EMPLOYEE ABSENTEEISM

Data Science Project

Singh, Ankit Kumar

Contents

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 Data

Our task is to build a regression model which helps us predict the Absenteeism time depending on various employee-related personal and professional factors. As you can see in the table below we have the following 20 variables, using which we have to correctly predict the absenteeism time.

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

IV. Endocrine, nutritional and metabolic diseases

V. Mental and behavioral 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,

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)

Chapter 2- Methodology

2.1 Pre-Processing

2.1.1. Exploratory Data Analysis

Before we start modelling, we need to look at the data, which basically means exploring and cleaning the data so as to make the data ready for consumption by the model. This is often called as **Exploratory Data Analysis**. To start this process, we look at the number of unique values each variable takes and divide the data into categorical and continuous variables. Next, we try to look at all the probability distributions of all continuous variables.
Most analysis (for example regression) requires the data to be normally distributed. The probability density function (shown below in Figure 2.1) helps us to visualize the distribution of data as well as if the variable is normally distributed or not. The blue lines indicate Kernel Density Estimations (KDE) of the variable. We can also analyze the same using **qqnorm**. Next, as a part of cleaning the data, we look at the range of each columns and if each variable is taking acceptable values

List of columns and their number of unique values -

```
ID                                36
Reason for absence                28
Month of absence                  13
Day of the week                    5
Seasons                            4
Disciplinary failure               2
Education                          4
Son                                5
Social drinker                     2
Social smoker                      2
Pet                                6

Transportation expense            24
Distance from Residence to Work   25
Service time                      18
Age                               22
Work load Average/day             38
Hit target                        13
Weight                            26
Height                            14
Body mass index                   17
Absenteeism time in hours         19
```

We have categorized them into categorical variable (top 11) and continuous (bottom 10)

Few observations:

1. Month should range between 1 to 12 . But for last three rows of the dataset, it's 0.
2. Additionally, for last three rows Disciplinary Failure and Reason for absence are not in trend. Except these three, rest of the rows follow the following trend
   Disciplinary failure =1, if Reason for Absence=0
   =0, if Reason for Absence! =0
3. There are lots of missing values which are unique to each employee. For ex- An employee will have unique height, Age, Weight, etc. and thus any of the personal missing values can be easily imputed.
4. The data is quite skewed

Due to 1. and 2. , we decided to drop the last three rows.

Next in the process are performing missing value analysis and outlier analysis described below.

2.1.2 Missing value analysis:

Here, missing value occurs when data is missing (or no data stored for any variable). We need to decide between deleting the row vs imputation, depending on how it could impact the model. Generally, if a column has more than 30% of missing values- either we ignore the whole column or we ignore those missing values. The below graph (Figure 1.) depicts the missing values for all the variables. We can easily observe that body mass index has maximum missing values which is less than 5% and thus we decided to impute the value. Since employee personal details are unique as mentioned earlier in this chapter, we have used **simputation** library to impute them using their ID. For remaining variable I used KNN. We generally use mean, median or KNN to impute the values. After analyzing, I found that KNN is more suitable to impute missing values for the given data set.

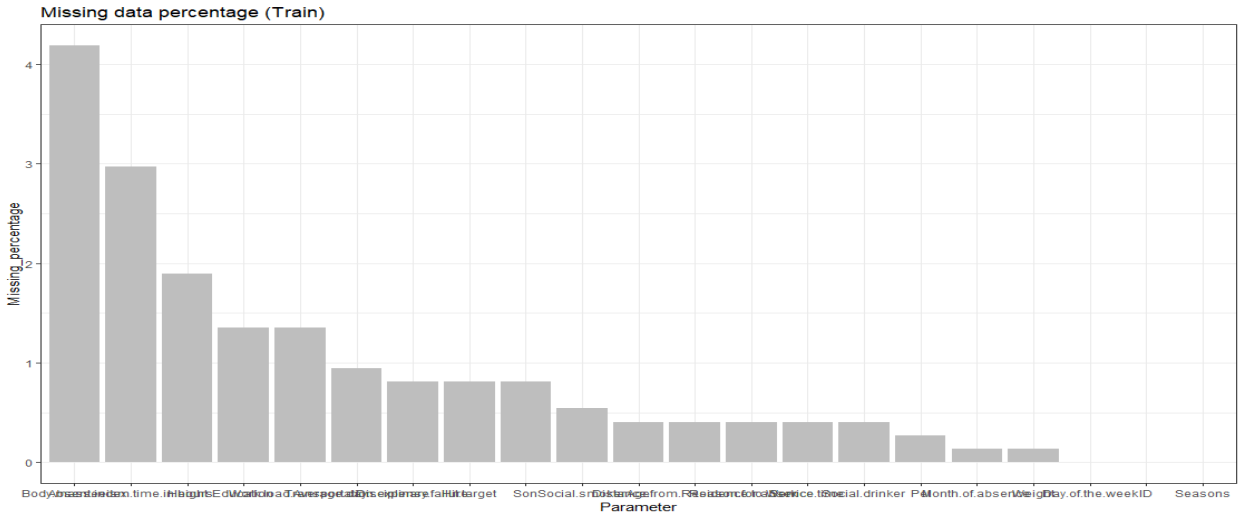
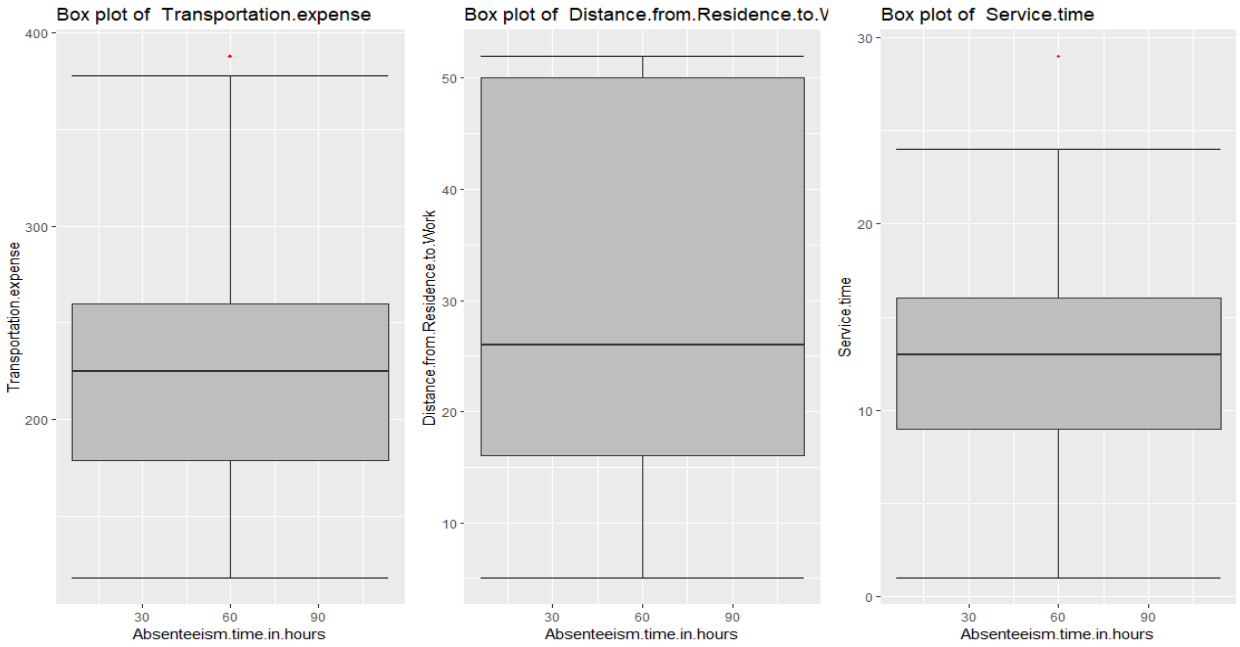


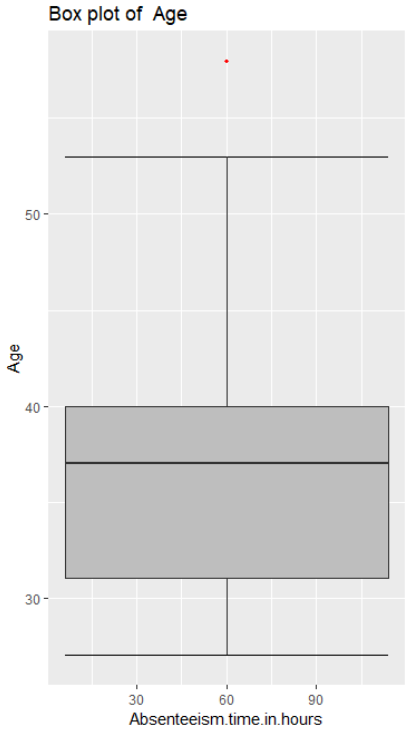
Figure 2.1. Missing value data percentage as a % of total rows

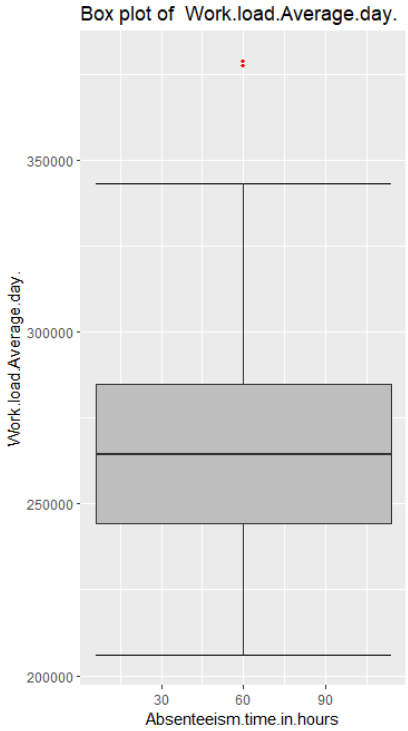
2.1.3 Outlier Analysis

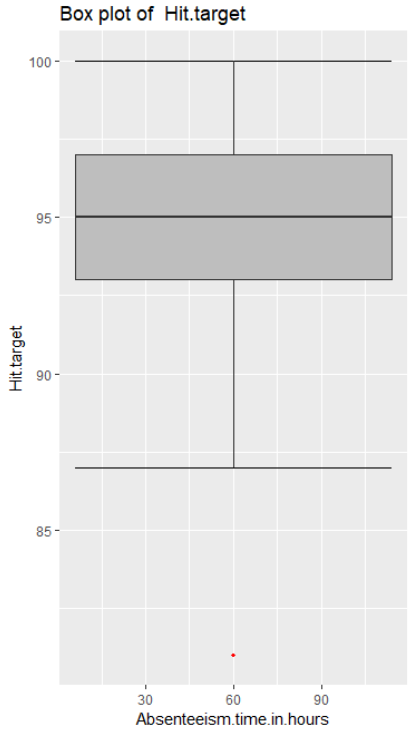
The next step in pre-processing is to check the presence of outlier. Outliers result in skewed distribution. In this case, we can clearly observe the skewness in data from probability distribution function. Outliers could be due to incorrect recording/observation or large variability in data. If we don't remove the outliers, it may negatively impact our model especially linear models

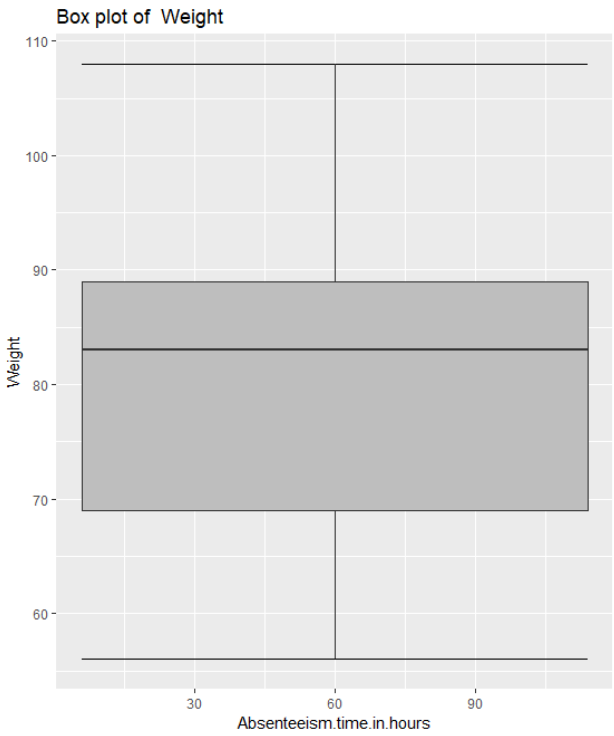
In this project, we have used a classic approach of removing the outliers- Tukey's method. We use boxplot method to visualize the outliers. In fig 2.2, we have plotted the boxplots of all the continuous variables

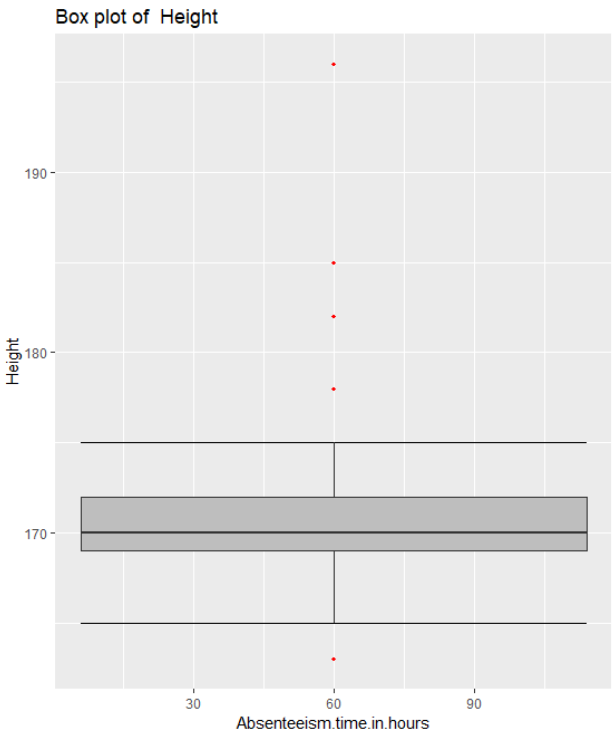



Box plot of Age
Age
50
40
30
30
60
90
Absenteeism.time.in.hours


Box plot of Work.load.Average.day.
Work.load.Average.day.
350000
300000
250000
200000
30
60
90
Absenteeism.time.in.hours


Box plot of Hit.target
Hit.target
100
95
90
85
30
60
90
Absenteeism.time.in.hours


Box plot of Weight
Weight
110
100
90
80
70
60
30
60
90
Absenteeism.time.in.hours


Box plot of Height
Height
190
180
170
30
60
90
Absenteeism.time.in.hours

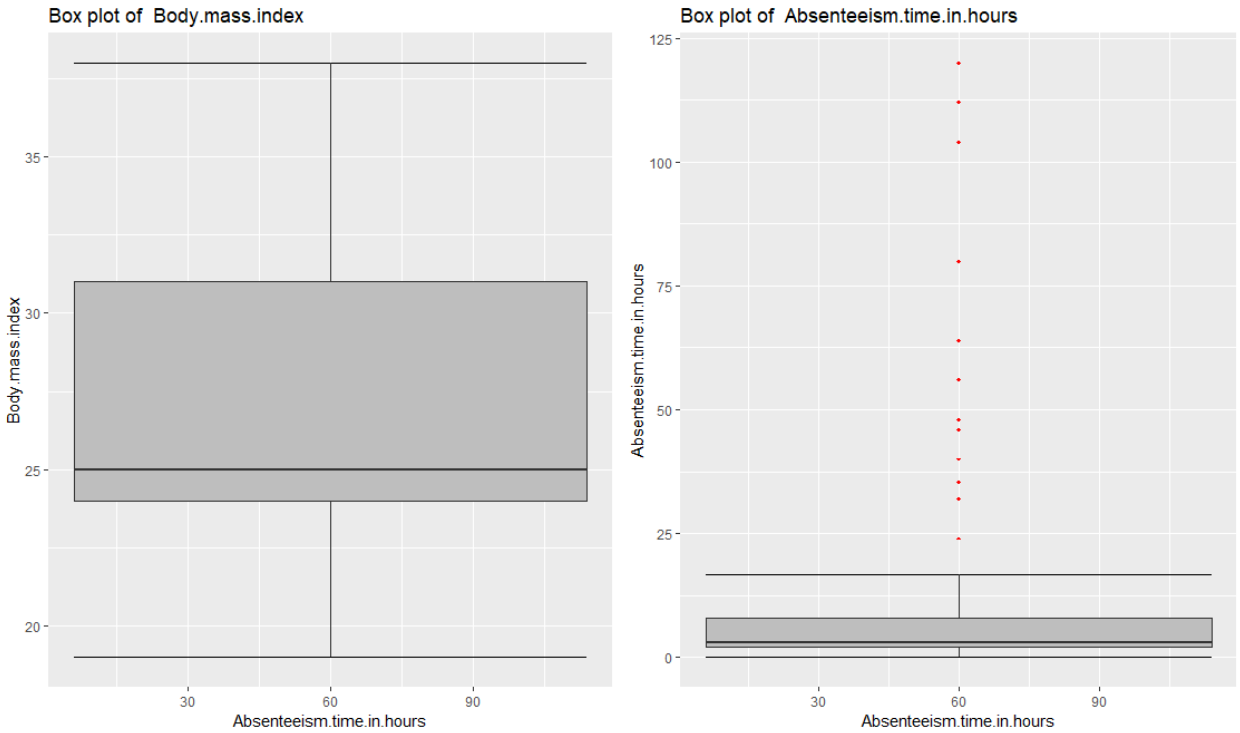


Fig 2.2 Boxplot for all continuous variable

From the boxplots, we can clearly observe that Height, Absenteeism and Pet has maximum number of outliers whereas Hit. Target, Distance from residence to Work, Weight, and Body mass index have minimum number of outliers. The above figure shows the boxplot of all continuous variables

2.1.4 Feature Selection

Before we perform modelling, it's important to assess the significance of each independent variable in predicting the target variable. There is a possibility that many variables in our analysis are not important to the problem of prediction. Feature selection is basically selecting relevant features for the model construction. We need to perform feature selection as some variables may carry same information or may be information not relevant in predicting the target variable. This in turn increases the overheads. Thus, feature selection helps us to deal with relevant variables and avoid the problem of multi collinearity. In this case we have used **Correlation Analysis** for continuous variable and **ANOVA** (Analysis of variance) for categorical variable.

a. Correlation

From correlation analysis, we observed that weight and BMI are highly correlated (shown in below) and thus would be carrying the same information. Thus, I have excluded the variable "Body Mass Index" as Weight is the basic variable.

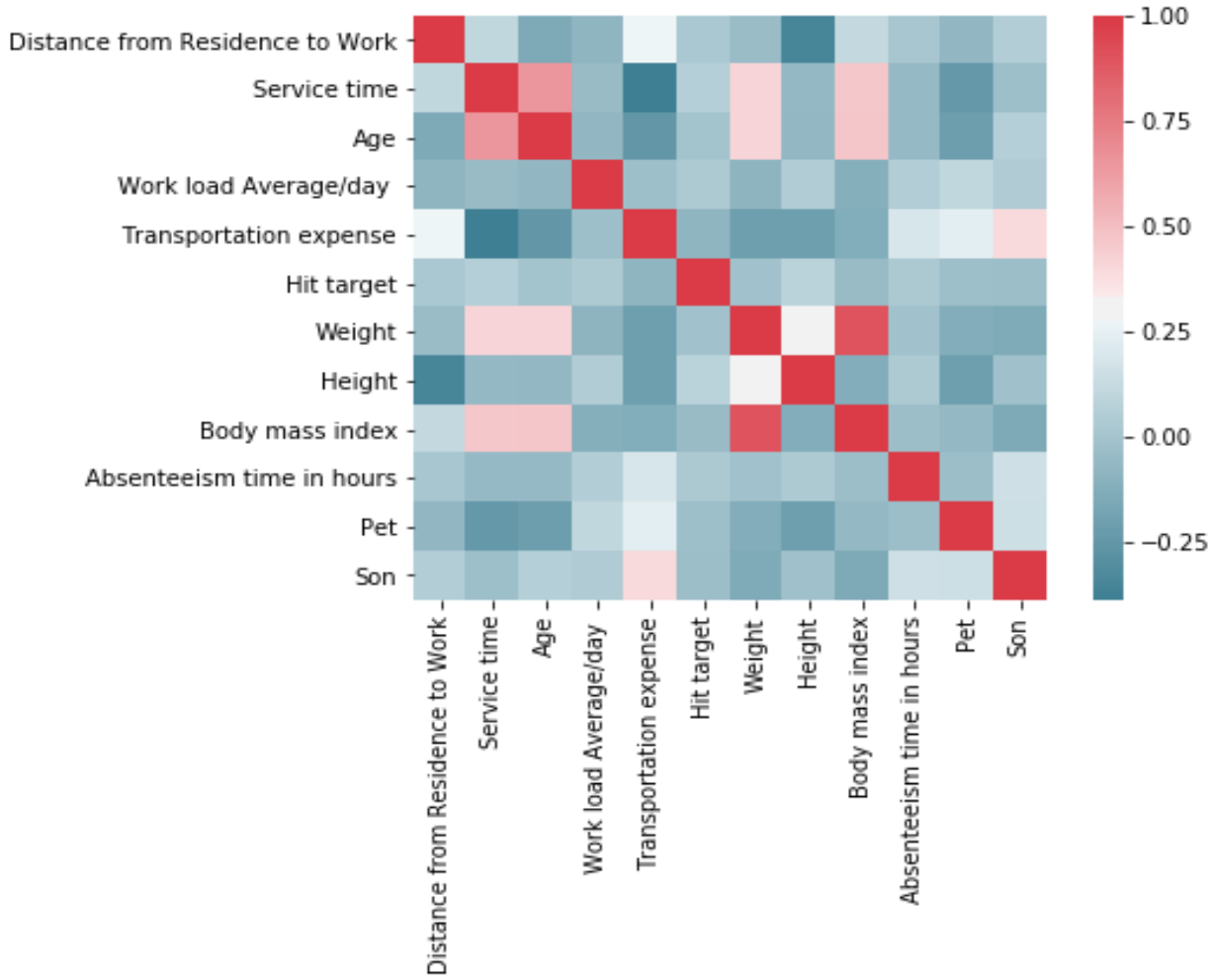


Figure 2.3 Correlation analysis of all continuous variables

b. ANOVA:

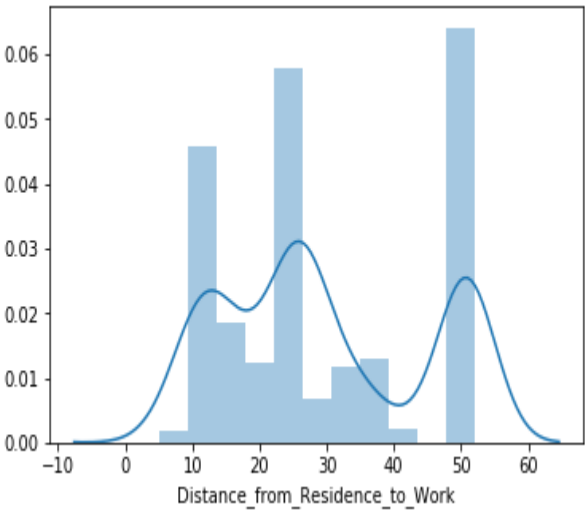
The below table represents summary of ANOVA

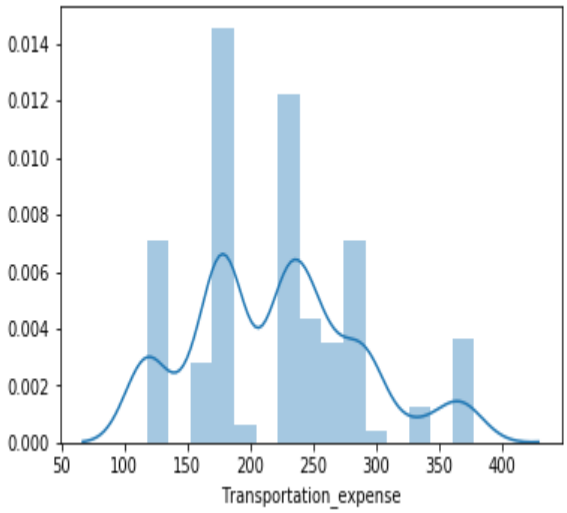
```
                    Df Sum Sq Mean Sq F value   Pr(>F)
Day.of.the.week      1     35    34.7   4.051 0.044516 *
Education            1      3     3.0   0.346 0.556808
Social.smoker        1     10    10.2   1.195 0.274781
Social.drinker       1    120   119.6  13.979 0.000199 ***
Reason.for.absence   1    182   182.3  21.310 4.62e-06 ***
Seasons              1     38    38.3   4.474 0.034745 *
Month.of.absence     1      0     0.1   0.017 0.897691
Disciplinary.failure 1   1901  1901.1 222.247  < 2e-16 ***
Son                  1    149   149.4  17.464 3.28e-05 ***
Pet                  1      9     9.0   1.050 0.305823
Residuals          726   6210     8.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

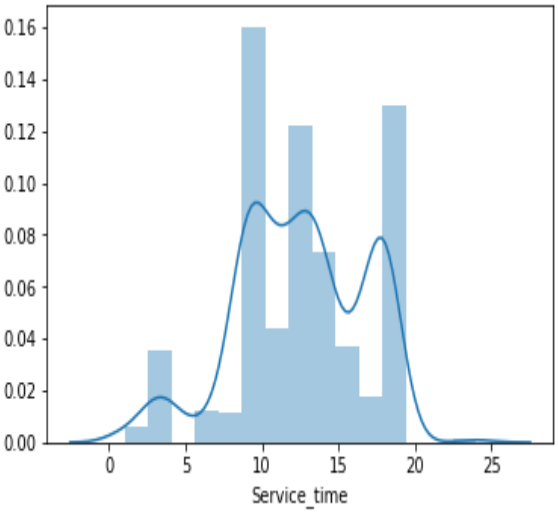
From the above table we observe that Education, Social Smoker, Month of absence and Pet are not significant variables (p values >0.05 and can be ignored.

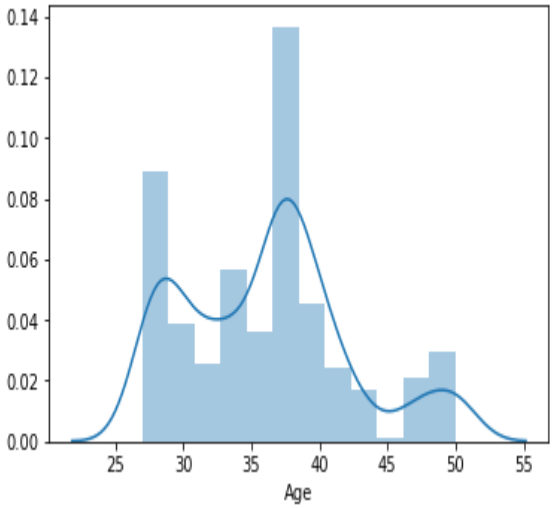
2.1.5 Feature Scaling

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 As range of values of raw data varies widely, the objective functions may not work properly without normalization. Thus, is important to go for feature scaling. There are two ways of feature scaling: Standardization or normalization. The choice between two depends on distribution of data if our data is normally distributed, we go for standardization, else we scale it from 0 to 1.









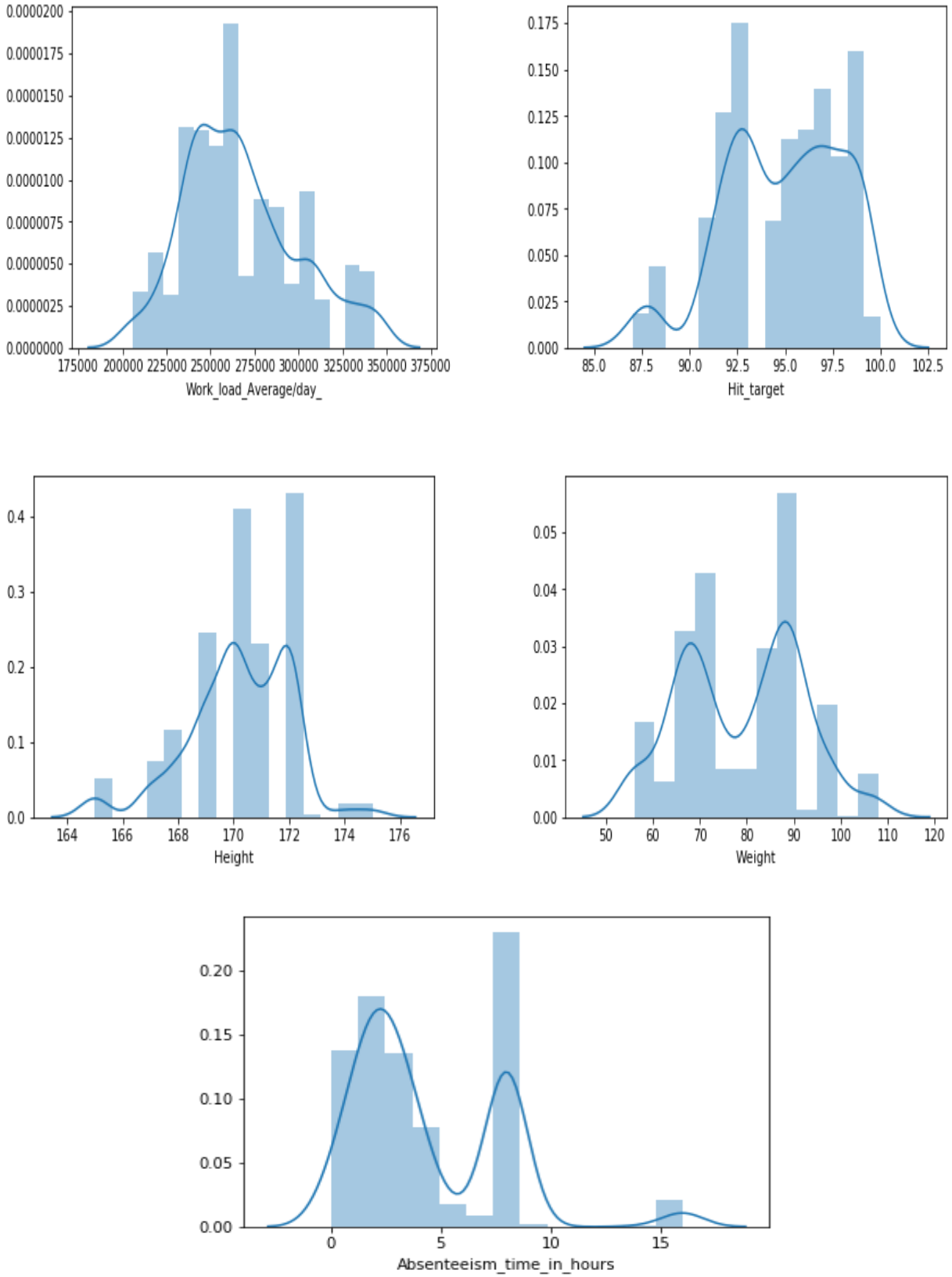


Figure 2.5 Probability distribution function of continuous variables

We plotted the probability density function to check the distribution of data.
In our project, our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

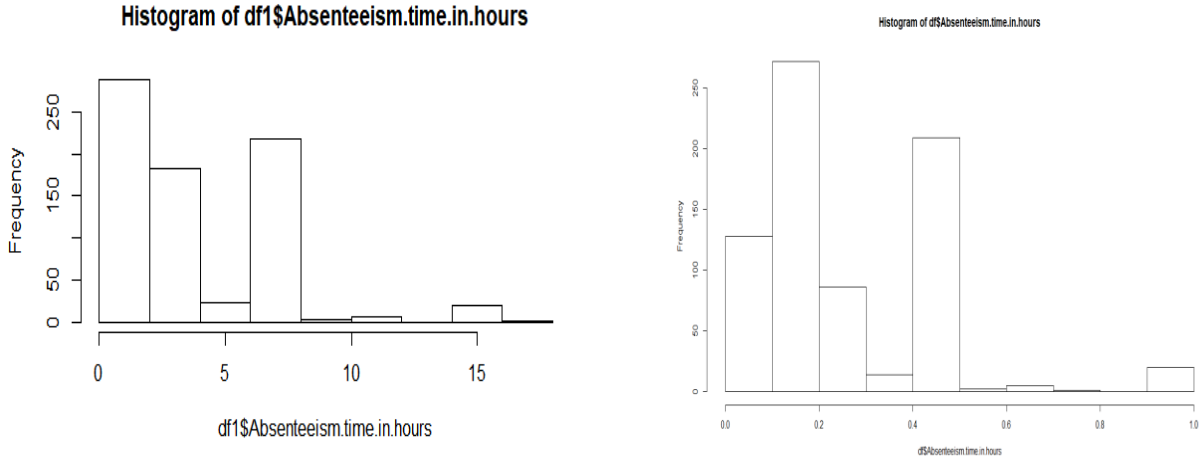


Fig 2.5 the impact of normalization shown as an example for the target variable Absenteeism in hours

In the above figure, the range changes from 0-15 to 0-1

2.1.6 Dimensionality reduction using PCA

We have a few categorical variables with a large number of levels. For example, Reason of Absence has 28 levels and with these many levels, it would be driving the whole model. We need to extract only important variables which can provide us meaning information and are linearly uncorrelated. Principal Component Analysis (PCA) is thus used as a dimensionality reduction technique which performs orthogonal transformation and provide us the linearly uncorrelated components which are basically combination of our features.
As we can observe in the below graph, 40 variables explain almost 99% of variance.

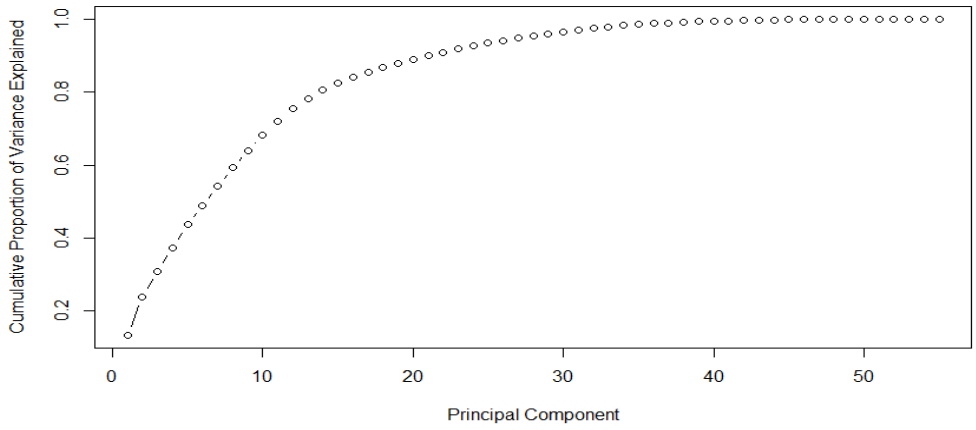


Figure 2.6 Cumulative Proportion Variance graph

2.2 Modeling

Once we are done with the preprocessing, the next step is to build the models using the processed data. As our target variable is continuous, we are using regression models to predict the target variable. Following are the models which we have built –

2.2.1 Decision Tree

A decision tree is a supervised learning approach that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. The advantage of this model is that it can be easily understood by the business users. The RMSE value and R^2 value for our project in R and Python are –

Decision Tree	R	PYTHON
RMSE Train	0.09591796	0.03601
RMSE Test	0.1208695	0.03691
R^2 Test	0.73957300	0.9688666

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 randomly to have more accuracy in dataset. The randomness in building the number of trees and choosing the variable leads to the name random forest. The RMSE value and R^2 value for this project in R and Python is as follows:

Random Forest	R	PYTHON
RMSE Train	0.0345808	0.002233
RMSE Test	0.09673609	0.0054
R^2 Test	0.8809532	0.99931

2.2.3 Linear Regression

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. It can only be applied if the data follows some assumptions. The most important assumption is to check the multicollinearity. While analyzing the data using VIF, we found that data doesn't have multicollinearity problem.

The RMSE and R-square values are shown below

Linear Regression	R	PYTHON
RMSE Train	0.004479807	4.318e-16
RMSE Test	0.004615743	4.407944e-06
R^2 Test	0.999614947	0.999999

2.2.4 Gradient boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Gradient boosting	R	PYTHON
RMSE Train	0.05618817	0.00030414
RMSE Test	0.07755080	0.00030740
R^2	0.90837434	0.9999

Chapter 3 - Conclusion

The next step in the process is to evaluate the model and decide which model is suitable for our data.

3.1 Model Evaluation

The few criteria that we need to look at is RMSE and R-square. As our data is time-series data, RMSE would be the better indicator of errors. In the previous chapter we observed the **Root Mean Square Error** (RMSE) and **R-Squared** Value for different models. **Root Mean Square Error** (RMSE) is the standard deviation of the residuals (prediction **errors**). Residuals are measure of spread of data from regression line data points and RMSE is the SD of residuals. **R-squared** is a relative measure of fit, **RMSE** is an absolute measure of fit
Lower values of **RMSE** and higher value of **R-Squared Value** indicate better fit.

3.2 Model Selection

Before PCA: Initially, we built the model without using PCA as a dimensionality reduction technique. Though the RMSE value was small, R-square was also small (~40%). Thus we decided to go for PCA to reduce dimensions

After PCA:

We observed that for linear regression, R-square was maximum and RMSE minimum and we chose that model for our further analysis.

Also, we can see that RMSE of train and test data don't differ significantly and thus there is no overfitting.

3.3 Answers of asked questions

A. **The Changes which company should bring to reduce the number of absenteeism –**

Methodology to extract important variables: Since PCA extracts information in form of components which is not very intuitive, we used following three methods to see which the major are components impacting our target variable.

1. **Boruta:** We used Boruta package to extract important variables. Below is the output and graphical representation of the same.

"Boruta performed 99 iterations in 48.37362 secs.
Tentatives roughfixed over the last 99 iterations.

31 attributes confirmed important: Age, Day.of.the.week6, Disciplinary.failure0, Disciplinary.failure1, Distance.from.Residence.to.Work and 26 more;
23 attributes confirmed unimportant: Day.of.the.week2, Day.of.the.week3, Day.of.the.week4, Day.of.the.week5, Reason.for.absence11 and 18 more;

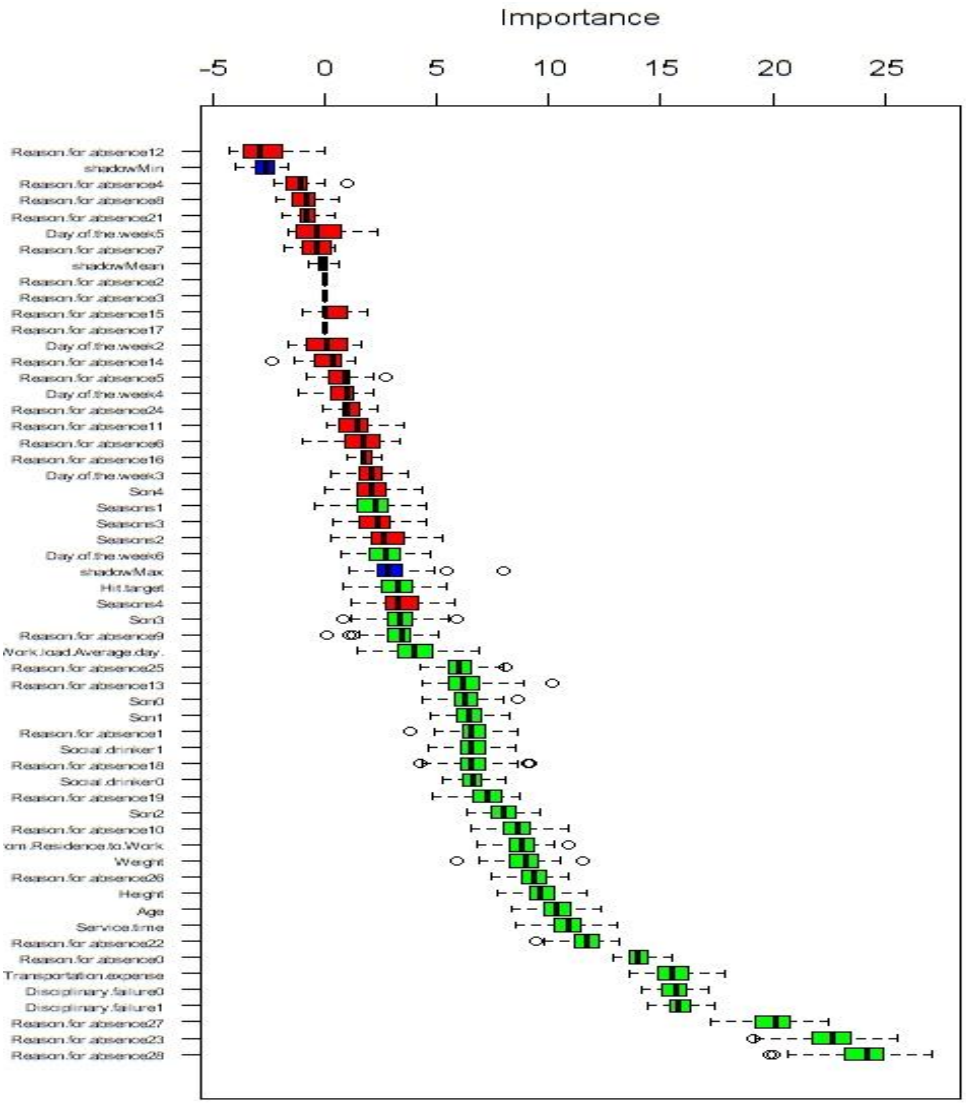


Fig 3.1 Boruta package output to predict important variables

Inference: *We can see Reason of Absence 28, 23, 27 (highlighted in green), Disciplinary failure and Transportation expenses are our important variables.*

2. **Random Forest:**

We also ran random forest to find out the important features and below is the snapshot of our important variables

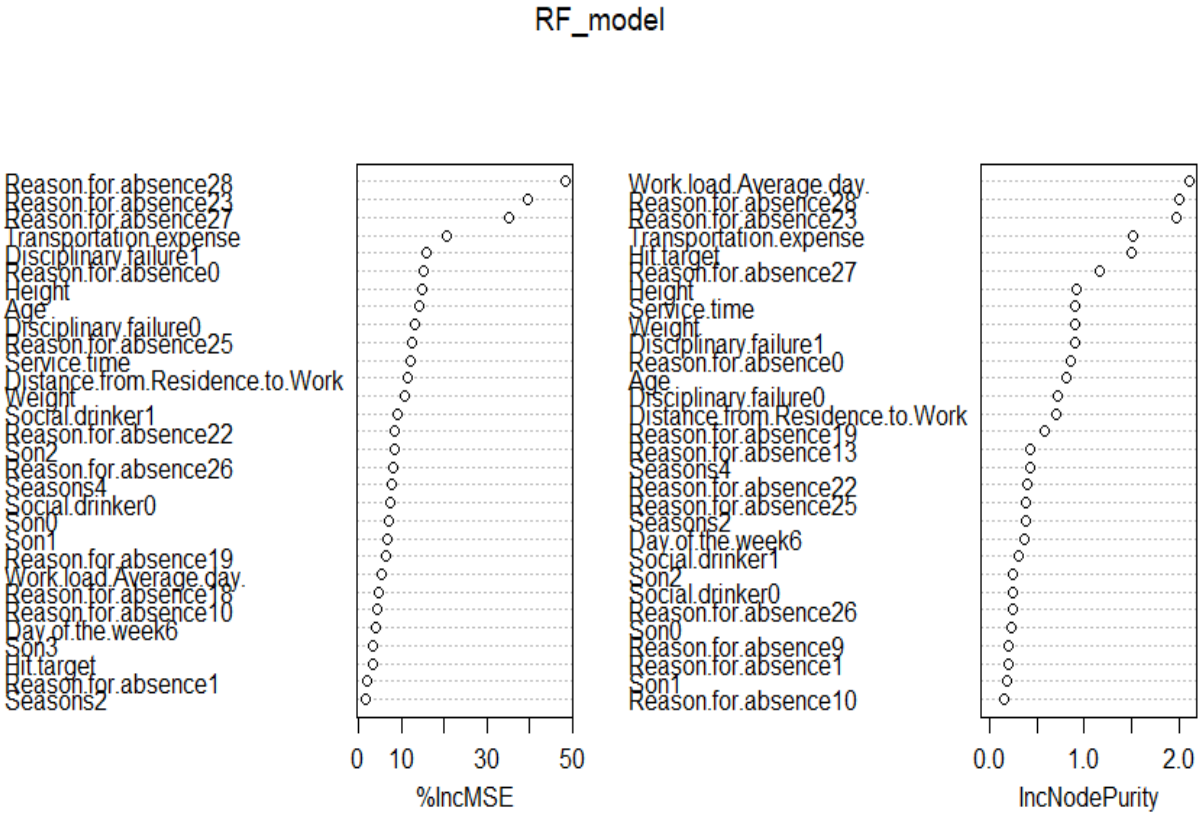


Fig 3.2 Important variables predicted by Random Forest model

Inference: *We can see Reason of Absence 28, 23, 27, Disciplinary failure and Transportation expenses are our important variables*

Linear Regression model

Residuals:

Min	1Q	Median	3Q	Max
-0.41576	-0.07575	-0.00209	0.05670	0.77887

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	0.40626	0.05792	7.014	6.72E-12	***
Reason.for.absence0	-0.236032	0.157026	-1.503	0.13337	
Reason.for.absence1	0.10313	0.046016	2.241	0.02541	*
Reason.for.absence9	0.143263	0.090071	1.591	0.11228	
Reason.for.absence10	0.075033	0.037279	2.013	0.04462	*
Reason.for.absence13	0.057192	0.029434	1.943	0.05251	.
Reason.for.absence18	0.061336	0.038211	1.605	0.10901	
Reason.for.absence19	0.083224	0.031944	2.605	0.00942	**
Reason.for.absence22	0.107293	0.037786	2.84	0.00468	**
Reason.for.absence23	-0.155087	0.022179	-6.993	7.74E-12	***
Reason.for.absence25	-0.145026	0.035994	-4.029	6.37E-05	***
Reason.for.absence26	0.074056	0.036191	2.046	0.0412	*
Reason.for.absence27	-0.192584	0.030341	-6.347	4.54E-10	***
Reason.for.absence28	-0.165403	0.023691	-6.982	8.32E-12	***
Day.of.the.week6	-0.018823	0.017215	-1.093	0.27468	
Seasons2	0.021851	0.018502	1.181	0.23812	
Seasons4	-0.003604	0.01666	-0.216	0.8288	
Transportation.expense	0.096789	0.038838	2.492	0.01299	*
Distance.from.Residence.to.Work	-0.063639	0.031413	-2.026	0.04326	*
Service.time	0.045709	0.061651	0.741	0.45876	
Age	-0.057466	0.038745	-1.483	0.13859	
Work.load.Average.day.	0.006833	0.029638	0.231	0.81775	
Hit.target	-0.016503	0.02997	-0.551	0.5821	
Disciplinary.failure0	NA	NA	NA	NA	
Disciplinary.failure1	-0.096021	0.158043	-0.608	0.54373	

Son0	-0.068829	0.035945	-1.915	0.05603	.
Son1	-0.114457	0.036439	-3.141	0.00177	**
Son2	-0.069078	0.040174	-1.719	0.08609	.
Son3	-0.104306	0.064504	-1.617	0.10643	
Social.drinker0	-0.034003	0.018762	-1.812	0.07048	.
Social.drinker1	NA	NA	NA	NA	
Weight	0.029814	0.035853	0.832	0.40601	
Height	0.01812	0.039494	0.459	0.64655	

Coefficients: (2 not defined because of singularities)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1527 on 558 degrees of freedom
Multiple R-squared: 0.4554, Adjusted R-squared: 0.4261
F-statistic: 15.55 on 30 and 558 DF, p-value: < 2.2e-16

Answer to Question 1

i) From all the above three methods, we can infer that Reason 23, 27, 28 are the three most important variables. Thus, to reduce the absenteeism, company should provide dental consultation, physiotherapy and medical consultation in-house.

ii) Employees with Disciplinary Failure only 0 get absent. So that's another way to control absenteeism.

iii) Employees with less number of sons get more absent.

iv) Employees who have more Distance from work are more absent. So company may work on providing Bus services to pick up employees.

Answer to Question 2

As Month of Absence is not a significant variable, the losses wont vary with the months.
Assuming an employee works 8 hours per day,
Work load average per hour= (Work load average/day)/8
Yearly Losses- Work load average per hour* Absenteesim time in hours.
Since it doesn't vary each month,
Loss each month = Sum(Work load average per hour* Absenteesim time in hours)/12
Answer is 9115291