**PROJECT REPORT EMPLOYEE ABSENTEEISM Shashank Paliwal**

**01-07-2019**

# Contents

1. Introduction
   1. [Problem Statement](#_bookmark0) 3
   2. [Variables](#_bookmark1) 3
   3. [Sample Data](#_bookmark2) 4
   4. [Unique Count](#_bookmark3) 5
2. Methodology
   1. [Pre – Processing](#_bookmark4) 6
   2. [Missing Value Analysis](#_bookmark5) 6
   3. [Outlier Analysis](#_bookmark6) 7
   4. [Distribution of Continuous variables](#_bookmark7) 9
   5. [Distribution of Categorical variables](#_bookmark8) 10
   6. [Feature Selection](#_bookmark9) 12
   7. [Feature Scaling](#_bookmark10) 12
   8. [Principal Component Analysis (PCA)](#_bookmark11) 13
3. Modelling
   1. [Model Selection](#_bookmark12) 14
   2. [Decision Tree](#_bookmark13) 14
   3. [Random Forest](#_bookmark14) 15
   4. [Linear Regression](#_bookmark15) 15
4. Conclusion
   1. [Model Evaluation](#_bookmark16) 17
   2. [Model Selection](#_bookmark17) 17
   3. [Solution of Problem Statement](#_bookmark18) 17
5. Appendix
   1. [Figures](#_bookmark19) 22
6. [R code](#_bookmark20) 28

# Chapter 1: Introduction

## 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?

## Variables

There are 21 variables in our data in which 20 are independent variables and 1 (Absenteeism time in hours) is dependent variable. Since the type of target variable is continuous, this is a regression problem.

Variable Information:

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

1. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
2. 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

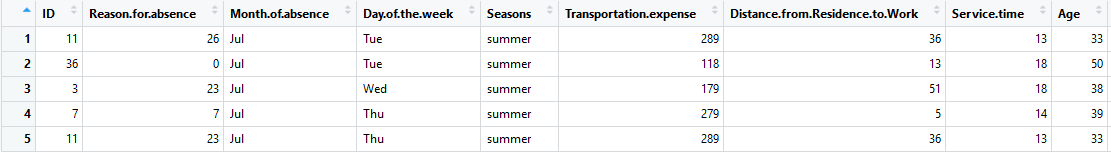
1. Diseases of the skin and subcutaneous tissue
2. Diseases of the musculoskeletal system and connective tissue XIV Diseases of the genitourinary system
3. Pregnancy, childbirth and the puerperium
4. Certain conditions originating in the perinatal period
5. Congenital malformations, deformations and chromosomal abnormalities
6. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified XIX Injury, poisoning and certain other consequences of external causes
7. External causes of morbidity and mortality
8. Factors influencing health status and contact with health services.

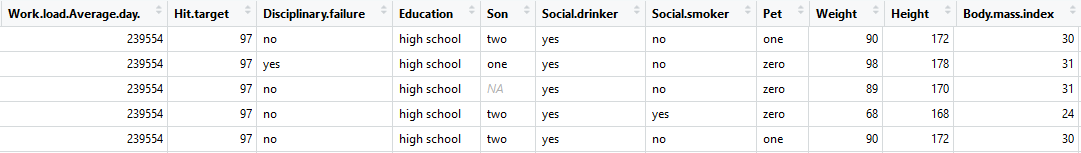
And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation

(28).

1. Month of absence
2. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
3. Seasons (summer (1), autumn (2), winter (3), spring (4))
4. Transportation expense
5. Distance from Residence to Work (KMs)
6. Service time
7. Age
8. Work load Average/day
9. Hit target
10. Disciplinary failure (yes=1; no=0)
11. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
12. Son (number of children)
13. Social drinker (yes=1; no=0)
14. Social smoker (yes=1; no=0)
15. Pet (number of pet)
16. Weight
17. Height
18. Body mass index
19. Absenteeism time in hours (target)

## Sample Data







### Fig 1.3 – First five rows of data

## Unique count

Below figure shows the unique count of all the variables present in the data.



### Fig 1.4 – Unique Count of data

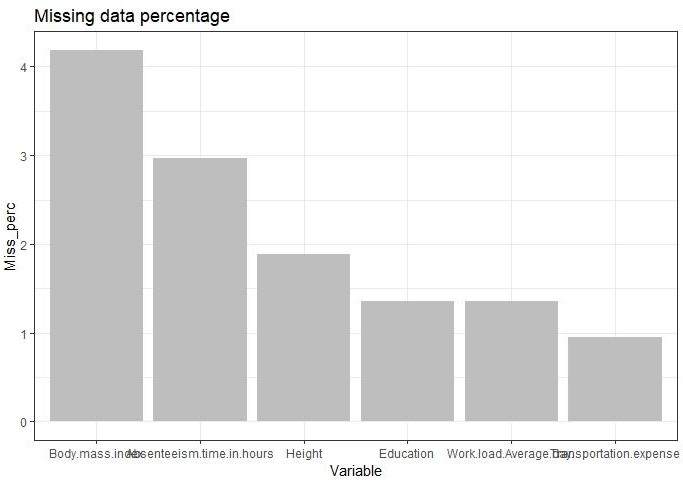
**Chapter 2: Methodology**

## Pre – Processing

A predictive model requires that we look at the data before we start to create a model. However, in data mining, looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is known as Exploratory Data Analysis. In this project we look at the distribution of categorical variables and continuous variables. We also look at the missing values in the data and the outliers present in the data.

## Missing Value Analysis

In statistics, missing data or missing values occur when no data value is stored for the variable in an observation. Missing values are a common occurrence in data analysis. These values can have a significant impact on the results or conclusions that would be drawn from these data. If a variable has more than 30% of its values missing, then those values can be ignored, or the column itself is ignored. In our case, none of the columns have a high percentage of missing values. The maximum missing percentage is 4.18% i.e., Body Mass Index column. The missing values have been computed using KNN computation method.

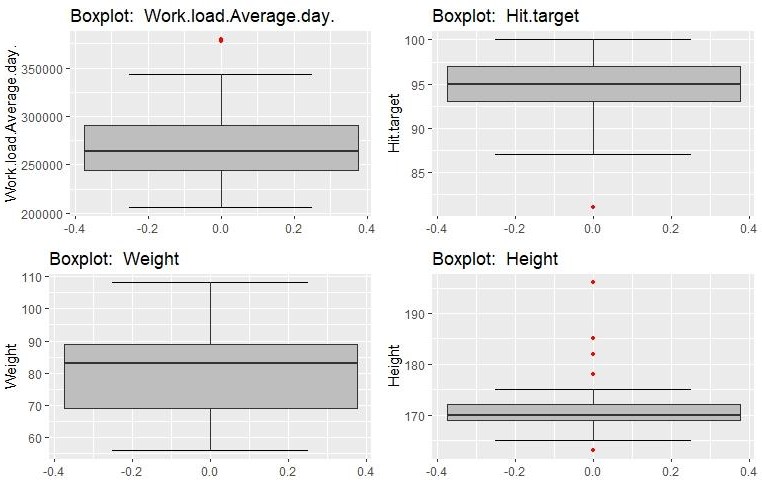
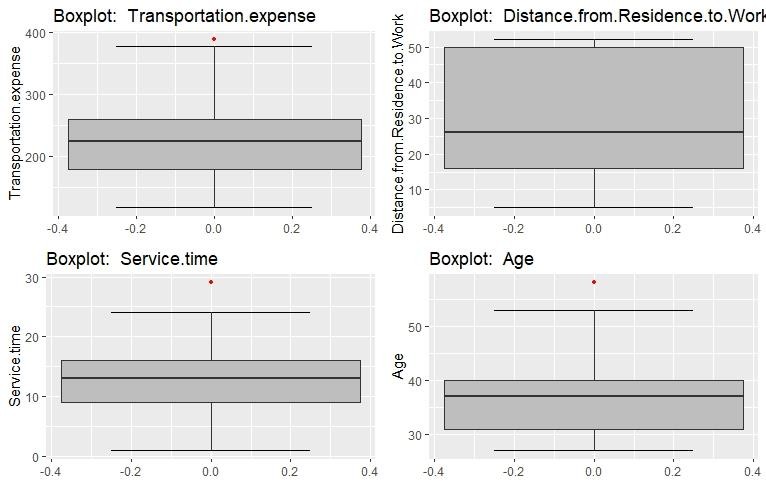


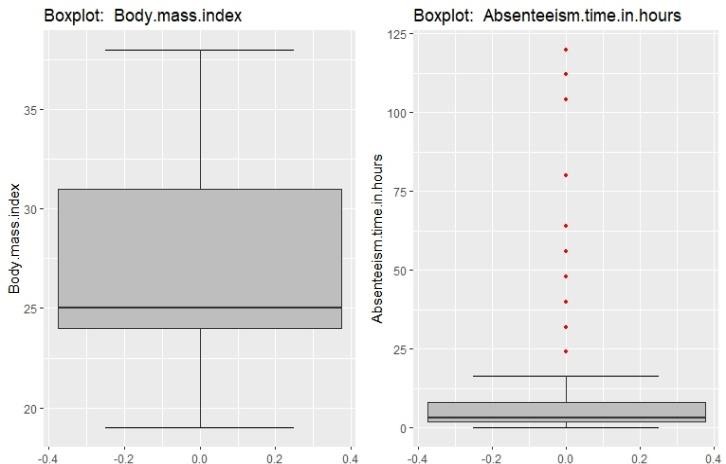
### Fig 2.2 – Missing value Percentage

## Outlier Analysis

It can be observed from the distribution of variables that almost none of the variables are normally distributed. The skew in these distributions can be explained by the presence of outliers and extreme values in the data. One of the steps in pre-processing involves the detection and removal of such outliers. In this project, we use boxplot to visualize and remove outliers. Any value lying outside of the lower and upper whisker of the boxplot are outliers.

Variables excluding Distance from residence to work, Weight and Body mass index, contain outliers.



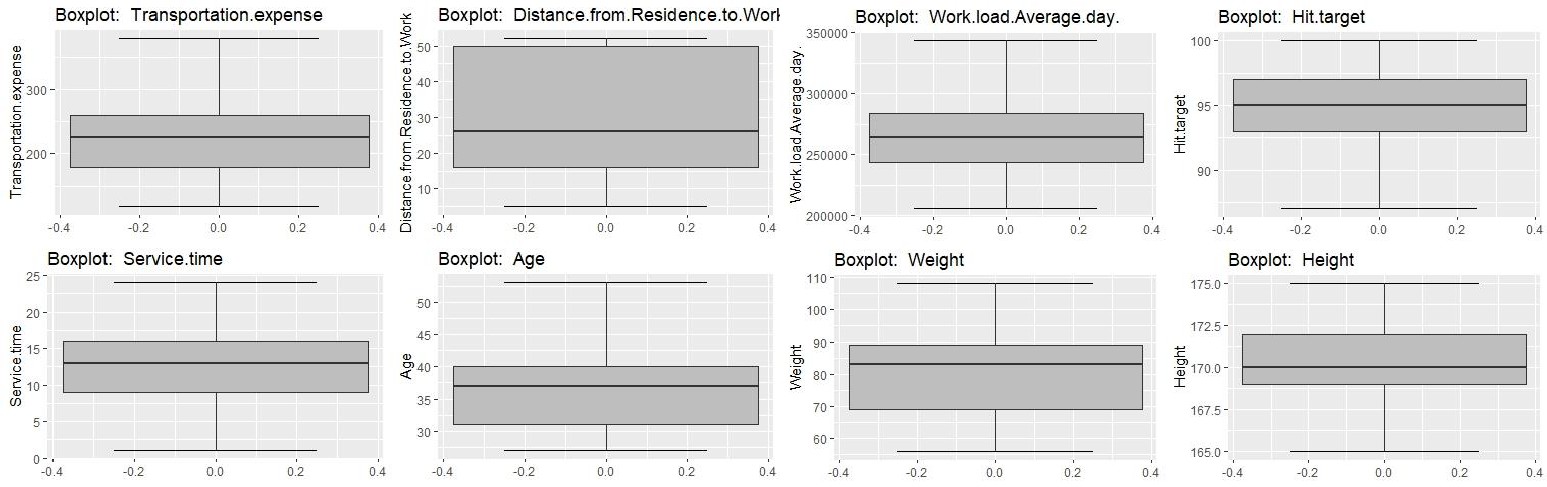


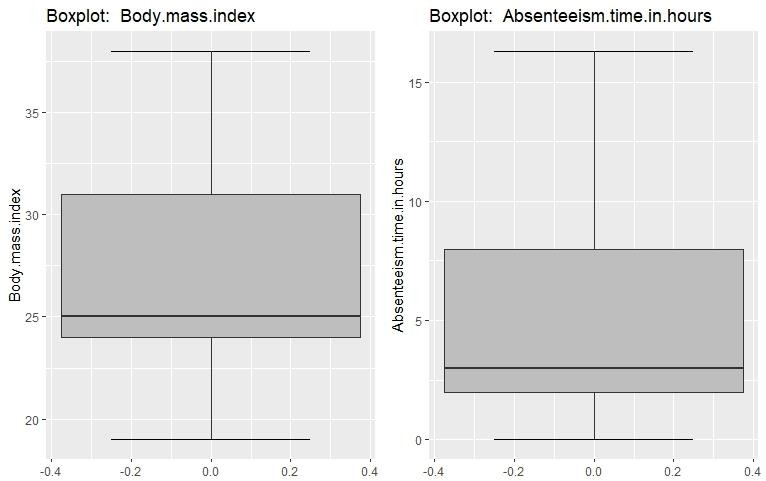
### Fig 2.3.1 – Boxplots of continuous variables with outliers

Imputing outlier values:

Missing values obtained from boxplots are first converted to have NA values. Then these missing values are imputed using KNN imputation method.

Below figure shows the boxplots of variables after removing outliers.

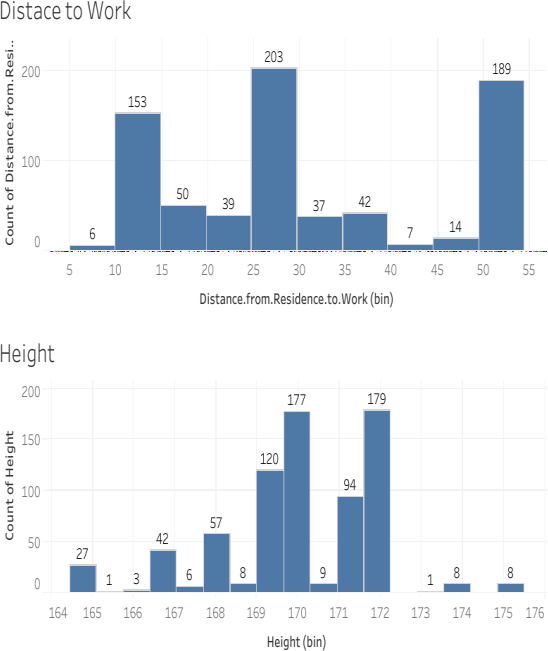
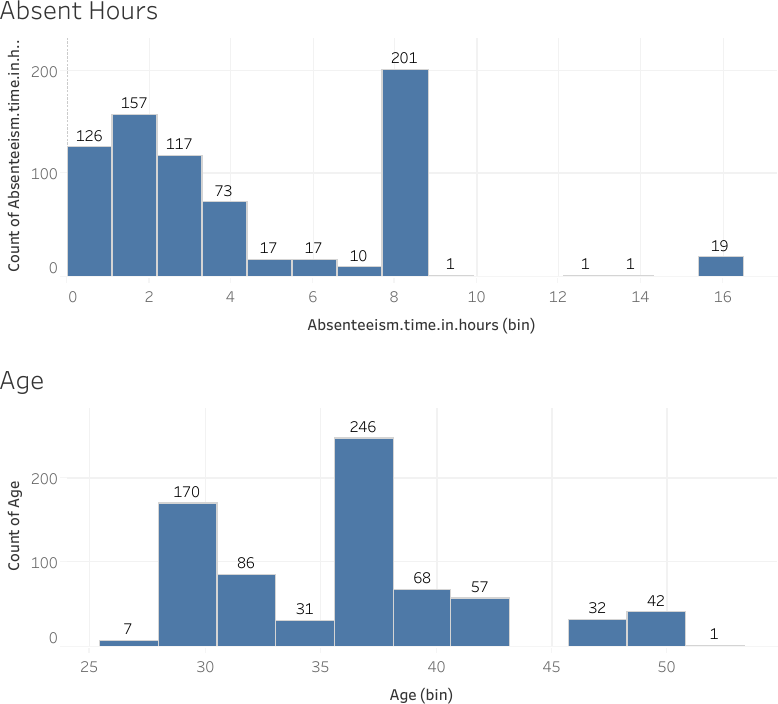


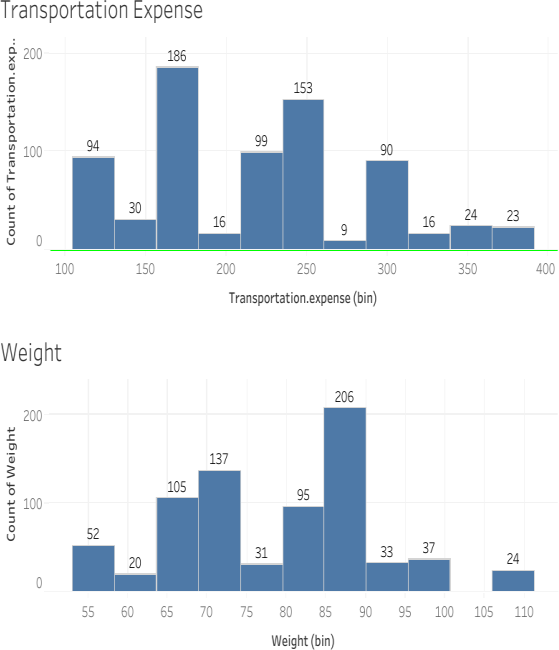
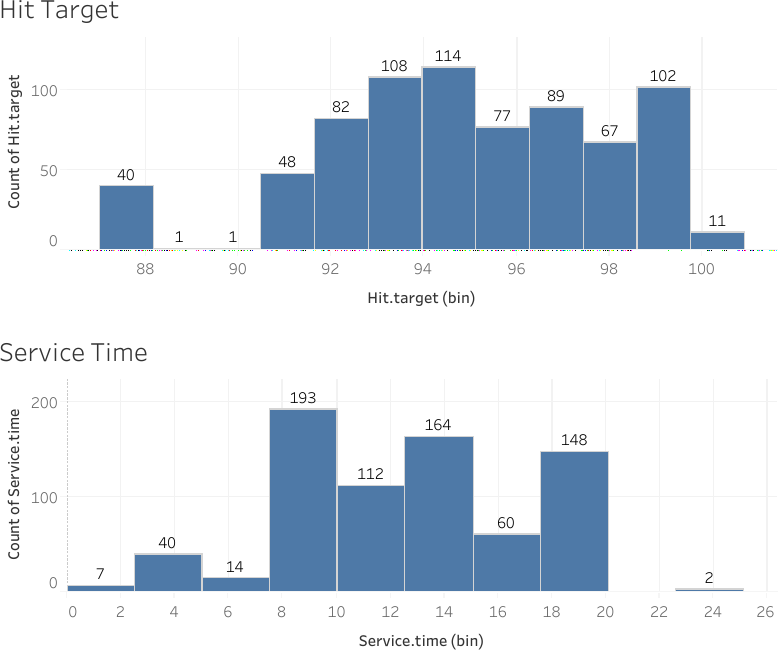


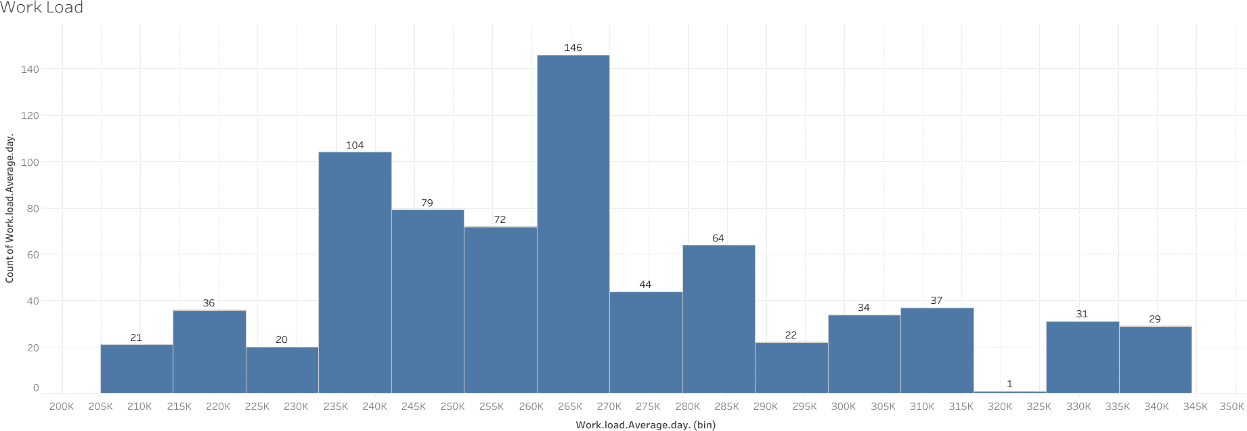
### Fig 2.3.2 – Boxplots of continuous variables without outliers

## Distribution of Continuous variables

By looking at the distribution of continuous variables, it can be observed that the variables are not normally distributed. Histograms are used to observe the distribution of continuous variables.







### Fig 2.4 – Distribution of Continuous variables using Histogram

## Distribution of Categorical Variables

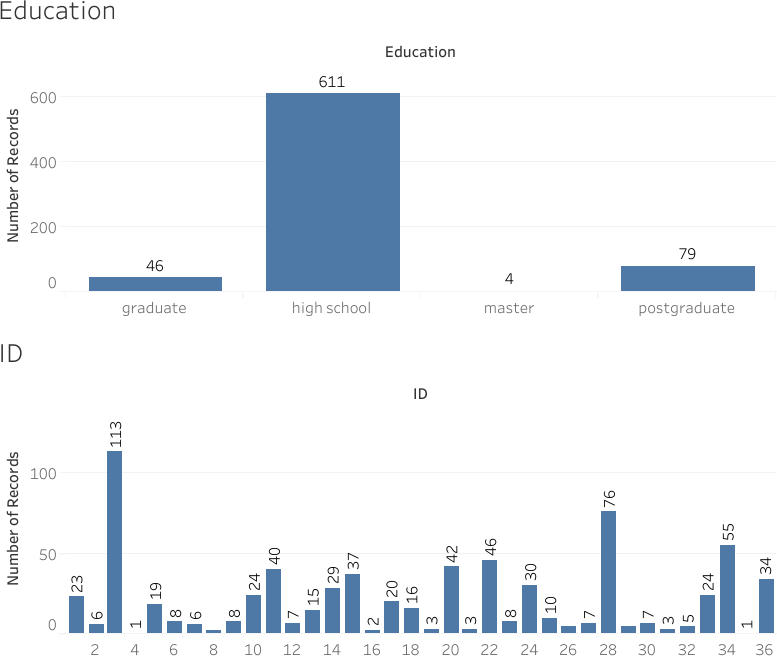
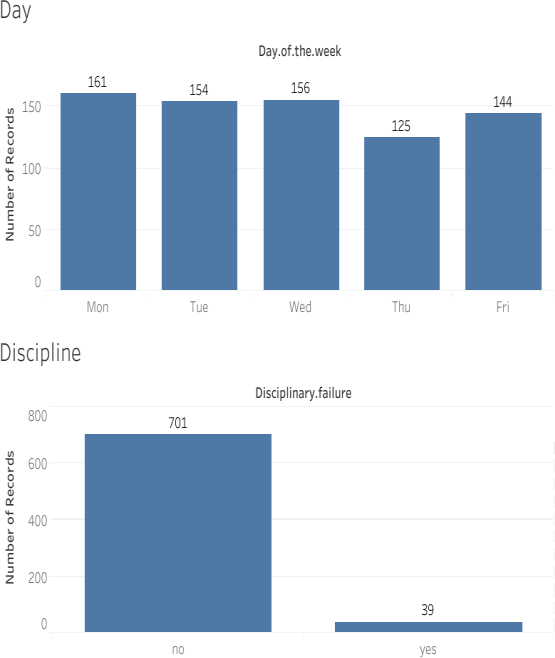
Bar graphs are used to visualize the distribution of categorical variables.

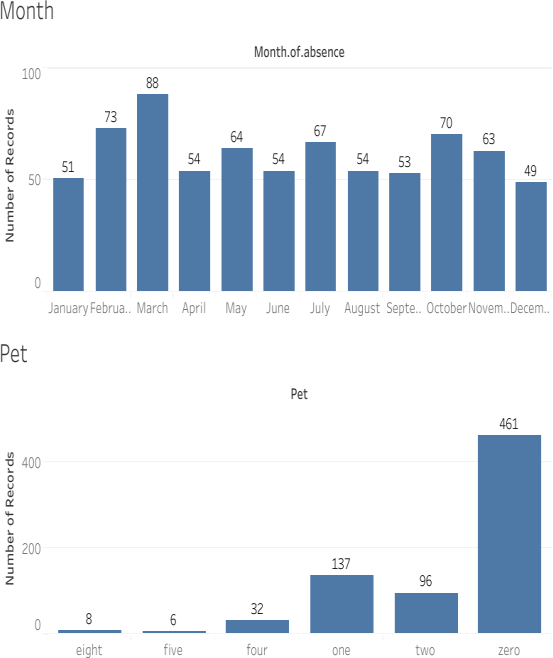
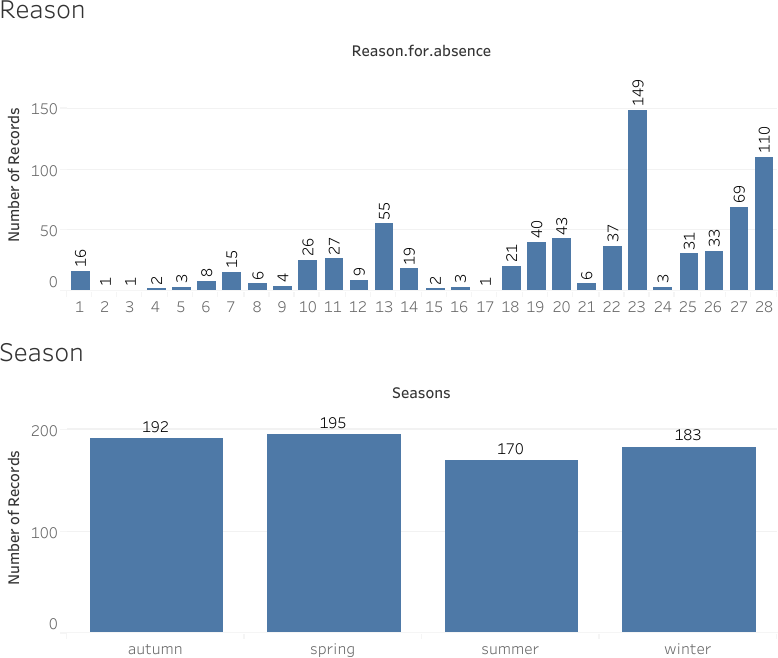
Employees who are social drinkers have more absent hours than those who do not drink. Employees having zero, one or two children have more absent hours.

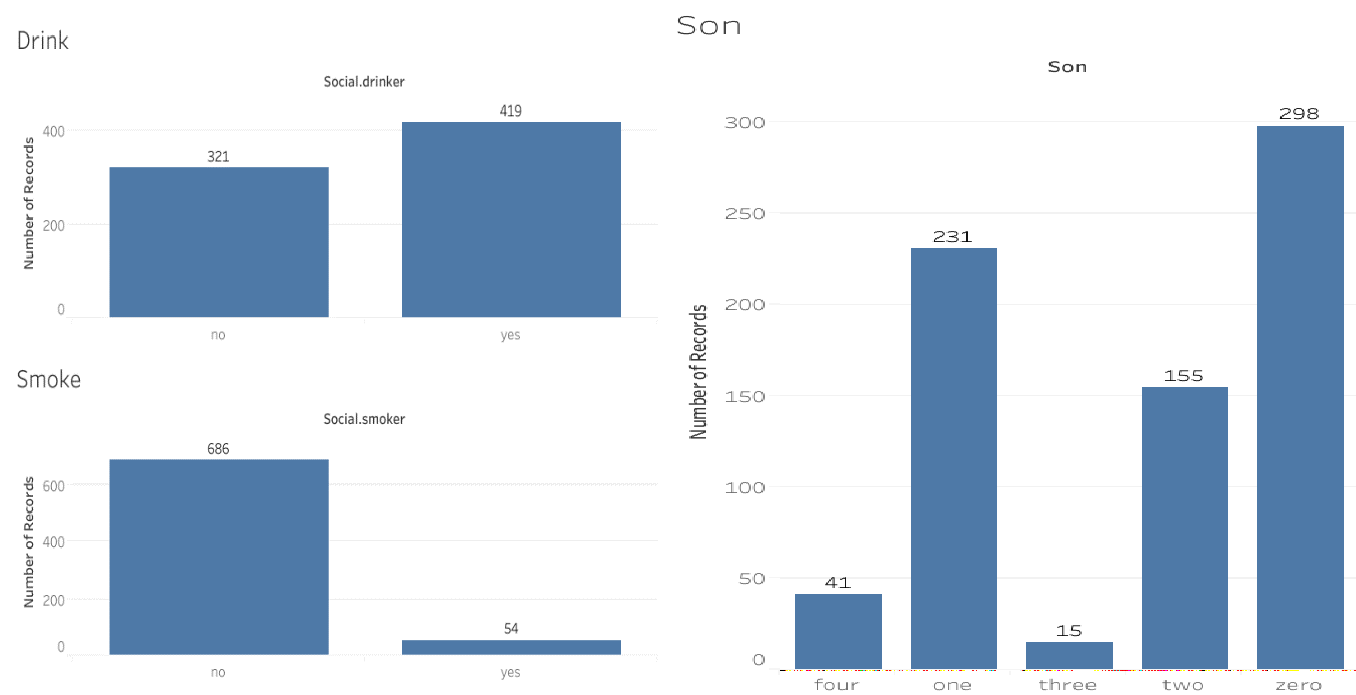
Employees with ID number 3 and 28 are absent the most.

Employees are absent the most on Mondays and the least on Thursdays. Reason 23 and 28 are the reasons employee give the most for being absent.

Employees who have completed only high school education are absent more than others. Employees are absent the most in the month of March.



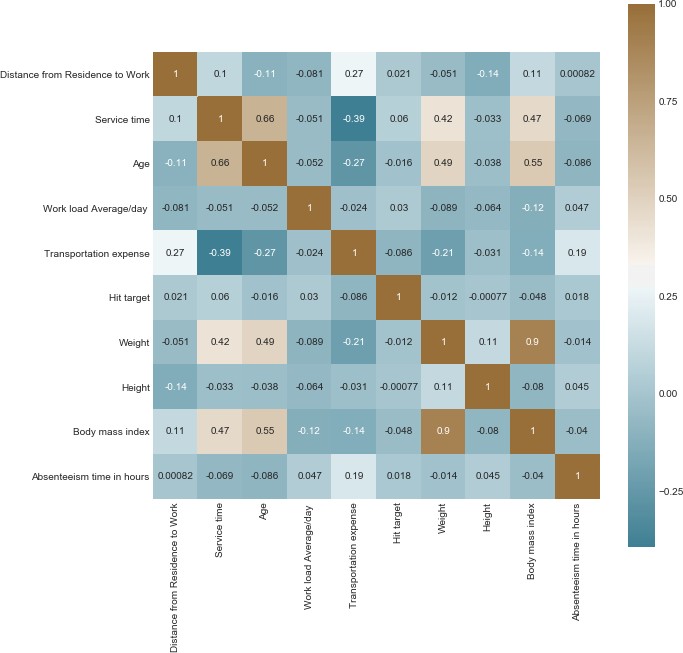


### Fig 2.5 – Distribution of Categorical variables using Bar graph

## Feature Selection

Feature Selection reduces the complexity of a model and makes it easier to interpret. It also reduces overfitting. Features are selected based on their scores in various statistical tests for their correlation with the outcome variable. Correlation plot is used to find out if there is any multicollinearity between variables. The highly collinear variables are dropped and then the model is executed.

From correlation analysis we have found that Weight and Body Mass Index has high correlation (>0.7), so we have excluded the Body Mass Index column.



### Fig 2.6 – Correlation plot of Continuous variables

## Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

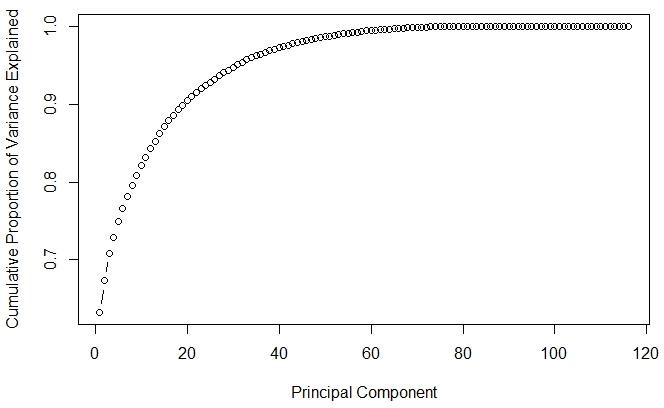
Most classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this feature. Therefore, the range of all features should be normalized so that each feature contributes proportionately to the

final distance. Since our data is not uniformly distributed, we will use Normalization as Feature Scaling Method.

## Principal Component Analysis (PCA)

Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible.

After creating dummy variable of categorical variables, the data would have 116 columns and 740 observations. This high number of columns leads to bad accuracy.



### Fig 2.8 – Cumulative Scree Plot of Principal Components

After applying PCA algorithm and observing the above Cumulative Scree Plot, it can be observed that almost 95% of the data can be explained by 45 variables out of 116. Hence, we choose only 45 variables as input to the models.

# Chapter 3: Modelling

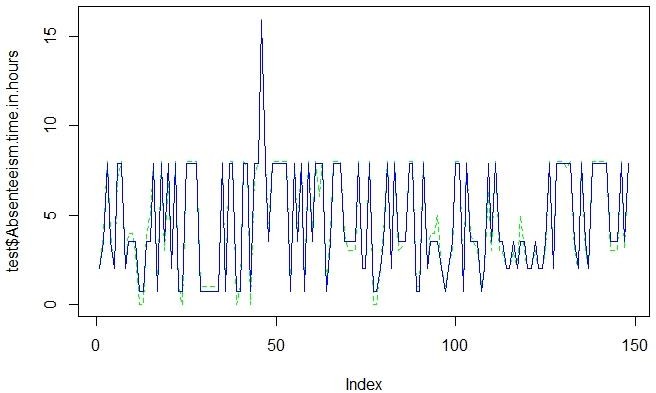
## Model Selection

After a thorough pre-processing we will be using some regression models on our processed data to predict the target variable. The target variable in our model is a continuous variable i.e., Absenteeism time in hours. Hence the models that we choose are Linear Regression, Decision Tree and Random Forest. The error metric chosen for the given problem statement is Root Mean Square Error (RMSE).

## Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Decision trees are used for both classification and regression problems.

A decision tree is a tree where each node represents a feature(attribute), each link(branch) represents a decision(rule) and each leaf represents an outcome (categorical or continues value). The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data).



### Fig 3.2 – Plot of actual values vs predicted values for Decision Tree

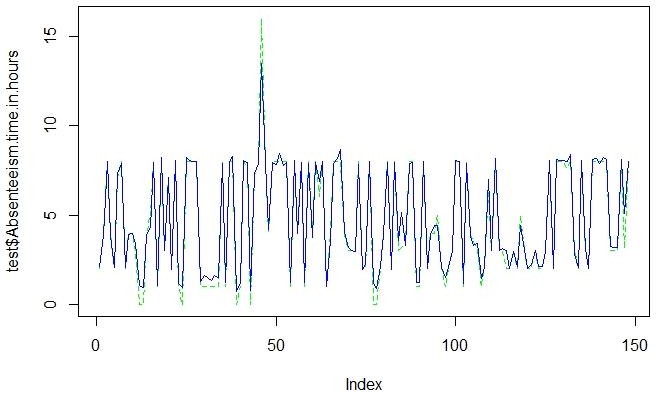
The RMSE values and R^2 values for the given project in R and Python are:

|  |  |  |
| --- | --- | --- |
| **DECISION TREE** | RMSE | R^2 |
| R | 0.442 | 0.978 |
| PYTHON | 0.0353 | 0.9998 |

## Random Forest

Random Forest is a supervised learning algorithm. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. It can be used for both classification and regression problems. The method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner.

The number of decision trees used for prediction in the forest is 500.



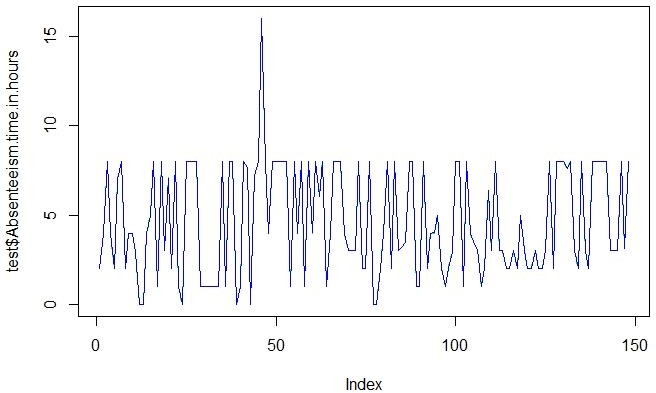
### Fig 3.3 – Plot of actual values vs predicted values for Random Forest

|  |  |  |
| --- | --- | --- |
| **RANDOM FOREST** | RMSE | R^2 |
| R | 0.480 | 0.978 |
| PYTHON | 0.0445 | 0.9998 |

## Linear Regression

Multiple linear regression is the most common form of linear regression analysis. Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

|  |  |  |
| --- | --- | --- |
| **LINEAR REGRESSION** | RMSE | R^2 |
| R | 0.003 | 0.9999 |
| PYTHON | 0.0013 | 0.9999 |



### Fig 3.5 – Plot of actual values vs predicted values for Linear Regression

**Chapter 4: Conclusion**

## Model Evaluation

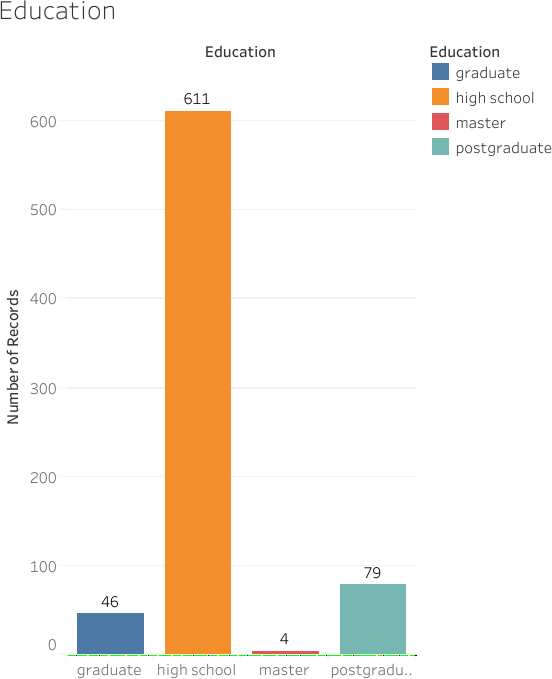
In the previous chapter we have seen the Root Mean Square Error (RMSE) and R-Squared Value of different models. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable. Lower values of RMSE and higher value of R-Squared Value indicate better fit.

## Model Selection

From the observation of all RMSE Value and R-Squared Value we have concluded that **Linear Regression Model** has minimum value of RMSE and its R-Squared Value is also maximum.

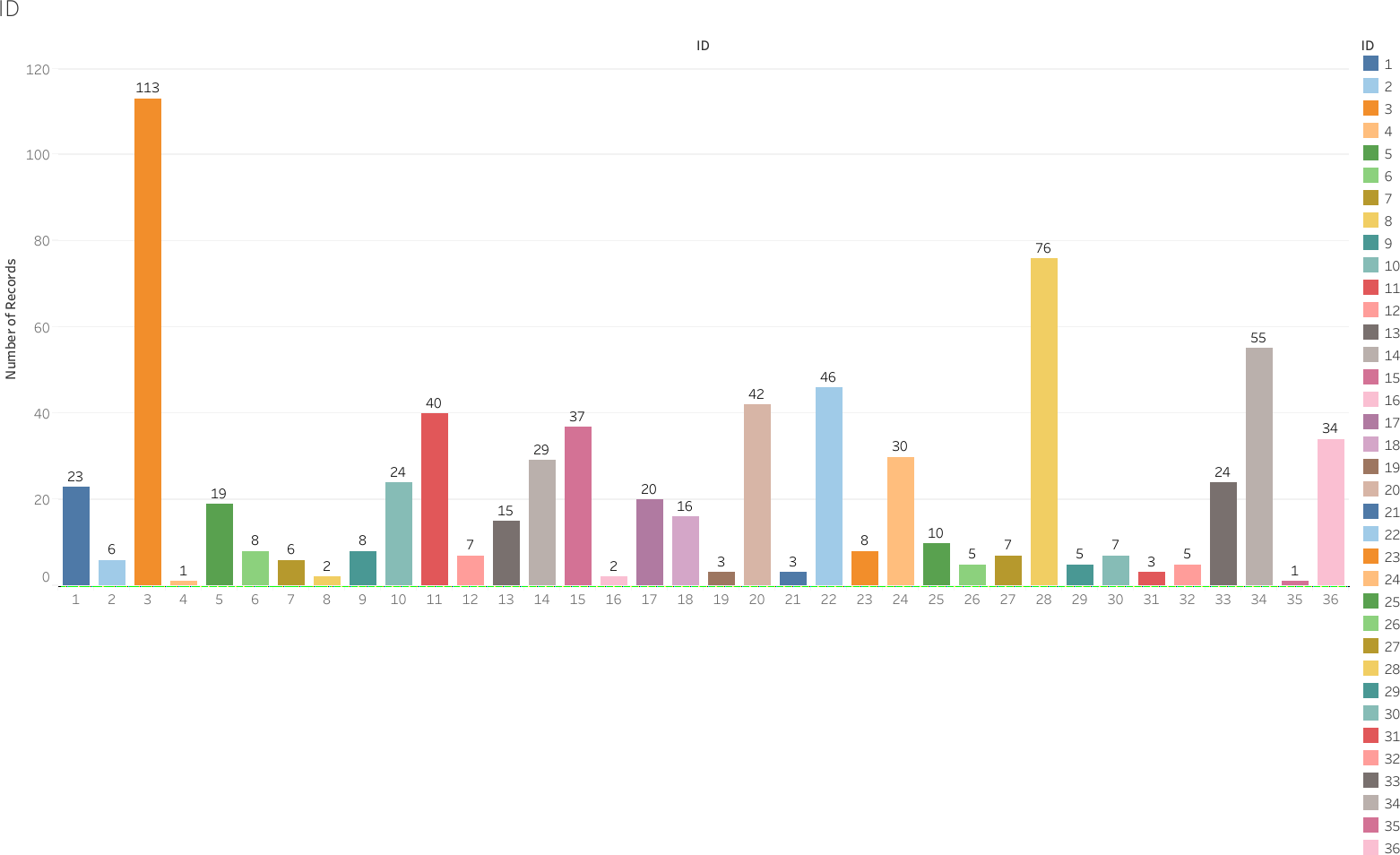
## Solutions of Problem Statement

* + 1. What changes company should bring to reduce the number of absenteeism? Solution:
       1. It can be observed that employees having education only till high school tend to be absent more than others. So, the company can either hire employees who have at least graduated from college or inform those employees who have completed only their high school education to reduce the number of hours they are absent.

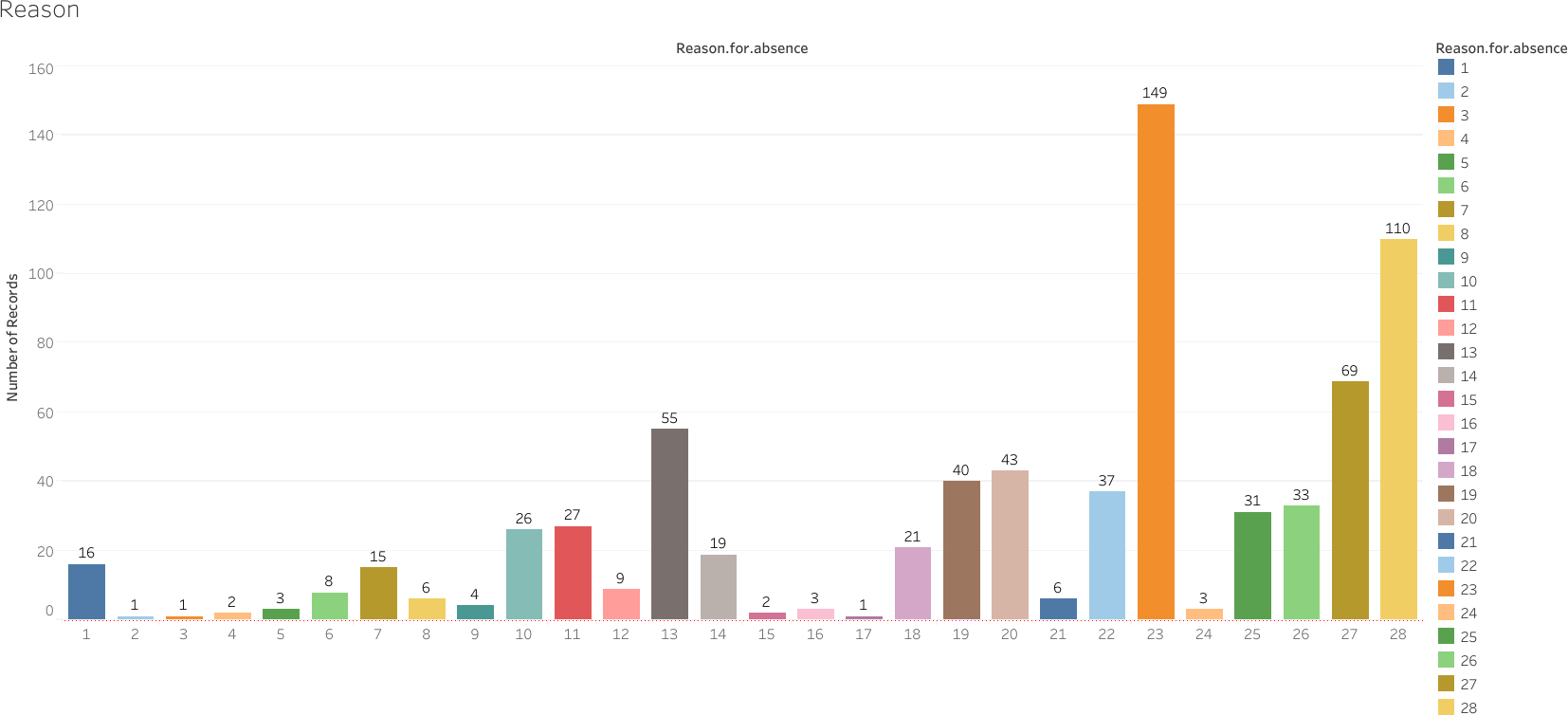


### Fig 4.3.1 – Plot of Education vs Absent Hours

* + - 1. Employees with ID 3, 28 and 34 are some of the employees who are absent the most. The company may act warn such employees to reduce being absent a lot or if repeated further, it can against them if necessary.

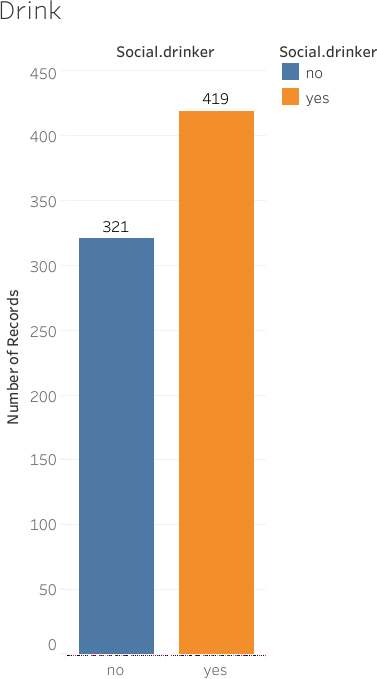


### Fig 4.3.2 – Plot of ID vs Absent Hours

* + - 1. The reasons most used by employees to be absent are reason 13, 20, 23 and 28. These reasons include Medical consultation, Dental appointments, morbidity, mortality and diseases of musculoskeletal system and connective tissue. The company XYZ can help in informing employees on how to keep themselves healthier by having monthly campus consultations.

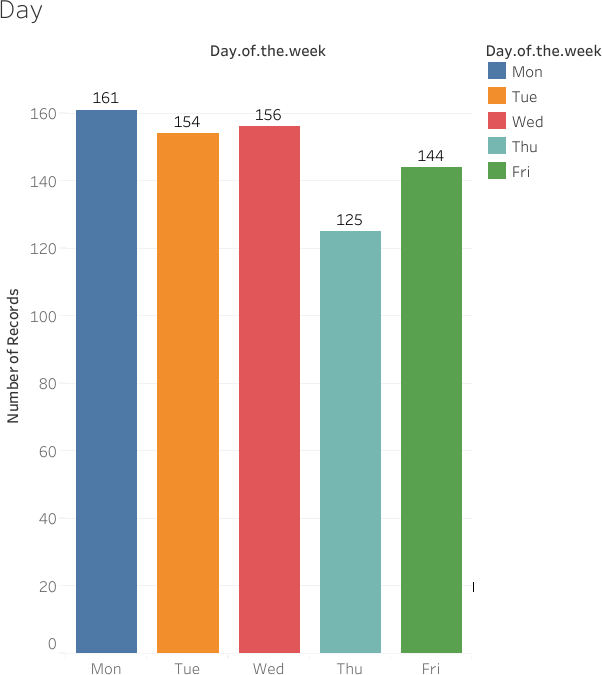
### Fig 4.3.3 – Plot of Reason of Absence vs Absent Hours

* + - 1. People who tend to be social drinkers tend to be more absent than those who don’t drink. XYZ can keep a track of those people and inform those employees to reduce the intake of alcohol during working days.



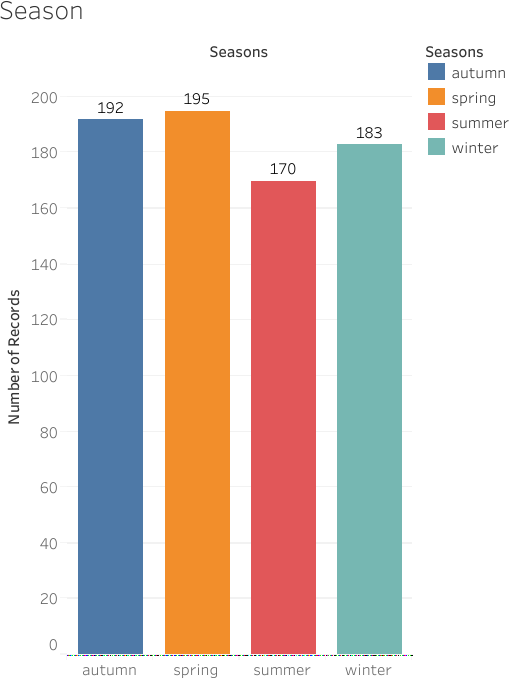
### Fig 4.3.4 – Plot of Social Drinker vs Absent Hours

* + - 1. Employees are absent the most on Mondays with absent hours equal to 1426 and Tuesdays with absent hours equal to 1322.4. XYZ can inform employees to not take as many absent hours on such days.



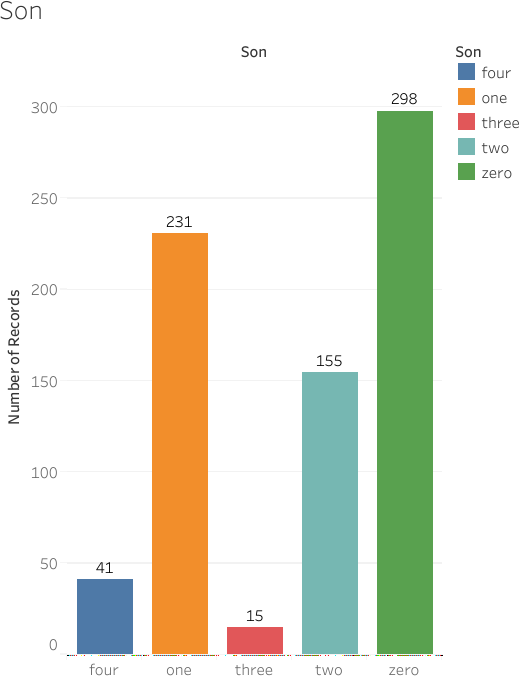
### Fig 4.3.5 – Plot of Day of the Week vs Absent Hours

* + - 1. Employees are mostly absent during Spring Season.



### Fig 4.3.6 – Plot of Season vs Absent Hours

* + - 1. Employees having a maximum of two children or no child at all are absent the most.



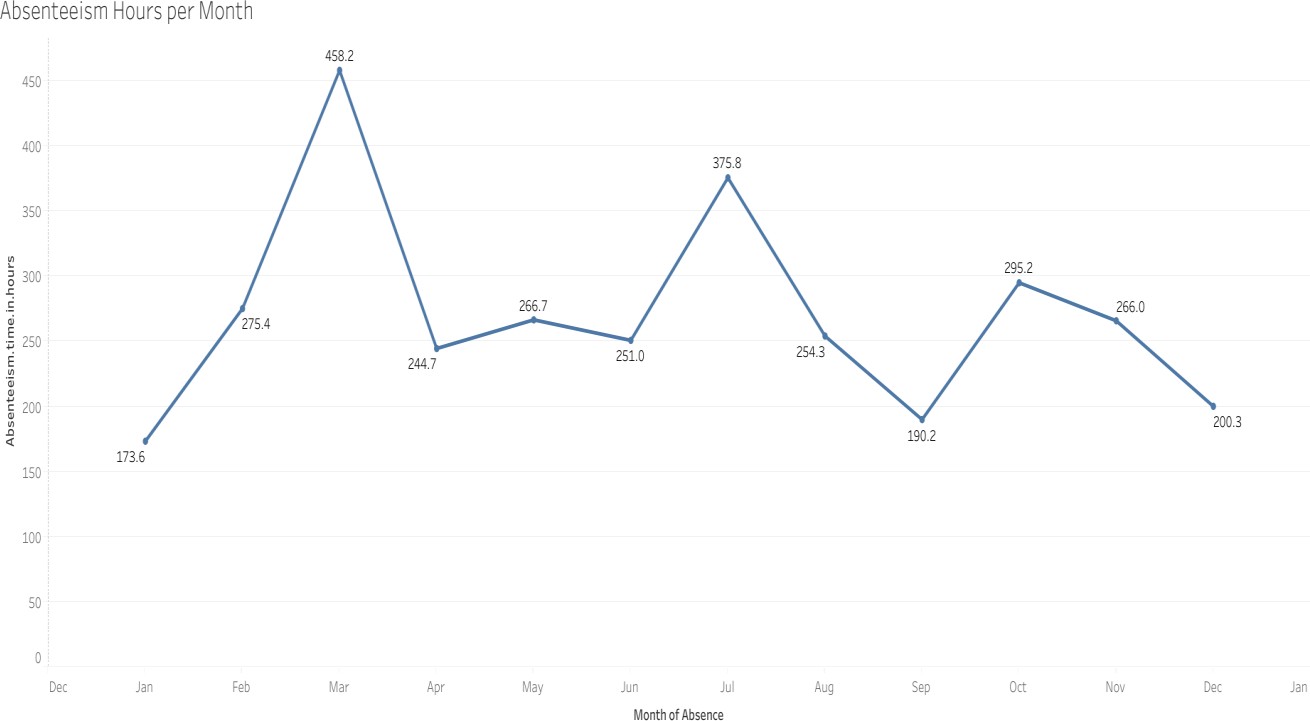
### Fig: 4.3.7 – Plot of Sons vs Absent Hours

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

Solution:

Considering the losses to be the absenteeism time in hours, if the same trend of absenteeism continues, then the total total losses per month is as shown in the graph below.

Employees are absent the most in the month of March, with total Absenteeism hours equal to 458.2 hours. Employees are absent the least in the month of January, with total Absenteeism hours equal to 173.6.



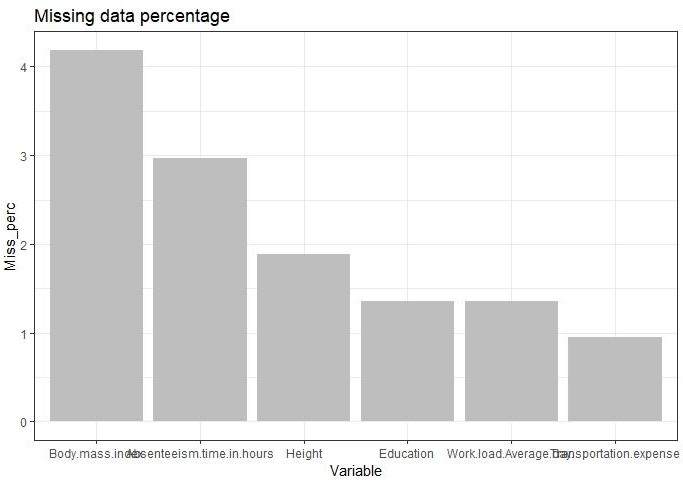
### Fig 4.3.2 – Absenteeism Hours per Month

Below table shows the monthly losses of absenteeism hours:

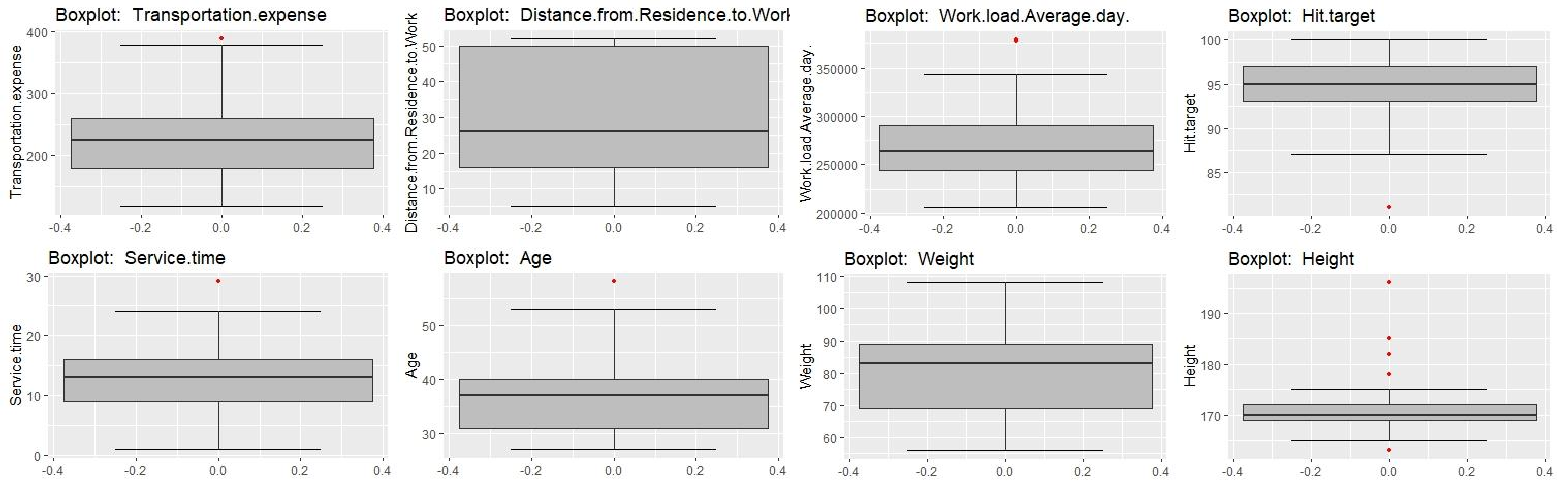
|  |  |
| --- | --- |
| **Month** | **Absent Hours** |
| January | 173.6 |
| February | 275.4 |
| March | 458.2 |
| April | 244.7 |
| May | 266.7 |
| June | 251 |
| July | 375.8 |
| August | 254.3 |
| September | 190.2 |
| October | 295.2 |
| November | 266 |
| December | 200.3 |

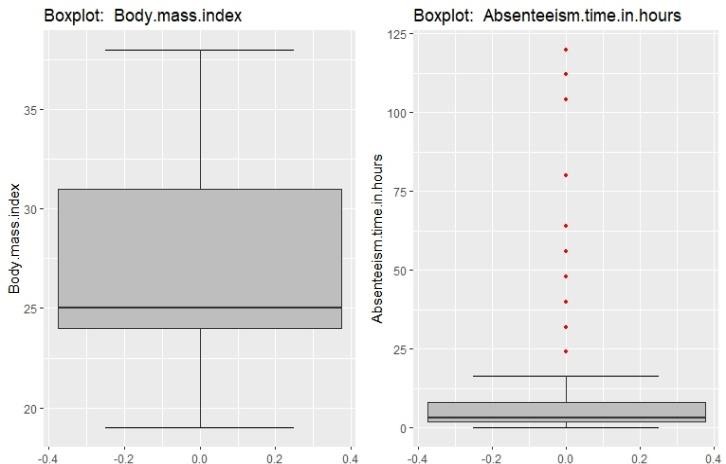
# Chapter 5: Appendix

## 5.1 Figures

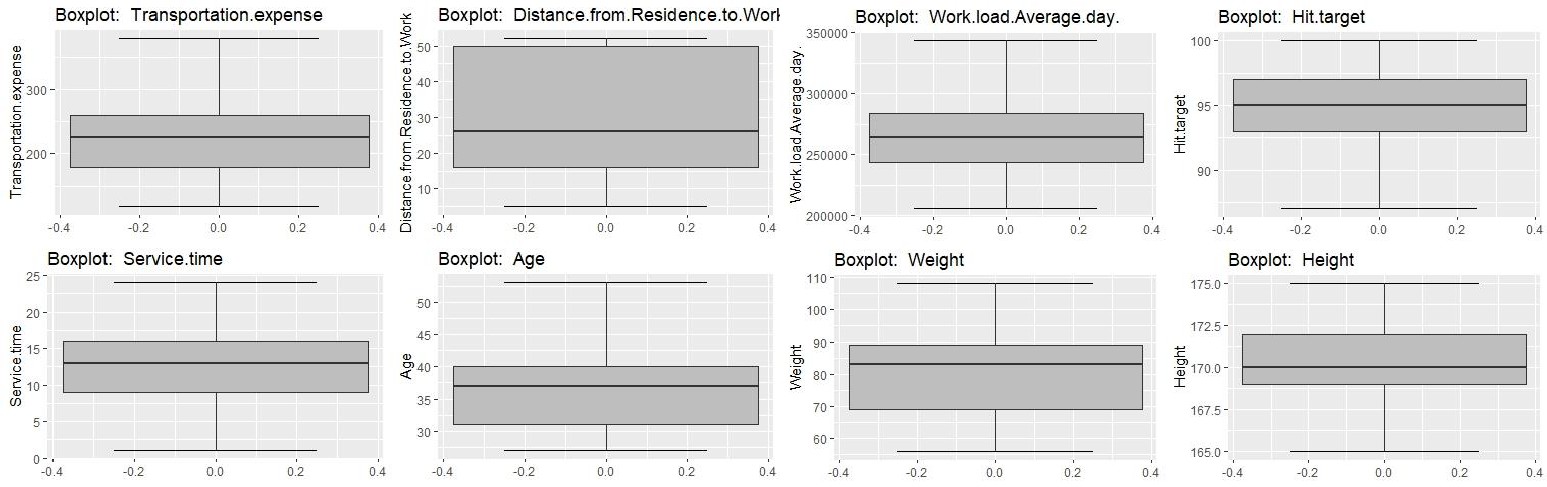


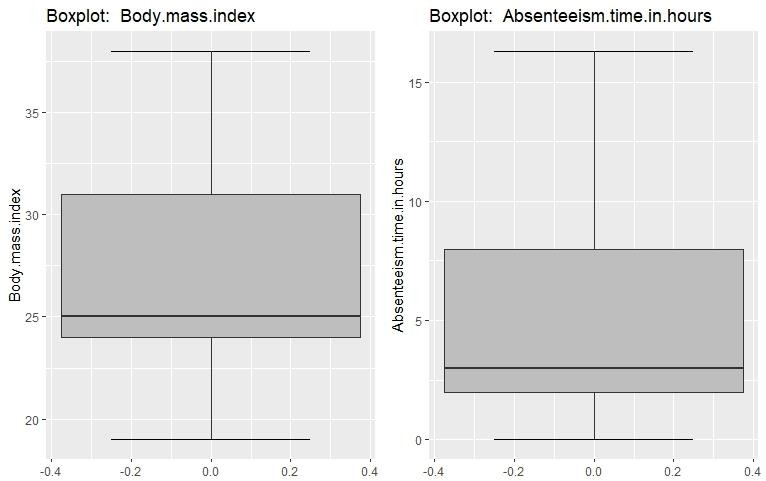
### Fig 2.2 – Missing value Percentage



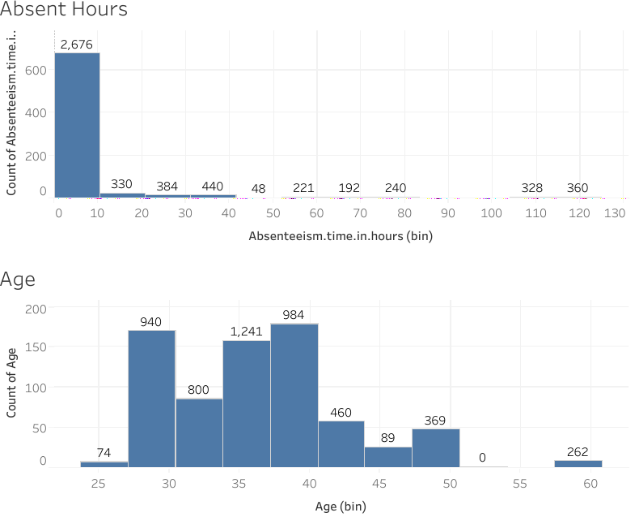
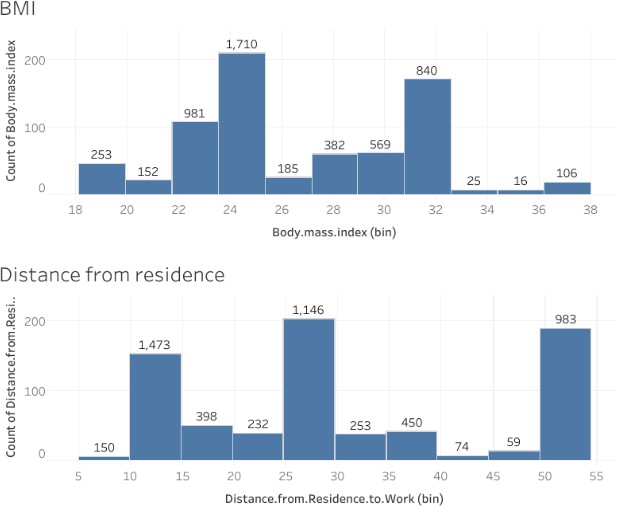


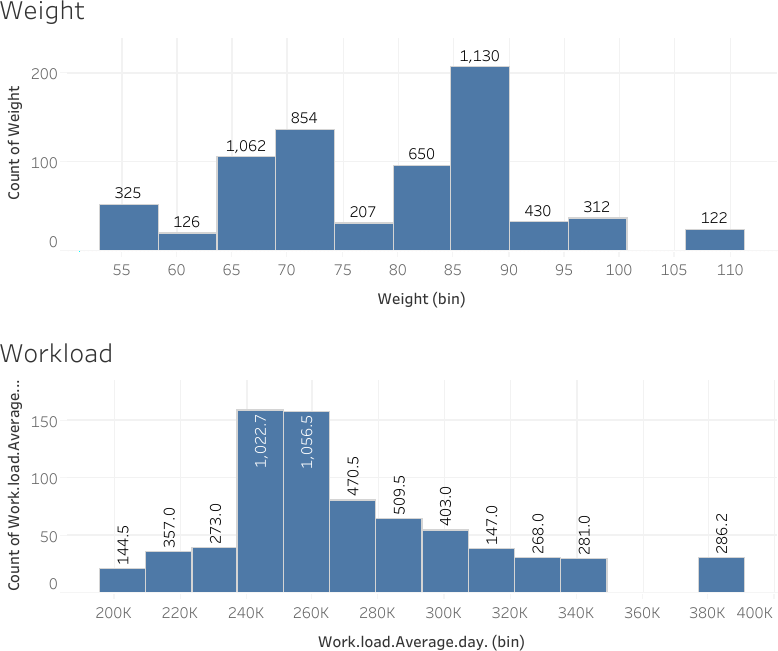
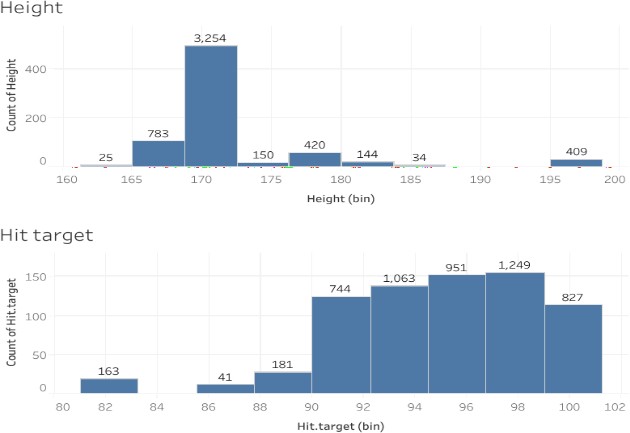
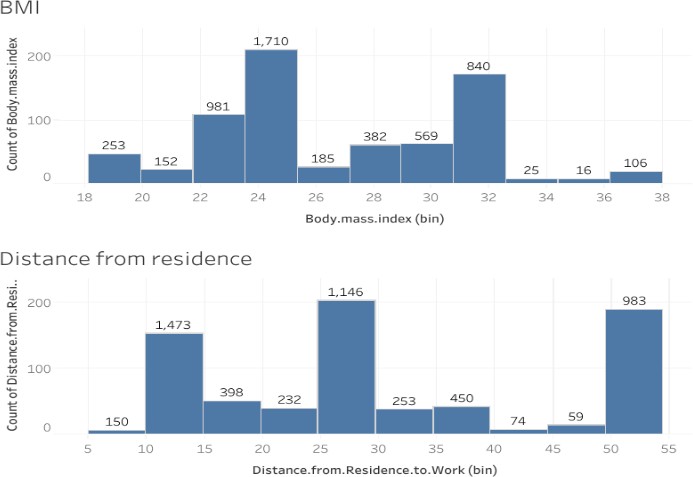
**Fig 2.3.1 – Boxplots of continuous variables with outliers**



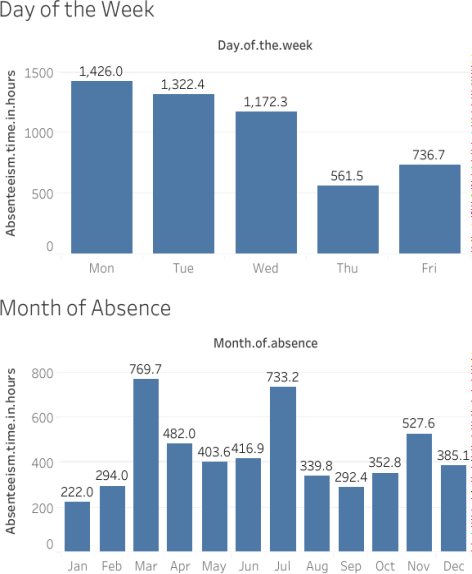
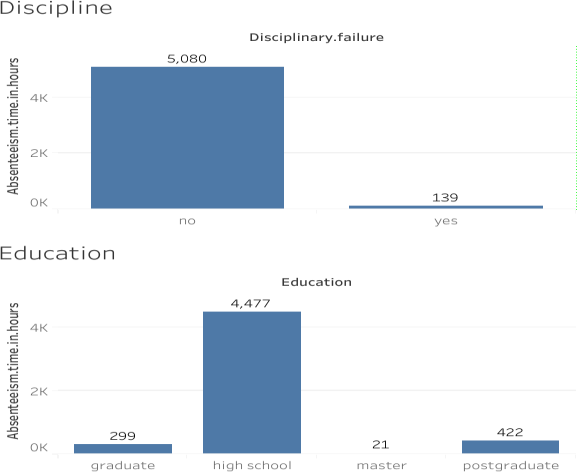


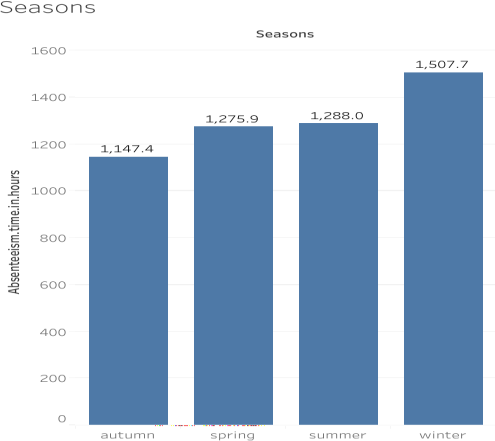
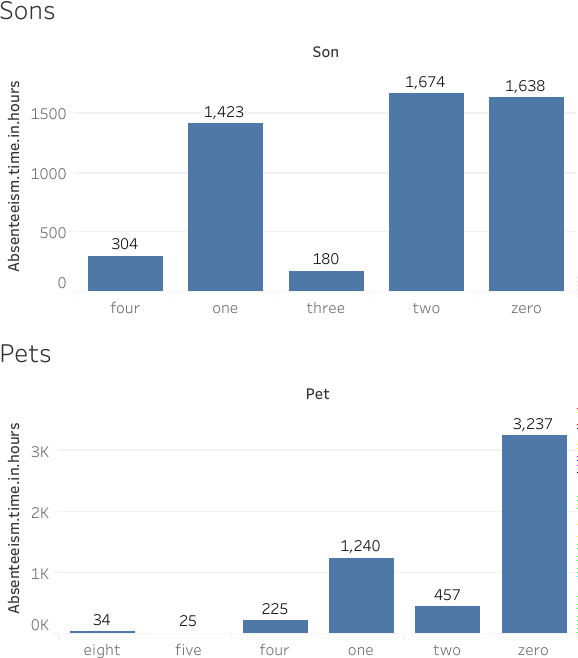
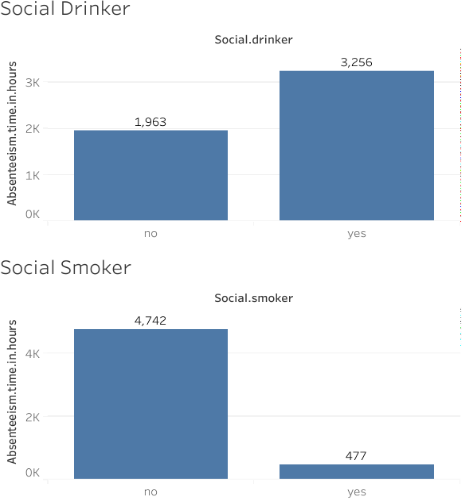
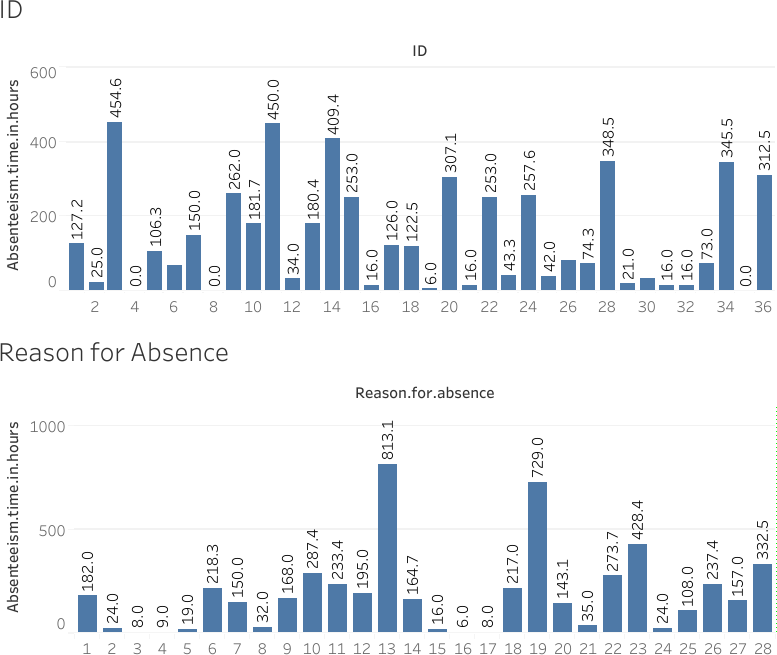
**Fig 2.3.2 – Boxplots of continuous variables without outliers**

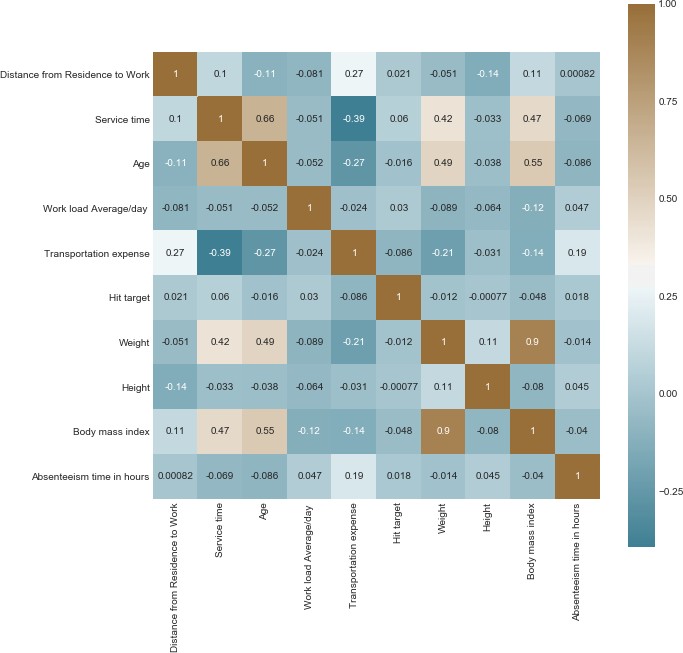


### Fig 2.4 – Distribution of Continuous variables using Histogram

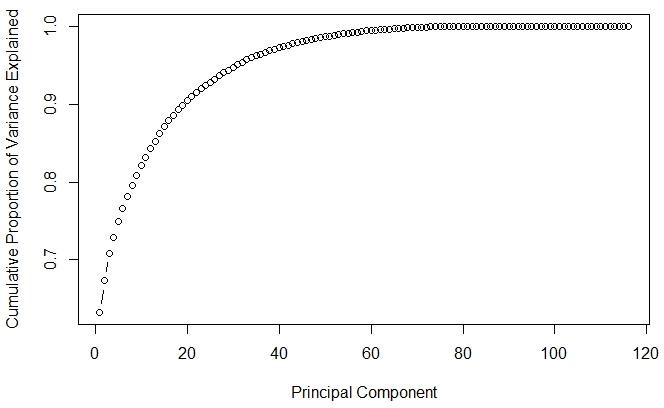
 



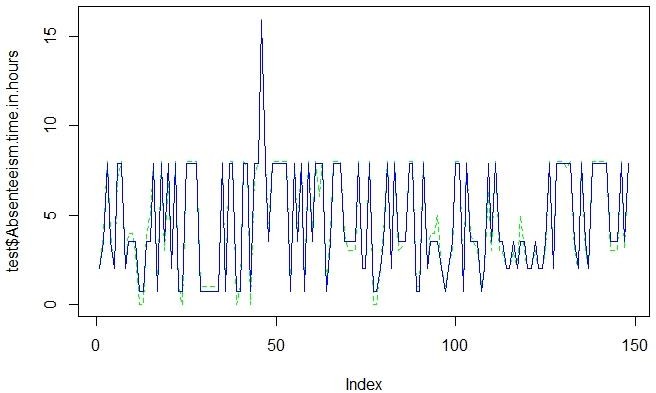
### Fig 2.5 – Distribution of Categorical variables using Bar graph



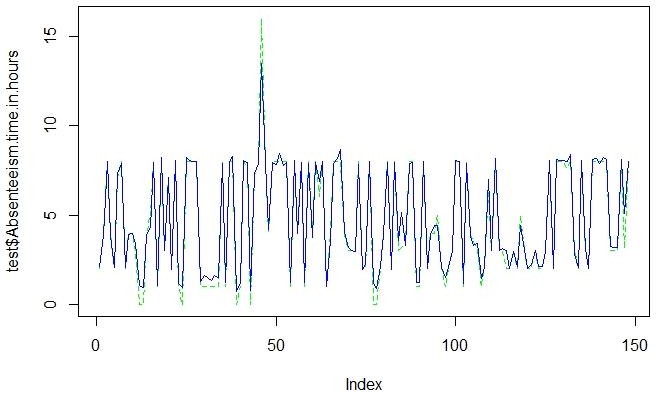
**Fig 2.6 – Correlation plot of Continuous variables**



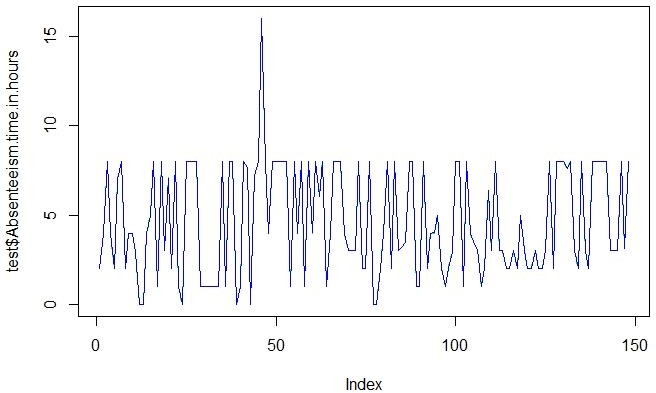
**Fig 2.8 – Cumulative Scree Plot of Principal Components**



**Fig 3.2 – Plot of actual values vs predicted values for Decision Tree**



**Fig 3.3 – Plot of actual values vs predicted values for Random Forest**



**Fig 3.5 – Plot of actual values vs predicted values for Linear Regression**

**Chapter 6:** **R coSde**

#Read the csv file

emp\_absent = read.xlsx(file = "Absenteeism\_at\_work.xls", header = T, sheetIndex = 1) #################################EXPLORE THE DATA#################################

#Check number of rows and columns dim(emp\_absent)

#Observe top 5 rows head(emp\_absent)

#Structure of variables str(emp\_absent)

#Transform data types

emp\_absent$ID = as.factor(as.character(emp\_absent$ID)) emp\_absent$Reason.for.absence[emp\_absent$Reason.for.absence %in% 0] = 20 emp\_absent$Reason.for.absence = as.factor(as.character(emp\_absent$Reason.for.absence)) emp\_absent$Month.of.absence[emp\_absent$Month.of.absence %in% 0] = NA emp\_absent$Month.of.absence = as.factor(as.character(emp\_absent$Month.of.absence)) emp\_absent$Day.of.the.week = as.factor(as.character(emp\_absent$Day.of.the.week)) emp\_absent$Seasons = as.factor(as.character(emp\_absent$Seasons)) emp\_absent$Disciplinary.failure = as.factor(as.character(emp\_absent$Disciplinary.failure)) emp\_absent$Education = as.factor(as.character(emp\_absent$Education))

emp\_absent$Son = as.factor(as.character(emp\_absent$Son)) emp\_absent$Social.drinker = as.factor(as.character(emp\_absent$Social.drinker)) emp\_absent$Social.smoker = as.factor(as.character(emp\_absent$Social.smoker)) emp\_absent$Pet = as.factor(as.character(emp\_absent$Pet))

#Structure of variables str(emp\_absent)

#Make a copy of data df = emp\_absent

#############################MISSING VALUE ANALYSIS#############################

#Get number of missing values sapply(df,function(x){sum(is.na(x))})

missing\_values = data.frame(sapply(df,function(x){sum(is.na(x))}))

#Get the rownames as new column missing\_values$Variables = row.names(missing\_values)

#Reset the row names row.names(missing\_values) = NULL

#Rename the column names(missing\_values)[1] = "Miss\_perc"

#Calculate missing percentage

missing\_values$Miss\_perc = ((missing\_values$Miss\_perc/nrow(emp\_absent)) \*100)

#Reorder the columns

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

#Sort the rows according to decreasing missing percentage missing\_values = missing\_values[order(-missing\_values$Miss\_perc),]

#Create a bar plot to visualie top 5 missing values

ggplot(data = missing\_values[1:5,], aes(x=reorder(Variables, -Miss\_perc),y = Miss\_perc)) + geom\_bar(stat = "identity",fill = "grey")+xlab("Parameter")+

ggtitle("Missing data percentage") + theme\_bw()

#Create missing value and impute using mean, median and knn df[["Body.mass.index"]][3] = NA

df = knnImputation(data = df, k = 5)

#Check if any missing values sum(is.na(df))

# Saving output result into excel file

write.xlsx(missing\_values, "Missing\_perc\_R.xlsx", row.names = F) ###################EXPLORE DISTRIBUTION USING GRAPHS###################

#Get numerical data

numeric\_index = sapply(df, is.numeric) numeric\_data = df[,numeric\_index]

#Distribution of factor data using bar plot

bar1 = ggplot(data = df, aes(x = ID)) + geom\_bar() + ggtitle("Count of ID") + theme\_bw() bar2 = ggplot(data = df, aes(x = Reason.for.absence)) + geom\_bar() +

ggtitle("Count of Reason for absence") + theme\_bw()

bar3 = ggplot(data = df, aes(x = Month.of.absence)) + geom\_bar() + ggtitle("Count of Month") + theme\_bw()

bar4 = ggplot(data = df, aes(x = Disciplinary.failure)) + geom\_bar() + ggtitle("Count of Disciplinary failure") + theme\_bw()

bar5 = ggplot(data = df, aes(x = Education)) + geom\_bar() + ggtitle("Count of Education")

+ theme\_bw()

bar6 = ggplot(data = df, aes(x = Son)) + geom\_bar() + ggtitle("Count of Son") + theme\_bw() bar7 = ggplot(data = df, aes(x = Social.smoker)) + geom\_bar() +

ggtitle("Count of Social smoker") + theme\_bw()

gridExtra::grid.arrange(bar1,bar2,bar3,bar4,ncol=2) gridExtra::grid.arrange(bar5,bar6,bar7,ncol=2)

#Check the distribution of numerical data using histogram

hist1 = ggplot(data = numeric\_data, aes(x =Transportation.expense)) + ggtitle("Transportation.expense") + geom\_histogram(bins = 25)

hist2 = ggplot(data = numeric\_data, aes(x =Height)) + ggtitle("Distribution of Height") + geom\_histogram(bins = 25)

hist3 = ggplot(data = numeric\_data, aes(x =Body.mass.index)) + ggtitle("Distribution of Body.mass.index") + geom\_histogram(bins = 25)

hist4 = ggplot(data = numeric\_data, aes(x =Absenteeism.time.in.hours)) + ggtitle("Distribution of Absenteeism.time.in.hours") + geom\_histogram(bins = 25)

gridExtra::grid.arrange(hist1,hist2,hist3,hist4,ncol=2) #########################OUTLIER ANALYSIS#########################

#Get the data with only numeric columns numeric\_index = sapply(df, is.numeric) numeric\_data = df[,numeric\_index]

#Get the data with only factor columns factor\_data = df[,!numeric\_index]

#Check for outliers using boxplots for(i in 1:ncol(numeric\_data)) {

assign(paste0("box",i), ggplot(data = df, aes\_string(y = numeric\_data[,i])) + stat\_boxplot(geom = "errorbar", width = 0.5) + geom\_boxplot(outlier.colour = "red", fill = "grey", outlier.size = 1) + labs(y = colnames(numeric\_data[i])) + ggtitle(paste("Boxplot: ",colnames(numeric\_data[i]))))

}

#Arrange the plots in grids gridExtra::grid.arrange(box1,box2,box3,box4,ncol=2) gridExtra::grid.arrange(box5,box6,box7,box8,ncol=2) gridExtra::grid.arrange(box9,box10,ncol=2)

#Replace all outlier data with NA for(i in numeric\_columns){

val = df[,i][df[,i] %in% boxplot.stats(df[,i])$out] print(paste(i,length(val)))

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

}

#Check number of missing values sapply(df,function(x){sum(is.na(x))})

#Get number of missing values after replacing outliers as NA missing\_values\_out = data.frame(sapply(df,function(x){sum(is.na(x))})) missing\_values\_out$Columns = row.names(missing\_values\_out) row.names(missing\_values\_out) = NULL names(missing\_values\_out)[1] = "Miss\_perc"

missing\_values\_out$Miss\_perc = ((missing\_values\_out$Miss\_perc/nrow(emp\_absent)) \*100) missing\_values\_out = missing\_values\_out[,c(2,1)]

missing\_values\_out = missing\_values\_out[order(-missing\_values\_out$Miss\_perc),] missing\_values\_out

#Compute the NA values using KNN imputation df = knnImputation(df, k = 5)

#Check if any missing values sum(is.na(df))

#############################FEATURE SELECTION#############################

#Check for multicollinearity using VIF vifcor(numeric\_data)

#Check for multicollinearity using corelation graph

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

#Variable Reduction

df = subset.data.frame(df, select = -c(Body.mass.index)) #########################FEATURE SCALING#########################

#Normality check hist(df$Absenteeism.time.in.hours)

#Remove dependent variable numeric\_index = sapply(df,is.numeric) numeric\_data = df[,numeric\_index] numeric\_columns = names(numeric\_data) numeric\_columns = numeric\_columns[-9]

#Normalization of continuous variables for(i in numeric\_columns){

print(i)

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

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

}

#Get the names of factor variables factor\_columns = names(factor\_data)

#Create dummy variables of factor variables df = dummy.data.frame(df, factor\_columns)

rmExcept(keepers = c("df","emp\_absent"))

########################DIMENSION REDUCTION USING PCA########################

#Principal component analysis prin\_comp = prcomp(train)

#Compute standard deviation of each principal component pr\_stdev = prin\_comp$sdev

#Compute variance pr\_var = pr\_stdev^2

#Proportion of variance explained prop\_var = pr\_var/sum(pr\_var)

#Cumulative scree plot

plot(cumsum(prop\_var), xlab = "Principal Component", ylab = "Cumulative Proportion of Variance Explained", type = "b")

#Add a training set with principal components

train.data = data.frame(Absenteeism.time.in.hours = train$Absenteeism.time.in.hours, prin\_comp$x)

# From the above plot selecting 45 components since it explains almost 95+ % data variance train.data =train.data[,1:45]

#Transform test data into PCA

test.data = predict(prin\_comp, newdata = test) test.data = as.data.frame(test.data)

#Select the first 45 components test.data=test.data[,1:45]

########################DECISION TREE########################

#Build decsion tree using rpart

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

#Predict the test cases

dt\_predictions = predict(dt\_model,test.data)

#Create data frame for actual and predicted values

df\_pred = data.frame("actual"=test[,116], "dt\_pred"=dt\_predictions) head(df\_pred)

#Calcuate MAE, RMSE, R-sqaured for testing data

print(postResample(pred = dt\_predictions, obs = test$Absenteeism.time.in.hours))

#Plot a graph for actual vs predicted values plot(test$Absenteeism.time.in.hours,type="l",lty=2,col="green") lines(dt\_predictions,col="blue")

###########################RANDOM FOREST###########################

#Train the model using training data

rf\_model = randomForest(Absenteeism.time.in.hours~., data = train.data, ntrees = 500)

#Predict the test cases

rf\_predictions = predict(rf\_model,test.data)

#Create dataframe for actual and predicted values df\_pred = cbind(df\_pred,rf\_predictions) head(df\_pred)

#Calcuate MAE, RMSE, R-sqaured for testing data

print(postResample(pred = rf\_predictions, obs = test$Absenteeism.time.in.hours))

#Plot a graph for actual vs predicted values plot(test$Absenteeism.time.in.hours,type="l",lty=2,col="green") lines(rf\_predictions,col="blue")

###########################LINEAR REGRESSION###########################

#Train the model using training data

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

#Get the summary of the model summary(lr\_model)

#Predict the test cases

lr\_predictions = predict(lr\_model,test.data)

#Create dataframe for actual and predicted values df\_pred = cbind(df\_pred,lr\_predictions) head(df\_pred)

#Calcuate MAE, RMSE, R-sqaured for testing data

print(postResample(pred = lr\_predictions, obs =test$Absenteeism.time.in.hours))

#Plot a graph for actual vs predicted values plot(test$Absenteeism.time.in.hours,type="l",lty=2,col="green") lines(lr\_predictions,col="blue")