NAME : KRISHNA KUMAR

PROJECT NAME : EMPLOYEE ABSENTEEISM

DATE OF SUBMISSION : 28 – 08 - 2018

Contents

Chapter 1

Introduction

1.1 Problem Statement

XYZ is a courier company facing employee Absenteeism.
Absenteeism refers to an employee's intentional or habitual absence from work. It's a major problem faced by almost all employers of today. The aim of this project is to help the company to reduce the absenteeism. We are going to predict the amount of loss that the company will face in future if the same trend continues.

1.2 Brief explanation of Absenteeism

Absenteeism of employees from work leads to back logs, piling of work and thus work delay. While employers expect workers to miss a certain number of workdays each year, excessive absences can equate to decreased productivity and can have a major effect on company finances, morale and other factors. Employees are absent from work and thus the work suffers.

Absenteeism is if two types :-

- **Innocent absenteeism -** Is one in which the employee is absent from work due to genuine cause or reason. It may be due to his illness or personal family problem or any other real reason

- **Culpable Absenteeism -** is one in which a person is absent from work without any genuine reason or cause. He may be pretending to be ill or just wanted a holiday and stay at home. The employers have got every right to enquire as to why an employee is absent from work. If an employee is absent because of illness he should be able to produce a doctor's letter as and when demanded.

1.3 Data

Let's take a look at our dataset. Our dataset consists of 740 observations and 21 variables. Of these variables, 20 are independent variables and 1 is our dependent variable, that is target variable . So we take a look at the sample data of our original dataset.

Table 1.1 Employee absenteeism sample data (Columns 1-7) - Predictor Variables

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work
11	26	7	3	1	289	36
36	0	7	3	1	118	13
3	23	7	4	1	179	51
7	7	7	5	1	279	5
11	23	7	5	1	289	36
3	23	7	6	1	179	51
10	22	7	6	1		52

Table 1.2 Employee absenteeism sample data (Columns 8-14) - Predictor Variables

Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Education	Son
13	33	239,554	97	0	1	2
18	50	239,554	97	1	1	1
18	38	239,554	97	0	1	0
14	39	239,554	97	0	1	2
13	33	239,554	97	0	1	2
18	38	239,554	97	0	1	0
3	28	239,554	97	0	1	1

Table 1.3 Employee absenteeism sample data (Columns 15-21)

Column 21 – Dependent variable (target)

Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
1	0	1	90	172	30	4
1	0	0	98	178	31	0
1	0	0	89	170	31	2
1	1	0	68	168	24	4
1	0	1	90	172	30	2
1	0	0	89	170	31	
1	0	4	80	172	27	8

From the above sample data we can have an idea about the dataset we are dealing with. This dataset also gives us a basic insight about employees personal habits such as alcohol consumption and smoking. Also we can know about the height, weight, BMI and other characteristics of the employees from this dataset.

We can also note that this dataset consists of **missing values**. And the characteristics of this dataset is **" Timeseries multivariate "** .

Of the 20 predictor variables (independent variables), there is a combination of both continuous variables and categorical variables . The continuous and categorical variables are listed below.

Continuous variables:

- Individual identification (ID)
- Transportation expense
- Distance from Residence to Work (kilometers)
- Service time

- Age
- Work load Average/day
- Hit target
- Son (number of children)
- Pet (number of pet)
- Weight
- Height
- Body mass index

Categorical variables :

- Reason for absence (28 categories)
- Month of absence (12 months)
- Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
- Seasons (summer (1), autumn (2), winter (3), spring (4))
- Disciplinary failure (yes=1; no=0)
- Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
- Social drinker (yes=1; no=0)
- Social smoker (yes=1; no=0)

Target variable (Dependent variable) :

- **Absenteeism time in hours**

We had a basic look at the data so let us proceed to the next step .

Chapter 2

Methodology

2.1 Pre Processing

Before proceeding to process the dataset to give us a valuable result there are a few steps left. In data analytics terms it is called Exploratory Data Analysis (EDA) . Here we use a combination of both graphical and numerical techniques to :

- Visualize the data
- Detect the missing values
- Detect outliers
- Remove outliers
- Find interesting patterns in the data
- Finding the trends
- Relationship between the variables
- Feature selection / Feature engineering
- Dimensionality reduction

2.1.1 Missing value analysis

Probably the first step we need to perform before diving any further. Not all dataset is 100 % complete. There is a good chance that any dataset may contain missing values. Missing data can be attributed due to various reasons. It could be due to human error , unwillingness to answer and many other reasons. Our dataset too consists of missing values .

There are many ways to deal with the missing values . We can either drop the variable if the number of missing values is too large . Or we can

impute the missing values when the number of missing values is less than 30%.

So in our dataset we are first determining the percentage of missing values . Below is a percentage of missing values in each variable .

Body.mass.index	4.1891892
Absenteeism.time.in.hours	2.9729730
Height	1.8918919
Work.load.Average.day.	1.3513514
Education	1.3513514
Transportation.expense	0.9459459
Hit.target	0.8108108
Disciplinary.failure	0.8108108
Son	0.8108108
Social.smoker	0.5405405

	Reason.for.absence	0.4054054
	Distance.from.Residence.to.Work	0.4054054
	Service.time	0.4054054
	Age	0.4054054
	Social.drinker	0.4054054
	Pet	0.2702703
	Month.of.absence	0.1351351
	Weight	0.1351351
	ID	0.0000000
	Day.of.the.week	0.0000000
	Seasons	0.0000000

So we decide to impute the missing values . So we are using KNN Imputation to impute the missing values. After imputing , there is no missing values in our dataset . We can now proceed further .

2.1.2 Data types

So after performing missing value analysis we have to make sure that the data types are in correct format. That is categorical variables need to be in factor type and continuous variables in numeric type.

2.1.3 Outlier analysis

Next thing in our process is Outlier analysis. In statistics, an outlier is an observation point that is distant from other observations. An outlier can cause serious problems in statistical analyses. Outliers can totally affect the accuracy of the model and the outcome. So before proceeding any further it is necessary to remove the outliers in our dataset. Below are the box plots showing the outliers in predictor variables.

Fig 1.1 R Boxplot - ID , Transportation expense

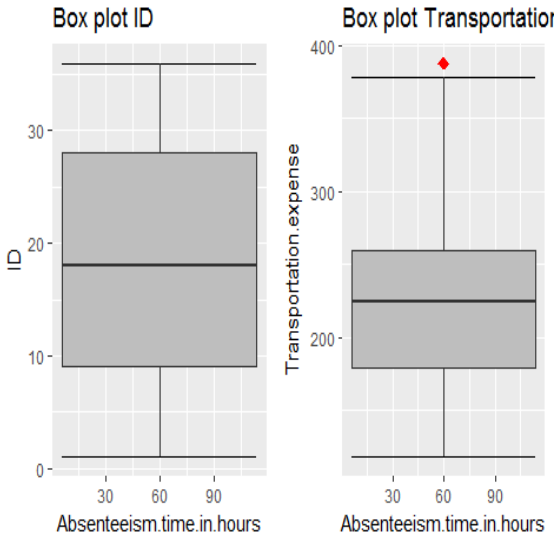


Fig 1.2 R Boxplot – Distance from Residence to work , Service time , Age

Fig 1.3 R Boxplot – Work load Average / day , Hit target , Son

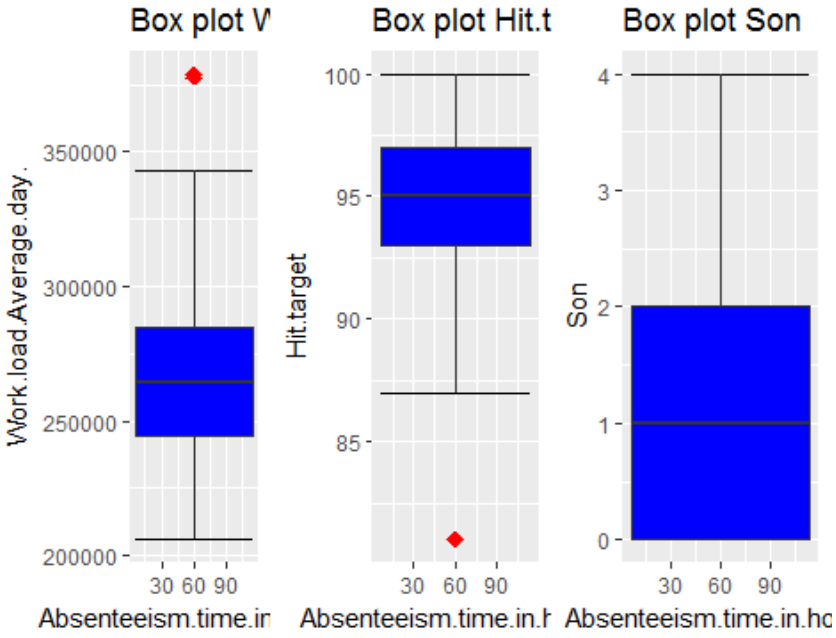


Fig 1.4 R – Boxplot – Pet , Weight , Height

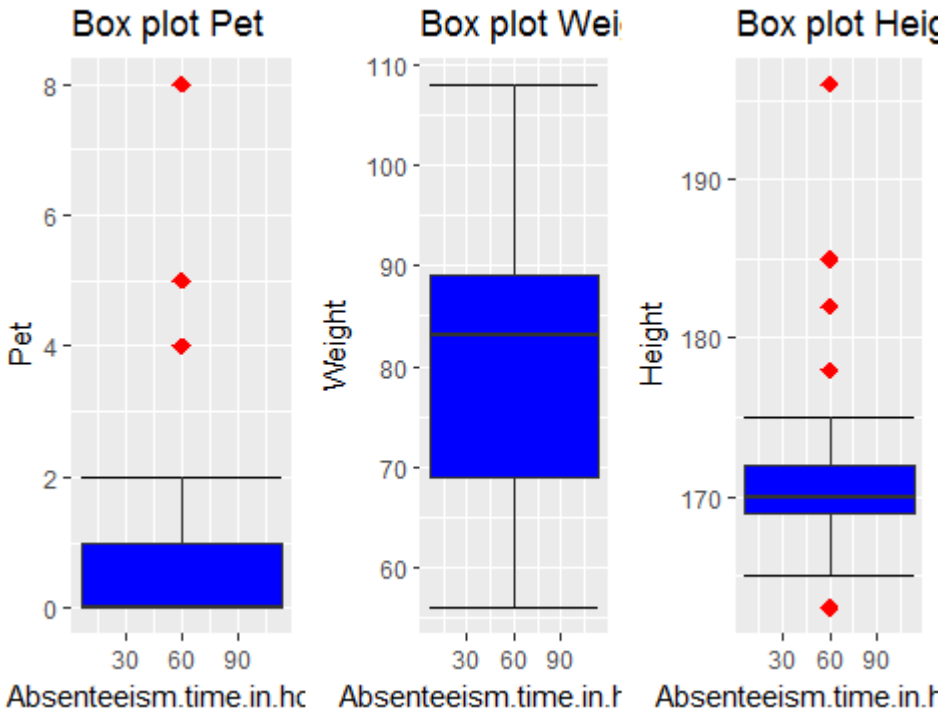
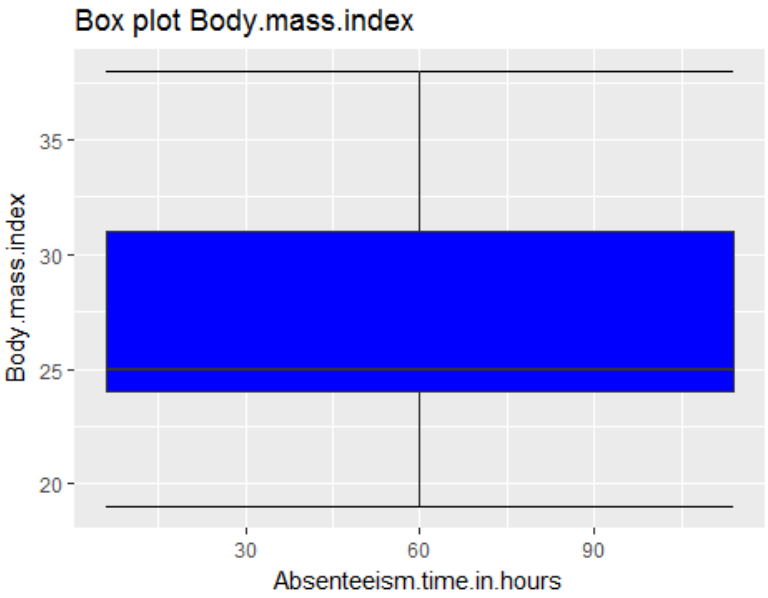


Fig 1.4 R Boxplot – Body Mass Index



In the above figures **" red dots "** indicate the outliers . Outliers are those extreme values present in our data . They are represented by red dots in the above figures. As we can see, that not all variables consist of outliers. But some variables have outliers.

So before proceeding it is necessary to remove the outliers . The code for removing the outliers in R and python can be found in the code file itself. Below are the figure of boxplot after removing the outliers .

Fig 1.5 Boxplot after outlier removal

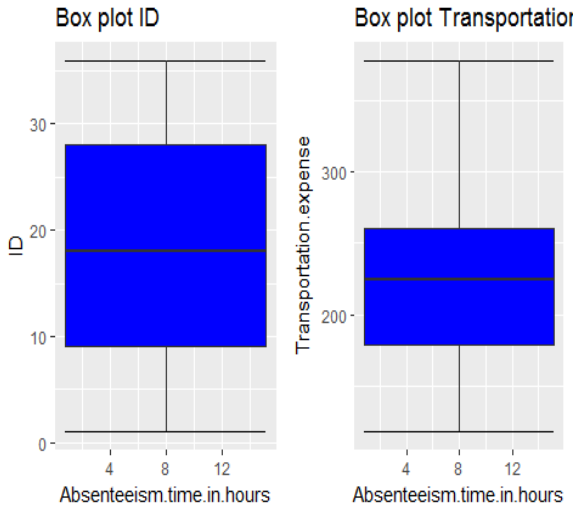


Fig 1.6 Boxplot after outlier removal

Fig 1.7 Boxplot after outlier removal

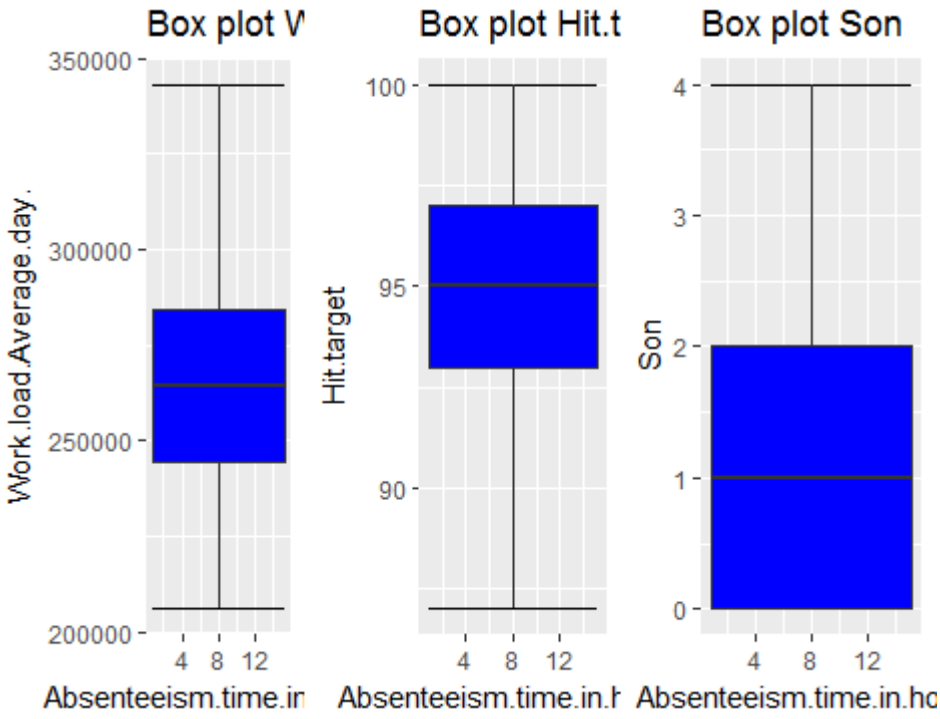


Fig 1.8 Boxplot after outlier removal

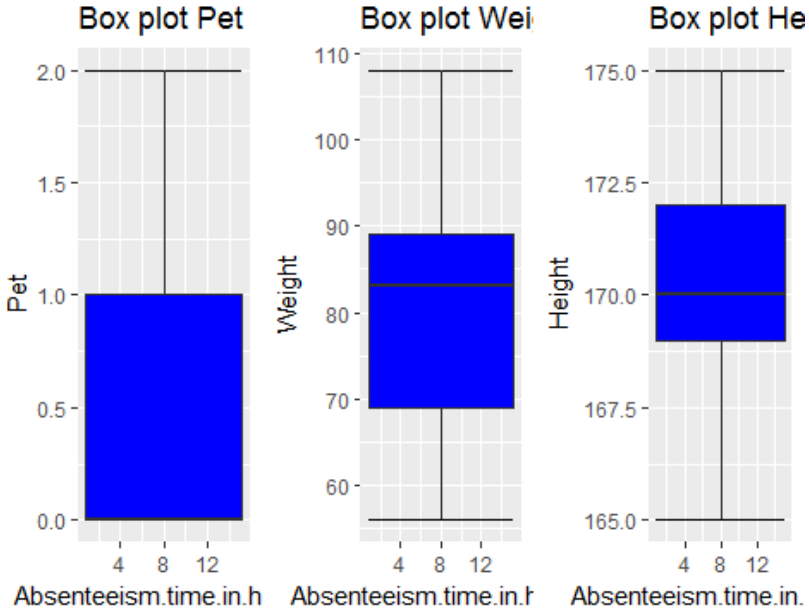
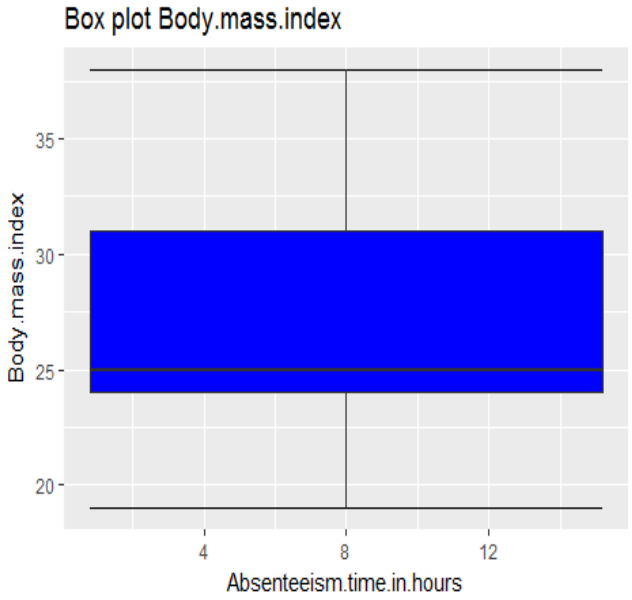


Fig 1.9 Boxplot after outlier removal



As we can see that there no outliers in our data now.

2.1.4 Feature Selection

Feature selection is the next step of our project . It is also called Dimensionality reduction . This is important because there may be many highly correlated independent variables in our dataset . This is called multicollinearity. If not taken care of it may affect the model . Also many variables may not be necessary for the model .

So if we have a really large dataset consisting of thousands or millions of rows and few hundred columns of data then in that situation there is a huge amount of data to be processed . It takes significant computer resources to process the data. Also it is time consuming . So we need to choose only the variables that are contributing to our model .

In the basic level we can perform correlation analysis for continuous variables to detect the highly correlated variables.

Similarly chi square test can be performed on categorical variables to know which variables have high prediction power and we can ignore the rest of the variables.

Below we have performed the random forest method on this dataset to determine the variable importance . The results are shown below.

```
> rf = randomForest(Absenteeism.time.in.hours ~ . , data = df , subset =
sample(1:nrow(df) , 0.8 * nrow(df)) ,
+                   mtry=5, ntree=500 , importance = TRUE)
>
>   importance(rf, type = 1)
```

	%IncMSE
ID	15.641575
Reason.for.absence	51.872120
Month.of.absence	8.738899
Day.of.the.week	4.294115
Seasons	5.468648
Transportation.expense	10.078785
Distance.from.Residence.to.Work	9.518719
Service.time	8.470779
Age	11.705812
Work.load.Average.day.	5.267288
Hit.target	2.338659
Disciplinary.failure	17.470034
Education	4.486265
Son	7.057093
Social.drinker	4.308822
Social.smoker	2.763356
Pet	6.055432
Weight	8.497052
Height	11.063537
Body.mass.index	10.455192

So according to the above data random forest seems to suggest that **Reason for absence** is the most important variable in our dataset . This variable seems to have the highest prediction power of any variable.

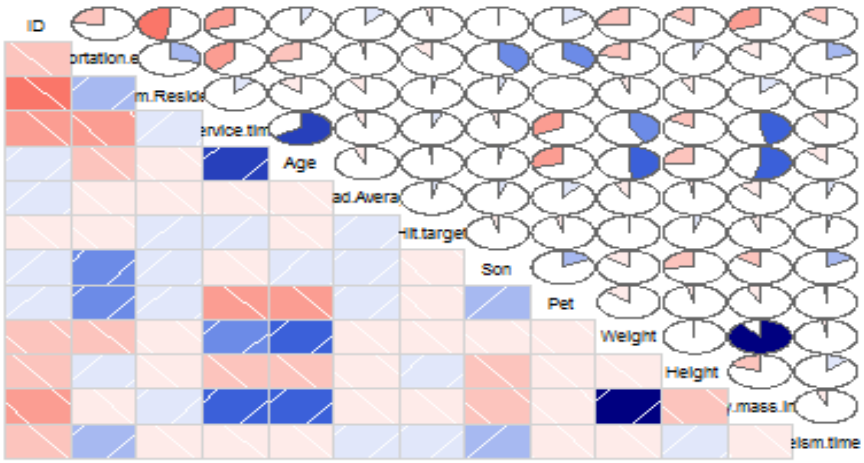
Next we check for correlation between the variables. Note that this can only be performed on the continuous variables.

```
symnum(cor(Z_1))
                                I T D S. Ag W. H. Sn P Wg Hg B A.
ID                              1
Transportation.expense            1
Distance.from.Residence.to.Work .   1
Service.time                    . .   1
Age                                 ,  1
Work.load.Average.day.                    1
Hit.target                                   1
Son                               .             1
Pet                               .   .  .         1
Weight                                .  .           1
Height                                                  1
Body.mass.index                 .     .  .           *     1
Absenteeism.time.in.hours                                     1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

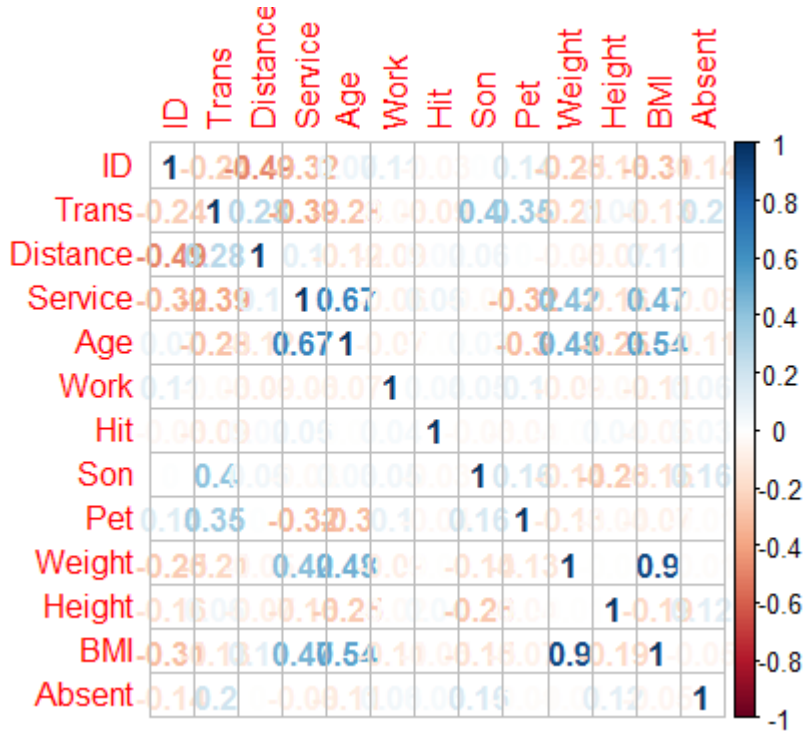
From the above data we can see that Body Mass Index and Weight are highly correlated. (0.9) . Below is a corrgram **(**Correlogram) showing the correlation visually .

```
> corrgram(Z_1 , order = F , upper.panel = panel.pie , text.panel = panel.txt
,
+          main = "Correlation plot")
```

Fig 2.0 Correlation plot




```
> corrplot(X2 ,   method="number" )
```



2.2 Modeling

2.2.1 Model Selection

Model selection depends on many factors . No model is 100 % perfect or accurate . Something which works for a model may not work for others. Some models may be more resistant to outliers and produce better results .Some models may have high accuracy but they may suffer from the presence of outliers. So we have to select our model accordingly .

Also model selection depends on the type of dependent variable or the target variable. If the target variable is categorical we can only perform classification . If the target variable is continuous we can perform regression. Below are some of the models :

Supervised machine learning:

- Decision Tree - Classification and Regression
- Random Forest - Classification and Regression
- Linear Regression – Regression for continuous target variable
- Logistic Regression - Regression for categorical target variable
- KNN - Classification and Regression
- Naïve Bayes - Classification

Unsupervised machine learning :

- Cluster analysis

Here we have decided to use Decision Trees , Random Forest and Linear Regression . We are going to perform Regression using all the models since our target variable is continuous.

2.2.2 Decision Tree (Regression)

We performed decision tree regression on our dataset . And the results are below :

```
> fit = rpart(Absenteeism.time.in.hours ~ . , data = train , method =
"anova")
> predictions = predict(fit , test[,-21])
```

```
> fit
n= 592

node), split, n, deviance, yval
      * denotes terminal node

 1) root 592 6673.82200 4.4232770
   2) Reason.for.absence=0,2,4,16,23,25,27,28 324 1586.35200 2.5997350
     4) Reason.for.absence=0,4 36  115.18670 0.8166949 *
     5) Reason.for.absence=2,16,23,25,27,28 288 1342.40700 2.8226150
      10) Transportation.expense< 295.5 275 1003.91600 2.6834760 *
      11) Transportation.expense>=295.5 13  220.54650 5.7659370 *
   3) Reason.for.absence=1,3,5,6,7,8,9,10,11,12,13,14,15,17,18,19,21,22,24,26
268 2707.54300 6.6278580
     6) ID>=25.5 67  682.07020 5.0522800
      12) Month.of.absence=1,4,5,6,8,10,11,12 47  245.95150 3.9989020 *
      13) Month.of.absence=2,3,7 20  261.41110 7.5277200
        26) Reason.for.absence=7,9,14,26 8   56.93734 4.8193010 *
        27) Reason.for.absence=1,5,11,15,18,22 12  106.66670 9.3333330 *
     7) ID< 25.5 201 1803.70800 7.1530510
      14) Reason.for.absence=7,9,11 22  172.09250 5.0229100 *
      15) Reason.for.absence=1,3,5,6,8,10,12,13,14,15,17,18,19,21,22,24,26
179 1519.52100 7.4148560
        30) Body.mass.index< 22 22  277.60860 5.7118190
          60) Reason.for.absence=13,21 8   35.50000 2.7500000 *
          61) Reason.for.absence=1,12,18,19,22 14  131.82730 7.4042870 *
        31) Body.mass.index>=22 157 1169.16400 7.6534980 *
```

2.2.3 Random Forest (Regression)

Next we perform Random forest Regression on our dataset . It is shown below :

```
> RF_mod = randomForest(Absenteeism.time.in.hours ~ . , data = df , subset =
sample(1:nrow(df) , 0.8 * nrow(df)))
```

```
Call:
 randomForest(formula = Absenteeism.time.in.hours ~ ., data = df,      subset
= sample(1:nrow(df), 0.8 * nrow(df)))
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 6

          Mean of squared residuals: 8.012457
                    % Var explained: 32.24
```

```
>   rf = randomForest(Absenteeism.time.in.hours ~ . , data = df , subset =
sample(1:nrow(df) , 0.8 * nrow(df)) ,
+                   mtry=5, ntree=500 , importance = TRUE)
> pred = predict(rf,test[,-21])
```

2.2.4 Linear Regression

Our third model is Linear Regression.

```
> summary(lm_model)

Call:
lm(formula = Absenteeism.time.in.hours ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8396 -1.4481 -0.0311  0.8991 12.5752

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -1.182e+00  1.593e+01  -0.074  0.94090
ID                            -4.358e-02  1.973e-02  -2.209  0.02761 *
Reason.for.absence1            4.489e+00  2.884e+00   1.557  0.12010
Reason.for.absence2           -3.484e-01  3.959e+00  -0.088  0.92991
Reason.for.absence3            5.152e+00  3.396e+00   1.517  0.12987
Reason.for.absence4            2.471e+00  3.422e+00   0.722  0.47050
Reason.for.absence5            6.028e+00  3.438e+00   1.753  0.08013 .
Reason.for.absence6            2.981e+00  3.047e+00   0.978  0.32831
Reason.for.absence7            1.476e+00  2.886e+00   0.512  0.60918
Reason.for.absence8            1.417e+00  3.247e+00   0.436  0.66267
Reason.for.absence9            9.616e+00  3.388e+00   2.838  0.00471 **
Reason.for.absence10           3.755e+00  2.862e+00   1.312  0.19013
Reason.for.absence11           2.623e+00  2.851e+00   0.920  0.35803
Reason.for.absence12           3.928e+00  2.967e+00   1.324  0.18607
Reason.for.absence13           3.522e+00  2.825e+00   1.246  0.21317
Reason.for.absence14           2.317e+00  2.898e+00   0.799  0.42442
Reason.for.absence15           5.091e+00  3.371e+00   1.510  0.13159
Reason.for.absence16          -1.073e+00  3.422e+00  -0.313  0.75408
Reason.for.absence17           3.570e+00  3.951e+00   0.903  0.36669
Reason.for.absence18           3.602e+00  2.853e+00   1.263  0.20726
Reason.for.absence19           3.650e+00  2.825e+00   1.292  0.19695
Reason.for.absence21           2.993e+00  3.019e+00   0.992  0.32188
Reason.for.absence22           3.902e+00  2.829e+00   1.379  0.16841
```

```
Reason.for.absence23              1.823e-01  2.793e+00   0.065  0.94800
Reason.for.absence24              4.607e+00  3.384e+00   1.361  0.17397
Reason.for.absence25              6.784e-01  2.846e+00   0.238  0.81168
Reason.for.absence26              4.328e+00  2.842e+00   1.523  0.12841
Reason.for.absence27             -7.946e-01  2.834e+00  -0.280  0.77930
Reason.for.absence28             -1.684e-01  2.800e+00  -0.060  0.95205
Month.of.absence1                 1.537e+00  3.518e+00   0.437  0.66249
Month.of.absence2                 2.527e+00  3.501e+00   0.722  0.47072
Month.of.absence3                 3.124e+00  3.495e+00   0.894  0.37175
Month.of.absence4                 7.432e-01  3.526e+00   0.211  0.83314
Month.of.absence5                 2.487e-01  3.532e+00   0.070  0.94390
Month.of.absence6                 8.852e-01  3.496e+00   0.253  0.80024
Month.of.absence7                 2.833e+00  3.458e+00   0.819  0.41311
Month.of.absence8                 2.585e+00  3.469e+00   0.745  0.45654
Month.of.absence9                 2.553e+00  3.444e+00   0.741  0.45878
Month.of.absence10                2.699e+00  3.588e+00   0.752  0.45223
Month.of.absence11                1.892e+00  3.579e+00   0.529  0.59733
Month.of.absence12                2.138e+00  3.547e+00   0.603  0.54689
Day.of.the.week3                 -7.672e-02  3.485e-01  -0.220  0.82585
Day.of.the.week4                  1.458e-01  3.511e-01   0.415  0.67817
Day.of.the.week5                  4.025e-01  3.789e-01   1.062  0.28865
Day.of.the.week6                 -2.381e-01  3.794e-01  -0.628  0.53051
Seasons2                          4.175e-01  8.894e-01   0.469  0.63898
Seasons3                          2.202e+00  8.469e-01   2.600  0.00959 **
Seasons4                         -5.117e-02  7.480e-01  -0.068  0.94548
Transportation.expense            2.887e-03  2.965e-03   0.974  0.33072
Distance.from.Residence.to.Work -1.427e-02  1.162e-02  -1.228  0.21991
Service.time                     -2.470e-02  5.252e-02  -0.470  0.63834
Age                              -3.986e-02  3.651e-02  -1.092  0.27539
Work.load.Average.day.            2.521e-06  4.567e-06   0.552  0.58128
Hit.target                       -4.405e-02  5.546e-02  -0.794  0.42734
Disciplinary.failure1            -2.654e+00  2.815e+00  -0.943  0.34618
Education2                       -6.863e-01  6.750e-01  -1.017  0.30975
Education3                       -7.689e-01  5.615e-01  -1.369  0.17147
Education4                        1.453e+00  1.505e+00   0.965  0.33475
Son                               2.230e-01  1.444e-01   1.544  0.12321
Social.drinker1                  -5.054e-01  4.841e-01  -1.044  0.29705
Social.smoker1                    2.178e-01  5.659e-01   0.385  0.70048
Pet                              -6.199e-01  2.356e-01  -2.631  0.00875 **
Weight                           -8.936e-03  3.279e-02  -0.273  0.78532
Height                            5.160e-02  8.644e-02   0.597  0.55080
Body.mass.index                  -7.600e-03  1.074e-01  -0.071  0.94362
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.668 on 527 degrees of freedom
Multiple R-squared:  0.4702,   Adjusted R-squared:  0.4058
F-statistic: 7.307 on 64 and 527 DF,  p-value: < 2.2e-16
```

Chapter 3

Conclusion

3.1 Model Evaluation

After training the model and predicting the outcomes it is now time to evaluate the performance of the model. This stage is key since we have to choose a model which have the most accuracy . Also other things in consideration are resource usage . Resource usage varies from model to model. For example KNN is a very lazy algorithm . We say this because it takes a long time to complete the process. So we have to choose a model which is also computationally less expensive.

3.1.1 Error metrics

When it comes to performance of a model we judge a model by its accuracy . Accuracy of a model can be obtained from Error Metrics. Following are the error metrics we are testing on our model to evaluate the performance.

- RMSE – Root Mean Square Error
- MSE – Mean Square Error
- MAPE – Mean Absolute Percentage Error
- MAE – Mean Absolute Error

3.1.2 Decision Tree Regression

Error metrics of this model is given as below :

```
> regr.eval(test[,21] , predictions_DT , stats = c("mse" , "rmse" , "mae" , "mape"))
```

```
     mse     rmse      mae     mape
5.391046 2.321863 1.565675      Inf
```

3.1.3 Random Forest Regression

Error metrics of this model is given as below :

```
> regr.eval(test[,21] , pred , stats = c("mse" , "rmse" , "mae" , "mape"))
     mse     rmse      mae     mape
4.839468 2.199879 1.644935      Inf
```

3.1.4 Linear Regression

Error metrics of this model is given as below :

```
> regr.eval(test[,21] , predictions_LR , stats = c("mse" , "rmse" , "mae" ,
"mape"))
     mse     rmse      mae     mape
6.915637 2.629760 1.849507      Inf
```

3.2 Model Selection

From the above results Decision Tree and Random Forest produce better results than Linear Regression but it is not enough. One thing to note here is because train and test data are randomly selected from the dataset, these errors vary each time when we run the model.

Also we saw previously saw that Body Mass Index is highly correlated with Weight . Removing the variable from the dataset did not yield any significant improvement . So all models are trained and evaluated with all predictor variables.

Chapter 4

Answering the Questions

So now we have come to the most important part of this project . That is answering the questions . So far we have performed Exploratory Data Analysis , Feature selection , model preparation , model evaluation . The company which gave us the dataset wants the answers for some questions . It is the time now to answer the questions .

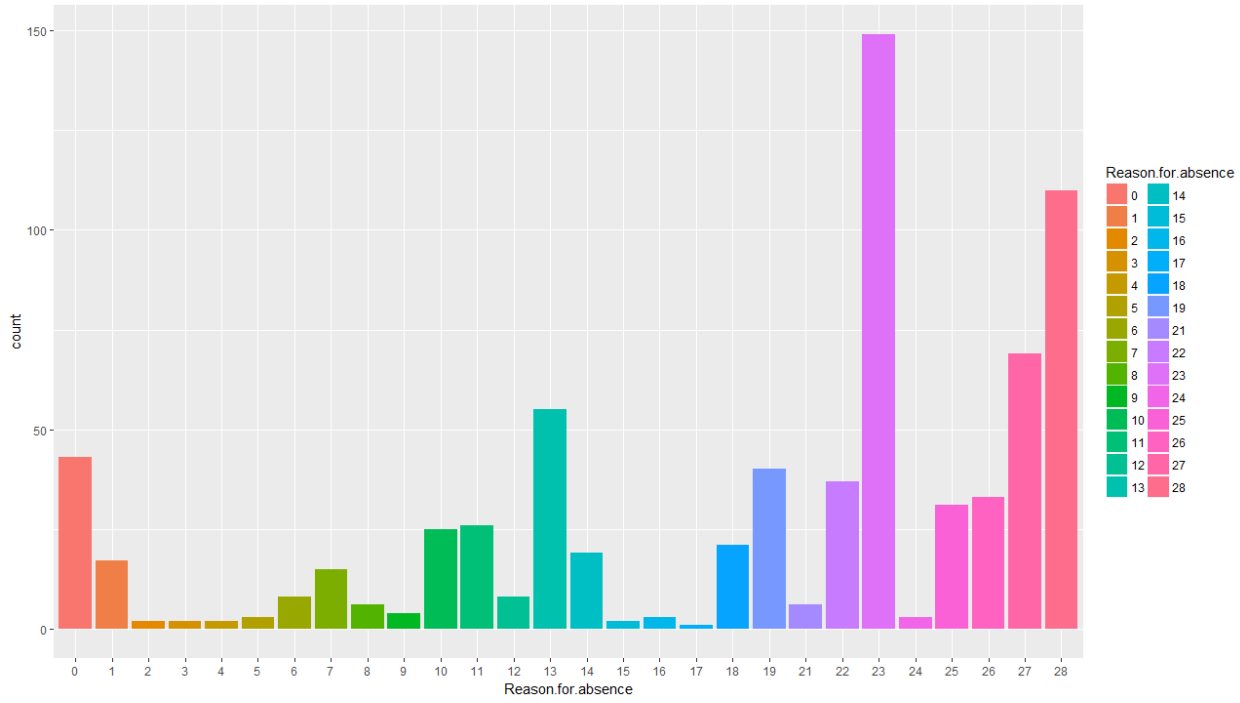
1. What changes company should bring to reduce the number of absenteeism?

To answer this question we have to look at which is the most important predictor variable . Because from this we can understand that which variable contributes a lot to the outcome . So as we discussed before in the Feature selection that the variable Reason for absence is the most significant variable . So by closely examining this variable we came to know the following .

The below figure shows the most occurring category of Reason for absence , that is which reason occurs the most in the variable . Looking at the below figure we can understand :

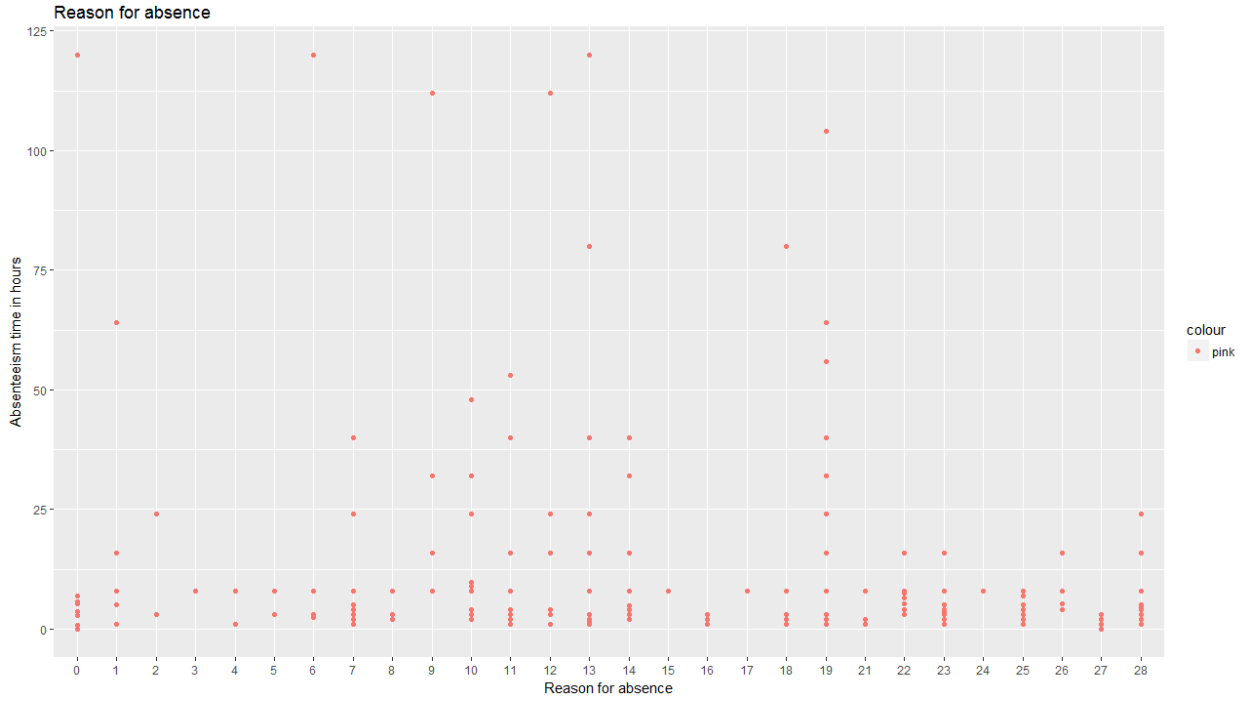
- Category – 23 , 28 , 27 and 13 occur the most in the dataset indicating that the employee absenteeism is related to these categories . Let's look at the categories :
- Category 23 - Medical consultation
- Category 28 - Dental consultation
- Category 27 – Physiotherapy
- Category 13 - Diseases of the musculoskeletal system and connective tissue

```
> ggplot(X, aes(Reason.for.absence, fill = Reason.for.absence ) ) +
+   geom_bar()
```




```
> qplot(x = X$Reason.for.absence, y = X$Absenteeism.time.in.hours,
data= X,

+       color= "pink", xlab= "Reason for absence",
+       ylab= "Absenteeism time in hours", main= "Reason for absence")
```



So it is clear that these four categories occur the most in the Reason for absence . These four categories together contribute to around 50 % of the Absenteeism . So this shows that the employees are facing serious health issues .

Barring dental consultation , other categories like medical consultation , physiotheraphy and Diseases of the musculoskeletal system and connective tissue shows that health of the employees are not good . Especially physiotherapy and Diseases of the musculoskeletal system and connective tissue show they are very much physically affected .

Another interesting fact is that employees are in the age group of 30 – 40 years the most . This combined with the above data shows that employees are both physically not fit and at the same time they are aged.

So the company can provide employee health care benefits , Medicare facilities and take care of the well being of the employees . Also there may be a reason that these employees are over working and may be under the

stress a lot . So , if this is true then the company must reduce the workload of the employees. The company may provide paid vacation to reduce the stress of the employees.

Or else the company may recruit younger workforce who don't suffer from illness.

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

To answer this question , first there is no financial information provided here . Losses in the sense can be attributed to back logs , additional pressure on other employees which further incresase their stress . And since this is a courier company which consists of roles of the employees like collection, transportation and delivery everything suffers . If collection is affected then the rest of the workflow is also affected . Employee morale is also affected .

And it even leads to loss of customers which leads to reduced profits . So in many ways huge loss can be expected if the same trend of Absenteeism continues .

Notes

1. The full code for the project can be found in R and python files . (Project 1.R and Project1.ipynb)
2. R code and python code are not provided here since they can be found in the code files itself.
3. In codes the necessary information or notes (if required) is provided in comments (#) .
4. The code file submitted is the final code .
5. Full care has been taken to ensure that the code runs properly . So if any error is encountered during code execution kindly re - run the code .

R Code

Please find the code file attached with this project report (Project 1.R)

Python Code

Please find the code file attached with this project report (Project 1.ipynb)