
# Plot Price Prediction - A Comparative Study

**By :**
Adrija Karmakar

# INTRODUCTION

**Lands or plots have always been in high demand in India. It is an excellent money saving investment opportunity that will guarantee higher returns in the future. This is because, a piece of land remains in good condition and only increases in value.**

**Also, as many people would say, "*Owning a Piece of Land Gives Peace of Mind*". Owning land not only provides financial security but also gives a peace of mind. Experts also recommend raw land investing and buying land for future development, such as housing or building. Hence, we can always consider plots/ lands to be tangible long term investments that one can benefit from over time.**

# PROBLEM STATEMENT

Nowadays finding a good plot is as difficult as distinguishing colors in the darkness. Consumers are facing lots of technical faults and issues like,

1. High cost of the land as compared to the conventional price

2. Transportation problems

3. Unavailability of markets, malls, and other amenities required for the day to day living

4. Surroundings not being up to the mark, etc.

A *multinational real estate* company has been trying to offer justified prices of properties, by using innovative statistical algorithms, technology and sophisticated business ideas. There have been various studies to detect the determinants of land prices to this day. Here, we are trying to examine the impacts of structural, locational, and environmental attributes of the land/plots. So, we have observed *12 major features* that deeply affect the *average price per sq ft of buying the plots* from different localities scattered throughout West Bengal, India.

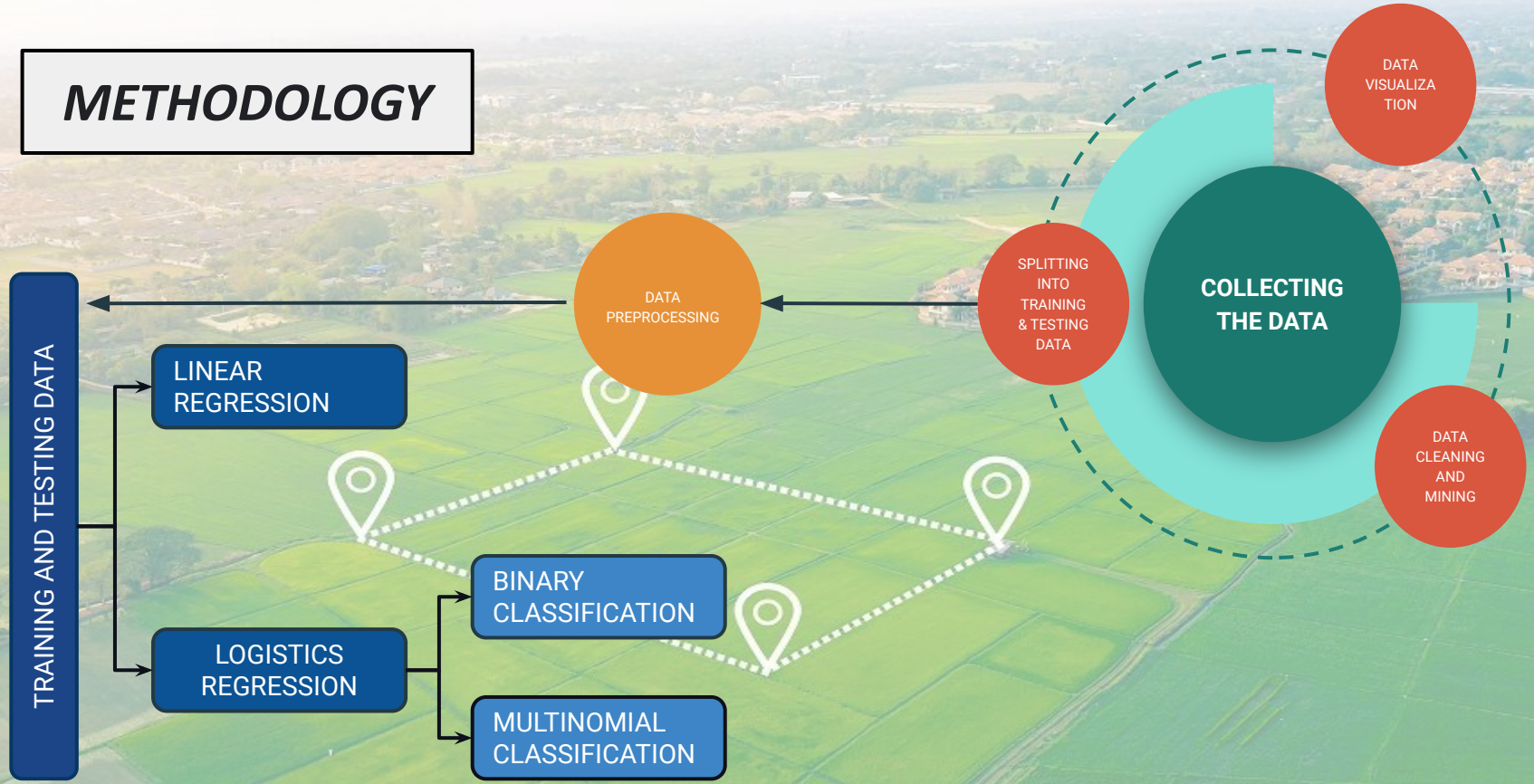
# AIM

Buying a home is a dream for most people, but there is always a dilemma when it comes to choosing the plot. There are various factors that play major roles in choosing the ideal land for investing. We have considered 12 such important factors among them.

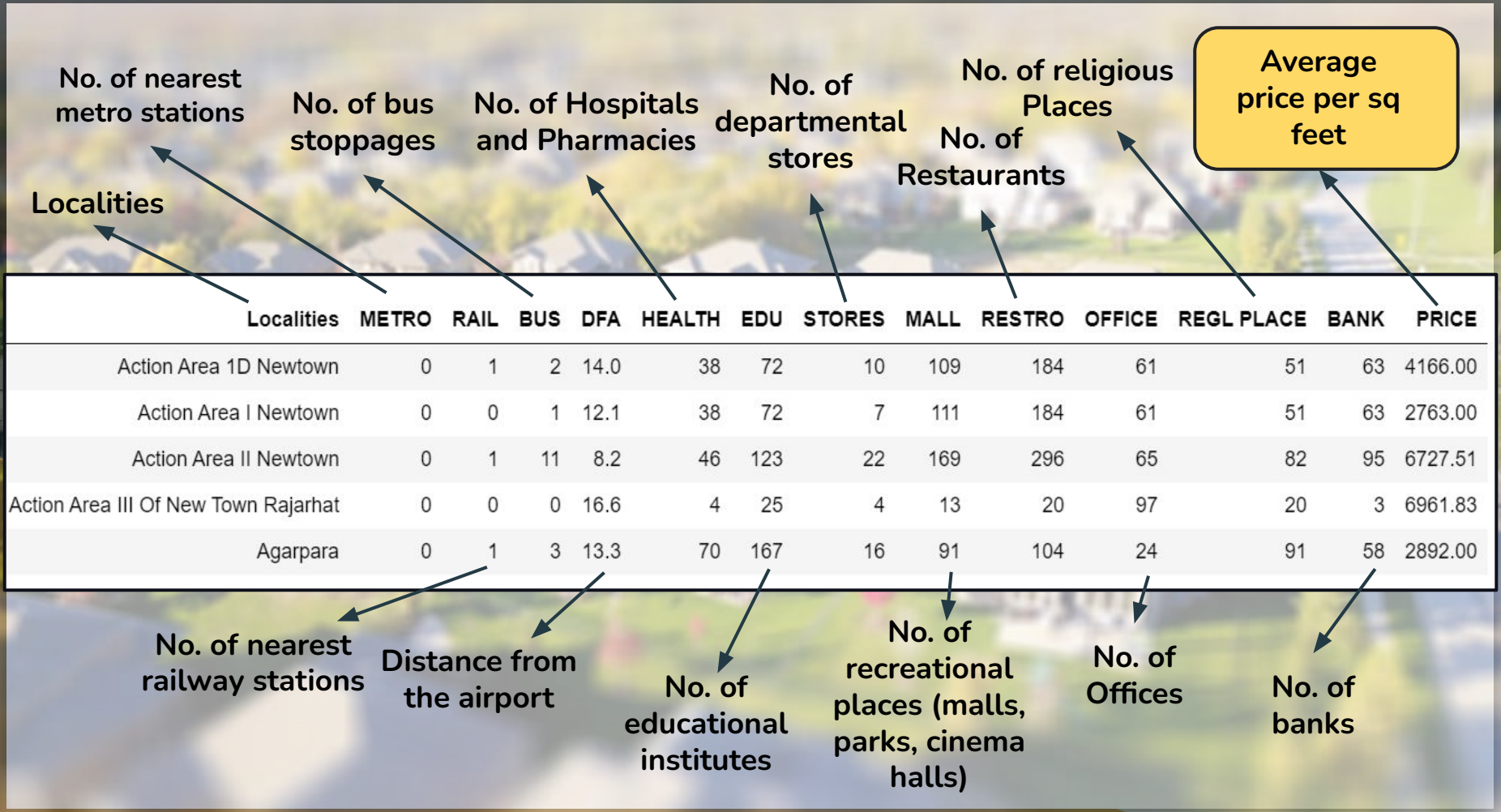
Our primary aim is to :

1. Predict the average price of buying plots in the localities based on the different factors.
2. Build a relationship between each of the different factors and the average price of buying plots in those localities to understand the effect of the factors on the price.

# METHODOLOGY



# DATASET




| Localities | METRO | RAIL | BUS | DFA | HEALTH | EDU | STORES | MALL | RESTRO | OFFICE | REGL PLACE | BANK | PRICE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Action Area 1D Newtown | 0 | 1 | 2 | 14.0 | 38 | 72 | 10 | 109 | 184 | 61 | 51 | 63 | 4166.00 |
| Action Area I Newtown | 0 | 0 | 1 | 12.1 | 38 | 72 | 7 | 111 | 184 | 61 | 51 | 63 | 2763.00 |
| Action Area II Newtown | 0 | 1 | 11 | 8.2 | 46 | 123 | 22 | 169 | 296 | 65 | 82 | 95 | 6727.51 |
| Action Area III Of New Town Rajarhat | 0 | 0 | 0 | 16.6 | 4 | 25 | 4 | 13 | 20 | 97 | 20 | 3 | 6961.83 |
| Agarpara | 0 | 1 | 3 | 13.3 | 70 | 167 | 16 | 91 | 104 | 24 | 91 | 58 | 2892.00 |

# DATA DESCRIPTION

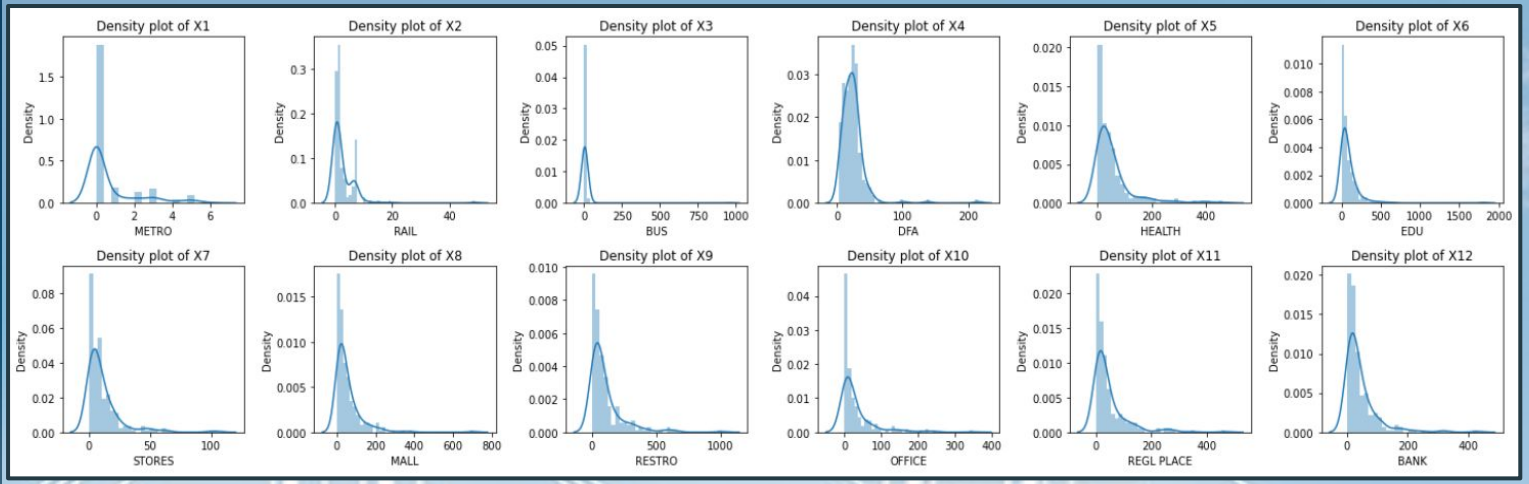
```
In [4]: data1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 225 entries, 0 to 224
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   METRO        225 non-null    int64
 1   RAIL         225 non-null    int64
 2   BUS          225 non-null    int64
 3   DFA          225 non-null    float64
 4   HEALTH       225 non-null    int64
 5   EDU          225 non-null    int64
 6   STORES       225 non-null    int64
 7   MALL         225 non-null    int64
 8   RESTRO       225 non-null    int64
 9   OFFICE       225 non-null    int64
 10  REGL PLACE   225 non-null    int64
 11  BANK         225 non-null    int64
 12  PRICE        225 non-null    float64
dtypes: float64(2), int64(11)
memory usage: 23.0 KB
```

Here the '***Distance from the airport***' and the '***Average price per sq feet of the plot***' contains float values. The rest are integer valued.

| | Count | Missing | No. of Unique | Dtype | Numeric | Mode | Mean | Min | 25% | Median | 75% | Max | Std | Skewness | Kurtosis | Coeff of Variation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **BUS** | 225 | 0 | 24 | int64 | True | 1 | 9.408889 | 0 | 1.00 | 3.00 | 7.00 | 958 | 63.773048 | 14.823377 | 221.426840 | 6.777957 |
| **METRO** | 225 | 0 | 7 | int64 | True | 0 | 0.653333 | 0 | 0.00 | 0.00 | 0.00 | 6 | 1.341241 | 2.121583 | 3.619205 | 2.052921 |
| **RAIL** | 225 | 0 | 13 | int64 | True | 1 | 2.533333 | 0 | 0.00 | 1.00 | 3.00 | 49 | 4.328065 | 6.147756 | 59.604069 | 1.708447 |
| **EDU** | 225 | 0 | 129 | int64 | True | 17 | 89.066667 | 0 | 22.00 | 46.00 | 108.00 | 1800 | 149.888649 | 7.363271 | 76.844555 | 1.682882 |
| **OFFICE** | 225 | 0 | 69 | int64 | True | 0 | 30.977778 | 0 | 5.00 | 12.00 | 33.00 | 349 | 48.888123 | 3.095375 | 11.933522 | 1.578168 |
| **REGL PLACE** | 225 | 0 | 92 | int64 | True | 10 | 46.337778 | 0 | 10.00 | 24.00 | 52.00 | 464 | 63.486582 | 3.045119 | 12.112754 | 1.370083 |
| **MALL** | 225 | 0 | 108 | int64 | True | 9 | 55.066667 | 0 | 13.00 | 30.00 | 65.00 | 703 | 74.299166 | 4.201541 | 27.793890 | 1.349258 |
| **STORES** | 225 | 0 | 43 | int64 | True | 1 | 11.017778 | 0 | 3.00 | 7.00 | 14.00 | 105 | 14.650989 | 3.289404 | 14.516808 | 1.329759 |
| **HEALTH** | 225 | 0 | 105 | int64 | True | 1 | 54.000000 | 0 | 11.00 | 32.00 | 63.00 | 464 | 71.737244 | 2.983851 | 10.744337 | 1.328467 |
| **BANK** | 225 | 0 | 92 | int64 | True | 5 | 43.337778 | 0 | 11.00 | 25.00 | 55.00 | 430 | 56.362995 | 3.318449 | 14.932840 | 1.300551 |
| **RESTRO** | 225 | 0 | 134 | int64 | True | 11 | 103.960000 | 0 | 25.00 | 61.00 | 119.00 | 1009 | 133.384625 | 3.006884 | 12.367078 | 1.283038 |
| **DPA** | 225 | 0 | 164 | float64 | True | 22.1 | 23.968444 | 3.0 | 13.30 | 22.60 | 28.50 | 215.0 | 19.307572 | 5.612031 | 48.022740 | 0.805541 |
| **PRICE** | 225 | 0 | 222 | float64 | True | 5555.56 | 5472.381867 | 1000.0 | 3247.86 | 4324.84 | 6154.04 | 29166.0 | 4229.597983 | 2.597275 | 8.434736 | 0.772899 |

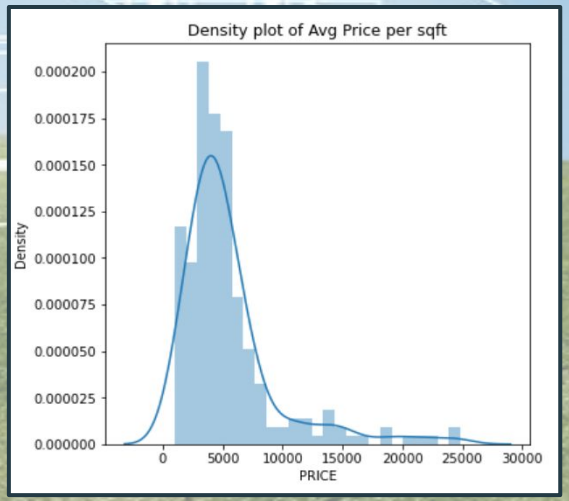
# EDA & DATA PRE-PROCESSING




***The Explanatory Variables - (X1,X2,...X12)***

***Avg price of plots per sq feet - (Y)***

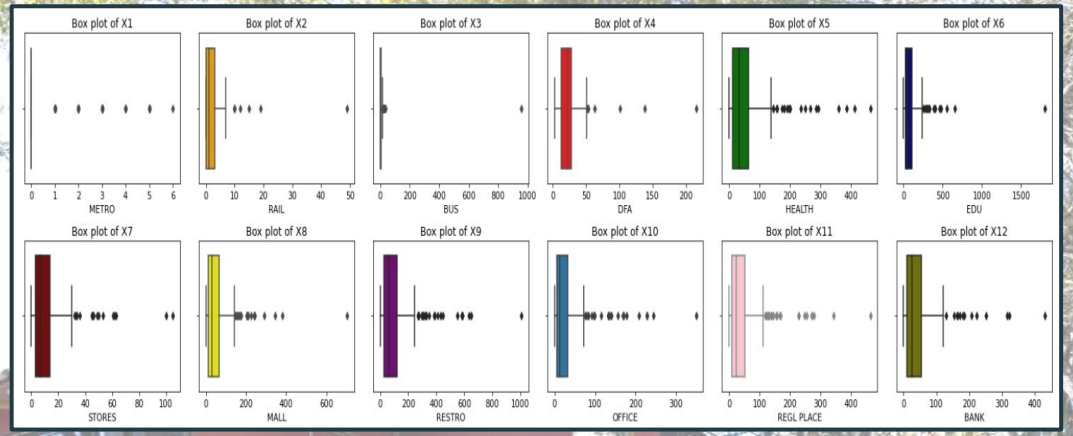
Usually, most of the plots have a lower value (within Rs. 10,000), while only a few of them are incredibly expensive (Rs. 10,000-25,000). Hence, statistically, it clearly forms a positively skewed distribution.




Each of our Explanatory variables show positive skewness. By positive skewness we can say that the mean of our distribution has a positive value and is present on the right side of the median and mode of the data.

# *DATA PREPROCESSING*

While reviewing a boxplot, an outlier is defined as a **datapoint that is located outside the whiskers of the boxplot**, i.e outside 1.5 times the Interquartile range above the upper quartile and below the lower quartile (Q1-1.5*IQR or Q3+1.5*IQR).



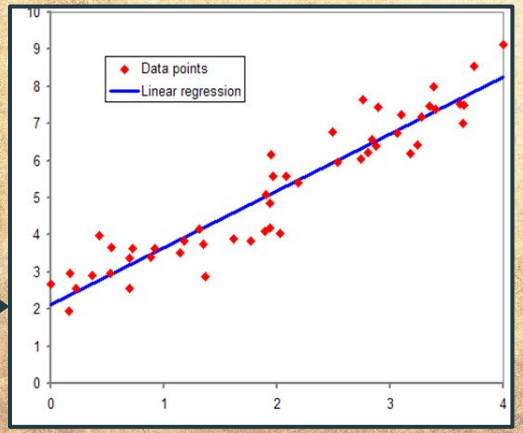

Therefore, from our boxplot we have identified the **outliers.** Further, we have applied the **TRAIN-TEST** technique to split our dataset.

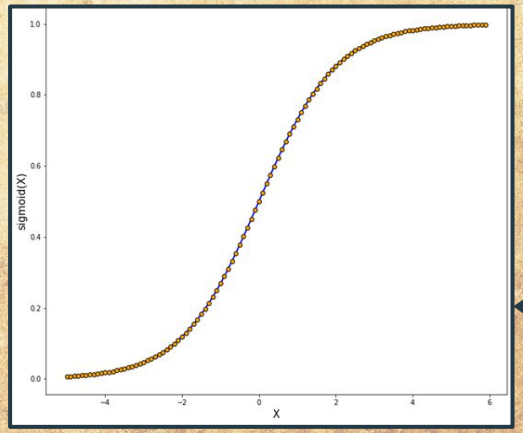
Since our data is mostly **positively skewed** and it doesn't satisfy the required assumptions, hence we have performed scaling as well as transformations on our data before fitting our model.

# WHAT IS REGRESSION ?

Regression is a statistical method used in house prices, finance, investing and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and series of other variables(known as independent variables), and also help in predicting the value of the dependent variable based on the independent variables.

***Linear Regression***




***Logistic Regression***

***Assumptions before performing regression:***

1. Normality
2. Linearity
3. Homoscedasticity
4. Independence of the Explanatory variables

***QQ plot***

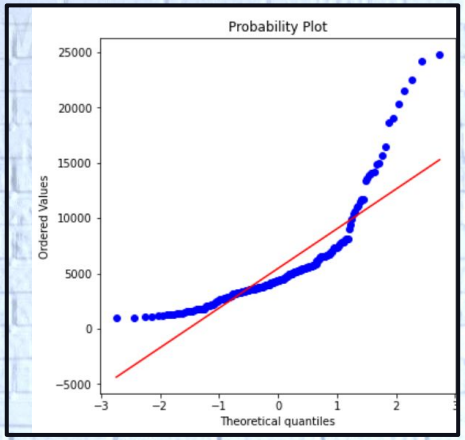
***1.Normality***

For our dataset, the normality assumption is violated.

We have come to this conclusion with the help of -

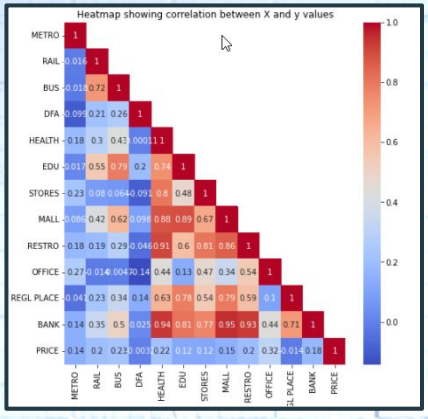
***Shapiro wilk test***

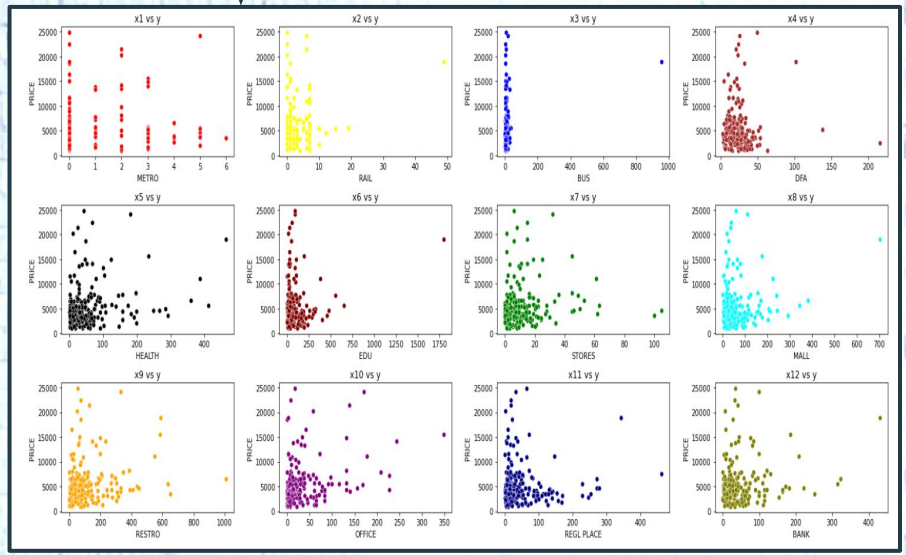
```
Test statistic: 0.7490413188934326
p value: 3.2666115124076486e-18
Our data does not follow normal distribution
```




***2.Linearity***

The ***correlation*** between the covariates and the response variable is checked and the data doesn't satisfy linearity assumption.



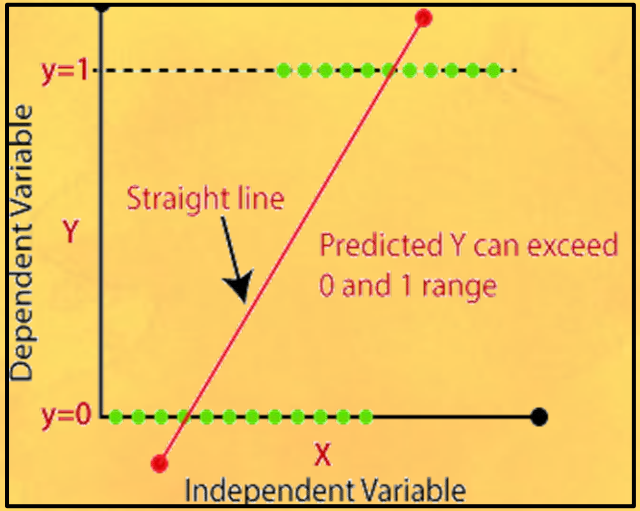



# LINEAR REGRESSION

**Objective - Predicting the average price of the localities on the basis of the 12 feature variables**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

***Where, y is the dependent variable and x1, x2 ... and xn are the explanatory variables.***

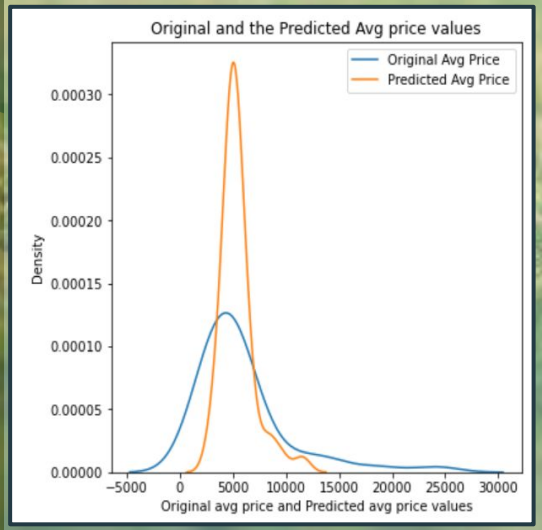


# LINEAR REGRESSION

## APPROACH-1

*Model with the original dataset including the 12 feature variables for the 'Avg Price' as the response variable*

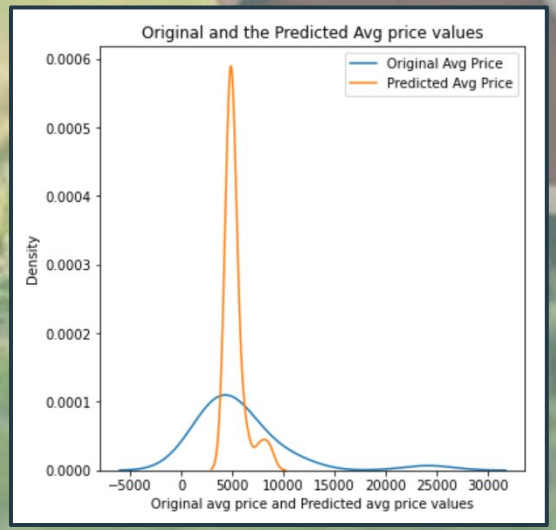
**R squared - 0.2547**




## APPROACH-2

*Model after removing the feature variables with high VIF and the 'Avg Price' as the response variable*

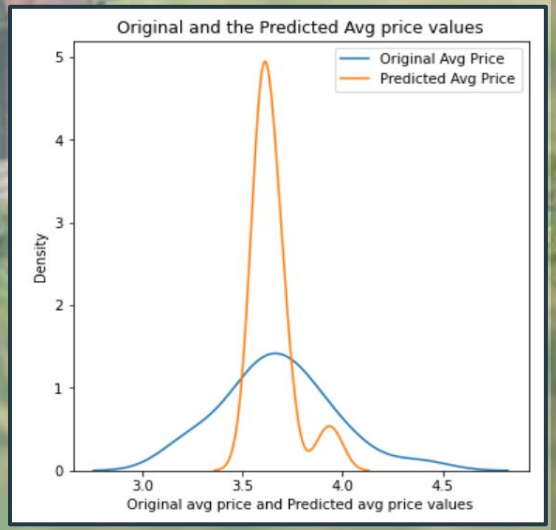
**R squared - 0.2733**



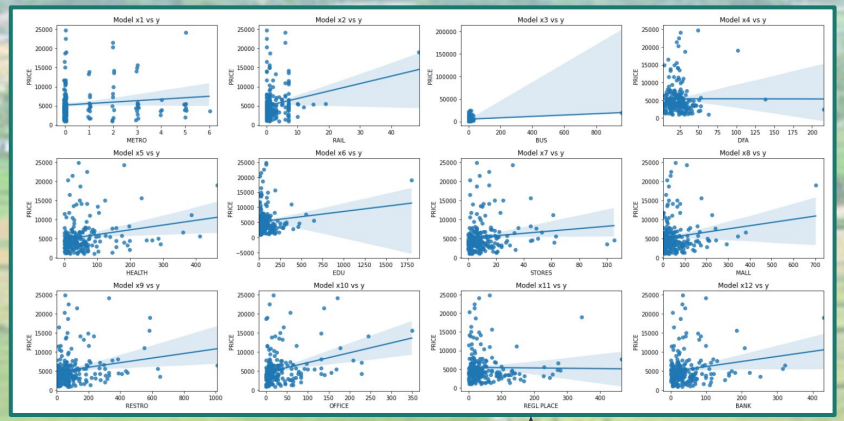

## APPROACH-3

*Model after scaling the feature variables with log transformed 'Avg Price' as the response variable*

**R squared - 0.4115**

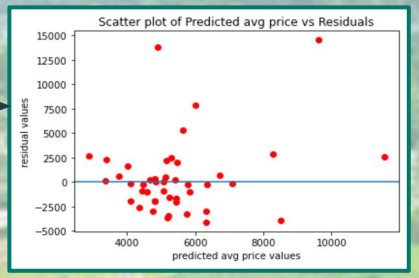


## APPROACH 1 - Model with the original dataset including the 12 feature variables for the 'Avg Price' as the response variable




**R squared - 0.2547**
**RMSE - 3815.65**

**Res vs Pred**




**Homoscedasticity**

**Breuschpagan test**

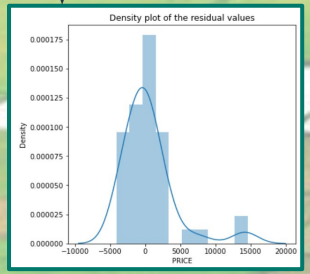
```
The p value of our test is 0.94
We accept the null hypothesis
Our errors are Homoscedastic
```

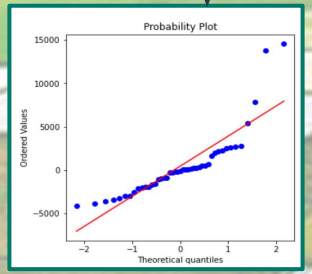
**Multicollinearity**

| | Feature Variables | VIF |
|---|---|---|
| 0 | METRO | 1.352500 |
| 1 | RAIL | 2.600902 |
| 2 | BUS | 10.336880 |
| 3 | DFA | 1.701081 |
| 4 | HEALTH | 17.941006 |
| 5 | EDU | 36.262525 |
| 6 | STORES | 7.058258 |
| 7 | MALL | 52.728018 |
| 8 | RESTRO | 38.736811 |
| 9 | OFFICE | 2.297559 |
| 10 | REGL PLACE | 12.128615 |
| 11 | BANK | 56.203100 |

**Regression model plot**

**Residuals**



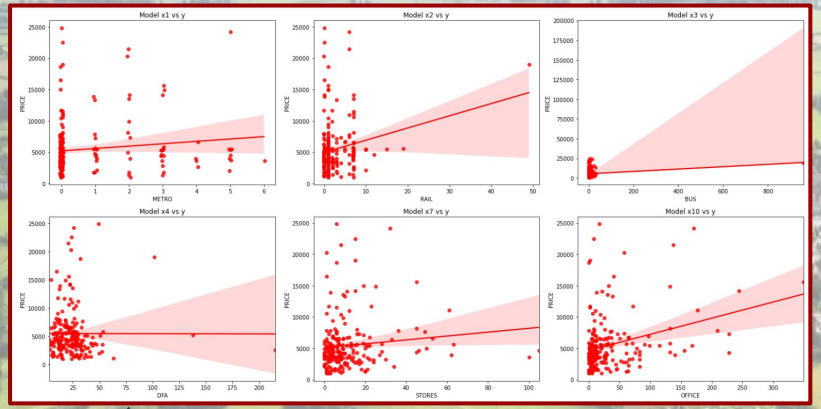




| 5-fold CV Scores | |
|---|---|
| 1 | -0.161936 |
| 2 | -0.600092 |
| 3 | 0.101201 |
| 4 | 0.065183 |
| 5 | 0.055003 |

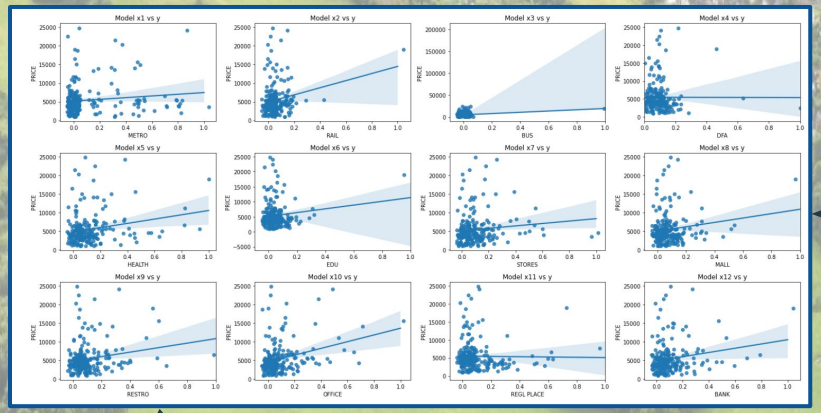
**5 fold cross validation scores**

**Average - -0.1801**
**Std. deviation - 0.2628**

## APPROACH 2 - *Model after removing the feature variables with high VIF and the 'Avg Price' as the response variable*



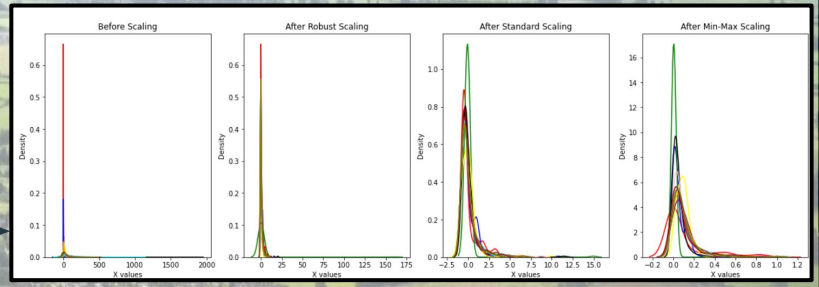

**R squared - 0.2733 RMSE - 3963.05**



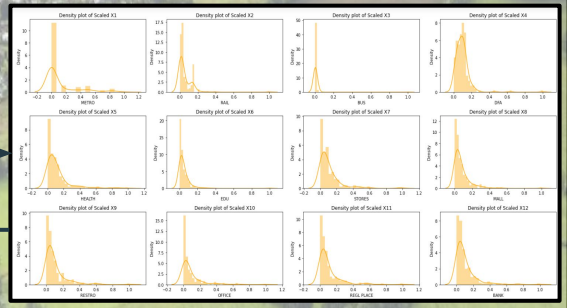

**R squared - 0.4115 RMSE - 0.2071**

***Different types of Scaling***

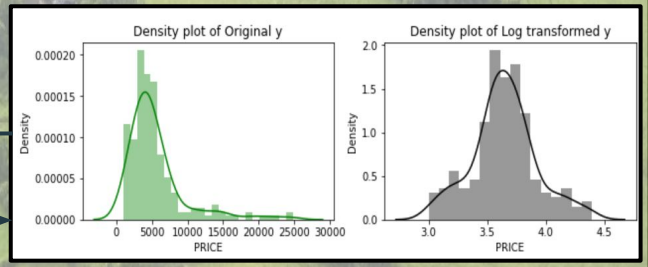
## APPROACH 3 - *Model after scaling the feature variables with log transformed 'Avg Price' as the response variable*




***Minmax scaled features variables***




***Log transformed response variable***



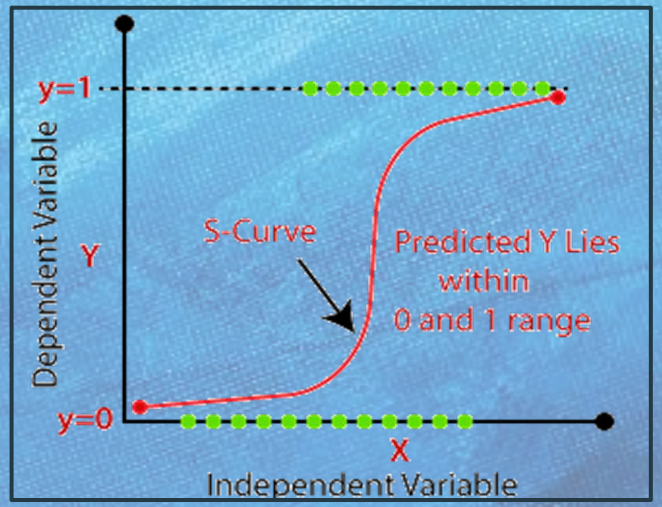
# LOGISTIC REGRESSION

Objective - Predicting the average price of the localities in the,
- 'High' and 'Low' price range
- 'High', 'Medium', 'Low' price range

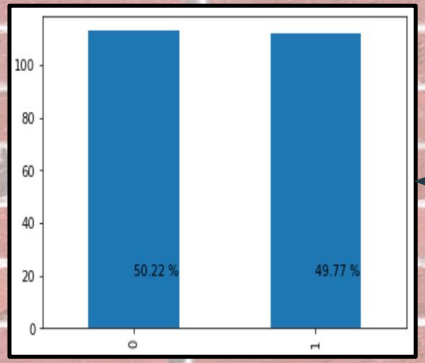
on the basis of the 12 feature variables

$$y = \log(p/(1-p))$$

*Where, y is the dependent variable and p is the probability of success.*

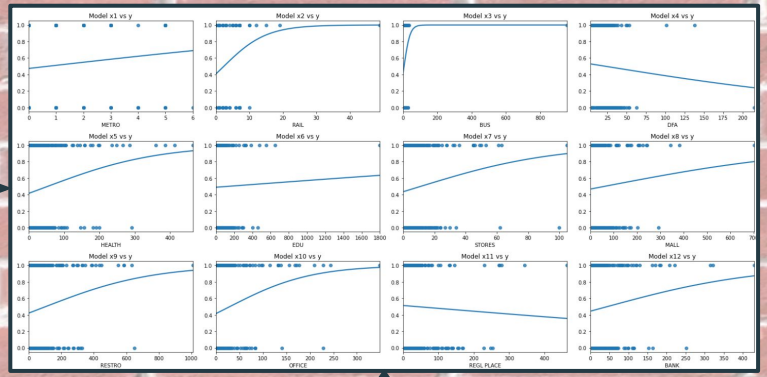


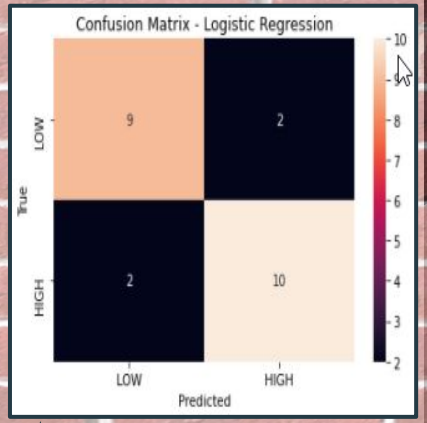
***BINARY LOGISTIC REGRESSION*** - *Binary logistic regression (LR) is a regression model where the response variable is binary, that is, it can take only two values, 0 or 1. The classification has been done based on its median value.*

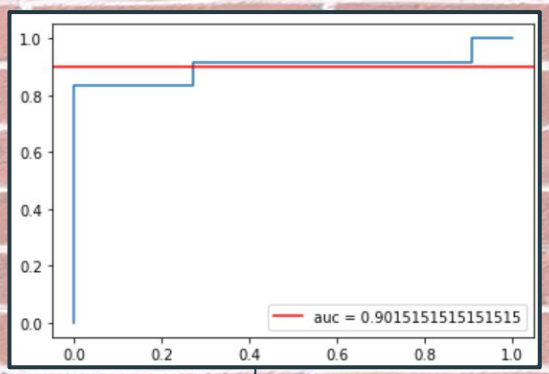



*0,1 distribution*

*Regression plot*




*AUC and ROC curve*

*F1 Score - 0.8260*

*Confusion matrix :*

```
[[ 9  2]
 [ 2 10]]
```

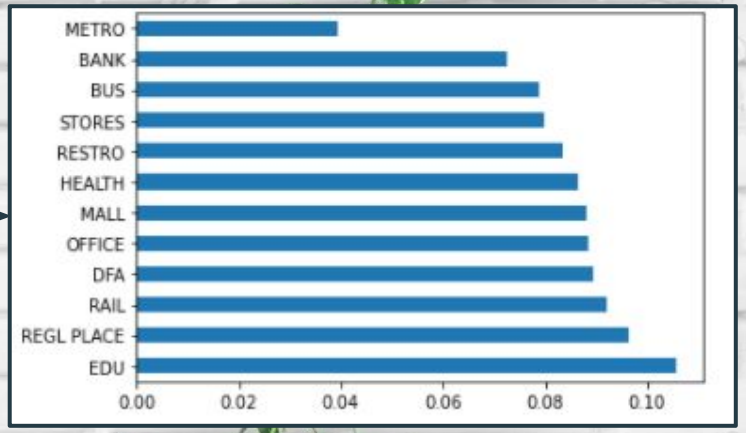
*Classification Report*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.82 | 0.82 | 11 |
| 1 | 0.83 | 0.83 | 0.83 | 12 |
| accuracy | | | 0.83 | 23 |
| macro avg | 0.83 | 0.83 | 0.83 | 23 |
| weighted avg | 0.83 | 0.83 | 0.83 | 23 |

*In our data set f1 score for classifier 0(low) is 0.82 and for classifier 1(high) is 0.83.*

```
   obs freq  exp freq
0
1         12      11.5
0         11      11.5
The test stat is 0.04
The p value is 0.83
We accept null
Each outcome has equal probability of occuring
```

**Goodness of Fit**




**EXTRATREESCLASSIFIER**

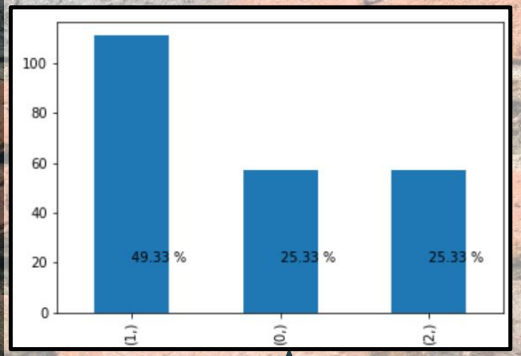
| | Coefficients |
|---|---|
| METRO | -0.034704 |
| RAIL | 0.159894 |
| BUS | 0.023582 |
| DFA | -0.003550 |
| HEALTH | 0.023298 |
| EDU | -0.013448 |
| STORES | -0.000175 |
| MALL | -0.022597 |
| RESTRO | 0.006044 |
| OFFICE | 0.011953 |
| REGL PLACE | 0.015954 |
| BANK | -0.007075 |

**Coefficient Estimates for each Feature Variables**

**5 fold cross validation scores**

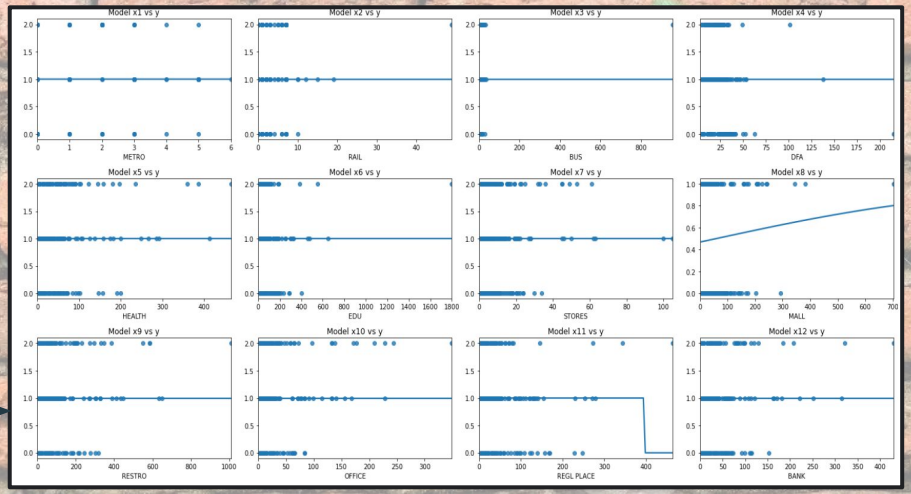
| | 5-fold CV scores |
|---|---|
| 1 | 0.600000 |
| 2 | 0.777778 |
| 3 | 0.666667 |
| 4 | 0.688889 |
| 5 | 0.711111 |

***MULTICLASS LOGISTIC REGRESSION*** - *In this regression model the response variable is dividing into 3 classes, that is, it can take three values, 0, 1, or 2. The classification has been done based on its 3 quantile values (Q1, Q2, Q3).*




*0,1,2 distribution*

*Regression plot*




| | 5-fold CV scores |
|---|---|
| 1 | 0.622222 |
| 2 | 0.755556 |
| 3 | 0.666667 |
| 4 | 0.711111 |
| 5 | 0.733333 |

*5 fold cross validation scores*

```
   obs freq  exp freq
0
1        13  7.666667
2         6  7.666667
0         4  7.666667
The test stat is 5.83
The p value is 0.05
We accept null
Each outcome has equal probability of occuring
```

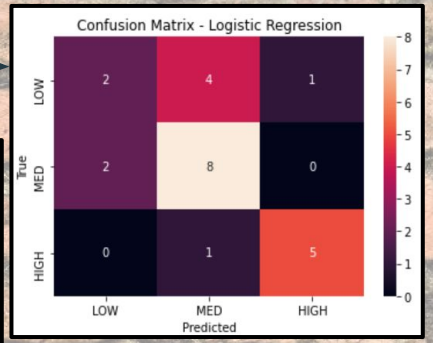
*Goodness of Fit*

*Classification Report*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.50 | 0.29 | 0.36 | 7 |
| 1 | 0.62 | 0.80 | 0.70 | 10 |
| 2 | 0.83 | 0.83 | 0.83 | 6 |
| accuracy | | | 0.65 | 23 |
| macro avg | 0.65 | 0.64 | 0.63 | 23 |
| weighted avg | 0.64 | 0.65 | 0.63 | 23 |

*F1 Score - 0.6521*

*Confusion matrix :*



# CONCLUSION

# SUMMARY

| | Original y values (y) | Predicted y values (y^) |
|---|---|---|
| 0 | 1 | 1 |
| 1 | 1 | 1 |
| 2 | 0 | 0 |
| 3 | 1 | 1 |
| 4 | 1 | 1 |

***Binary Classification***

***Multiclass Classification***

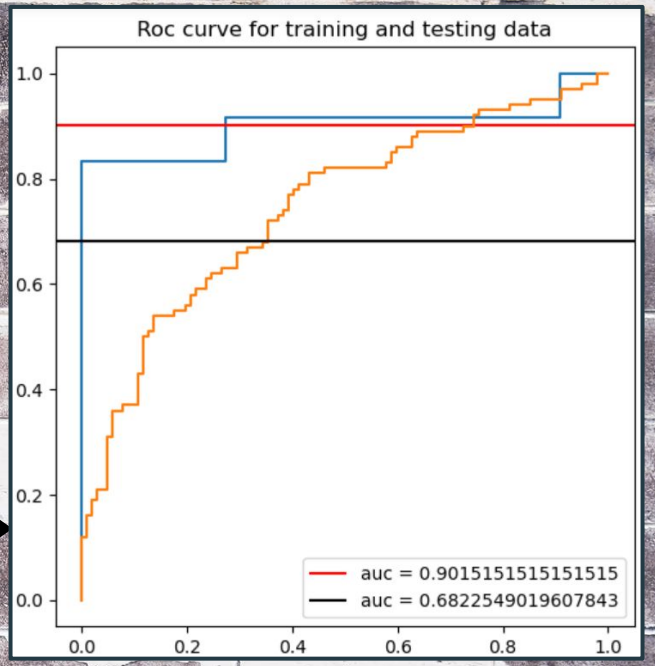
| | Original y values (y) | Predicted y values (y^) |
|---|---|---|
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 2 | 2 | 2 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |

*By analysing our data for the **plot prices of localities within WEST BENGAL**, along with the 12 relevant features, we have established some interesting patterns and trends. Our primary aim was to **predict the price of the plots based on the above mentioned features**. However, as we were unable to do that using Linear Regression, due to the lack of resources and other limitations that we will discuss soon, we proceeded with Logistic Regression.*

*We have tried to predict the price of the plots after dividing them into 2 and 3 ranges; **'High', 'Low'** & **'High', 'Medium', 'Low'** respectively.*

*Further, we have also seen that we have got a better model that has higher accuracy and precision when we perform Binary Classification than that in the case of Multiclass classification.*

*We have also checked the **efficiency of our training and testing datasets** while performing Binary Classification. Both our models give us almost accurate predicted price values based on the 'High' and 'Low' range.*




*We have also shown the most significant feature variable through the **Extratreesclassifier method. Hence,** using different statistical techniques, we were then able to identify the **prices that were predicted correctly,** and also establish a **relationship among our feature variables and our response variable**.*

# LIMITATIONS

## DATA SHORTAGE

1. *The total number of localities considered are not enough to fit an accurate regression model, be it linear model or a logistic model*
2. *The data is not enough for training and testing. Hence, after training the dataset, there are less number of localities left for testing and predicting the accurate prices.*
3. *While classifying the data for logistic regression, it is not enough to split the response variables into the price ranges.*

## FEATURE VARIABLES SHORTAGE

1. *Due to the absence of other important features variables, accurate price prediction was not successful.*
2. *Linear regression could not be done accurately due to the lack of the explanatory power of the feature variables.*
3. *Hence, we have had to shift to classify our response variable(Avg. Price) and perform logistic Regression*

## FUTURE SCOPE

1. *If the data of more localities can be collected, and more features can be recorded, then we will be able to do further study to predict the average prices.*
2. *We can also use different algorithms like PCA, PLS , Ridge regression, Lasso and many more such techniques to make our Linear Regression model fit better.*

# BIBLIOGRAPHY


***Data Source ~***

https://housing.com/in/buy/plots-in-kolkata
https://www.makaan.com/
https://www.google.com/maps/

**Book Reference ~**

ISLRv2_website.pdf

***Documentations ~***

https://numpy.org/doc/
https://pandas.pydata.org/docs/
https://pandas.pydata.org/docs/
https://matplotlib.org/stable/users/index.html
https://seaborn.pydata.org/
https://scikit-learn.org/stable/
https://devdocs.io/statsmodels/

# THANK YOU!!

We would like to say thank you to our honorable VC**, Prof. Saikat Mitra** and respected DIRECTOR of the University**, Prof. Sukhendu Samajdar** & HEAD OF THE DEPT**., Prof. Prasanta Narayan Dutta** for giving us this great opportunity.

We would especially like to thank our SUPERVISOR **Prof. Taranga Mukherjee** for his continuous support and guidance and we are also grateful to the others faculty members.

**Thank you to everyone…**