePrice')
sns.distplot(num_features['SalePrice'])
plt.show()
```


Search ENG IN 14:47 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

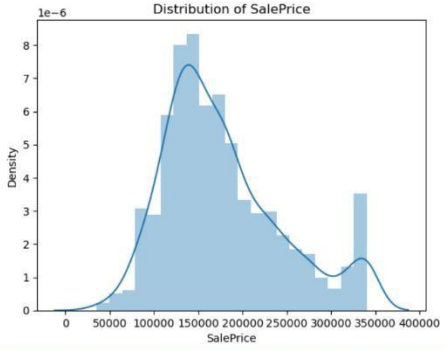
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a minute ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [44]:
```
# we have already seen the distribution of sales price before. Lets check it again
plt.title('Distribution of SalePrice')
sns.distplot(num_features['SalePrice'])
plt.show()
```




Search ENG IN 14:47 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a minute ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

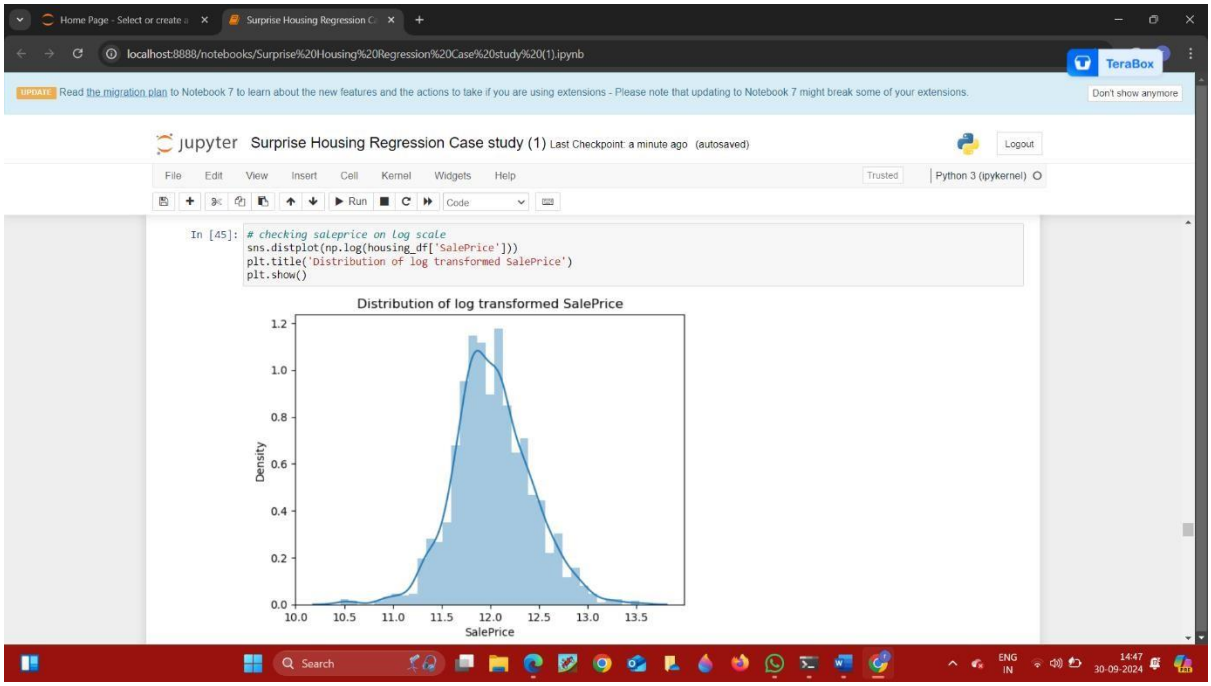
```
y = np.log(num_features['SalePrice'])
print(y)
```

```
0       12.247694
1       12.109011
2       12.317167
3       11.849398
4       12.429216
          ...
1455    12.072541
1456    12.254863
1457    12.493130
1458    11.864462
1459    11.901583
Name: SalePrice, Length: 1460, dtype: float64
```

In [47]:
```
# Splitting into train and test data

# We specify this so that the train and test data set always have the same rows, respectively
np.random.seed(0)
X_train, X_test, y_train, y_test = train_test_split(X,y, train_size = 0.7, test_size = 0.3, random_state = 100)
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(1021, 311)
(438, 311)
(1021,)
(438,)
```

Search ENG IN 14:47 30-09-2024



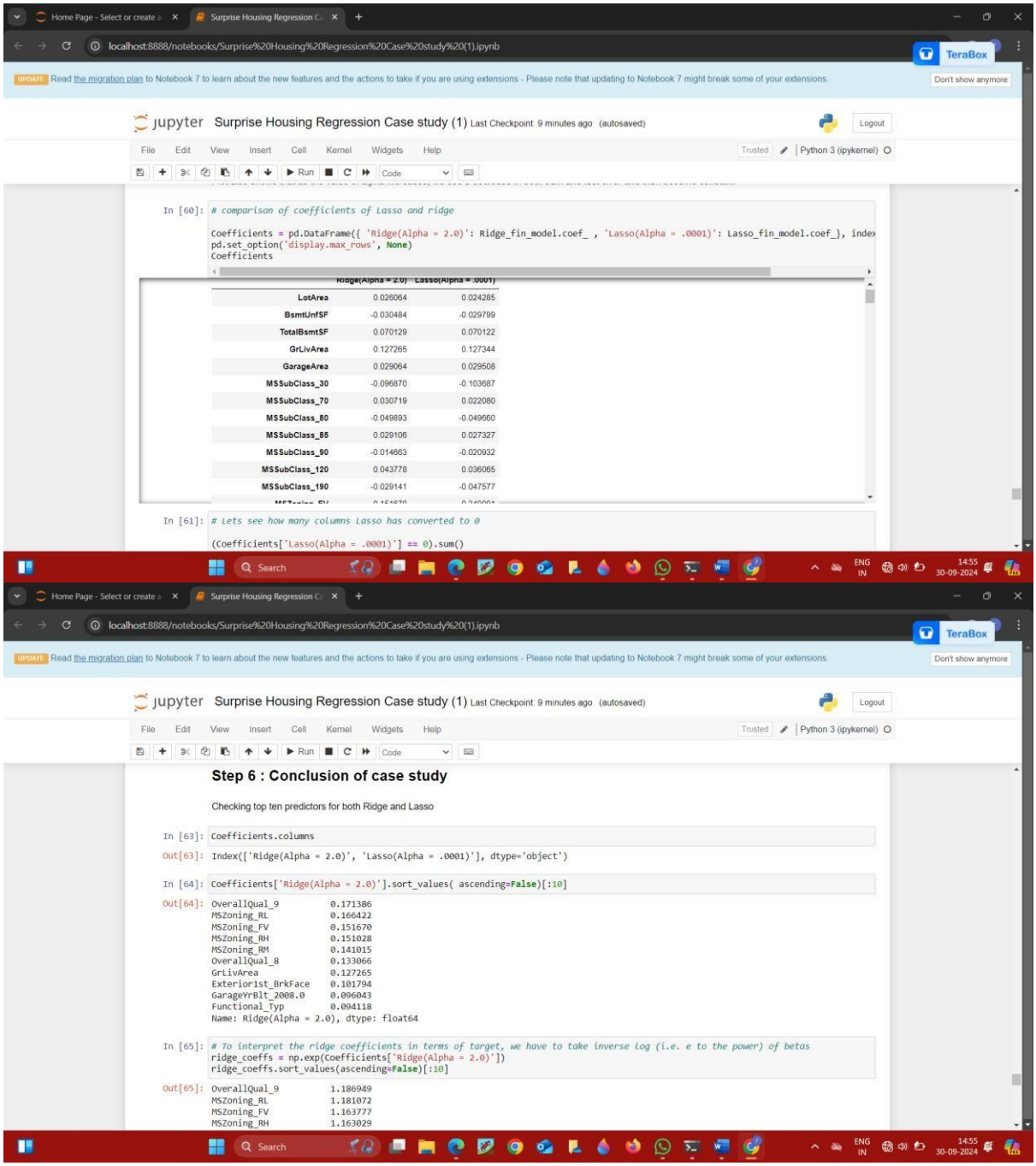

**Rescaling the Features**

It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might cause issues at the time of model evaluation. We should therefore use standardization or normalization so that units of the coefficients obtained are all on the same scale. Here we have two common ways of rescaling:

1. Min-Max scaling ( normailzation) ( x-xmin)/(xmax - xmin)
2. Standardisation (mean-0, sigma-1) ( x- mU)/sigma

This time, we will use Standardisation scaling.

from sklearn.preprocessing import StandardScaler


Home Page - Select or create
Surprise Housing Regression C
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 9 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [60]:
# comparison of coefficients of Lasso and ridge
Coefficients = pd.DataFrame({ 'Ridge(Alpha = 2.0)': Ridge_fin_model.coef_ , 'Lasso(Alpha = .0001)': Lasso_fin_model.coef_}, index
pd.set_option('display.max_rows', None)
Coefficients
Ridge(Alpha = 2.0) Lasso(Alpha = .0001)
LotArea 0.026064 0.024285
BsmtUnfSF -0.030484 -0.029799
TotalBsmtSF 0.070129 0.070122
GrLivArea 0.127265 0.127344
GarageArea 0.029064 0.029508
MSSubClass_30 -0.096870 -0.103687
MSSubClass_70 0.030719 0.022080
MSSubClass_80 -0.049893 -0.049660
MSSubClass_85 0.029106 0.027327
MSSubClass_90 -0.014663 -0.020932
MSSubClass_120 0.043778 0.036065
MSSubClass_190 -0.029141 -0.047577
In [61]:
# Lets see how many columns Lasso has converted to 0
(Coefficients['Lasso(Alpha = .0001)'] == 0).sum()
Search
ENG IN
14:55 30-09-2024


Home Page - Select or create
Surprise Housing Regression C
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 9 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Step 6 : Conclusion of case study
Checking top ten predictors for both Ridge and Lasso
In [63]:
Coefficients.columns
Out[63]:
Index(['Ridge(Alpha = 2.0)', 'Lasso(Alpha = .0001)'], dtype='object')
In [64]:
Coefficients['Ridge(Alpha = 2.0)'].sort_values( ascending=False)[:10]
Out[64]:
OverallQual_9 0.171386
MSZoning_RL 0.166422
MSZoning_FV 0.151670
MSZoning_RH 0.151028
MSZoning_RM 0.141015
OverallQual_8 0.133066
GrLivArea 0.127265
Exterior1st_BrkFace 0.101794
GarageYrBlt_2008.0 0.096043
Functional_Typ 0.094118
Name: Ridge(Alpha = 2.0), dtype: float64
In [65]:
# To interpret the ridge coefficients in terms of target, we have to take inverse log (i.e. e to the power) of betas
ridge_coeffs = np.exp(Coefficients['Ridge(Alpha = 2.0)'])
ridge_coeffs.sort_values(ascending=False)[:10]
Out[65]:
OverallQual_9 1.186949
MSZoning_RL 1.181072
MSZoning_FV 1.163777
MSZoning_RH 1.163029
Search
ENG IN
14:55 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: a minute ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
y = np.log(num_features['SalePrice'])
print(y)
```

```
0       12.247694
1       12.109011
2       12.317167
3       11.849398
4       12.429216
          ...
1455    12.072541
1456    12.254863
1457    12.493130
1458    11.864462
1459    11.901583
Name: SalePrice, Length: 1460, dtype: float64
```

In [47]:
```
# Splitting into train and test data

# We specify this so that the train and test data set always have the same rows, respectively
np.random.seed(0)
X_train, X_test, y_train, y_test = train_test_split(X,y, train_size = 0.7, test_size = 0.3, random_state = 100)
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(1021, 311)
(438, 311)
(1021,)
(438,)
```

Search ENG IN 14:47 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

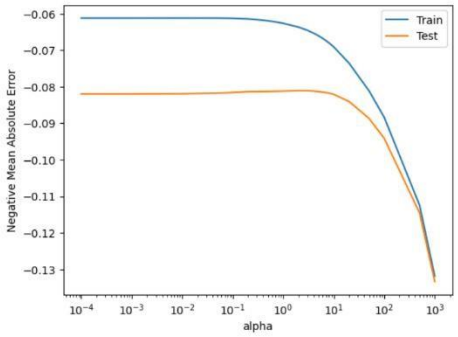
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 5 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
R-Square for test:0.8659669311749246
RMSE for train:0.08972875378486463
RMSE for test:0.13994770973204124
```




**Inferences**

Search ENG IN 14:51 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 5 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [48]:
```
# creating object of standard scaler
scaler = StandardScaler()
```

In [49]:
```
# Apply scaler() to all the columns except the 'yes-no' and 'dummy' variables i.e. all numerical variables
# here we get the no. of columns
num_cols = list(X_train.select_dtypes(include=['int64', 'float64']).columns)

# Scaling both test and train data set
X_train[num_cols] = scaler.fit_transform(X_train[num_cols])
X_test[num_cols] = scaler.transform(X_test[num_cols])
```

In [50]:
```
X_train.head()
```

Out[50]:

| | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtUnfSF | TotalBsmtSF | 2ndFlrSF | GrLivArea | BsmtFullBath | FullBath | ... | Age_46 | Age_47 | Age_48 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **210** | -0.108024 | -1.117166 | -0.665113 | 0.057374 | -0.375237 | -0.463273 | -0.788850 | -1.338981 | 1.072205 | -1.026085 | ... | False | False | False |
| **318** | 1.168096 | 0.081094 | 1.227470 | 1.252830 | -0.457448 | 0.725402 | 2.146453 | 2.394669 | 1.072205 | 0.786428 | ... | False | False | False |
| **239** | -0.940276 | -0.242180 | -0.665113 | -0.804090 | 0.184253 | -0.780745 | 0.798610 | -0.033670 | -0.839972 | -1.026085 | ... | False | False | False |
| **986** | -0.551892 | -1.199170 | -0.665113 | -1.020608 | -0.171994 | -1.396001 | 0.671889 | 0.276419 | -0.839972 | -1.026085 | ... | False | False | False |
| **1416** | -0.496408 | 0.482745 | -0.665113 | -1.020608 | 0.494827 | -0.677382 | 1.616532 | 1.648777 | -0.839972 | 0.786428 | ... | False | False | False |

5 rows × 311 columns

In [51]:
```
X_train.describe()
```

Out[51]:

| | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtUnfSF | TotalBsmtSF | 2ndFlrSF | GrLivArea | BsmtFullBath | FullBath |
|---|---|---|---|---|---|---|---|---|---|---|

Search ENG IN 14:50 30-09-2024

Home Page - Select or create a | Surprise Housing Regression C

localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

TeraBox

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions. Don't show anymore

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 5 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

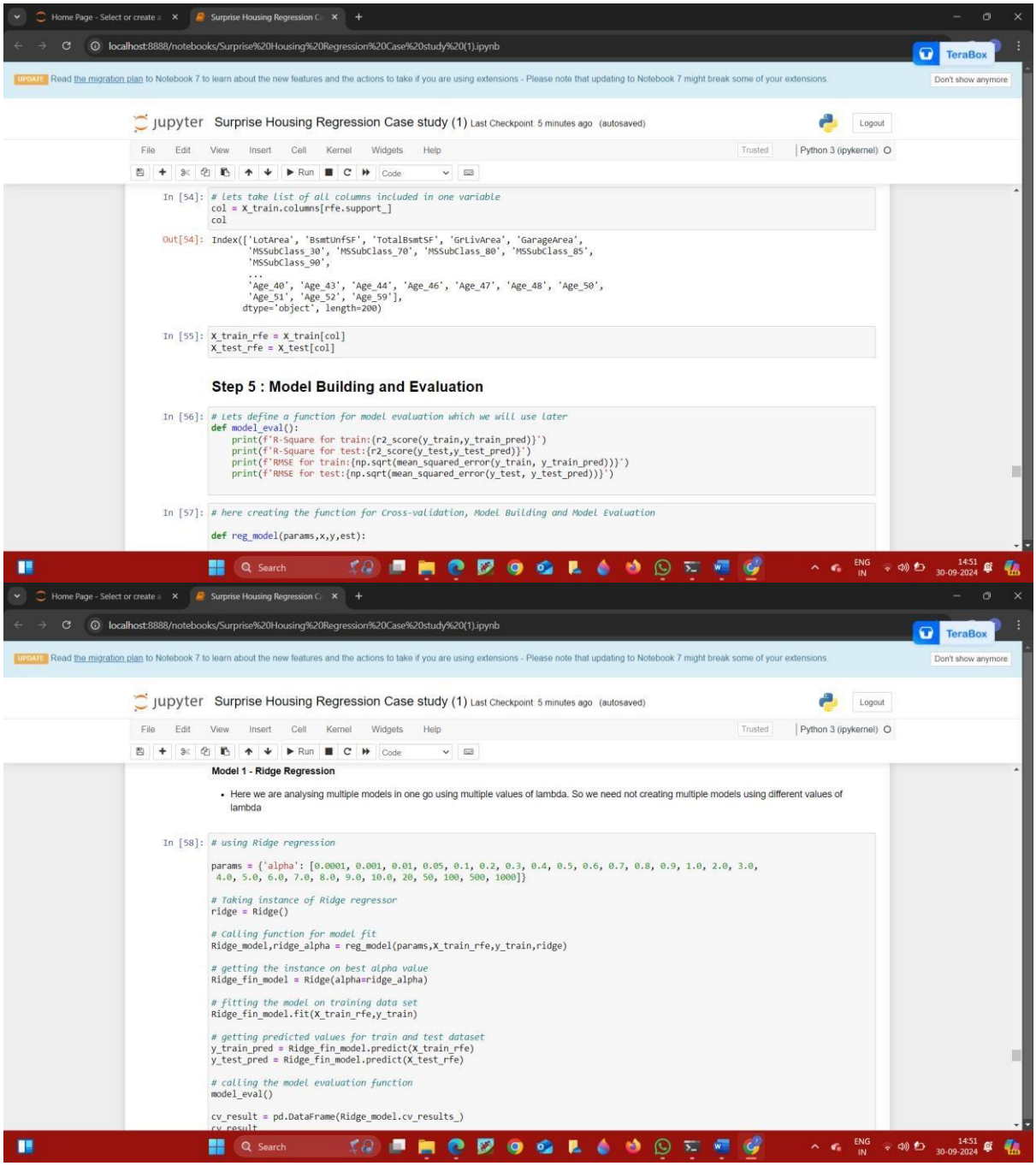
In [52]:
```
# Running RFE with output number of variables as 200

lmodel = LinearRegression()
lmodel.fit(X_train,y_train)

rfe = RFE(lmodel , n_features_to_select=200)
rfe = rfe.fit(X_train, y_train)
```

In [53]:
```
# list of all the columns, their inclusion in model, ranking

list = pd.DataFrame({'Columns' : X_train.columns, 'Included': rfe.support_, 'Ranking' :rfe.ranking_})
list.sort_values(by = 'Ranking')
```

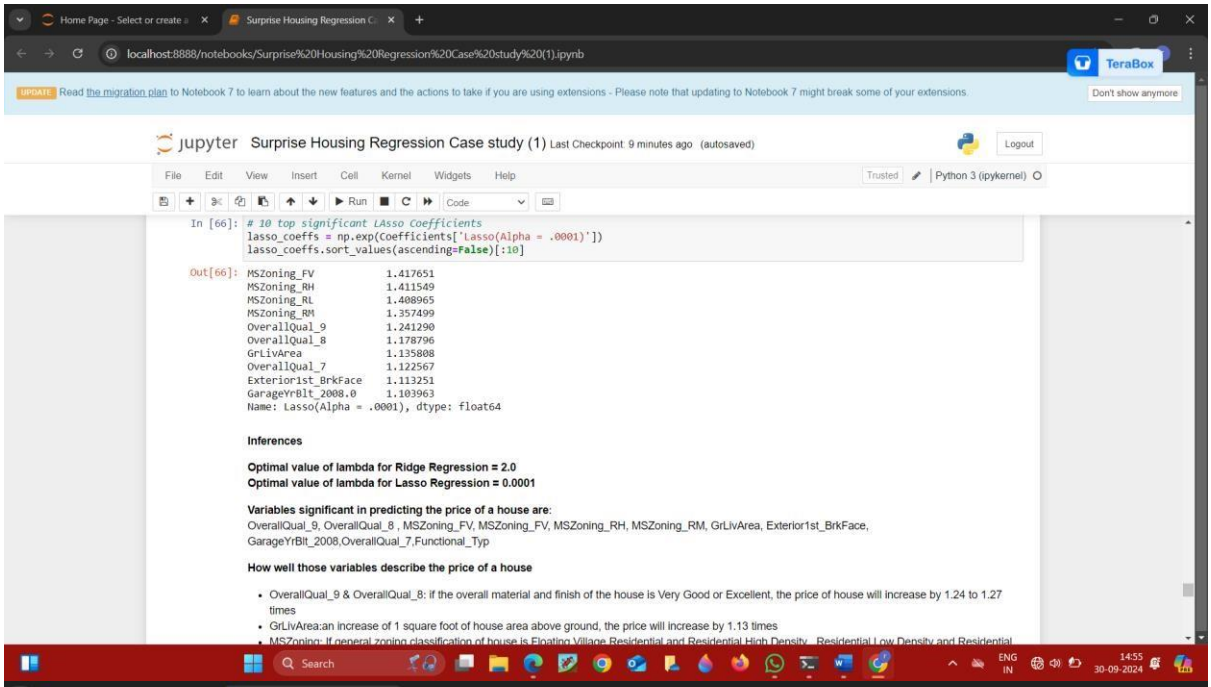
Out[53]:

| | Columns | Included | Ranking |
|---|---|---|---|
| **155** | BsmtQual_No Basement | True | 1 |
| **166** | BsmtFinType1_LwQ | True | 1 |
| **167** | BsmtFinType1_No Basement | True | 1 |
| **168** | BsmtFinType1_Rec | True | 1 |
| **169** | BsmtFinType1_Unf | True | 1 |
| ... | ... | ... | ... |
| **182** | Electrical_SBrkr | False | 108 |
| **0** | LotFrontage | False | 109 |
| **140** | Exterior2nd_VinylSd | False | 110 |
| **11** | BedroomAbvGr | False | 111 |

Search ENG IN 14:51 30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 5 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [54]:
# Lets take list of all columns included in one variable
col = X_train.columns[rfe.support_]
col
Out[54]:
Index(['LotArea', 'BsmtUnfSF', 'TotalBsmtSF', 'GrLivArea', 'GarageArea',
'MSSubClass_30', 'MSSubClass_70', 'MSSubClass_80', 'MSSubClass_85',
'MSSubClass_90',
...
'Age_40', 'Age_43', 'Age_44', 'Age_46', 'Age_47', 'Age_48', 'Age_50',
'Age_51', 'Age_52', 'Age_59'],
dtype='object', length=200)
In [55]:
X_train_rfe = X_train[col]
X_test_rfe = X_test[col]
Step 5 : Model Building and Evaluation
In [56]:
# Lets define a function for model evaluation which we will use later
def model_eval():
print(f'R-Square for train:{r2_score(y_train,y_train_pred)}')
print(f'R-Square for test:{r2_score(y_test,y_test_pred)}')
print(f'RMSE for train:{np.sqrt(mean_squared_error(y_train, y_train_pred))}')
print(f'RMSE for test:{np.sqrt(mean_squared_error(y_test, y_test_pred))}')
In [57]:
# here creating the function for Cross-validation, Model Building and Model Evaluation
def reg_model(params,x,y,est):
Search
ENG
IN
14:51
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 5 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Model 1 - Ridge Regression
Here we are analysing multiple models in one go using multiple values of lambda. So we need not creating multiple models using different values of lambda
In [58]:
# using Ridge regression
params = {'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0,
4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000]}
# Taking instance of Ridge regressor
ridge = Ridge()
# Calling function for model fit
Ridge_model,ridge_alpha = reg_model(params,X_train_rfe,y_train,ridge)
# getting the instance on best alpha value
Ridge_fin_model = Ridge(alpha=ridge_alpha)
# fitting the model on training data set
Ridge_fin_model.fit(X_train_rfe,y_train)
# getting predicted values for train and test dataset
y_train_pred = Ridge_fin_model.predict(X_train_rfe)
y_test_pred = Ridge_fin_model.predict(X_test_rfe)
# calling the model evaluation function
model_eval()
cv_result = pd.DataFrame(Ridge_model.cv_results_)
cv_result
Search
ENG
IN
14:51
30-09-2024



Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 9 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help | Trusted | Python 3 (ipykernel)

In [66]:
```
# 10 top significant LAsso Coefficients
lasso_coeffs = np.exp(Coefficients['Lasso(Alpha = .0001)'])
lasso_coeffs.sort_values(ascending=False)[:10]
```

Out[66]:
```
MSZoning_FV            1.417651
MSZoning_RH            1.411549
MSZoning_RL            1.408965
MSZoning_RM            1.357499
OverallQual_9          1.241290
OverallQual_8          1.178796
GrLivArea              1.135808
OverallQual_7          1.122567
Exterior1st_BrkFace    1.113251
GarageYrBlt_2008.0     1.103963
Name: Lasso(Alpha = .0001), dtype: float64
```

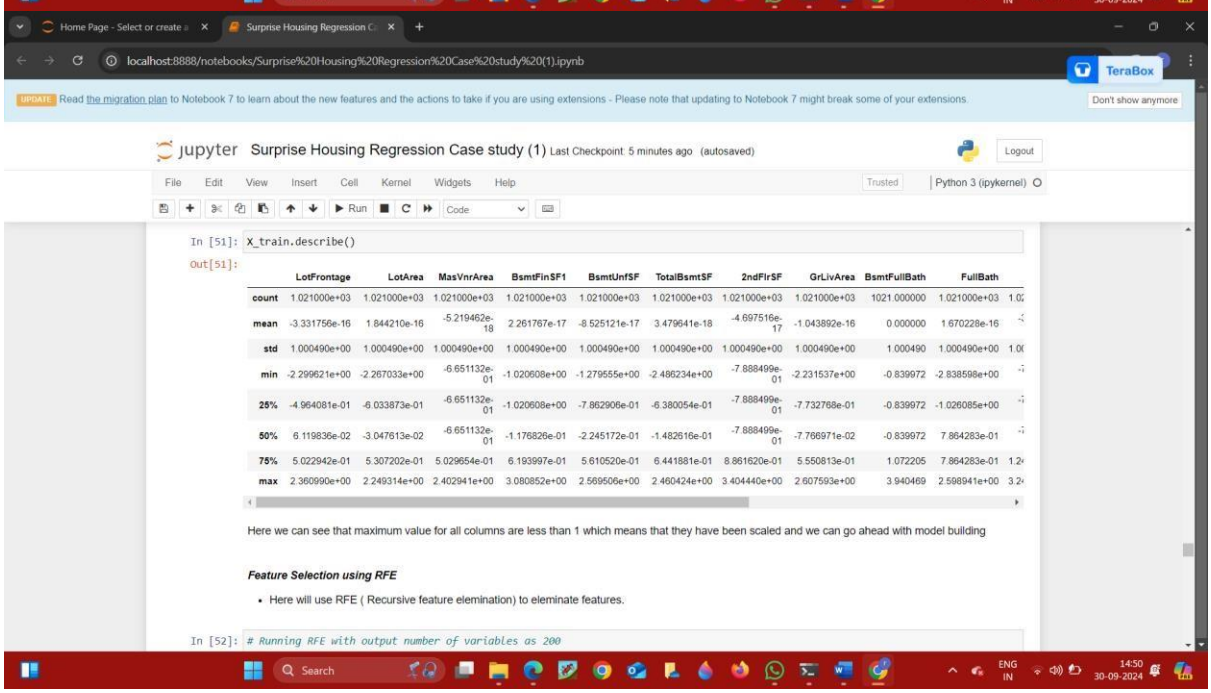
**Inferences**

**Optimal value of lambda for Ridge Regression = 2.0**
**Optimal value of lambda for Lasso Regression = 0.0001**

**Variables significant in predicting the price of a house are:**
OverallQual_9, OverallQual_8 , MSZoning_FV, MSZoning_FV, MSZoning_RH, MSZoning_RM, GrLivArea, Exterior1st_BrkFace, GarageYrBlt_2008,OverallQual_7,Functional_Typ

**How well those variables describe the price of a house**

- OverallQual_9 & OverallQual_8: if the overall material and finish of the house is Very Good or Excellent, the price of house will increase by 1.24 to 1.27 times
- GrLivArea:an increase of 1 square foot of house area above ground, the price will increase by 1.13 times
- MSZoning: If general zoning classification of house is Floating Village Residential and Residential High Density , Residential Low Density and Residential



Home Page - Select or create a | Surprise Housing Regression C | localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb

UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.

jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 5 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help | Trusted | Python 3 (ipykernel)

In [51]:
```
X_train.describe()
```

Out[51]:

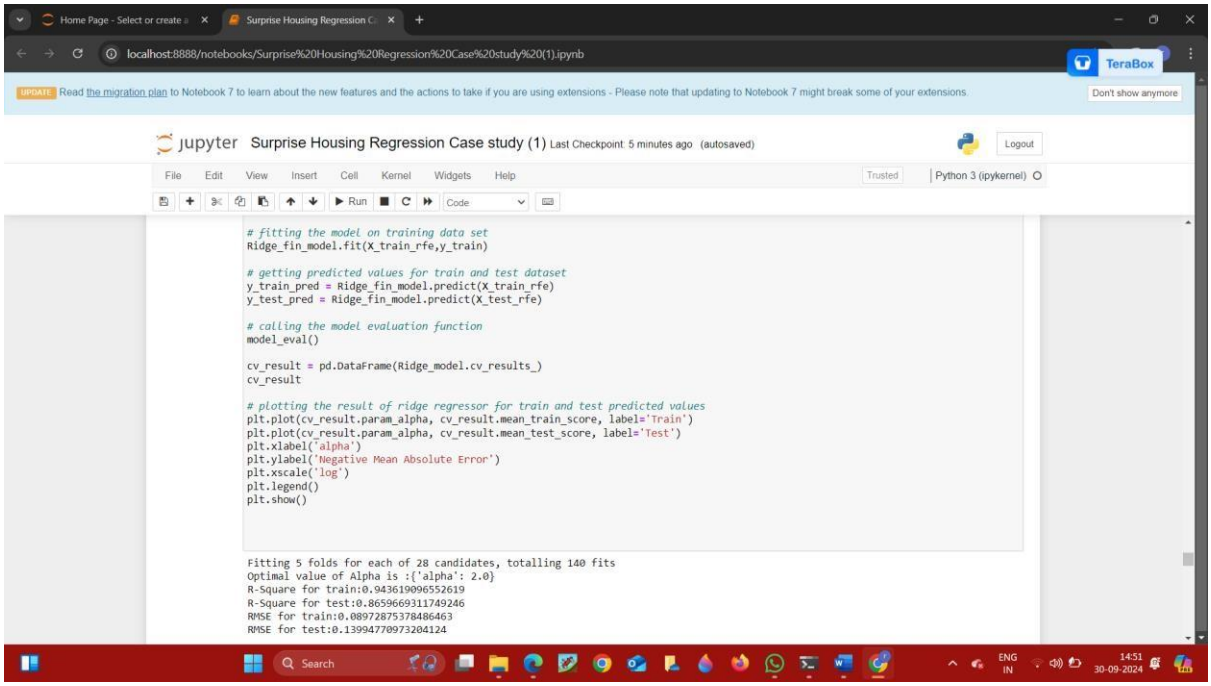
| | LotFrontage | LotArea | MasVnrArea | BsmtFinSF1 | BsmtUnfSF | TotalBsmtSF | 2ndFlrSF | GrLivArea | BsmtFullBath | FullBath | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 1.021000e+03 | 1.021000e+03 | 1.021000e+03 | 1.021000e+03 | 1.021000e+03 | 1.021000e+03 | 1.021000e+03 | 1.021000e+03 | 1021.000000 | 1.021000e+03 | 1.02 |
| mean | -3.331756e-16 | 1.844210e-16 | -5.219462e-18 | 2.261767e-17 | -8.525121e-17 | 3.479641e-18 | -4.697516e-17 | -1.043892e-16 | 0.000000 | 1.670228e-16 | -2 |
| std | 1.000490e+00 | 1.000490e+00 | 1.000490e+00 | 1.000490e+00 | 1.000490e+00 | 1.000490e+00 | 1.000490e+00 | 1.000490e+00 | 1.000490 | 1.000490e+00 | 1.00 |
| min | -2.299621e+00 | -2.267033e+00 | -6.651132e-01 | -1.020608e+00 | -1.279555e+00 | -2.486234e+00 | -7.888499e-01 | -2.231537e+00 | -0.839972 | -2.838598e+00 | -2 |
| 25% | -4.964081e-01 | -6.033873e-01 | -6.651132e-01 | -1.020608e+00 | -7.862906e-01 | -6.380054e-01 | -7.888499e-01 | -7.732768e-01 | -0.839972 | -1.026085e+00 | -2 |
| 50% | 6.119836e-02 | -3.047613e-02 | -6.651132e-01 | -1.176826e-01 | -2.245172e-01 | -1.482616e-01 | -7.888499e-01 | -7.766971e-02 | -0.839972 | 7.864283e-01 | -2 |
| 75% | 5.022942e-01 | 5.307202e-01 | 5.029654e-01 | 6.193997e-01 | 5.610520e-01 | 6.441881e-01 | 8.861620e-01 | 5.550813e-01 | 1.072205 | 7.864283e-01 | 1.2 |
| max | 2.360990e+00 | 2.249314e+00 | 2.402941e+00 | 3.080852e+00 | 2.569506e+00 | 2.460424e+00 | 3.404440e+00 | 2.607593e+00 | 3.940469 | 2.598941e+00 | 3.2 |

Here we can see that maximum value for all columns are less than 1 which means that they have been scaled and we can go ahead with model building

***Feature Selection using RFE***

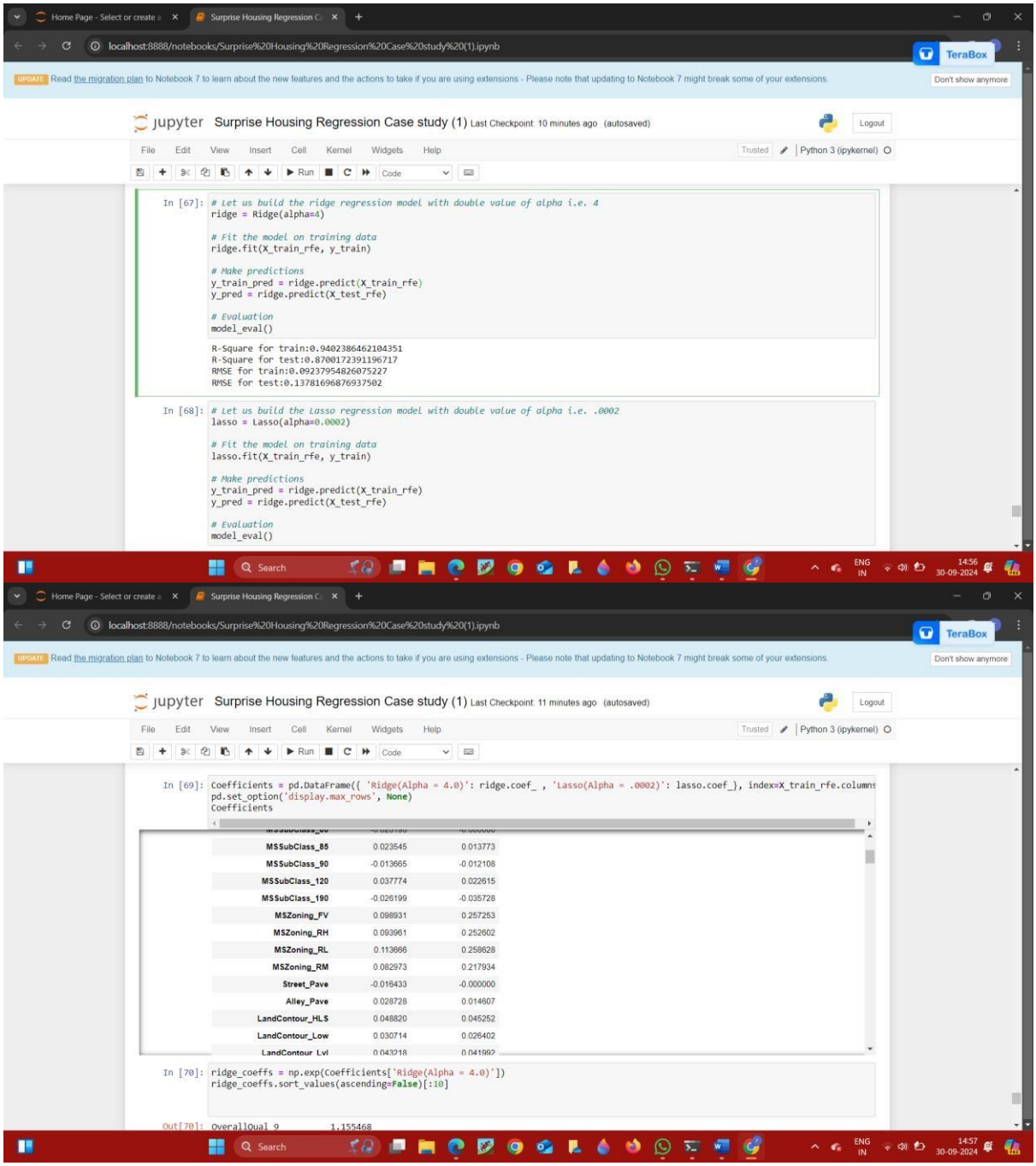
- Here will use RFE ( Recursive feature elemination) to eleminate features.

In [52]:
```
# Running RFE with output number of variables as 200
```

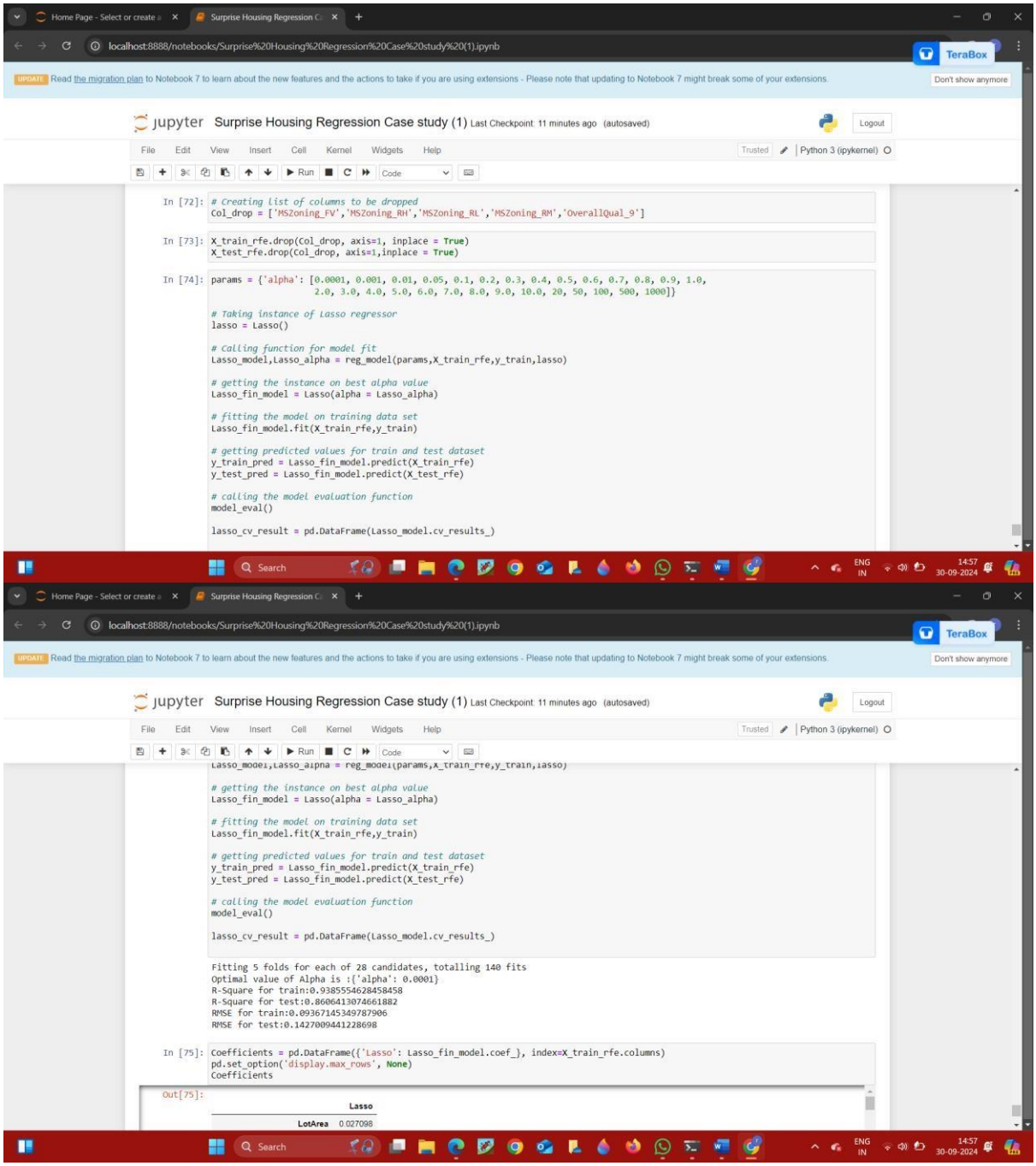



**Step 7 : Summary**

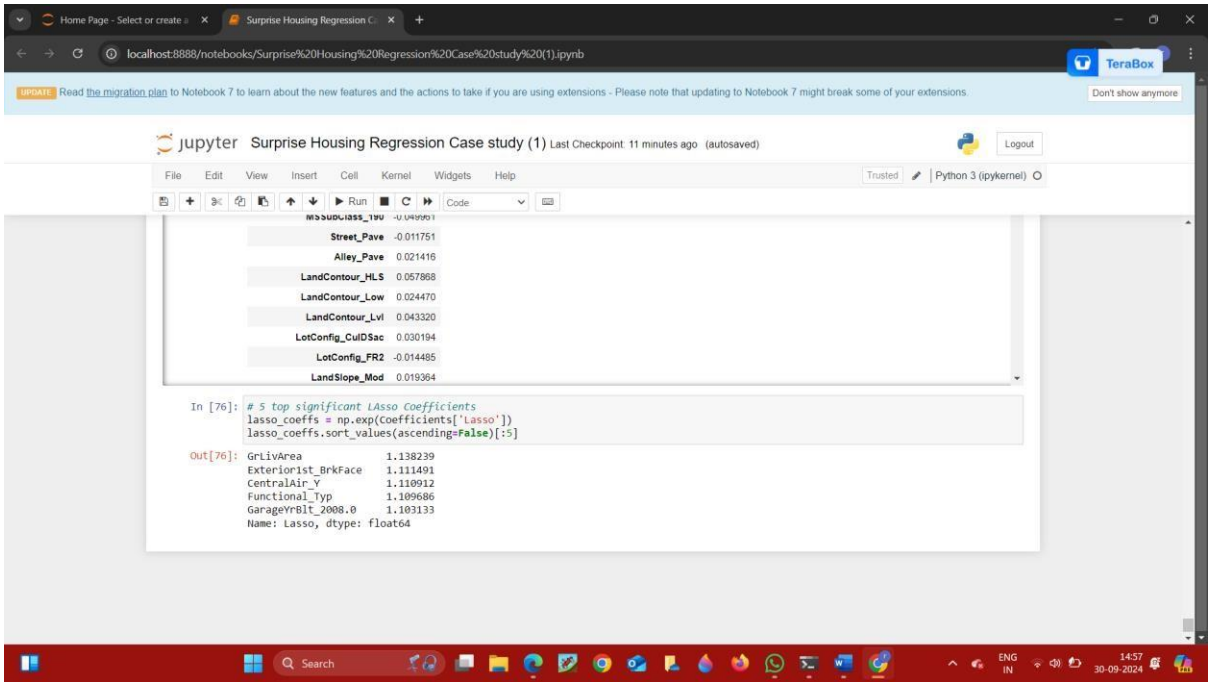
- Housing data is read and analyzed
- Data set is divided into numerical and categorical types.
- SalePrice is the target column here.
- Numerical features are analysed, visualized and cleaned for missing data, outliers
- Categorical features are also analysed, visualized and cleaned for NA values and
- Multiple features are also dropped
- New feature 'Age' is also derived
- Dummy variables are created
- Imbalanced dummy variables are dropped
- Data is divided into train and test dataset
- feature selection done based on RFe using 200 features
- Ridge and Lasso model was build and analysed


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 10 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [67]:
# Let us build the ridge regression model with double value of alpha i.e. 4
ridge = Ridge(alpha=4)
# Fit the model on training data
ridge.fit(X_train_rfe, y_train)
# Make predictions
y_train_pred = ridge.predict(X_train_rfe)
y_pred = ridge.predict(X_test_rfe)
# Evaluation
model_eval()
R-Square for train:0.9402386462104351
R-Square for test:0.8700172391196717
RMSE for train:0.09237954826075227
RMSE for test:0.13781696876937502
In [68]:
# Let us build the Lasso regression model with double value of alpha i.e. .0002
lasso = Lasso(alpha=0.0002)
# Fit the model on training data
lasso.fit(X_train_rfe, y_train)
# Make predictions
y_train_pred = ridge.predict(X_train_rfe)
y_pred = ridge.predict(X_test_rfe)
# Evaluation
model_eval()
Search
ENG
IN
14:56
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 11 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [69]:
Coefficients = pd.DataFrame({ 'Ridge(Alpha = 4.0)': ridge.coef_ , 'Lasso(Alpha = .0002)': lasso.coef_}, index=X_train_rfe.columns
pd.set_option('display.max_rows', None)
Coefficients
MSSubClass_85 0.023545 0.013773
MSSubClass_90 -0.013665 -0.012108
MSSubClass_120 0.037774 0.022615
MSSubClass_190 -0.026199 -0.035728
MSZoning_FV 0.098931 0.257253
MSZoning_RH 0.093961 0.252602
MSZoning_RL 0.113666 0.258628
MSZoning_RM 0.082973 0.217934
Street_Pave -0.016433 -0.000000
Alley_Pave 0.028728 0.014607
LandContour_HLS 0.048820 0.045252
LandContour_Low 0.030714 0.026402
In [70]:
ridge_coeffs = np.exp(Coefficients['Ridge(Alpha = 4.0)'])
ridge_coeffs.sort_values(ascending=False)[:10]
Out[70]: OverallQual 9 1.155468
Search
ENG
IN
14:57
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 11 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
In [72]:
# Creating list of columns to be dropped
Col_drop = ['MSZoning_FV','MSZoning_RH','MSZoning_RL','MSZoning_RM','OverallQual_9']
In [73]:
X_train_rfe.drop(Col_drop, axis=1, inplace = True)
X_test_rfe.drop(Col_drop, axis=1,inplace = True)
In [74]:
params = {'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0,
2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000]}
# Taking instance of Lasso regressor
lasso = Lasso()
# Calling function for model fit
Lasso_model,Lasso_alpha = reg_model(params,X_train_rfe,y_train,lasso)
# getting the instance on best alpha value
Lasso_fin_model = Lasso(alpha = Lasso_alpha)
# fitting the model on training data set
Lasso_fin_model.fit(X_train_rfe,y_train)
# getting predicted values for train and test dataset
y_train_pred = Lasso_fin_model.predict(X_train_rfe)
y_test_pred = Lasso_fin_model.predict(X_test_rfe)
# calling the model evaluation function
model_eval()
lasso_cv_result = pd.DataFrame(Lasso_model.cv_results_)
Search
ENG
IN
14:57
30-09-2024


Home Page - Select or create a
Surprise Housing Regression Ca
localhost:8888/notebooks/Surprise%20Housing%20Regression%20Case%20study%20(1).ipynb
TeraBox
UPDATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.
Don't show anymore
jupyter Surprise Housing Regression Case study (1) Last Checkpoint: 11 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Lasso_model,Lasso_alpha = reg_model(params,X_train_rfe,y_train,lasso)
# getting the instance on best alpha value
Lasso_fin_model = Lasso(alpha = Lasso_alpha)
# fitting the model on training data set
Lasso_fin_model.fit(X_train_rfe,y_train)
# getting predicted values for train and test dataset
y_train_pred = Lasso_fin_model.predict(X_train_rfe)
y_test_pred = Lasso_fin_model.predict(X_test_rfe)
# calling the model evaluation function
model_eval()
lasso_cv_result = pd.DataFrame(Lasso_model.cv_results_)
Fitting 5 folds for each of 28 candidates, totalling 140 fits
Optimal value of Alpha is :{'alpha': 0.0001}
R-Square for train:0.9385554628458458
R-Square for test:0.8606413074661882
RMSE for train:0.09367145349787906
RMSE for test:0.14270094412286 98
In [75]:
Coefficients = pd.DataFrame({'Lasso': Lasso_fin_model.coef_}, index=X_train_rfe.columns)
pd.set_option('display.max_rows', None)
Coefficients
Out[75]:
Lasso
LotArea 0.027098
Search
ENG
IN
14:57
30-09-2024




## 6. CONCLUSION

In this case study, machine learning models were successfully applied to predict housing prices based on a variety of features. The most accurate model was identified based on RMSE and R-squared values, and it demonstrated that features like the number of rooms, house size, and location have a significant impact on pricing. Future work could involve expanding the dataset, using more complex algorithms, or including additional market variables such as economic indicators or proximity to amenities.