**Department of Statistics**
**Jahangirnagar University**

**Professional Masters in Applied Statistics and Data Science (ASDS)**
**Course Title: Introduction to Data Science with Python**
**Course No.: PM-ASDS04**

# Assignment – 1

# EDA Report on Boston housing Data Set

**Submit to –**
**Dr. Rumana Rois**
**Associate Professor, Department of Statistics**
**Professional Masters in Applied Statistics and Data Science (ASDS)**
**Jahangirnagar University**

**Submitted by –**
**Md. Shamimul Islam**
**ID:20204012**
**Section-B, 4th Batch**
**Professional Masters in Applied Statistics and Data Science (ASDS)**
**Jahangirnagar University**

**December 13, 2020**

# EDA Report on Boston house Data

**Research Problem:** The main purpose of this study is to predict the price of house. Having a housing price prediction model can be a very important tool for both the seller and the buyer as it can aid them in making well informed decision. For sellers, it may help them to determine the average price at which they should put their house for sale while for buyers, it may help them find out the right average price to purchase the house.

***The current goal is to analysis the data which is the process of understanding, cleaning, transforming and modeling data for discovering useful information, deriving conclusions and making data decisions. We will make a cursory investigation of the Boston housing data. Data analysis is prerequisite for building a model.***

In this work, we will make a complete Exploratory Data Analysis(EDA) which is initial investigations on data, to discover patterns, to spot anomalies, to test hypothesis and to check assumptions. It is a very important step before training the model. The Exploratory Data Analysis (EDA) performed on the Housing data set employed a variety of statistical analysis and visualization methods to gain insight into the data and attempt to understand the story the data are telling. The following steps represent Exploratory Data Analysis (EDA) for our data set.

## 1.1 Data Description:

The Boston housing data was collected from various suburbs in Boston, Massachusetts in 1978 by Harrison and Rubinfeld. There are 506 samples and 13 feature variables in this dataset.
To find the dataset description

from sklearn import datasets
boston= datasets.load_boston ()
####Now transform the data as a pandas's DATAFRAME
import pandas as pd
df = pd.DataFrame(boston.data ,columns = boston.feature_names)
df['price']=boston.target

# List of data series
datarowsSeries = [pd.Series([0.069,10,2.3,0,0.53,6.5,65.2,4.01,1,290,15,395,4.9,24.0],
index=df.columns ), pd.Series([0.69+12, 10+12, 2.3+.12, 0, 0.5+.12, 6.5+.12, 65.2+.12, 4.1+.12, 1, 290+12, 15+12, 395+12, 4.9+.12, 24.3+12],
index=df.columns ), pd.Series([0.68+12, 11+12, 2.4+.12, 0, 0.6+.12, 6.6+.12, 65.1+.12, 4.0+.12, 1, 291+12, 13+12, 390+12, 4.2+.12, 24.2+12],
index=df.columns ), pd.Series([0.67+12, 12+12, 2.5+.12, 0, 0.4+.12, 6.5+.12, 65.3+.12, 4.2+.12, 1, 292+12, 14+12, 392+12, 4.3+.12, 24.1+12],
index=df.columns ), pd.Series([0.66+12, 13+12, 2.4+.12, 0, 0.7+.12, 6.5+.12, 65.4+.12, 4.1+.12, 1, 293+12, 16+12, 391+12, 4.4+.12, 24.2+12],
index=df.columns ) ]
new_data = df.append(datarowsSeries , ignore_index=True)

Code to print decription of our data

***print(boston['DESCR'])***

```
shamim@shamim:~/Dropbox/Data_Science_JU/First Semester/Introduction to Data Science with Python/Lab/Assignment_1$ python3 eda_analysis_boston_data.py
~~~~~~~~~~~~~~~~~~~~~~~~~~~~To print the dataset description~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
.. _boston_dataset:

Boston house prices dataset
---------------------------

**Data Set Characteristics:**

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

    :Attribute Information (in order):
        - 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 bounds river; 0 otherwise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - 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 proportion of blacks by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's

    :Missing Attribute Values: None

    :Creator: Harrison, D. and Rubinfeld, D.L.

This is a copy of UCI ML housing dataset.
https://archive.ics.uci.edu/ml/machine-learning-databases/housing/


This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.

The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic
prices and the demand for clean air', J. Environ. Economics & Management,
vol.5, 81-102, 1978.   Used in Belsley, Kuh & Welsch, 'Regression diagnostics
...', Wiley, 1980.   N.B. Various transformations are used in the table on
pages 244-261 of the latter.

The Boston house-price data has been used in many machine learning papers that address regression
problems.

.. topic:: References
```

*Figure 1: The description of dataset*

***print(boston.feature_names)***

```
shamim@shamim:~/Dropbox/Data_Science_JU/First Semester/Introduction to Data Science with Python/Lab/Assignment_1$ python3 eda_analysis_boston_data.py
~~~~~~~~~~~~~~~~~~~~~~~~~~~~To print feature names of dataset~~~~~~~~~~~~~~~~~~~~~~~~~~~~
['CRIM' 'ZN' 'INDUS' 'CHAS' 'NOX' 'RM' 'AGE' 'DIS' 'RAD' 'TAX' 'PTRATIO'
 'B' 'LSTAT']
```

*Figure 2: The feature names of dataset*

***print("Type of boston dataset:", boston.data.shape)***

```
shamim@shamim:~/Dropbox/Data_Science_JU/First Semester/Introduction to Data Science with Python/Lab/Assignment_1$ python3 eda_analysis_boston_data.py
~~~~~~~~~~~~~~~~~~~~~~~~~~~To print data shape of dataset~~~~~~~~~~~~~~~~~~~~~~~~~~~
Type of boston dataset: (506, 13)
```

*Figure 3: The data shape*

Table 1 represents all the variables description with their value level, level of measurements, and suitable measures.

Table 1: Variables summary information of Haberman’s Survival Dataset.

| Variable Name | Variable Description | Value level | Level of Measurements | Appropriate measures |
|---|---|---|---|---|
| CRIM | Per capita crime rate by town. Since CRIM gauges the threat to well-being that households perceive in various neighborhood of the Boston(assuming that crimes rates are generally proportional to people’s perceptions of denger) , it should have a bad effect on housing values. | | Ratio | Mean, median, mode |
| ZN | Proportion of a town’s residential land zoned for lots greater than 25,000 square feet. Since such zoning restricts construction of small lot houses, we expect ZS to be positively related to housing values. A positive coefficient may also arise because zoning proxies the exclusivity, social | | Ratio | Mean, median, mode |

| | | | | |
|---|---|---|---|---|
| | class, and outdoor amenities of a community. | | | |
| INDUS | Proportion nonretail business acres per town. ISDUS serves as a proxy for the externalities associated with industry-noise, heavy traffic, and unpleasant visual effects. | | Ratio | Mean, median, mode |
| CHAS | Charles River dummy : = 1 if tract bounds the Charles River; =0 if ot,herwise. CHAS captures the amenities of a riverside location and thus the coefficient should be positive. | 1 if tract bounds river; 0 otherwise | Nominal | Mode |
| NOX | Nitric oxides concentration (parts per 10 million) | | Ratio | Mean, median, mode |
| RM | Average number of rooms in owner units. RM represents spaciousness and, in a certain sense, quantity of housing. It should be positively related to housing value. The $RM^2$ form was found to provide a better fit than either the linear or logarithmic forms. | | Ratio | Mean, median, mode |
| AGE | Proportion of owner units built prior to 1940. Unit age is generally related to structure quality. | | Ratio | Mean, median, mode |
| DIS | Weighted distances to five Boston employment centres | | Ratio | Mean, median, mode |
| RAD | Index of accessibility to radial highways. Good road acress variables are needed so that auto pollution variables do not | | Ordinal | Mode |

| | | | | |
|---|---|---|---|---|
| | capture the locational advantages of roadways. RAD captures other sorts of locational advantages besides nearness to workplace. It is entered in logarithmic form ; the experted sign is positive. | | | |
| TAX | Full-value property-tax rate per USD 10,000 | | Ratio | Mean, median, mode |
| PTRATIO | Pupil-teacher ratio by town school district. Measures public sector benefits in each town. The relation of the pupil&teacher ratio to school quality is not entirely clear, although a low ratio should imply each student receives more individual attention. We expect the sign on PTRATIO to be negative. | | Ratio | Mean, median, mode |
| B | Black proportion of population. At low to moderate levels of B, an increase in B should have a negative influence on housing value if Blacks are regarded as undesirable neighbors by Whites. However, market discrimination means that housing values are higher at very high levels of B. One expects, therefore, a parabolic relationship between proportion Black in a neighborhood and housing values. | | Ratio | Mean, median, mode |
| LSTAT | Proportion of population that is lower status = ½ (proportion of adults without, some high | | Ratio | Mean, median, mode |

| | | | | |
|---|---|---|---|---|
| | school education and proportion of male workers classified as laborers). The logarithmic specification implies that socioeconomic status distinctions mean more in the upper brackets of society than in the lower classes. | | | |
| MEDV | Median value of owner-occupied homes in USD 1000's | | Ratio | Mean, median, mode |

## *1.2 Objectives for EDA of Boston Housing Dataset:*

Price of house is the main response variable for our study, as we have to predict that using remaining variables. Hence, to establish the main objective of our study, we could specify the following objectives.

- To specify the distribution of CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B and LSTAT
- To reveal the different proportion of price of house after investigation of remaining variables.
- To calculate different summary measures of CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B and LSTAT.
- To find the association between Price with other variables.
- To illustrate correlation matrix of different variables.

## *1.3 Data Cleaning:*

A survey found that Data scientists spend 60% of their time on cleaning and organizing data. Data cleaning and preparation is the most critical first step in any Data Science project. Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. To clean the data set, you need to handle missing values, categorical features and others steps, because the mathematics underlying most machine learning models assumes that the data is numerical and contains no missing values.

### *1.3.1 Handling missing data*

Missing data is perhaps the most common trait of unclean data. These values usually take the form of NaN or None. Here are several causes of missing values: sometimes values are missing because they do not exist, or because of improper collection of data or poor data entry. Missing values need to be handled carefully because they reduce the quality of any of our performance matrix. It can also lead to wrong prediction or classification and can also cause a high bias for any given model being used.

Now, we need to search for any missing data. The missing data is normally converted into NaN values by the Pandas Dataframe. *df.isnull().sum()* returns the amount of null values in a particular column or feature.

***printf(df.isnull().sum())***

```
shamim@shamim:~/Dropbox/Data_Science_JU/First Semester/Introduction to Data Science with Python/Lab/Assignment_1$ python3 eda_analysis_boston_data.py
~~~~~~~~~~~~~~~~~~~~~~~~~~~~To print the missing data~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
price      0
```

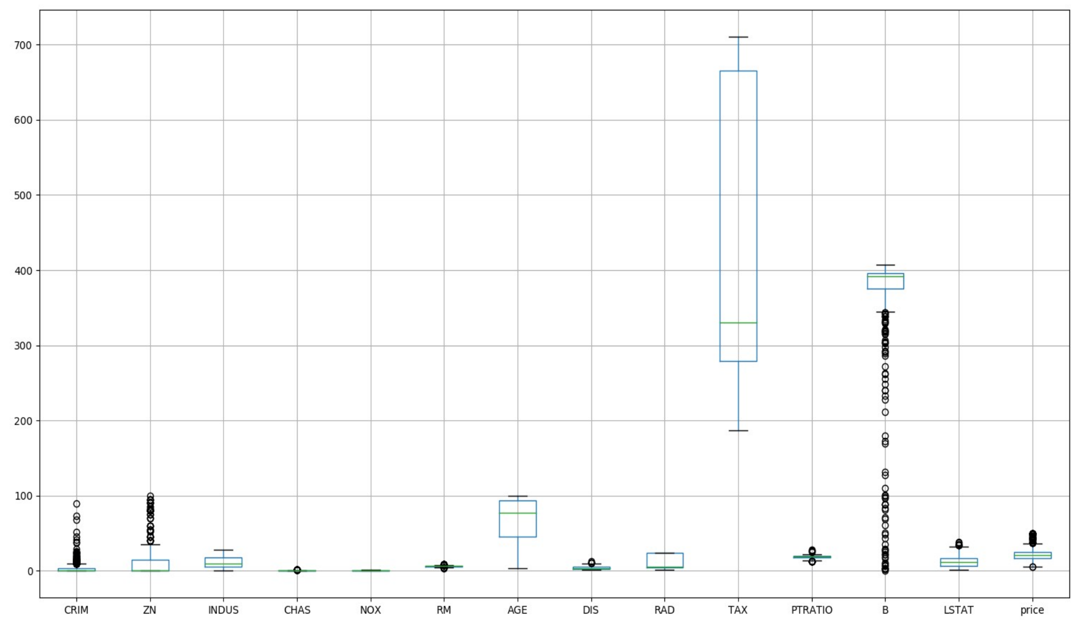
*Figure 4: The amount of null values in Data set*

The above code indicates that there are no null values in our data set

### *1.3.2 Dealing with outliers*

An outlier is something which is separate or different from the crowd. Outliers can be a result of a mistake during data collection or it can be just an indication of variance in your data. By using outliers, we can easily detect outliers.

Now we will construct the boxplot to found the outliers in this dataset.
The code for box plot all variables




***df.boxplot()***
***plt.show()***
*Figure 5: Boxplot of all of variables of Boston Housing Dataset*

From the above code , we found that CRIM, ZN, RM, DIS, PTRTIO, B and LSTAT, do have some outliers but variables, INDUS, CHAS, NOX, AGE, RAD and TAX do not have outlier. We may delete all the observations along with CRIM, ZN, RM, DIS, PTRTIO, B, and LSTAT outliers, hereafter, we may use some robust statistics to calculate different features of this dataset. However, at this point we will consider entire dataset to calculate various measures (as we are not aware of different robust statistics).

### *1.3.3 Handling categorical data*

Data sets often contain the object data type than needs to be transformed into numeric. In this session, you will check data-types of our dataset.

***print(df.dtypes)***

```
shamim@shamim:~/Dropbox/Data_Science_JU/First Semester/Introduction to Data Science with Python/Lab/Assignment_1$ python3 eda_analysis_boston_data.py
~~~~~~~~~~~~~~~~~~~~~~~~~~~Print datatypes of variables~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
CRIM       float64
ZN         float64
INDUS      float64
CHAS       float64
NOX        float64
RM         float64
AGE        float64
DIS        float64
RAD        float64
TAX        float64
PTRATIO    float64
B          float64
LSTAT      float64
price      float64
dtype: object
```

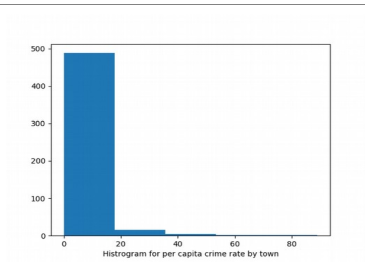
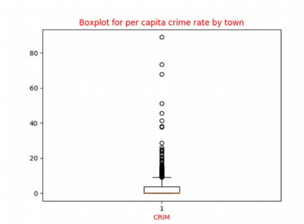
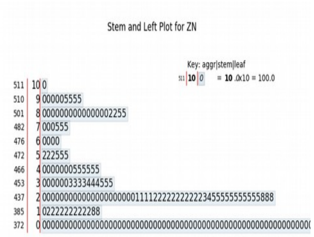
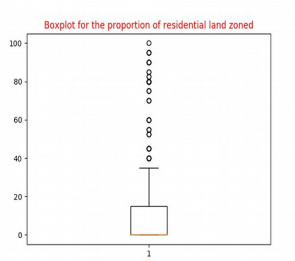
*Figure 6: The* data-types of our dataset

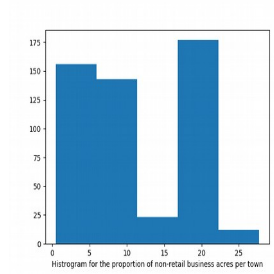
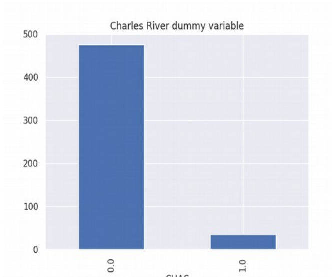
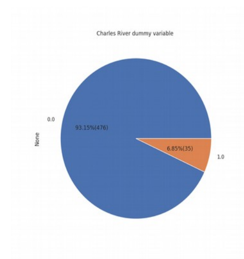
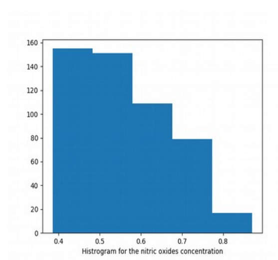
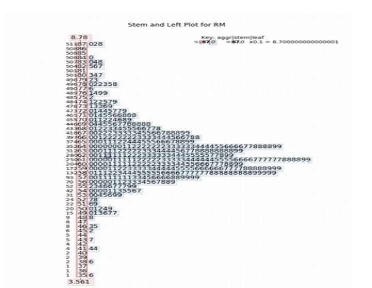
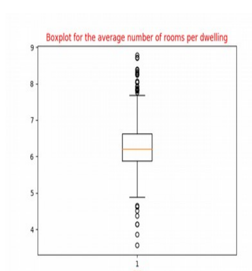
This represents that all the fields is of float type and therefore most likely they are a continuous variable, including our target.

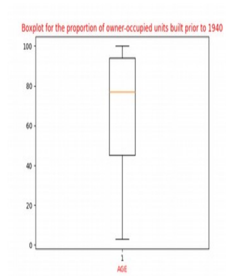
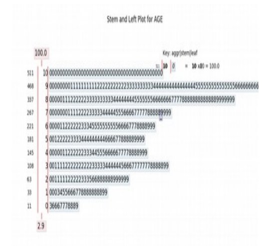
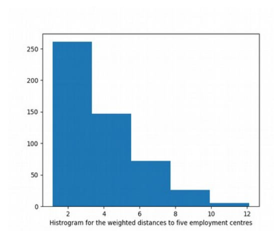
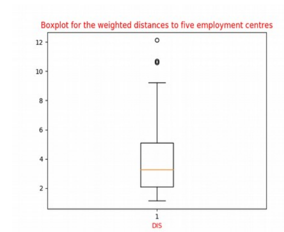
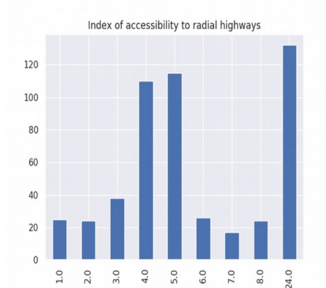
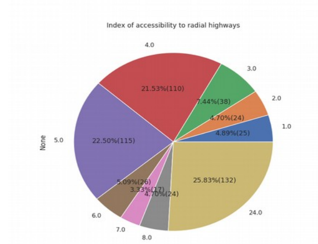
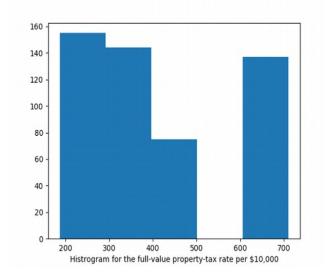
## *1.4 Univariate Analysis:*

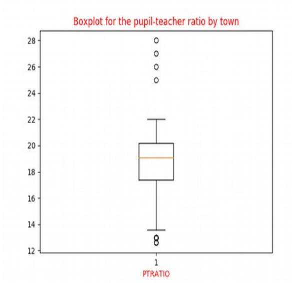
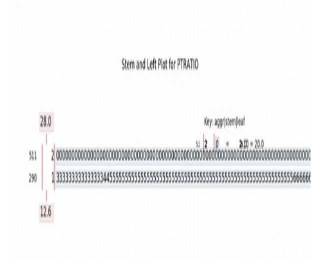
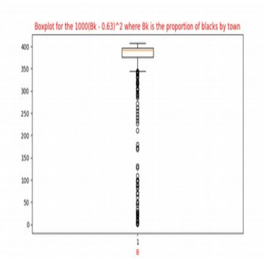
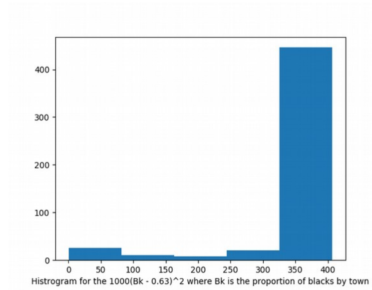
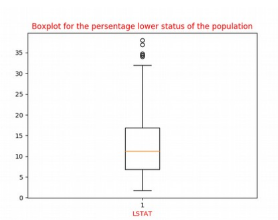
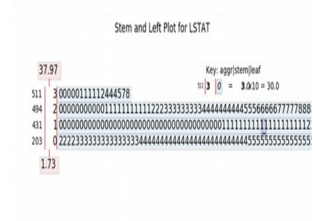
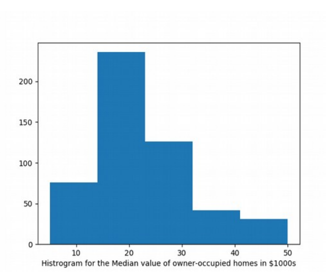
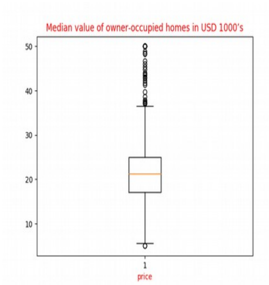
In this Boston Housing data, variables “CRIM”, “ZN”, “RM”, “DIS”, “PTRTIO”, “B”, “LSTAT”, “INDUS”, “NOX”, “AGE” and “TAX” are ratio, the variable ‘RAD’ is ordinal, and the variable ‘CHAS’ is nominal. Hence, we could make a pie chart or a bar diagram for the variable ‘CHAS’, and a histogram, a boxplot or a stem-and-leaf plot for the variables “CRIM”, “ZN”, “RM”, “DIS”, “PTRTIO”, “B”, “LSTAT”, “INDUS”, “NOX”, “AGE” and “TAX” and a bar diagram for the variable “RAD”.

## *1.4.1 Graphical Representations:*

| Variable Name | Types of Variable | Code of Graphical Representation | Graphical Representation | Observation |
|---|---|---|---|---|
| CRIM | Ratio | #for histogram<br>plt.hist(df.CRIM, bins=5)<br>plt.xlabel('Histrogram for per capita crime rate by town')<br>plt.show()<br><br>#for boxplot<br>plt.boxplot(df.CRIM)<br>plt.title('Boxplot for per capita crime rate by town', color='RED')<br>plt.xlabel('CRIM', color='RED')<br>plt.show() | <br>Histogram<br><br>Boxplot | 1. Positively skewed distribution.<br>2. It has outliers. |
| ZN | Ratio | #For Stem and Leaf<br>fig, ax = stemgraphic.stem_graphic(df.ZN, scale=10)<br>ax.set_title("Stem and Left Plot for ZN")<br>plt.show()<br><br>#for Boxplot<br>plt.boxplot(df.ZN)<br>plt.title('Boxplot for the proportion of residential land zoned', color='RED') | <br>Stem and Left Plot<br><br>Boxplot | 1. Skewed right so positively skewed distribution.<br>2. It has outliers. |

| | | | | |
|---|---|---|---|---|
| | | plt.xlabel('ZN',<br>color='RED')<br>plt.show() | | |
| INDUS | Ratio | *#for Boxplot*<br>*plt.boxplot(df.INDUS)*<br>*plt.title('Boxplot for the proportion of non-retail business acres per town', color='RED')*<br>*plt.xlabel('INDUS', color='RED')*<br>*plt.show()*<br><br>*#For Histogram*<br>*plt.hist(df.INDUS, bins=5)*<br>*plt.xlabel('Histrogram for the proportion of non-retail business acres per town')*<br>*plt.show()* |  <br>Boxplot<br> <br>Histogram | 1. Skewed right so positively skewed distribution.<br>2. It has no outliers. |
| CHAS | Nominal | #for bar chart<br>s=df.groupby('CHAS').size()<br>sns.set()<br>s.plot(kind='bar', title='Charles River dummy variable', stacked=True)<br>plt.show()<br><br>#for pie chart<br>s=df.groupby('CHAS').size()<br>sns.set()<br>s.plot(kind='pie', title='Charles River dummy variable ', figsize=[8,8],<br>autopct=lambda p: '{:.2f}%<br>({:.0f})'.format(p,(p/100)*s.sum()))<br>plt.show() |  <br>Bar Chart<br> <br>Pie Chart | 1 is 6.85% means tract bounds river<br>0 is 93.15% |
| NOX | Ratio | *#For Box Plot*<br>*plt.boxplot(df.NOX)*<br>*plt.title('Boxplot for the nitric oxides concentration ', color='RED')*<br>*plt.xlabel('NOX', color='RED')*<br>*plt.show()*<br><br>#For Histogram<br>plt.hist(df.NOX, bins=5)<br>plt.xlabel('Histrogram for the nitric oxides concentration')<br>plt.show() |  <br>Boxplot<br> <br>Histogram | 1. Skewed right so positively skewed distribution.<br>2. It has no outliers. |
| RM | Ratio | *#for Stem and left*<br>*fig, ax = stemgraphic.stem_graphic(df.RM)*<br>*ax.set_title("Stem and Left Plot for RM")*<br>*plt.show()*<br><br>*#for Boxplot*<br>*plt.boxplot(df.RM)*<br>*plt.title('Boxplot for the average number of rooms per dwelling', color='RED')*<br>*plt.xlabel('RM', color='RED')*<br>*plt.show()* |  <br>Step and Left<br> <br>Boxplot | 1. It has symmetric distribution.<br>2. It has outliers. |
| AGE | Ratio | *#for box plot*<br>*plt.boxplot(df.AGE)* | | 1. Skewed left so negatively skewed |

| | | | | |
|---|---|---|---|---|
| | | *plt.title('Boxplot for the proportion of owner-occupied units built prior to 1940', color='RED')*<br>*plt.xlabel('AGE', color='RED')*<br>*plt.show()*<br><br>*#For Stem and Left plot*<br>*fig, ax = stemgraphic.stem_graphic(df.AGE, scale=10)*<br>*ax.set_title("Stem and Left Plot for AGE")*<br>*plt.show()* |   Boxplot   Step and Left | distribution.<br>2. It has no outliers. |
| DIS | Ratio | *#For Histogram*<br>*plt.hist(df.DIS, bins=5)*<br>*plt.xlabel('Histrogram for the weighted distances to five employment centres')*<br>*plt.show()*<br><br>*#for box plot*<br>*plt.boxplot(df.DIS)*<br>*plt.title('Boxplot for the weighted distances to five employment centres', color='RED')*<br>*plt.xlabel('DIS', color='RED')*<br>*plt.show()* |   Step and Left   Boxplot | 1. Skewed right so positively skewed distribution.<br>2. It has some outliers. |
| RAD | Ordinal | *#for bar chart*<br>*s=df.groupby('RAD').size()*<br>*sns.set()*<br>*s.plot(kind='bar', title='Index of accessibility to radial highways', stacked=True)*<br>*plt.show()*<br><br>*#for pie chart*<br>*s=df.groupby('RAD').size()*<br>*sns.set()*<br>*s.plot(kind='pie', title='Index of accessibility to radial highways', figsize=[8,8],*<br>*autopct=lambda p: '{:.2f}%({:.0f})'.format(p,(p/100)*s.sum()))*<br>*plt.show()* |   Bar Chart   Pie Chart | There are nine index of accessibility to radial highways. Index 1 is 4.89%, Index 2 is 4.70%, Index 3 is 7.44%, Index 4 is 21.53%, Index 5 is 22.50%, Index 6 is 5.09%, Index 7 is 3.33%, Index 8 is 4.70%, Index 24 is 25.83%. |
| TAX | Ratio | *#for Box Plot*<br>*plt.boxplot(df.TAX)*<br>*plt.title('Boxplot for the full-value property-tax rate per $10,000', color='RED')*<br>*plt.xlabel('TAX', color='RED')*<br>*plt.show()*<br><br>*#For Histogram*<br>*plt.hist(df.TAX, bins=5)*<br>*plt.xlabel('Histrogram for the full-value property-tax rate per $10,000')*<br>*plt.show()* |   Boxplot   Histogram | 1. Skewed right so positively skewed distribution.<br>2. It has no outliers. |

| | | | | |
|---|---|---|---|---|
| PTRATIO | Ratio | *#for Box plot<br>plt.boxplot(df.PTRATIO)<br>plt.title('Boxplot for the pupil-teacher ratio by town', color='RED')<br>plt.xlabel('PTRATIO', color='RED')<br>plt.show()<br><br>#For Stem and Left plot<br>fig, ax = stemgraphic.stem_graphic(df.PTRATIO, scale=10)<br>ax.set_title("Stem and Left Plot for PTRATIO")<br>plt.show()* | <br>Boxplot<br><br>Stem and left | 1. Skewed left so negatively skewed distribution.<br>2. It has some outliers. |
| B | Ratio | *#for Box plot<br>plt.boxplot(df.B)<br>plt.title('Boxplot for the 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town', color='RED')<br>plt.xlabel('B', color='RED')<br>plt.show()<br><br>#For Histogram<br>plt.hist(df.B, bins=5)<br>plt.xlabel('Histrogram for the 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town')<br>plt.show()* | <br>Boxplot<br><br>Histogram | 1. Skewed left so negitively skewed distribution.<br>2. It has outliers. |
| LSTAT | Ratio | *#for Box plot<br>plt.boxplot(df.LSTAT)<br>plt.title('Boxplot for the persentage lower status of the population', color='RED')<br>plt.xlabel('LSTAT', color='RED')<br>plt.show()<br><br>#For Stem and Left plot<br>fig, ax = stemgraphic.stem_graphic(df.LSTAT, scale=10)<br>ax.set_title("Stem and Left Plot for LSTAT")<br>plt.show()* | <br>Boxplot<br><br>Step and Left | 1. Skewed Right so positively skewed distribution.<br>2. It has some outliers. |
| price | Ratio | *#For Histogram<br>plt.hist(df.price, bins=5)<br>plt.xlabel('Histrogram for the Median value of owner-occupied homes in $1000s')<br>plt.show()<br><br>#for Box Plot<br>plt.boxplot(df.price)<br>plt.title('Median value of owner-occupied homes in USD 1000's', color='RED')<br>plt.xlabel('price', color='RED')<br>plt.show()* | <br>Histogram<br><br>Box Plot | 1. It has symmetric distribution.<br>2. It has some outliers. |

## 1.4.2 Summary Measures:

*Using df.describe(), df.mode(), df.median() commands we can make the following table.*

*Table 3: Summary measures of different variables of Haberman's Survival Dataset.*

| Variable Name | Mean | Median | Mode | Standard Deviation |
|---|---|---|---|---|
| CRIM | 3.67 | 0.26169 | 0.01501 | 8.598009 |
| ZN | 11.455969 | 0.00000 | 0.0 | 23.232827 |
| INDUS | 11.052035 | 9.69000 | 18.1 | 6.879777 |
| CHAS | 0.068493 | 0.00000 | 0.0 | 0.252838 |
| NOX | 0.555549 | 0.53800 | 0.538 | 0.116184 |
| RM | 6.287877 | 6.21100 | 5.713 | 0.699960 |
| AGE | 68.543209 | 76.90000 | 100.0 | 28.012357 |
| DIS | 3.798790 | 3.26280 | 3.4952 | 2.095728 |
| RAD | 9.465753 | 5.00000 | 24.0 | 8.705323 |
| TAX | 407.185910 | 330.00000 | 666.0 | 168.043548 |
| PTRATIO | 18.511742 | 19.10000 | 20.2 | 2.275664 |
| B | 357.119491 | 391.70000 | 396.9 | 90.957762 |
| LSTAT | 12.574618 | 11.28000 | 6.36 | 7.149791 |
| Price | $22.642661 | $21.20000 | $50.0 | $9.231178 |

## 1.5 Bi-variate Analysis:

### 1.5.1 Graphical Representations:

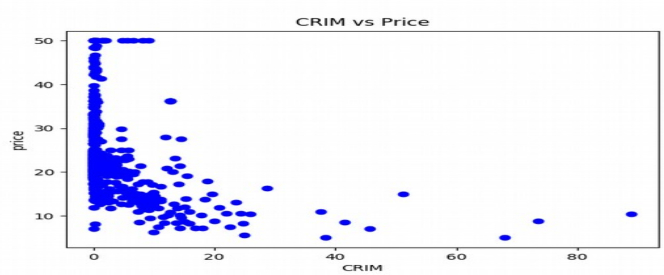



Figure 12: Scatter diagram of CRIM and Price.

There are negative weak and non linear relationship between CRIM and Price which appears in figure 12. Outliers are also in here.

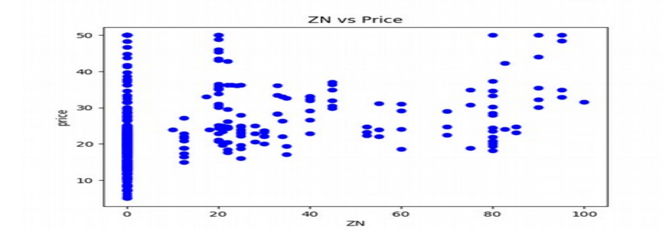



Figure 13: Scatter diagram of ZN and Price.

In figure 13, it show that there is non linear correlation between Zn(Proportion of a town’s residential land zoned) and house price. We found also outlines in here.

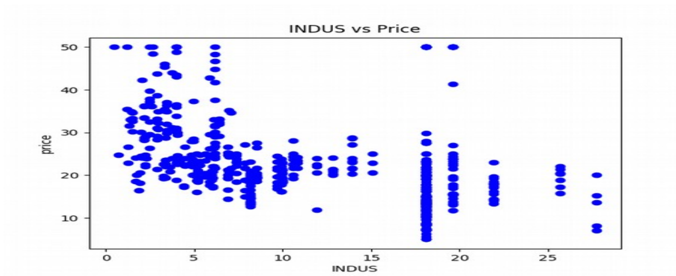



Figure 13: Scatter diagram of INDUS and Price.

Figure 13 shows that there is a very week positive non linear correlation between INDUS and price of house. We found some outliers here.

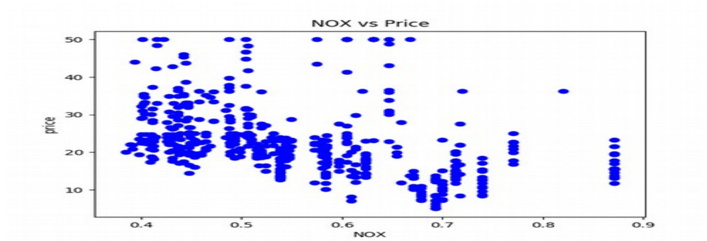



Figure 13: Scatter diagram of NOX and Price.

Figure 13 represents that there is a very week positive non linear correlation between NOX and price of house. We found some outliers here.

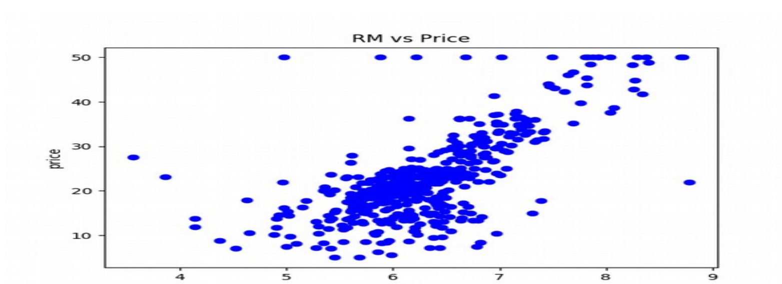



Figure 14: Scatter diagram of RM and Price.

Figure 14 explores that there is a very week positive linear correlation between RM and price of house. RM' is the average number of rooms among homes in the neighborhood. For a higher RM, one would expect to observe a higher price. This is because more rooms would imply more space, thereby costing more, taking all other factors constant. We found some outliers here.

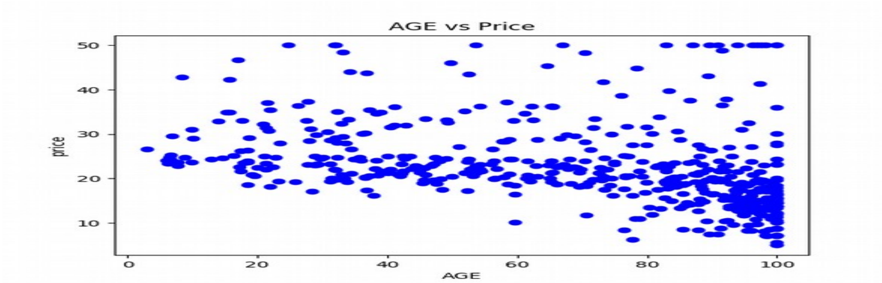



Figure 15: Scatter diagram of Age and Price.

Figure 15 illustrates that there is a very week positive linear correlation between AGE and price of house that means the prices is decreasing with aged house. We found many outliers here.

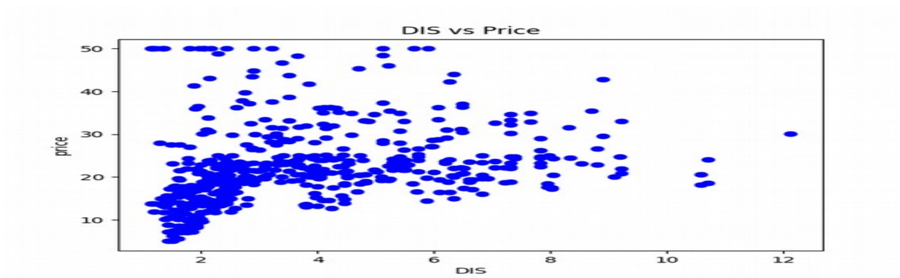



Figure 16: Scatter diagram of DIS and Price.

Figure 16 represents that there is a very week positive non linear correlation between DIS and price of house. We found many outliers here.


Figure 17: Scatter diagram of DIS and Price.

Figure 17 illustrates that there is a very week non linear correlation between TAX and price of house. We found many outliers here.

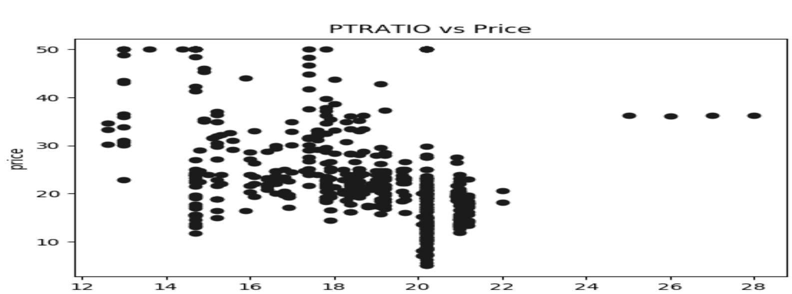


Figure 18: Scatter diagram of PTRATIO and Price.

Figure 18 shows that there is a very week non linear correlation between PTRATIO and price of house. We found many outliers here. 'PTRATIO' is the ratio of students to teachers in primary and secondary schools in the neighborhood. For a higher PTRATIO, one would expect to observe a a lower price.

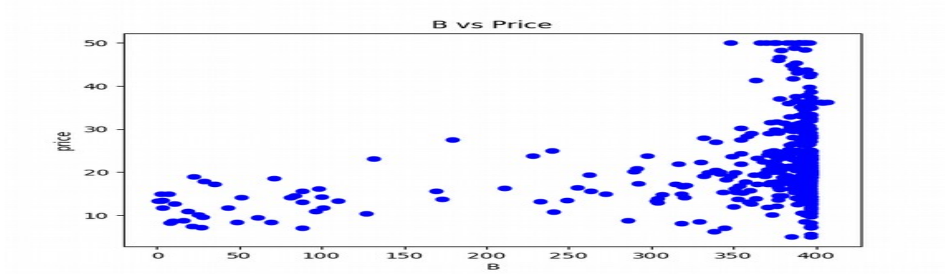



Figure 19: Scatter diagram of B and Price.

Figure 19 shows that there is a very week positive non linear correlation between B and price of house. We found many outliers here.

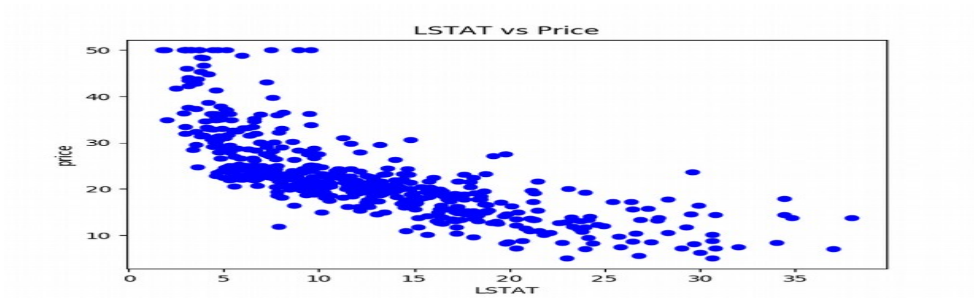



Figure 20: Scatter diagram of LSTAT and Price.

Figure 20 shows that there is a week negative linear correlation between LSTAT and price of house. We found many outliers here. 'LSTAT' is the percentage of homeowners in the neighborhood considered "lower class" (working poor). For a higher LSTAT, one would expect to observe a lower house price. The social milieux in an area dominated by "lower class" citizens may not be conducive for young children. It may also be relatively unsafe compared to an area dominated by "upper class" citizens. Hence an area with more "lower class" citizens would lower demand, hence lower prices.

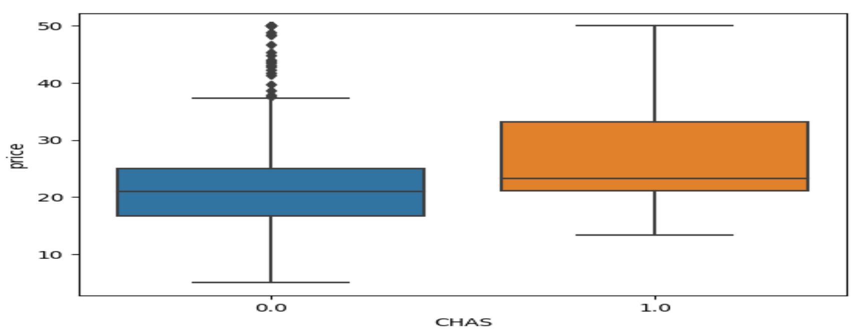



Figure 21: Scatter diagram of CHAS and Price.

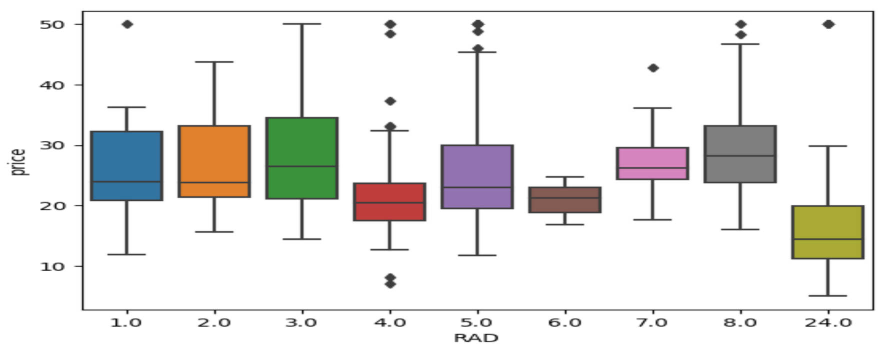



Figure 22: Scatter diagram of RAD and Price.

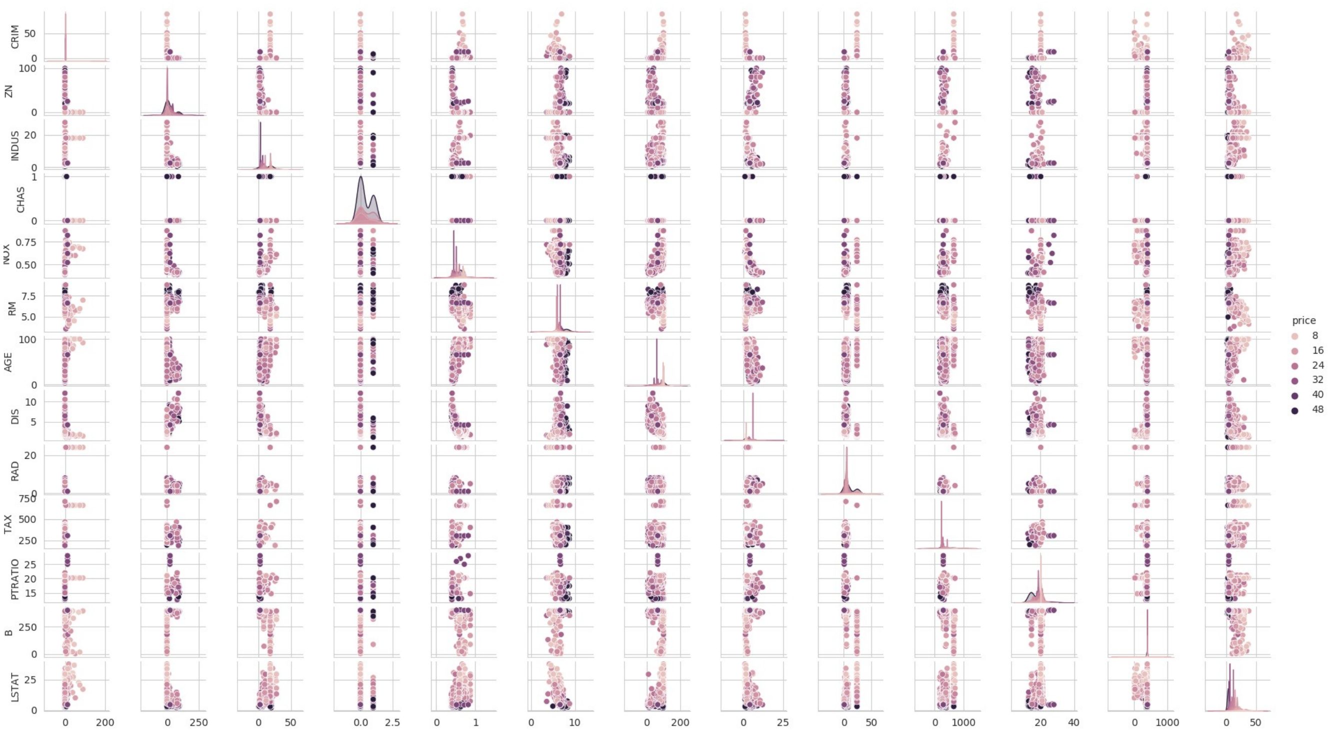



Figure 23: Pairplots of CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT at Price of house.

Figure 23 explores histograms for each variables and scatter diagrams between each pair of variables at price of house '8' with Rose Fog, '16' with careys Pink, '24' with Turkish Rose, '32' with Strikemaster, '40' with Voodoo and at price '48' with Bossanova dots.

### *1.5.2 Summary Measures:*

Using df.corr() commands we can make the following output.

`shamim@shamim:~/Dropbox/Data_Science_JU/First Semester/Introduction to Data Science with Python/Lab/Assignment_1$ python3 eda_analysis_boston_data.py`

| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CRIM | 1.000000 | -0.195016 | 0.392403 | -0.057640 | 0.424232 | -0.214038 | 0.350283 | -0.376314 | 0.612513 | 0.574447 | 0.303508 | -0.378936 | 0.442403 | -0.371215 |
| ZN | -0.195016 | 1.000000 | -0.534068 | -0.043714 | -0.507951 | 0.313350 | -0.569337 | 0.664394 | -0.314023 | -0.316045 | -0.355759 | 0.177176 | -0.414494 | 0.362964 |
| INDUS | 0.392403 | -0.534068 | 1.000000 | 0.065773 | 0.742893 | -0.393984 | 0.641170 | -0.704663 | 0.599764 | 0.721587 | 0.329399 | -0.359898 | 0.609163 | -0.490665 |
| CHAS | -0.057640 | -0.043714 | 0.065773 | 1.000000 | 0.088487 | 0.089857 | 0.086788 | -0.099608 | -0.004723 | -0.033805 | -0.121697 | 0.047382 | -0.050601 | 0.170462 |
| NOX | 0.424232 | -0.507951 | 0.742893 | 0.088487 | 1.000000 | -0.295603 | 0.725082 | -0.761936 | 0.596819 | 0.657174 | 0.206959 | -0.372959 | 0.574516 | -0.409069 |
| RM | -0.214038 | 0.313350 | -0.393984 | 0.089857 | -0.295603 | 1.000000 | -0.240507 | 0.205836 | -0.213138 | -0.294048 | -0.323102 | 0.130092 | -0.614531 | 0.694624 |
| AGE | 0.350283 | -0.569337 | 0.641170 | 0.086788 | 0.725082 | -0.240507 | 1.000000 | -0.747905 | 0.454955 | 0.506133 | 0.244780 | -0.273742 | 0.599866 | -0.375046 |
| DIS | -0.376314 | 0.664394 | -0.704663 | -0.099608 | -0.761936 | 0.205836 | -0.747905 | 1.000000 | -0.493923 | -0.534395 | -0.214721 | 0.292009 | -0.495859 | 0.250104 |
| RAD | 0.612513 | -0.314023 | 0.599764 | -0.004723 | 0.596819 | -0.213138 | 0.454955 | -0.493923 | 1.000000 | 0.910248 | 0.413826 | -0.446557 | 0.494093 | -0.388169 |
| TAX | 0.574447 | -0.316045 | 0.721587 | -0.033805 | 0.657174 | -0.294048 | 0.506133 | -0.534395 | 0.910248 | 1.000000 | 0.420387 | -0.443482 | 0.546535 | -0.470970 |
| PTRATIO | 0.303508 | -0.355759 | 0.329399 | -0.121697 | 0.206959 | -0.323102 | 0.244780 | -0.214721 | 0.413826 | 0.420387 | 1.000000 | -0.154639 | 0.324155 | -0.436289 |
| B | -0.378936 | 0.177176 | -0.359898 | 0.047382 | -0.372959 | 0.130092 | -0.273742 | 0.292009 | -0.446557 | -0.443482 | -0.154639 | 1.000000 | -0.368847 | 0.336308 |
| LSTAT | 0.442403 | -0.414494 | 0.609163 | -0.050601 | 0.574516 | -0.614531 | 0.599866 | -0.495859 | 0.494093 | 0.546535 | 0.324155 | -0.368847 | 1.000000 | -0.740179 |
| price | -0.371215 | 0.362964 | -0.490665 | 0.170462 | -0.409069 | 0.694624 | -0.375046 | 0.250104 | -0.388169 | -0.470970 | -0.436289 | 0.336308 | -0.740179 | 1.000000 |

Figure 14: Correlation Matrix of different variables of Boston Housing Dataset.

Observe that RAD and TAX are highly correlated with each other (Correlation score: 0.92) while there are a couple of features which are somewhat correlated with one another with a correlation score of around 0.70 (INDUS and TAX, NOX and INDUS, AGE and DIS, AGE and INDUS).

We observe that both RM and LSTAT are correlated with price with a correlation score of 0.66 and 0.74 respective.

Using df.cov() commands we can make the following output.

`shamim@shamim:~/Dropbox/Data_Science_JU/First Semester/Introduction to Data Science with Python/Lab/Assignment_1$ python3 eda_analysis_boston_data.py`

| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CRIM | 73.925755 | -38.955687 | 23.211568 | -0.125303 | 0.423788 | -1.288140 | 84.365722 | -6.780837 | 45.845642 | 829.984324 | 5.938492 | -296.348645 | 27.196255 | -29.463250 |
| ZN | -38.955687 | 539.764234 | -85.363616 | -0.256782 | -1.371106 | 5.095719 | -370.528172 | 32.349134 | -63.510825 | -1233.878073 | -18.808992 | 374.408801 | -68.851483 | 77.843647 |
| INDUS | 23.211568 | -85.363616 | 47.331329 | 0.114409 | 0.593810 | -1.897258 | 123.565580 | -10.159928 | 35.920266 | 834.228327 | 5.157084 | -225.213047 | 29.964101 | -31.161397 |
| CHAS | -0.125303 | -0.256782 | 0.114409 | 0.063927 | 0.002599 | 0.015903 | 0.614682 | -0.052780 | -0.010395 | -1.436288 | -0.070021 | 1.089662 | -0.091474 | 0.397857 |
| NOX | 0.423788 | -1.371106 | 0.593810 | 0.002599 | 0.013499 | -0.024040 | 2.359848 | -0.185524 | 0.603635 | 12.830679 | 0.054719 | -3.941374 | 0.477246 | -0.438733 |
| RM | -1.288140 | 5.095719 | -1.897258 | 0.015903 | -0.024040 | 0.489945 | -4.715752 | 0.301946 | -1.298731 | -34.587036 | -0.514661 | 8.282574 | -3.075462 | 4.488284 |
| AGE | 84.365722 | -370.528172 | 123.565580 | 0.614682 | 2.359848 | -4.715752 | 784.692164 | -43.906724 | 110.943875 | 2382.516892 | 15.603903 | -697.477840 | 120.142682 | -96.982047 |
| DIS | -6.780837 | 32.349134 | -10.159928 | -0.052780 | -0.185524 | 0.301946 | -43.906724 | 4.392077 | -9.011132 | -188.199916 | -1.024044 | 55.663551 | -7.429956 | 4.838514 |
| RAD | 45.845642 | -63.510825 | 35.920266 | -0.010395 | 0.603635 | -1.298731 | 110.943875 | -9.011132 | 75.782648 | 1331.577948 | 8.198050 | -353.591272 | 30.752943 | -31.193438 |
| TAX | 829.984324 | -1233.878073 | 834.228327 | -1.436288 | 12.830679 | -34.587036 | 2382.516892 | -188.199916 | 1331.577948 | 28238.633997 | 160.760558 | -6778.567788 | 656.649179 | -730.587555 |
| PTRATIO | 5.938492 | -18.808992 | 5.157084 | -0.070021 | 0.054719 | -0.514661 | 15.603903 | -1.024044 | 8.198050 | 160.760558 | 5.178646 | -32.008714 | 5.274169 | -9.165149 |
| B | -296.348645 | 374.408801 | -225.213047 | 1.089662 | -3.941374 | 8.282574 | -697.477840 | 55.663551 | -353.591272 | -6778.567788 | -32.008714 | 8273.314435 | -239.871839 | 282.379694 |
| LSTAT | 27.196255 | -68.851483 | 29.964101 | -0.091474 | 0.477246 | -3.075462 | 120.142682 | -7.429956 | 30.752943 | 656.649179 | 5.274169 | -239.871839 | 51.119515 | -48.852568 |
| price | -29.463250 | 77.843647 | -31.161397 | 0.397857 | -0.438733 | 4.488284 | -96.982047 | 4.838514 | -31.193438 | -730.587555 | -9.165149 | 282.379694 | -48.852568 | 85.214647 |

Figure 15: Co-variance Matrix of different variables of Boston Housing Dataset.

## *1.6 Discussion:*

From the EDA, it is exhibited that Price of house has an impact on CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT. The features 'RM', 'LSTAT', 'PTRATIO', and 'MEDV' are more essential. The survival status is a variable with . So we could use a Linear Regressionn model to predict the price of house, whether price based upon CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B and LSTAT.