PROJECT REPORT EMPLOYEE ABSENTEEISM Gourav Nandy

# 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. [Feature Selection](#_bookmark9) 12
   5. [Feature Scaling](#_bookmark10) 12
   6. [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. [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

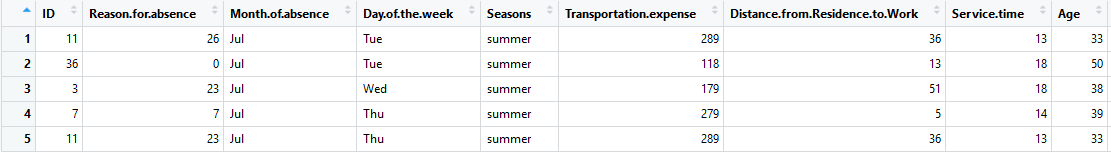
1. Diseases of the skin and subcutaneous tissue
2. Diseases of the musculoskeletal system and connective tissue XIV Diseases of the genitourinary system
3. Pregnancy, childbirth and the puerperium
4. Certain conditions originating in the perinatal period
5. Congenital malformations, deformations and 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.

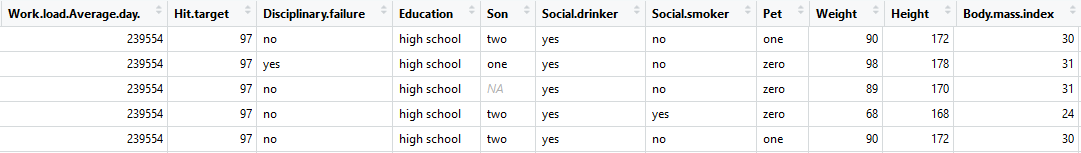
And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation

(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**

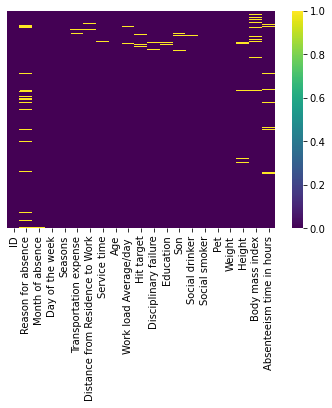
## Pre – Processing

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. The missing values have been computed using KNN computation method.

### Fig 2.2 – Missing value Columns and density

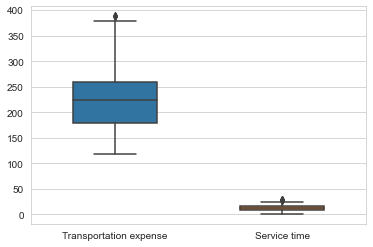
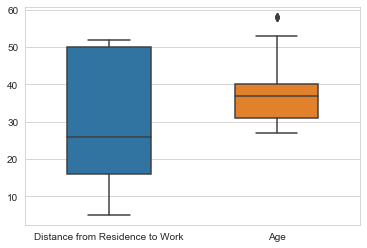


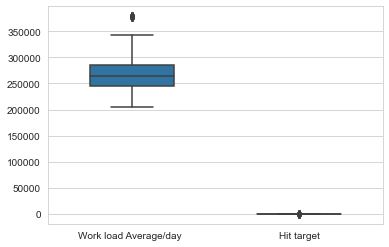
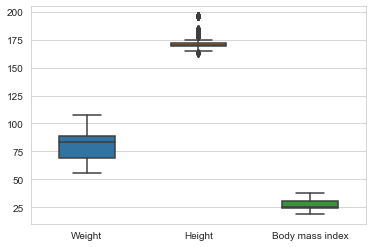
## Outlier Analysis

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

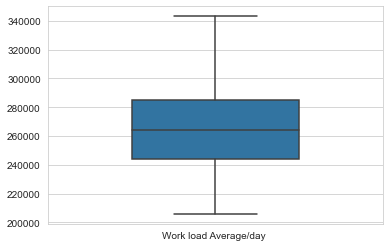
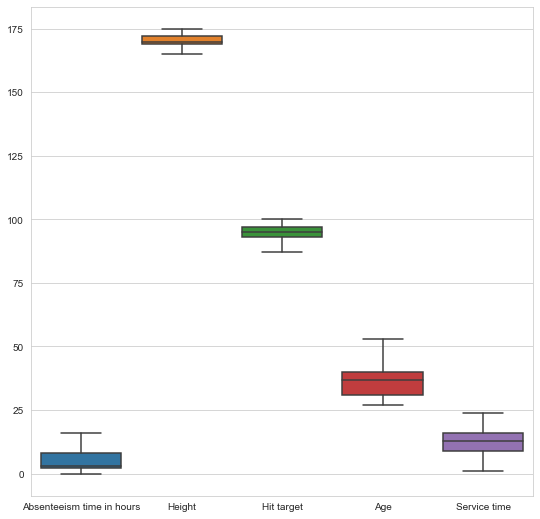
 

Imputing outlier values:

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.

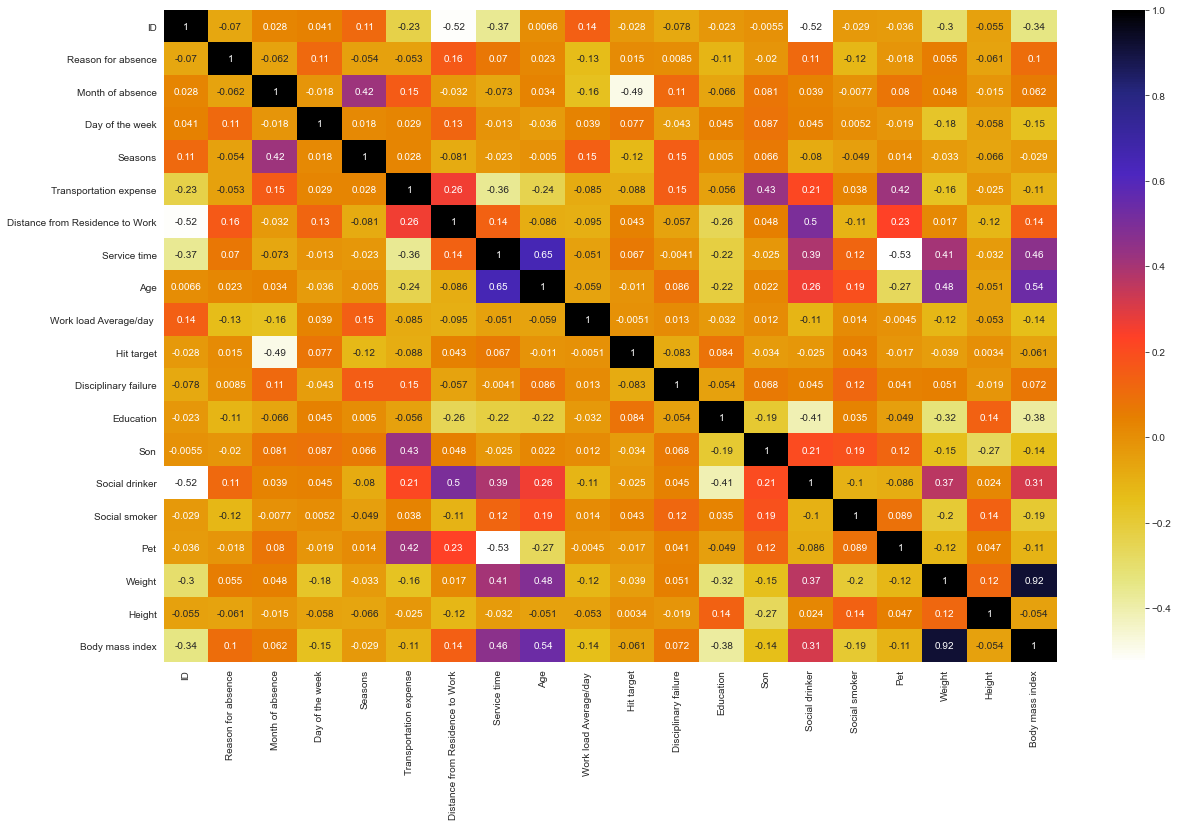
### Fig 2.3.2 – Boxplots of continuous variables without outliers

## 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 (>0.9), 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.

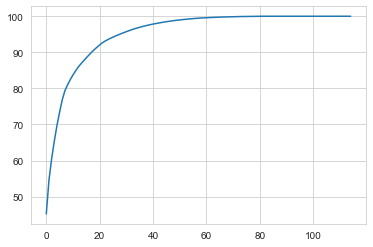
Most classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this feature. Therefore, the range of all features should be normalized so that each feature contributes proportionately to the

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 98+% of the data can be explained by 42 variables out of 116. Hence, we choose only 42 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).

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

|  |  |
| --- | --- |
| **DECISION TREE** | R squared value |
| R | 0.9877370 |
| PYTHON | 0.9613241597122618 |

## 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** | R^2 |
| R | 0.9809813 |
| PYTHON | 0.9999741345674866 |

## 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** | R^2 |
| R | 0.9999 |
| PYTHON | 0.9999 |

**Chapter 4: Conclusion**

## Model Evaluation

The coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

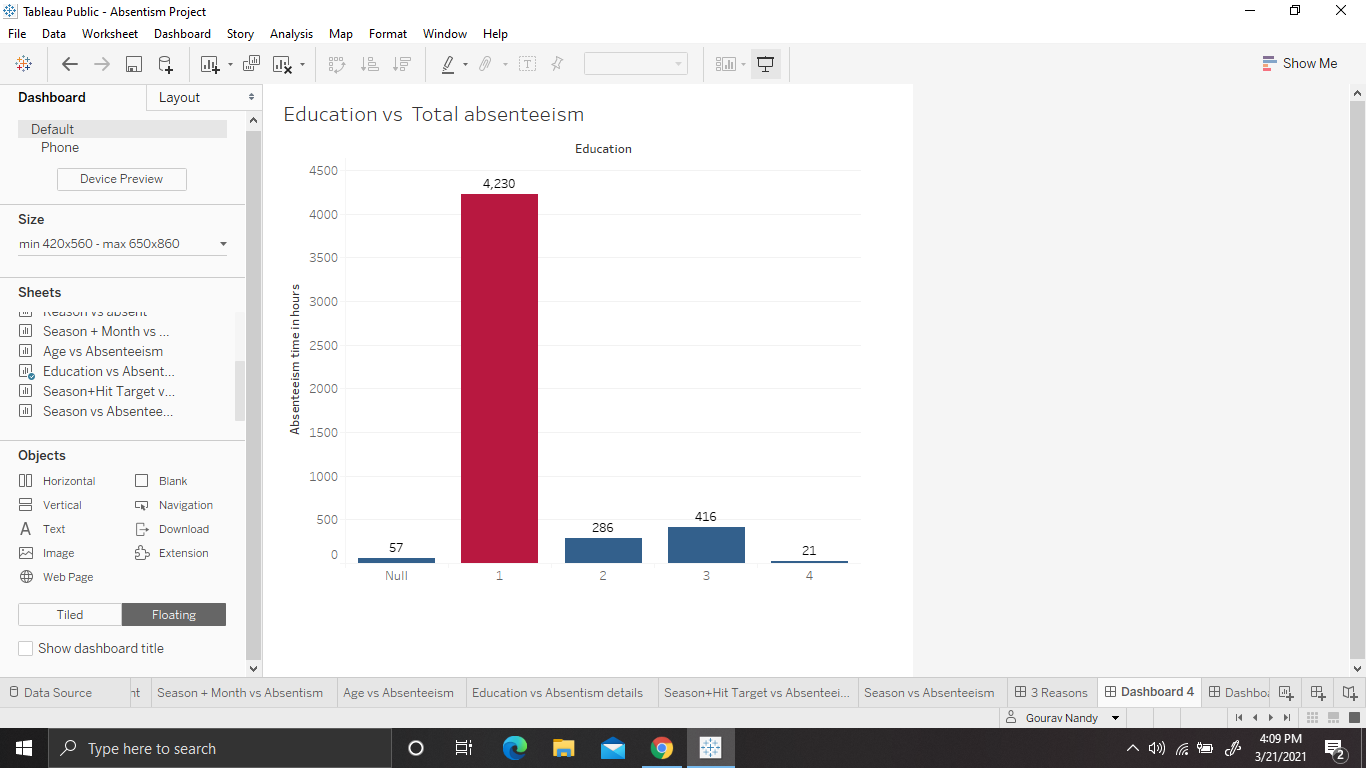
R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R^2 of 1 indicates that the regression predictions perfectly fit the data.

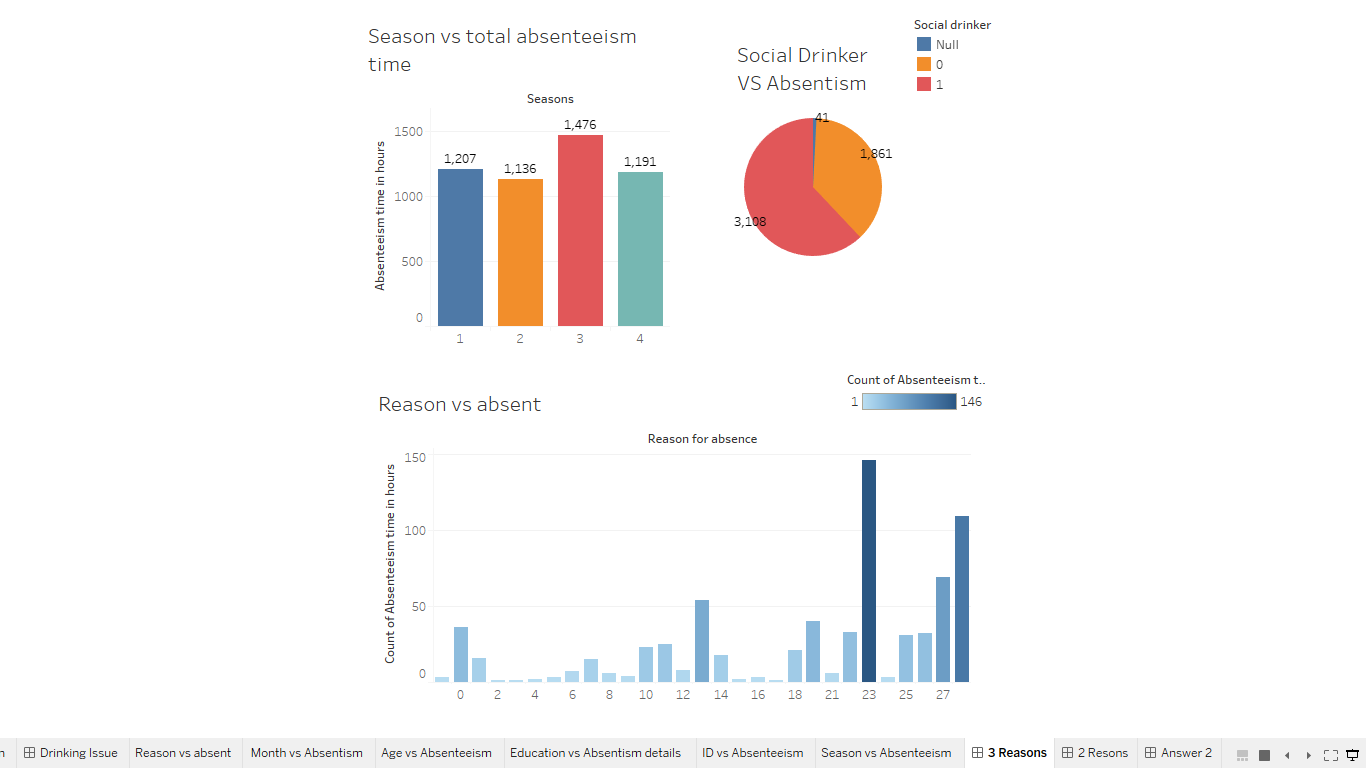
## Model Selection

From the observation of R-Squared Value we have concluded that **Linear Regression Model** has maximum value of its R-Squared Value.

## 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.
       2. 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.
       3. The reasons most used by employees to be absent are reason 13, 23 and 28. These reasons include Medical consultation, Dental appointments 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.
       4. 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.
       5. This data also shows that people tend to be absent more during winters.



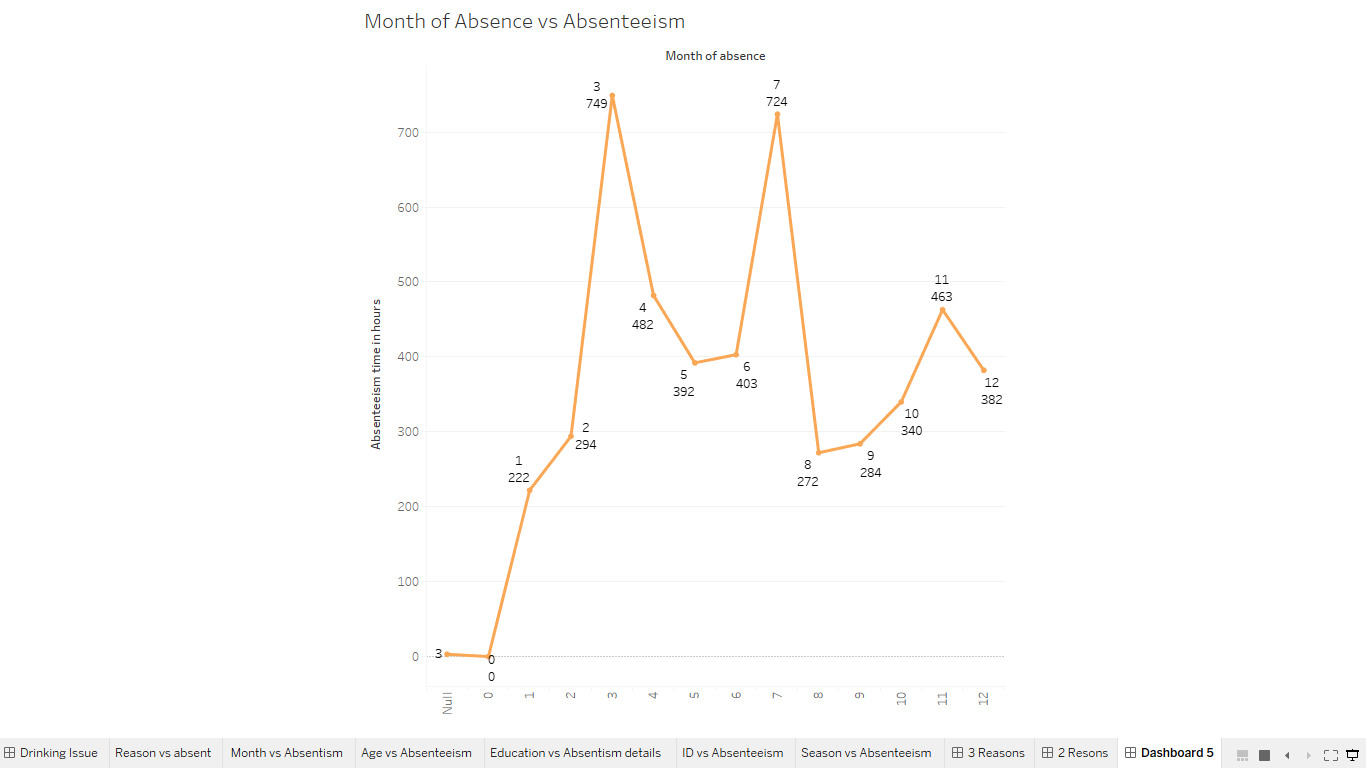


* + 1. How much losses every month can we project in 2011 if same trend of absenteeism continues?

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 749 hours. Employees are absent the least in the month of January, with total Absenteeism hours equal to 222.



### Fig 4.3.2 – Absenteeism Hours per Month

**Chapter 6:** **R code**

rm(list= ls())

setwd("E:/My career/Data Science/EDwisor/Projects/Employee Absentism")

getwd()

library(rpart)

library(readxl)

df <- read\_excel("Absenteeism\_at\_work\_Project.xls",

sheet = "Absenteeism\_at\_work")

View(df)

summary(df)

df$`Reason for absence`[df$`Reason for absence`== 0] <-NA

df$`Month of absence`[df$`Month of absence`== 0] <-NA

#\*\*\*\*\*\* Datatype Conversion\*\*\*\*\*\*

str(df)

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

df$`Reason for absence` = as.factor(df$`Reason for absence`)

df$`Month of absence` = as.factor(df$`Month of absence`)

df$`Day of the week`=as.factor(df$`Day of the week`)

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

df$`Disciplinary failure`=as.factor(df$`Disciplinary failure`)

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

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

df$`Social drinker`=as.factor(df$`Social drinker`)

df$`Social smoker`=as.factor(df$`Social smoker`)

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

#\*\*\*\*\*\*\*Missing Value Analysis

colSums(is.na(df))

total\_msv = sum(colSums(is.na(df)))

total\_value = nrow(df)\*ncol(df)

percentage\_msv = (total\_msv/total\_value)\*100.00

print(percentage\_msv)

library("VIM")

df = kNN(df, k=3, imp\_var = FALSE)

total\_msv = sum(colSums(is.na(df)))

total\_value = nrow(df)\*ncol(df)

percentage\_msv = (total\_msv/total\_value)\*100.00

print(percentage\_msv)

#Outlier Analys

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

numeric\_data = df[,numeric\_index]

numeric\_col = colnames(numeric\_data)

factor\_data = df[,!numeric\_index]

factor\_col = colnames(factor\_data)

for(i in numeric\_col){

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

print(paste(i,length(val)))

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

}

df = kNN(df, k=3, imp\_var = FALSE)

View(df)

#Multicolinearity

library("corrgram")

corrgram(df[,numeric\_col], order = F,upper.panel=panel.fill, text.panel=panel.txt, main = "Correlation Plot")

df$`Body mass index`=NULL

View(df)

#Feature Scaling

#Standardization

numeric\_col = numeric\_col[!numeric\_col %in% 'Body mass index'] #Removing colinear variable

numeric\_col = numeric\_col[-9]# dropping dependent variable

for(i in numeric\_col){

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

sd(df[,i])

}

View(df)

#Feature Engineering

library('data.table')

library('mltools')

df = one\_hot(as.data.table(df))

df = as.data.frame(df)

View(df)

str(df)

#\*\*\*\*\*\*\*\*\* Splitting data\*\*\*\*\*\*\*\*\*\*

library("caret")

set.seed(1234)

train.index = createDataPartition(df$`Absenteeism time in hours`, p = .80, list = FALSE)

train = df[ train.index,]

test = df[-train.index,]

#\*\*\*\*\*\*\*\*\* Dimension reduction \*\*\*\*\*\*\*

prin\_comp = prcomp(train)

pr\_stdev = prin\_comp$sdev

pr\_var = pr\_stdev^2

prop\_var = pr\_var/sum(pr\_var)

#Add a training set with principal components

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

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

ylab = "Cumulative Proportion of Variance Explained",

type = "b")

# From the above plot selecting 52 components since it explains almost 97+ % data variance

train.data =train.data[,1:52]

#Transform test data into PCA

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

test.data = data.frame(test.data)

#Select the first 52 components

test.data=test.data[,1:52]

#\*\*\*\*\*\*\*\*\* Model Development \*\*\*\*\*\*

#Linear Regression

print('Linear Regression')

lr\_model = lm(train.data$Absenteeism\_time\_in\_hours~. , data = train.data)

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

df\_pred = data.frame("actual"=test[,114], "lr\_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`))

# Decision Tree

print('Decision Tree')

library(rpart)

dt\_model = rpart(Absenteeism\_time\_in\_hours ~., data = train.data, method = "anova")

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

df\_pred = data.frame("actual"=test[,114], "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`))

#Random Forest

print('Random Forest')

library(randomForest)

rf\_model = randomForest(Absenteeism\_time\_in\_hours~., data = train.data, ntrees = 500)

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

df\_pred = data.frame("actual"=test[,114], "rf\_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`))

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

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