4/2/2019

Employee Absenteeism

Gohel Ravi Vashrambhai

Contents Page No.

1.Introduction\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2

1.1 Problem Statement\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2

1.2 Data\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2

1.3 Variable Description\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_3

2.Methodology\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_5

2.1Pre Processing\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_5

2.1.1Missing Value Analysis\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_5

2.1.2Outlier Analysis\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_7

2.1.3Feature Selection\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10

2.1.4Feature scaling\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_12

2.1.5 Dataset Sampling\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14

2.1.6 Principal Component Analysis\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_15

2.2. Modeling\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17

2.2.1. Decision tree for Regression\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17

2.2.2. Random Forest\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17

2.2.5 Linear Regression\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17

3. Conclusion \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_18

4. Question 1 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_19

5.UQuestion 2\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_22

**1.Introduction**

1.1 Problem Statement

XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The company has shared it dataset and requested to have an answer on the following areas:

1. What changes company should bring to reduce the number of absenteeism?

2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

1.2 Data



Figure:1 :: Column(1 to 7) of the dataset



Figure:2 :: Column(8 to 14) of the dataset



Figure:3 :: Column(15 to 21) of the dataset

**1.2.1 Variable Description**

1. Individual identification (ID)

2. Reason for absence (ICD). Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioral disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities

XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX Injury, poisoning and certain other consequences of external causes

XX External causes of morbidity and mortality

XXI Factors influencing health status and contact with health services.

And 7 categories without (CID)

patient follow-up (22),

medical consultation (23),

blood donation (24),

laboratory examination (25),

unjustified absence (26),

physiotherapy (27),

dental consultation (28).

3. Month of absence

4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons (summer (1), autumn (2), winter (3), spring (4))

6. Transportation expense

7. Distance from Residence to Work (kilometers)

8. Service time

9. Age

10. Work load Average/day

11. Hit target

12. Disciplinary failure (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

14. Son (number of children)

15. Social drinker (yes=1; no=0)

16. Social smoker (yes=1; no=0)

17. Pet (number of pet)

18. Weight

19. Height

20. Body mass index

21. Absenteeism time in hours (target)

**2.Methodology**

**2.1: Pre Processing**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process, we will first try and look at all the probability distributions of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable.

**2.1.1: Missing Value Analysis**

Following table shows the percentage of missing value of each variable.



Figure: 4:: Variable wise missing value percentage

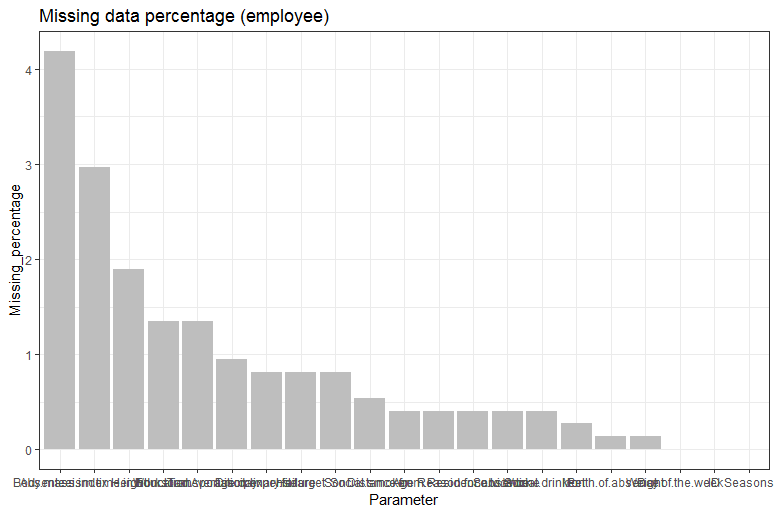


Figure:5:: Bar graph of the variable wise missing value percentage

Here we are using simple technique to impute missing value, very first we will create the missing value at some data point, then we will impute that missing value using various imputation method and we will compare that imputed value with the actual value. We will select the imputation method which perform well or whose result is much closer to the actual value.

Our actual value is :24

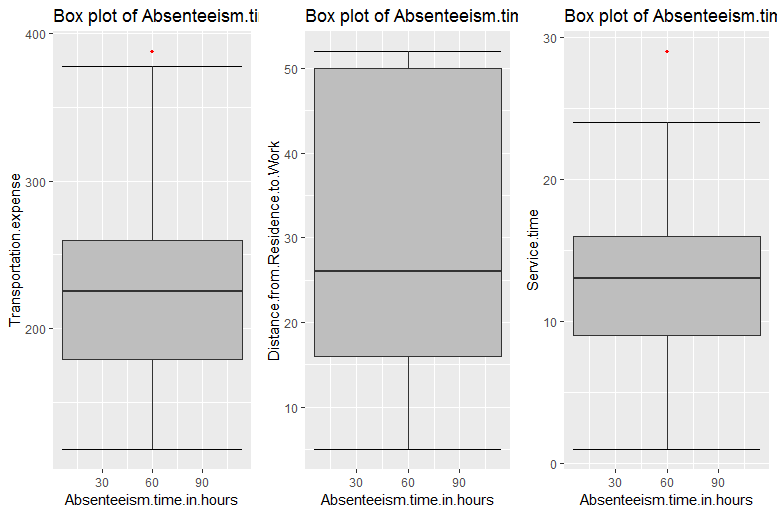
|  |  |  |
| --- | --- | --- |
| No. | Method | Result |
| 1. | Imputation by mean value | 26.68785 |
| 2. | Imputation by Median Value | 25 |
| 3. | KNN imputation Method | 24.32342 |

Figure:6:: results of imputation methods

From the above table we can compare the results of all the imputation method, we can clearly see that here KNN imputation method is performing well. We are using KNN imputation method to impute the missing values.

**2.1.2. Outlier Analysis:**

Here we are using a simple method to detect outliers which is box plot method. After detecting them we will replace them using appropriate method. Following are the box plot graphs of continuous variables.

Figure:7:: Box Plot of’ Transportation expense’ ,’Distance from Residence to Work (kilometers)’ ,’Service time’ variables.

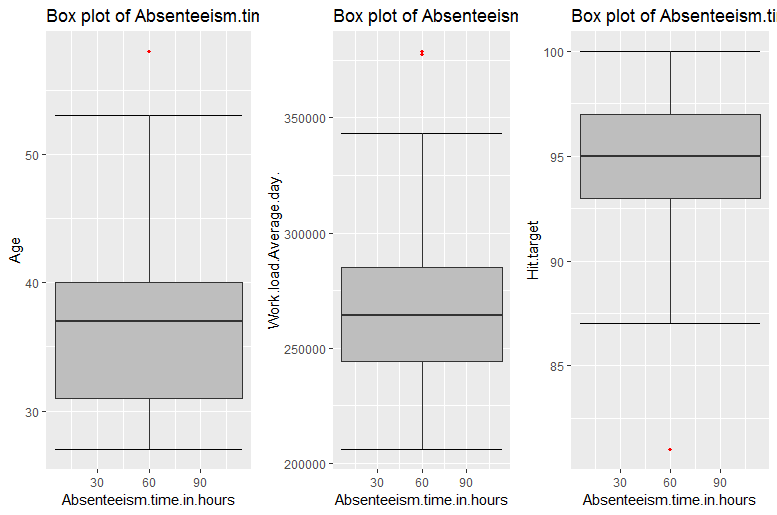


Figure: 8:: Boxplot of ‘Age’,’Work.load.Average.day ‘and ‘ hit,tarrget’ variables.

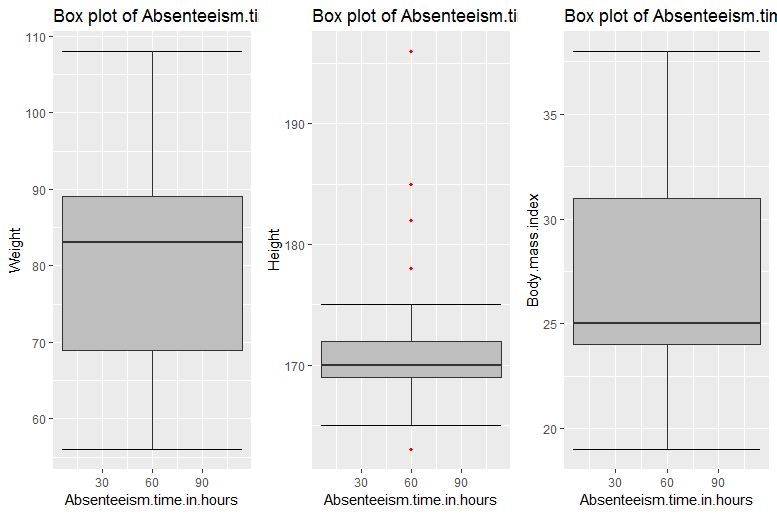


Figure:9:: Box plot of ‘Weight’,’Height’, and ‘Body.mass.index’ variables.

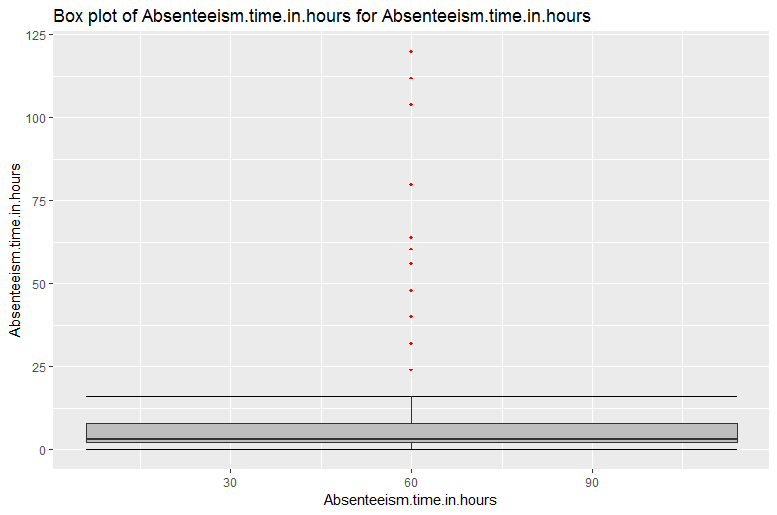


Figure: 10:: Box plot of ‘’Absenteeism.time.hours.

In above figures, the red dots indicate the outliers of the particular variable. Here, we detect that outliers and assign them “NA” values, again we are creating missing values in the data set then we impute that missing value with KNN imputation method. thus we detect and remove the outliers from the dataset.

**2.1.3 Feature Selection:**

If the two or more features of the data set carry the same information and If we use both of them to develop a model, then it will consume the more time and memory so here feature selection must be needed. Thumb rule for feature selection is: there must be high co-relation between dependent and independent variable and there must be low co relation between two independent variables. Following are the co relation graph for the all continuous variables.

Using chi-square test we can check the co relation between categorical variable.

By both the test we will drop some variables for further analysis.

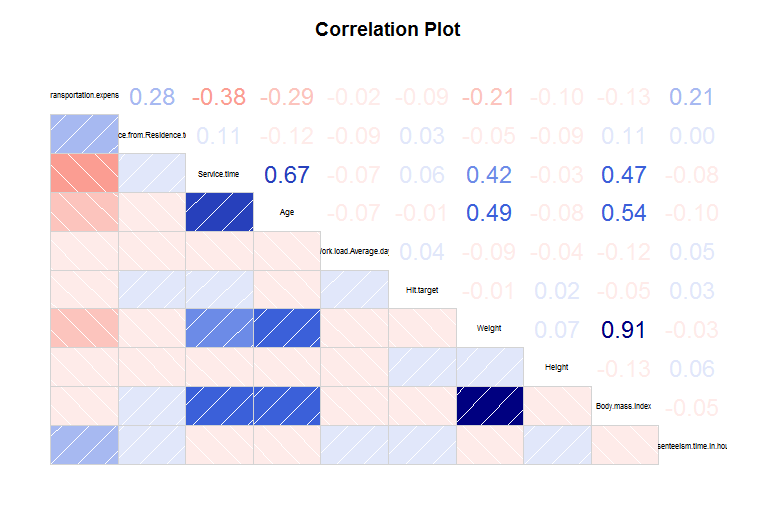


Figure:11:: Correlation Graph of the continuous variables.(R)

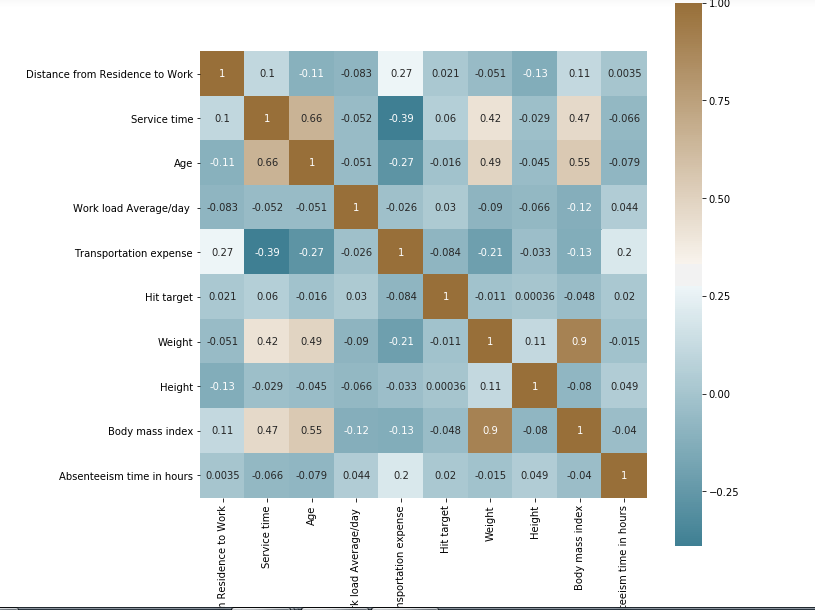


Figure: 12:: Correlation plot of continuous variables of then dataset.(python)

From the above two correlation plots we can clearly notice that “weight” variable have collinearity issue with the variable ”Body.mass.index”. In simple words we can say that they both variables are carrying almost the same information. So we can drop one of them variable for further analysis. Here, we are dropping or deleting “weight” variable for our further analysis.

**2.1.4Feature Scaling**

Data distribution also affects the performance the model. We can use normalization or standardization method to scale the data.by plotting histogram graph of the data we can see the data distribution. following are the histogram of the all continuous data.

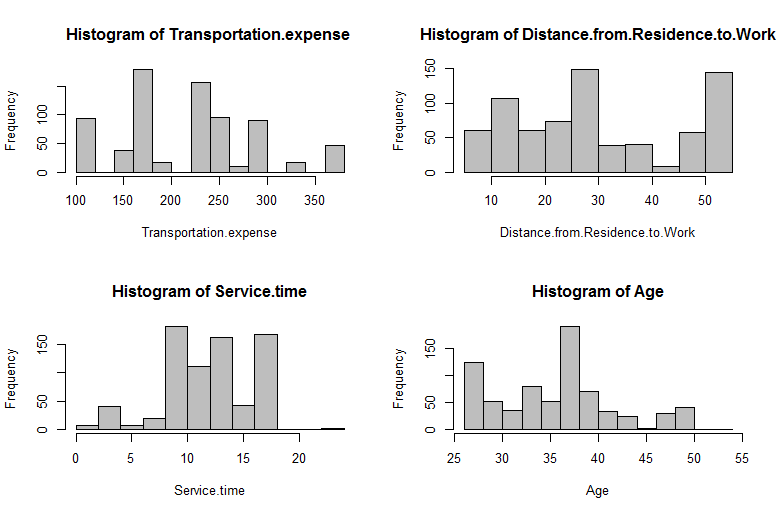


Figure:13:: Histogram of “Transportaion.expense”,”Distance.from.Residence.to.Work”,”Service.Time”and “Age” variables.

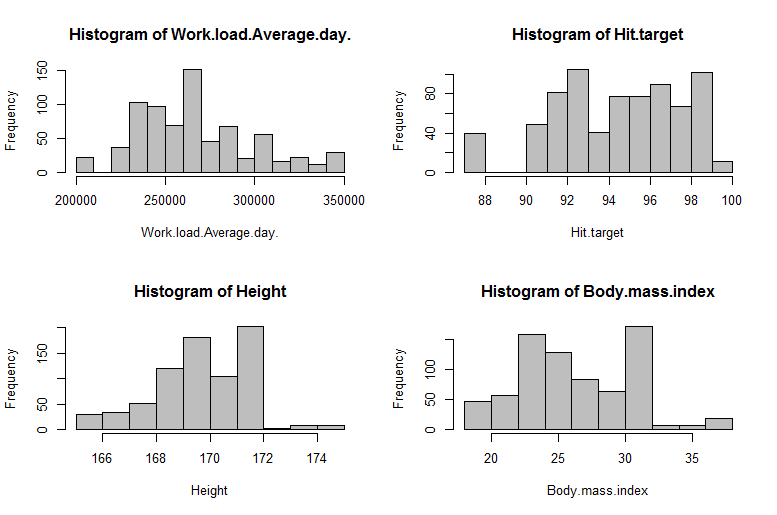


Figure:14:: Histogram of “Work.load.Average.day”,”Hit.target”, “Height” and “Body.mass.index” variables.

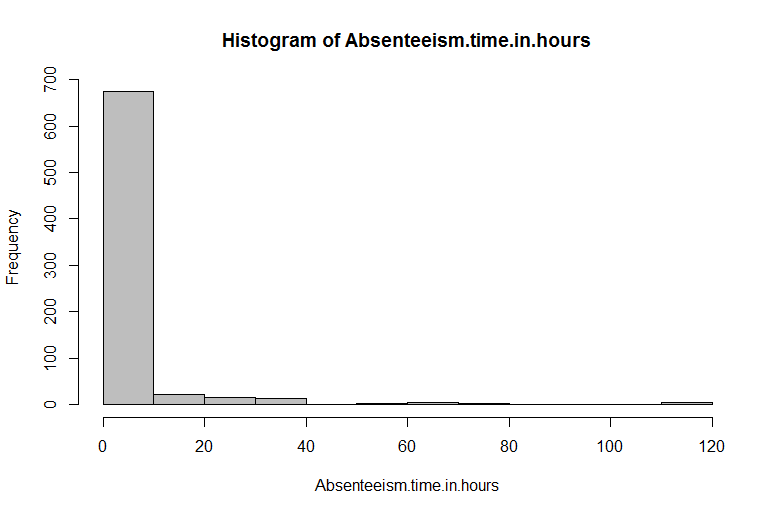


Figure:15:: Histogram of “Absenteeism.time.in.hours”variable.

We can see clearly from the above graphs that our data is not normally distributed, so here we are applying the Standardization method to scale the data.

**2.1.5 Dataset sampling:**

After feature scaling to build a model we are splitting dataset into train and test dataset. Train dataset contains 80% of the data which will be used to train the machine learning algorithm. Test dataset contains 20 % of the data which will be used to test the outcome of the machine learning algorithm.

**2.2 Principal Component Analysis**

After creating the dummies for the categorical variable, now numbers of the dataset increased to the 116, which is very high. And when we apply the model on the data set its not giving acceptable accuracy. It happens because of not all the variable are important, or may be multi collinearity among the variables.so here we are performing **Principal Component Analysis** to extract the important variables from the dataset for further model building.

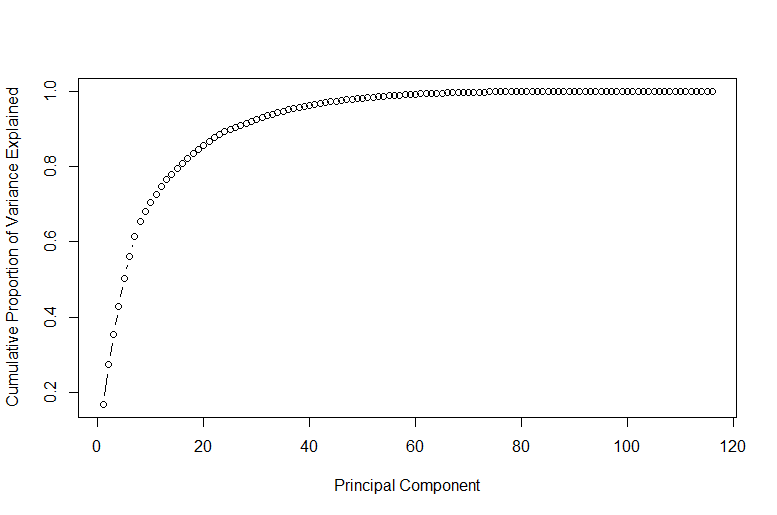


Figure:16:: cumulative scree plot(R)

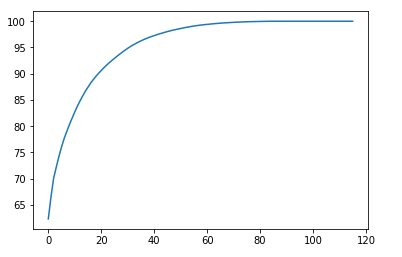


Figure:16:: cumulative scree plot(Python)

From the above graph we can say that first 50 variables are explain more that 95% of the data.so we will select first 50 variables to feed our model.

**2.2 Modeling**

**2.2.1 Decision Tree for classification:**

|  |  |  |
| --- | --- | --- |
|  | R | Python |
| RMSE for train dataset | 0.4094358 | 3.189275569398857 |
| RMSE for test dataset | 0.5339116 | 2.928523586869602 |

**2.2.2 Random Forest:**

|  |  |  |
| --- | --- | --- |
|  | R | Python |
| RMSE for train dataset | 0.1547309 | 1.0224014514777564 |
| RMSE for test Dataset | 0.3848523 | 2.763885291077596 |

**2.2.3 Linear Regression:**

|  |  |  |
| --- | --- | --- |
|  | R | Python |
| RMSE for train dataset | 0.004599576 | 2.9042796533645328e-15 |
| RMSE for test Dataset | 0.005653529 | 0.000565994936306012 |

**3. Conclusion**

By comparing the Root Mean Square Error of the different model, we can conclude that Linear Regression model performs the best on the given dataset.

4. What changes company should bring to reduce the number of absenteeism?

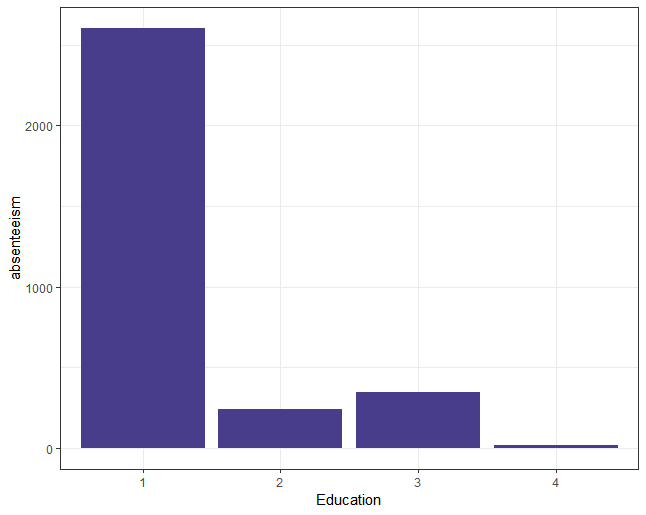


Figure: Absenteeism :: Education.

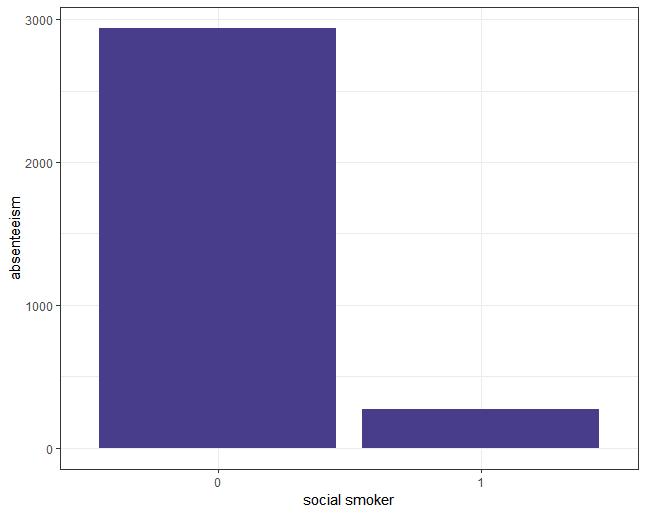


Figure: Absenteeism:: Social Smoker

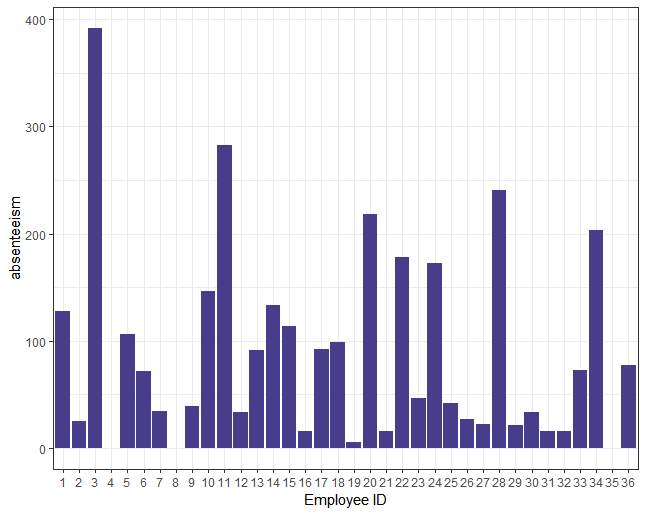


Figure: absenteeism ::Employee ID

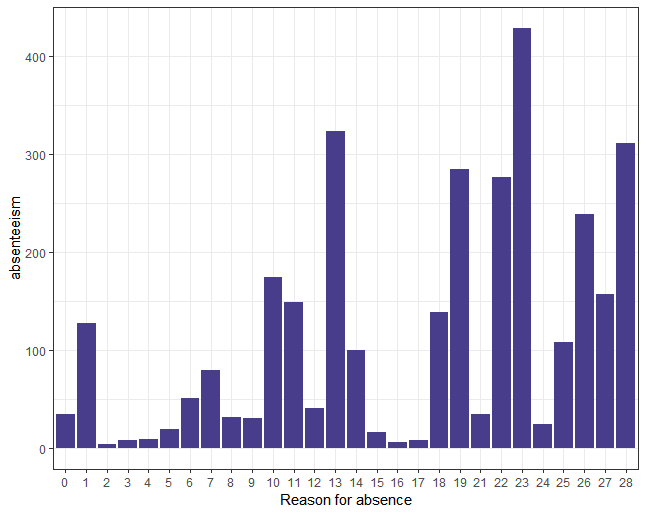


Figure: Absenteeism :: Reason for absence

From the above visualization we can see that relationship of various variable with our target variable. For example,

absenteeism is very high by the employee whose education is low. So company should take care of this, by hiring educated employee company can reduce the absenteeism time.

Some employee’s absenteeism hours are very high with respect to others, by this kind of visualization company can easily detect that employee and can take proper action against them.

There is high absenteeism due to some particular disease, so company can reduce absenteeism by providing proper medical facilities regarding that particular disease.

Thus, company can build new strategies or policies to reduce the absenteeism.

5.How much losses every month can we project in 2011 if same trend of absenteeism continues?

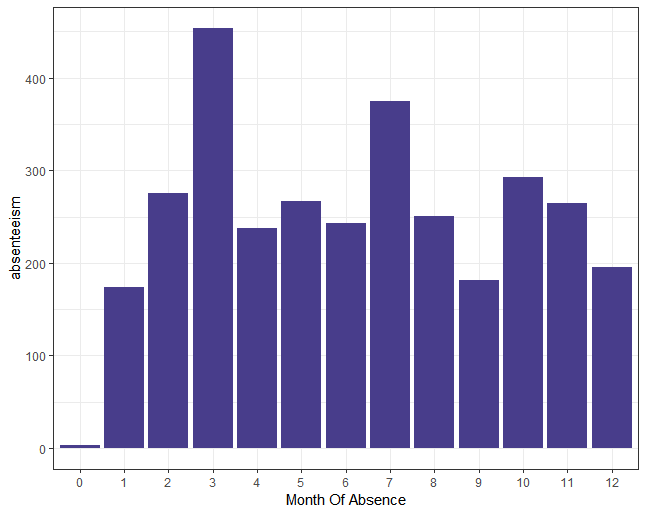


Figure: Month wise absenteeism

If the absenteeism trends continue we can project the month wise absenteeism from the above graph.