EMPLOYEE ABSENTEEISM

*Sumeet Patel*

*28 November 2018*

Contents

[Chapter 1 3](#_Toc531208785)

[**Introduction** 3](#_Toc531208786)

[1.1 Problem Description 3](#_Toc531208787)

[1.2 Problem Statement 3](#_Toc531208788)

[1.3 Data 3](#_Toc531208789)

[Chapter 2 5](#_Toc531208790)

[**Methodology** 5](#_Toc531208791)

[2.1 Pre Processing 5](#_Toc531208792)

[2.1.1 Missing value analysis 7](#_Toc531208793)

[2.1.2 Outlier analysis 9](#_Toc531208794)

[2.1.3 Feature Selection 12](#_Toc531208795)

[2.1.4 Feature Scaling 12](#_Toc531208796)

[2.1.5 Sampling 14](#_Toc531208797)

[2.2 Modelling 14](#_Toc531208798)

[2.2.1 Model Selection 14](#_Toc531208799)

[Chapter 3 15](#_Toc531208800)

[**Conclusion** 15](#_Toc531208801)

[3.1 Model Evaluation 15](#_Toc531208802)

[3.1 Model Selection 15](#_Toc531208803)

[3.1 Answers 15](#_Toc531208804)

[**Appendix A – R codes** 19](#_Toc531208805)

[**References** 26](#_Toc531208806)

Chapter 1

**Introduction**

# Problem Description

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.

# Problem Statement

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?

# Data

Data consist 740 observations and 21 variables. Following is the sample of the data set.

**Table 1.1 Employee Absenteeism (columns 1-7)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Reason of Absence** | **Month of Absence** | **Day of the week** | **Season** | **Transportation Expense** | **Distance from Residence to Woek** |
| 11 | 26 | 7 | 3 | 1 | 289 | 36 |
| 36 | 0 | 7 | 3 | 1 | 118 | 13 |
| 3 | 23 | 7 | 4 | 1 | 179 | 51 |
| 7 | 7 | 7 | 5 | 1 | 279 | 5 |

**Table 1.2 Employee Absenteeism (columns 8-14)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Service time** | **Age** | **Work load Average/day** | **Hit target** | **Disciplinary failure** | **Education** | **Son** |
| 13 | 33 | 2,39,554 | 97 | 0 | 1 | 2 |
| 18 | 50 | 2,39,554 | 97 | 1 | 1 | 1 |
| 18 | 38 | 2,39,554 | 97 | 0 | 1 | 0 |
| 14 | 39 | 2,39,554 | 97 | 0 | 1 | 2 |

**Table 1.3 Employee Absenteeism (columns 15-21)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Social drinker** | **Social smoker** | **Pet** | **Weight** | **Height** | **Body mass index** | **Absenteeism time in hours** |
| 1 | 0 | 1 | 90 | 172 | 30 | 4 |
| 1 | 0 | 0 | 98 | 178 | 31 | 0 |
| 1 | 0 | 0 | 89 | 170 | 31 | 2 |
| 1 | 1 | 0 | 68 | 168 | 24 | 4 |

Following are the given 21 independent 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:

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 behavioural 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)

Chapter 2

**Methodology**

# 2.1 Pre Processing

After checking the unique values we concluded that following variables are categorical. So converted them from numeric to object.

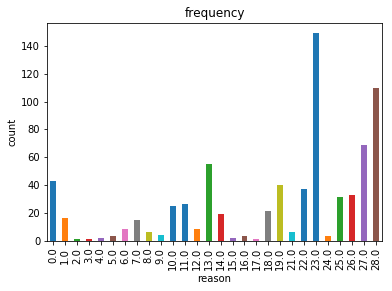
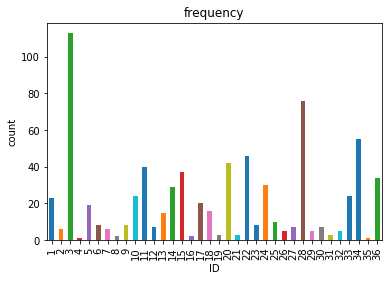
1. Individual identification (ID)
2. Reason for absence
3. Month of absence
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
5. Seasons (summer (1), autumn (2), winter (3), spring (4))
6. Disciplinary failure (yes=1; no=0)
7. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
8. Son (number of children)
9. Social drinker (yes=1; no=0)
10. Social smoker (yes=1; no=0)
11. Pet (number of pet)

Let’s have a look at all the continuous variables and their description.

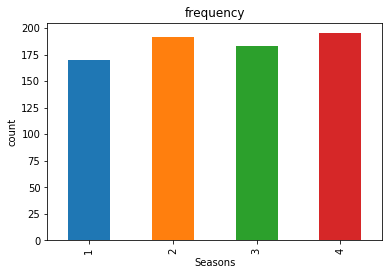
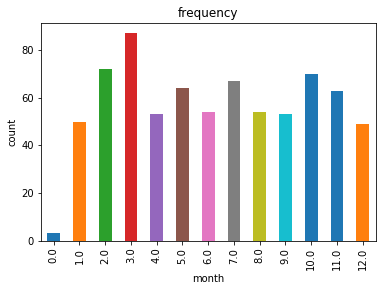
**Table 2.1 Summary before treating outliers**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | Std | Mean | Median | Min | Max | Count | Unique |
| Expense | 66.95 | 221.03 | 225 | 118 | 388 | 733 | 24 |
| Distance | 14.84 | 29.66 | 26 | 5 | 52 | 737 | 25 |
| Time | 4.38 | 12.56 | 13 | 1 | 29 | 737 | 18 |
| Age | 6.48 | 36.44 | 37 | 27 | 58 | 737 | 22 |
| Workload | 38981.88 | 271188.86 | 264249 | 205917 | 378884 | 730 | 38 |
| Hit target | 3.79 | 94.58 | 95 | 81 | 100 | 734 | 13 |
| Weight | 12.86 | 79.06 | 83 | 56 | 108 | 739 | 26 |
| Height | 6.08 | 172.15 | 170 | 163 | 196 | 726 | 14 |
| BMI | 4.29 | 26.68 | 25 | 19 | 38 | 709 | 17 |
| Absence time | 13.47 | 6.97 | 3 | 0 | 120 | 718 | 19 |

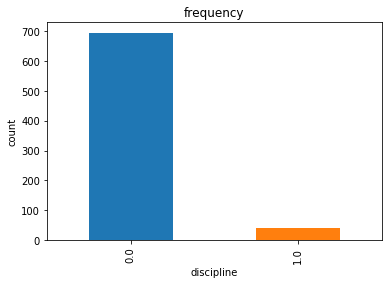
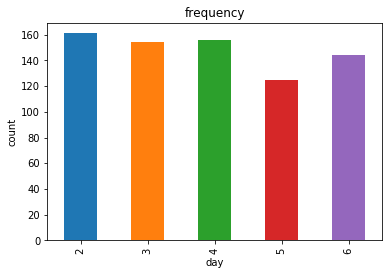
We will see the frequency distribution of categorical variables.



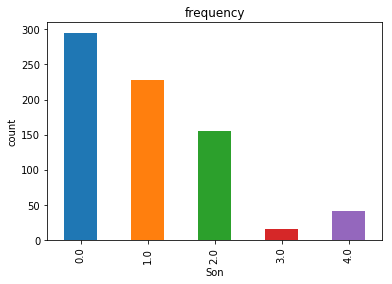
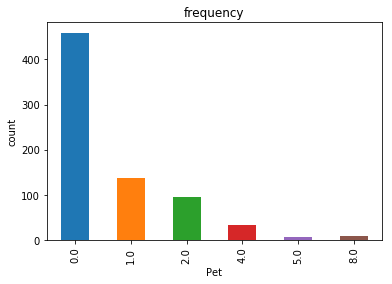
As per the distribution of the IDs we see that ID no 3 is more frequently appearing followed by ID no 28. When it comes to reasons, reason no 23 is the topmost reason for employee absenteeism.



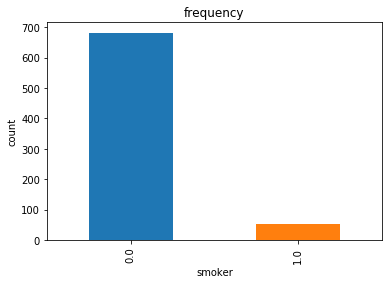
3rd month has more number of observations and 0 indicates that there are cases where there was no absenteeism. Almost all the seasons have approximately equal number of observations.



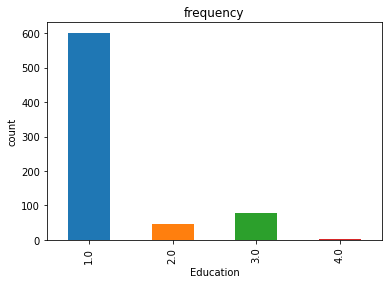
Data looks approximately equally distributed among the days, slightly less observations of Day 6 followed by Day 5. Disciplinary failure is very less.



Either most of the employees don’t have pet or son or the employees who don’t have pet or son are more frequent in the dataset.

There are more number of observations for drinkers than smokers.



So education wise more number of observations are for high school and least for PhD holders.

## 2.1.1 Missing value analysis

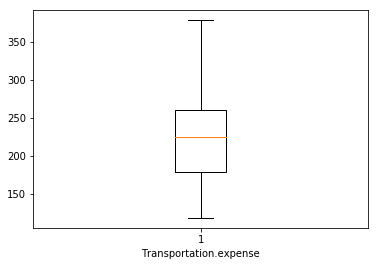
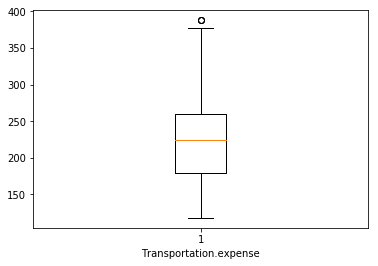
If a columns has more than 30% of data as missing value either we ignore the entire column or we ignore those observations. In the given data the maximum percentage of missing value is 4.189% for **body mass index** column. So we will compute missing value for all the columns.

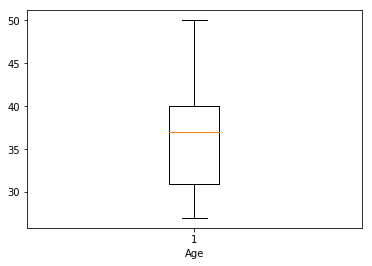
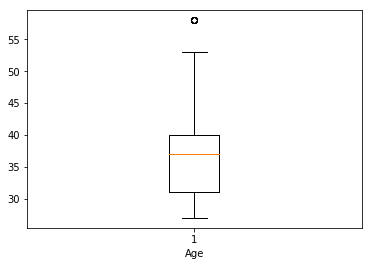
|  | **Variables** | **Missing\_percentage** |
| --- | --- | --- |
| **0** | bmi | 4.189189 |
| **1** | absentime | 2.972973 |
| **2** | Height | 1.891892 |
| **3** | Work load Average/day | 1.351351 |
| **4** | Education | 1.351351 |
| **5** | expense | 0.945946 |
| **6** | Son | 0.810811 |
| **7** | discipline | 0.810811 |
| **8** | target | 0.810811 |
| **9** | smoker | 0.540541 |
| **10** | Age | 0.405405 |
| **11** | reason | 0.405405 |
| **12** | time | 0.405405 |
| **13** | distance | 0.405405 |
| **14** | drinker | 0.405405 |
| **15** | Pet | 0.270270 |
| **16** | Weight | 0.135135 |
| **17** | month | 0.135135 |
| **18** | Seasons | 0.000000 |
| **19** | day | 0.000000 |
| **20** | ID | 0.000000 |

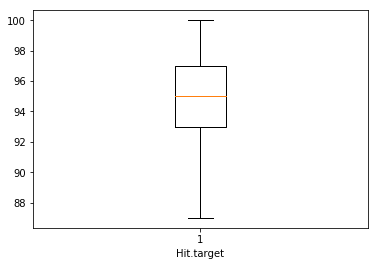
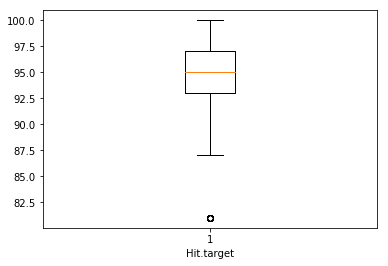
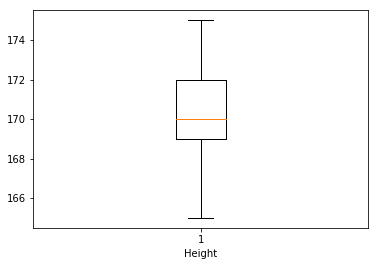
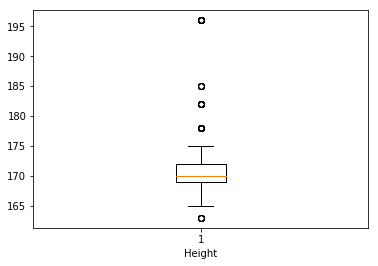
## 2.1.2 Outlier analysis

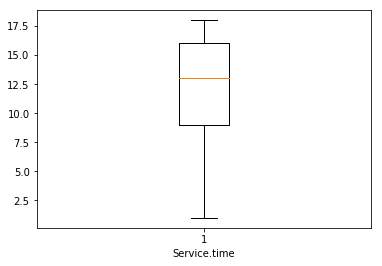
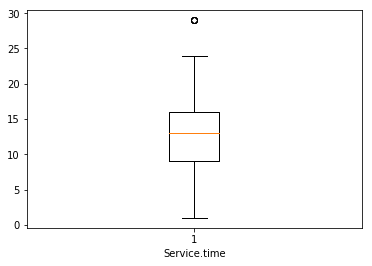
To detect the outliers in the continuous variables we have plotted boxplots. As we concluded according to the histogram, there are no outliers in the variables ‘Distance from Residence to Work’, ‘Weight’ and ‘Body mass index’. While those skewed ones had outliers. So we capped the outliers with 1st and 99th percentile values in most of the variables.

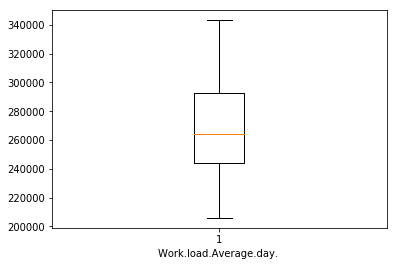
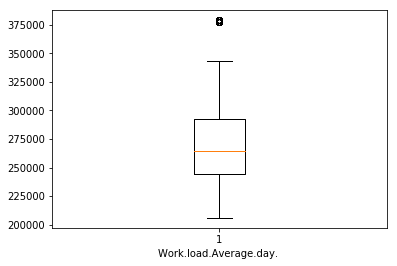
Following is the summary in the form of boxplot of the dataset before and after treating the outliers.

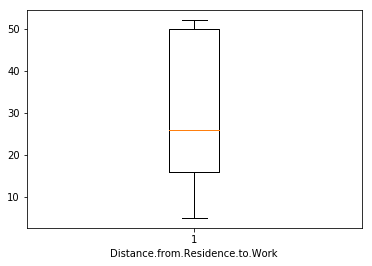
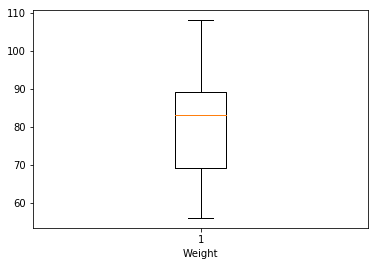


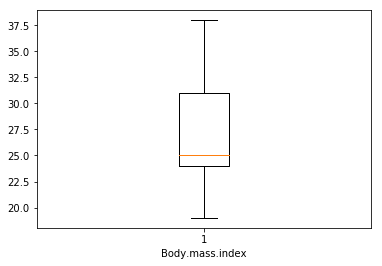






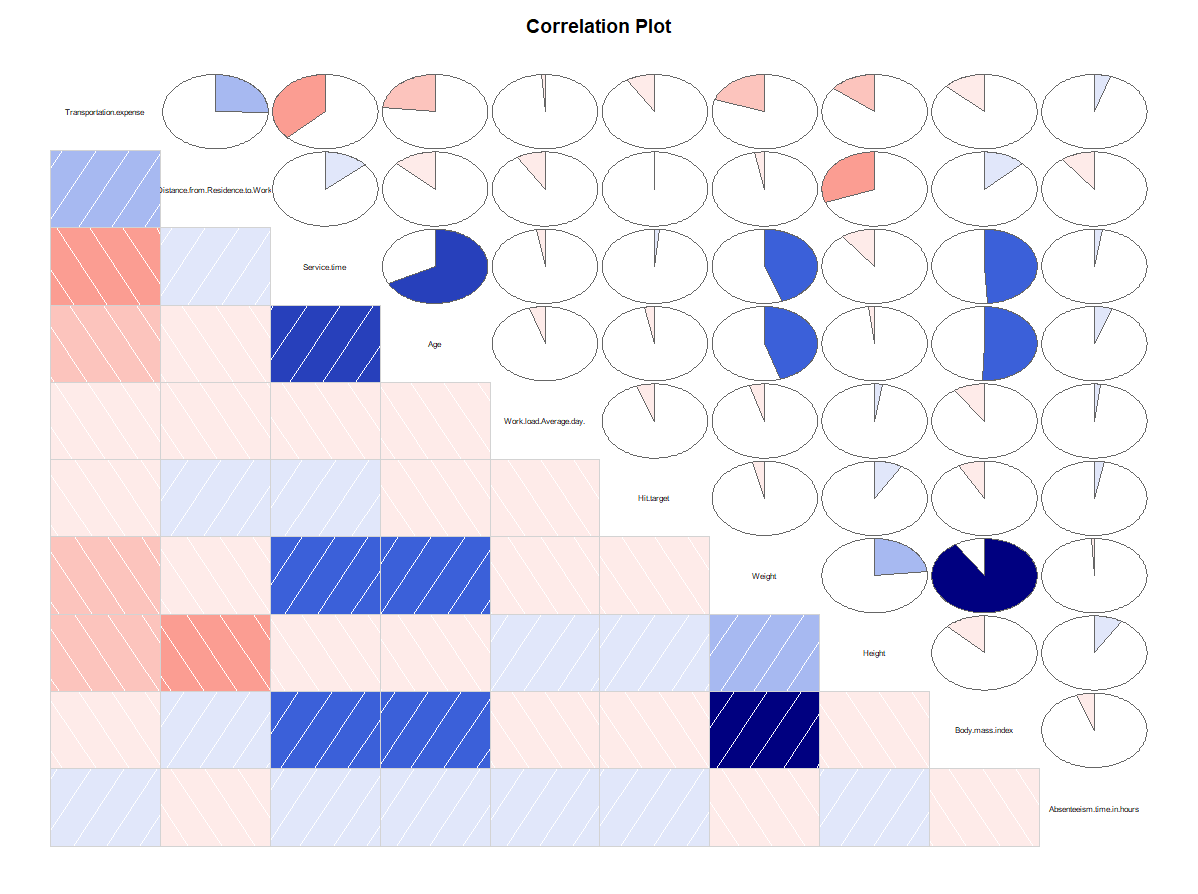




## 2.1.3 Feature Selection

Here we will check the correlation within the continuous variables through correlation plot. If the continuous predictive variables are correlated to any other then we will try to eliminate in between those variables which are highly correlated. We will plot the correlation among all the variables as all are now in continuous form.

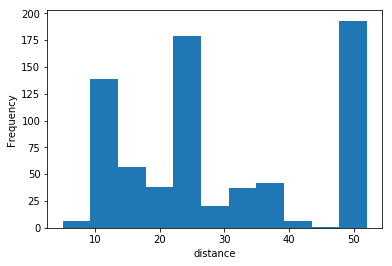
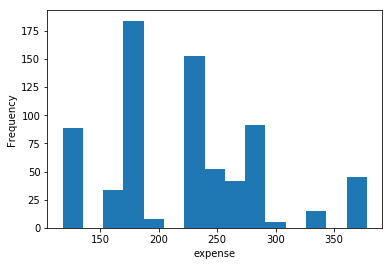


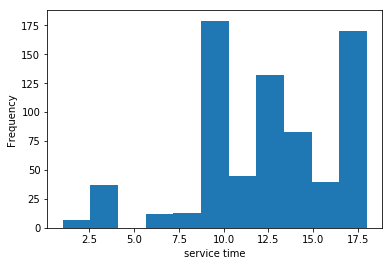
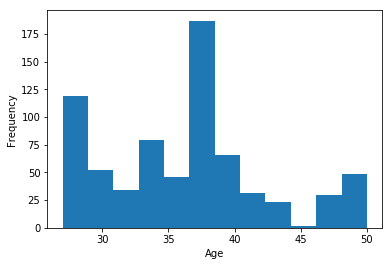
From the above plot we can see that *’Weight’* and *‘Body mass index’* are highly positively correlated. So we will drop ‘Body mass index' between the two.

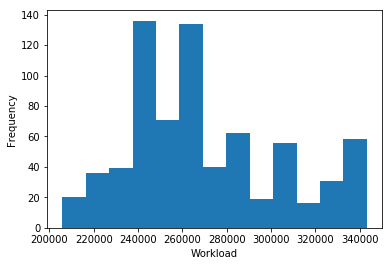
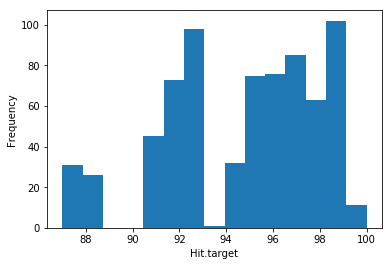
For categorical variable we will do the ANOVA test to see the dependency of the target variable on the independent categorical variables.

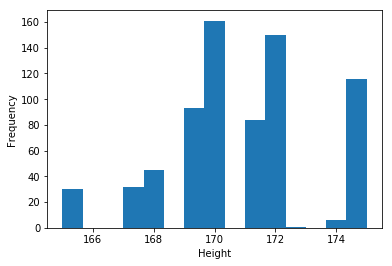
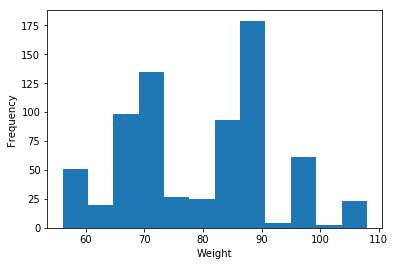
## 2.1.4 Feature Scaling

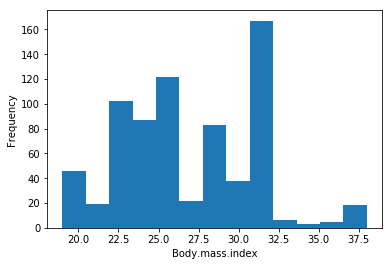
Now we will plot histogram to understand the numerical distribution of the continuous variables.











After looking at the plots we concluded that the continuous variables are not normally distributed while skewness is clearly visible in most of them. Since there is no normally distributed curve we will use Normalization for Feature Scaling.

## 2.1.5 Sampling

As our target variable is continuous we will use the simple random sampling method and create two datasets out of the existing data namely – *train* and *test*. It will be divided so that *train* will contain 80% of the dataset. On *train* dataset we will train the algorithm and *test* dataset we will perform the test.

# 2.2 Modelling

## 2.2.1 Model Selection

After all the pre-processing of the data now we will build our model using different machine learning algorithms. The dependent variable is *continuous*, so the predictive analysis that we can perform is **Regression**. We will use following machine learning algorithms to develop our model.

1. Decision tree – it is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
2. Random Forest – it operates by constructing a multitude of decision trees
3. Linear regression – it is a linear approach to modelling the relationship between a continuous dependent variable and one or more explanatory variables.

Chapter 3

**Conclusion**

# 3.1 Model Evaluation

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following error matrix:

1. MAE – mean absolute error
2. MAPE – mean absolute percentage error
3. MSE – mean square error
4. RMSE – root mean square error

# 3.1 Model Selection

We will compare all the models on different parameters.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter** | **Decision tree** | **Linear regression** | **Random Forest** |
| **MSE** | 307.82 | 370.95 | 258.23 |
| **RMSE** | 17.544 | 19.26 | 16.06 |

MSE = (1/n){(Actual-Predicted)^2}

RMSE = (√MSE)

n=Total number of observations

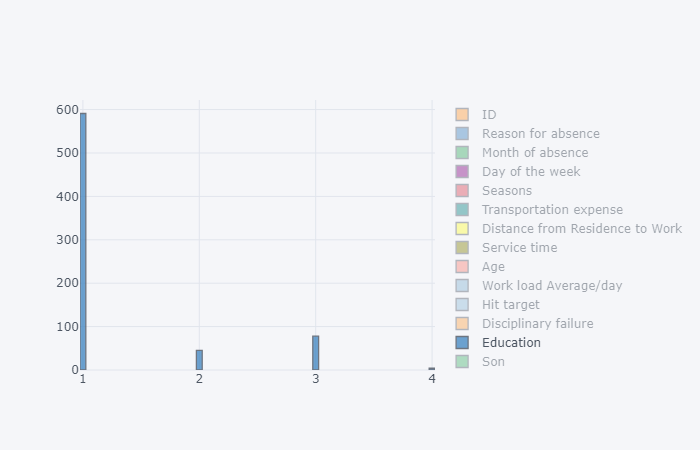
From the observation of all **MSE** and **RMSE Value** we have concluded that **Random Forest Model** has minimum value of RMSE. It gives us values which are nearer to actual values.

# 3.1 Answers

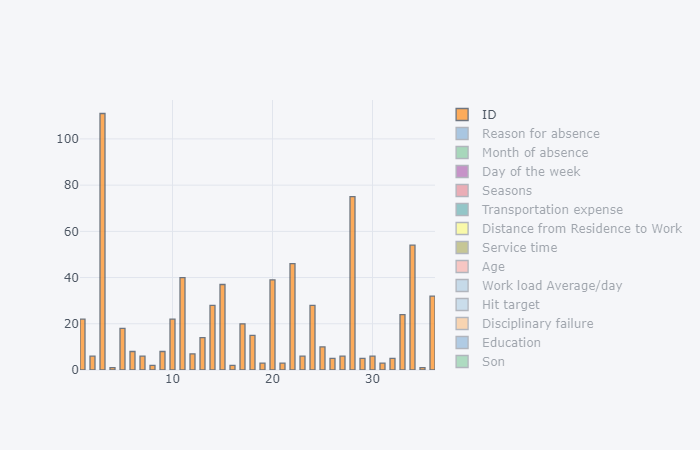
1. What changes company should bring to reduce the number of absenteeism?

The Changes which company should bring to reduce the number of absenteeism –

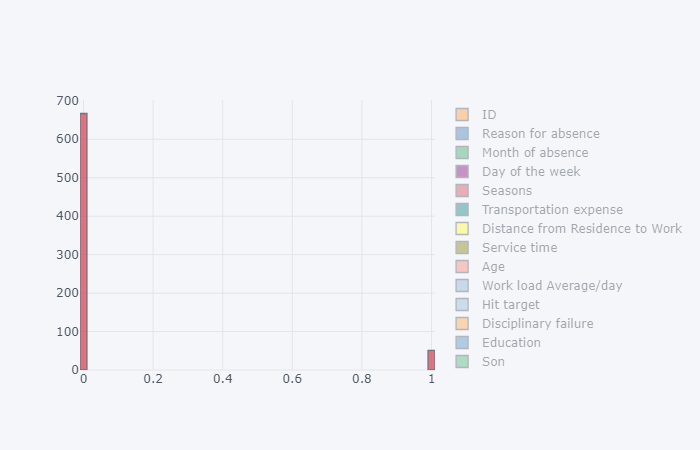
1. It is observed that employee with low education have maximum absentee time.

****

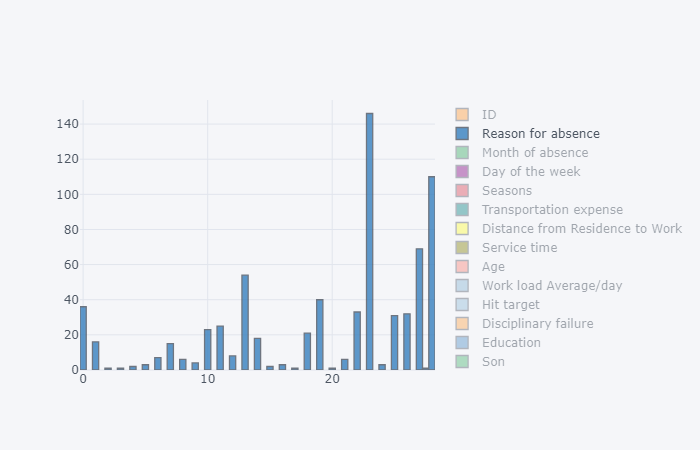
1. Some employee with **ID 3, 28, 34** are often absent from work, company should take action against them.

****

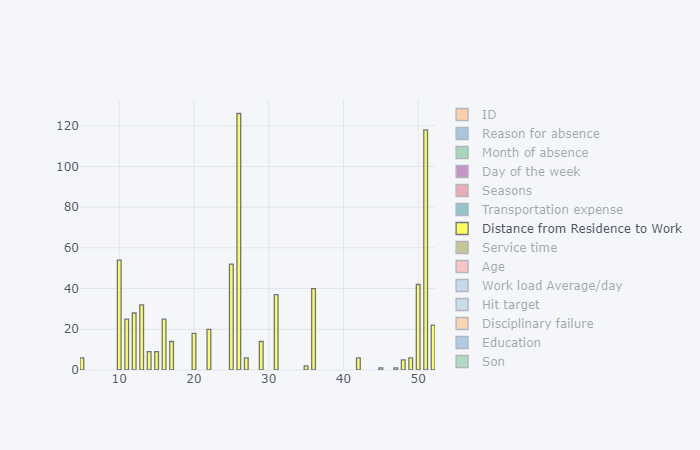
1. Employees who are social smoker have more absentee hour than who are not social smoker.

****

1. Most often Reason for absence are medical consultation and dental consultation, company should take care of it.

****

1. Employees who has Distance from Residence to Work high more tends to absent more.

****

**Appendix A – R codes**

#remove all data

rm(list=ls(all=T))

#set working directory

setwd("G:/edwisor")

#Current working directory

getwd()

#Load Libraries

x = c("ggplot2", "corrgram", "DMwR", "caret", "randomForest", "unbalanced", "C50", "dummies", "e1071", "Information",

"MASS", "rpart", "gbm", "ROSE", 'sampling', 'DataCombine', 'inTrees')

#install.packages(x)

lapply(x, require, character.only = TRUE)

rm(x)

##Load data in R

#reading Excel sheet

library(xlsx)

data = read.xlsx("data.xlsx", sheetIndex = 1, header = T)

#view data

View(data)

#check datatype

class(data)

#summary of data

summary(data)

#column names

colnames(data)

#number of variables

length(unique(data))

#change class from numerical to factor

data$ID = as.factor(data$ID)

data$Month.of.absence = as.factor(data$Month.of.absence)

data$Seasons = as.factor(data$Seasons)

data$Reason.for.absence = as.factor(data$Reason.for.absence)

data$Day.of.the.week = as.factor(data$Day.of.the.week)

data$Disciplinary.failure = as.factor(data$Disciplinary.failure)

data$Education = as.factor(data$Education)

data$Son = as.factor(data$Son)

data$Social.drinker = as.factor(data$Social.drinker)

data$Social.smoker = as.factor(data$Social.smoker)

data$Pet = as.factor(data$Pet)

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

missing\_val$Columns = row.names(missing\_val)

names(missing\_val)[1] = "Missing\_percentage"

missing\_val$Missing\_percentage = (missing\_val$Missing\_percentage/nrow(data)) \* 100

missing\_val = missing\_val[order(-missing\_val$Missing\_percentage),]

row.names(missing\_val) = NULL

missing\_val = missing\_val[,c(2,1)]

# Checking for missing value

sum(is.na(new))

new = data[complete.cases(data[ , 21]),]

#Mean Method

new$Age[is.na(new$Age)] = mean(new$Age, na.rm = T)

#Median Method

new$Age[is.na(new$Age)] = median(new$Age, na.rm = T)

# kNN Imputation

new = knnImputation(new, k = 3)

new[71,9]

#Actual=28

#mean = 36.48815

#median = 37

#knn = 28

new[71,9] = NA

#Box plot

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Age,

fill = new$Age)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("Age") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Distance.from.Residence.to.Work,

fill = new$Distance.from.Residence.to.Work)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("Distance") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Service.time,

fill = new$Service.time)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("time") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Work.load.Average.day.,

fill = new$Work.load.Average.day.)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("workload") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Weight,

fill = new$Weight)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("weight") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Height,

fill = new$Height)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("height") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Body.mass.index,

fill = new$Body.mass.index)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("bmi") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Transportation.expense,

fill = new$Transportation.expense)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("Expense") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Absenteeism.time.in.hours, y = new$Hit.target,

fill = new$Hit.target)) +

geom\_boxplot(outlier.colour = "red", outlier.size = 3) +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

guides(fill=FALSE) + theme\_bw() + xlab("hours") + ylab("Target") +

ggtitle("Outlier Analysis") +

theme(text=element\_text(size=20))

quantile(new$Height, c(.01))

quantile(new$Height, c(.85))

new[which(new$Height<165),("Height")] = 165

new[which(new$Height>175),("Height")] = 175

quantile(new$Work.load.Average.day., c(.96))

new[which(new$Work.load.Average.day.>343253),("Work.load.Average.day.")] = 343253

quantile(new$Age, c(.99))

new[which(new$Age>50),("Age")] = 50

quantile(new$Service.time, c(.99))

new[which(new$Service.time>18),("Service.time")] = 18

quantile(new$Transportation.expense, c(.99))

new[which(new$Transportation.expense>378),("Transportation.expense")] = 378

quantile(new$Hit.target, c(.04))

new[which(new$Hit.target<87),("Hit.target")] = 87

#correlation

#Load Libraries

library(corrgram)

numeric\_index = sapply(new,is.numeric)

numeric\_data = new[,numeric\_index]

corrgram(new[,numeric\_index], order = F,

upper.panel=panel.pie, text.panel=panel.txt, main = "Correlation Plot")

## ANOVA test for Categprical variable

summary(aov(formula = Absenteeism.time.in.hours~ID,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Reason.for.absence,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Month.of.absence,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Day.of.the.week,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Seasons,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Disciplinary.failure,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Education,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Social.drinker,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Social.smoker,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Son,data = new))

summary(aov(formula = Absenteeism.time.in.hours~Pet,data = new))

#Histogram

ggplot(new, aes\_string(x = new$Transportation.expense)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("expense") + ylab("Frequency") + ggtitle('transportation expense') +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Distance.from.Residence.to.Work)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("distance") + ylab("Frequency") + ggtitle('distance from home to office') +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Work.load.Average.day.)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("workload") + ylab("Frequency") + ggtitle('workload') +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Hit.target)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("target") + ylab("Frequency") + ggtitle('Hit target') +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Weight)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("weight") + ylab("Frequency") + ggtitle('weight') +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Height)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("Height") + ylab("Frequency") + ggtitle('Height') +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Body.mass.index)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("BMI") + ylab("Frequency") + ggtitle('BMI') +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Service.time)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("service time") + ylab("Frequency") + ggtitle('service time') +

theme(text=element\_text(size=20))

ggplot(new, aes\_string(x = new$Age)) +

geom\_histogram(fill="cornsilk", colour = "black") + geom\_density() +

scale\_y\_continuous(breaks=pretty\_breaks(n=10)) +

scale\_x\_continuous(breaks=pretty\_breaks(n=10))+

theme\_bw() + xlab("Age") + ylab("Frequency") + ggtitle('Age') +

theme(text=element\_text(size=20))

#Normalisation

cnames = c('Distance.from.Residence.to.Work', 'Service.time', 'Age',

'Work.load.Average.day.', 'Transportation.expense',

'Hit.target', 'Height', 'Weight',

'Body.mass.index')

for(i in cnames){

print(i)

new[,i] = (new[,i] - min(new[,i]))/

(max(new[,i] - min(new[,i])))

}

copy=new

new=copy

# dummify the data

# Creating dummy variables for categorical variables

library(mlr)

catagorical = c('ID', 'Reason.for.absence', 'Month.of.absence',

'Seasons', 'Day.of.the.week',

'Disciplinary.failure', 'Education', 'Son',

'Social.drinker','Social.smoker','Pet')

new = dummy.data.frame(new, catagorical)

new = subset(new,

select = -c(ID36,Reason.for.absence0,Month.of.absence0,Day.of.the.week2,Seasons4,Education4,Disciplinary.failure1,Son4,Social.drinker0,Social.smoker0,Pet8))

#check multicollearity

library(usdm)

vif(new[,-106])

vifcor(new[,-106], th = 0.9)

new = subset(new,

select = -c(Son3,Pet5,Weight)

#Decision treemodel

#Load Libraries

library(rpart)

#Divide the data into train and test

#set.seed(123)

train\_index = sample(1:nrow(new), 0.8 \* nrow(new))

train = new[train\_index,]

test = new[-train\_index,]

# ##rpart for regression

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

#Predict for new test cases

predictions\_DT = predict(fit, test[,-106])

#calculate MAPE

MAPE = function(y, yhat){

mean(abs((y - yhat)/y))

}

MAPE(test[,106], predictions\_DT)

#calculate MSE

MSE = function(m, o){

(mean((m - o)^2))

}

MSE(test[,106], predictions\_DT)

#calculate RMSE

RMSE = function(m, o){

sqrt(mean((m - o)^2))

}

RMSE(test[,106], predictions\_DT)

#Linear Regression

#run regression model

lm\_model = lm(Absenteeism.time.in.hours ~., data = train)

#Summary of the model

summary(lm\_model)

#Predict

predictions\_LR = predict(lm\_model, test[,1:105])

#Calculate MAPE,MSE,RMSE

MAPE(test[,106], predictions\_LR)

MSE(test[,106], predictions\_LR)

RMSE(test[,106], predictions\_LR)

#Randomforest

RF\_model = randomForest(Absenteeism.time.in.hours ~ ., train, importance = TRUE, ntree = 100)

#transform rf object to an inTrees' format

treeList = RF2List(RF\_model)

#Extract rules

exec = extractRules(treeList, train[,-106]) # R-executable conditions

exec[1:2,]

# #Make rules more readable:

readableRules = presentRules(exec, colnames(train))

readableRules[1:2,]

ruleMetric = getRuleMetric(exec, train[,-106], train$cnt) # get rule metrics

#

# #evaulate few rules

ruleMetric[1:2,]

#Presdict test data using random forest model

RF\_Predictions = predict(RF\_model, test[,-106])

#Calculate MAPE,MSE,RMSE

MAPE(test[,106], RF\_Predictions)

MSE(test[,106], RF\_Predictions)

RMSE(test[,106], RF\_Predictions)

**References**

<https://edwisor.com/>

<https://www.analyticsvidhya.com/>

<http://www.statisticssolutions.com/>

<https://en.wikipedia.org/wiki/>

<https://towardsdatascience.com/>

<https://github.com/>

<https://www.datacamp.com/>

<https://stackoverflow.com/>

<https://www.khanacademy.org/>

<https://www.listendata.com/>

<https://www.hackerearth.com/>

<https://stackabuse.com/>