HOUSING PRICE PREDICTION

Submitted by:

AVINASH PATEL

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my teacher as well as my intern company FLIPROBO who gave me this golden opportunity to do this wonderful project on the topic PRICE OF HOUSING, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them. Secondly I want to mention some sites which helped me when I got stuck somewhere while completing my projects.

Those sites are :

http://scikit-learn.org

https://www.w3schools.com

https://www.youtube.com

https://www.kaggle.com

INTRODUCTION

- Business Problem Framing

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

- Conceptual Background of the Domain Problem

I think we have to get some knowledge on how price is affected by all of the columns provided so we have to ask someone with better understanding in this field.

- Review of Literature

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling,

recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

- Motivation for the Problem Undertaken

  As I am an intern I want to work on as many project as possible for me. So ,I am highly motivated to do this project and to learn new thing and also learn from my mistakes.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

  First of all I checked that how many of the column are categorical columns and how many of them are numerical columns. By doing so I find out that there are thirty five columns with categorical data i.e. 'MSZoning', 'Street', 'LotShape', 'LandContour','LotConfig','LandSlope', 'Neighborhood', 'Condition1', 'BldgType', 'HouseStyle', 'RoofStyle', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'ExterQual', 'ExterCond', 'Foundation', 'BsmtQual', 'BsmtCond', 'BsmtExposure','BsmtFinType1', 'BsmtFinType2', 'HeatingQC', 'CentralAir', 'Electrical','KitchenQual', 'Functional', 'FireplaceQu', 'GarageType','GarageFinish', 'GarageQual', 'GarageCond', 'PavedDrive', 'SaleType', 'SaleCondition'. So I have to convert these data into numerical column.

  Now I started to visualise the column individually.

  Countplot:

  I created all the countplots for all the categorical columns and then I realised that many column have highly uneven data which could affect the final model as we do not have much data of the other column to feed the model.

  Catplot:

  Now I make all the catplots for all the columns and and understand that which acolumns have high amount of data and which are more significant than others.

  Heatmap:
  I also created a heatmap to see he correlation between all the columns and the other columns also that we can determine which columns are related to our target variable and which are not related.

- Data Sources and their formats

  The data is provided by our client who is in US-based housing company named Surprise Housing. The data has 1168 rows and 81 columns. The data is provided in csv(comma separated value) format. There are 35 categorical columns and 46 numerical columns.

- Data Pre-processing

For Numerical Data:

For MSSubClass: As per our data description this column coluld have only 20,30,40,45,50,60,70,75,80,90,120,150,160,180,190.So any thing other than these values is a outlies or bad datapoint but here there are no such datapoints that means all the values are good.

For LotFrontage: Many of the rows have null value so we are going to give those rows the mean of all the rows.

Treating OverallQual column: As per our data description this column coluld have only 1,2,3,4,5,6,7,8,9,10 . So any thing other than these values is a outlies or bad datapoint but here there are no such datapoints that means all the values are good.

Treating LotFrontagecolumn : Many of the rows have null value so we are going to give those rows the mean of all the rows.

Treating MasVnrArea column: Checking how many rows are 0, Hence I am checking how many 0 there are and 75% of data are zeros. So, I am going to delete this column.

Treating BsmtFinSF2 columns: Checking how many columns are 0 Hence I am checking how many 0 there are and 95% of data are zeros. So, I am going to delete this column.

Treating 2ndFlrSF column: Checking how many rows are 0, Hence I am checking how many 0 there are and 75% of data are zeros. So, I am going to delete this column.

Treating LowQualFinSF column: Checking how many rows are 0, Hence I am checking how many 0 there are and 90% of data are zeros. So, I am going to delete this column.

Treating BsmtFullBath,BsmtHalfBath,FullBath,HalfBath column: Among all these three columns BsmtFullBath,BsmtHalfBath and the HalfBath are the only columns does have zeros in there rows which are more than 70%. So we are going the keep the FullBAth column only and delete the rest of the columns as we already do not have much data to begin with and on top of that all the rows are already empty. This could effect our model.

Treating WoodDeckSF, OpenPorchSFcolumn: These columns also do not have much data as most of the data is empty to begin with.

Treating nclosedPorch,3SsnPorch,ScreenPorch,PoolArea,MiscVal: More than 75% of the data is null so we have to delete these columns as well.

For Categorical Data:

Treating MSZoning: In this column one of the values is C(All) so we are going to replace all these values to C for better understanding.

Treating Utilities: This column has only one type of value which is AllPub ,hence only one value is not sufficient for building a model as no variety is there to give predictions upon.

Treating Condition2, RoofMatl, Heating: All these columns have lots of null value in them so we are going to delete all of these columns.

Now after encoding our data we are going to find the correlation between our target variable and the categorical variables.

After doing that we discovered that columns 'ExterQual', 'BsmtQual', 'KitchenQual', 'SalePrice' have more than 50% of correlation with our target variable and columns 'Neighborhood', 'RoofStyle','Foundation','BsmtExposure','HeatingQC','GarageType',' GarageFinish' have corelation between 20% to 50% and columns 'MSZoning', 'Street', 'LotShape', 'LandContour', 'LotConfig', 'LandSlope', 'Condition1', 'BldgType', 'HouseStyle', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'ExterCond', 'BsmtCond', 'BsmtFinType1', 'BsmtFinType2', 'CentralAir', 'Electrical', 'Functional', 'FireplaceQu', 'GarageQual', 'GarageCond', 'PavedDrive', 'SaleType', 'SaleCondition' have correlation lower than 20%.

- Data Inputs- Logic- Output Relationships

  The Remaining Data after all the pre-processing acts as an input data for the Model(in our case Ridge Regressor) the already trained model gives us an output of the estimated price for the given input.

- Hardware and Software Requirements and Tools Used

  Ram:16GB
  ROM:200MB
  Processor : Ryzen 7 5800h(8 cores 16 threads)
  Tool : Jupiter Notebook
  Language Used: Python

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  First I imported all the basic libraries of python used in the problem solving. Then I read carefully the data and the description given to me. After that I imported the provided csv file into the Jupiter notebook. Now first I checked for any nan values in the dataset after that I checked how many number of rows and columns I am dealing with and I also done some more analysis of the raw data to get a general idea of there max and mins and there inter quartile ranges etc.

  After analysing the data I started cleaning the data as explained above.

  After cleaning I did some visualisation of the data to get some insights from our cleaned data.

  I removed skewness and some excess outliers. On the completion of all these steps I moved on to the model selection part. Here I tested many models and choose one of the model to proceed.

  After I selected one of the models I did hyperparameter tuning and finally I saved my model.

- Testing of Identified Approaches (Algorithms)
- - LinearRegression
- - DecisionTreeRegressor
- - Ridge
- - Lasso
- - RandomForestRegressor
- - KNeighborsRegressor

After running all the models the lowest difference between the accuracy score and the cross validation is for Ridge Regressor.

```
LR=LinearRegression()
regress(LR,x,y)
```

```
R2: 76.71751252186519
CV Score: 58.36564589058313
Diffrence: 18.351866631282057
```

```
DTR = DecisionTreeRegressor()
regress(DTR,x,y)
```

```
R2: 100.0
CV Score: 54.76100194045216
Diffrence: 45.23899805954784
```

```
RDG=Ridge()
regress(RDG,x,y)
```

```
R2: 76.71720114073729
CV Score: 58.42669056832239
Diffrence: 18.290510572414895
```

```
RFR=RandomForestRegressor()
regress(RFR,x,y)
```

```
R2: 96.88242312247824
CV Score: 73.83329117207478
Diffrence: 23.049131950403464
```

```
KNR=KNeighborsRegressor()
regress(KNR,x,y)
```

```
R2: 70.20541086436118
CV Score: 52.74704774626057
Diffrence: 17.458363118100614
```

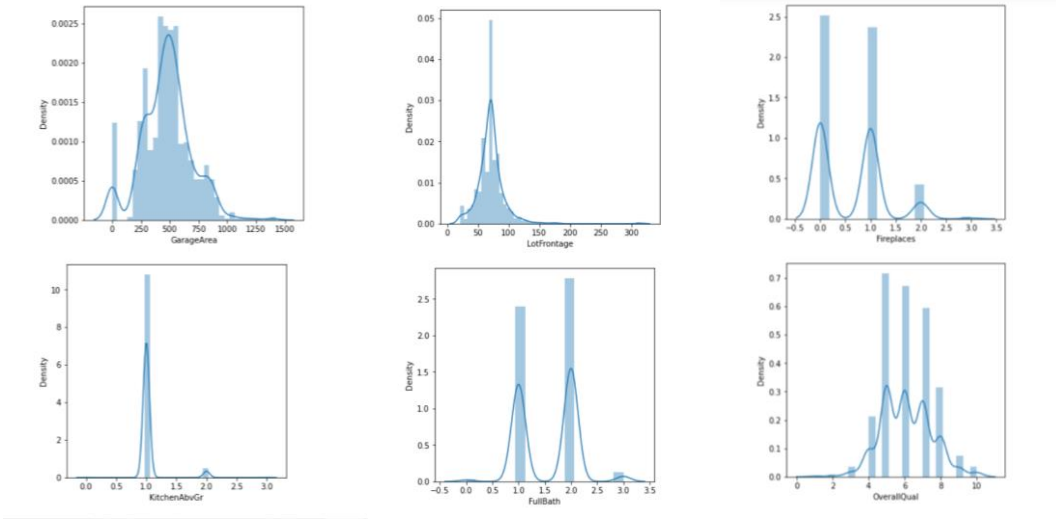
```
XGB=XGBRegressor()
regress(XGB,x,y)
```

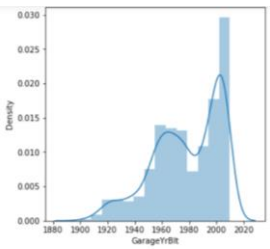
```
R2: 99.99919816778146
CV Score: 78.67989979336065
Diffrence: 21.31929837442081
```

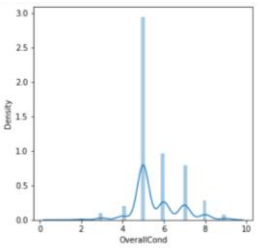
That means ridge regressor is not overfitted nor underfitted.

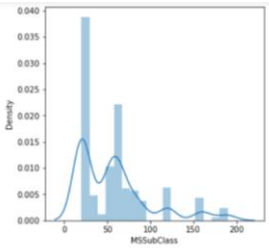
- Visualizations

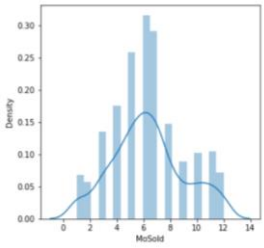
Distplot:

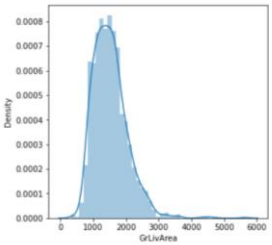


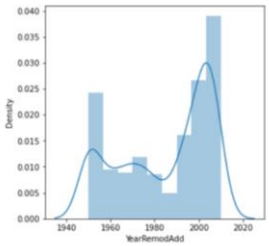

Density
GarageYrBlt

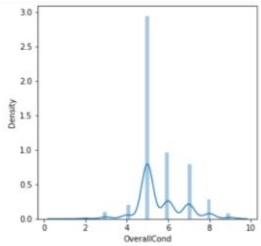

Density
OverallCond

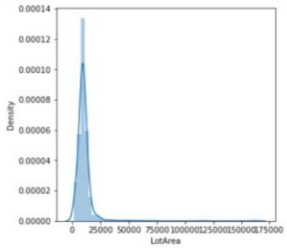

Density
MSSubClass

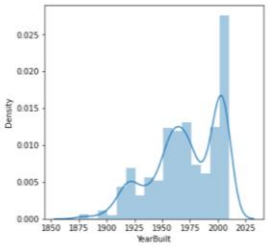

Density
MoSold

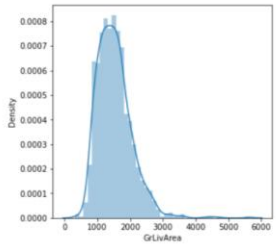

Density
GrLivArea

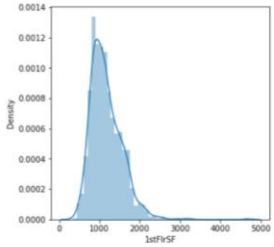

Density
YearRemodAdd

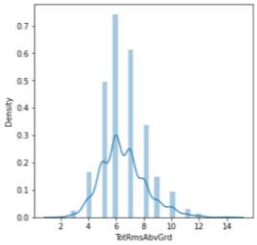

Density
OverallCond

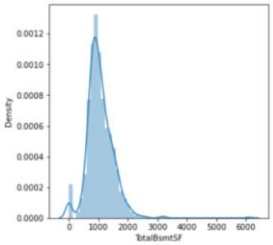

Density
LotArea

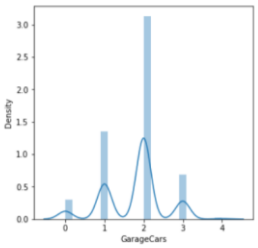

Density
YearBuilt

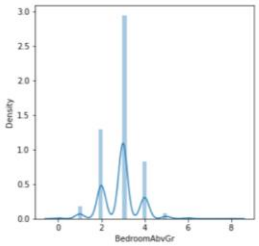

Density
GrLivArea

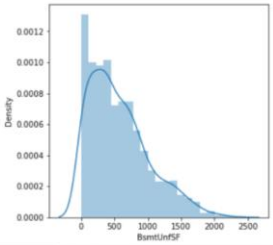

Density
1stFlrSF

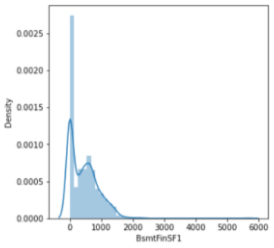

Density
TotRmsAbvGrd

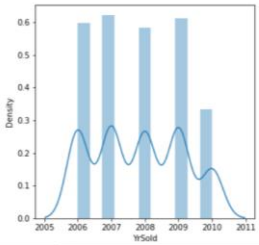

Density
TotalBsmtSF


Density
GarageCars

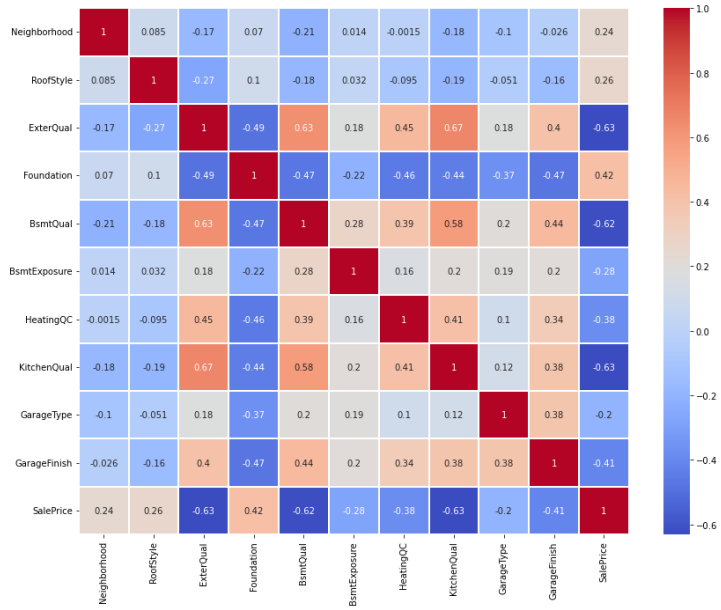

Density
BedroomAbvGr

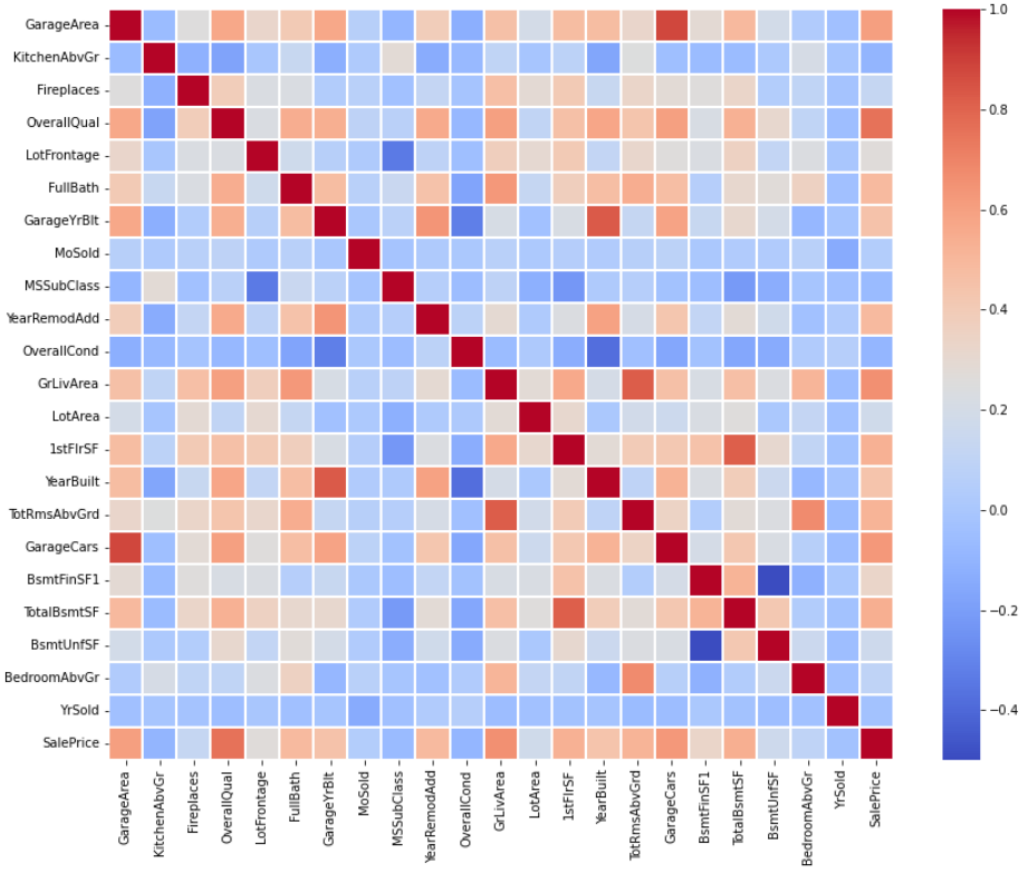

Density
BsmtUnfSF


Density
BsmtFinSF1

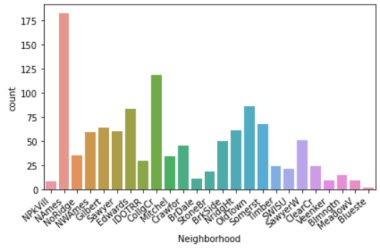

Density
YrSold

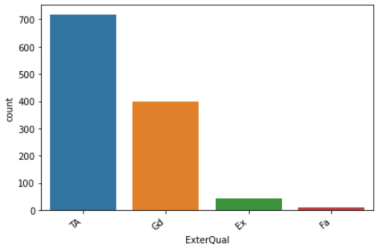
Heatmap:

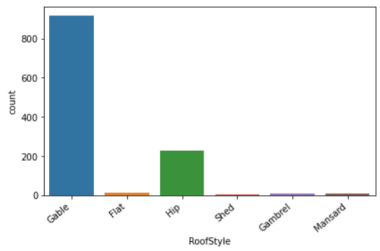

Neighborhood 1 0.085 -0.17 0.07 -0.21 0.014 -0.0015 -0.18 -0.1 -0.026 0.24
RoofStyle 0.085 1 -0.27 0.1 -0.18 0.032 -0.095 -0.19 -0.051 -0.16 0.26
ExterQual -0.17 -0.27 1 -0.49 0.63 0.18 0.45 0.67 0.18 0.4 -0.63
Foundation 0.07 0.1 -0.49 1 -0.47 -0.22 -0.46 -0.44 -0.37 -0.47 0.42
BsmtQual -0.21 -0.18 0.63 -0.47 1 0.28 0.39 0.58 0.2 0.44 -0.62
BsmtExposure 0.014 0.032 0.18 -0.22 0.28 1 0.16 0.2 0.19 0.2 -0.28
HeatingQC -0.0015 -0.095 0.45 -0.46 0.39 0.16 1 0.41 0.1 0.34 -0.38
KitchenQual -0.18 -0.19 0.67 -0.44 0.58 0.2 0.41 1 0.12 0.38 -0.63
GarageType -0.1 -0.051 0.18 -0.37 0.2 0.19 0.1 0.12 1 0.38 -0.2
GarageFinish -0.026 -0.16 0.4 -0.47 0.44 0.2 0.34 0.38 0.38 1 -0.41
SalePrice 0.24 0.26 -0.63 0.42 -0.62 -0.28 -0.38 -0.63 -0.2 -0.41 1
Neighborhood RoofStyle ExterQual Foundation BsmtQual BsmtExposure HeatingQC KitchenQual GarageType GarageFinish SalePrice
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6

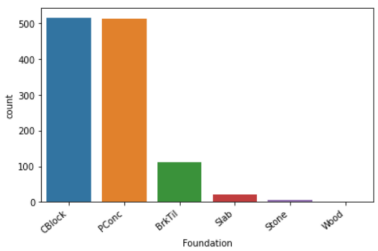

GarageArea
KitchenAbvGr
Fireplaces
OverallQual
LotFrontage
FullBath
GarageYrBlt
MoSold
MSSubClass
YearRemodAdd
OverallCond
GrLivArea
LotArea
1stFlrSF
YearBuilt
TotRmsAbvGrd
GarageCars
BsmtFinSF1
TotalBsmtSF
BsmtUnfSF
BedroomAbvGr
YrSold
SalePrice
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

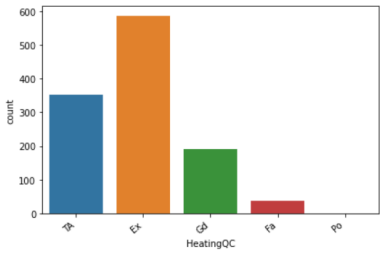
Countplot:

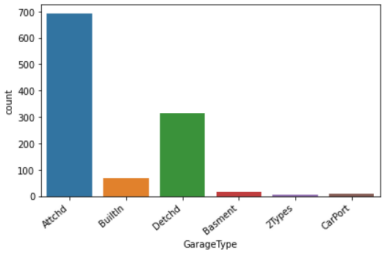

count
Neighborhood
NAmes
NoRidge
NWAmes
Gilbert
Sawyer
Edwards
IDOTRR
CollgCr
Mitchel
Crawfor
SWISU
BrkSide
Blueste
NPkVill
Veenker
Blmngtn
MeadowV
Timber
Somerst
OldTown
NridgHt
SawyerW
ClearCr

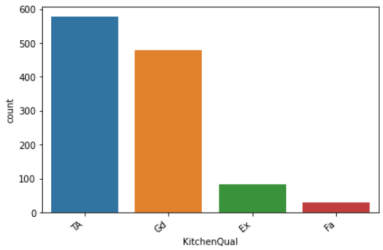

count
ExterQual
TA
Gd
Ex
Fa

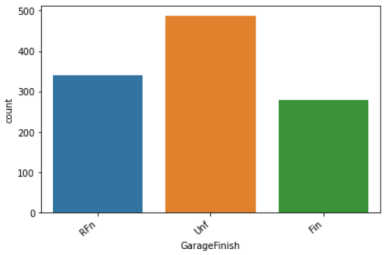

count
RoofStyle
Gable
Flat
Hip
Shed
Gambrel
Mansard

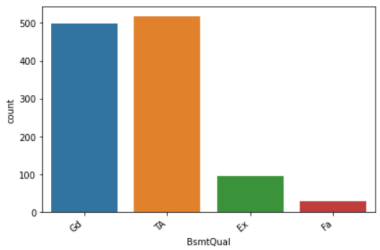

count
Foundation
CBlock
PConc
BrkTil
Slab
Stone
Wood

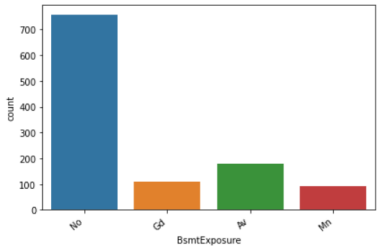

count
HeatingQC
TA
Ex
Gd
Fa
Po


count
GarageType
Attchd
Builtin
Detchd
Basment
2Types
CarPort


count
KitchenQual
TA
Gd
Ex
Fa


count
GarageFinish
RFn
Unf
Fin


count
BsmtQual
Gd
TA
Ex
Fa


count
BsmtExposure
No
Gd
Av
Mn

- Interpretation of the Results

  After all the pre processing and then visualising the data I find out that although there are many columns to the dataset ,but most of the columns are useless and the some of the remaining columns do have a significant correlation with our Target column.

CONCLUSION

- Key Findings and Conclusions of the Study

  I find out that few columns are correlated to our target variable that means predicting our target variable is difficult for any model because there is no direct correlation between them. Also, many of the columns have values which are unrealistic. Many of the columns have highly skewed values and many columns also have high amounts of outlies in them. Many columns have zeros as a value in them and we do not have much data to begin with.

- Learning Outcomes of the Study in respect of Data Science

  As I have already told that this dataset has a lot of outliers and I have to look into each and every column to get some of the data cleaned. As per the visualisation part I plotted boxplots to see the number of outliers and in which column they exist and in how much amount.

  I have also plotted Distplot to get a general idea of the columns which have skewness to further treat them.

  I plotted heatmap to see which columns are in a correlation with our target variable.

  I have plotted countplot to see that our target variable is balanced or not.

- Limitations of this work and Scope for Future Work

  As I am not able to fully understand the meaning of all the columns and if I could get someone who is already working then the Telecom company then I could get a better idea of how all the column are working and which outer columns needs to be in the dataset. As per the future scope, yes there is definitely more room to grow as I could get the model efficiency to 76.71% it could be more on further research.