Employee Absenteeism

*Akash Siripuram*

*12-07-2019*

**Contents**

1. **Introduction**  ……………………………………………………………………….3

1.1 Problem Statement …………………………………………………………............3

1.2 Data…………………………………………………………………… ……….3

1. **Methodology**……………………………………………………………………..…9
   1. Pre-Processing……………………………………………………………………….9
      1. Missing Value Analysis …………………………………………………………….9
      2. Outlier Analysis……………………………………………………………..……...10
      3. Feature Selection……………………………………………………….......……….14
      4. Feature Scaling…………………………………………………… ……… ………14
   2. Modeling …………………………………………………………………………….18
      1. Model Selection……………………………………………………………..………18
2. **Conclusion**………………………………………………….………… ……….….19
   1. Model Evaluation…………………………………………………..……….………19
   2. RMSE…………………………………………………………..………….……….20

**Appendix –A** R- code ………………….……………………………………………….21

**Appendix -B** Python Code …………………………………………………...………30

**Chapter 1**

**Introduction**

* 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?

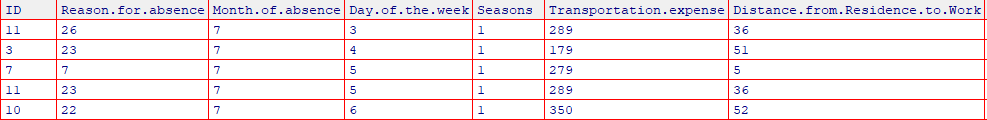
**1.2 Data**

Our task is to build a model so that from that model we can bring the solution to both the problems

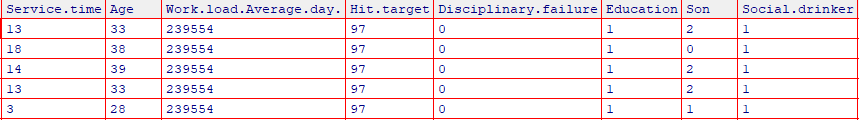
mentioned above.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

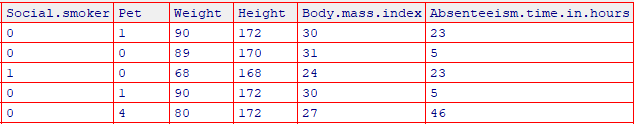
**Table for Columns (1-7)**



**Table for Columns (8-15)**

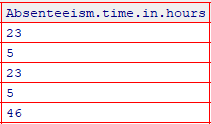


**Table for Columns (16-21)**



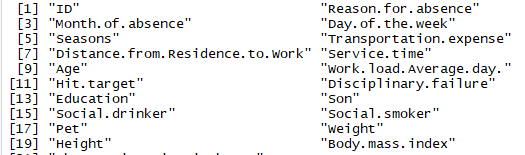
**Dependent Variable:**

It is nothing but the variable that we what to predict or classify. Here as we are having 21 variables. The Absenteeism.time.in.hours variable will become our Dependent Variable. There will be only one dependent Variable for whole data.



**Independent Variables:**

The variables that helps in predicting the dependent variables are known as independent variables. They carry information that helps in predicting the dependent variables. There can be more than 1 independent variables in the data.



Let’s see the type of variables we have :

***Categorical Variables:***

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:
   1. Certain infectious and parasitic diseases
   2. Neoplasms
   3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
   4. Endocrine, nutritional and metabolic diseases
   5. Mental and behavioral disorders
   6. Diseases of the nervous system
   7. Diseases of the eye and adnexa
   8. Diseases of the ear and mastoid process
   9. Diseases of the circulatory system
   10. Diseases of the respiratory system
   11. Diseases of the digestive system
   12. Diseases of the skin and subcutaneous tissue
   13. Diseases of the musculoskeletal system and connective tissue
   14. Diseases of the genitourinary system
   15. Pregnancy, childbirth and the puerperium
   16. Certain conditions originating in the perinatal period
   17. Congenital malformations, deformations and chromosomal abnormalities
   18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
   19. Injury, poisoning and certain other consequences of external causes
   20. External causes of morbidity and mortality
   21. Factors influencing health status and contact with health services.

1. 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).
2. Month of absence
3. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
4. Seasons (summer (1), autumn (2), winter (3), spring (4))
5. Disciplinary failure (yes=1; no=0)
6. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
7. Social drinker (yes=1; no=0)
8. Social smoker (yes=1; no=0))

Let’s see the continuous variables:

***Continuous Variables:***

1. Son (number of children)

2. Pet (number of pet)

3. Transportation expense

4. Distance from Residence to Work (kilometers)

5. Service time

6. Age

7. Work load Average/day

8. Hit target

9. Weight

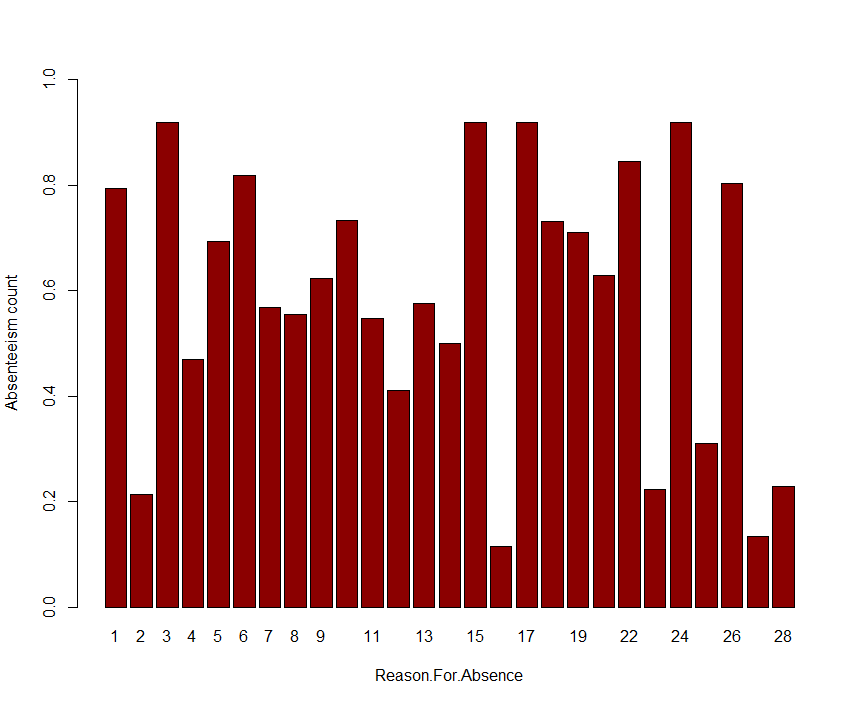
10. Height

11. Body mass index

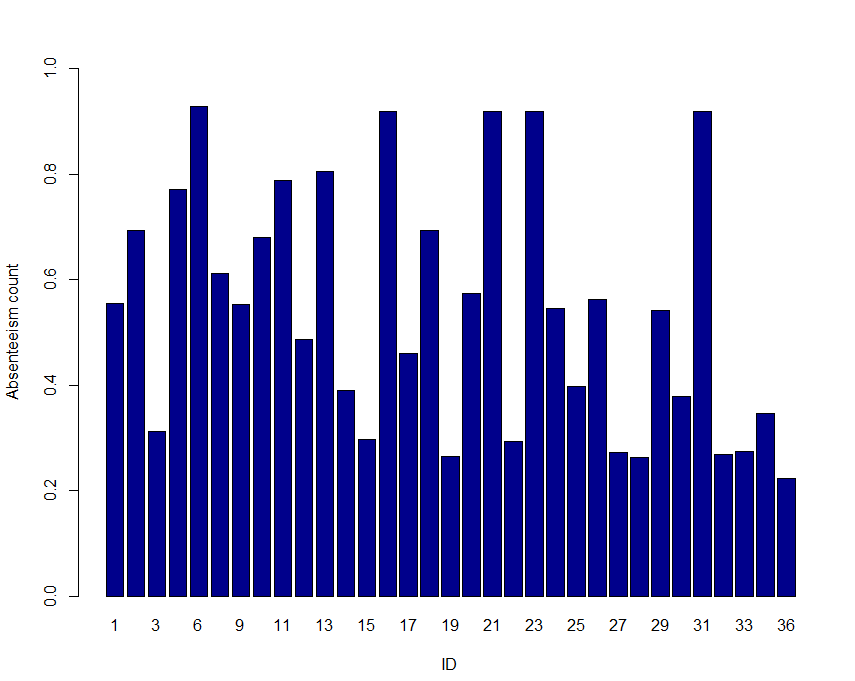
12. Absenteeism time in hours (target)

Let’s also see some visualizations of the categorical variables w.r.t dependent variable:

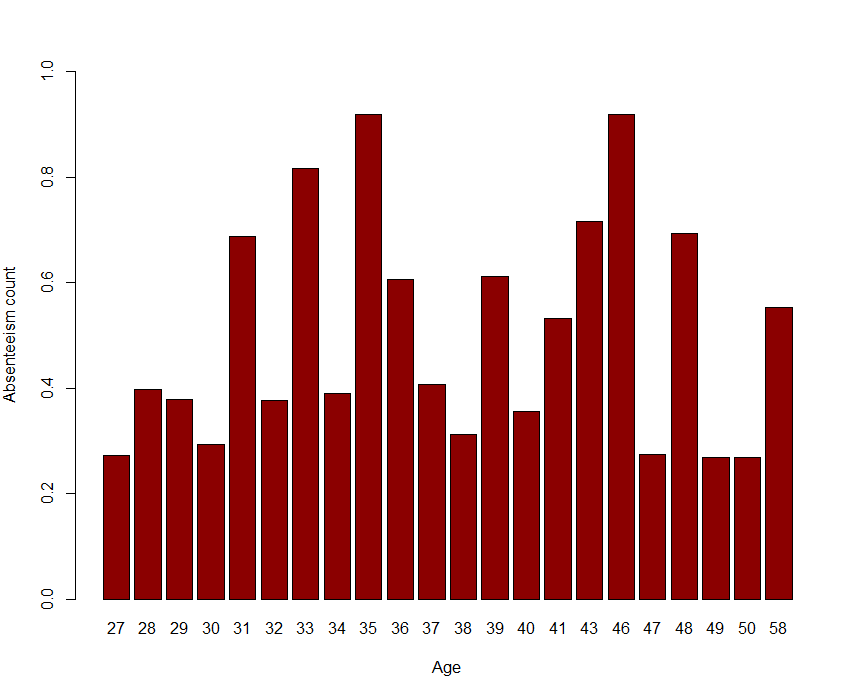
**Absenteeism w.r.t Reason for absence**



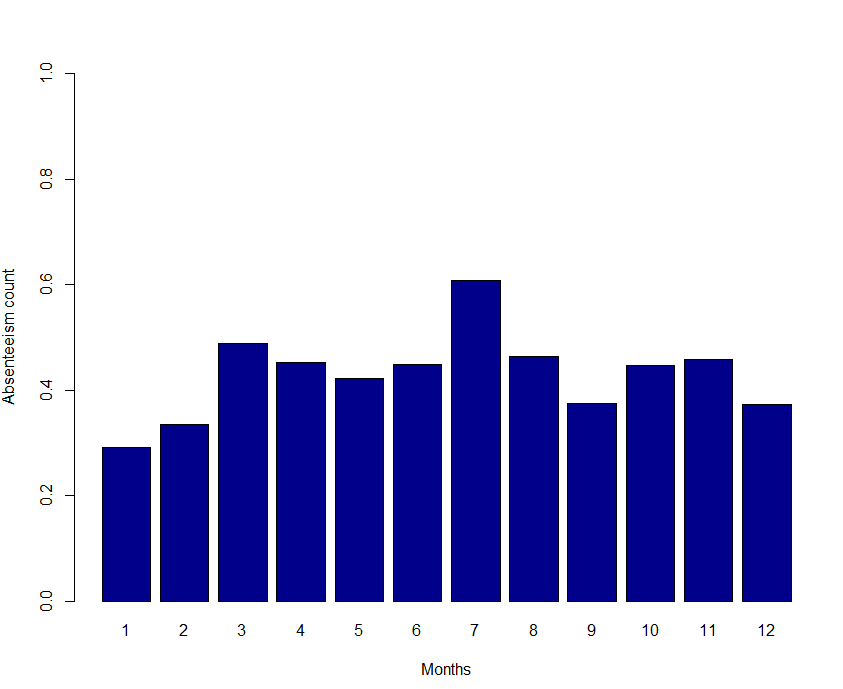
**Absenteeism w.r.t ID**



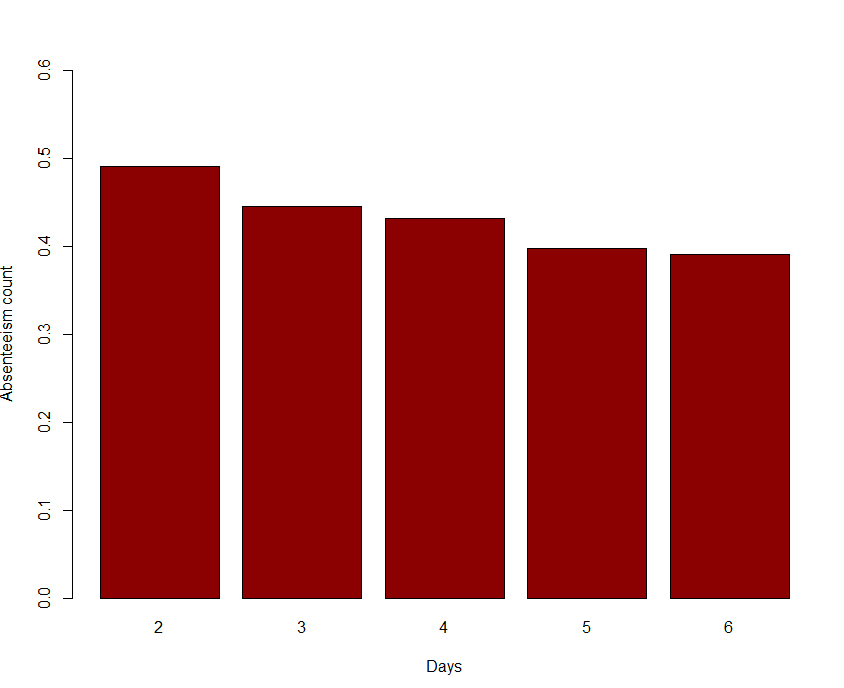
**Absenteeism w.r.t Age**



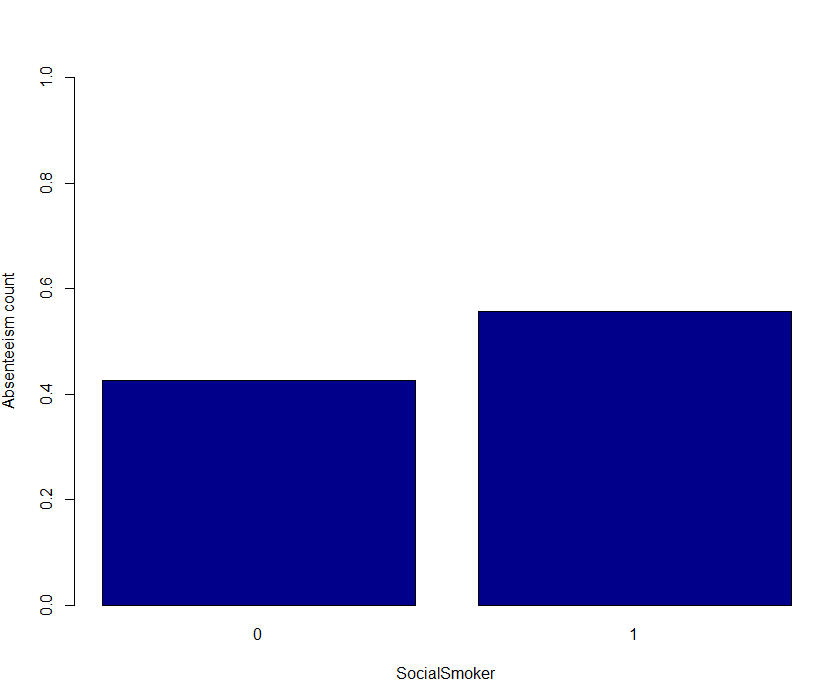
**Absenteeism w.r.t Month of Absence**



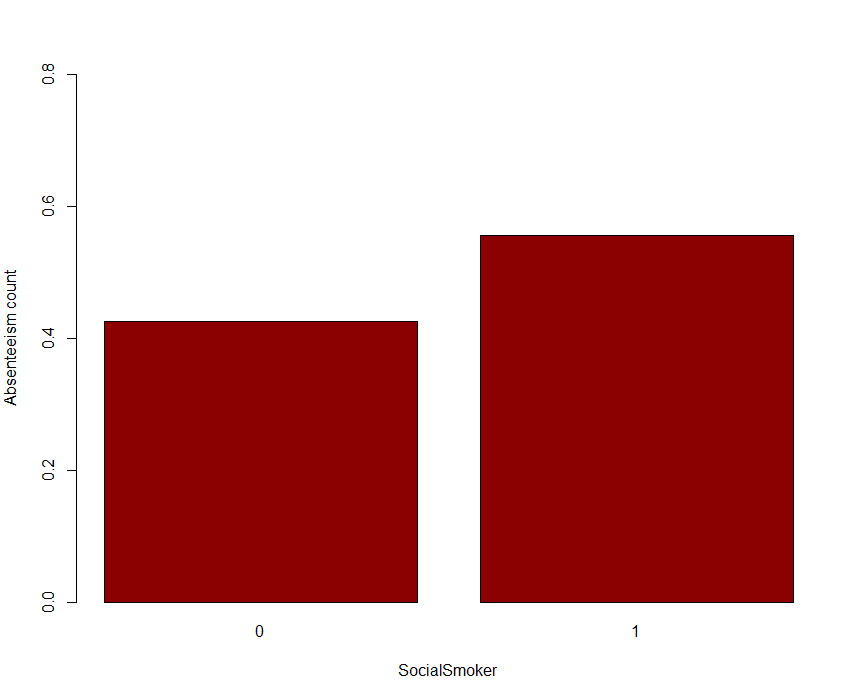
**Absenteeism w.r.t Days of the week:**



**Absenteeism w.r.t Social Smoker**



**Absenteeism w.r.t Social Drinker**



**Chapter 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.

Now, here we have done Missing value analysis, Outlier Analysis, Feature Selection and Feature Scaling. We have also checked them using the Visualizations. I have done visualizations in R, Python.

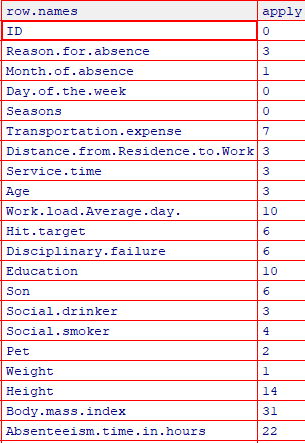
We will be showing each and every step one after another so, that you can understand one after another.

**2.1.1 Missing Value Analysis:**

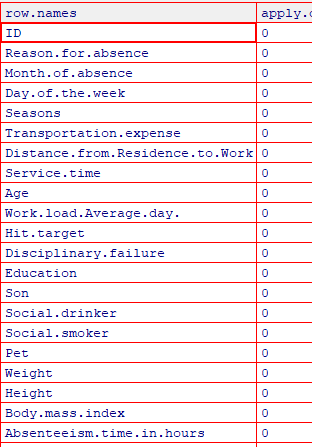
Here we will be checking for the missing values i.e. Na Values or blank spaces , so that it won’t affect our model.

We found that we have a lot of missing values.

So Let’s see them:



After Imputation using KNN:



**2.1.2 Outlier Analysis:**

Actually, why we do outlier analysis is that to find the outliers and remove them.

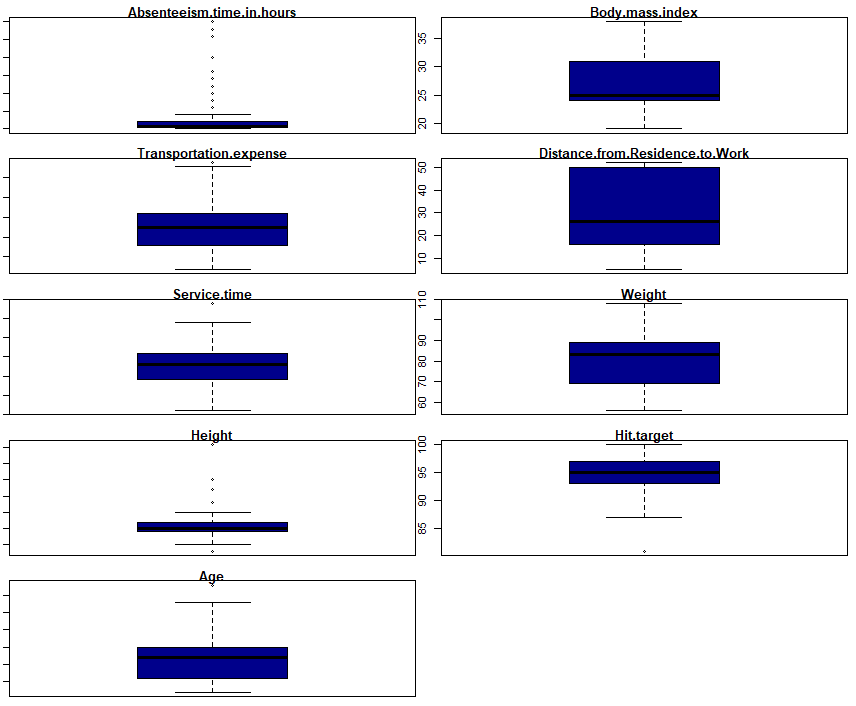
So, what are outliers. In simple words let’s take an example:

If I have a flower shop and I want to total the average sales in a month. But in the same month on a particular day I have sold many flowers that I didn’t sell in other days. So, when counting the average, I will be having a problem because on that particular day I sold many flowers. So, when counting for each normal day average I will have the effect of that particular day on other days. Such the no of flowers sold on that day are known as outliers.

So, we have to remove them so that they won’t affect our model.So, Outlier analysis works only on Continous variables let’s check them.

Let’s see the outliers for each and every continuous variables:

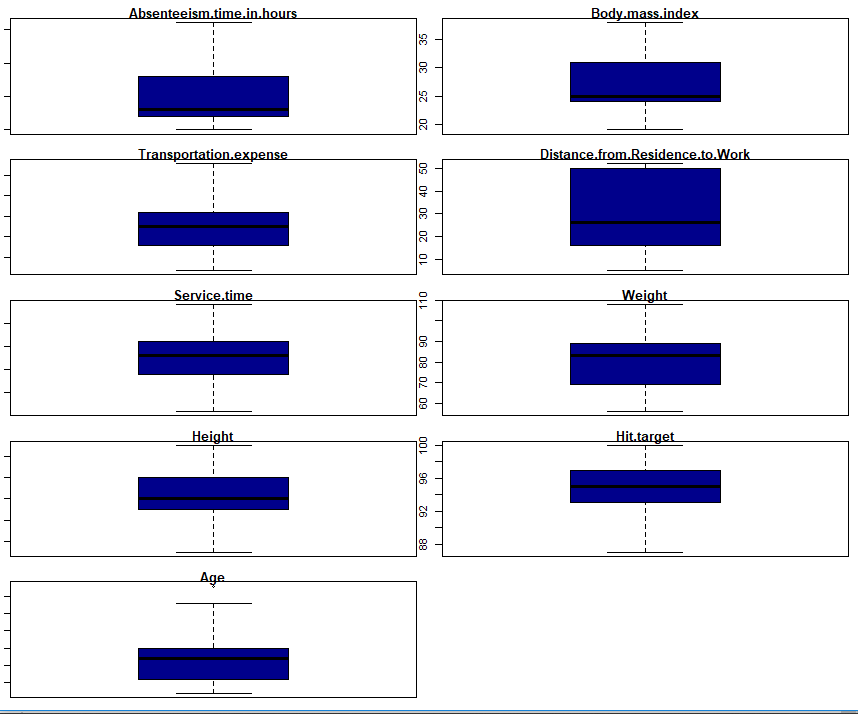
**Boxplot For All the continuous variables:**



Here we have found many outliers so, we should remove them because it will affect our model.

We used KNN imputation for removing the outliers and after removing of outliers we can again visualize the variables using boxplots in next graph:

**After Outlier Analysis:**



When we have seen the boxplots of continuous variables, we found that the outliers are present in variables named Target variable, Transportation Expense ,Service Time, Height, Age, Hit target variables.

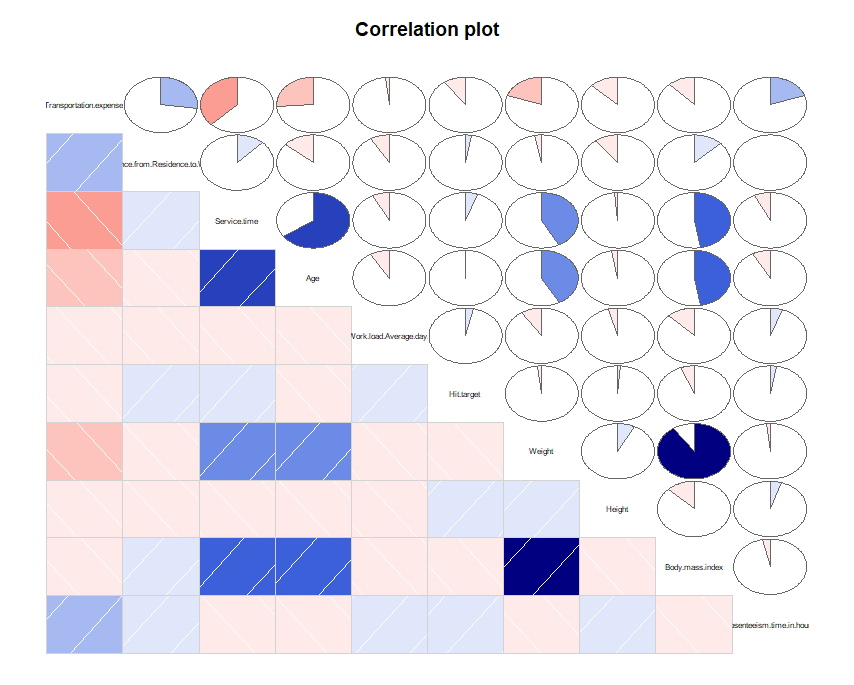
So, when we have done outlier analysis, we have also shown you the boxplots after outlier analysis.

We have used the best methods to reduce the outliers present in that particular variable.

**2.1.3 Feature Selection:**

Here what we do is that we will be finding the variables that give more information about out dependent variable. If any variable is not giving meaning then, it can be removed so that it won’t decrease the model’s accuracy.

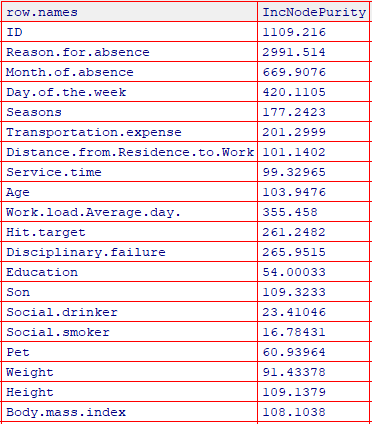
We have done it by using correllogram graph:



Here above we can see the variables, Body Mass Index is having multicollinearity issue with weight, Age, Service time. So, we thought to remove Body Mass Index.

We have also used random forest so that we can check the importance of variables:

We found that:



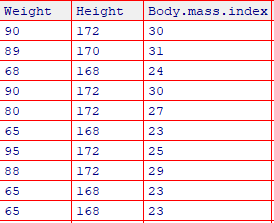
**2.1.4 Feature Scaling:**

Why do we do this?

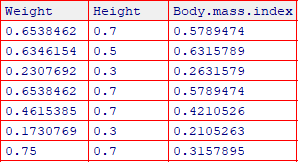
We have the data which is of different ranges or units. So, when we wanted to predict the model. This really causes a huge problem to the model so, we will convert all the values in between the range of 0 and 1. We will be using Normalization to do this instead of Standardization.

Let’s see with an example:

Before Feature Scaling:



After Feature Scaling:



**2.2 Modeling**

**2.2.1 Model Selection**

We have reached here We have done with Exploratory data analysis and we need a better model.

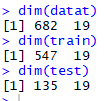
So, to develop a model we should be having two datasets named train and test.

We train our model with the train dataset and test that model using the RMSE values.

So, in the whole project we should not touch even the test data after doing Sampling.

Sampling is a technique that helps to take a sample of data from the whole data by considering all the characteristics of the data.

So, we have our train and test data now.



We have the train data of 80% and test data of 20%.

So, 547 is 80% of Actual data and 135 is the 20% of Actual data

We don’t remove or change the variables during Sampling because we will be needing all the Actual dataset variables in both train and test datasets.

So, now how to do the training on the train dataset.

We have many different MLA’s(Machine Learning Algorithms) as well as Statistical models that helps in

making a better model.

**Chapter 3**

**Conclusion**

**3.1 Model Evaluation**

We will be using the RMSE for evaluating the model.

We used both Decision Tree and Random Forest for evaluating the model.

We found both the algorithms have worked well and gave an

RMSE of Decision Tree is 3.14

RMSE of Random Forest is 3.42

So, we considered that Decision Tree is working well so we selected Decision Tree to be used for further Analysis.

* 1. **Problem outcomes:**

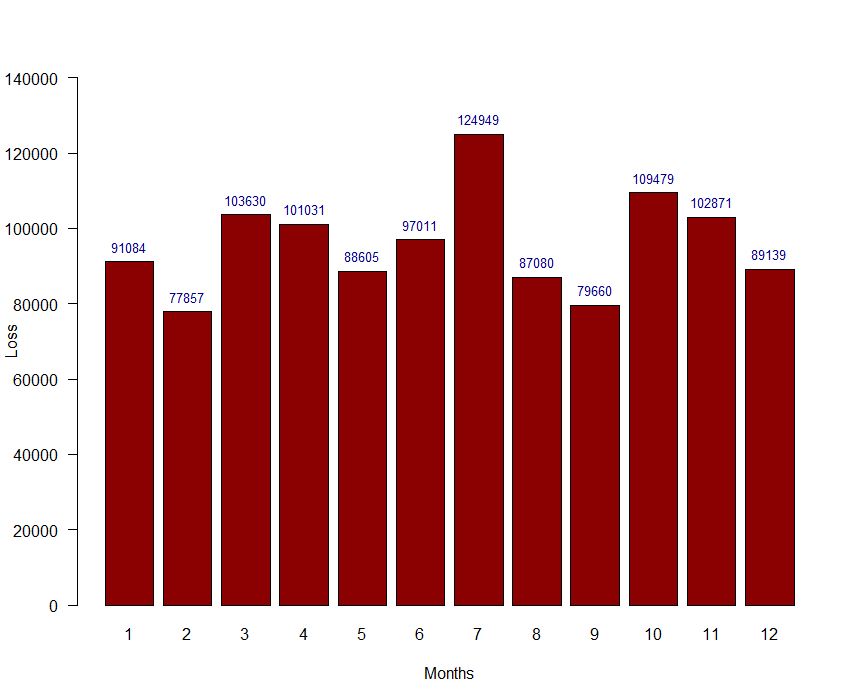
1. What changes company should bring to reduce the number of absenteeism?

As we have visualized the variable w.r.t the target variable we have found many reasons and causes that resulted in absenteeism.

We found that that ID’s 6,16,21,23,31 are having much absenteeism when compared with others and the reason for absence is also high for these reasons. like 3,15,17,23. These reasons are common for the absentees. Reason number 26 shows that there are a greater number of people who did not give any response why they are absent. So certain action should be taken on those which will help in further analysis. The age group below 30 years are having less absenteeism. we have found that months march and July are having high absenteeism and January is having low absenteeism. we found that Monday is highest because it is followed by week days. Social Drinker an Social Smoker are also affecting absenteeism but in a slight manner.

It won't affect much even if the employees are having the habits of smoking and drinking.

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



**Here we have used the formula :**

**(Absenteeism in hours/Service time) \* Workload average**

**So we found that July is having the much loss next to October and March if the same trend of this absenteeism continues. The least loss is in the month of February.**

**R code:**

rm(list=ls())

#install.packages("RColorBrewer")

#install.packages("class")

library("xlsx")

library("corrgram")

library(caret)

library(rpart)

library(MASS)

library(DMwR)

library(randomForest)

library(Metrics)

library(dplyr)

library(usdm)

#-----------------------------------------------------------------------------------------------------------------------------------------------------------------

#Getting data

data=read.xlsx("Absent.xls",sheetIndex=1)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Changing the datatype

str(data)

colnames(data)

for(i in c(1,2,3,4,5,12,13,14,15,16,17)){

data[,i]=as.factor(data[,i])

}

#converting the datatype of Continous variables into integer

for(i in c(6,7,8,9,10,11,18,19,20,21)){

data[,i]=as.integer(data[,i])

}

#----------------------------------------------------------------------------------------------------------------------------------------------------------------------

int\_var=data[,c(6,7,8,9,10,11,18,19,20,21)]

cat\_var=data[,-c(6,7,8,9,10,11,18,19,20,21)]

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#missing value analysis:

missing\_val=data.frame(apply(data,2,function(x){sum(is.na(x))}))

sum(missing\_val)

# df=data[rowSums(is.na(data))>0,]

# df

#------------------------------------------------------------------------------------------------------------------------------------------------------------------str(data)

data=data[!is.na(data$Absenteeism.time.in.hours),]

dim(data)

#Imputing the values using KnnImputation:

data=knnImputation(data,k=5)

data

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Outlier Analysis:

par('mar')

par(mar=c(1,1,1,1))

par(mfrow=c(5,2))

boxplot(data$Absenteeism.time.in.hours,col="Darkblue",main="Absenteeism.time.in.hours")

boxplot(data$Body.mass.index,col="Darkblue",main="Body.mass.index")

boxplot(data$Transportation.expense,col="Darkblue",main="Transportation.expense")

boxplot(data$Distance.from.Residence.to.Work,col="Darkblue",main="Distance.from.Residence.to.Work")

boxplot(data$Service.time,col="Darkblue",main="Service.time")

boxplot(data$Weight,col="Darkblue",main="Weight")

boxplot(data$Height,col="Darkblue",main="Height")

boxplot(data$Hit.target,col="Darkblue",main="Hit.target")

boxplot(data$Age,col="Darkblue",main="Age")

str(data)

colnames(data)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Replacing the outliers with NA

for (i in c(6,7,8,10,11,18,19,20,21)){

val=data[,i][data[,i]%in%boxplot.stats(data[,i])$out]

data[,i][data[,i]%in% val]=NA

}

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Imputation using KnnImputation

data=knnImputation(data,k=5)

#Checking the missing values:

missing\_val=data.frame(apply(data,2,function(x){sum(is.na(x))}))

missing\_val

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Checking the outliers again with boxplot after removing outliers

par('mar')

par(mar=c(1,1,1,1))

par(mfrow=c(5,2))

boxplot(data$Absenteeism.time.in.hours,col="Darkblue",main="Absenteeism.time.in.hours")

boxplot(data$Body.mass.index,col="Darkblue",main="Body.mass.index")

boxplot(data$Transportation.expense,col="Darkblue",main="Transportation.expense")

boxplot(data$Distance.from.Residence.to.Work,col="Darkblue",main="Distance.from.Residence.to.Work")

boxplot(data$Service.time,col="Darkblue",main="Service.time")

boxplot(data$Weight,col="Darkblue",main="Weight")

boxplot(data$Height,col="Darkblue",main="Height")

boxplot(data$Hit.target,col="Darkblue",main="Hit.target")

boxplot(data$Age,col="Darkblue",main="Age")

colnames(data)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Feature Selection:

#correlation plot:

Num\_data=data[,c(6,7,8,9,10,11,18,19,20,21)]

corrgram(Num\_data,order=F,upper.panel=panel.pie,text.panel =panel.txt,main="Correlation plot")

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

# Checking importance of Catagorical variables using random Forest:

fit\_rf = randomForest(data$Absenteeism.time.in.hours~., data=data)

# Create an importance

Rf\_Imp=importance(fit\_rf)

#fix(Rf\_Imp)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Removing farcical info:

data= data[!data$Month.of.absence==0 & !data$Reason.for.absence==0, ]

# data=data[!data$Reason.for.absence==0,]

# data=data[!data$Month.of.absence==0,]

#Creating a copy

datat=data

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Checking the dimensions and structure:

dim(datat)

str(datat)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Feature Scaling:

str(datat)

colnames(datat)

# Applying feature selecton:

for(i in c(6,7,8,9,10,11,18,19,20,21)){

datat[,i] = (datat[,i] - min(datat[,i]))/(max(datat[,i] - min(datat[,i])))

}

dim(datat)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Removing the ID and Body.Mass.Index variables:

data\_cpy=data

datat\_cpy=datat

datat = subset(datat, select = -c(1,20))

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#visualization the data:

par(mar=c(5.1,4.1,4.1,2.1))

library(RColorBrewer)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Visualizing for reason for absence data

ReasonForAbsencevisual=aggregate(datat\_cpy[,21], list(datat\_cpy$Reason.for.absence), mean)

colnames(ReasonForAbsencevisual)=c("Reason.for.Absence","Abseenteism")

barplot(ReasonForAbsencevisual$Abseenteism,names.arg =ReasonForAbsencevisual$Reason.for.Absence,col="Darkred",xlab="Reason.For.Absence",ylab="Absenteeism count",ylim=c(0,0.6))

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Visualizing for ID data

IDVisual=aggregate(datat\_cpy[,21], list(datat\_cpy$ID), mean)

colnames(IDVisual)=c("ID","Abseenteism")

par(mar=c(5.1,4.1,4.1,2.1))

barplot(IDVisual$Abseenteism,names.arg =IDVisual$ID,col="Darkblue",xlab="ID",ylab="Absenteeism count",ylim=c(0,0.6))

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

###Visualizing for age data

data\_cpy$Age=as.factor(data\_cpy$Age)

AgeVisual=aggregate(datat\_cpy[, 21], list(data\_cpy$Age), mean)

colnames(AgeVisual)=c("Age","Abseenteism")

barplot(AgeVisual$Abseenteism,names.arg =AgeVisual$Age,col="Darkred",xlab="Age",ylab="Absenteeism count",ylim=c(0,0.6))

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

###visualizing month.of.absence w.r.t absenteeism count

hm=aggregate(datat\_cpy[,21],list(datat\_cpy$Month.of.absence),mean)

colnames(hm)=c("Month.of.absence","Absenteeism")

barplot(hm$Absenteeism,names.arg =hm$Month.of.absence,col="Darkblue",xlab="Months",ylab="Absenteeism count",ylim = c(0,0.4))

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

###visualizing month.of.absence w.r.t absenteeism count

MOAVisual=aggregate(datat\_cpy[,21],list(datat\_cpy$Month.of.absence),mean)

colnames(MOAVisual)=c("Month.of.absence","Absenteeism")

barplot(MOAVisual$Absenteeism,names.arg =MOAVisual$Month.of.absence,col="Darkblue",xlab="Months",ylab="Absenteeism count",ylim = c(0,1.0))

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

###visualizing Day.of.the.week w.r.t absenteeism

DOWVisual=aggregate(datat\_cpy[,21],list(datat\_cpy$Day.of.the.week),mean)

colnames(DOWVisual)=c("Day.of.the.week","Absenteeism")

barplot(DOWVisual$Absenteeism,names.arg =DOWVisual$Day.of.the.week,col="Darkred",xlab="Days",ylab="Absenteeism count",ylim=c(0,0.4))

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

###Visualizing the Social.drinker w.r.t absenteeism

SDVisual=aggregate(datat\_cpy[,21],list(datat\_cpy$Social.smoker),mean)

colnames(SDVisual)=c("SocialSmoker","Absenteeism")

barplot(SDVisual$Absenteeism,names.arg =SDVisual$SocialSmoker,col="Darkblue",xlab="SocialSmoker",ylab="Absenteeism count",ylim=c(0,1.0))

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

###Visualizing the Social.Smoker w.r.t absenteeism

SSVisual=aggregate(datat\_cpy[,21],list(datat\_cpy$Social.smoker),mean)

colnames(SSVisual)=c("SocialSmoker","Absenteeism")

barplot(SSVisual$Absenteeism,names.arg =SSVisual$SocialSmoker,col="Darkred",xlab="SocialSmoker",ylab="Absenteeism count",ylim=c(0,0.8))

datat

colnames(datat)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Model Development

set.seed(321)

indexTrain = createDataPartition(datat$Absenteeism.time.in.hours, p = .80, list = FALSE)

train = datat[ indexTrain,]

test = datat[-indexTrain,]

dim(datat)

dim(train)

dim(test)

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Applying models:

#Applying Decision Tree on test data:

DT=rpart(Absenteeism.time.in.hours~ .,data=train,method="anova")

DTPredictions=predict(DT,test)

#Calculating rmse for test

rmse(test$Absenteeism.time.in.hours,DTPredictions)

#RMSE value: 0.190

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

# Applying RandomForest Model:

# Fitting model

fit <- randomForest(train$Absenteeism.time.in.hours ~ .,train,ntree=500)

rfpredicted= predict(fit,test)

#calculating rmse

rmse(test$Absenteeism.time.in.hours,rfpredicted)

#RMSE value: 0.18

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

# As we have seen random forest is providing us the less RMSE value we can say that we

# Can use random forest in future for any findings.

#------------------------------------------------------------------------------------------------------------------------------------------------------------------

#Answering the questions:

#####1. causes of absenteism:

###From the visualizations

###1. we came to know that ID no 6,16,21,23,31 are having high absenteeism When compared with others.

##2. when reason.for.absenteeism is taken these levels have highest absenteeism count

# 1,3,6,9,15,17,22,24,26

# 1. Certain infectious and parasitic diseases

# 3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

# 6. Diseases of the nervous system

# 9. Diseases of the circulatory system

# 15. Pregnancy, childbirth and the puerperium

# 17. Congenital malformations, deformations and chromosomal abnormalities

# 22. patient follow-up

# 24. blood donation

# 26. unjustified absence

# Here we must take necessary action on employees who won't mark reason for absence field.

##3. when the visualization is done for Age we found that age below 30 has less absenteeism,

# 31,33,35,36,39,41,43,46,48 age group has high absenteeism so,

# Taking action on these age grouped employees will solve some part of the prooblem.

##4. when done visualization on Month.of.absence

# we have found that months march and july are having high absenteeism and

# January is having low absenteeism.

##5. when done visualization on Days.of.absence

# we found that monday is highest because it is followed by week days.

##6. Social Drinker an Social Smoker are also affecting absenteeism but in a slight manner.

# It won't affect much even if the employees are having the habits of smoking and drinking.

#-----------------------------------------------------------------------------------------------------------------------------------------------------------

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

fix(data)

colnames(data)

data$Absenteeism.time.in.hours=as.integer(data$Absenteeism.time.in.hours)

G\_RFA=aggregate(data[,21],list(data$Month.of.absence),mean)

colnames(G\_RFA)=c("Months","Absenteeism")

G\_MOA=aggregate(data$Service.time,list(data$Month.of.absence),mean)

colnames(G\_MOA)=c("Months","service.time")

G\_WLD\_MOA=aggregate(data$Work.load.Average.day.,list(data$Month.of.absence),mean)

colnames(G\_WLD\_MOA)=c("Months","work.load.average.day")

MergeOne=merge(G\_RFA,G\_MOA,by="Months")

MergedData=merge(MergeOne,G\_WLD\_MOA,by="Months")

MergedData$Losses=round((MergedData$Absenteeism/(MergedData$service.time))\*(MergedData$work.load.average.day))

options(scipen = 10)

par(las=1)

MergedData=MergedData[order(MergedData$Months),]

FinalBar=barplot(MergedData$Losses,names.arg =MergedData$Months,col="Darkred",xlab="Months",ylab="Loss",ylim=c(0,140000))

text(x = FinalBar, y = MergedData$Losses, label = MergedData$Losses, pos = 3, cex = 0.8, col = "Darkblue")

**PYTHON CODE**

