**Case Problem**

**Predictive Analytics**

**Module 1 – Lesson 2**

**Professor: Lee**

**Rajiv Kumar**

**#1: Thought Question: A dataset has 1000 records and 50 variables with 5% of the values missing, spread randomly throughout the records and variables. An analyst decides to remove records that have missing values. What percentage of records would you expect would be removed?**

**To be discussed**

**#2: Perform a Data Audit on the Boston Housing Data. Follow these guidelines about assessments to make and graphs/cross tabs to create. List each variable and perform the assessment. Note that you won’t do the final part for your target variables! Note also that you have a description of the variables in the excel file. List that information as part of your audit.**

**Continuous Variable Assessment**

* Good summary table for continuous variables contains: Variable Name, Minimum, Maximum, Mean, Standard Deviation, # Missing

1. **CRIM-per capita crime rate by town**

* ***Statistics***
  + Count 506
  + Mean 3.614
  + Min 0.006
  + Max 88.976
  + Range 88.970
  + Variance 73.987
  + Standard Deviation 8.602
  + Standard Error of Mean 0.382
  + Skeweness 5.223
  + #Missing 0

1. **ZN-  proportion of residential land zoned for lots over 25,000 sq.ft.**

* **Statistics**
  + Count 506
  + Mean 11.364
  + Min 0.000
  + Max 100.000
  + Range 100.000
  + Variance 543.937
  + Standard Deviation 23.322
  + Standard Error of Mean 1.037
  + Skeweness 2.226
  + #Missing 0

1. **INDUS-proportion of non-retail business acres per town.**

* **Statistics**
  + Count 506
  + Mean 11.137
  + Min 0.460
  + Max 27.740
  + Range 27.280
  + Variance 47.064
  + Standard Deviation 6.860
  + Standard Error of Mean 0.305
  + Skeweness .295
  + #Missing 0

1. **NOX-nitric oxides concentration (parts per 10 million)**

* **Statistics**
  + Count 506
  + Mean 0.555
  + Min 0.385
  + Max 0.871
  + Range 0.486
  + Variance 0.013
  + Standard Deviation 0.116
  + Standard Error of Mean 0.005
  + Skewness .729
  + #Missing 0

1. **RM-average number of rooms per dwelling**

* **Statistics**
  + Count 506
  + Mean 6.285
  + Min 3.561
  + Max 8.780
  + Range 5.219
  + Variance 0.494
  + Standard Deviation 0.703
  + Standard Error of Mean 0.031
  + Skewness .404
  + #Missing 0

1. **AGE-proportion of owner-occupied units built prior to 1940**

* **Statistics**
  + Count 506
  + Mean 68.575
  + Min 2.900
  + Max 100.000
  + Range 97.100
  + Variance 792.358
  + Standard Deviation 28.149
  + Standard Error of Mean 1.251
  + Skewness -0.599

1. **DIS-weighted distances to five Boston employment centres**

* **Statistics**
  + Count 506
  + Mean 3.795
  + Min 1.130
  + Max 12.126
  + Range 10.997
  + Variance 4.434
  + Standard Deviation 2.106
  + Standard Error of Mean 0.094
  + Skewness 1.012
  + #Missing Values 0

1. **TAX-full-value property-tax rate per $10,000**

* **Statistics**
  + Count 506
  + Mean 408.237
  + Min 187.000
  + Max 711.000
  + Range 524.000
  + Variance 28404.759
  + Standard Deviation 168.537
  + Standard Error of Mean 7.492
  + Skeweness .670
  + #Missing Values

1. **PTRATIO-pupil-teacher ratio by town**

* **Statistics**
  + Count 506
  + Mean 18.456
  + Min 12.600
  + Max 22.000
  + Range 9.400
  + Variance 4.687
  + Standard Deviation 2.165
  + Standard Error of Mean 0.096
  + Skewness -0.802
  + #Missing Values 0

1. **B-1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town**

* **Statistics**
  + Count 506
  + Mean 356.674
  + Min 0.320
  + Max 396.900
  + Range 396.580
  + Variance 8334.752
  + Standard Deviation 91.295
  + Standard Error of Mean 4.059
  + Skewness -2.89
  + #Missing Values 0

1. **LSTAT-% lower status of the population**

* **Statistics**
  + Count 506
  + Mean 12.653
  + Min 1.730
  + Max 37.970
  + Range 36.240
  + Variance 50.995
  + Standard Deviation 7.141
  + Standard Error of Mean 0.317
  + Skewness .906
  + #Missing Values 0

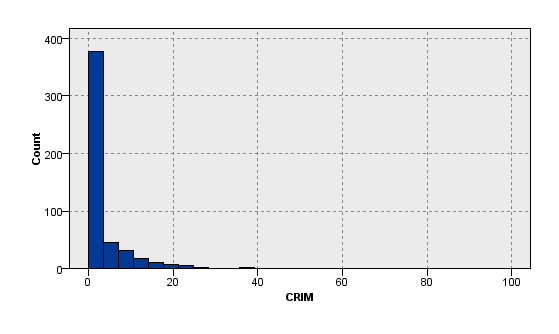
1. **MEDV-Median value of owner-occupied homes in $1000**

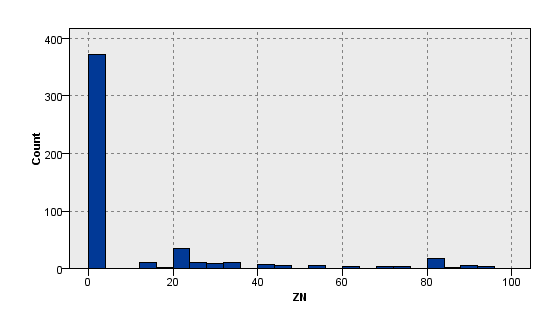
* **Statistics**
  + Count 506
  + Mean 22.533
  + Min 5.000
  + Max 50.000
  + Range 45.000
  + Variance 84.587
  + Standard Deviation 9.197
  + Standard Error of Mean 0.409
  + Skeweness 1.108
  + #Missing Values 0
* Do the min and max values make sense? Are there any unexpected values?
* Are data normally distributed?
  + Answer :

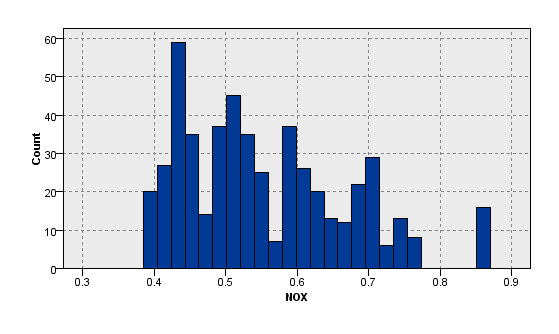
*Data is highly skewed for CRIM (Skewness: 5.223),ZN(skewness:2.226) ,B(Skewness:-2.89). None of the variables follows a perfect normal distribution curve. However RM , Indus and Age are very close to normal distribution.*

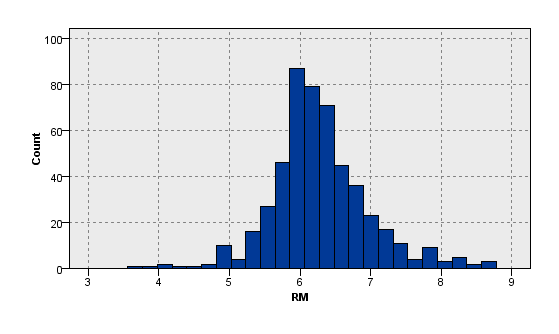
* Are there missing values? How many? How are they coded?

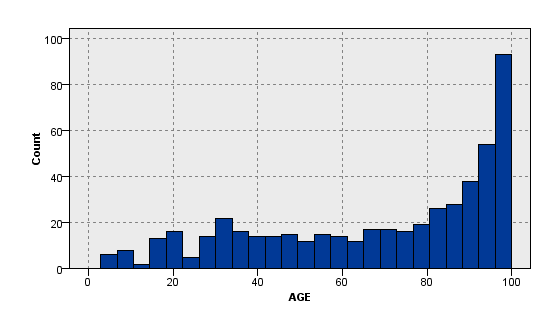
Answer: There are not any missing values

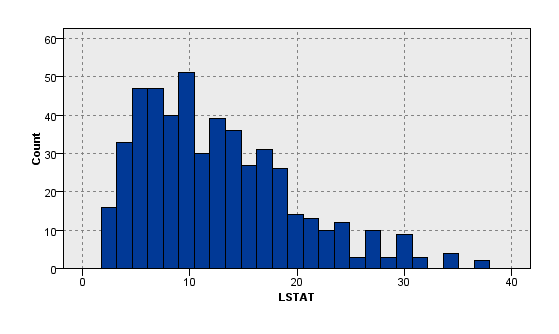
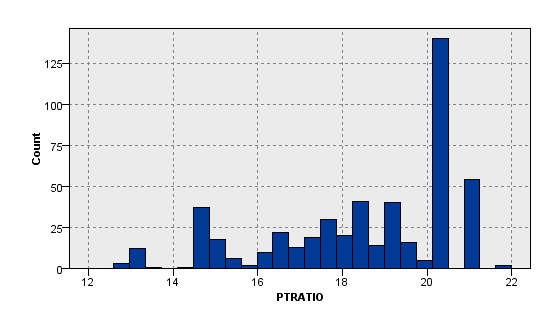
* **Graphs: Histograms**
* 

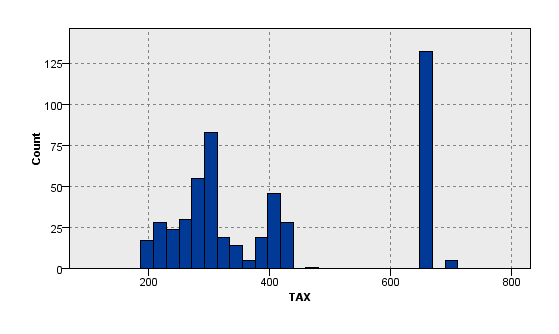


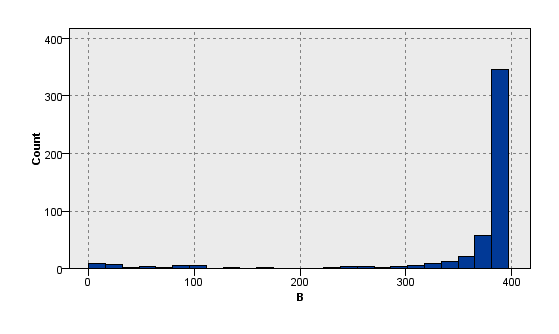


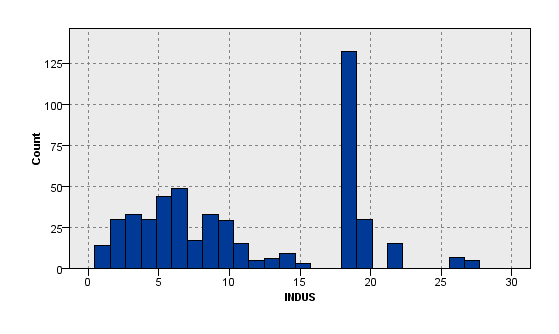


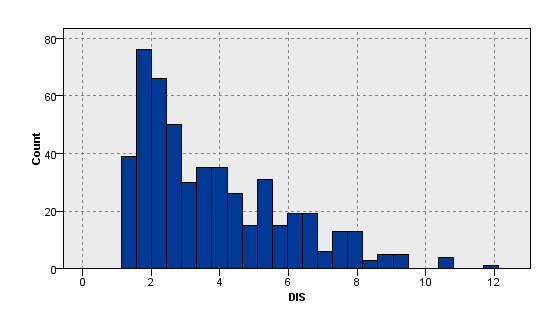


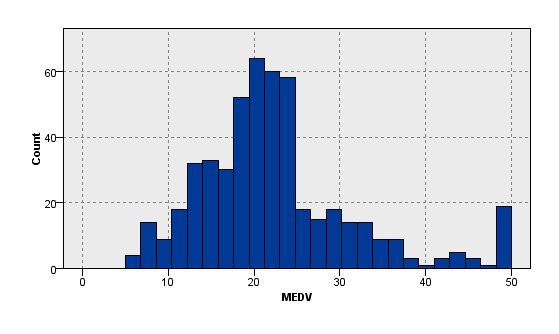
  






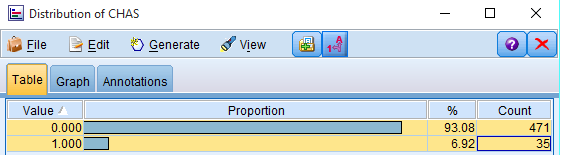


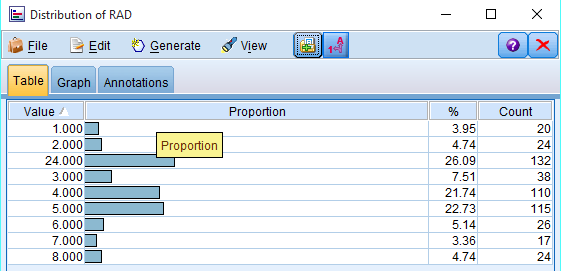


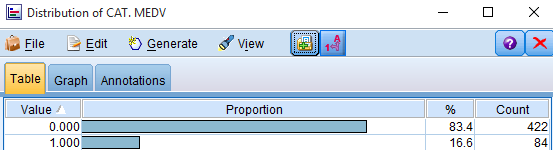


**Categorical Variable Assessment**

* Good summary table for categorical variables contains: Variable Name, # of Levels, Mode, N (Sample Size), # Missing, % Missing





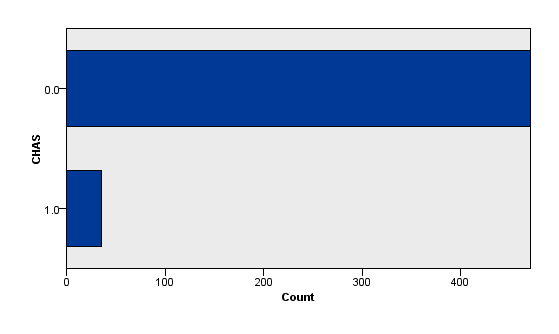
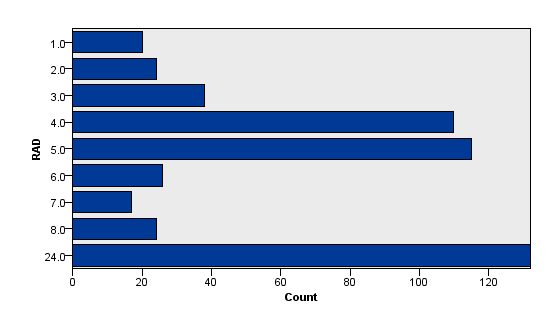
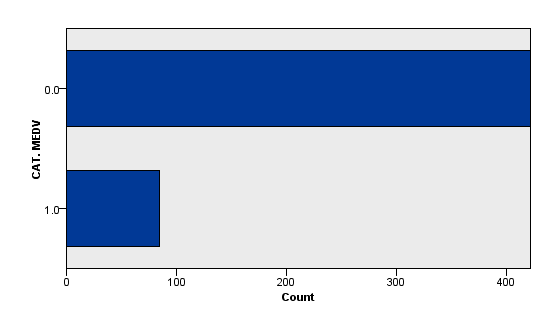


* Do the values make sense? Are there values that are unexpected, erroneous, or indicators of missing or unknown values

Value 24 For one of the variable RAD seem little high as compare to other values in the column. It would need additional analysis to confirm it is correct.

* Are there missing values? How many? How are they coded?

Answer:

* + There are no missing values
* How many levels are there? Is there only one value? Are there more than 50 or 100 levels? What is the mode of the levels?
  + CAT.MEDV
    - Levels=2 (0,1)
    - Only One Value? = No
    - Are there more than 50 or 100 levels? No
    - Mode =0
  + CHAS
    - Levels=2 (0,1)
    - Only One Value? = No
    - Are there more than 50 or 100 levels? No
    - Mode =0
  + RAD
    - Levels=9 (1,2,3,4,5,6,7,8,24)
    - Only One Value? = No
    - Are there more than 50 or 100 levels? No
    - Mode =24
* Bar/Column Plot (Distribution)
* 
* 
* 

**Use multiple variable summaries to help see how each variable is related to your target variable.**

* Does it appear that the variable will be predictive of the target variables?

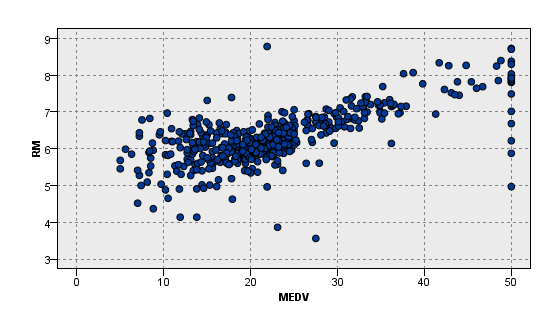
Answer :

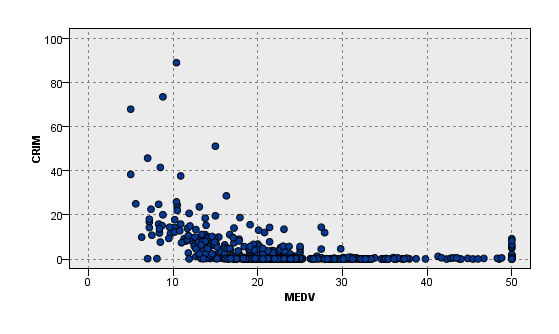
* + From the matrix , it appears that variable RM,LSTAT,TAX are the most important predictor variable for our Target variable MEDV.

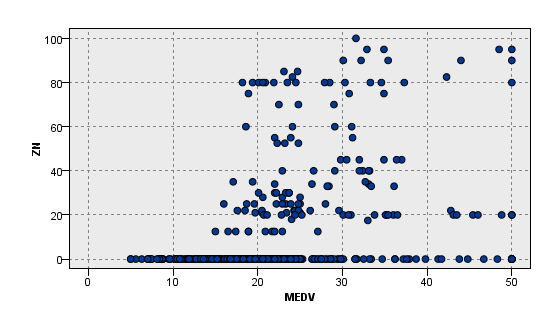
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nodes** | **Importance** | **Importance** | **V4** | **V5** |
| DIS | 0 | 0 | DIS | 0 |
| RAD | 0 | 0 | RAD | 0 |
| INDUS | 0.0209 | 0.0209 | INDUS | 0.0209 |
| ZN | 0.0223 | 0.0223 | ZN | 0.0223 |
| AGE | 0.0259 | 0.0259 | AGE | 0.0259 |
| CRIM | 0.0438 | 0.0438 | CRIM | 0.0438 |
| B | 0.0645 | 0.0645 | B | 0.0645 |
| PTRATIO | 0.0658 | 0.0658 | PTRATIO | 0.0658 |
| CHAS | 0.0667 | 0.0667 | CHAS | 0.0667 |
| NOX | 0.0787 | 0.0787 | NOX | 0.0787 |
| TAX | 0.1186 | 0.1186 | TAX | 0.1186 |
| LSTAT | 0.2201 | 0.2201 | LSTAT | 0.2201 |
| RM | 0.2727 | 0.2727 | RM | 0.2727 |

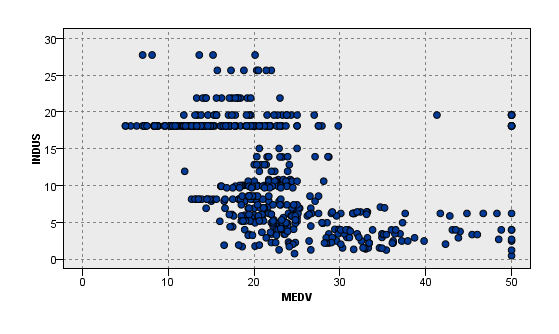
* Scatterplots (Plot), Cross Tabs (Matrix) and Overlays (Collection & Multiplot)

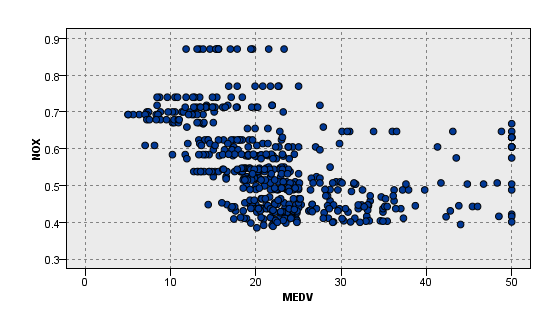
**Scatterplots:**

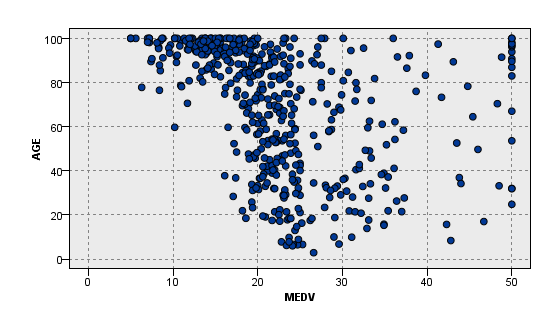
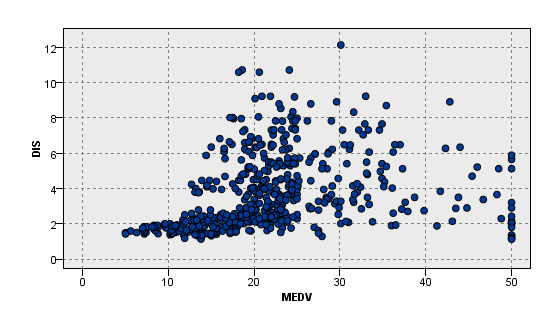


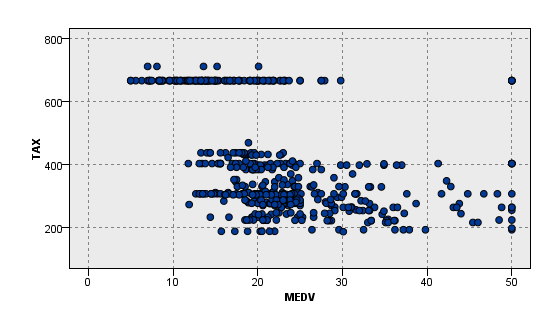


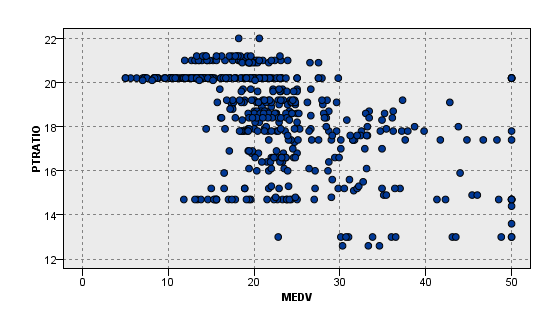


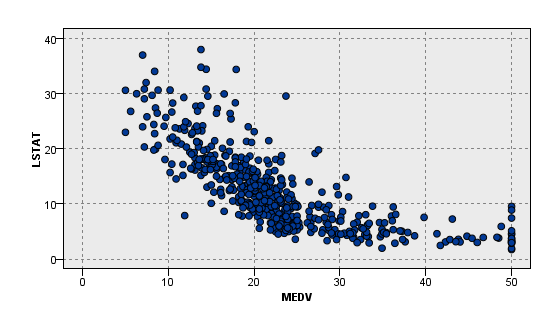


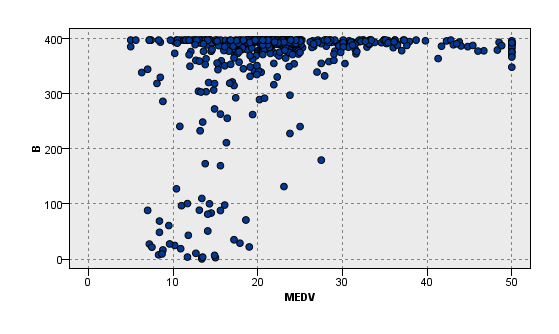










**CROSS TABS:**

|  |  |  |
| --- | --- | --- |
| **CAT.MEDV** | | |
| **CHAS** | 0.0 | 1.0 |
| 0.0 | 398 | 73 |
| 1.0 | 24 | 11 |

**RAD**

**CHAS** 1.0 2.0 24.0 3.0 4.0 5.0 6.0 7.0 8.0

|  |
| --- |
|  |

0.0 19 24 124 36 102 104 26 17 19

|  |
| --- |
| 1.0 1 0 8 2 8 11 0 0 5  **Overlays:** |
|  |