Employee Absenteeism

Srijith Rajeev

4th Dec 2018

Contents

Chapter 1

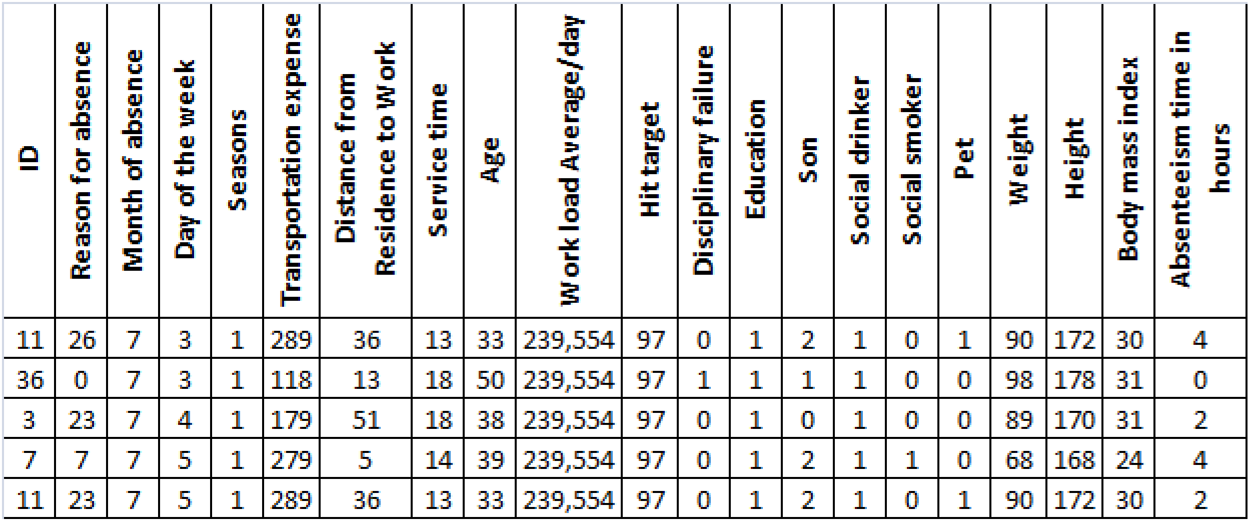
Introduction

1.1 Problem Statement

The objective of this project is to suggest the changes the company should bring to reduce the number of absenteeism and to project the losses month wise.

1.2 Data

A sample of the data set is shown below.



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	Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
11	26	7	3	1	289	36	13	33	239,554	97	0	1	2	1	0	1	90	172	30	4
36	0	7	3	1	118	13	18	50	239,554	97	1	1	1	1	0	0	98	178	31	0
3	23	7	4	1	179	51	18	38	239,554	97	0	1	0	1	0	0	89	170	31	2
7	7	7	5	1	279	5	14	39	239,554	97	0	1	2	1	1	0	68	168	24	4
11	23	7	5	1	289	36	13	33	239,554	97	0	1	2	1	0	1	90	172	30	2

Here we are given 21 variables. Of these, 20 variables are predictor variables and 1 variable is dependent variables. The dependent variable is ‘Absenteeism time in hours’. All other variables are independent/predictor variables

The details of the data attributes in the dataset are:

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)

The data is having numerical and categorical variables. The numeric variables are

- Transportation expense
- Distance from Residence to work
- Service time
- Age
- Work load average per day
- Hit target
- Pet
- Weight
- Height
- Body mass index
- Absenteeism time in hours

Chapter 2

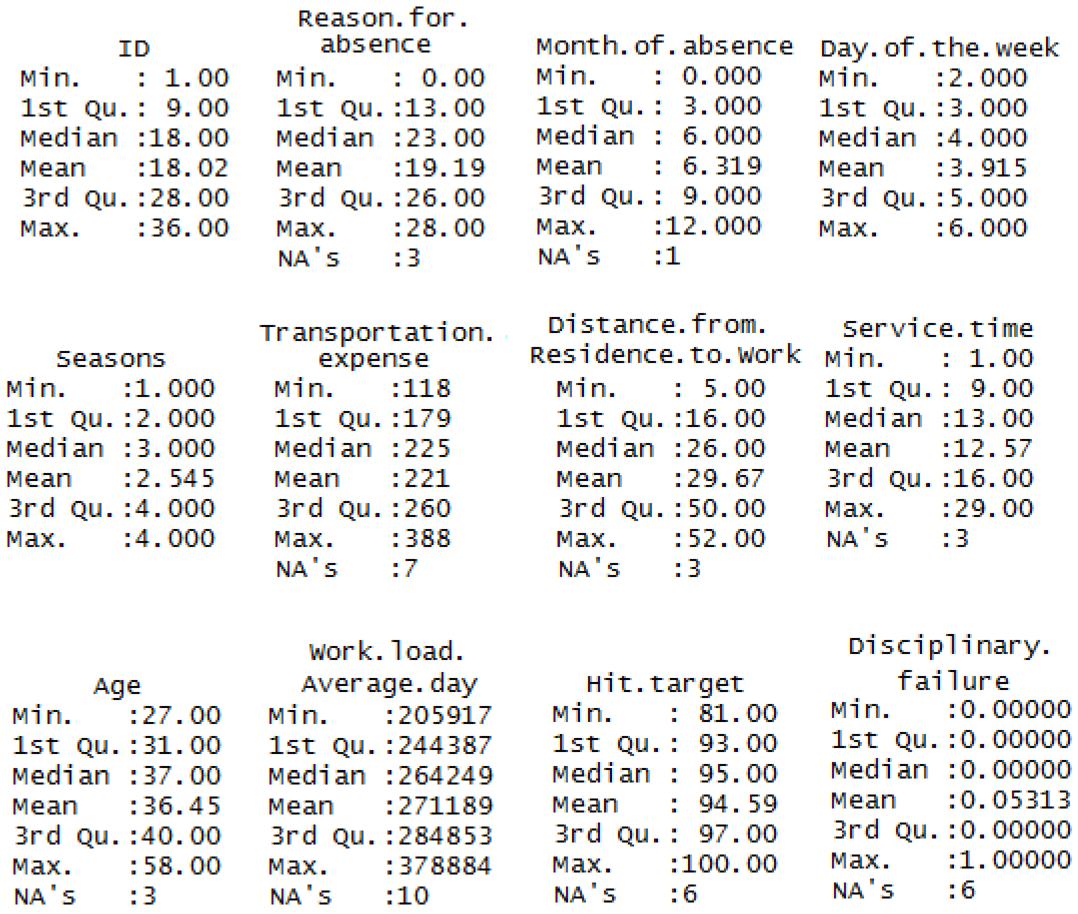
Methodology

2.1 Pre Processing

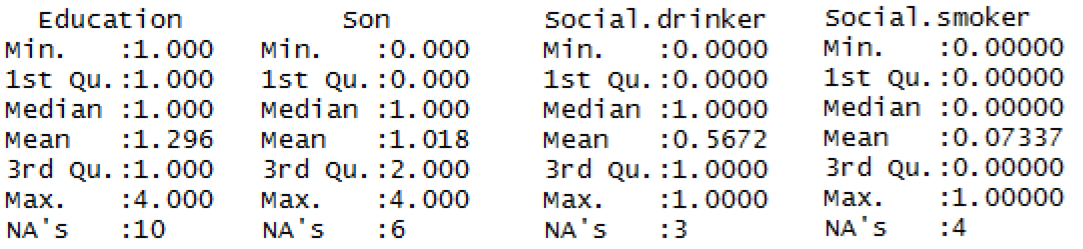
The data has to be explored, cleaned, and visualized before doing predictive modeling, which is often, termed Exploratory data analysis.

2.1.1 Data **Summary**

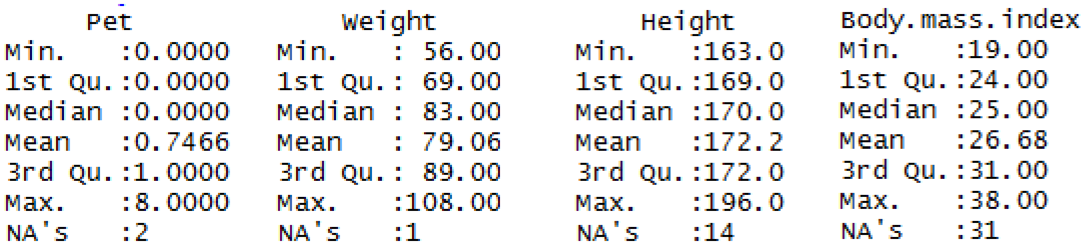
The summary of the data is shown below:



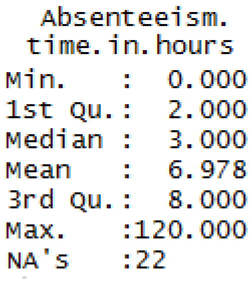
```
                    Reason.for.
       ID             absence       Month.of.absence  Day.of.the.week
 Min.   : 1.00   Min.   : 0.00   Min.   : 0.000   Min.   :2.000
 1st Qu.: 9.00   1st Qu.:13.00   1st Qu.: 3.000   1st Qu.:3.000
 Median :18.00   Median :23.00   Median : 6.000   Median :4.000
 Mean   :18.02   Mean   :19.19   Mean   : 6.319   Mean   :3.915
 3rd Qu.:28.00   3rd Qu.:26.00   3rd Qu.: 9.000   3rd Qu.:5.000
 Max.   :36.00   Max.   :28.00   Max.   :12.000   Max.   :6.000
                 NA's   :3       NA's   :1

                 Transportation.   Distance.from.      Service.time
    Seasons         expense       Residence.to.Work  Min.   : 1.00
 Min.   :1.000   Min.   :118       Min.   : 5.00     1st Qu.: 9.00
 1st Qu.:2.000   1st Qu.:179       1st Qu.:16.00     Median :13.00
 Median :3.000   Median :225       Median :26.00     Mean   :12.57
 Mean   :2.545   Mean   :221       Mean   :29.67     3rd Qu.:16.00
 3rd Qu.:4.000   3rd Qu.:260       3rd Qu.:50.00     Max.   :29.00
 Max.   :4.000   Max.   :388       Max.   :52.00     NA's   :3
                 NA's   :7         NA's   :3

                                                    Disciplinary.
                   Work.load.                         failure
      Age         Average.day       Hit.target     Min.   :0.00000
 Min.   :27.00   Min.   :205917   Min.   : 81.00   1st Qu.:0.00000
 1st Qu.:31.00   1st Qu.:244387   1st Qu.: 93.00   Median :0.00000
 Median :37.00   Median :264249   Median : 95.00   Mean   :0.05313
 Mean   :36.45   Mean   :271189   Mean   : 94.59   3rd Qu.:0.00000
 3rd Qu.:40.00   3rd Qu.:284853   3rd Qu.: 97.00   Max.   :1.00000
 Max.   :58.00   Max.   :378884   Max.   :100.00   NA's   :6
 NA's   :3       NA's   :10       NA's   :6
```



```
   Education          Son          Social.drinker    Social.smoker
 Min.   :1.000   Min.   :0.000   Min.   :0.0000   Min.   :0.00000
 1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.00000
 Median :1.000   Median :1.000   Median :1.0000   Median :0.00000
 Mean   :1.296   Mean   :1.018   Mean   :0.5672   Mean   :0.07337
 3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.00000
 NA's   :10      NA's   :6       NA's   :3        NA's   :4
```



```
      Pet             Weight           Height       Body.mass.index
 Min.   :0.0000   Min.   : 56.00   Min.   :163.0   Min.   :19.00
 1st Qu.:0.0000   1st Qu.: 69.00   1st Qu.:169.0   1st Qu.:24.00
 Median :0.0000   Median : 83.00   Median :170.0   Median :25.00
 Mean   :0.7466   Mean   : 79.06   Mean   :172.2   Mean   :26.68
 3rd Qu.:1.0000   3rd Qu.: 89.00   3rd Qu.:172.0   3rd Qu.:31.00
 Max.   :8.0000   Max.   :108.00   Max.   :196.0   Max.   :38.00
 NA's   :2        NA's   :1        NA's   :14      NA's   :31
```



```
  Absenteeism.
 time.in.hours
Min.   :  0.000
1st Qu.:  2.000
Median :  3.000
Mean   :  6.978
3rd Qu.:  8.000
Max.   :120.000
NA's   :22
```

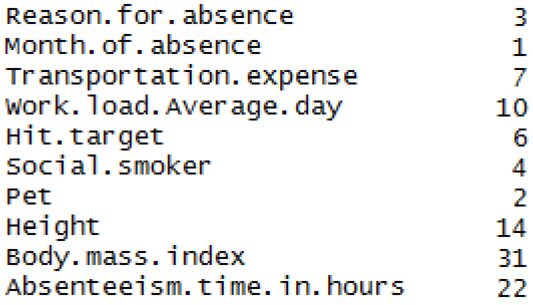
The mean, median, minimum, maximum and quartiles of the numeric variables are shown in the summary. For categorical variables, the number of observations in each category is shown.

It is observed that the the following varibles are employee specific (ID) specific variables, and is a constant for an employee. These variables are

- Transportation expense
- Distance from Residence to work
- Service Time
- Age
- Education
- Son
- Social Drinker
- Social Smoker
- Pet
- Height
- Weight
- Body mass index

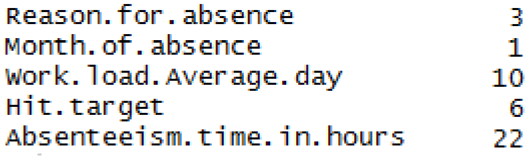
2.1.2 Missing Value Analysis

In the given data, there are missing values. The variable names and corresponding number of missing values are given below.



```
Reason.for.absence              3
Month.of.absence                1
Transportation.expense          7
Work.load.Average.day          10
Hit.target                      6
Social.smoker                   4
Pet                             2
Height                         14
Body.mass.index                31
Absenteeism.time.in.hours      22
```

Missing values in employee specific variables are imputed directly as these values are known. The missing data in other variables and the corresponding count is



```
Reason.for.absence              3
Month.of.absence                1
Work.load.Average.day          10
Hit.target                      6
Absenteeism.time.in.hours      22
```

One observation having parameters of employee ID 28 had a typo. It has been corrected.

The missing data is imputed by Mode.

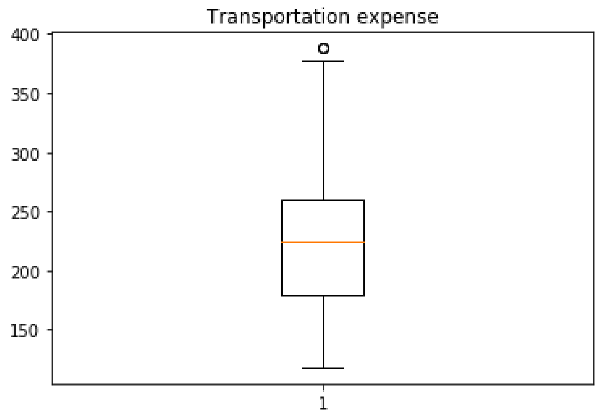
2.1.3 Outlier Analysis

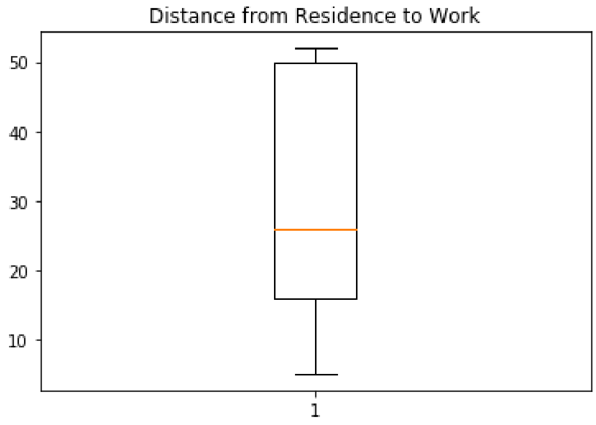
We visualize the outliers using boxplot. The boxplot of the numeric predictor variables are shown below. Here the variables having outliers are list below.

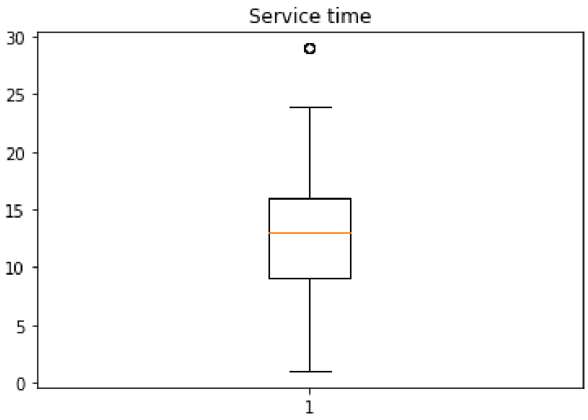
- Transportation Expense
- Service Time
- Age
- Work load average per day
- Hit target
- Pet
- Height
- Absenteeism time in hours

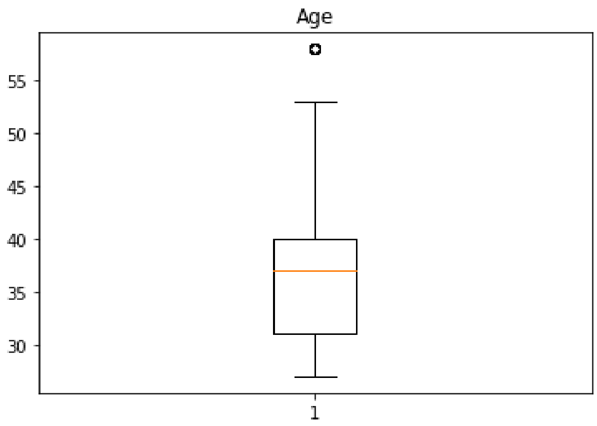
Outliers in employee specific variables are ignored, while observations having outliers in Work load average per day, Hit target, Absenteeism time in hours are dropped from the data.

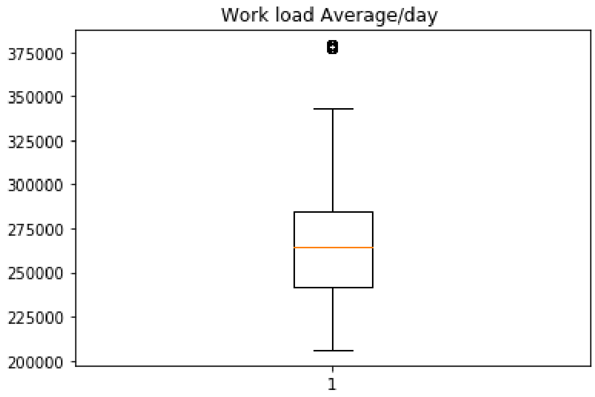
The outliers are replaced with ‘NA’ and imputed by KNN imputation. The histogram of all numeric variables is plotted.

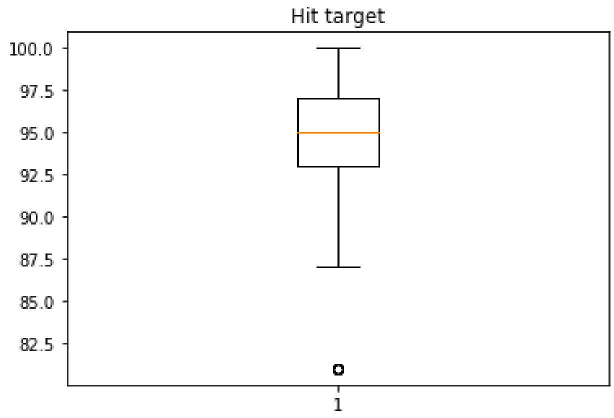

Transportation expense
400
350
300
250
200
150
1

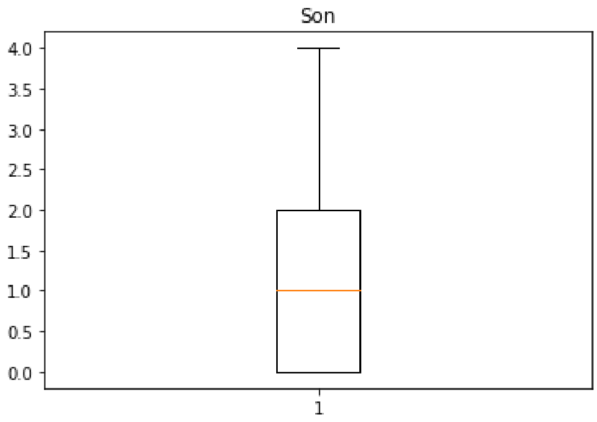

Distance from Residence to Work
50
40
30
20
10
1

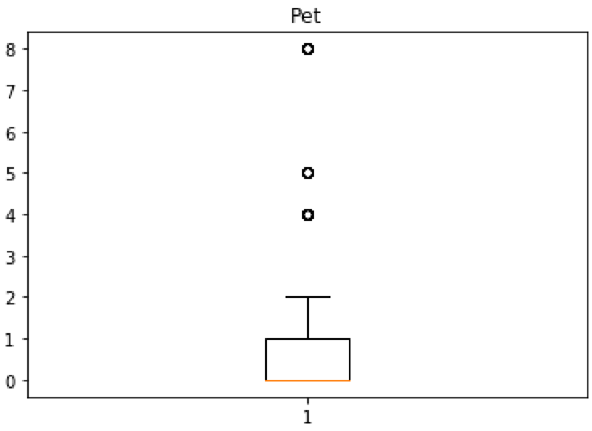

Service time
30
25
20
15
10
5
0
1

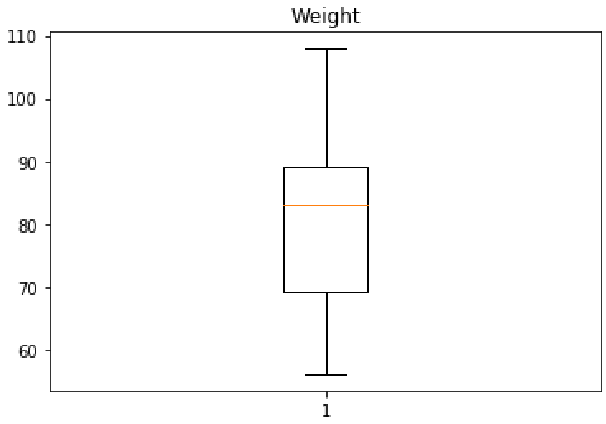

Age
55
50
45
40
35
30
1

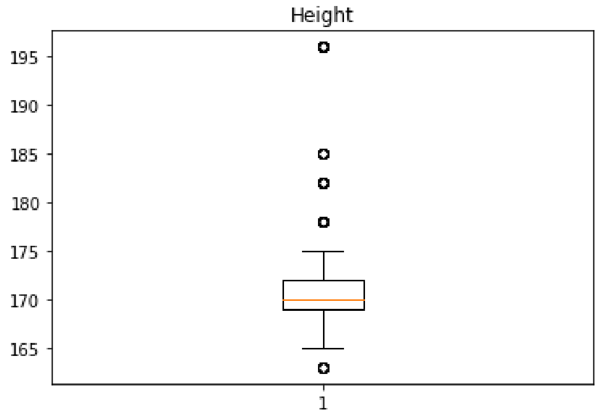

Work load Average/day
375000
350000
325000
300000
275000
250000
225000
200000
1

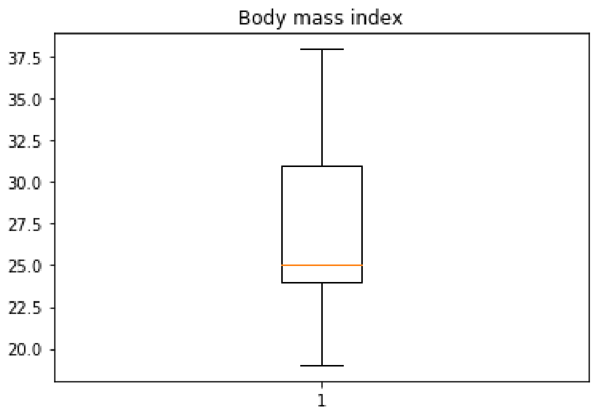

Hit target
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
1

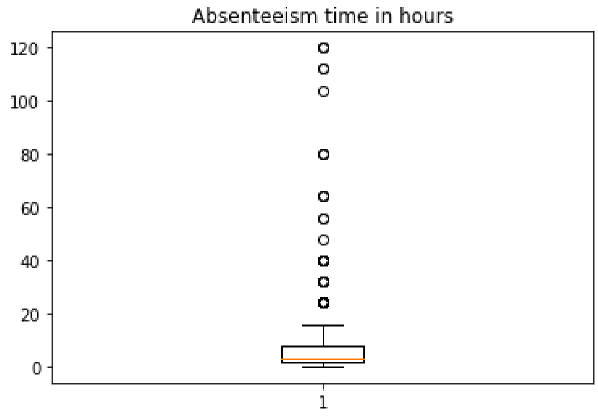

Son
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1


Pet
8
7
6
5
4
3
2
1
0
1

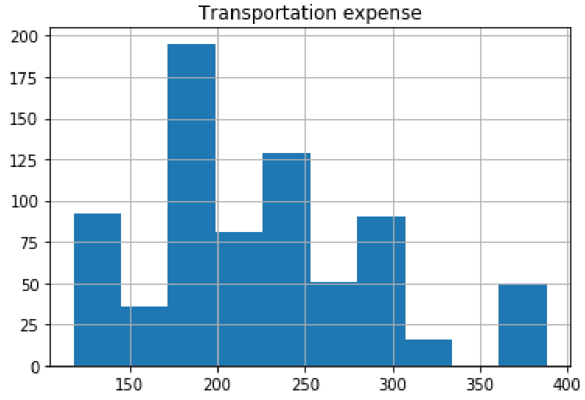

Weight
110
100
90
80
70
60
1

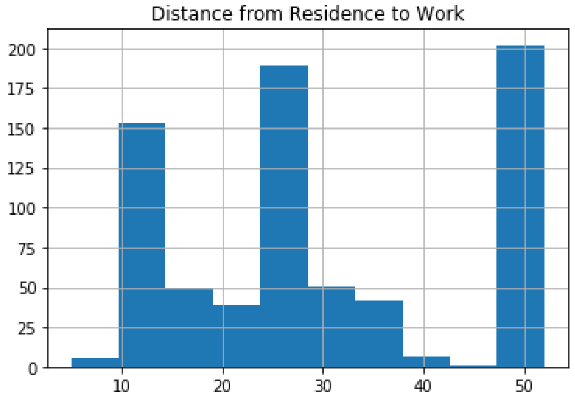

Height
195
190
185
180
175
170
165
1

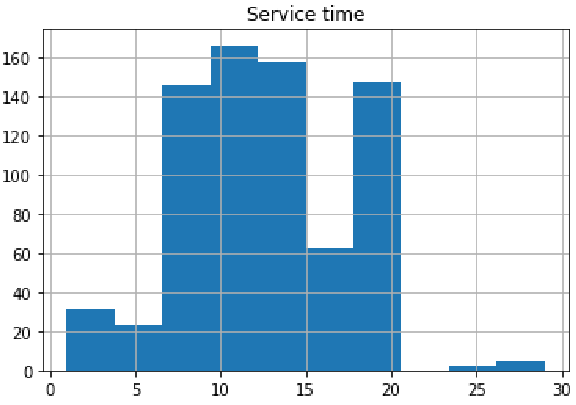

Body mass index
37.5
35.0
32.5
30.0
27.5
25.0
22.5
20.0
1

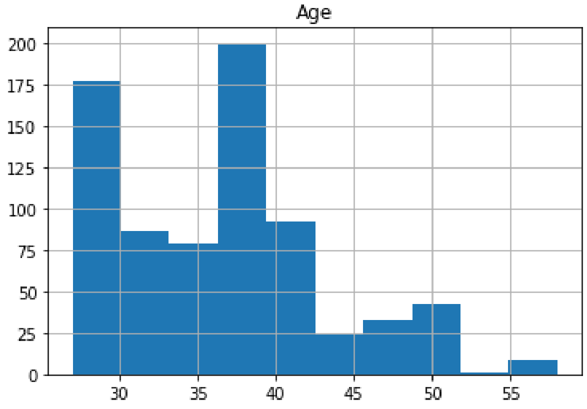

Absenteeism time in hours
120
100
80
60
40
20
0
1

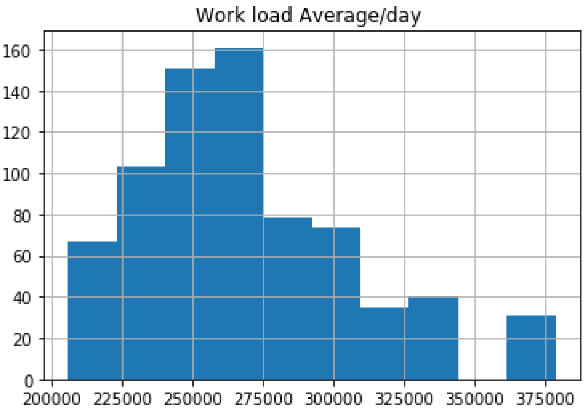
The histogram of the variables are as follows:

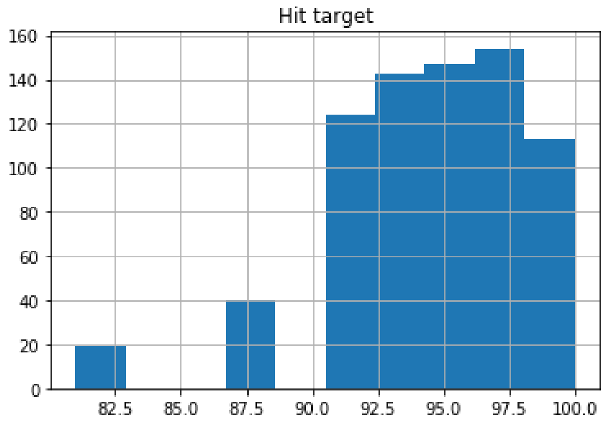


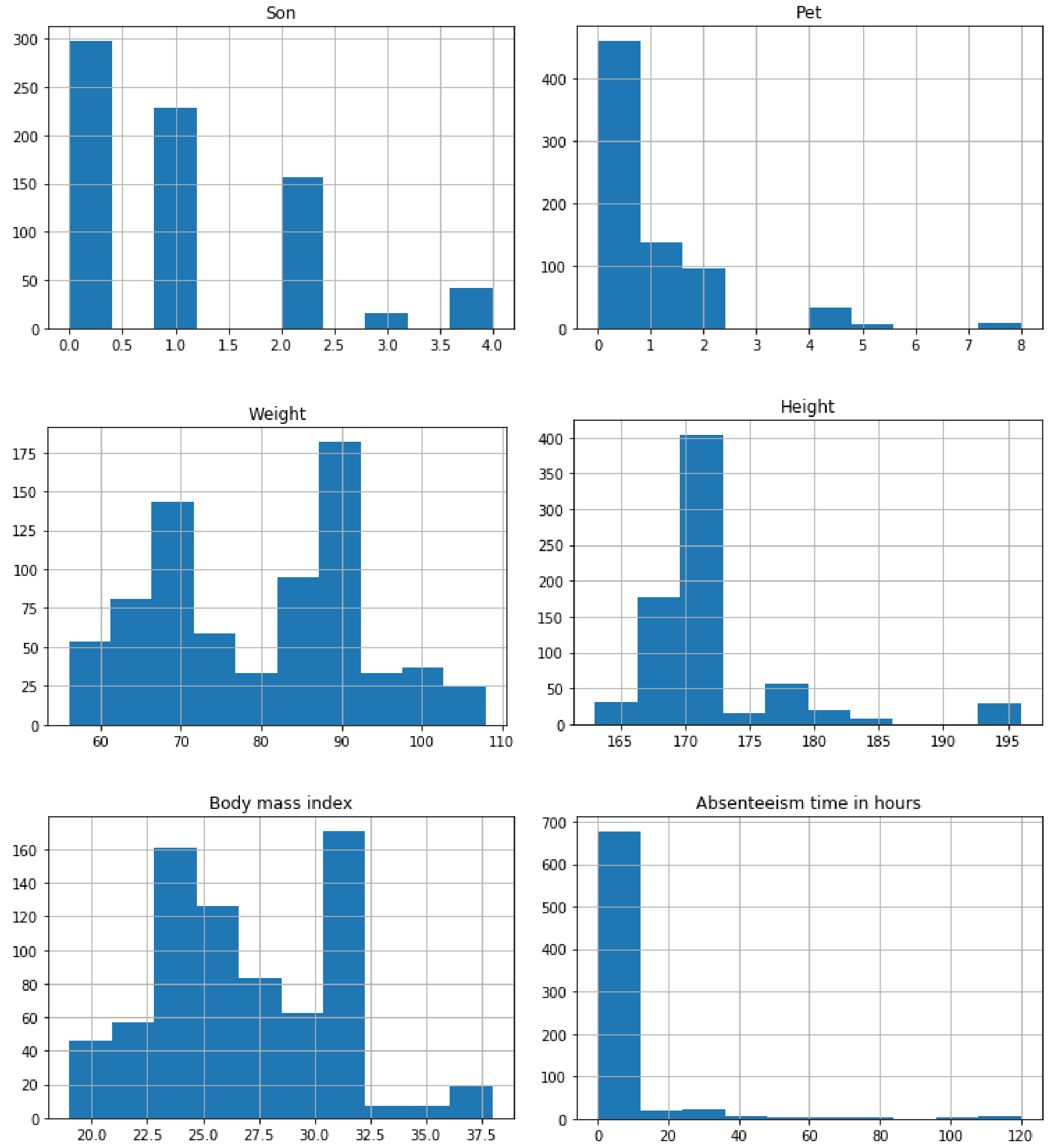






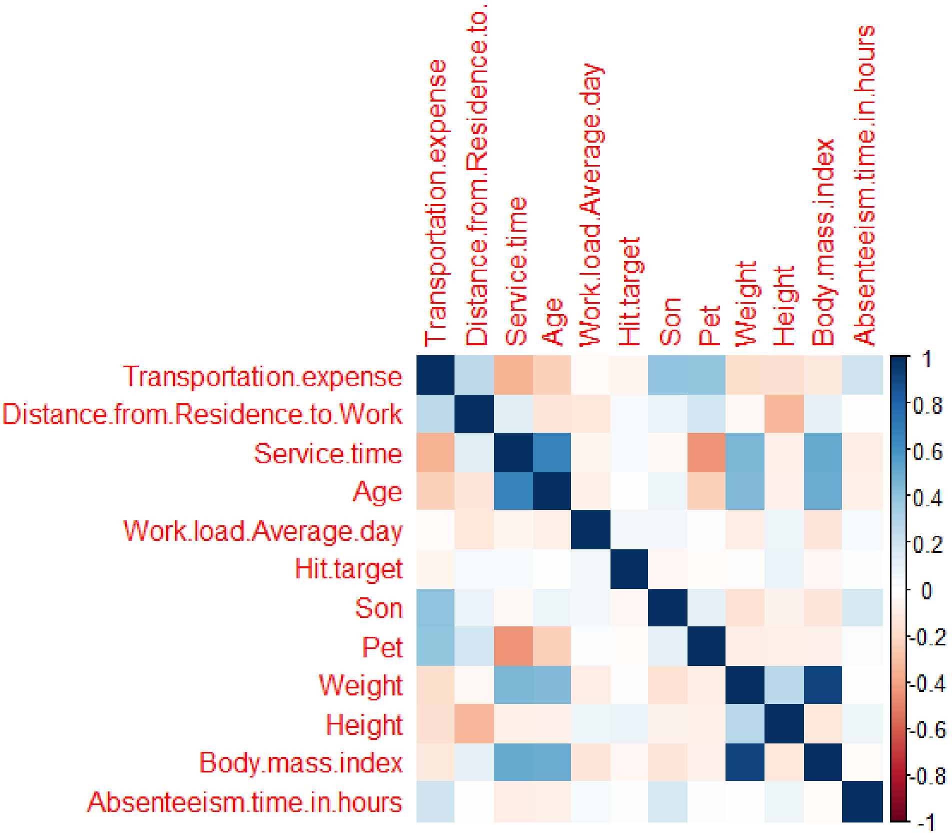






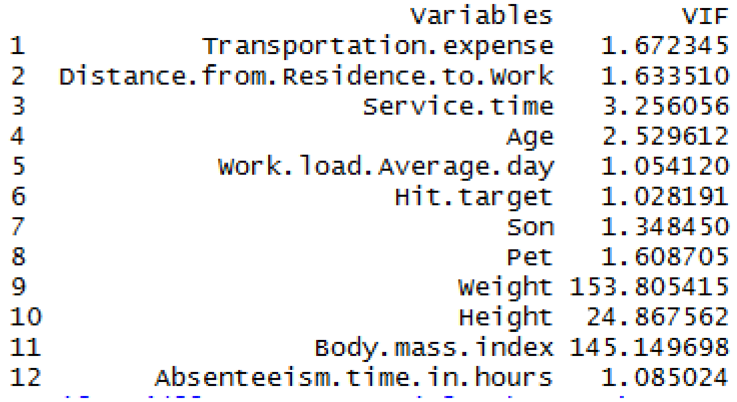
2.1.4 Correlation Analysis

The correlation between the numeric variables is studied. The correlation matrix shows that variables ‘weight’ and ‘body mass index’ are highly correlated in the order of 0.916. The correlation plot is shown below:

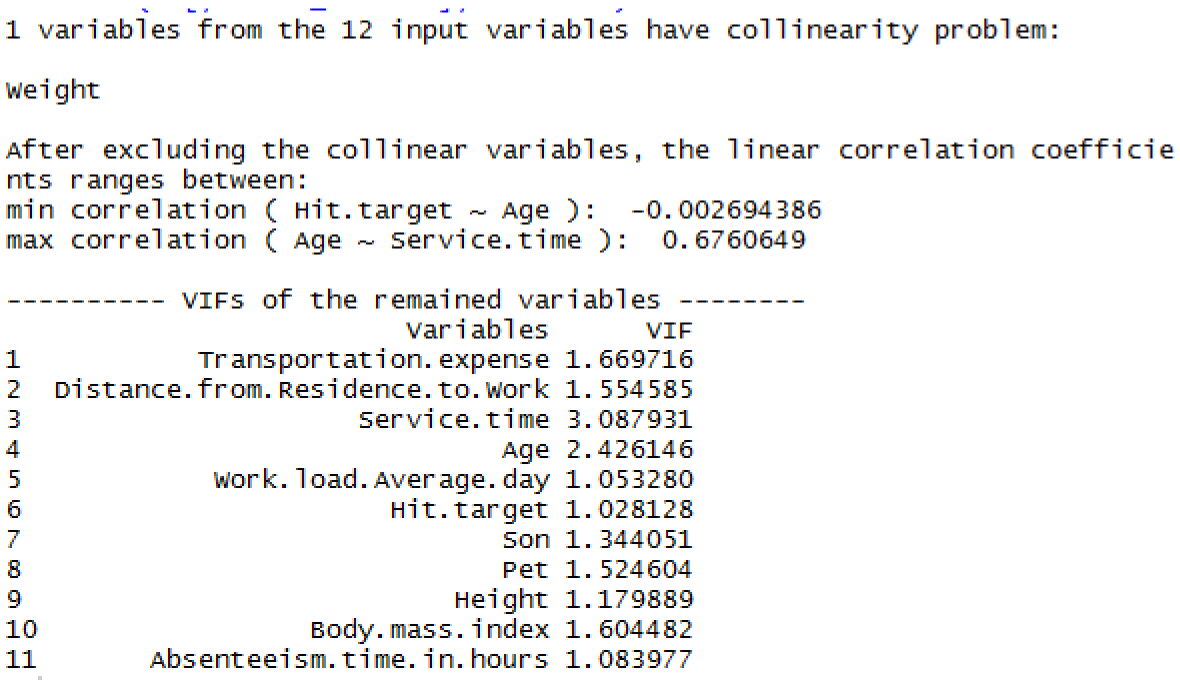


From the correlation analysis, it's evident that the variable 'weight' and 'body mass index' are having a high degree of correlation. So for further analysis, the variable 'weight' is neglected.

VIF Analysis



```
                    Variables        VIF
1          Transportation.expense   1.672345
2  Distance.from.Residence.to.Work   1.633510
3                    Service.time   3.256056
4                             Age   2.529612
5           Work.load.Average.day   1.054120
6                      Hit.target   1.028191
7                             Son   1.348450
8                             Pet   1.608705
9                          Weight 153.805415
10                         Height  24.867562
11                Body.mass.index 145.149698
12      Absenteeism.time.in.hours   1.085024
```



```
1 variables from the 12 input variables have collinearity problem:

Weight

After excluding the collinear variables, the linear correlation coefficie
nts ranges between:
min correlation ( Hit.target ~ Age ):  -0.002694386
max correlation ( Age ~ Service.time ):  0.6760649

---------- VIFs of the remained variables --------
                         Variables      VIF
1            Transportation.expense 1.669716
2   Distance.from.Residence.to.Work 1.554585
3                      Service.time 3.087931
4                               Age 2.426146
5             Work.load.Average.day 1.053280
6                        Hit.target 1.028128
7                               Son 1.344051
8                               Pet 1.524604
9                            Height 1.179889
10                  Body.mass.index 1.604482
11        Absenteeism.time.in.hours 1.083977
```

2.1.5 Train – Test Data

For further analysis after model fitting, the data is divided into train data, cross validation data and test data. The model is trained on train data and its performance is evaluated on test data.

2.2 Modeling

2.2.1 Model Selection

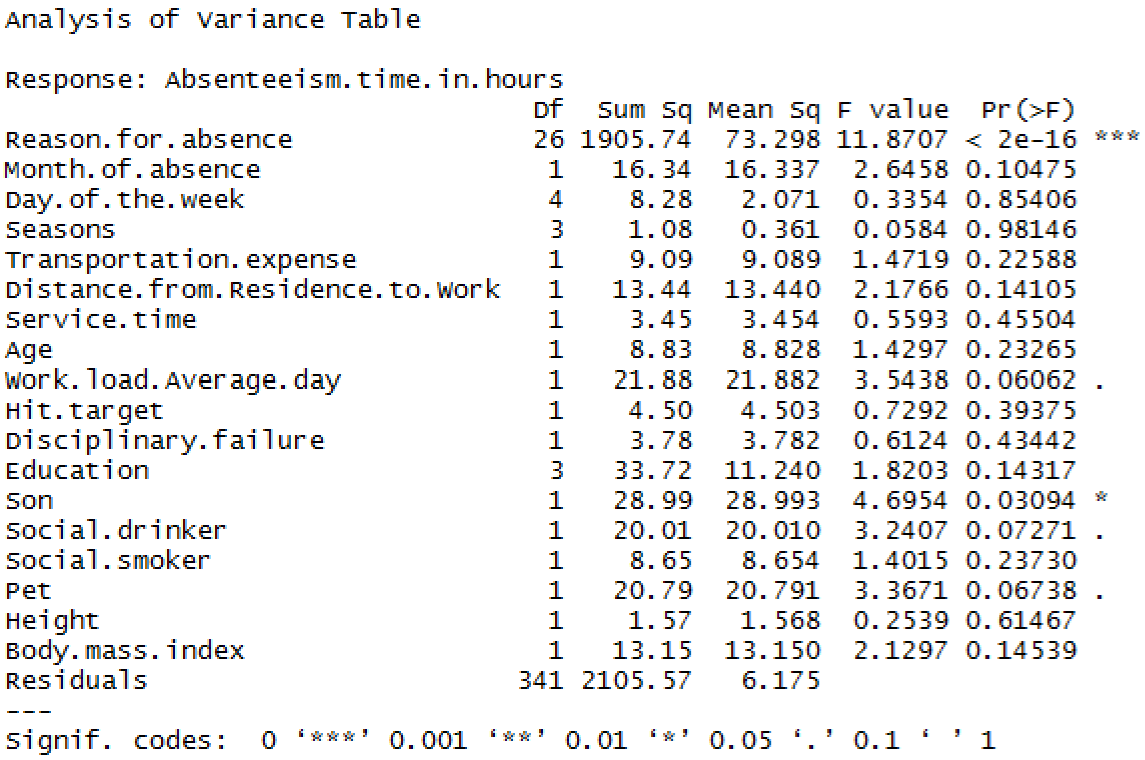
The objective is to predict the hours of employee absenteeism. This is a case of Regression Problem. The models to be fitted on this dataset are

- Multiple Linear Regression
- Decision Tree
- Random Forest

Train data is inputted to the regression model and separate analysis ais done to predict the variable ‘Absenteeism time in hours’.

2.2.2 Multiple Linear Regression

- Multiple linear Regression model is fitted and anova table is prepared.



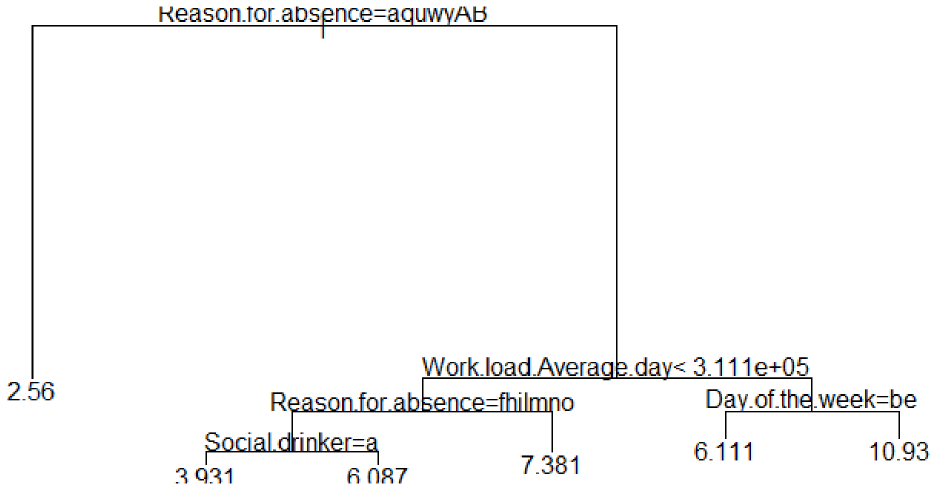
```
Analysis of Variance Table

Response: Absenteeism.time.in.hours
                                 Df  Sum Sq Mean Sq F value   Pr(>F)    
Reason.for.absence               26 1905.74  73.298 11.8707 < 2e-16 ***
Month.of.absence                  1   16.34  16.337  2.6458 0.10475    
Day.of.the.week                   4    8.28   2.071  0.3354 0.85406    
Seasons                           3    1.08   0.361  0.0584 0.98146    
Transportation.expense            1    9.09   9.089  1.4719 0.22588    
Distance.from.Residence.to.Work   1   13.44  13.440  2.1766 0.14105    
Service.time                      1    3.45   3.454  0.5593 0.45504    
Age                               1    8.83   8.828  1.4297 0.23265    
Work.load.Average.day             1   21.88  21.882  3.5438 0.06062 .  
Hit.target                        1    4.50   4.503  0.7292 0.39375    
Disciplinary.failure              1    3.78   3.782  0.6124 0.43442    
Education                         3   33.72  11.240  1.8203 0.14317    
Son                               1   28.99  28.993  4.6954 0.03094 *  
Social.drinker                    1   20.01  20.010  3.2407 0.07271 .  
Social.smoker                     1    8.65   8.654  1.4015 0.23730    
Pet                               1   20.79  20.791  3.3671 0.06738 .  
Height                            1    1.57   1.568  0.2539 0.61467    
Body.mass.index                   1   13.15  13.150  2.1297 0.14539    
Residuals                       341 2105.57   6.175                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above anova table, it’s evident the variables which are significant as p values are less than 0.05 are

- Reason for absence
- Work load average per day
- Son
- Social drinker
- Pet

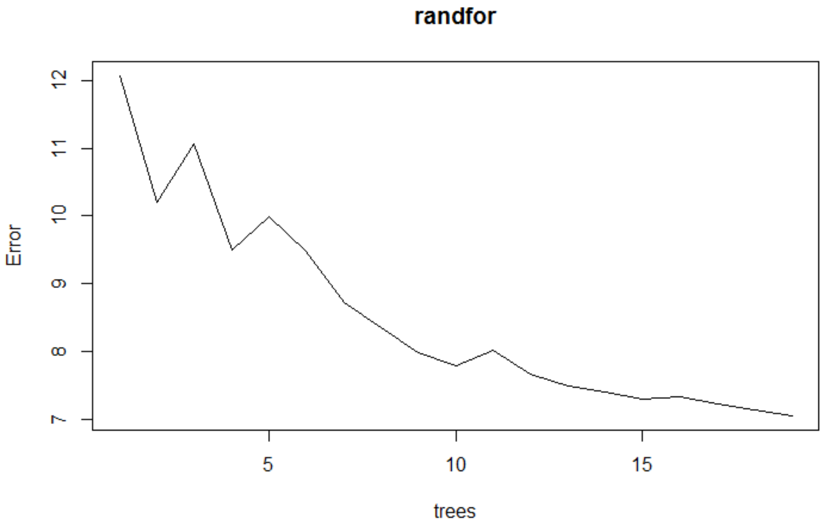
All factors except Pet contribute to increase in Absenteeism time. The R squared value obtained is 0.502

2.2.3 Decision Tree


Reason.for.absence=aquwyAB
2.56
Work.load.Average.day< 3.111e+05
Reason.for.absence=fhilmno
Day.of.the.week=be
Social.drinker=a
3.931
6.087
7.381
6.111
10.93

2.2.4 Random Forest

Similar analysis is done in Random Forest also, with max number of trees limited to 19 as error is minimized with 19 trees.



Chapter 3

Conclusion

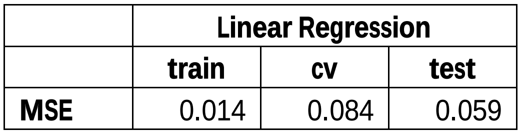
3.1 Model Evaluation

Model evaluation is done by predicting the test data values, using the model which is trained in train data. MSE (Mean Square Error) is the error matrices used for the model evaluation.

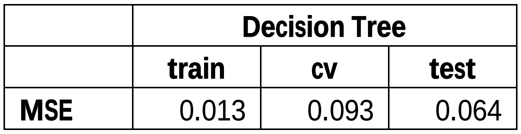
Train data and test data are randomly generated in R and Python. So the results slightly vary. When entire data was used for training, the linear regression results were the same. The trained model is validated in cv data and tested in test data.

3.1.1 Results

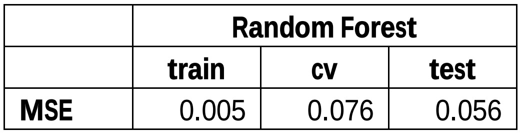
The table having MSE of the three models namely Linear Regression, Decision Tree and Random forest is shown below.



	Linear Regression		
	train	**cv**	**test**
MSE	0.014	0.084	0.059



	Decision Tree		
	train	**cv**	**test**
MSE	0.013	0.093	0.064



	Random Forest		
	train	**cv**	**test**
MSE	0.005	0.076	0.056

The performances of all models are good.

3.2 Suggestions

The factors contributing to increased Employee absenteeism are

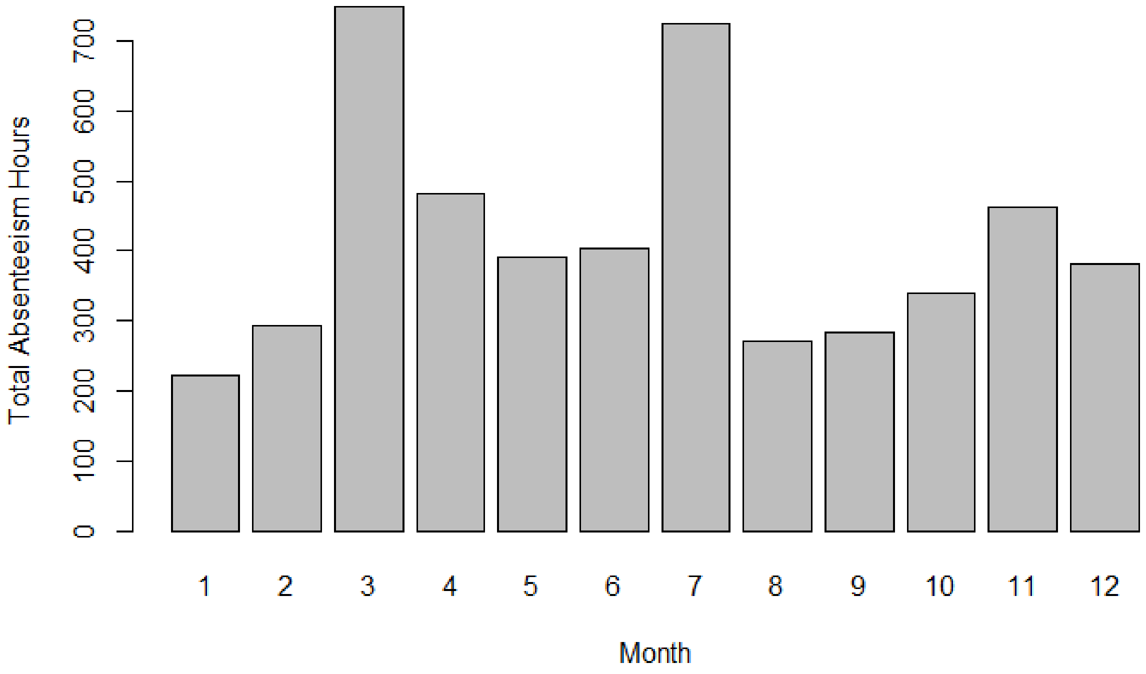
- Medical Reasons for absence
- Work load average per day
- Son
- Social drinker

The variable ‘Pet’ contribute to decrease Employee absenteeism.

An ideal employee is one who is not a social drinker, have lesser number of children, have more pets and having less work load average per day.

3.3 Month wise absentee projection

The month wise distribution (trend) of total hours of Employee absenteeism is



Employee absenteeism trend is more in the months of March and July.