Project Report

Predicting Employee Absenteeism

NITISH ROHILLA

15TH MARCH 2019

Contents

1.1 PROBLEM STATMENT

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 DATA

Our task is to build regression models which will predict the employees absenteeism in hours depending on various factors. Given below is the sample of dataset using to predict the employees absenteeism in hours.

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	W
0	11	26.0	7.0	3	1	289.0	36.0	13.0	33.0	239554.0	...	0.0	1.0	2.0	1.0	0.0	1.0	
1	36	25.0	7.0	3	1	118.0	13.0	18.0	50.0	239554.0	...	1.0	1.0	1.0	1.0	0.0	0.0	
2	3	23.0	7.0	4	1	179.0	51.0	18.0	38.0	239554.0	...	0.0	1.0	0.0	1.0	0.0	0.0	
3	7	7.0	7.0	5	1	279.0	5.0	14.0	39.0	239554.0	...	0.0	1.0	2.0	1.0	1.0	0.0	
4	11	23.0	7.0	5	1	289.0	36.0	13.0	33.0	239554.0	...	0.0	1.0	2.0	1.0	0.0	1.0	

Weight	Height	Body mass index	Absenteeism time in hours
90.0	172.0	30.0	4.0
98.0	178.0	31.0	0.0
89.0	170.0	31.0	2.0
68.0	168.0	24.0	4.0
90.0	172.0	30.0	2.0

1. Individual identification (ID)

2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioural disorders
VI Diseases of the nervous system
VII Diseases of the eye and adnexa
VIII Diseases of the ear and mastoid process
IX Diseases of the circulatory system
X Diseases of the respiratory system
XI Diseases of the digestive system
XII Diseases of the skin and subcutaneous tissue
XIII Diseases of the musculoskeletal system and connective tissue
XIV Diseases of the genitourinary system
XV Pregnancy, childbirth and the puerperium
XVI Certain conditions originating in the perinatal period
XVII Congenital malformations, deformations and chromosomal abnormalities
XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX Injury, poisoning and certain other consequences of external causes
XX External causes of morbidity and mortality
XXI Factors influencing health status and contact with health services.

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

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

2.Methodologies

2.1 Pre-processing

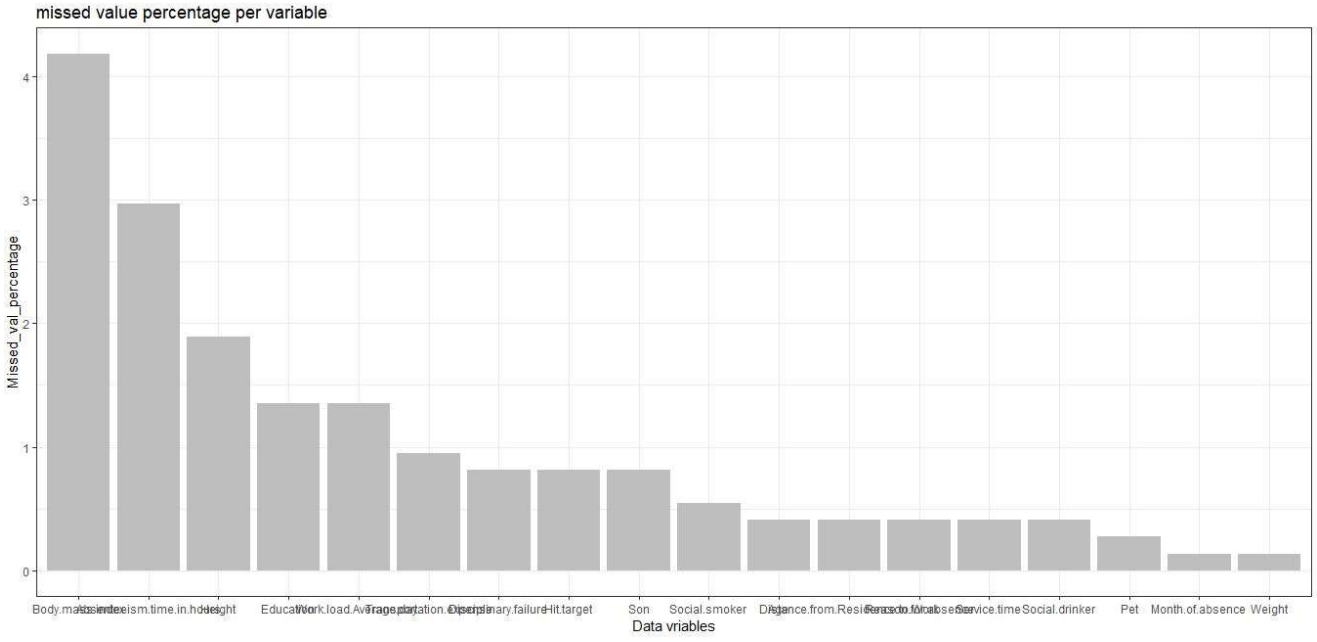
Exploratory Data Analysis is analysing the data sets to extract their characterstics. It is much more than just looking at the data but also to analyse, clean and to visualize through graphs and plots There are 740 observations and 21 columns in our data set. Missing value is also present in our data

```
'data.frame':   740 obs. of  21 variables:
 $ ID                              : num  11 36 3 7 11 3 10 20 14 1 ...
 $ Reason.for.absence              : num  26 0 23 7 23 23 22 23 19 22 ...
 $ Month.of.absence                : num  7 7 7 7 7 7 7 7 7 7 ...
 $ Day.of.the.week                 : num  3 3 4 5 5 6 6 6 2 2 ...
 $ Seasons                         : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Transportation.expense          : num  289 118 179 279 289 179 NA 260 155 235 ...
 $ Distance.from.Residence.to.Work: num  36 13 51 5 36 51 52 50 12 11 ...
 $ Service.time                    : num  13 18 18 14 13 18 3 11 14 14 ...
 $ Age                             : num  33 50 38 39 33 38 28 36 34 37 ...
 $ Work.load.Average.day.          : num  239554 239554 239554 239554 239554 ...
 $ Hit.target                      : num  97 97 97 97 97 97 97 97 97 97 ...
 $ Disciplinary.failure            : num  0 1 0 0 0 0 0 0 0 0 ...
 $ Education                       : num  1 1 1 1 1 1 1 1 1 3 ...
 $ Son                             : num  2 1 0 2 2 0 1 4 2 1 ...
 $ Social.drinker                  : num  1 1 1 1 1 1 1 1 1 0 ...
 $ Social.smoker                   : num  0 0 0 1 0 0 0 0 0 0 ...
 $ Pet                             : num  1 0 0 0 1 0 4 0 0 1 ...
 $ Weight                          : num  90 98 89 68 90 89 80 65 95 88 ...
 $ Height                          : num  172 178 170 168 172 170 172 168 196 172 ...
 $ Body.mass.index                 : num  30 31 31 24 30 31 27 23 25 29 ...
 $ Absenteeism.time.in.hours       : num  4 0 2 4 2 NA 8 4 40 8 ...
```

. As we anlysed data there were some zero values in month variable and in the reason variable which were were manually imputed according to the their relationship with season and the variable info given respectively, as we had very less number of observation in our data set, our approach was remove least no. of rows to ensure the high performance by the models.

2.2.1 Missing Value Analysis

As we saw from the EDA that there were some missing values in between the observations of the variables, it may be due to human error, or just didn't have the information or etc. To calculate the missing value percentage of each variable , we do so that for us it is important to calculate the missing value if it exceeds 30%. To analyse this on the data set we created the following plot to understand it better.

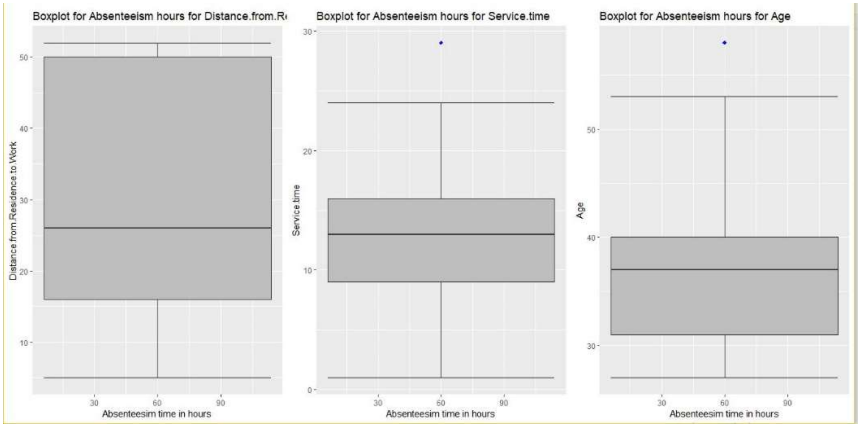


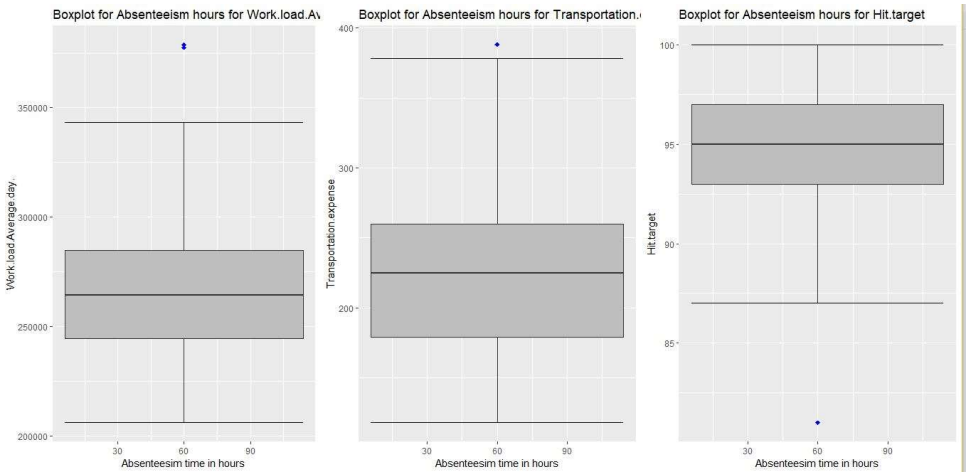
Now, to impute these missing values we choose one the imputation methods which suits the best for the given data set. To achieve this we manually remove one value and try imputation with each method i.e mean,median,KNN .The method which gives the closes value to the original value removed is chosen for the model

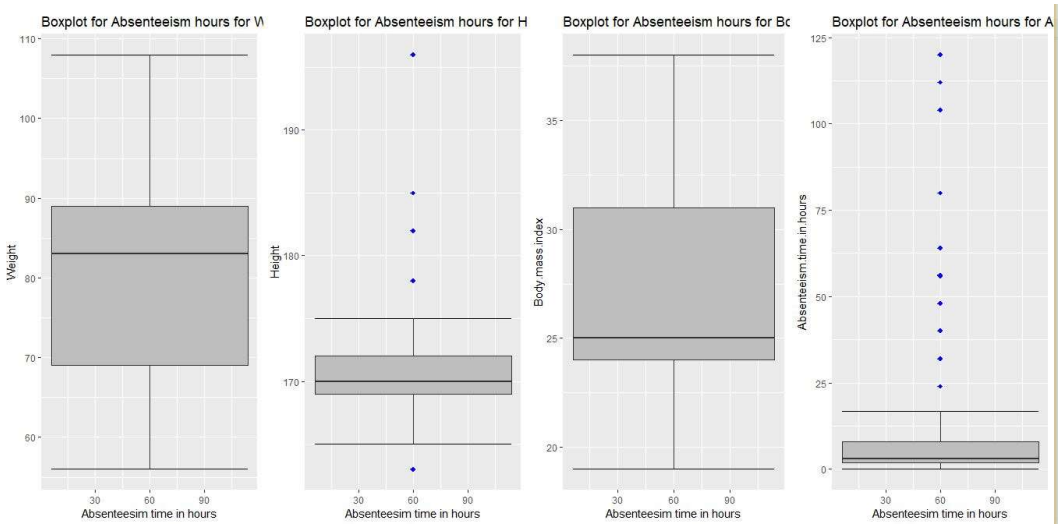
2.1.2 Outlier Analysis

Presence of outliers in the dataset can lead to the biased outcome of the model, which may effect the accuracy of the model developed.This is mostly explained by the presence of the extreme values, this could be solved by two approaches. These approached can be , eliminating the outliers out of the data set and Second is to replace the outliers with NA and then performing the imputation.

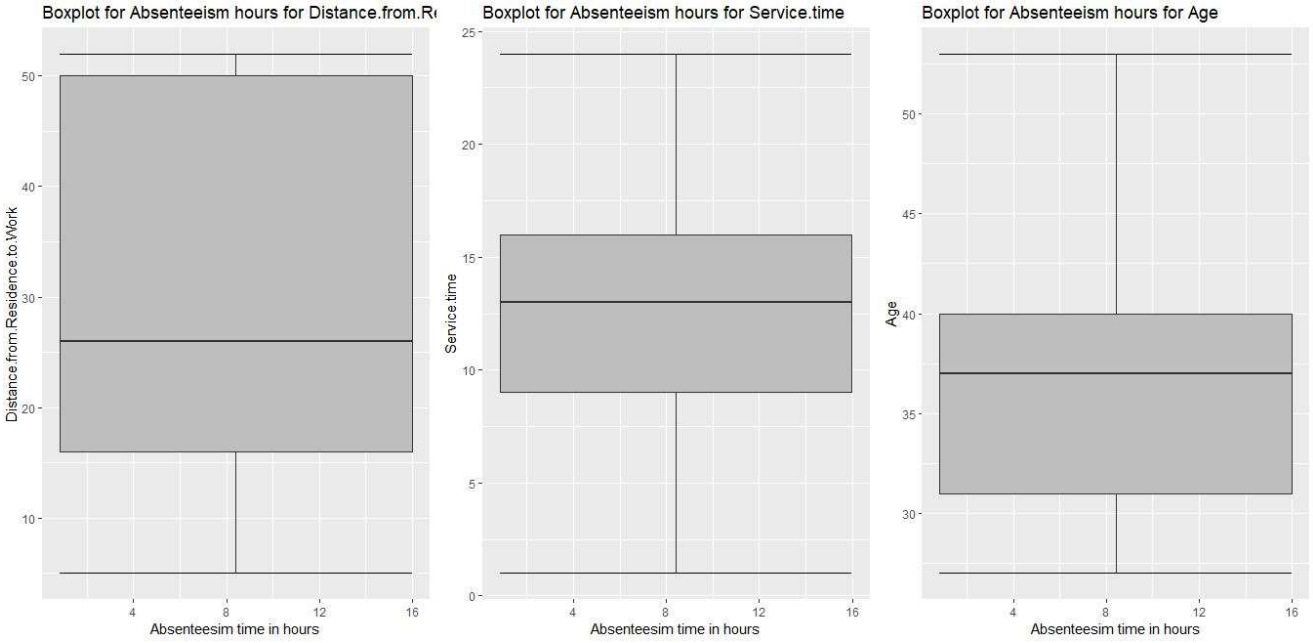
As you will see first we created the box plots for the variables with outliers.

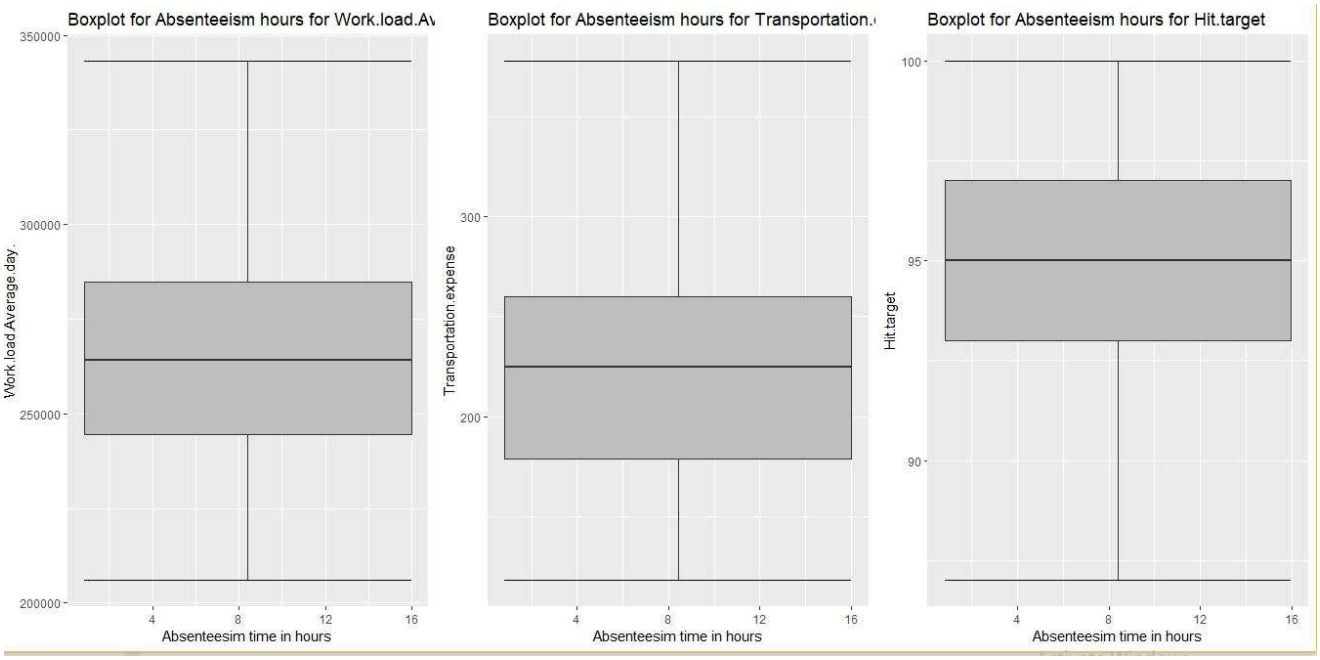


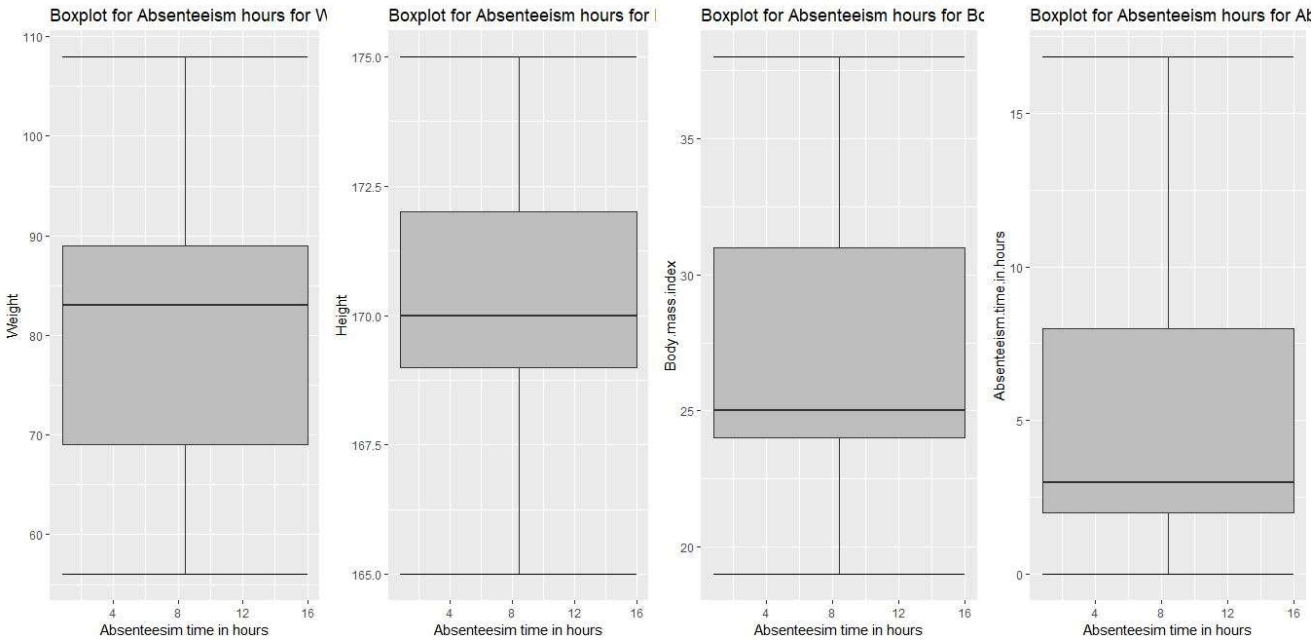




Now, forming the boxplots after replacing the outliers with NA and performing the imputation



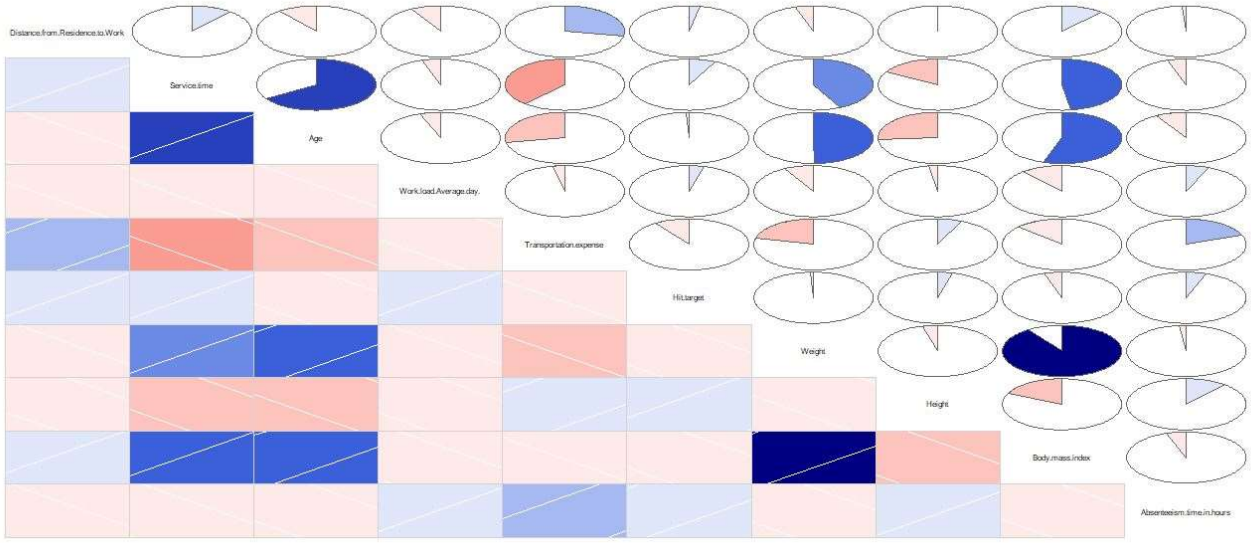




2.1.3 Feature Selection

This is the very important part of the pre-processing . It is very important to extract the meaningful components of the data that have been provided to increase the efficiency of the model also this may cost us redundency of data.There is no point of carrying the all the compenents along the way which provides same information to us which may cause increase in overhead. To deal it, we need to reduce the unnecessary components through selection technique of meaningfull variables out of data given. To achieve this we have chosen correlation analysis for the numerical variables and ANOVA for the categorical variables as the dependent variable in numerical.

correlation plot



For continuous variables it was clearly seen that weight had correlation more that 0.8 , so elimination took place . and for the categorical variables whose values were more than 0.05 i.e independent of dependent variable.

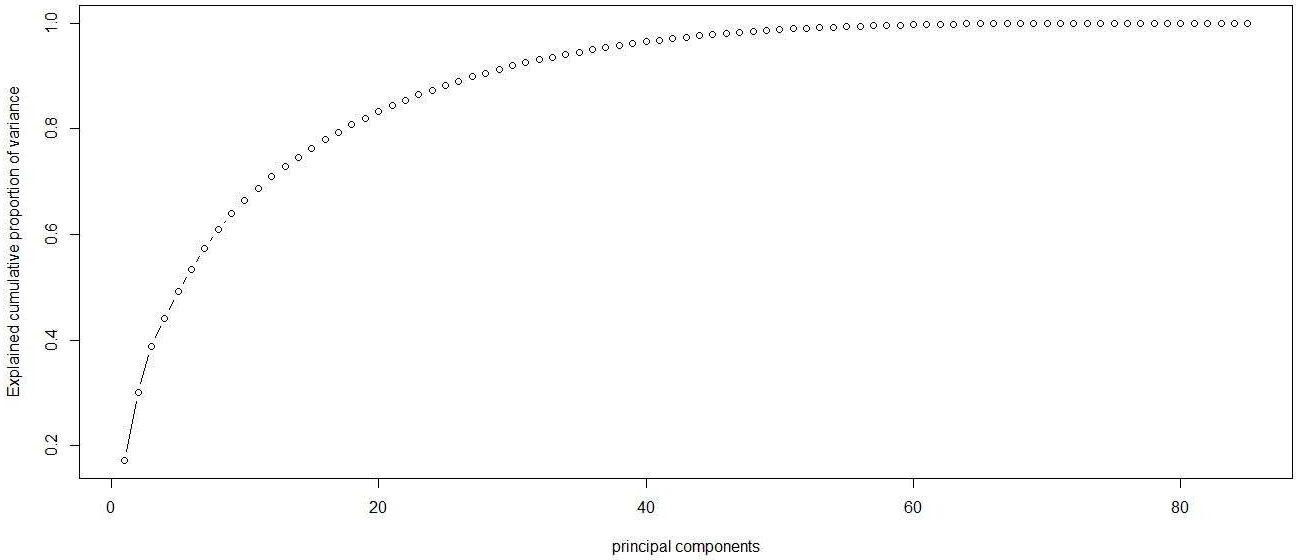
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
ID	1	154	153.69	13.23	0.000295	***
Residuals	738	8573	11.62			

2.1.4 Feature Scaling

To supress this effect, we need to bring all features to the same level of magnitudes. This can be acheived by scaling, It is a method to Standardize range of independent variable. It is also known as data normalization. and is generally performed during the data preprocessing step. For example, the majority of 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 particular feature. Therefore, it is necessary for range of features should be normalized so that every feature contributes proportionately to the final value

2.2.5 Principal Component Analysis

Principal component analysis (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called **principal components**. If there are observations with variables, then the number of distinct principal components is . This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.



We applied PCA algorithm on our data and concluded that 60 variables out of 84 explains more than 92% of data. So we have selected only those 60 variables.

2.2 Modeling

After a thorough preprocessing we will be variable. Following are the models which we have built –

2.2.1 Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each branch connects nodes with "and" and multiple branches are connected by "or". It can be used for classification and regression. It is a supervised machine learning algorithm. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users. Split of decision tree is seen in the below tree.

The RMSE value and R^2 value for our project in R and Python are –

Decision Tree	R	PYTHON
RMSE Test	0.146	0.368
R^2 Test	0.46	0.66

2.2.2 Random Forest

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data. The RMSE value and R^2 value for our project in R and Python are –

Random Forest	R	PYTHON
RMSE Test	0.0814	0.015
R^2 Test	0.8	0.86

2.2.3 Liner Regression

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. To build any model we have some assumptions to put on data and model. Here are the assumptions to the linear regression model.

Linear Regression	R	PYTHON
RMSE Test	0.002	3.14e-16
R^2 Test	0.999	1

2.2.4 KNN

Say we are given a data set of items, each having numerically valued features (like Height, Weight, Age, etc). If the count of features is *n*, we can represent the items as points in an *n*-dimensional grid. Given a new item, we can calculate the distance from the item to every other item in the set. We pick the *k* closest neighbors and we see where most of these neighbors are classified in.

Linear Regression	R	PYTHON
RMSE Test	0.12	0.376
R^2 Test	0.459	0.62

Chapter 3
Conclusion

3.1 Model Evaluation

In the model we have Root Mean Square Error (RMSE) and R-Squared Value of 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. And we conclude that lower values of RMSE and higher value of R-Squared Value indicate better fit.

3.2 Model Selection

We conclude from the model evaluation that the best suited model for the data set was linear regression, having lowest rmse value and highest r.sq value i.e it is not case of over-fitting.

3.2 Answers of asked questions

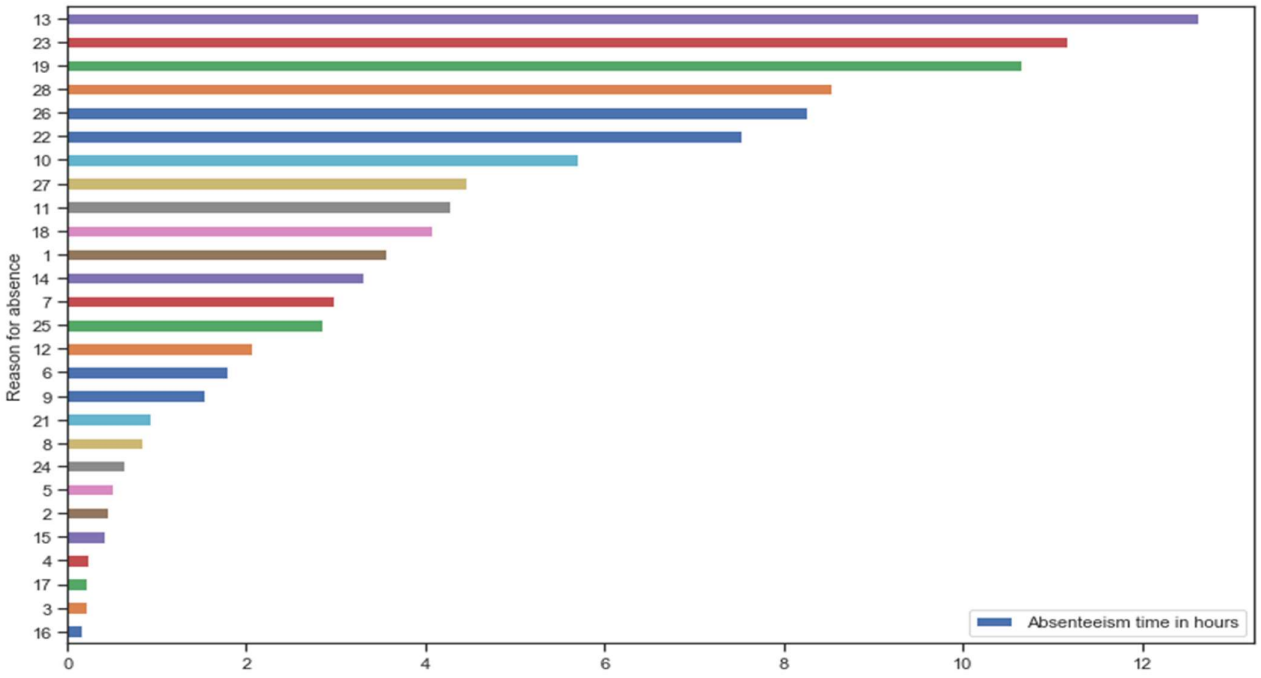
By the following graph we can se that company should look after the spreading awareness for the muskcaleton system and connective tissue.

Company should conduct the working posture awareness

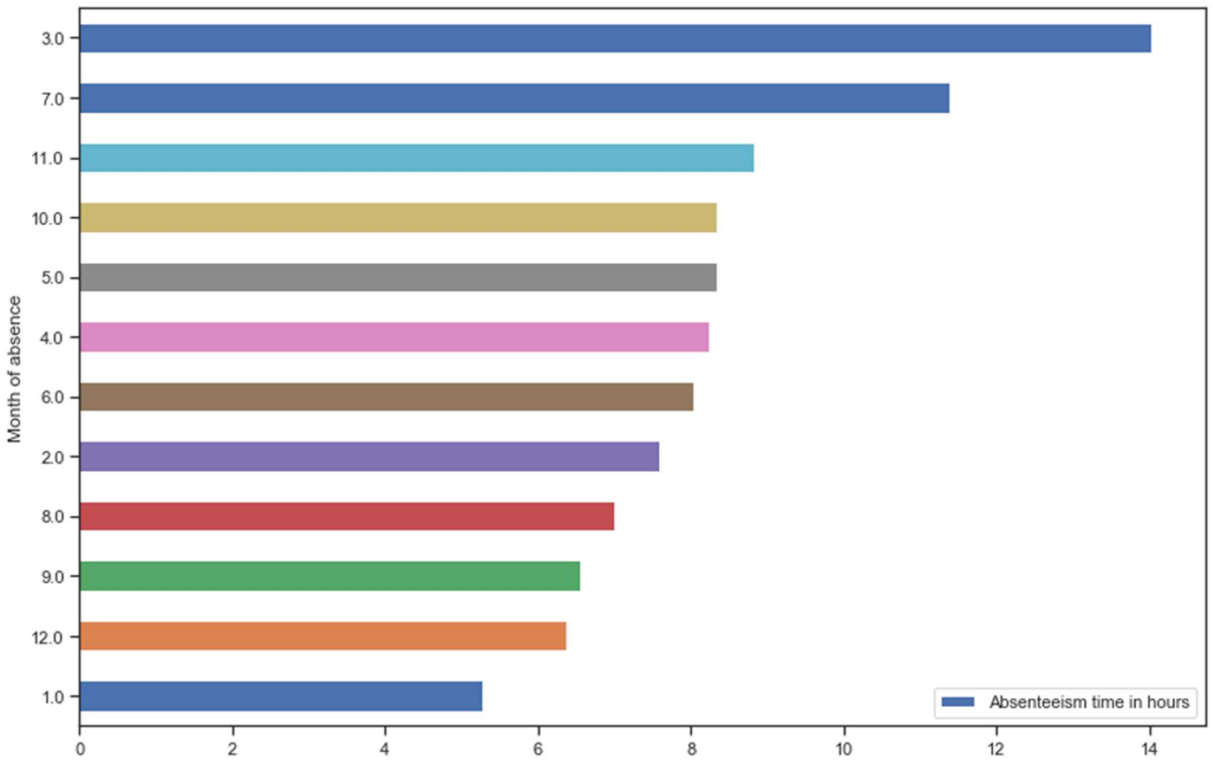
Injury incidence may be reduced by creating proper ergonomic working setup.

Medical consultation may be brought down by optimizing workloads

1. 'Reason for absence' Vs. 'Absenteeism time in hours'



2. 'Month of absence' Vs. 'Absenteeism time in hours'



If the same trend continues then in 2011 according the data

Month 3 : March - 13.02 % of total time

Month 7 : July - 11.37 % of total time

Month 11 : November - 8.81 % of total tim

APPENDIX:

R code for fig

```
ggplot(data=missed_val[1:18,], aes(x=reorder(Variables,-Missed_val_percentage),y=Missed_val_percentage))+
geom_bar(stat="identity",fill="grey")+xlab("Data vriables")+ggtitle("missed value percentage per
variable")+theme_bw()


for(i in 1:length(var_cont))
{
 assign(paste0("gn",i),ggplot(aes_string(y=(var_cont[i]),x="Absenteeism.time.in.hours"),data=subset(dt))+
        stat_boxplot(geom="errorbar",width=0.5)+
        geom_boxplot(outlier.colour="blue",fill="grey",outlier.shape=18,outlier.size=2,notch=FALSE)+
   theme(legend.position="bottom")+
    labs(y=var_cont[i],x="Absenteesim time in hours")+
    ggtitle(paste("Boxplot for Absenteeism hours for",var_cont[i])))


}
#plotting together all the plots generated
gridExtra::grid.arrange(gn1,gn2,gn3,ncol=3)
gridExtra::grid.arrange(gn4,gn5,gn6,ncol=3)
gridExtra::grid.arrange(gn7,gn8,gn9,gn10,ncol=4)
corr diag
corrgram(dt[,var_cont],order=F,upper.panel=panel.pie, text.panel=panel.txt,main="correlation plot")
PCA
an_cmp=prcomp(Train)
#calculation of SD and Var of each component
sd_cmp=an_cmp$sdev
var_cmp=sd_cmp^2
```

```
#visually analyzing the cumulative proportion of varinace with principal components

p_var_cmp=var_cmp/sum(var_cmp)

plot(cumsum(p_var_cmp),xlab="principal components",ylab = "Explained cumulative proportion of
variance",type="b")
```