Data Science

Project Name: Employee Absenteeism

Report

Submitted By: Akash Dash

6th January 2019

**Contents**

1 Introduction

1.1 Problem Statement 3

1.2 Data 3

2 Methodology

2.1 Pre Processing 4

2.1.1 Missing Value Analysis 7

2.1.2 Outlier Analysis 8

2.1.3 Feature Selection 9

2.1.4 Feature Scaling 10

2.1.5 Principal Component Analysis 10

2.2 Modelling

2.2.1 Preparing Data for Models 11

2.2.2 Decision Tree 11

2.2.3 Linear regression 12

2.2.4 Random Forest 12

3 Conclusion

3.1 Model Evaluation 12

3.2 Model Selection 12

A Suggestions to reduce absenteeism 13

B R Code 15

C Reference 21

# Introduction

Absenteeism is the unplanned absence from the duty without any good reason. It is one of the severe issues faced by many companies. We know that human capital plays an important role in any business. Being able to predict the employee absenteeism can prevent company from severe loss. It is very important to know the cause of absenteeism among the employee. One of the best ways to make the predictions is with the help of machine learning techniques.

# Problem Statement

# XYZ a courier company, acknowledges that human capital plays vital role in collection, transportation and delivery. The company is facing a pressing issue of absenteeism. The company shared the dataset to get insights on the major factors affecting Absenteeism.

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

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

# Data

The objective is to analyse the factors which lead to absenteeism and how the changes can help the company reduce it. We are provided with a dataset of 740 observations and 21 variables, of these 21 variables we have 9 categorical variables and 12 continuous variables (including ID). By plotting distribution, we can observe the variables are not normally distributed. Thus we will perform Normalization on the variables in feature scaling. Table 1 (Column 1 to 6)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Reason.for.absence** | **Month.of.absence** | **Day.of.the.week** | **Seasons** | **Transportation.expense** |
| 11 | 26 | 7 | 3 | 1 | 289 |
| 36 | 0 | 7 | 3 | 1 | 118 |
| 3 | 23 | 7 | 4 | 1 | 179 |
| 7 | 7 | 7 | 5 | 1 | 279 |

Table 2 (Column 7 to 12)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Distance.from.Residence.to.Work** | **Service.time** | **Age** | **Work.load.per.day** | **Hit.target** | **Disciplinary.failure** |
| 36 | 13 | 33 | 239554 | 97 | 0 |
| 13 | 18 | 50 | 239554 | 97 | 1 |
| 51 | 18 | 38 | 239554 | 97 | 0 |
| 5 | 14 | 39 | 239554 | 97 | 0 |

Table 3 (Column 13 to 18)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Education** | **Son** | **Social.drinker** | **Social.smoker** | **Pet** | **Weight** |
| 1 | 2 | 1 | 0 | 1 | 90 |
| 1 | 1 | 1 | 0 | 0 | 98 |
| 1 | 0 | 1 | 0 | 0 | 89 |
| 1 | 2 | 1 | 1 | 0 | 68 |

Table 4 (Column 19 to 21)

|  |  |  |
| --- | --- | --- |
| **Height** | **Body.mass.index** | **Absenteeism.time.in.hours** |
| 172 | 30 | 4 |
| 178 | 31 | 0 |
| 170 | 31 | 2 |
| 168 | 24 | 4 |

# Methodology

The backbone of many business processes today is data. The scientific approach to derive insights from data is called Data Science. Data is collected from different sources thus they are often messy, noisy and unstructured. Converting data into proper format is necessary. In any project 80% of time and effort is spent on understanding, cleaning and preparing the data as preparing the data according to problem. The whole process is divided into six phases termed in industry as CRISP-DM. They are:

1. Business understanding: When client brings a problem, we try to understand their business first, then the context of the problem in respect of the business. It helps us to get relevant data for better results.
2. Data understanding: Here we apply statistical techniques, graphs and visualizations to explore and understand the data so that we can acquire right set of data to provide relevant insights to the client.
3. Data Preparation: This process involves exploring the raw data, clean, organize, and process the data for further use. More than 80% of our time goes into data understanding, cleaning and preparation and 20% in model development and model evaluation. The quality of data used is directly proportional to the performance of the model.
4. Data modeling: From many machine learning algorithms, we select the most appropriate algorithm according to our problem statement.
5. Evaluation: In this step we evaluate our model to see whether it fulfills our objective. Error metrics are checked and the model is tweaked accordingly if need arises. By the end of this process, we ascertain whether our model is able to accomplish the business objective or not.
6. Deployment: This is the final phase in which we deploy our model in client premises.

# Pre-processing

Any predictive modelling requires that we process the data prior to modelling. In data mining terms, exploring the data includes cleaning, organizing, treating anomalies and visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis (EDA).

1. Missing value analysis
2. Outlier Analysis
3. Feature selection

3.1 Correlation Analysis

3.2 ANOVA

1. Normalization

Figure 1

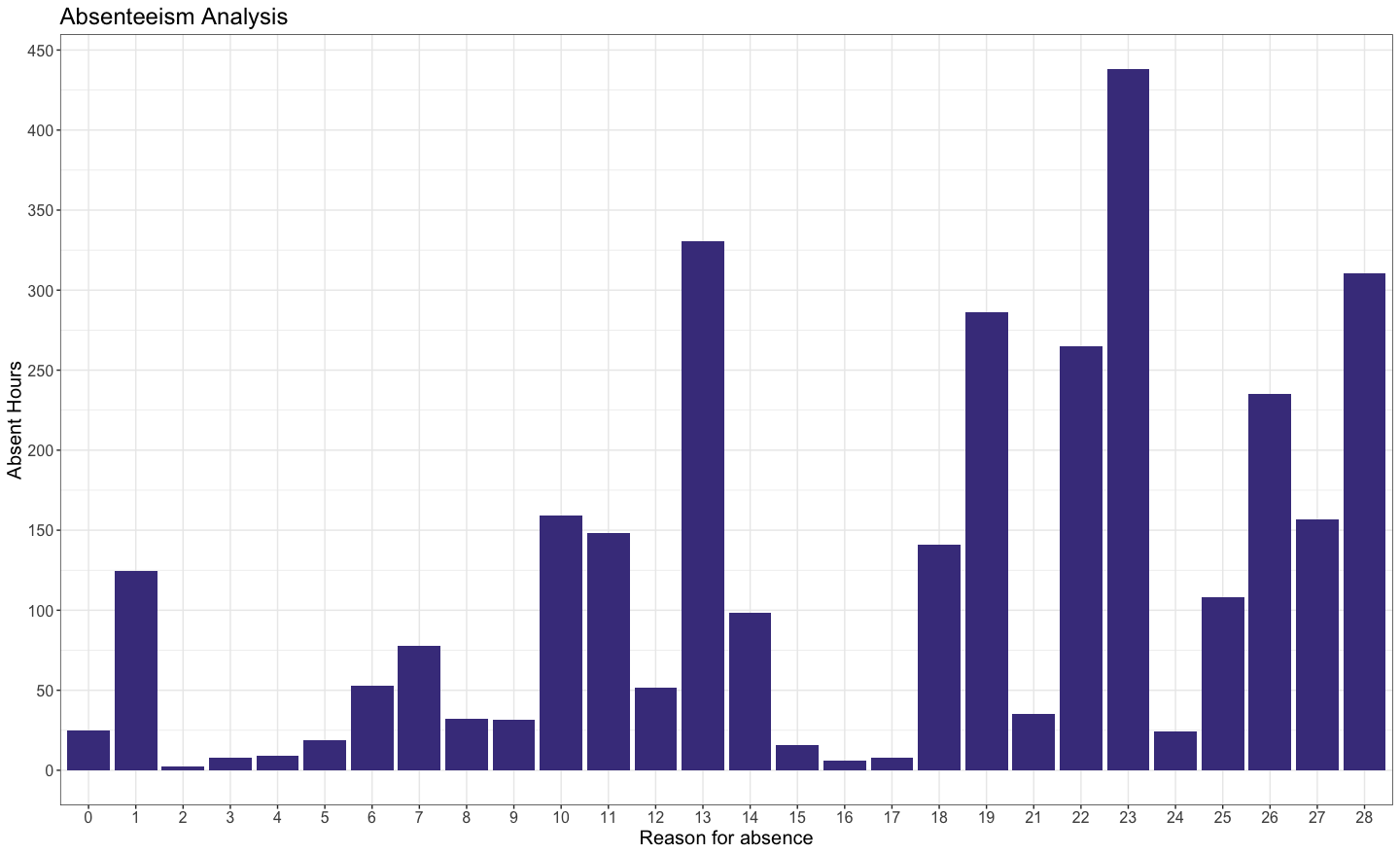
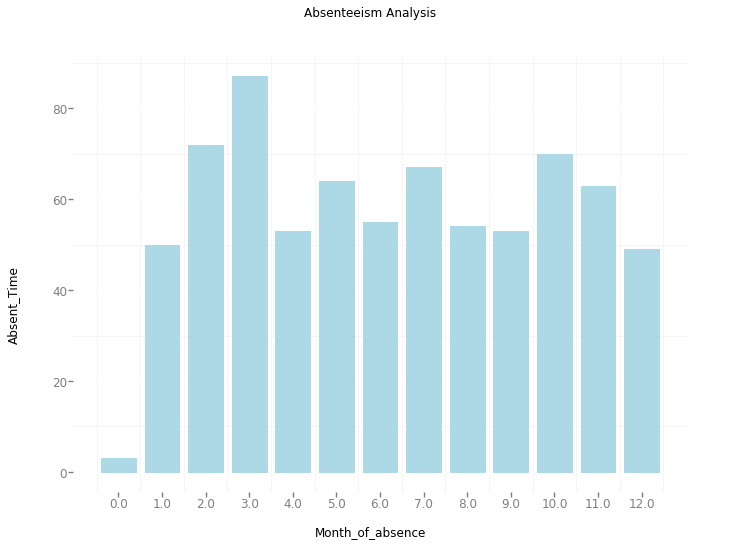
 

Figure 2

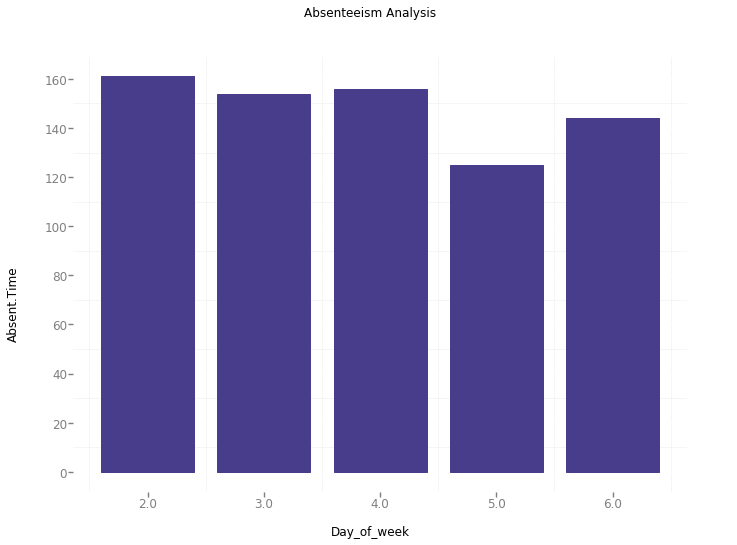
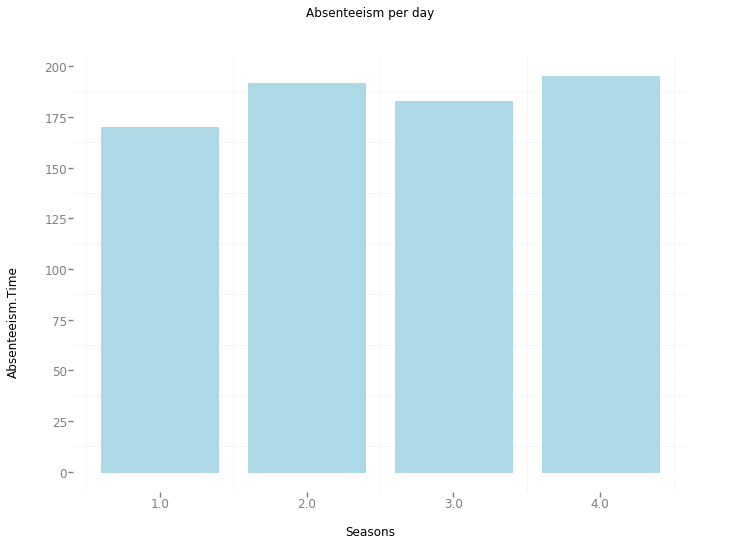
 

Figure 3

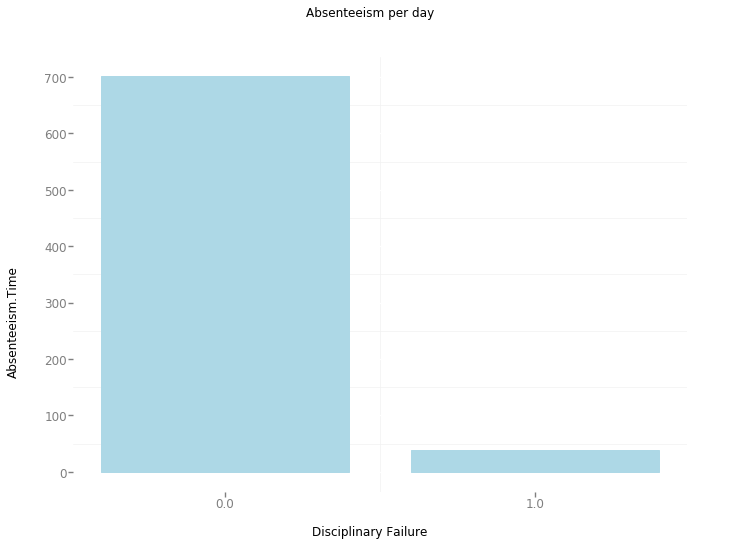
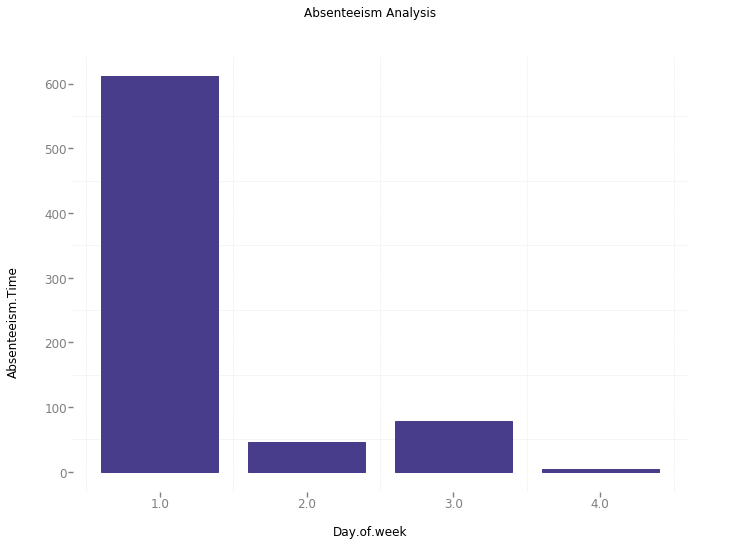
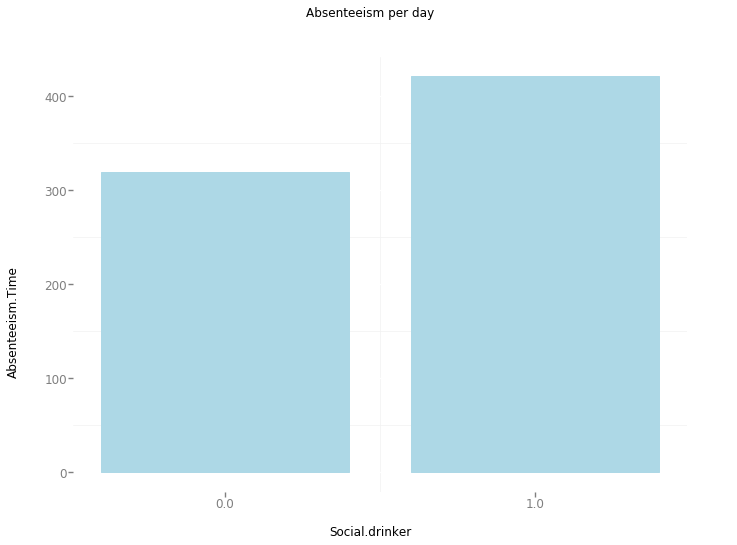
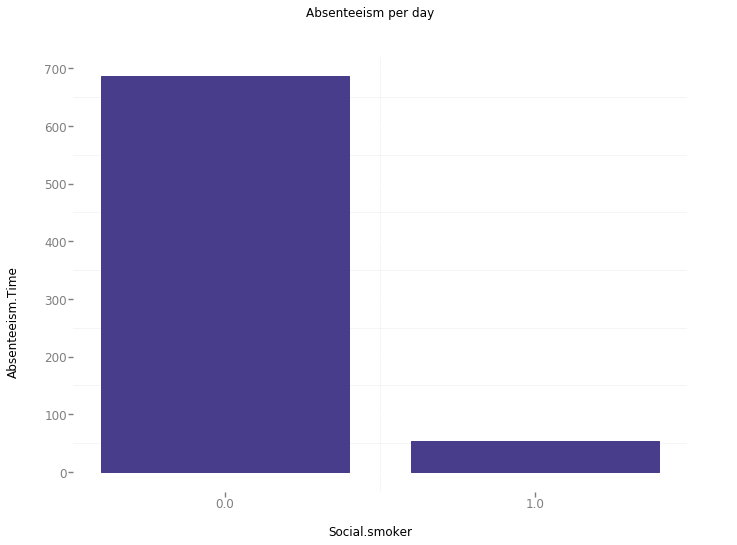
 

Figure 4

# Missing Value Analysis

Missing value is the values that are not present or missing from the dataset. Missing values may appear in the dataset because of human error, refusal to answer a questions during survey or options based questions. Missing values can be treated either by dropping the variable or by imputing the missing values. It is very essential to understand the cause missing value, it can be done with proper domain knowledge and understanding of business problem. Then it is possible to ascertain whether to ignore or impute the missing values.

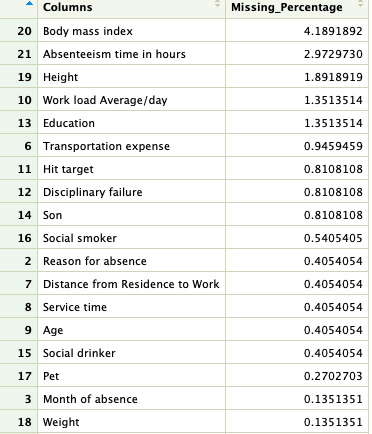
We drop the variables having more than 30% (according to industry standards) of missing values.

We impute those variables whose missing percentage is less than 30%. There are three methods to impute missing values:

a) Fill with central statistics method i.e. mean and median for continuous variable and mode (majority minority rule) for categorical variable.

b) Distance based or KNN imputation.

c) The last method is prediction method based on ML algorithms.



**We have 135 missing values. The variables have less than 10% of missing values so we impute them using median accordingly.** The graphs of missing values percentage are presented below.

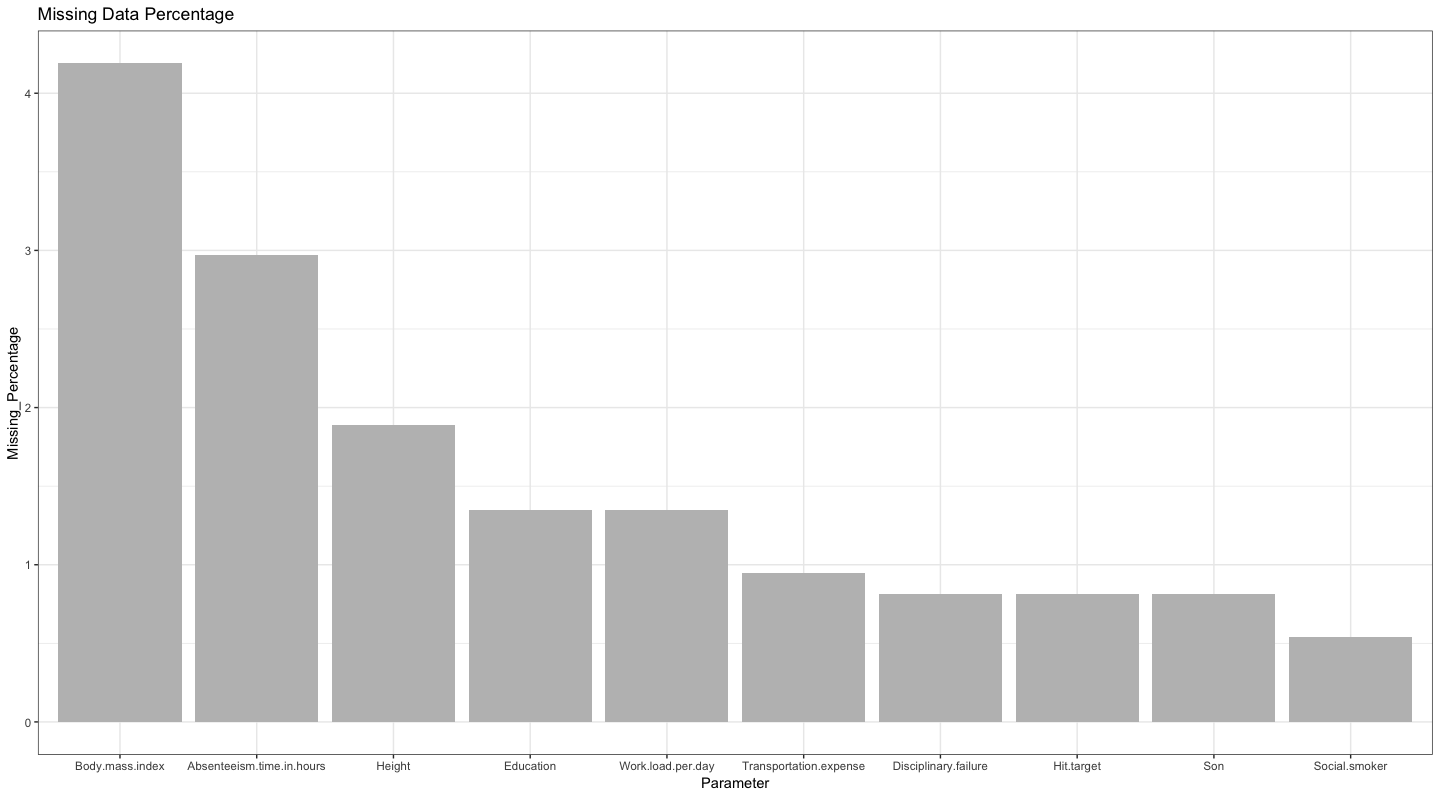
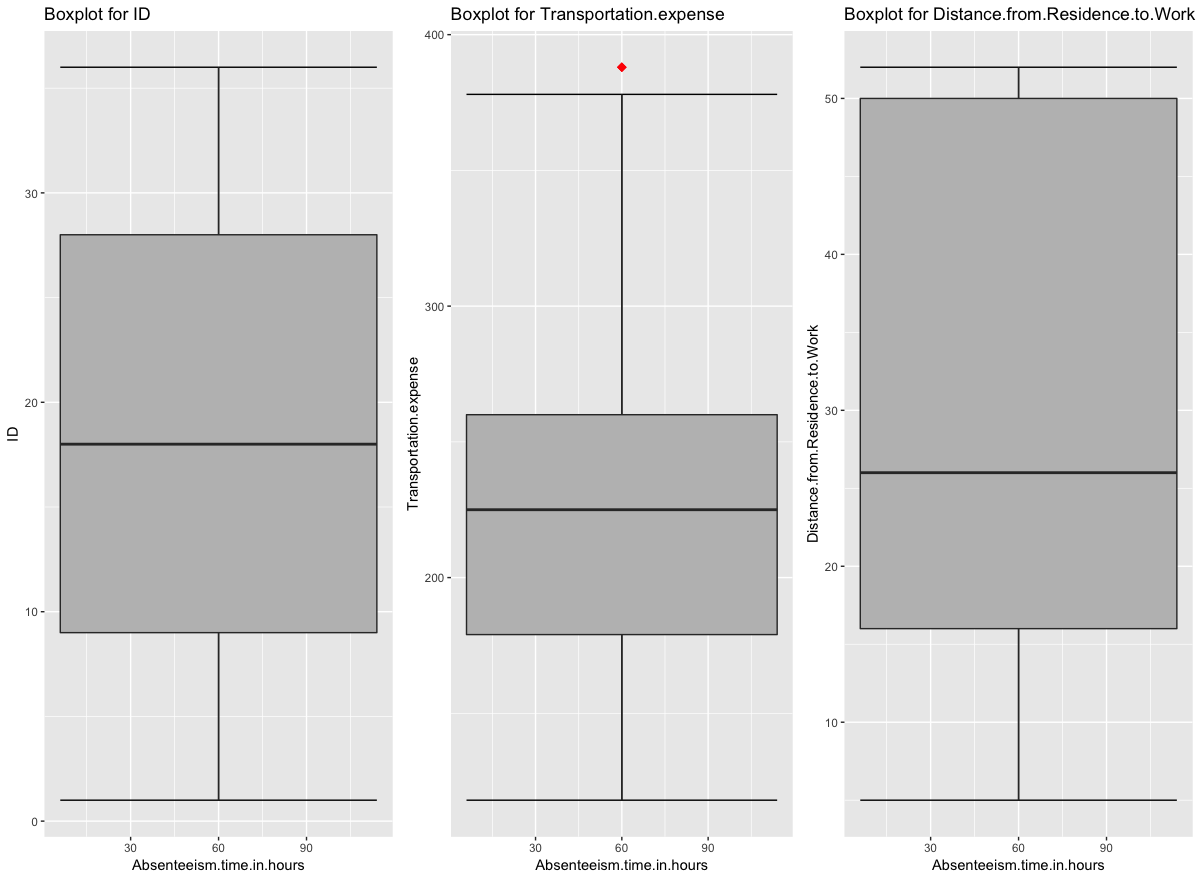
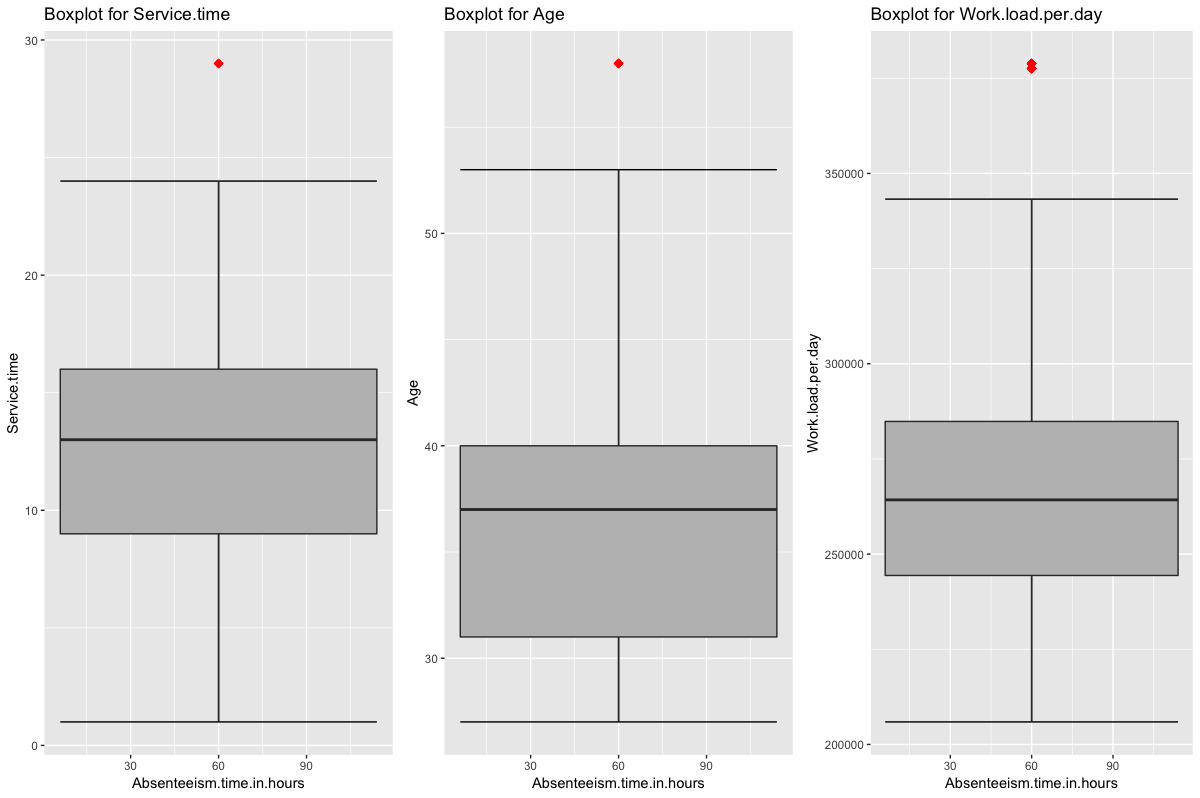


Figure 5

# Outlier Analysis

Outlier can be defined as the observations which are inconsistent with rest of the dataset, these observations are anomaly or abnormality in data. In certain cases, outliers are useful in fraud detection or unusual patterns in medical domain. However, in context of Absenteeism Analysis of the data provided by the XYZ courier company, the outliers should be treated as anomaly. We can observe outliers in Transportation expense, Distance from residence from work, Height etc. which are removed are removed by KNN imputation.

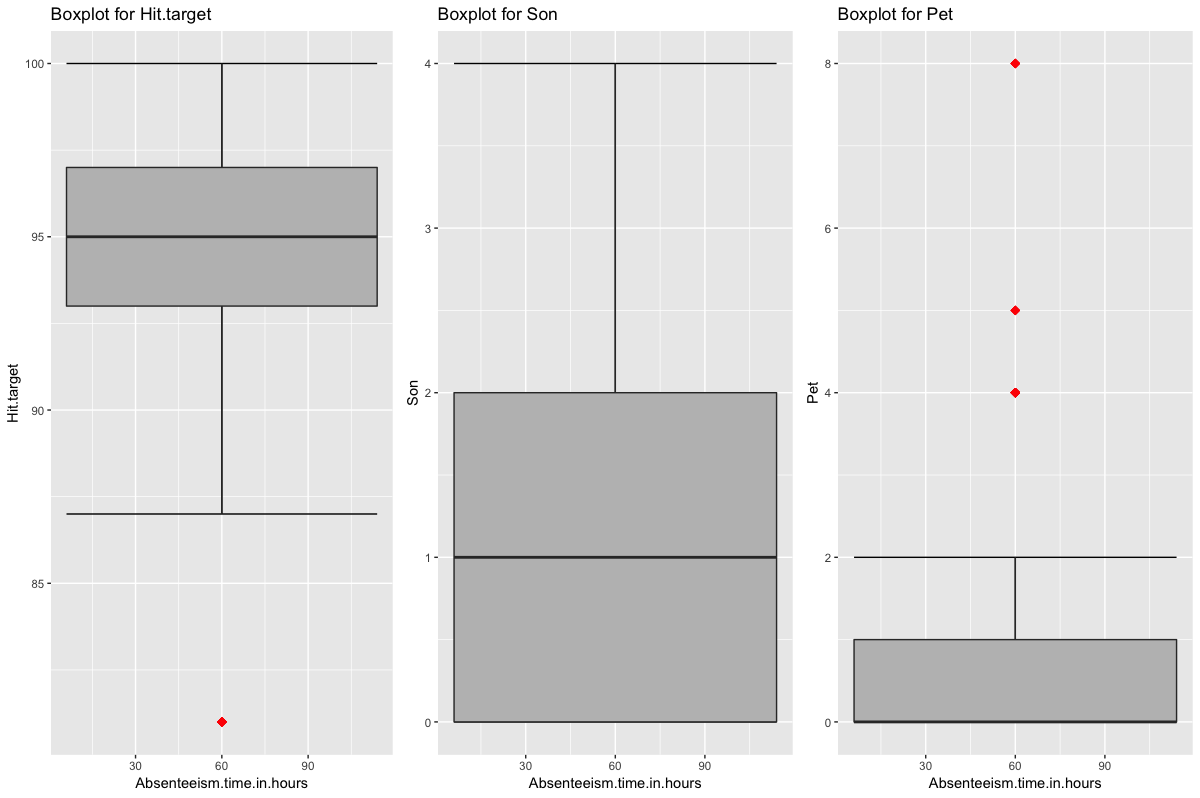
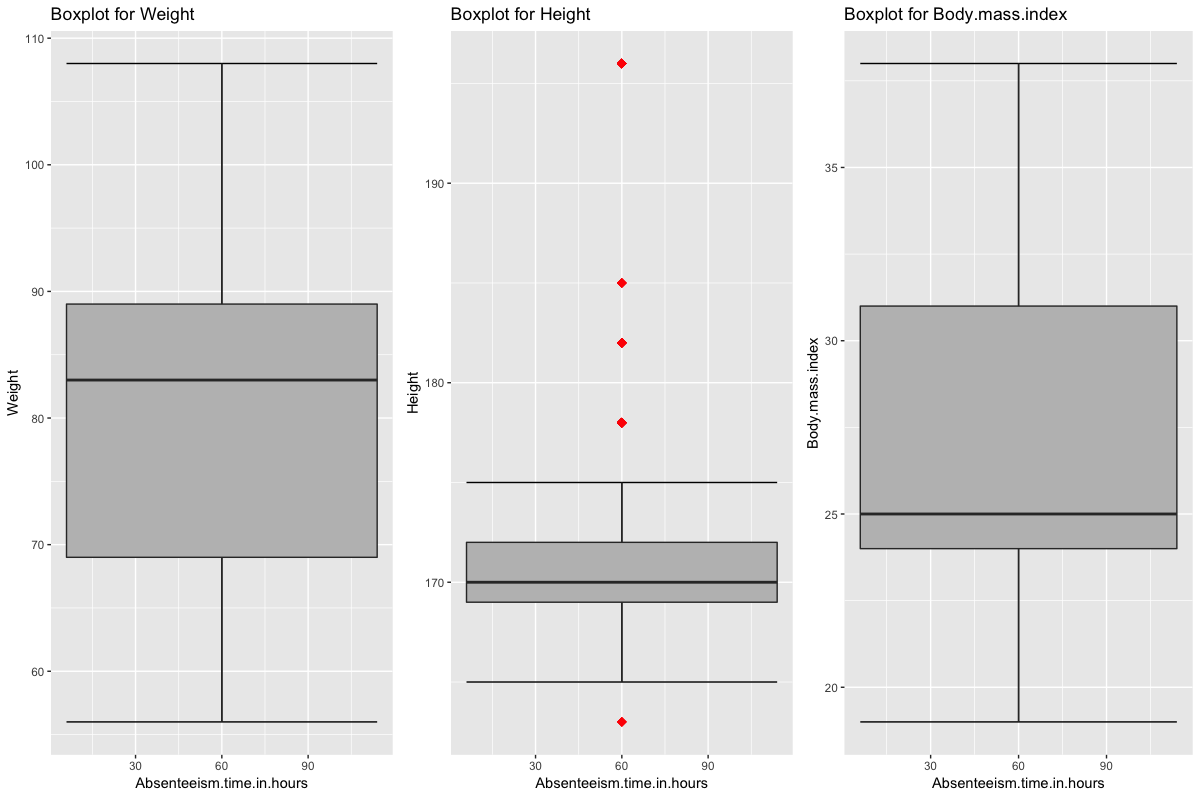
 

Figure 6

# Feature Selection

Before performing any type of modelling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi collinearity. In this project we have selected **Correlation Test** for numerical variables and **Anova Test** for categorical variables.



Figure: 7

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. As our target variable is numerical we will use ANOVA for feature selection technique to see whether any categorical variable is related to target variable. The result of anova is as follows:

|  |  |
| --- | --- |
| **Variable name** | **p- value** |
| Reason for absence | 2e-16 \*\*\* |
| Month of absence | 0.00211 \*\* |
| Day of the week | 0.116 |
| Seasons | 0.186 |
| Disciplinary Failure | 2.24e-13 \*\*\* |
| Education | 0.361 |
| Social smoker | 0.167 |
| Social drinker | 0.0344 \* |

If the p value of the categorical variable is less than 0.05 then we will consider that variable that is the target variable is dependent on the categorical variable for which we confirm the null hypothesis.

From the above result we can see that only one variable is very much related to target variable but we shouldn’t delete all the variables. So with some domain knowledge we delete the below mentioned variables

Therefore, from both the correlation analysis and ANOVA we got some variable which we shouldn’t consider for further processing. The variables that could be deleted are as follows

**Numerical**: "Weight"

**Categorical**: "Day of week", "Seasons", "Education", "Social smoker"

# Feature Scaling

Feature scaling is also called as variable scaling; it is one of the most important steps when we deal with continuous variables of different scales. As we can see, the variables like “ID”, “Transportation expense”, “Distance from residence to work”, “Service time” all are of different scales. For analysis, the variable should be scaled unitarily. For this purpose, we used Normalization of variables. It is a process to bring all the variables in proportion with one another.

Standardization is the process of converting each data point of a variable to a unit of standard deviation. However, we perform standardization only on the variables which are uniformly distributed, since the variables in our dataset are not normally distributed we have performed Normalization on the variables.

# Principal Component Analysis

Principal component analysis is defined as a technique used for identification of a smaller number of uncorrelated variables known as principal components from a larger set of data. The use of principal component analysis is to eliminate the number of variables or when there are too many predictors compared to number of observations or to avoid multicollinearity.

Simply put, the objective of PCA is

* To discover or to reduce the dimensionality of the data set.
* To identify new meaningful underlying variables.

It is important here to note that we have applied principal component analysis only in R programming to get a comprehensive result and compare the output of both approach.

# Modelling

Data modeling refers to selecting right machine learning algorithm according to problem statement and running it on the data to predict test cases. There are more than 100 algorithms, thus we need to select right ones.

# Preparing Data

Preparing data for modelling involves creating dummy variables, sampling, dividing dataset in training and test set. In R we have used principal components in train and test set.

# Decision Tree

Decision Tree is one of the basic supervised machine learning algorithm which can be used both for classification as well as regression problems. It is a predictive model based on branching series of Boolean tests. Decision tree is rule based and output of decision tree is in the form of simple business rules which are extremely easy to understand by business users.

Here we use this algorithm to predict the Absenteeism variable of test cases. We created decision tree model named DT.

|  |  |
| --- | --- |
| Python | R |
| RMSE: 0.1872 | RMSE: 0.1425 |
| Accuracy: 81.28 | Accuracy: 85.75 |

# Linear Regression

Linear regression model is one of the most widely used regression algorithm. It is the most basic type of regression and commonly used predictive analysis. Linear regression is an approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

|  |  |
| --- | --- |
| Python | R |
| RMSE: 0.1672 | RMSE: 0.1521 |
| Accuracy: 83.28 | Accuracy: 84.79 |
| R2: 0.793 | R2: 0.9656 |
| Adj R2: 0.776 | Adj R2: 0.963 |

* Note: The difference in R2 and Adj R2 in R and Python is due to PCA applied in R and not in Python to compare the results due to PCA.

# Random Forest

Random Forest or decision tree forests is an ensemble learning method for classification, regression and other tasks. It consists of an arbitrary number of simple trees, which are used to determine the final outcome. In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy (i.e., better ability to predict new data cases). The goal of using a large number of trees is to train enough that each feature has a chance to appear in several models. We have built the model taking 500 trees both in R and Python.

|  |  |
| --- | --- |
| Python | R |
| RMSE: 0.1553 | RMSE: 0.1396 |
| Accuracy: 84.47 | Accuracy: 86.04 |

# Conclusion

# Model Evaluation

After building regression models, there are criteria by which they can be evaluated and compared. Model evaluation tells us whether our model is able to accomplish the business objective or not.

RMSE is a popular metric to measure the error rate of time series or transition regression model. It can be only compared between models whose errors are measured in the same units. It can be calculated by squaring the errors, finding their average and taking their square root.

**Accuracy = 100 - RMSE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | R | | Python | |
| RMSE | Accuracy | RMSE | Accuracy |
| Decision tree | 0.1425 | 85.75 | 0.1872 | 81.28 |
| Linear Regression | 0.1521 | 84.79 | 0.1672 | 83.28 |
| **Random Forest** | **0.1396** | **86.04** | **0.1553** | **84.47** |

# Model Selection

From the above comparative study, we can see Random Forest model is outperforming other modelling algorithms. Thus we select Random Forest machine learning algorithm for prediction of future cases.

# A. Suggestions

# Q: What changes company should bring to reduce the number of absenteeism?

Ans: From the data set we were able to find some interesting insights on the absenteeism patterns of employees at XYZ courier company. We believe on acting upon these, the company will be able to reduce the absenteeism considerably.

* Introducing incentive for 100% attendance as bonus.
* Introducing policy for penalizing by salary deduction for un-informed or un-
* approved leaves.
* Health check-up camp once on the last Saturday of the month
* Sensitizing employees on ills of alcohol and smoking on health and benefits of yoga for personal development and increased productivity at work.
* Engaging employees in different activities apart from regular work on Monday morning.
* Family at work once a month to increase involvement.

The suggestions are based on the following information:

* Medical Consultation (> 430 hrs of absent time)
* Diseases of the musculoskeletal system and connective tissue (> 330 hrs of absent time)
* Dental Consultation (> 300 hrs of absent time)
* Injury, poisoning and certain other consequences of external causes (> 250 hrs of absent time)
* Social Drinker (> 1800 hrs of absent time)
* Education High School (> 2500 hrs of absent time)
* Work load highest for high school education
* March 450 hrs of absent hours
* Monday > 750 hrs of absent hours

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

Losses are caused because of low productivity as well as absenteeism from work. One of the major factors of work loss is employee absenteeism. In our project we have predicted the work loss faced by the company in year 2011. We computed the monthly work loss the company might face if the current trend follows in coming year. The losses are predicted with the help of the formula give below.

**Work loss = (Absenteeism time in hours \* Work load average/day)/ Service time**

|  | **Workload loss per month** |
| --- | --- |
| **No Absent** | 0 |
| **January** | 6270829 |
| **February** | 5938663 |
| **March** | 15787937 |
| **April** | 10620082 |
| **May** | 7854523 |
| **June** | 8335891 |
| **July** | 15641063 |
| **August** | 5325883 |
| **September** | 6049021 |
| **October** | 8410438 |
| **November** | 10504827 |
| **December** | 8832856 |

# B. R-Code

# Remove all objects from environment

rm(list=ls())

# Set Working Directory

setwd("/Users/ad/Desktop/Project 2")

# Check the set directory

getwd()

# Loading Required Libraries

x= c("ggplot2", "corrgram", "DMwR", "caret", "randomForest", "unbalanced", "C50", "dummies", "e1071", "MASS", "rpart", "gbm",

"ROSE","dummies", "inTrees", "e1071")

# Install.packages(x)

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

rm(x)

# Importing dataframe

data<- read.csv("Absenteeism\_at\_work\_Project.csv", header = TRUE)

# Exploring Data

colnames(data)

str(data)

summary(data)

#histogram(data$Reason.for.absence) # Factors influencing health status and contact with health services

#histogram(data$Month.of.absence)

#histogram(data$Day.of.the.week)

#histogram(data$Seasons)

#histogram(data$Transportation.expense)

#hist(data$Distance.from.Residence.to.Work)

#hist(data$Service.time)

#hist(data$Age)

#hist(data$Work.load.per.day)

#hist(data$Hit.target)

#hist(data$Disciplinary.failure)

#hist(data$Education)

#hist(data$Son)

#hist(data$Social.drinker)

#hist(data$Social.smoker)

#hist(data$Pet)

#hist(data$Weight)

#hist(data$Height)

#hist(data$Body.mass.index)

#hist(data$Absenteeism.time.in.hours)

### Missing Value Analysis ###

# Create dataframe with missing value count

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

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

row.names(missing\_val) = NULL

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

# Converting missing values to percentage

missing\_val$Missing\_Percentage <- missing\_val$apply.data..2..function.x...

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

# Rearranging Columns

missing\_val = missing\_val[,c(2,1,3)]

# Arranging missing\_percentage in decending order

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

# Plot bar-graph for missing values

ggplot(data = missing\_val[1:9,], aes(x= reorder(Columns, -Missing\_Percentage), y= Missing\_Percentage))+

geom\_bar(stat = "identity", fill= "grey")+ xlab("Parameter")+

ggtitle("Missing Data Percentage")+ theme\_bw()

# Changing data type of "Work.load.Average.per.day" to numeric

data$Work.load.per.day = as.numeric(data$Work.load.per.day)

## Imputing Missing Values

data$Body.mass.index[is.na(data$Body.mass.index)] = median(data$Body.mass.index, na.rm = T)

data$Absenteeism.time.in.hours[is.na(data$Absenteeism.time.in.hours)] = median(data$Absenteeism.time.in.hours, na.rm = T)

data$Height[is.na(data$Height)] = median(data$Height, na.rm = T)

data$Work.load.per.day[is.na(data$Work.load.per.day)] = median(data$Work.load.per.day , na.rm = T)

data$Education[is.na(data$Education)] = median(data$Education, na.rm = T)

data$Transportation.expense[is.na(data$Transportation.expense)] = median(data$Transportation.expense, na.rm = T)

data$Hit.target[is.na(data$Hit.target)] = median(data$Hit.target, na.rm = T)

data$Education[is.na(data$Education)] = median(data$Education, na.rm = T)

data$Disciplinary.failure[is.na(data$Disciplinary.failure)] = median(data$Disciplinary.failure, na.rm = T)

data$Son[is.na(data$Son)] = median(data$Son, na.rm = T)

data$Social.smoker[is.na(data$Social.smoker)] = median(data$Social.smoker , na.rm = T)

data$Reason.for.absence[is.na(data$Reason.for.absence)] = median(data$Reason.for.absence, na.rm = T)

data$Distance.from.Residence.to.Work[is.na(data$Distance.from.Residence.to.Work)] = median(data$Distance.from.Residence.to.Work, na.rm = T)

data$Service.time[is.na(data$Service.time)] = median(data$Service.time, na.rm = T)

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

data$Social.drinker[is.na(data$Social.drinker)] = median(data$Social.drinker, na.rm = T)

data$Pet[is.na(data$Pet)] = median(data$Pet, na.rm = T)

data$Month.of.absence[is.na(data$Month.of.absence)] = median(data$Month.of.absence, na.rm = T)

data$Weight[is.na(data$Weight)] = median(data$Weight, na.rm = T)

# Data Manipulation

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

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

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

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

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

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

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

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

# Separating continuous and categorical variables

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

numeric\_data = data[,numeric\_index]

factor\_index = sapply(data, is.factor)

factor\_data = data[,factor\_index]

### Outlier Analysis ###

cnames = colnames(numeric\_data)

for (i in 1:length(cnames)){

assign(paste0("gn", i), ggplot(aes\_string(y = (cnames[i]), x = "Absenteeism.time.in.hours"), data = numeric\_data)+

stat\_boxplot(geom = "errorbar", width = 0.5 ) +

geom\_boxplot(outlier.colour = "red", fill = "grey", outlier.shape = 18, outlier.size = 3, notch = FALSE) +

theme(legend.position = "bottom") +

labs(y=cnames[i], x= "Absenteeism.time.in.hours") +

ggtitle(paste("Boxplot for", cnames[i])))}

## Plotting plots together

gridExtra::grid.arrange(gn1,gn2,gn3,ncol=3)

gridExtra::grid.arrange(gn4,gn5,gn6, ncol=3)

gridExtra::grid.arrange(gn7,gn8,gn9, ncol=3)

gridExtra::grid.arrange(gn10,gn11,gn12, ncol=3)

# Saving dataframe as backup

df<- data

#data <- df

# Replace outliers with NA and imputing

for (i in cnames){

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

#print(length(val))

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

}

# Impute the NA's using KNN imputation

data = knnImputation(data, k=3)

### Feature Scaling ###

qqnorm(data$Distance.from.Residence.to.Work)

qqnorm(data$Transportation.expense)

qqnorm(data$Service.time)

qqnorm(data$Age)

qqnorm(data$Work.load.per.day) # Normal data

qqline(data$Work.load.per.day)

qqnorm(data$Hit.target)

qqnorm(data$Pet)

qqnorm(data$Weight)

qqnorm(data$Height)

qqnorm(data$Body.mass.index)

qqnorm(data$Absenteeism.time.in.hours)

### Feature Selection ###

## Correlation Plot

corrgram(data[,numeric\_index], order = FALSE,

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

## ANOVA Test

aov1 <- aov(data$Absenteeism.time.in.hours~data$Reason.for.absence)

aov2 <- aov(data$Absenteeism.time.in.hours~data$Month.of.absence)

aov3 <- aov(data$Absenteeism.time.in.hours~data$Day.of.the.week)

aov4 <- aov(data$Absenteeism.time.in.hours~data$Seasons)

aov5 <- aov(data$Absenteeism.time.in.hours~data$Disciplinary.failure)

aov6 <- aov(data$Absenteeism.time.in.hours~data$Education)

aov7 <- aov(data$Absenteeism.time.in.hours~data$Social.drinker)

aov8 <- aov(data$Absenteeism.time.in.hours~data$Social.smoker)

summary(aov1)

summary(aov2)

summary(aov3)

summary(aov4)

summary(aov5)

summary(aov6)

summary(aov7)

summary(aov8)

# Remove "Weight", "Day.of.week", "seasons","education", "social.smoker"

numeric\_data <- subset(numeric\_data, select = -c(Weight))

factor\_data <- subset(factor\_data, select= -c(Day.of.the.week, Seasons, Education, Social.smoker))

## Creating Dummy Variables for Catagorical Variables

factor\_new = dummy.data.frame(factor\_data, sep = "\_")

### Feature Scaling ###

## Normalizing the numeric variables

for (i in 1:12){

#print(i)

numeric\_data[,i] = (numeric\_data[,i] - min(numeric\_data[,i]))/

(max(numeric\_data[,i]) - min(numeric\_data[,i]))

}

### Sampling ###

df <- cbind(factor\_new, numeric\_data)

train\_index <- sample(1:nrow(df), 0.8 \* nrow(df))

train <- df[train\_index,]

test <- df[-train\_index,]

### Principal Component Analysis ###

pc <- prcomp(train,center = F,scale. = F )

names(pc)

# Output the mean of variables

pc$center

std\_dev <- pc$sdev

# Compute variance

pc\_var <- std\_dev^2

prop\_varex <- pc\_var/sum(pc\_var)

# Plot the resultant principal components.

biplot(pc, scale = 0)

# Scree plot

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

# Cumulative scree plot

plot(cumsum(prop\_varex), xlab = "Principal Component",

ylab = "Cumulative Proportion of Variance Explained",

type = "b")

train.data <- data.frame(Absenteeism.time.in.hours = train$Absenteeism.time.in.hours, pc$x)

train.data <- train.data[,1:43]

test.data <- data.frame(Absenteeism.time.in.hours = test$Absenteeism.time.in.hours, pc$x)

test.data <- test.data[,1:43]

### Decision Tree Model ###

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

# Predict test data

prediction\_DT <- predict(DT, test.data[,-1])

# Error Metric

regr.eval(test.data[,1], prediction\_DT, stats = c('mae','rmse','mape', 'mse'))

### Linear Regression Model ###

library(usdm)

vif(train.data[,-1])

vifcor(train.data[,-1], th = 0.9)

LM <- lm(Absenteeism.time.in.hours~., data= train.data)

summary(LM)

# Predict Test Data

prediction\_lm <- predict(LM, test.data[,-1])

# Error Metric

regr.eval(test.data[,1], prediction\_lm, stats = c('mae','rmse','mape', 'mse'))

### Random Forest Model ###

RF <- randomForest(Absenteeism.time.in.hours~., train.data, importance= TRUE, ntree = 500)

# Predict Test Data

prediction\_RF <- predict(RF, test.data[,-1])

# Error Metric

regr.eval(test.data[,1], prediction\_RF, stats = c('mae','rmse','mape', 'mse'))

library("scales")

library("psych")

library("gplots")

ggplot(data, aes\_string(x= data$Reason.for.absence, y= data$Absenteeism.time.in.hours)) +

geom\_bar(stat = "identity", fill = "Darkslateblue") + theme\_bw() +

xlab("Reason for absence") + ylab("Absent Hours") + scale\_y\_continuous(breaks = pretty\_breaks(n=10)) +

ggtitle("Absenteeism Analysis") + theme(text = element\_text(size = 15))

ggplot(data, aes\_string(x= data$Reason.for.absence)) +

geom\_bar(stat = "Count", fill = "Darkslateblue") + theme\_bw() +

xlab("Reason for absence") + ylab("Count") + scale\_y\_continuous(breaks = pretty\_breaks(n=10)) +

ggtitle("Absenteeism Analysis") + theme(text = element\_text(size = 15))

ggplot(data, aes\_string(x= data$Month.of.absence, y= data$Absenteeism.time.in.hours)) +

geom\_bar(stat = "identity", fill = "Darkslateblue") + theme\_bw() +

xlab("Months") + ylab("Absent Hours") + scale\_y\_continuous(breaks = pretty\_breaks(n=10)) +

ggtitle("Absenteeism Analysis") + theme(text = element\_text(size = 15))

ggplot(data, aes\_string(x= data$Day.of.the.week, y= data$Absenteeism.time.in.hours)) +

geom\_bar(stat = "identity", fill = "Darkslateblue") + theme\_bw() +

xlab("Day of Week") + ylab("Absent Hours") + scale\_y\_continuous(breaks = pretty\_breaks(n=10)) +

ggtitle("Absenteeism Analysis") + theme(text = element\_text(size = 15))

ggplot(data, aes\_string(x= data$Disciplinary.failure, y= data$Absenteeism.time.in.hours)) +

geom\_bar(stat = "identity", fill = "Darkslateblue") + theme\_bw() +

xlab("Disciplinary Failure") + ylab("Absent Hours") + scale\_y\_continuous(breaks = pretty\_breaks(n=5)) +

ggtitle("Absenteeism Analysis") + theme(text = element\_text(size = 12))

ggplot(data, aes\_string(x= data$Social.smoker, y= data$Absenteeism.time.in.hours)) +

geom\_bar(stat = "identity", fill = "Darkslateblue") + theme\_bw() +

xlab("Social Smoker") + ylab("Absent Hours") + scale\_y\_continuous(breaks = pretty\_breaks(n=5)) +

ggtitle("Absenteeism Analysis") + theme(text = element\_text(size = 12))

ggplot(data, aes\_string(x= data$Social.drinker, y= data$Absenteeism.time.in.hours)) +

geom\_bar(stat = "identity", fill = "Darkslateblue") + theme\_bw() +

xlab("Social Drinker") + ylab("Absent Hours") + scale\_y\_continuous(breaks = pretty\_breaks(n=5)) +

ggtitle("Absenteeism Analysis") + theme(text = element\_text(size = 12))

# C. Reference:

Edwisor- [www.edwisor.com/learning](http://www.edwisor.com/learning)

Analytics Vidhya- <https://www.analyticsvidhya.com/>

R-bloggers- <https://www.r-bloggers.com/>

Stack Overflow- <https://stackoverflow.com/>