Housing Data : Linear Regression

Report Submitted By

Chirag Chaudhari

Table of Content

1. Defining Objective and Understanding the Data

1.1 Objective:

The data set contains basic housing data such as proportion of residential land per zone, average number of rooms per dwelling etc. Linear regression is to be plotted to find out the relation between different variables considered in dataset.

1.2 Data Understanding:

Dataset for credit risk analysis consist of different variables. Following table contains the variable name or column name and its description.

Variable Name/ Column Name	Description
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bo unds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 milli on)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per 10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportio n of blacks by town
LSTAT	Percent lower status of the population
MEDV	Median value of owner-occupied homes in 1000's

2. Data Preparation and Exploratory Data Analysis

2.1 Data Preparation:

Data preparation consist of verifying the quality of data, checking missing values, missing value treatment, detection of outlier.

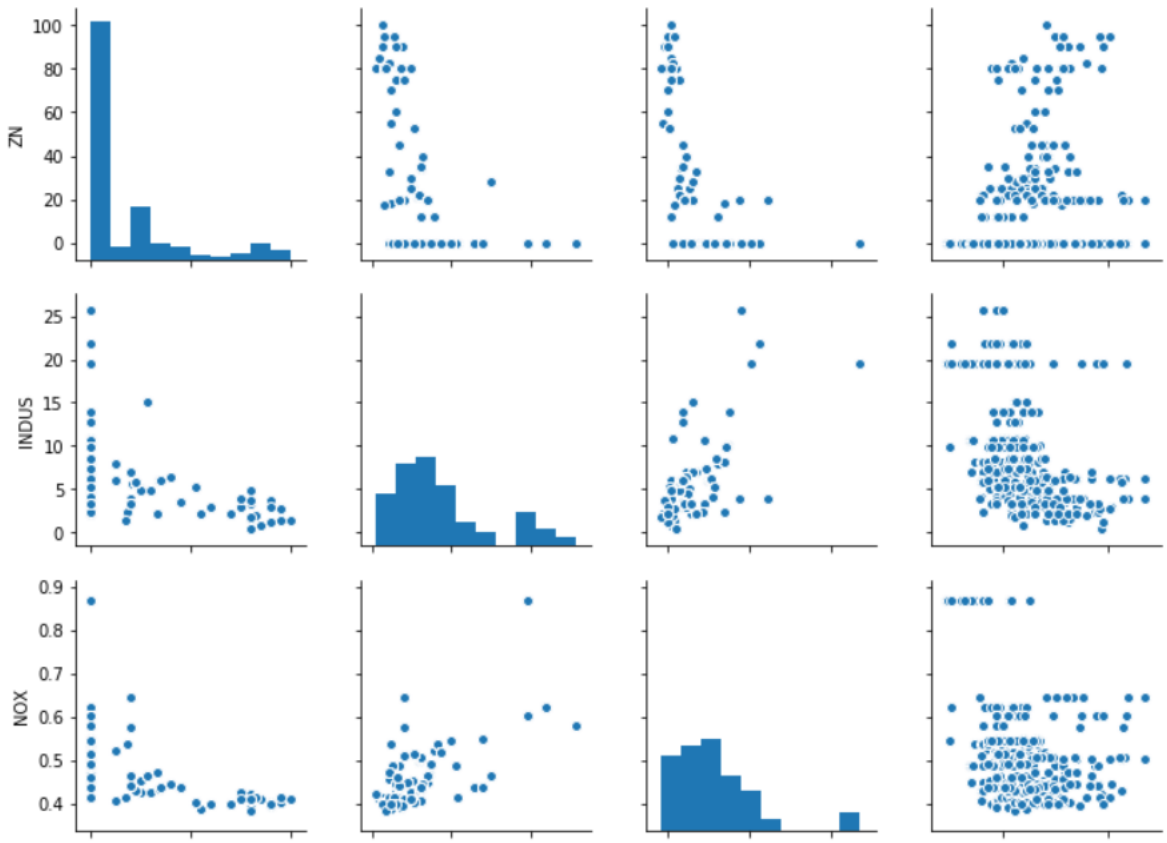
2.2 Missing Values:

The data set contains two missing values. One in LSTAT column and one in MDEV column. The missing values need to be handled. In current case study, missing values are replaced by the median of the data to complete the dataset for further analysis.

2.3 Exploratory data analysis:

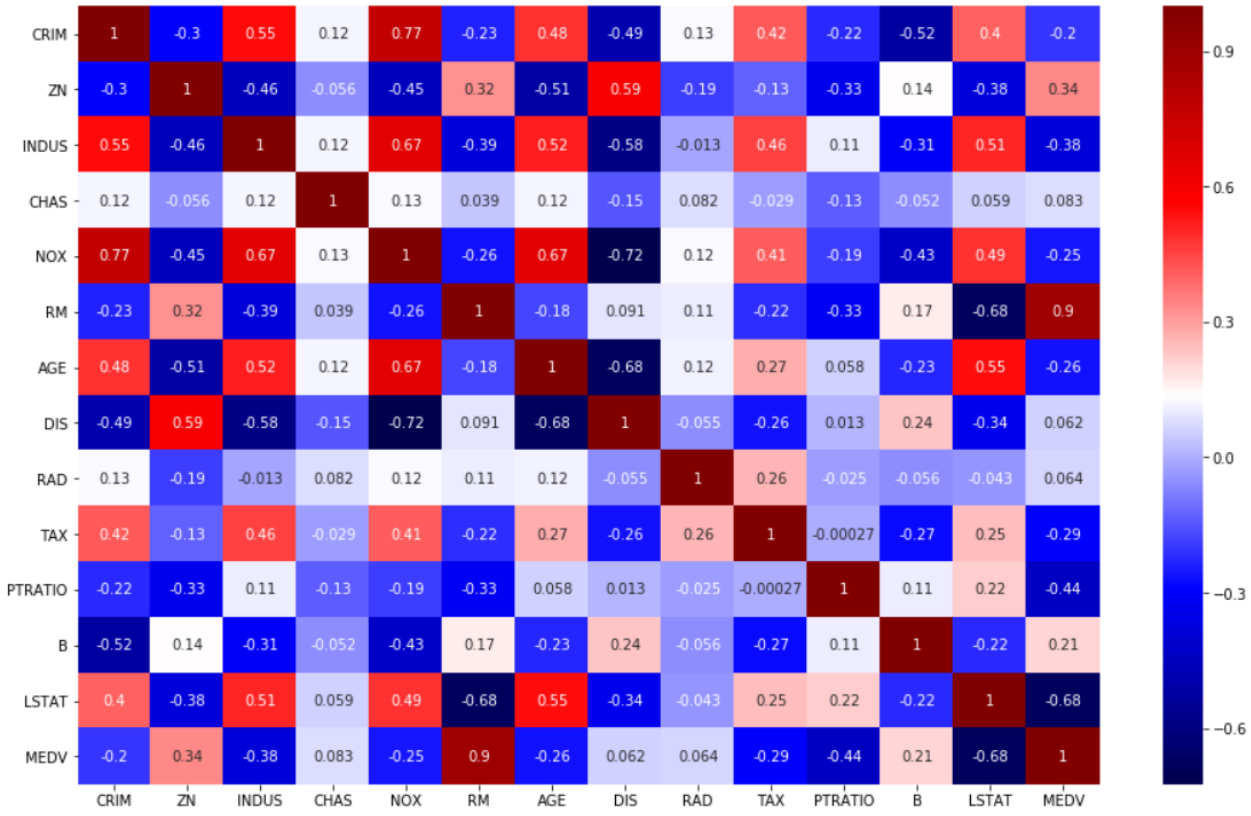
a. Scatter pair plots:

Scatter plots are plotted for each variable with all the variables of dataset to chech the relation between the variables. Form this plots some variables are selected to study their relation in depth. Below are some of the scatter plots which are plotted to find the relation between variables.



b. Correlation:

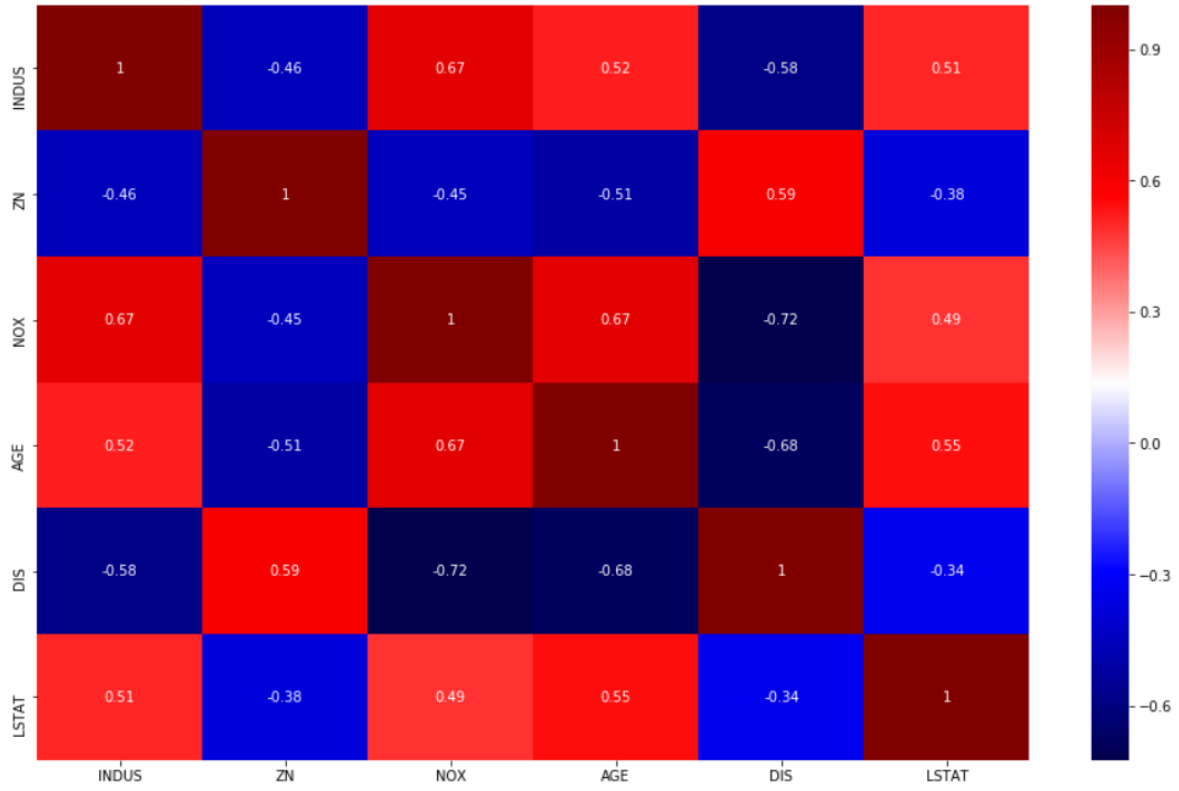
To find out the correlation between the variables, heatmap created using seaborn library. First heatmap is created for all variables, then some variables are selected having greater relation, based on scatter plot study, and heatmap is created for those variables. Below are some heatmaps.

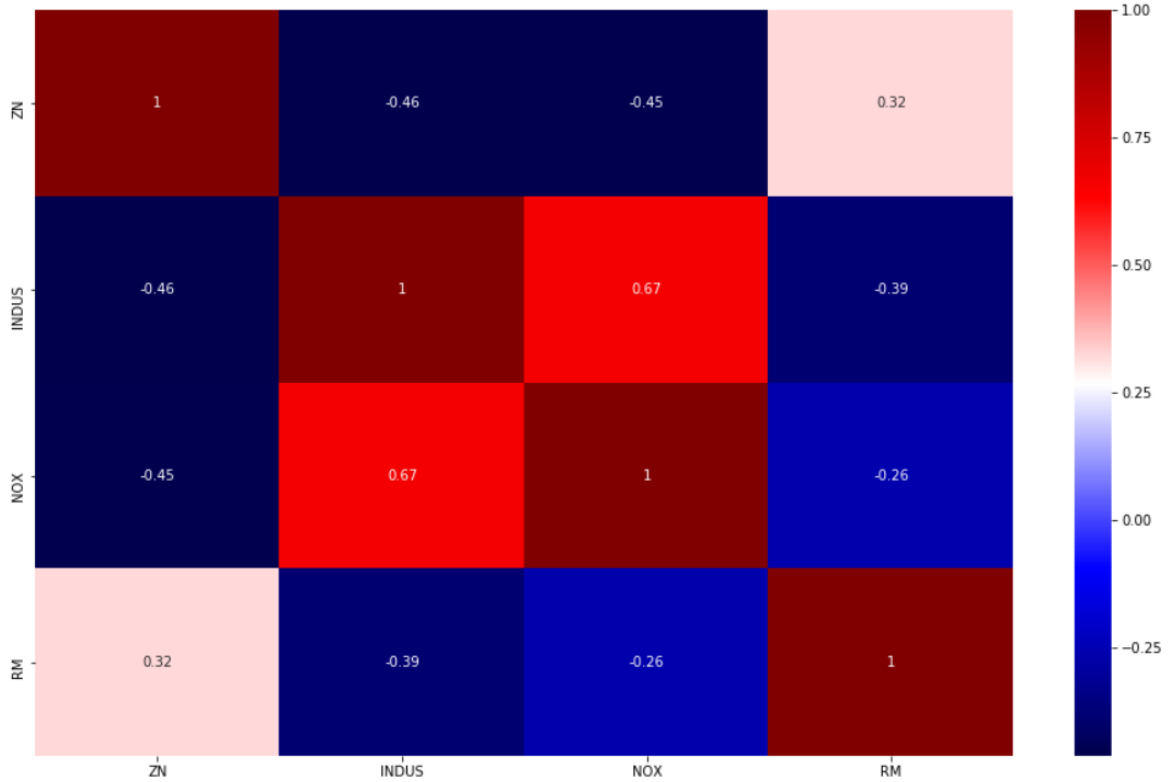


Above is the heatmap for all variables in dataset. Some variable shows positive correlation and some variables shows negative relation.

From above heatmap it is observed that variable RM (average number of rooms per dwelling) and MEDV (Median value of owner-occupied homes in 1000's) shows strong positive correlation whereas NOX (nitric oxides concentration) and DIS (weighted distance to five Boston employment centres) shows strong negative correlation.

Similar heatmaps are created for study purpose by selecting some variables from given dataset.


INDUS
ZN
NOX
AGE
DIS
LSTAT
1 -0.46 0.67 0.52 -0.58 0.51
-0.46 1 -0.45 -0.51 0.59 -0.38
0.67 -0.45 1 0.67 -0.72 0.49
0.52 -0.51 0.67 1 -0.68 0.55
-0.58 0.59 -0.72 -0.68 1 -0.34
0.51 -0.38 0.49 0.55 -0.34 1
0.9
0.6
0.3
0.0
−0.3
−0.6


ZN
INDUS
NOX
RM
1 -0.46 -0.45 0.32
-0.46 1 0.67 -0.39
-0.45 0.67 1 -0.26
0.32 -0.39 -0.26 1
1.00
0.75
0.50
0.25
0.00
−0.25

3. Model Building

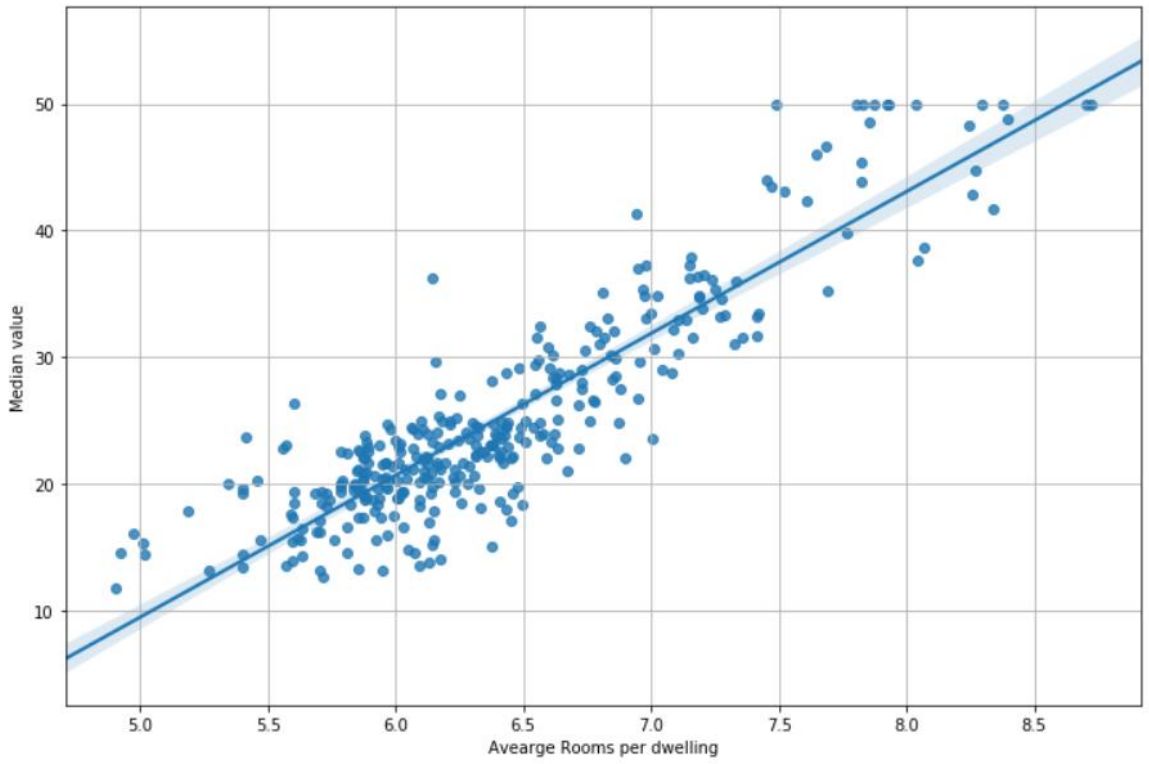
3.1 Linear Regression:

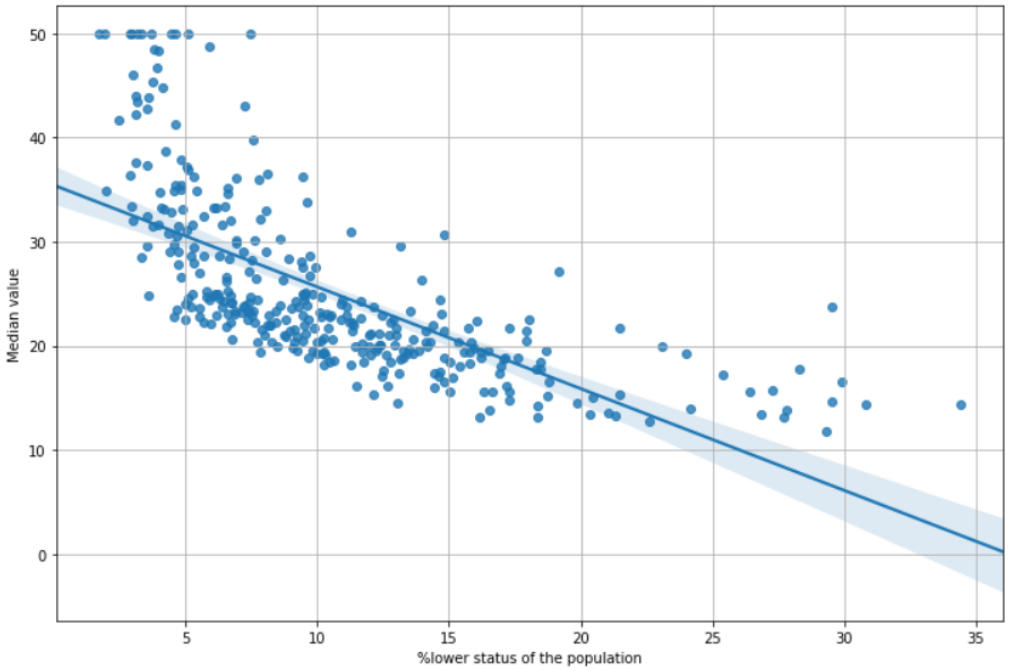
Linear regression is useful for finding relationship between continuous variables. One is predictor or independent variable and other is response or dependent variable. Linear regression looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

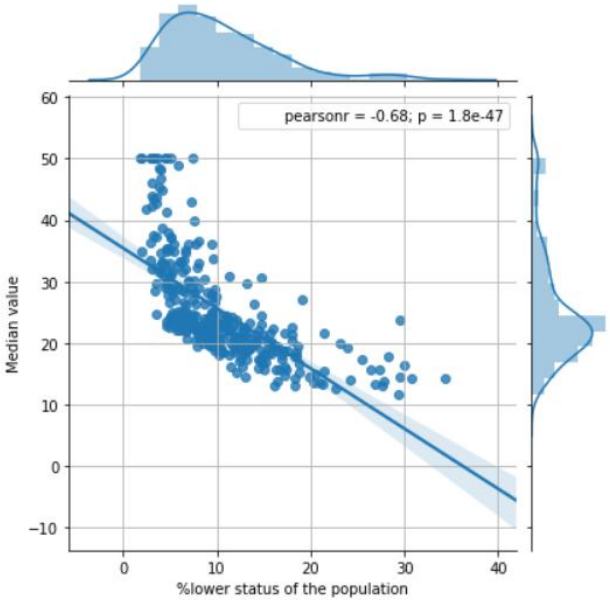
When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

For linear regression modelling sklearn liabrary is used in this case study.

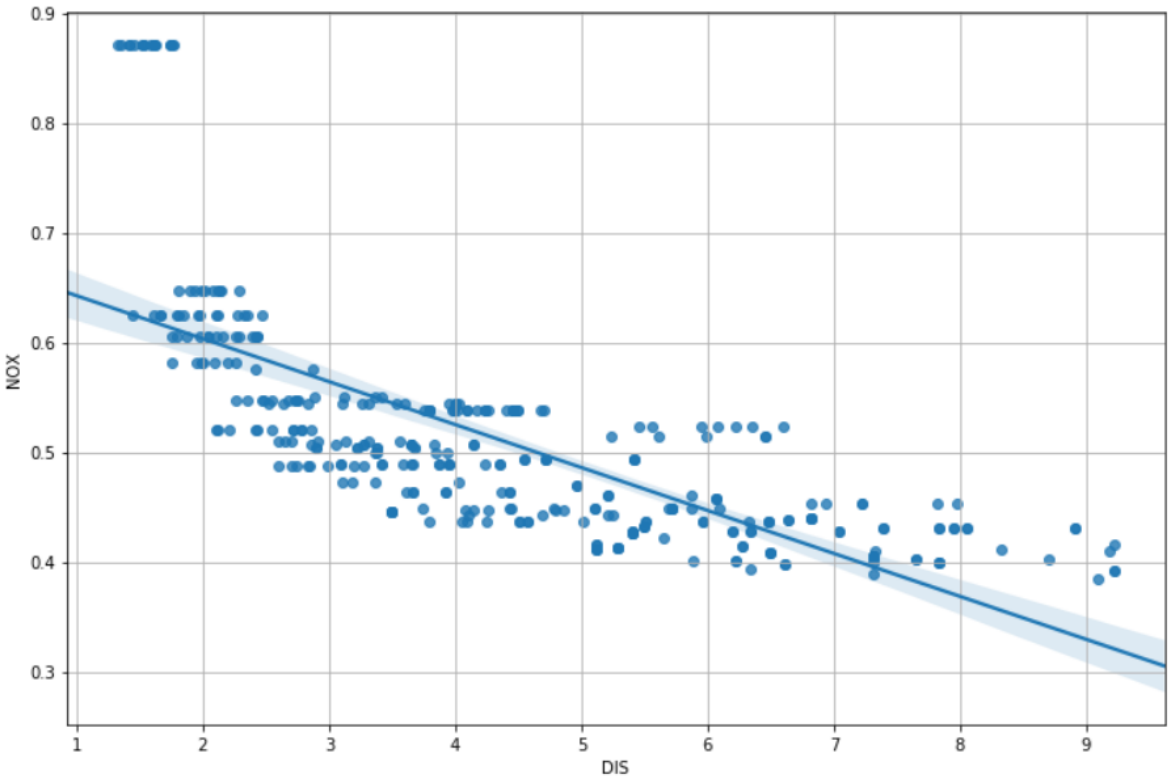
Below are some regression plots and joints plots showing correlation between the variables.

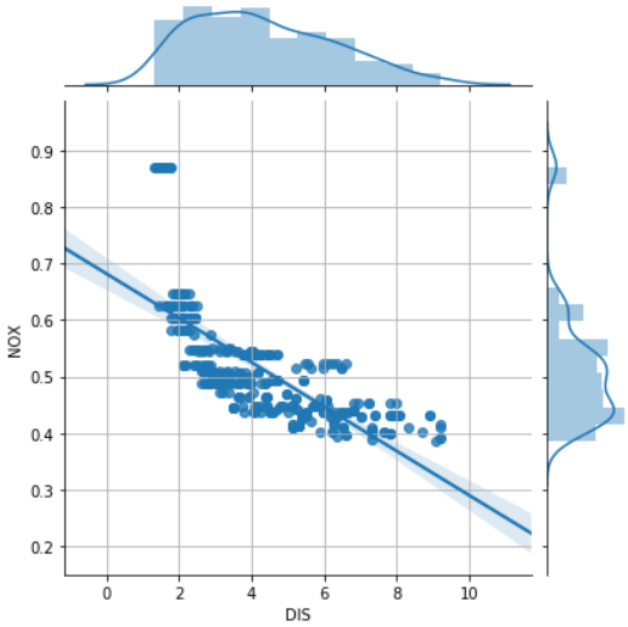






Above linear regression graph shows correlation between %lower status of population and median value.





Above linear regression graph shows correlation between weighted distance from five Boston employment centres and nitric oxide concentration. For the graph it is observed that both variable has negative correlation.