Assignment3

Sangamesh

21/09/2018

Perform the CRISP-DM analysis as per instructions (Data Preprocessing, Exploratory Data Analysis,and if required, Dimension-Reduction Methods)

CRISP-DM: The six Phases

Cross-Industry standard process for Data Mining.

1. Business Understanding
2. Data Understanding
3. Data Prepration
4. Modeling Phase
5. Evaluation Phase
6. Deployment Phase

The objective of this assignment is to perform first two phases of CRISP-DM.

Exercise 1

Business Understanding

- Data Mining Goal: For banks, Whenever an applicant applies for a loan, predicting whether the loan will be repaid is an important activity for any bank. High accuracy is beneficial for both the banks and the loan applicants. Our goal here is to extarct various parameters which can most accurately determine if an applicant would repay the loan or not.
- Data Mining Success Criteria: The sucess of data mining solely depends on the fact that whether we are able to identify feature which can accuratily predict an applicant defaulter or not.

Data Understanding

- Data Exploration Report : Following is just the Header Part of the previous bank record data.

```
## # A tibble: 6 x 9
##     age    ed employ address income debtinc creddebt othdebt default
##   <int> <int>  <int>   <int>  <dbl>   <dbl>    <dbl>   <dbl>   <int>
## 1    41     3     17      12    176     9.3   11.4     5.01        1
## 2    27     1     10       6     31    17.3    1.36    4.00        0
## 3    40     1     15      14     55     5.5    0.856   2.17        0
## 4    41     1     15      14    120     2.9    2.66    0.821       0
## 5    24     2      2       0     28    17.3    1.79    3.06        1
## 6    41     2      5       5     25    10.2    0.393   2.16        0
```

Size of Data is:

```
## [1] 850
```

We are dividing the data into two halves Traning and Test sets.

```
bankDetailsTrain <- bankDetails[1:700,1:9]
bankDetailsTest <- bankDetails[701:850,1:9]
```

Summery of the Traning data:

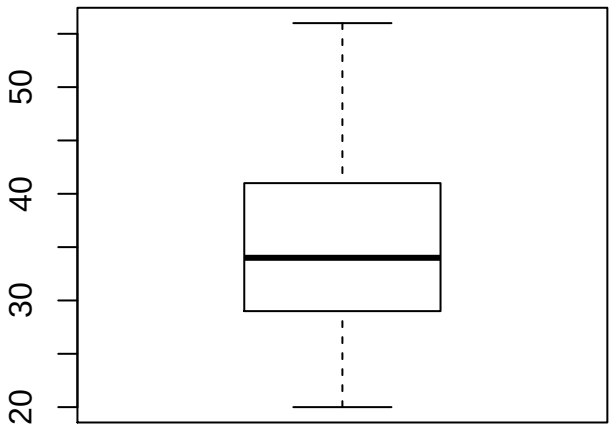
```
##       age              ed           employ          address
##  Min.   :20.00   Min.   :1.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:29.00   1st Qu.:1.000   1st Qu.: 3.000   1st Qu.: 3.000
##  Median :34.00   Median :1.000   Median : 7.000   Median : 7.000
##  Mean   :35.03   Mean   :1.711   Mean   : 8.566   Mean   : 8.372
##  3rd Qu.:41.00   3rd Qu.:2.000   3rd Qu.:13.000   3rd Qu.:12.000
##  Max.   :56.00   Max.   :5.000   Max.   :33.000   Max.   :34.000
##
##      income          debtinc         creddebt          othdebt
##  Min.   : 13.00   Min.   : 0.10   Min.   : 0.0117   Min.   : 0.04558
##  1st Qu.: 24.00   1st Qu.: 5.10   1st Qu.: 0.3822   1st Qu.: 1.04594
##  Median : 35.00   Median : 8.70   Median : 0.8851   Median : 2.00324
##  Mean   : 46.68   Mean   :10.17   Mean   : 1.5768   Mean   : 3.07879
##  3rd Qu.: 55.75   3rd Qu.:13.80   3rd Qu.: 1.8984   3rd Qu.: 3.90300
##  Max.   :446.00   Max.   :41.30   Max.   :20.5613   Max.   :35.19750
##
##     default
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2614
##  3rd Qu.:1.0000
##  Max.   :1.0000
##  NA's   :150
```

- Methods of identifing outliers:

1. Boxplot: Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution. The spacings between the different parts of the box indicate the degree of dispersion (spread) and skewness in the data, and show outliers.

Boxplot of age: Age in years of the loan applicant

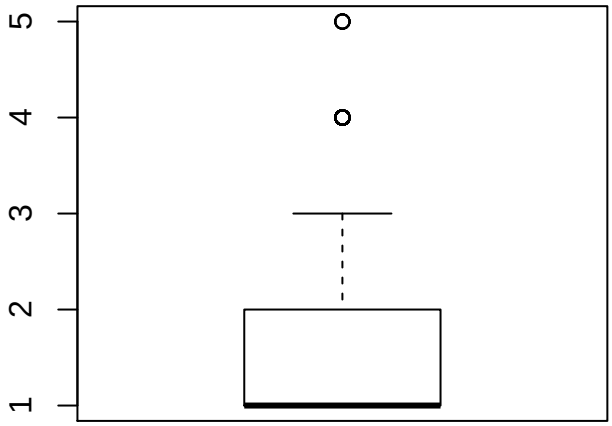
```
boxplot(bankDetails$age)
```



As the plot shows that the range of people who have taken loan are in the range of 20 to 55. Mean is also around 35. Age factore doesn't shows any outlier in data.

Boxplot of ed: Level of education of a loan application

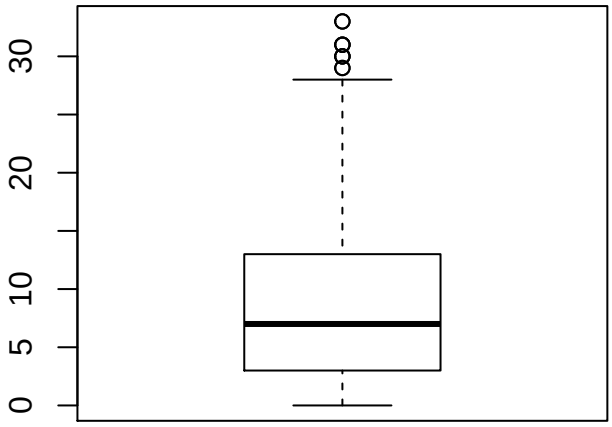
```
boxplot(bankDetails$ed)
```



Here We can observe that major population is not so educated about Loan. There are only 54 out of 850 people with higher level of understanding of Loan.

Boxplot of employ: Years with current employer

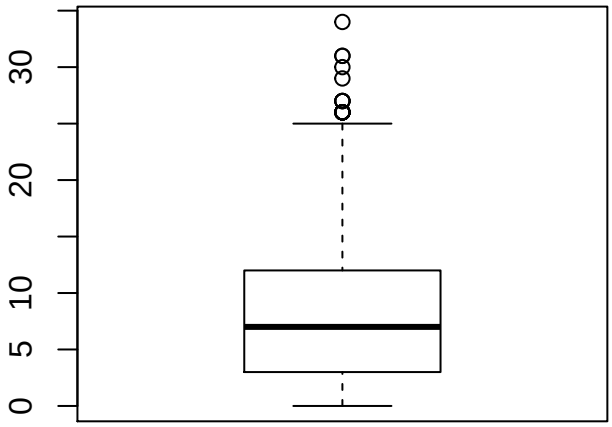
```
boxplot(bankDetails$employ)
```



Only 3 distinct people are employed with the same employer for more the 30 years and 10 are employed for more them 25 years. Major population has an employment history of around 8-10 years.

Boxplot of address: Years at current address

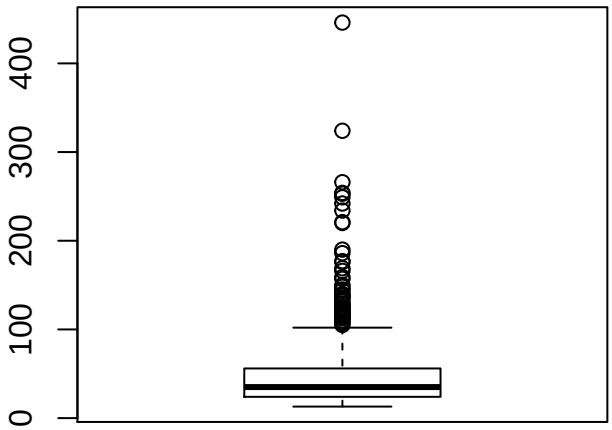
```
boxplot(bankDetails$address)
```



Major population is residing at there current address for 10-15 years. There are only 3 people who have lived at there current address for more then 30 year. only 15 records show someone residing at the current address for more then 25 years.

Boxplot of income: Household income in thousands

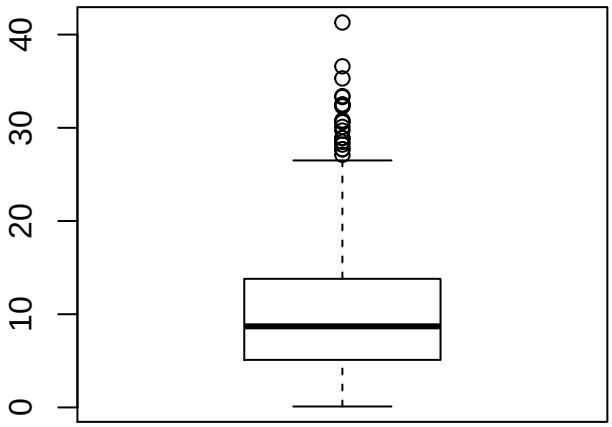
```
boxplot(bankDetails$income)
```



The point summary shows that the third point is at 55, which means that max should have been around 100 for the distribution to be uniform but Data shows that there are 58 recordes which shows income greater then 100. These point are the outliers.

Boxplot of debtinc: Debt to income ratio in thousands

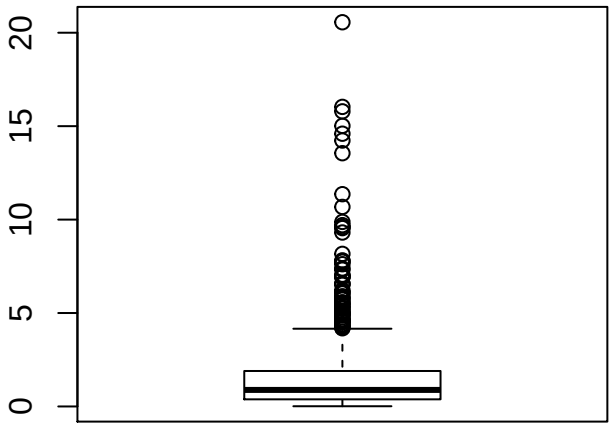
```
boxplot(bankDetails$debtinc)
```



We have 32 people who are having debt to income ratio greater then 25. It's not a surprise that most of these people are dafaulters.

Boxplot of creddebt: Credit card debt in thousands

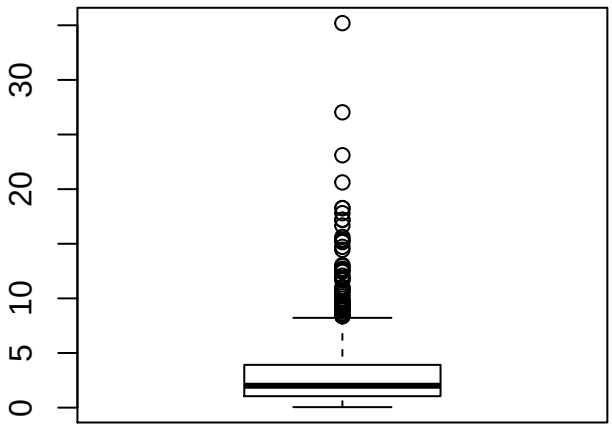
```
boxplot(bankDetails$creddebt)
```



There is a discrete number of population having dept greater then 4 or 5, The number is just 76. Those having Credit card debt greater then 10 are all defaulters and they are significantly less in number.

Boxplot of othdebt :Other debt in thousands

```
boxplot(bankDetails$othdebt)
```



This shows highest number of outliers as people having any kind of debt greater then 10,000 are 147 in number also there default status is equalily likely. but those whose debt is greater then 20 are mostly defaulters.