Employee Absenteeism Project

Pavan Kumar Nagula

Contents

Q2. How much loss every month can we project in 2011 if same trend of absenteeism continues?

Complete R Code

CHAPTER 1: INTRODUCTION

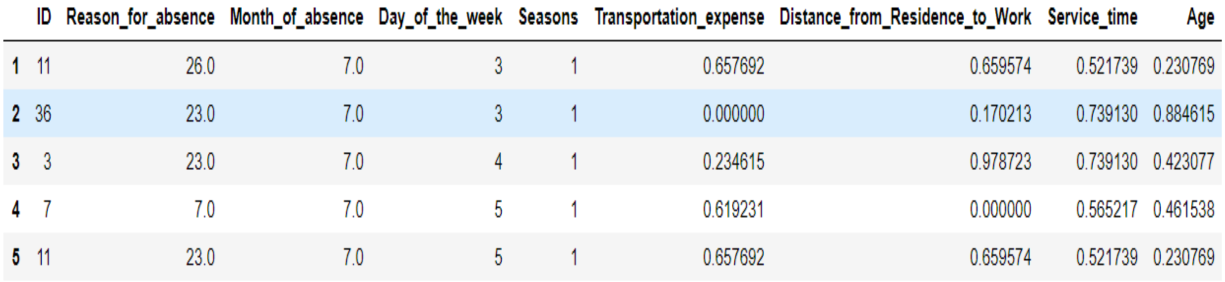
1.1 PROBLEM STATEMENT:

XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The company has shared it dataset and requested to have an answer on the following areas:

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

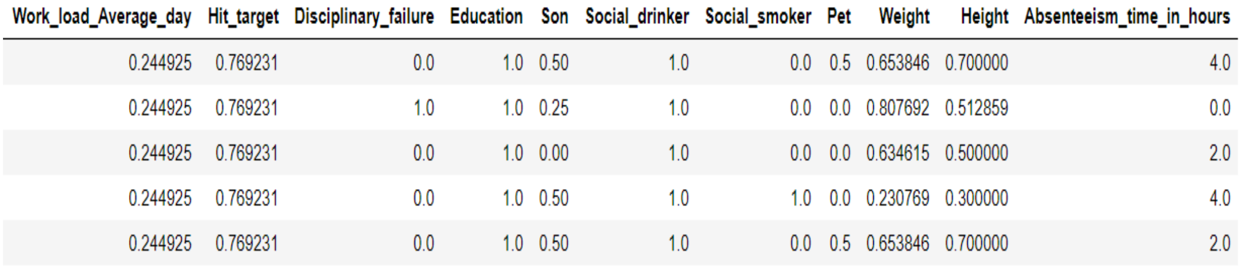
2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

1.2 OVERVIEW OF THE DATASET:



	ID	Reason_for_absence	Month_of_absence	Day_of_the_week	Seasons	Transportation_expense	Distance_from_Residence_to_Work	Service_time	Age
1	11	26.0	7.0	3	1	0.657692	0.659574	0.521739	0.230769
2	36	23.0	7.0	3	1	0.000000	0.170213	0.739130	0.884615
3	3	23.0	7.0	4	1	0.234615	0.978723	0.739130	0.423077
4	7	7.0	7.0	5	1	0.619231	0.000000	0.565217	0.461538
5	11	23.0	7.0	5	1	0.657692	0.659574	0.521739	0.230769

Fig 1.2 a First 1 to 9 columns from dataset



Work_load_Average_day	Hit_target	Disciplinary_failure	Education	Son	Social_drinker	Social_smoker	Pet	Weight	Height	Absenteeism_time_in_hours
0.244925	0.769231	0.0	1.0	0.50	1.0	0.0	0.5	0.653846	0.700000	4.0
0.244925	0.769231	1.0	1.0	0.25	1.0	0.0	0.0	0.807692	0.512859	0.0
0.244925	0.769231	0.0	1.0	0.00	1.0	0.0	0.0	0.634615	0.500000	2.0
0.244925	0.769231	0.0	1.0	0.50	1.0	1.0	0.0	0.230769	0.300000	4.0
0.244925	0.769231	0.0	1.0	0.50	1.0	0.0	0.5	0.653846	0.700000	2.0

Fig 1.2 b 10 to 21 columns from dataset

From the above figures we can see, that there are 8 categorical variables and 12 numeric variables. The target or dependent variable in the dataset is Absenteeism time in hours, which represents about the number of hours employees remaining absent.

Brief details about the data attributes in the dataset:

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 (kilometres)

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

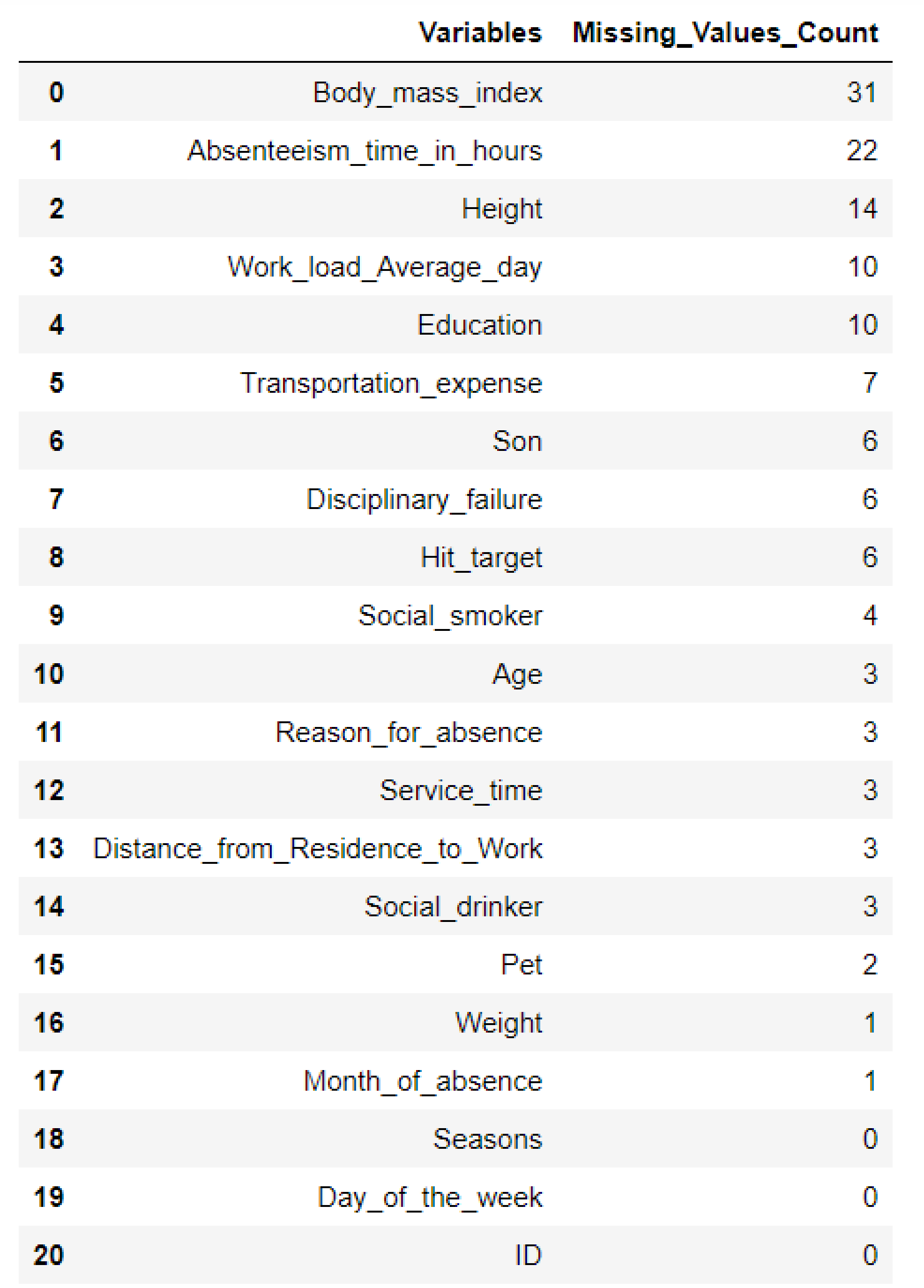
19. Height

20. Body mass index

21. Absenteeism time in hours (target)

CHAPTER 2: EXPLORATORY DATA ANALYSIS (EDA)

2.1 Missing Values Analysis:



	Variables	Missing_Values_Count
0	Body_mass_index	31
1	Absenteeism_time_in_hours	22
2	Height	14
3	Work_load_Average_day	10
4	Education	10
5	Transportation_expense	7
6	Son	6
7	Disciplinary_failure	6
8	Hit_target	6
9	Social_smoker	4
10	Age	3
11	Reason_for_absence	3
12	Service_time	3
13	Distance_from_Residence_to_Work	3
14	Social_drinker	3
15	Pet	2
16	Weight	1
17	Month_of_absence	1
18	Seasons	0
19	Day_of_the_week	0
20	ID	0

Fig 2.1: Missing Values in the variables of the Dataset

From the above fig we can see most of the missing values are observed in the BMI, Height and Absenteeism time in hours variables, and in rest of the variables we can see not more than 10 missing values,

1. After taking all the rows that contains missing values seems to be 90 rows contains all the 135 missing values, like one row may contains 1 or more number of missing values.

2. Dropping of these 90 rows will leads to loss of information, so instead of that performed correlation analysis and VIF to check the multi collinearity, results showed Body Mass Index is redundant variable, so dropped the Body Mass Index variable before imputing into any task.

3. I have used the Mice Method in R & mean and mode method in Python to replace these missing values.

2.2 Visualization of the data:

By using the visualization plots like barplots or histograms we can able to see the overview of the problem why employee absenteeism occurring and the reason behind it, And it can be solved easily by visualizing the data.

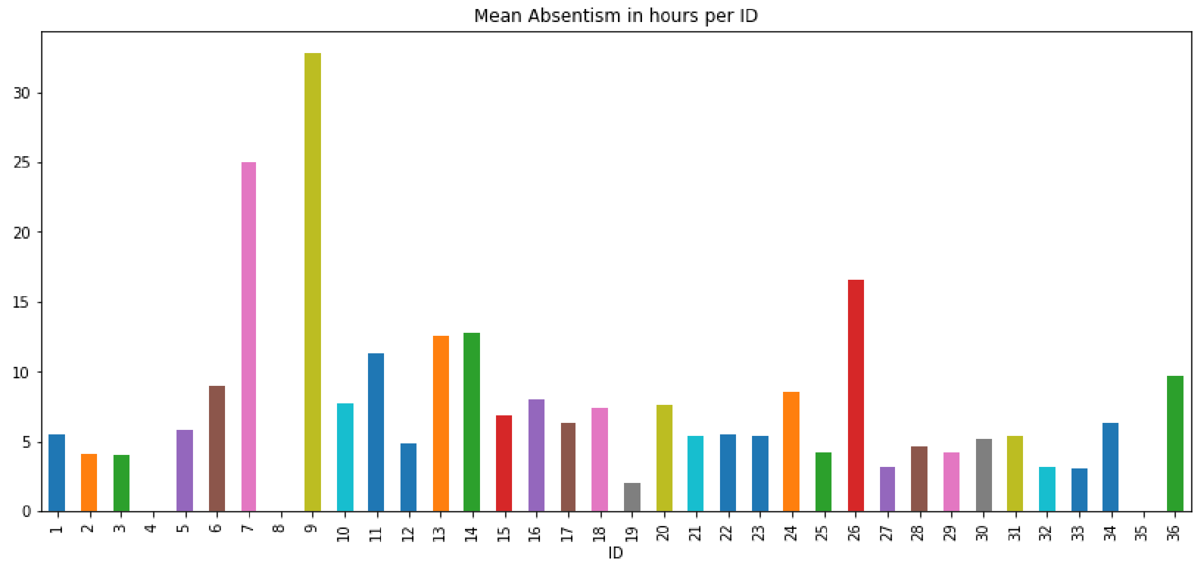


Fig 2.2 a Mean Absenteeism in hours per ID

The above fig represents the mean of hours that each particular employee was absent, from the fig we can say that employees with ID 9 and 7 had the greatest number of absent hours and Employees with ID 8 & 35 had the least number of absent hours.

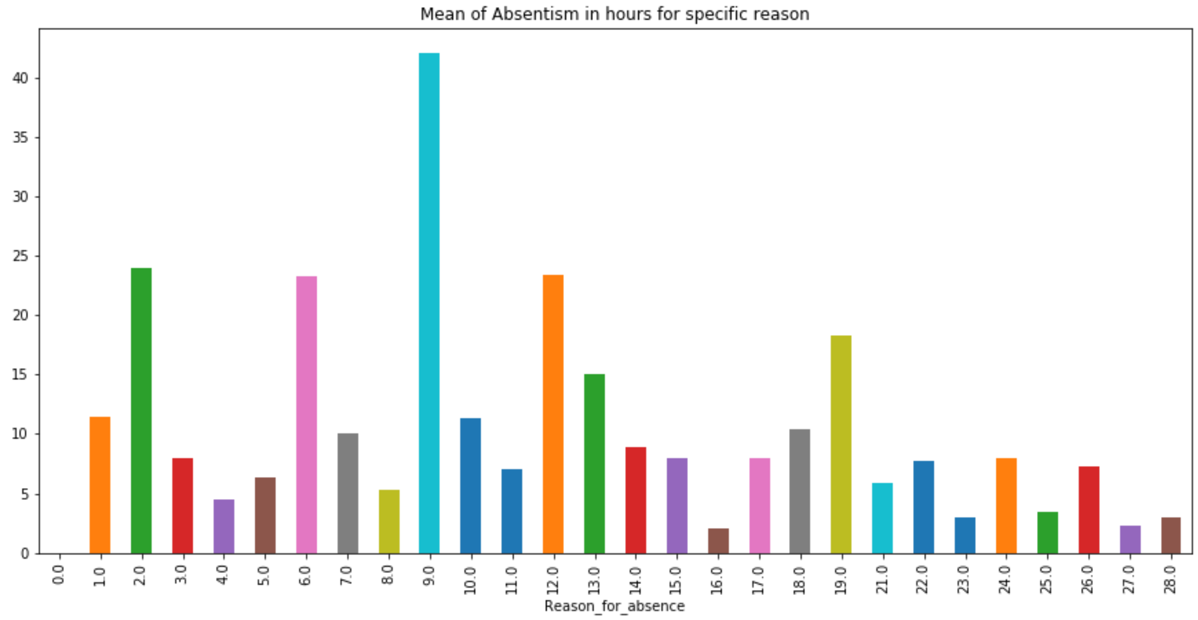


Fig. 2.2 c Mean of Absenteeism in hours for specific reason

The Above fig describes about the mean of hours each employee was absent for particular reason, we can say from the fig that reason 9(Diseases of the circulatory system) is the major reason that effected on the employee absent hours, and reason 16(Certain conditions originating in the perinatal period) will be the least reason specified by employees.

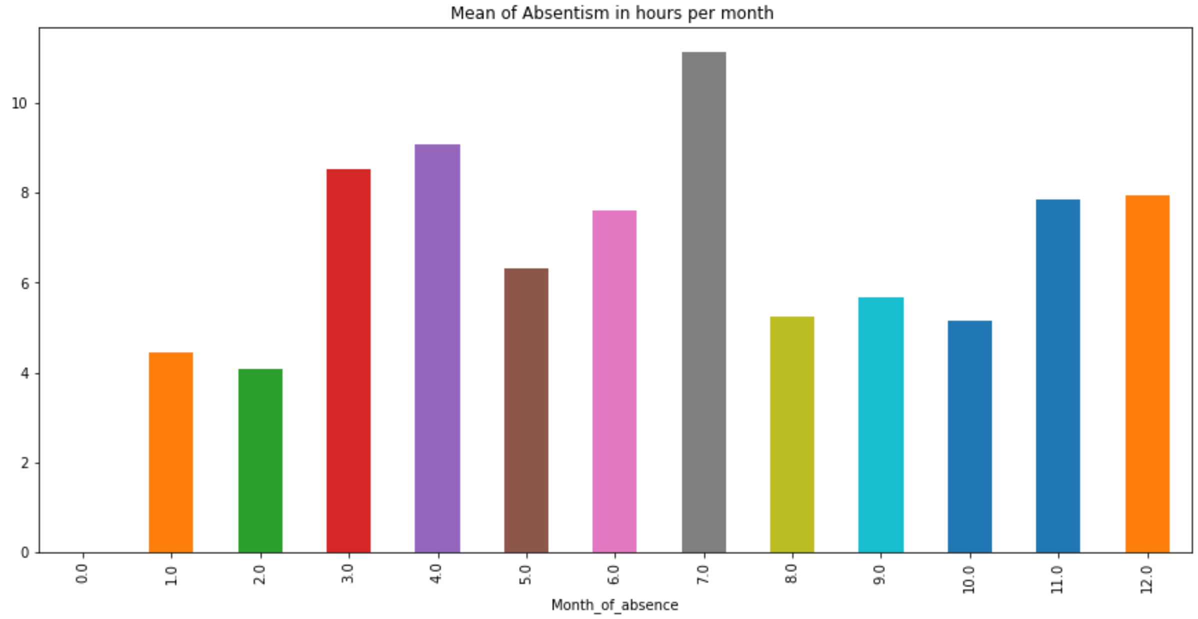


Fig 2.2 d Mean Absenteeism in hours per month

The above fig shows the mean of absenteeism in hours per month by the employees, from the fig It can be seen that most of the employees were absent in the month of July and followed by April and March.

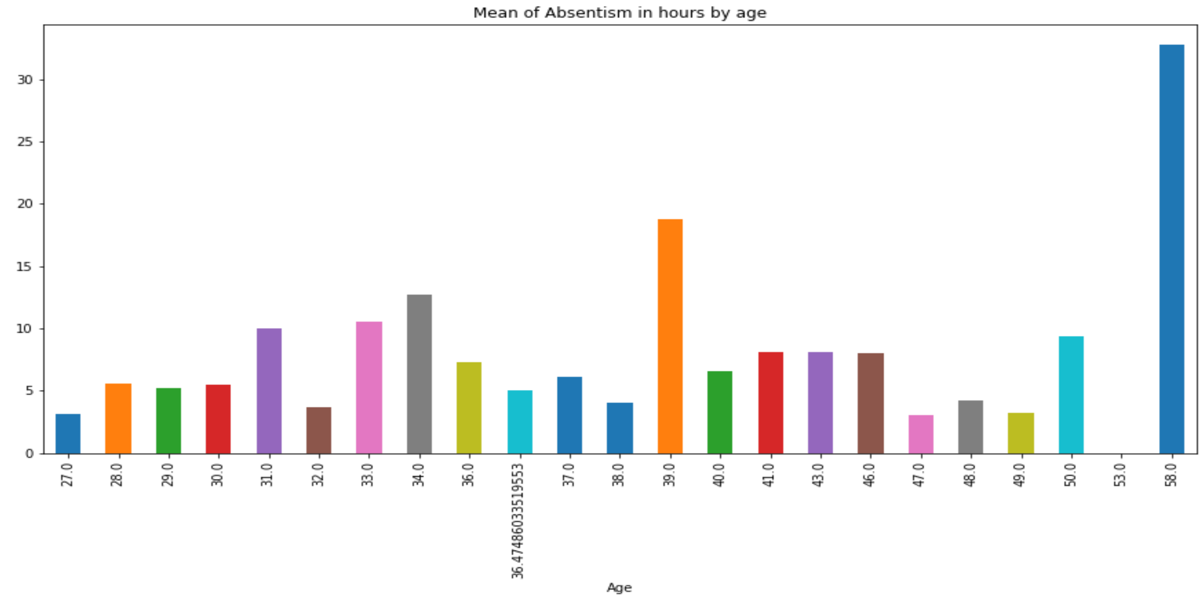


Fig 2.2 e Mean of Absenteeism in hours by Age

The above fig shows the different age group of employees and there absent hours, It seems to be the employees with age group of 58 are the highest in the absenteeism and least absenteeism hours would be the young age group employees.

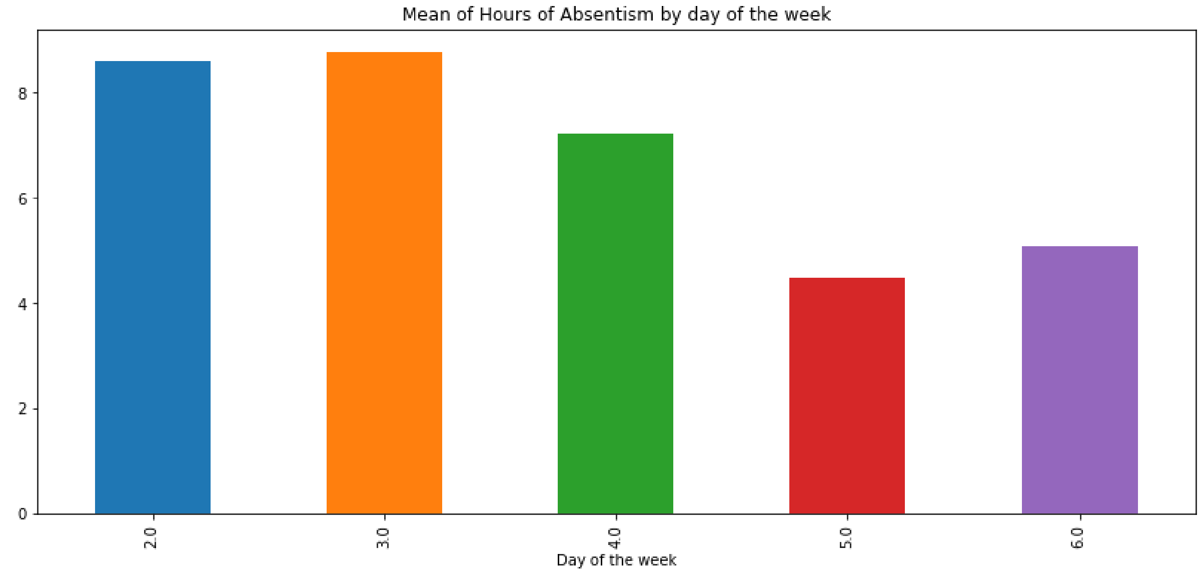


Fig 2.2 f Mean of hours of Absenteeism by day of the week

From the above fig we can say that most of the employees took leave or absent mostly in Monday or Tuesday of the week.

2.3 Feature Selection

In this section for selection of the features or variables which provides useful information about the target variable I have performed Correlation Analysis for numerical data, VIF for checking if any multicollinearity is found and for variable importance used extra tree regressor.

2.3.1 Correlation Analysis:

I have plotted the correlation graph between the numerical variables of the dataset. From the below mentioned fig we can detect that there is positive correlation between the Body Mass Index variable and weight variable. So before applying these data for further analysis dropping the Body Mass Index variable will helps in smooth process.

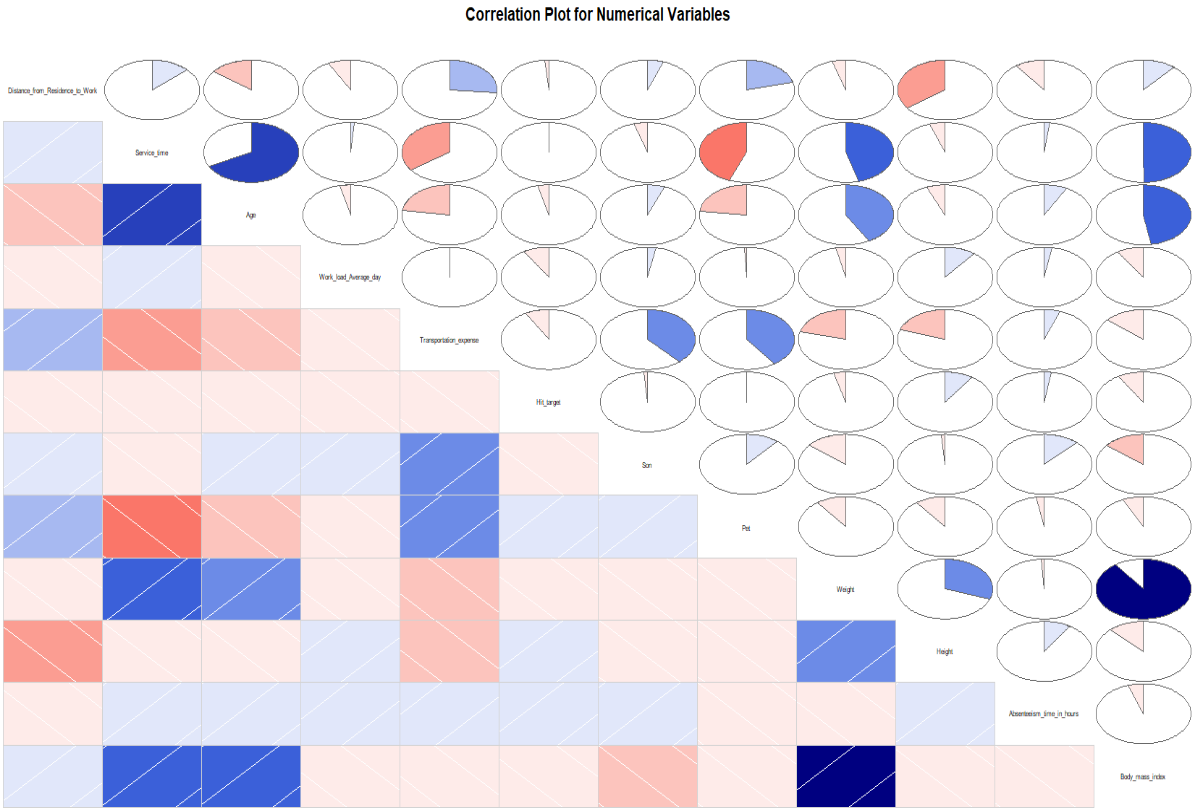
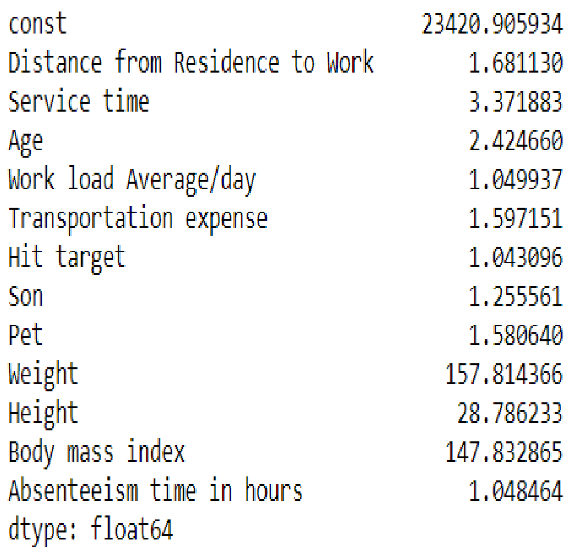


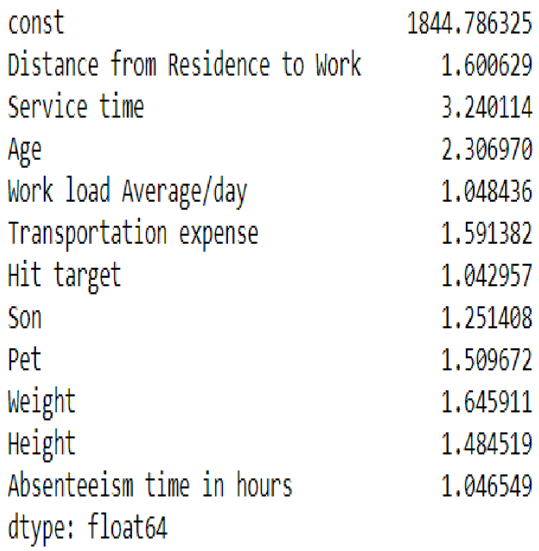
Fig 2.3 a Correlation Plot for numerical Variables

2.3.2 VIF (Variance Inflation Factor):

VIF is like extension of correlation analysis, in correlation analysis it can find only collinearity between only two variables, But, when it comes to VIF it can shows us the collinearity between two or more variables. Below figs will shows the results of before and after variable selection.



```
const                              23420.905934
Distance from Residence to Work        1.681130
Service time                           3.371883
Age                                    2.424660
Work load Average/day                  1.049937
Transportation expense                 1.597151
Hit target                             1.043096
Son                                    1.255561
Pet                                    1.580640
Weight                               157.814366
Height                                28.786233
Body mass index                      147.832865
Absenteeism time in hours              1.048464
dtype: float64
```



```
const                              1844.786325
Distance from Residence to Work       1.600629
Service time                          3.240114
Age                                   2.306970
Work load Average/day                 1.048436
Transportation expense                1.591382
Hit target                            1.042957
Son                                   1.251408
Pet                                   1.509672
Weight                                1.645911
Height                                1.484519
Absenteeism time in hours             1.046549
dtype: float64
```

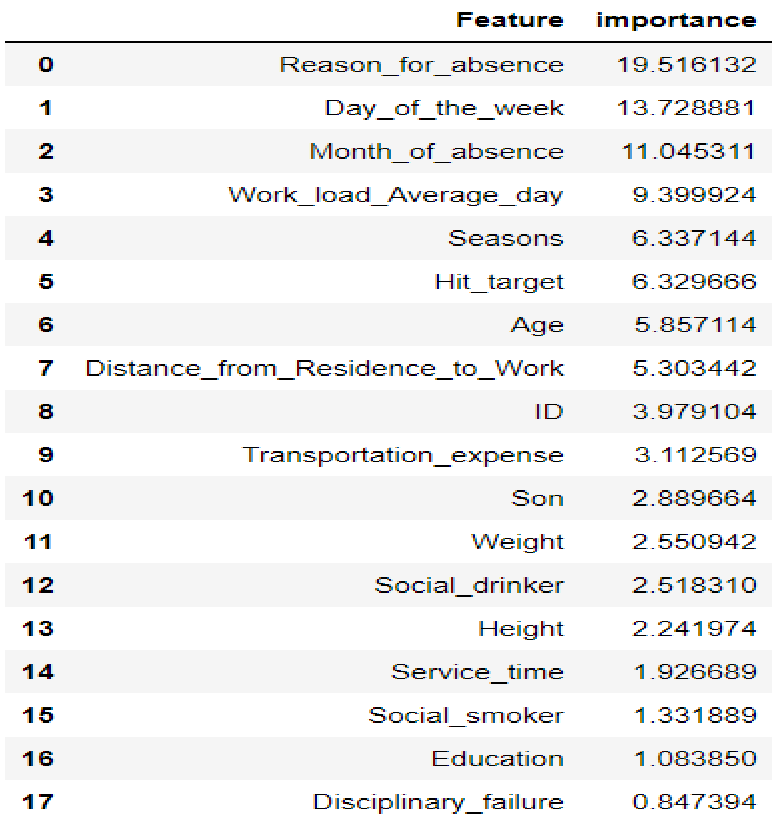
Fig 2.3(b)1 VIF before BMI removal

Fig 2.3(b)2 VIF after BMI removal

From the above fig we can say that the Body mass index variable, weight and height variables are correlated with each other, here VIF shows more than 2 variables correlation rather than the above correlation plot. So, after removal of Body mass index the r values inflation between the weight and height variables seems to levelled.

2.3.3 Extra Tree Regressor:

The extra tree regressor method used to know the variable importance, below the fig shows



	Feature	importance
0	Reason_for_absence	19.516132
1	Day_of_the_week	13.728881
2	Month_of_absence	11.045311
3	Work_load_Average_day	9.399924
4	Seasons	6.337144
5	Hit_target	6.329666
6	Age	5.857114
7	Distance_from_Residence_to_Work	5.303442
8	ID	3.979104
9	Transportation_expense	3.112569
10	Son	2.889664
11	Weight	2.550942
12	Social_drinker	2.518310
13	Height	2.241974
14	Service_time	1.926689
15	Social_smoker	1.331889
16	Education	1.083850
17	Disciplinary_failure	0.847394

that the reason for absence, Day of the week and month of absence are the important features among all the variables with respect to target variable. In the bottom disciplinary failure, education doesn't provide much of information about the target variable. And also, we excluded the body mass index variable for correlation effect and not containing much information about target variable.

2.4 Outlier Analysis:

The outliers are detected in our dataset using the box plot method in which it detects the values which are higher than the upper quartile range and lower than the lower quartile range. I have mentioned below the box plots of each variable which contains the outliers.

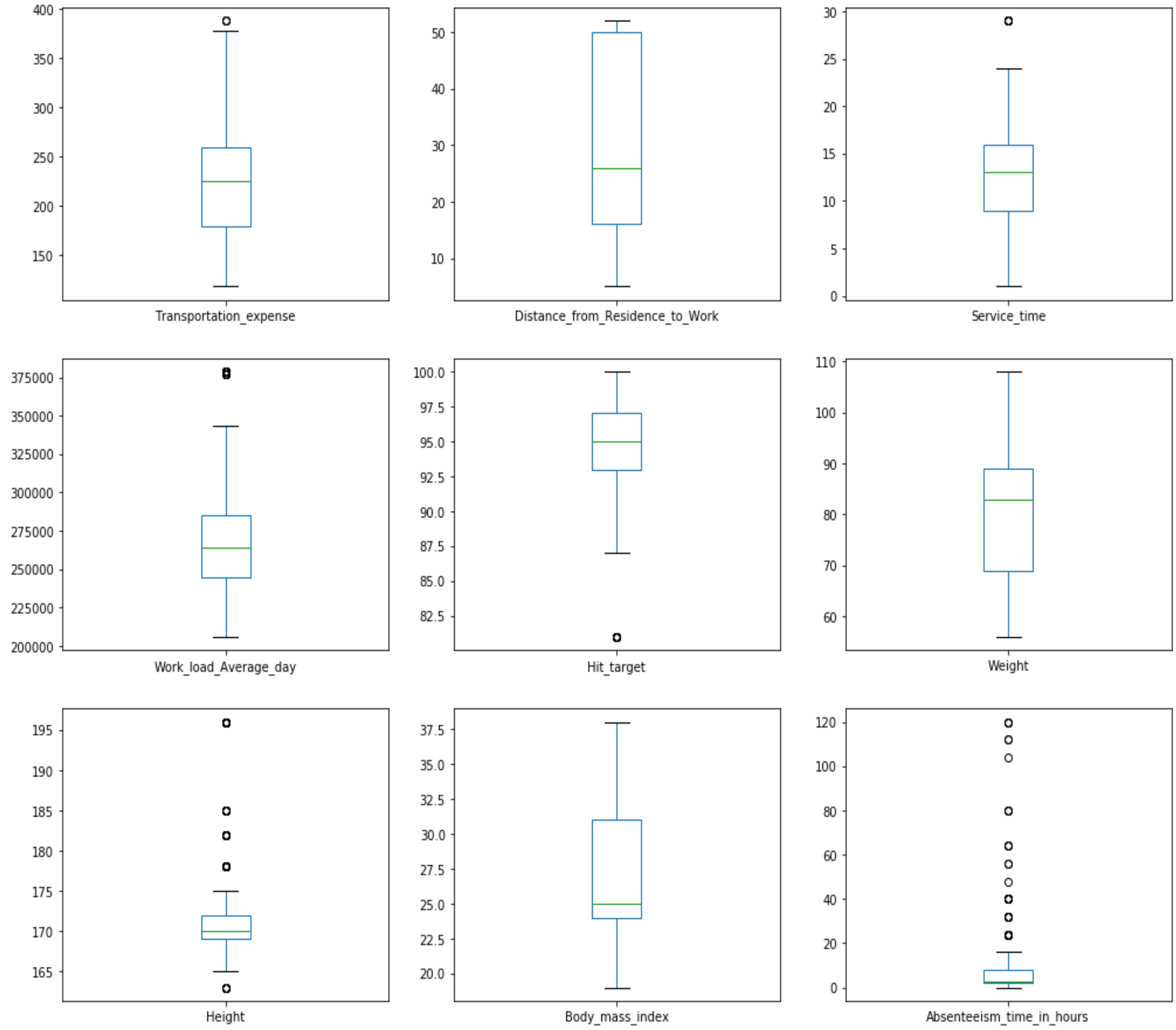


Fig 2.4 a Box plots of the numerical variables

From the above fig we can say that the outliers are highly present in the target variable and height variable. To deal with the outliers I have used Mean method in Python and MICE method in Rstudio.

2.5 Feature Selection:

In our dataset it is observed that many numerical variables having different scales of units which can’t be compared with the variables and also this data can’t be passed through models for prediction so I have used normalisation technique to make all the variables scale free and so that all the variables ranges between 0 and 1.

Normalization Formula:

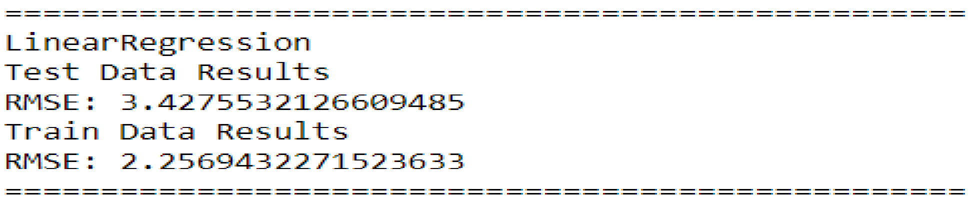
$$z = \frac{x \ \text{-} \min(x)}{\max(x) \text{ - } \min(x)}$$

CHAPTER 3: MODEL DEPLOYMENT

In the model deployment, Firstly I have split the data into test and train using random sampling method. And I have used all the regression models and to determine the performance of each model we can use error metrics in this project I have used RMSE error metric to distinguish the performance of all the models.

3.1 Linear Regression:

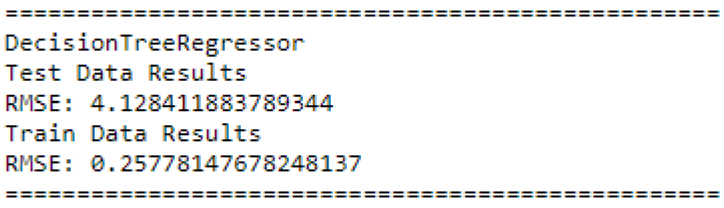
Linear Regression is one of the best regression model used for numeric data, below I have mentioned the results of linear regression and the results are satisfactory and acceptable values with good parameter values like R squared value is 0.483 and Adjusted R square value is 0.423 and RMSE value of 3.42.



```
==================================================
LinearRegression
Test Data Results
RMSE: 3.4275532126609485
Train Data Results
RMSE: 2.2569432271523633
==================================================
```

3.2 Decision Tree Regressor:

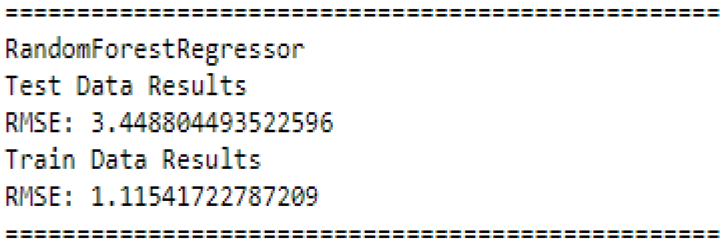
In the Decision Tree regressor the RMSE value obtained for train data is 0.2577 which is closest to the zero which seems to be that there is over-fitting problem in the model. So to overcome this problem we should go for regularised methods like Random Forest model.



```
==================================================
DecisionTreeRegressor
Test Data Results
RMSE: 4.128411883789344
Train Data Results
RMSE: 0.25778147678248137
==================================================
```

3.3 Random Forest Regressor:

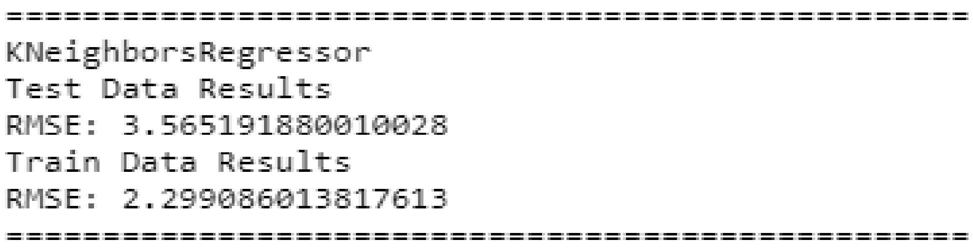
As we can see that there is over-fitting problem occurred in decision tree model, so to solve that problem we should go for regularised method which is Random Forest. And also regularised models are useful when we face over-fitting problems, although we didn’t get any over-fitting problems in linear regression model but the Random forest test and train rmse results are more accurate than LR model.



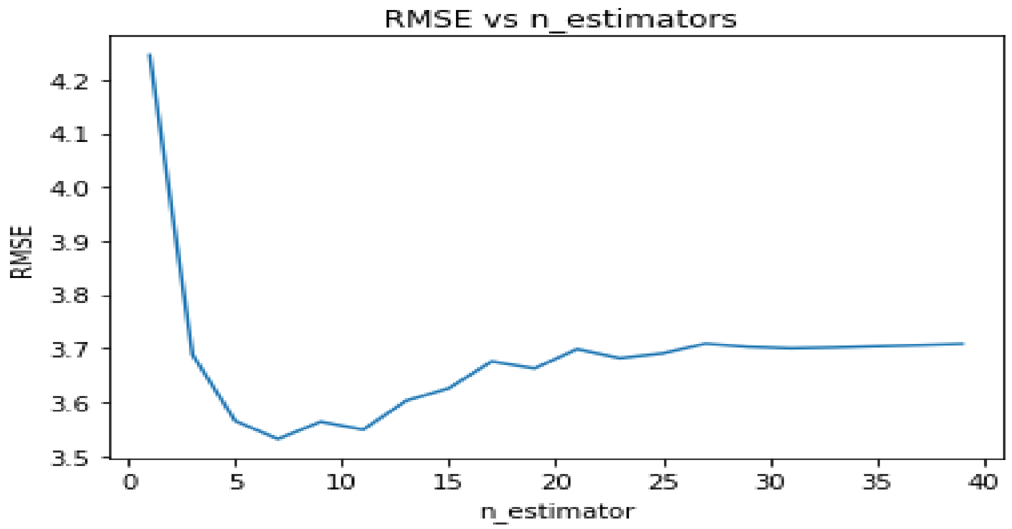
```
==================================================
RandomForestRegressor
Test Data Results
RMSE: 3.448804493522596
Train Data Results
RMSE: 1.11541722787209
==================================================
```

3.4 KNN

In KNN based on the K number of neighbours the accuracy of the model varies and I have mentioned the graph below between the RMSE values and K number of neighbours, As we can see for K = 6 we got the RMSE value of 3.565 and even though we increase the K value there is no improvement in the RMSE value and also we can consider this model result but not the best when compared with other model results.

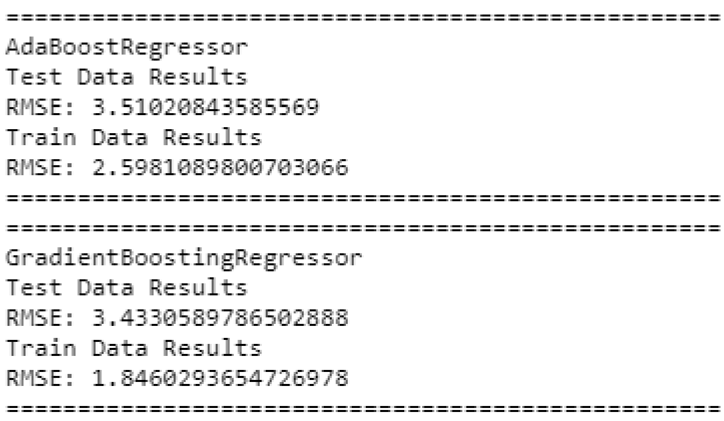


```
==================================================
KNeighborsRegressor
Test Data Results
RMSE:  3.565191880010028
Train Data Results
RMSE:  2.299086013817613
==================================================
```



3.5 ADA Boost Regressor and Gradient Boost Regressor

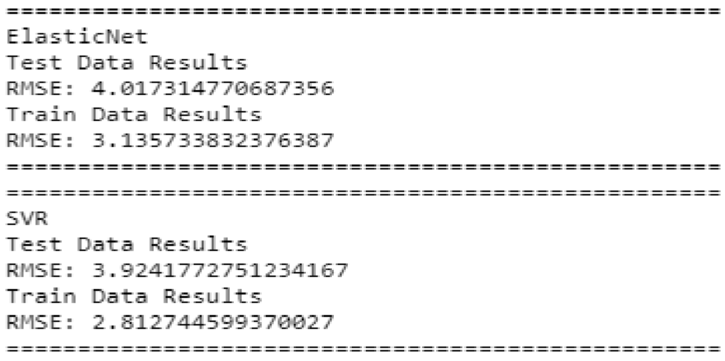
ADA boost and Gradient boost are boosting algorithms which means that they convert a set of weak learners into a single strong learner. We got RMSE value of 3.5 for ADA Boost Regressor and a RMSE value of 3.4 for Gradient Boost regressor which are also considerable.



```
==================================================
AdaBoostRegressor
Test Data Results
RMSE:  3.51020843585569
Train Data Results
RMSE:  2.5981089800703066
==================================================
==================================================
GradientBoostingRegressor
Test Data Results
RMSE:  3.4330589786502888
Train Data Results
RMSE:  1.8460293654726978
==================================================
```

3.6 SVR and Elastic Net Regressor

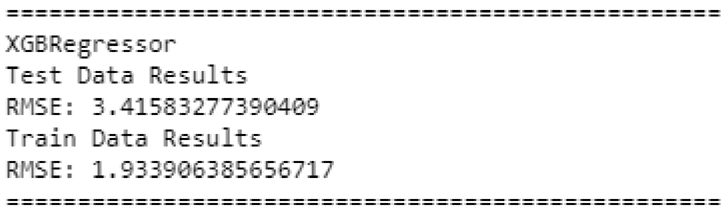
Although I have used both this models but the RMSE values of both these regressor are nearly ~4, So which can't be considered as when compared with other models RMSE values this values are higher.



```
==================================================
ElasticNet
Test Data Results
RMSE: 4.017314770687356
Train Data Results
RMSE: 3.135733832376387
==================================================
==================================================
SVR
Test Data Results
RMSE: 3.9241772751234167
Train Data Results
RMSE: 2.812744599370027
==================================================
```

3.7 XGB Regressor

XGB regressor is also one of the boost regressor which uses the computation techniques to speed up the gradient descent line component search; we have obtained RMSE value of 3.41 which is acceptable for this model.

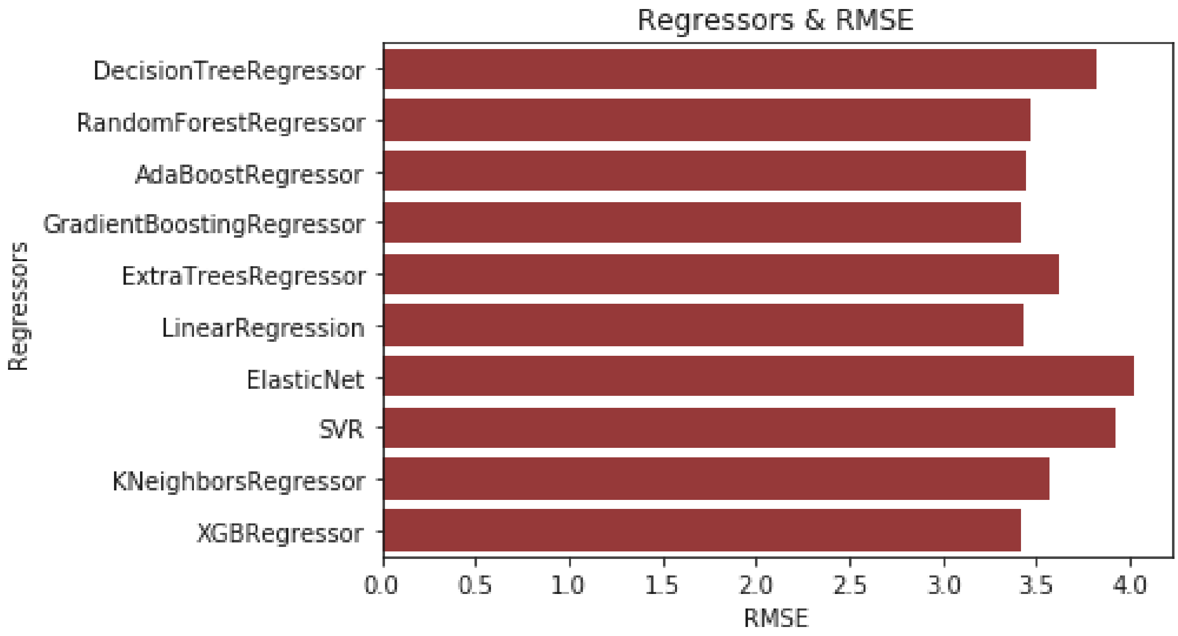


```
==================================================
XGBRegressor
Test Data Results
RMSE: 3.41583277390409
Train Data Results
RMSE: 1.933906385656717
==================================================
```

CHAPTER 4: Model Evaluation and Answers

4.1 Model Evaluation

Based upon all the above regression models we can come to the conclusion that Random Forest, boosting regressors and Linear Regression models performed well and based on their RMSE values we can consider any one of those models. And also, I have plotted graph between all the regressor models and there RMSE values.



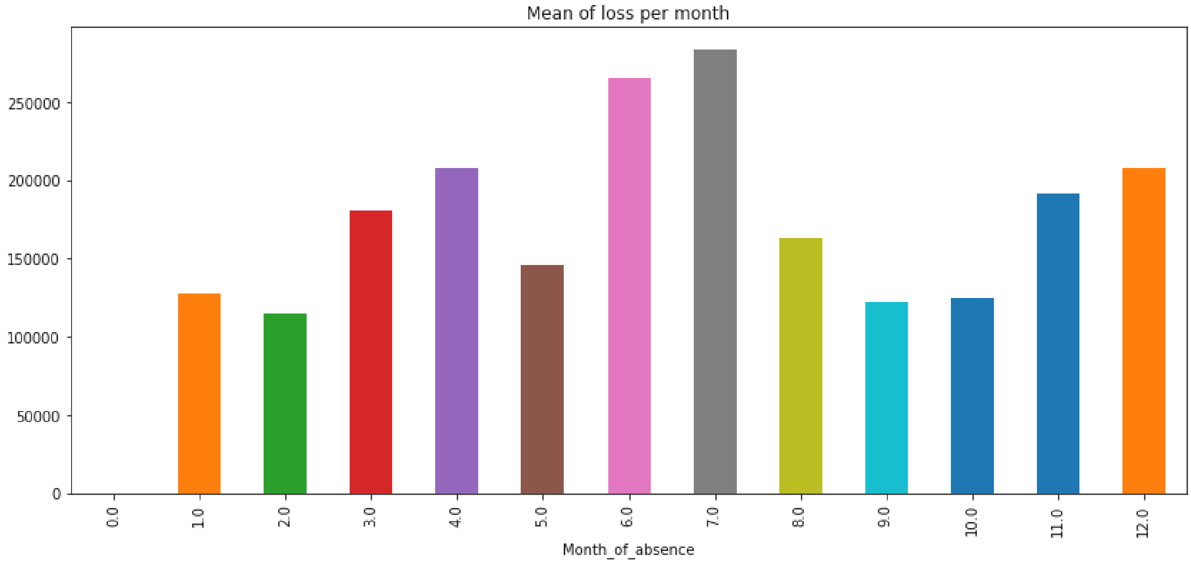
4.2 Answers:

Q1. What changes company should bring to reduce the number of absenteeism?

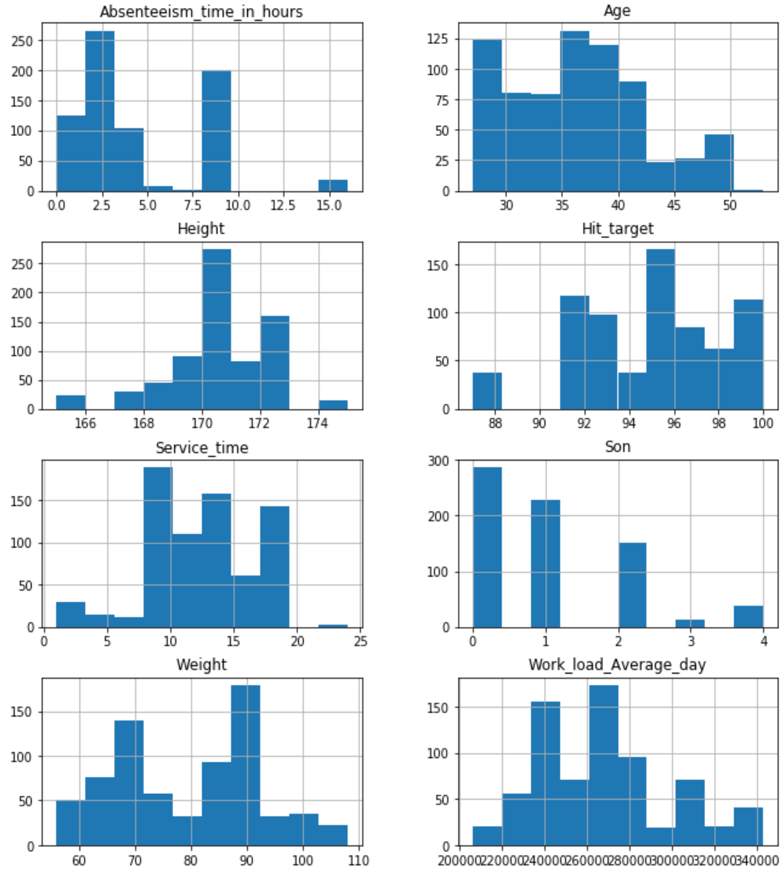
A: The company can reduce the number of absenteeism by making changes in the areas or sections where the employees got effected heavily like we can say from the visualisation graphs that major reason of absenteeism occurs due to health related issues, so if company provides any health check-ups or doctor's consultation for each employee twice every year will be able to reduce the major absenteeism issue and also It seems to be 50–58 years age group people were high in the absenteeism count when compared with the young age people. So, it is advisory to drop people whose age ranges form 50-58 yrs. and instead of them recruiting freshers or young people will decrease the number of absenteeism. The employees who are social drinkers are taking major leaves on the starting days of the week as it was followed by weekends, so to reduce it company should take necessary actions like implementing strict rules like coming in starting days of the week is mandatory. And other major thing is reducing the daily work load which will make each employee's work life balance and also will reduce the stress and reduces the absenteeism. There are situations like some of the employees haven't mentioned any reason for the absence so if the company takes necessary actions and implement strict rules against such type of employees will also be beneficial. These are major concerns I have spotted out and company will be able to reduce the number of absenteeism if they invest their time in solving the above-mentioned issues.

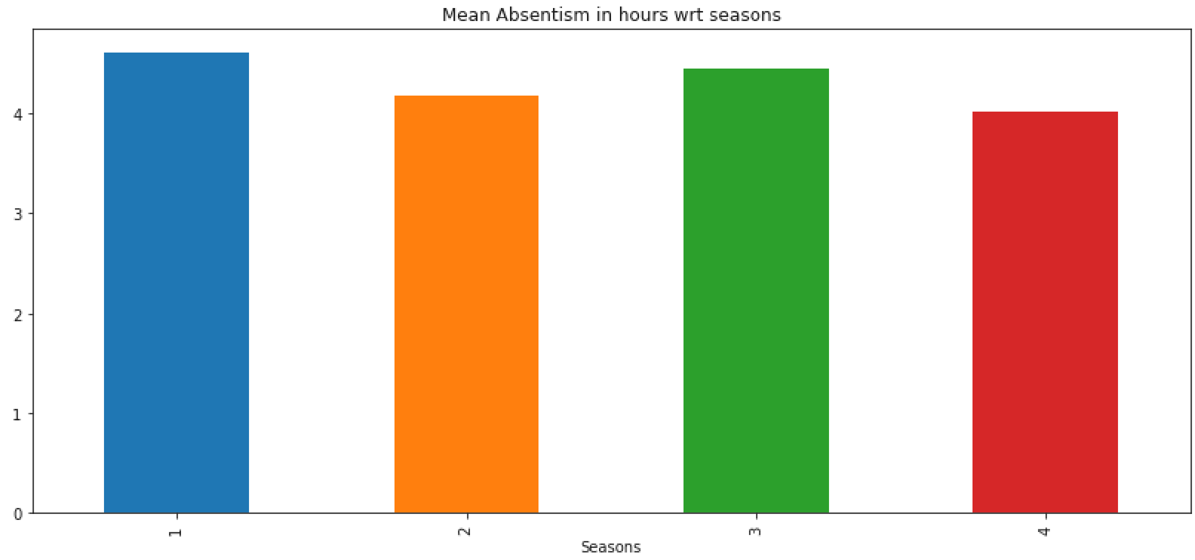
Q2. How much loss every month can we project in 2011 if same trend of absenteeism continues?

A: The loss per every month is plotted below if the same trend of absenteeism continuous and it seems to be that starting from June, July huge losses can be faced by the company. I Here I have calculated the loss by dividing the absenteeism time in hours by service time variable and multiplying it with the work load per day.



Extra Figures:





R CODE:

```
rm(list=ls())


x = c("ggplot2", "corrgram", "DMwR", "caret", "randomForest", "unbalanced",
"C50", "dummies", "e1071", "Information",
     "MASS", "rpart", "gbm", "ROSE", 'sampling', 'DataCombine', 'inTrees',
'readxl')


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

#Reading the Data
Empabs_data =
read_xls("C:/Users/npava/Desktop/Employee_Absentism_Project/Absenteeism_at_
work_Project.xls")

Empabs_data = as.data.frame(Empabs_data)

#Missing Value Analysis
missing_values = data.frame(apply(Empabs_data, 2, function(x)
{sum(is.na(x))}))
sum(missing_values)
new_DF <- Empabs_data[rowSums(is.na(Empabs_data)) > 0,]
Columns_DF <- Empabs_data[colSums(is.na(Empabs_data)) > 0,]
sum(is.na(Empabs_data))

cnames <- c('Distance_from_Residence_to_Work', 'Service_time', 'Age',
'Work_load_Average_day', 'Transportation_expense',
              'Hit_target', 'Son', 'Pet', 'Weight',
'Height','Absenteeism_time_in_hours','Body_mass_index')

cat_names <-
c('Social_smoker','Month_of_absence','Social_drinker','Reason_for_absence',
'Disciplinary_failure','ID','Education','Seasons','Day_of_the_week')


#Checking the summary
summary(Empabs_data[,cnames])
lapply(Empabs_data[,cnames], function(feat) length(unique(feat)))
str(Empabs_data)


for(i in cnames){
  Empabs_data[,i] = as.factor(Empabs_data[,i])
}

str(Empabs_data)

library(mice)

library('missForest')
Imputed_values = mice(Empabs_data, method = 'rf', seed = 1 )
Imputed_values_output = complete(Imputed_values)
anyNA(Imputed_values_output)
sum(is.na(Imputed_values_output))
```

```
write.csv(Imputed_values_output,
"C:/Users/npava/Desktop/Employee_Absentism_Project/Miceoutput.csv",
row.names = F )
miceOutput=read.csv
("C:/Users/npava/Desktop/Employee_Absentism_Project/Miceoutput.csv")

data = miceOutput


#Data exploration


library(ggthemes)
library(grid)
library(gridExtra)

p <- ggplot(data, aes(x = Pet, fill = Pet)) + geom_bar()
s <- ggplot(data, aes(x = Son, fill = Son)) + geom_bar()

SS <- ggplot(data, aes(x =  Social_smoker, fill =  Social_drinker)) +
geom_bar()

S <- ggplot(data, aes(x =   Seasons,fill = Seasons)) + geom_bar()

grid.arrange(p,s, nrow = 1)
grid.arrange(SS,S, nrow = 1)

library(dplyr)

absent <- as.data.frame( data %>% select(everything()) %>%
filter(Absenteeism_time_in_hours > 0))
Reason <-  as.data.frame(absent %>% group_by(Reason_for_absence) %>%
summarise(count= n(), percent = round(count*100/nrow(absent),1))%>%
arrange(desc(count)))
ggplot(Reason,aes(x = reorder(Reason_for_absence,percent), y= percent,
pos=3, xpd=NA, fill= Reason_for_absence)) + geom_bar(stat = 'identity') +
coord_flip() + theme(legend.position='none') +
  geom_text(aes(label = percent), vjust = 0.5, hjust = 1.1) + xlab('Reason
for absence')


a <- ggplot(data, aes(x = Age, fill = Son)) + geom_bar()
grid.arrange(a, nrow = 1)

##Outlier Analysis



for (i in 1:length(cnames))
{
  assign(paste0("gn",i), ggplot(aes_string(y = cnames[i]), data = data)+
           stat_boxplot(geom = "boxplot", width = 0.1)+
           geom_boxplot(outlier.colour="red", fill = "grey"
,outlier.shape=18,
                        outlier.size=2, notch=FALSE) +
           theme(axis.text.x =element_blank(),
                 axis.ticks.x=element_blank(),axis.title.y
=element_blank())+
           ggtitle(paste("Box plot of",cnames[i])))
}
```

```
gridExtra::grid.arrange(gn1,gn10,gn11,gn12,gn3,gn2,gn4,gn5,gn6,gn7,gn8,gn9,
nrow= 4, ncol=4)

for (i in cnames)
{
  print(i)
  value=data[,i][data[,i] %in% boxplot.stats(data[,i],coef=1.5)$out]
  print(value)
  data[,i][data[,i] %in% value] = NA
}

#Imputing outliers using MICE method
micemethod_2 <- mice(data, method="rf", seed=1)  # perform mice imputation,
based on random forests.
miceOutput_2 <- complete(micemethod_2)  # generate the completed data.
anyNA(miceOutput_2)
data = miceOutput_2

write.csv(miceOutput_2,
"C:/Users/npava/Desktop/Employee_Absentism_Project/Miceoutput_2.csv",
row.names = F )
write.csv(miceOutput,
"C:/Users/npava/Desktop/Employee_Absentism_Project/Miceoutput_1.csv",
row.names = F )


corrgram(data[,cnames], order = F, upper.panel=panel.pie,
text.panel=panel.txt, main = "Correlation Plot for Numerical Variables")

#ANOVA Analysis
anova_multi_way <- aov(Absenteeism_time_in_hours~(Social_smoker)+
                         (Month_of_absence)+
                         (Social_drinker)+
                         (Reason_for_absence)+
                         (Disciplinary_failure)+
                         (ID)+
                         (Education)+
                         (Seasons)+
                         (Day_of_the_week), data = data)
summary(anova_multi_way)

## Dimension Reduction
data = subset(data,select = -c(Body_mass_index,ID))


#Feature Scaling
cnames <- c('Distance_from_Residence_to_Work', 'Service_time', 'Age',
'Work_load_Average_day', 'Transportation_expense',
            'Hit_target', 'Son', 'Pet', 'Weight', 'Height')

cat_names <- c('Social_smoker',
               'Month_of_absence',
               'Social_drinker',
               'Reason_for_absence',
               'Disciplinary_failure',
               'Education',
               'Seasons',
               'Day_of_the_week')
for (i in cat_names){

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

#Scaling all numeric variables leaving our target variable untouched
for(i in cnames){
  print(i)
  data[,i] = as.numeric(data[,i])
  data[,i] = (data[,i] - min(data[,i]))/
    (max(data[,i] - min(data[,i])))
}
str(data)
anyNA(data)


#Removing the rows containing absurd information
data= data[!data$Month_of_absence==0 & !data$Reason_for_absence==0, ]

#Creating dummies for categorical variables
DFdummies <- as.data.frame(model.matrix(~. -1, data))
dim(DFdummies)

library(DataCombine)
rmExcept(c("marketing_train","DFdummies","miceoutput_2"))

wdf=data
data=DFdummies

#Creating train and test data
library(caret)
set.seed(1)
train.index = createDataPartition(data$Absenteeism_time_in_hours, p = .80,
list = FALSE)
X_train = data[ train.index,]
y_train  = data[-train.index,]
y_test=y_train$Absenteeism_time_in_hours
X_test=X_train$Absenteeism_time_in_hours

library(usdm)

#Creating a function to run all types of regression and comparing the
values
Show.RMSE = function(method, train_data, test_data){
  regressor_fit = caret::train(Absenteeism_time_in_hours~., data = X_train,
method = method)

  y_pred = predict(regressor_fit, y_train)
  print("RMSE value of test data")
  print(caret::RMSE(y_test, y_pred))
}
require(gbm)
library (ridge)
library(enet)
library(elasticnet)
library(h20)
regressors=c('lm','knn','svmLinear3', 'rpart2','rf','xgbTree','ridge')

#Running all the regressions and checking the performance on test data
for(i in regressors){
  print(i)
  Show.RMSE(i, X_train, y_test)
  print(strrep('-',50))
}
```