# Project Employee Absenteeism

**Garima**

**06 November 2018**

# Contents

# Chapter 1

# Introduction

## 1.1 Problem Statement :

Human capital plays an important role in courier companies for work like collection ,transportation and delivery.However,absenteeism poses serious threat to the profitability of the company.Our problem at hand is to assist XYZ courier company in :

**i) Formulating policies for reducing the number of changes.**

**ii)To project monthly loss in 2011,if same trend continues.**

## 1.2 Data

**Given :**

**Dataset Details:**

Dataset Characteristics: Timeseries Multivariant
Number of Attributes: 21
Missing Values : Yes

**Attribute Information:**
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)

Give below is the first few observations of the data set that we will be using :

Table 1.1 Absenteeism data set table

| 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 | Weight | Height | Body mass index | Absenteeism time in hours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 90.0 | 172.0 | 30.0 | 4.0 |
| 36 | 0.0 | 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 | 98.0 | 178.0 | 31.0 | 0.0 |
| 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 | 89.0 | 170.0 | 31.0 | 2.0 |
| 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 | 68.0 | 168.0 | 24.0 | 4.0 |
| 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 | 90.0 | 172.0 | 30.0 | 2.0 |
| 3 | 23.0 | 7.0 | 6 | 1 | 179.0 | 51.0 | 18.0 | 38.0 | 239554.0 | ... | 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 89.0 | 170.0 | 31.0 | NaN |
| 10 | 22.0 | 7.0 | 6 | 1 | NaN | 52.0 | 3.0 | 28.0 | 239554.0 | ... | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 4.0 | 80.0 | 172.0 | 27.0 | 8.0 |
| 20 | 23.0 | 7.0 | 6 | 1 | 260.0 | 50.0 | 11.0 | 36.0 | 239554.0 | ... | 0.0 | 1.0 | 4.0 | 1.0 | 0.0 | 0.0 | 65.0 | 168.0 | 23.0 | 4.0 |
| 14 | 19.0 | 7.0 | 2 | 1 | 155.0 | 12.0 | 14.0 | 34.0 | 239554.0 | ... | 0.0 | 1.0 | 2.0 | 1.0 | 0.0 | 0.0 | 95.0 | 196.0 | 25.0 | 40.0 |
| 1 | 22.0 | 7.0 | 2 | 1 | 235.0 | 11.0 | 14.0 | 37.0 | 239554.0 | ... | 0.0 | 3.0 | 1.0 | 0.0 | 0.0 | 1.0 | 88.0 | 172.0 | 29.0 | 8.0 |

# Chapter 2

# Methodology

## 2.1 Missing Value Analysis :

We can categorise our feature set into numerical and categorical feature set consisting of following features in the data:

**Numerical features set :**
"ID","Transportation.expense","Distance.from.Residence.to.Work","Service.time","Age","Work.load.Average.day.","Hit.target","Son","Pet","Height","Weight","Body.mass.index","Absenteeism.time.in.hours"

**Categorical features set** :
"Reason.for.absence","Month.of.absence","Day.of.the.week","Seasons","Disciplinary.failure","Education","Social.drinker","Social.smoker"

Before proceeding with any analysis ,we must get a feel of the data set at hand .We will first evaluate missing value in the data.

Many times, data set has missing value due to various reasons may be error in collection or error in reporting the data .Let us find out how many data points are missing in our data set.
Also in the data set it was observed that some predictors like

"Reason.for.absence","Month.of.absence","Day.of.the.week","Seasons","Education","ID","Age","Weight","Height","Body.mass.index" had '0' value in the observation .

Logically '0' values for these predictors are not acceptable and can be treated as missing values. We replace these values with NA in the data and then do missing value analysis.

Below is a summary of the missing value in our data set .

Table 2.1 Missing Value Table

| Features | NA_Sum | NA_Percent |
|---|---|---|
| ID | 0 | 0.0000000 |
| Reason.for.absence | 46 | 6.2162162 |
| Month.of.absence | 4 | 0.5405405 |
| Day.of.the.week | 0 | 0.0000000 |
| Seasons | 0 | 0.0000000 |
| Transportation.expense | 7 | 0.9459459 |
| Distance.from.Residence.to.Work | 3 | 0.4054054 |
| Service.time | 3 | 0.4054054 |
| Age | 3 | 0.4054054 |
| Work.load.Average.day. | 10 | 1.3513514 |
| Hit.target | 6 | 0.8108108 |

| | | |
|---|---|---|
| Disciplinary.failure | 6 | 0.8108108 |
| Education | 10 | 1.3513514 |
| Son | 6 | 0.8108108 |
| Social.drinker | 3 | 0.4054054 |
| Social.smoker | 4 | 0.5405405 |
| Pet | 2 | 0.2702703 |
| Weight | 1 | 0.1351351 |
| Height | 14 | 1.8918919 |
| Body.mass.index | 31 | 4.1891892 |
| Absenteeism.time.in.hours | 22 | 2.9729730 |

However we see that no column has more than 30 % of the missing data.
Thus we keep all the feature set for our further analysis.

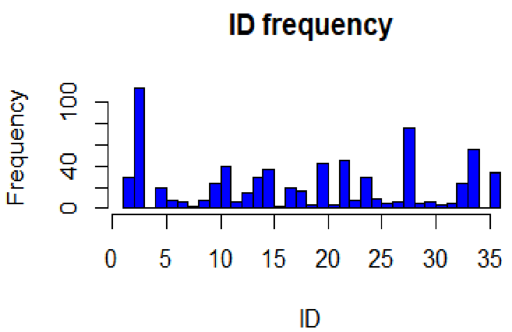
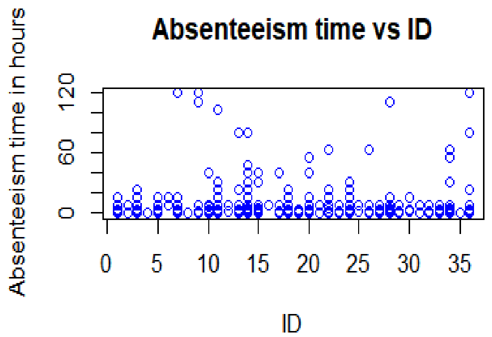
Out of the many methods ,we test following 3 methods i.e, mean mode method ,median mode method and knn method We see that median mode suits best in our dataset.Not only this ,we have also observed many features has ‘0’ as input which makes no sense .Thus we replace them as “NA” data point.

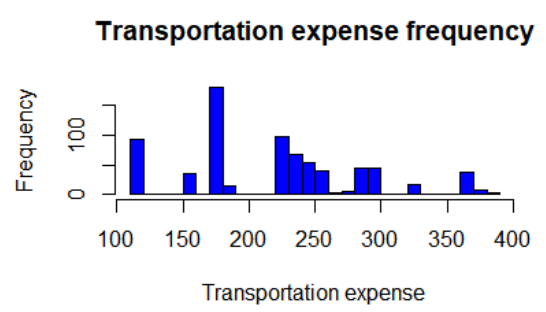
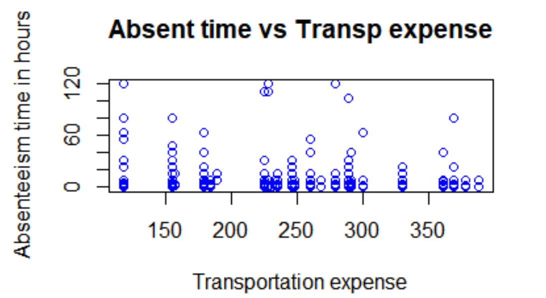
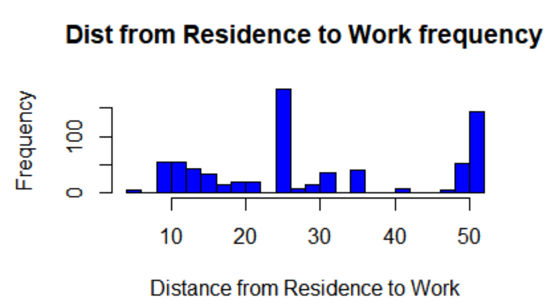
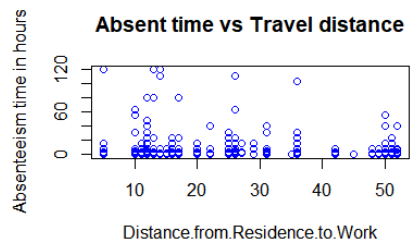
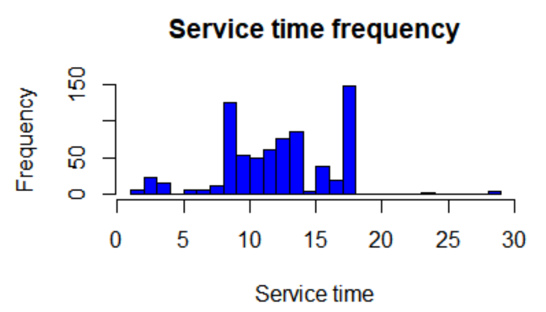
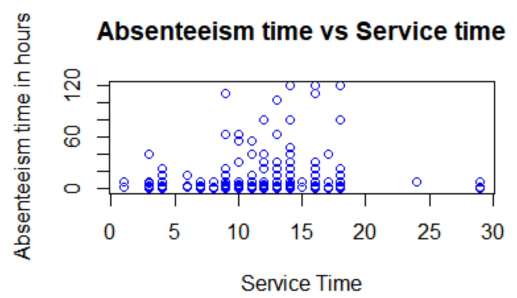
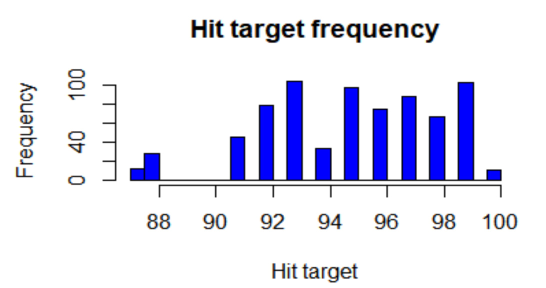
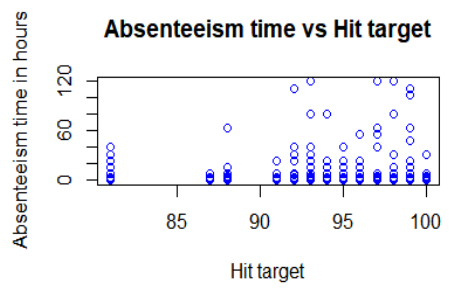
## 2.2 Data visualisation

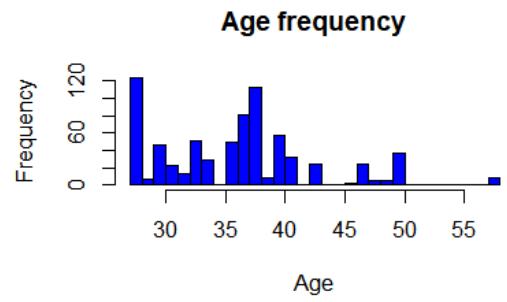
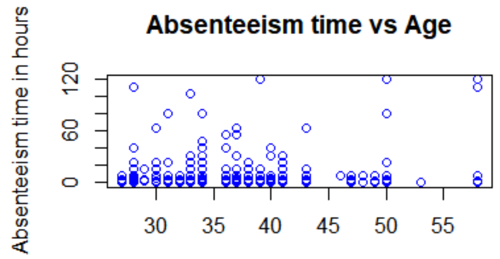
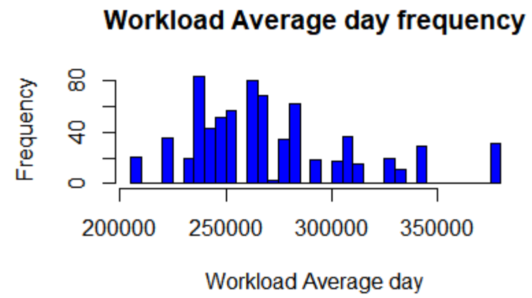
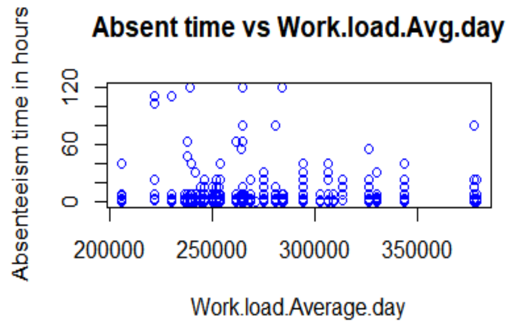
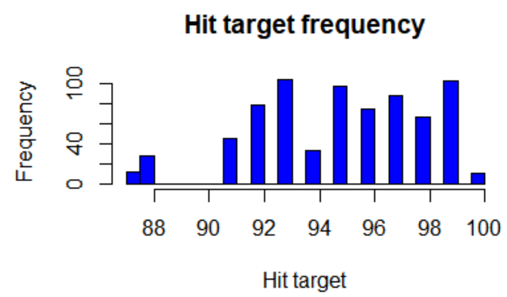
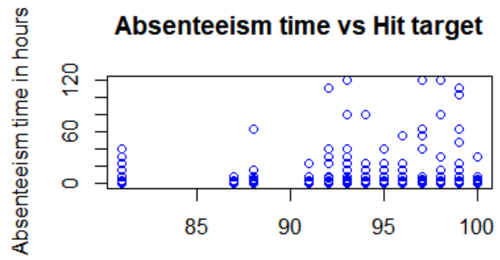
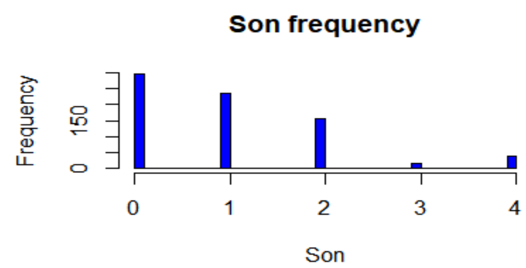
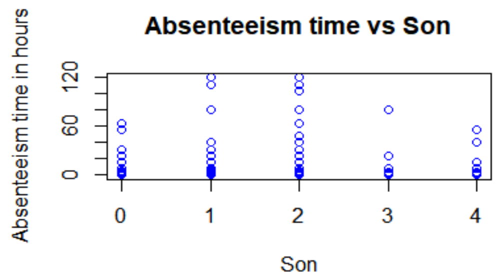
After imputation let us visualise our dataset ,to get a pictorial representation:

**Numerical set:**

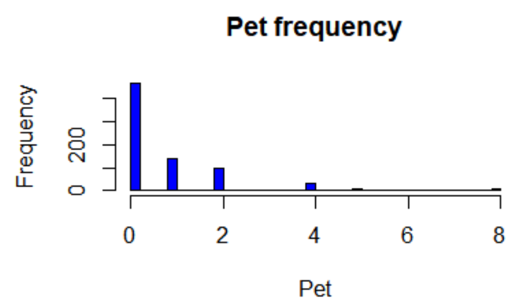
Table 2.2 Table showing distribution of each feature (left) and scatter plot of feature variable vs the target variable

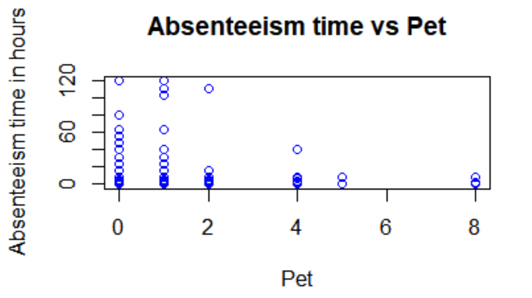
| S.no. | Feature variable frequency distrubution | Feature variable vsTarget variable |
|---|---|---|
| **1.** |  |  |
| | **ID :** The frequency distribution of the “ID” feature set shows a non uniform.However,around ID 3 is observed to be occuring most frequently as compared to others.On plotting this feature against the target variable absenteeism hours ,it can be seen that plot is scattered in the entire range of IDs .However ,the dots are concentrated in in less than 20 hours.This implies most of the implies remain absent for less than 20 hours. | |

<table>
<tr><td>2.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Transportation expense : Most of the transportation expense is around 175 and the distribution is non uniform.Also on comparing its trend with respect to the absenteeism hours we see that absenteeism is found for all ranges of transportation expense .Also the hours of absenteeism is less than 20hours.</td></tr>
<tr><td>3.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Distance from resident to work :Residential distance is found to occur mostll for around 25 units or more than 50 units.Absenteeism is found mostly for residential distance around 10 to 30 units and some concentrated around 50 units.</td></tr>
<tr><td>4.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Service time : The observation set has employees with service time around 7-17 hours.And maximum concentration of absenteeism is also found in this range.However ,there also exist some outliers seen around 25 and 30 units</td></tr>
<tr><td>5.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Hit target : Hit target is around 92 to 99 in our data set.On plotting ,it against our target variable ,we see there is some higher amount of absenteeism hours around 92 – 93 and around 98-99 units.Some concentration of absenteeism is also seen on the lower end of hit target</td></tr>
</table>

| | | |
|---|---|---|
| 6. |  |  |
| | **Age :** The sample set in our study is mainly in the age group of 25 – 40 as depicted in the age frequency plot.Also when we compare it against absenteeism hours we see that absenteeism hours is also concentrated in this range with peak around 33-34 years of age. | |
| 7. |  |  |
| | **Work load average day :**From the frequency plot we see that work load average is spread across the entire range with peak around 260000 – 270000 units.Absenteeism hours is concentrated around 240000 – 270000 units | |
| 8. |  |  |
| | **Hit target :** Hit target frequency is mainly between 91-99 units in our data set.Higher absenteeism hours can be seen around 93 unit and 97 units. | |
| 9. |  |  |
| | **Son frequency :** Most of the employees in the data set has less than or equal to 2 sons.Higher absenteeism hours is seen for employees with 1 and 2 sons**.** | |

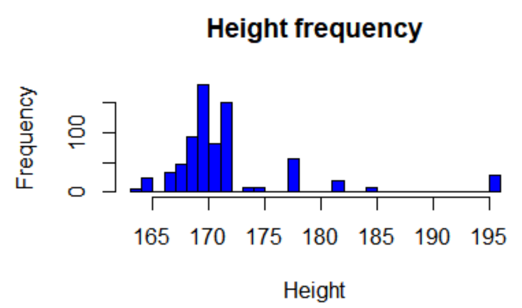
**10.**

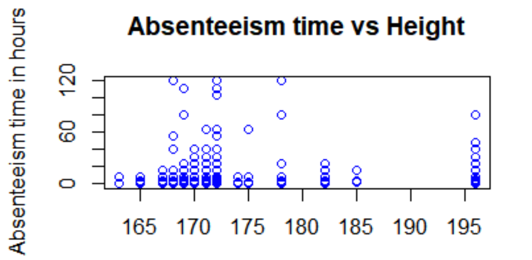



**Pet :** Most of the employees don't have any pets .However,some employees have 1 or 2 pet.We can see some data points around 4 and 8 which can be considered as outliers.Absenteeism hours is high around 0 and 1 pet.

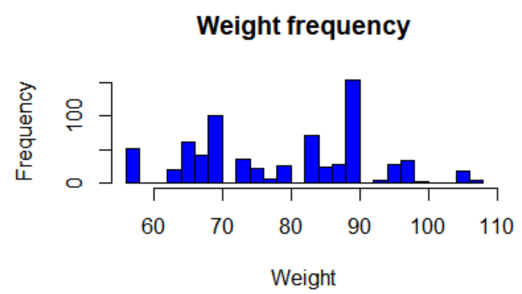
**11**

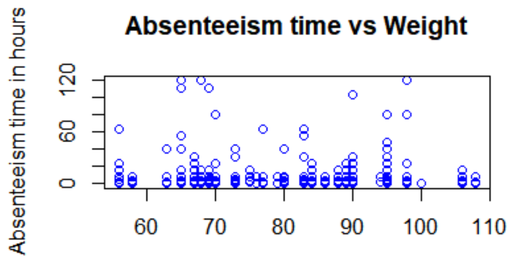



**Height :** The employees in our sample set are mostly of 170 units high.Absenteeism hours is mostly concentrated 168-172 units of height.Some outlier can be see for more than 195 units also.

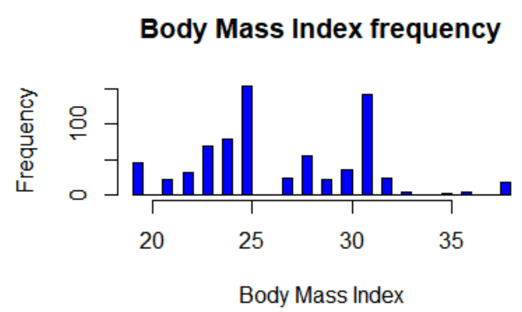
**12.**

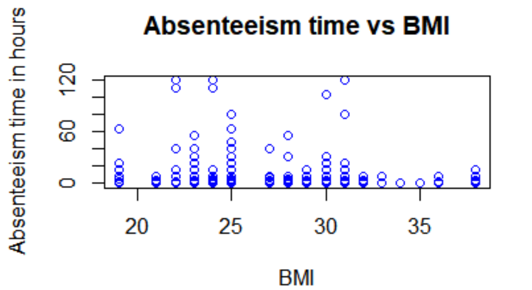



**Weight :**Higher concentration of weight in our sample set is seen around 90 and 70 units.Absenteeism hours is alsmost spread across all weight groups.With higher hourse of absenteeism in 65 unit and 95 unit.

**13.**

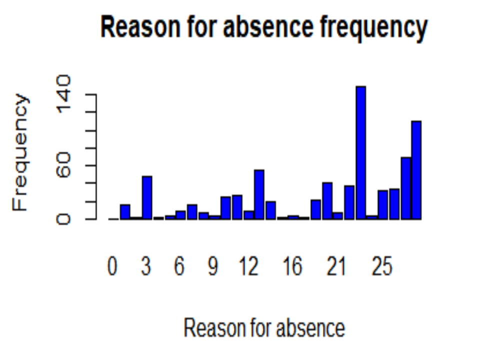
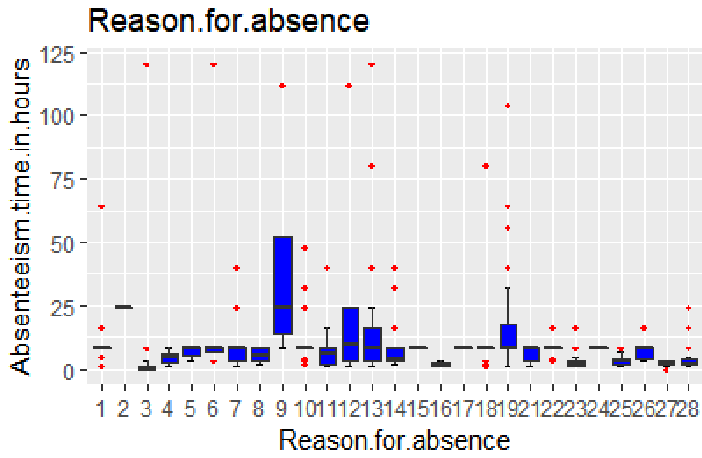
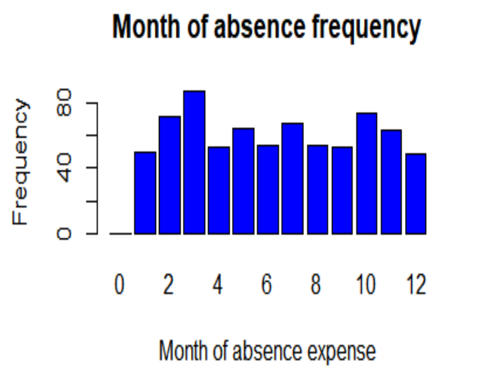
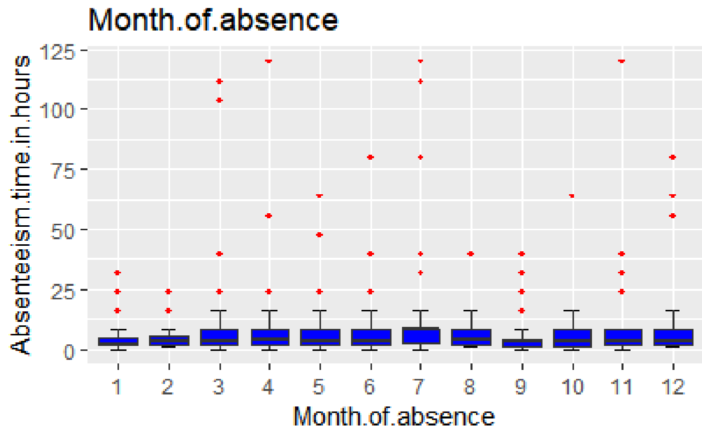
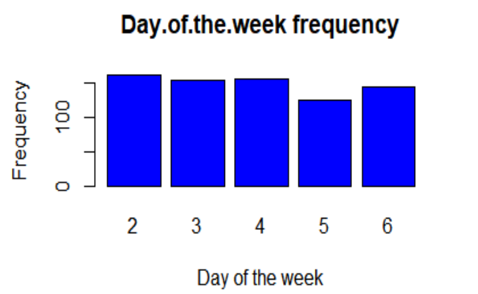
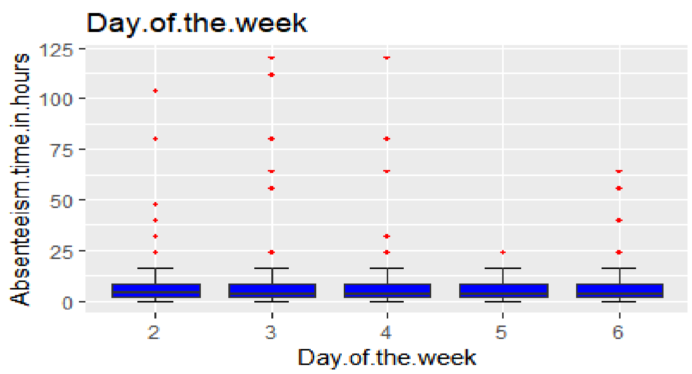


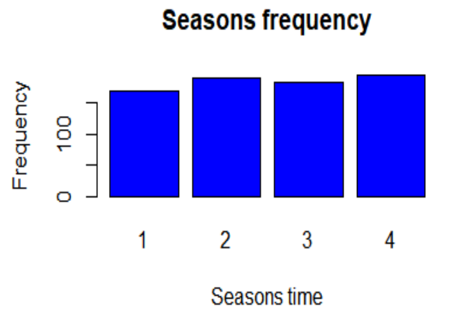
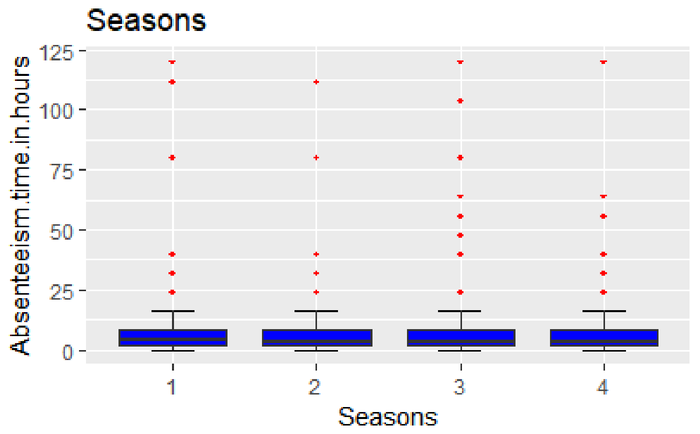
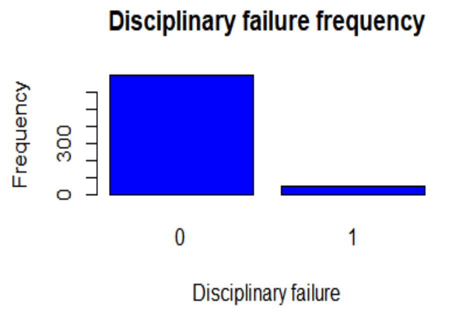
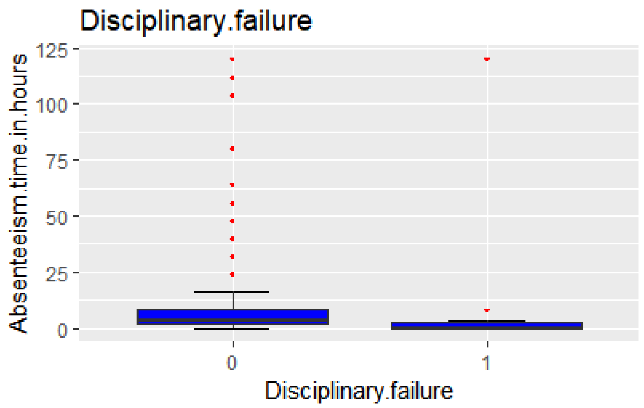
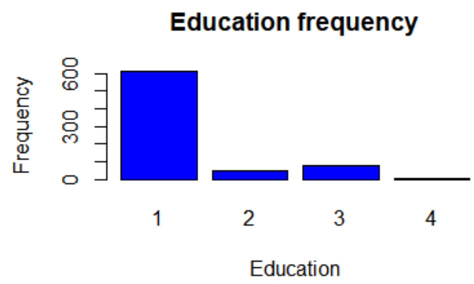
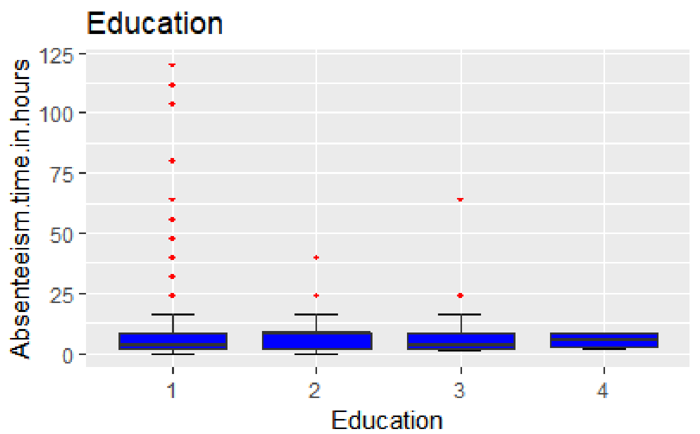



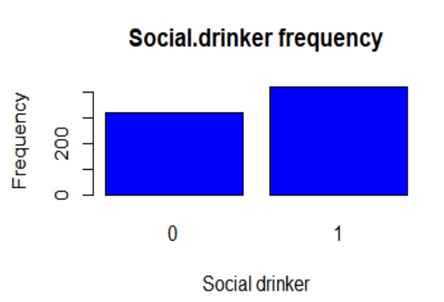
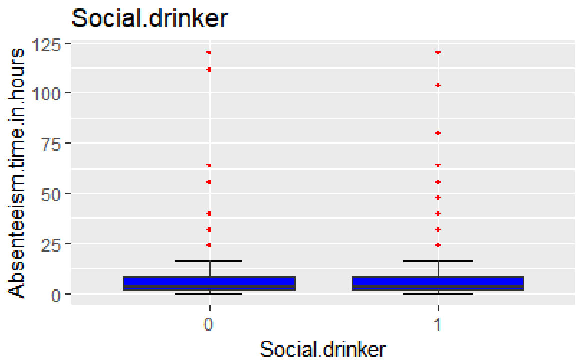
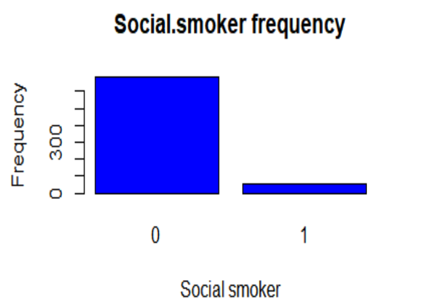
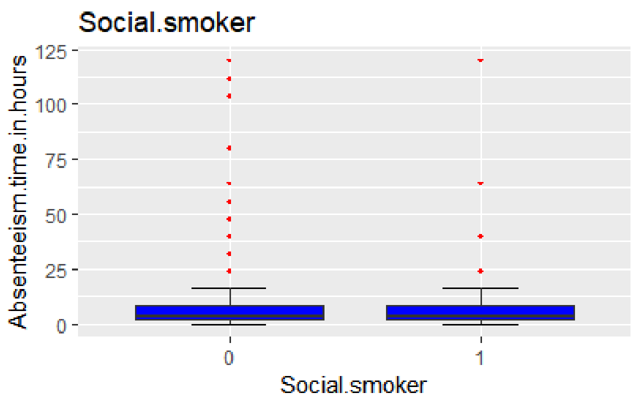

**Body Mass Index :** BMI of the employees in our data set is mostly around 25 and 31.Higher hours of absenteeism is found around 25 and 31.

**<u>Categorical features :</u>**

Table 2.3 Table showing distribution of each feature (left) and box plot of feature variable vs the target variable

<table>
<tr><td>1.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Reason for absence : We see maximum frequency of absent reason is reason no. 23(medical consultation) and 28 (dental consultation).maximum median is found for reason no. 9 (Disease of Circulatory system).We see that there is wide range of median absenteeism hours for this feature set</td></tr>
<tr><td>2.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Month of absence : Maximum frequency of absenteeism is in the March month.However on seeing the boxplot against absenteeism hours,we observe median value is more or less same ,specially around 3-8 months</td></tr>
<tr><td>3.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Day of the week : Lowest frequency of data in the given data set is observed on the 5$^{th}$ day i.e Friday.However the range and median values are almost uniform for all the days of the week.</td></tr>
</table>

<table>
<tr><td>4.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Season:The frequency of season in our data set is almost uniform and so is the median and range of absenteeism hours.</td></tr>
<tr><td>5.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Disciplinary failure : Occurrence of disciplinary failure is less in our data set.On comparing it against absenteeism hours,we see that absenteeism hours is found slightly higher in employees with no disciplinary failure.</td></tr>
<tr><td>6.</td><td>
</td><td>
</td></tr>
<tr><td></td><td colspan="2">Education : Sample set of the employees in XYZ company are mostly found to be high school educated.The range and median of absenteeism hours grouped by the education level is mostly uniform.Also high school educated employees show more number of absenteeism hours.</td></tr>
</table>

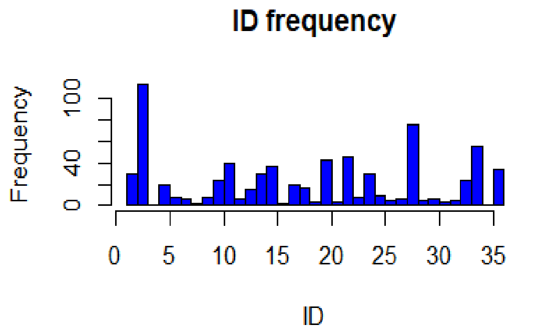
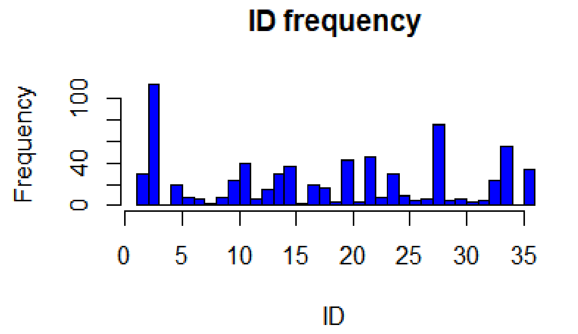
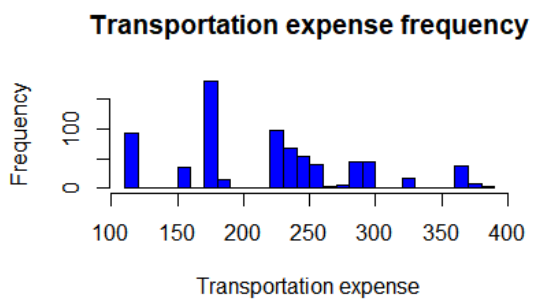
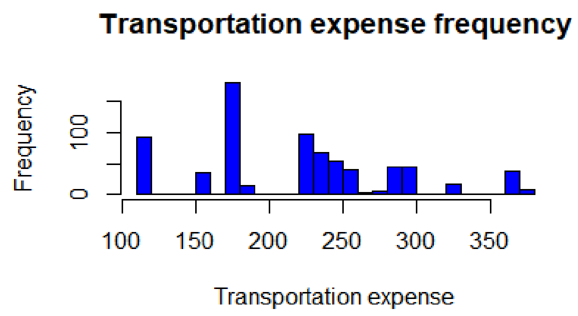
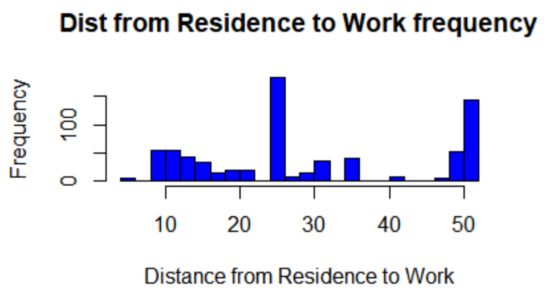
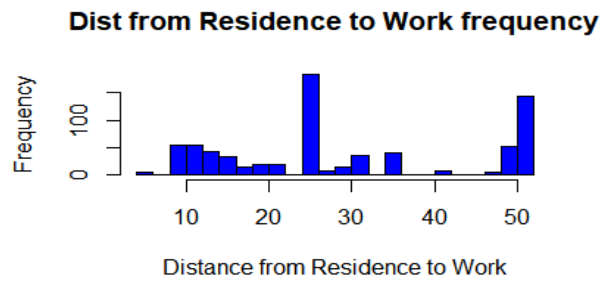
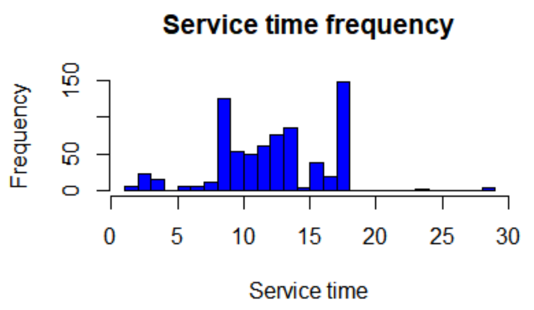
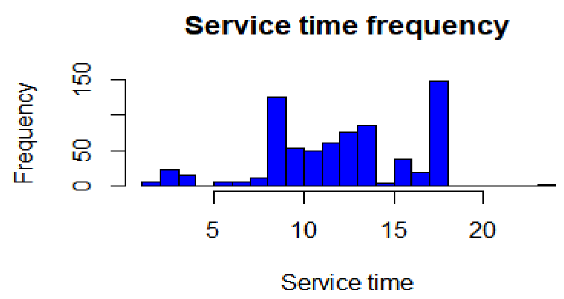
| | | |
|---|---|---|
| 7. |   |   |
| | **Social drinker :** There are slightly more social drinkers in our data set.However,when plotted against absenteeism hours ,they tend to show similar range and median values. | |
| 8. |   |   |
| | **Social smoker :** There are mostly no social smoker in our data set.When we observe absenteeism hours grouped by social smoker ,the median value and the range are almost the same. | |

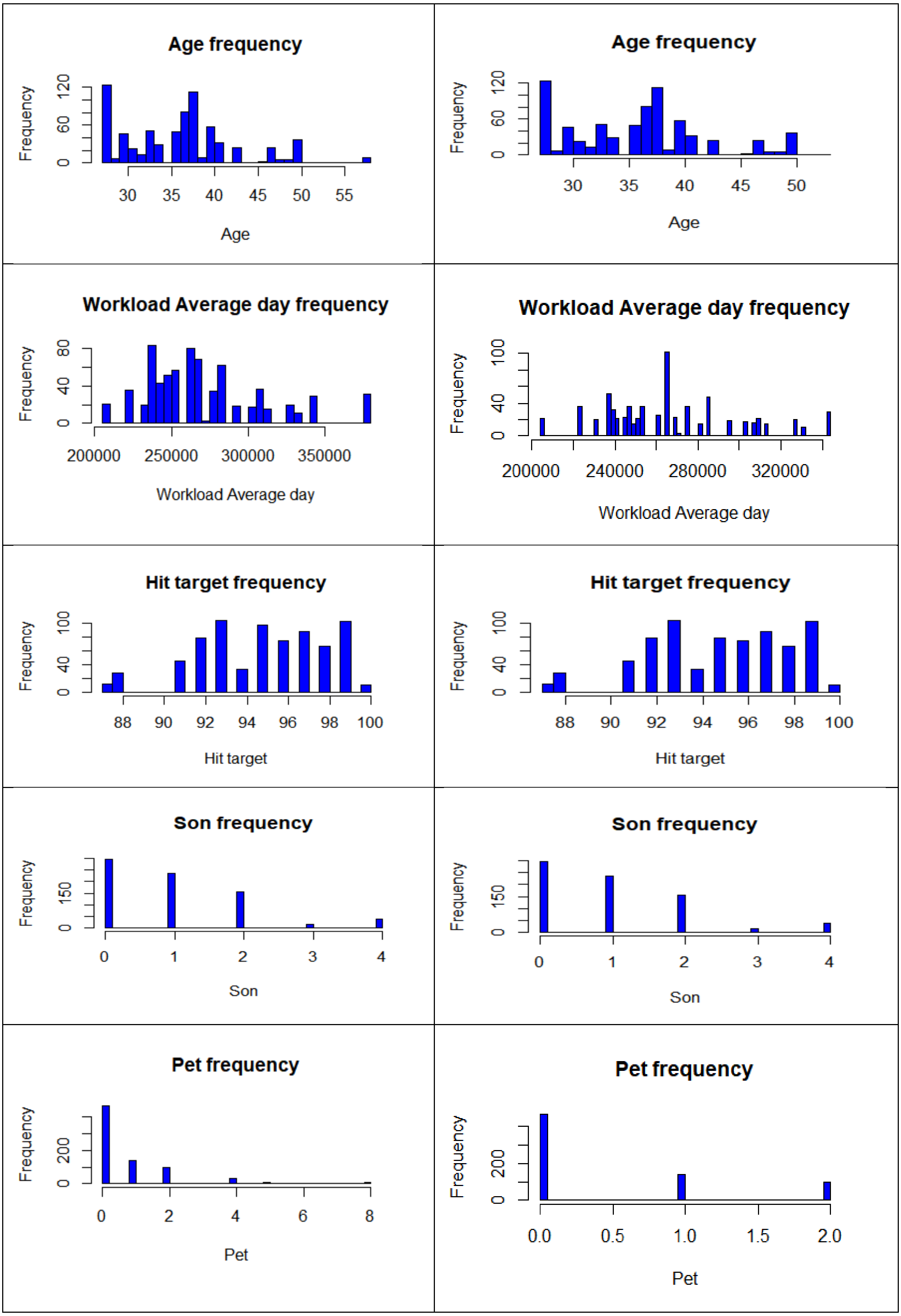
## 2.3 Outlier Analysis :

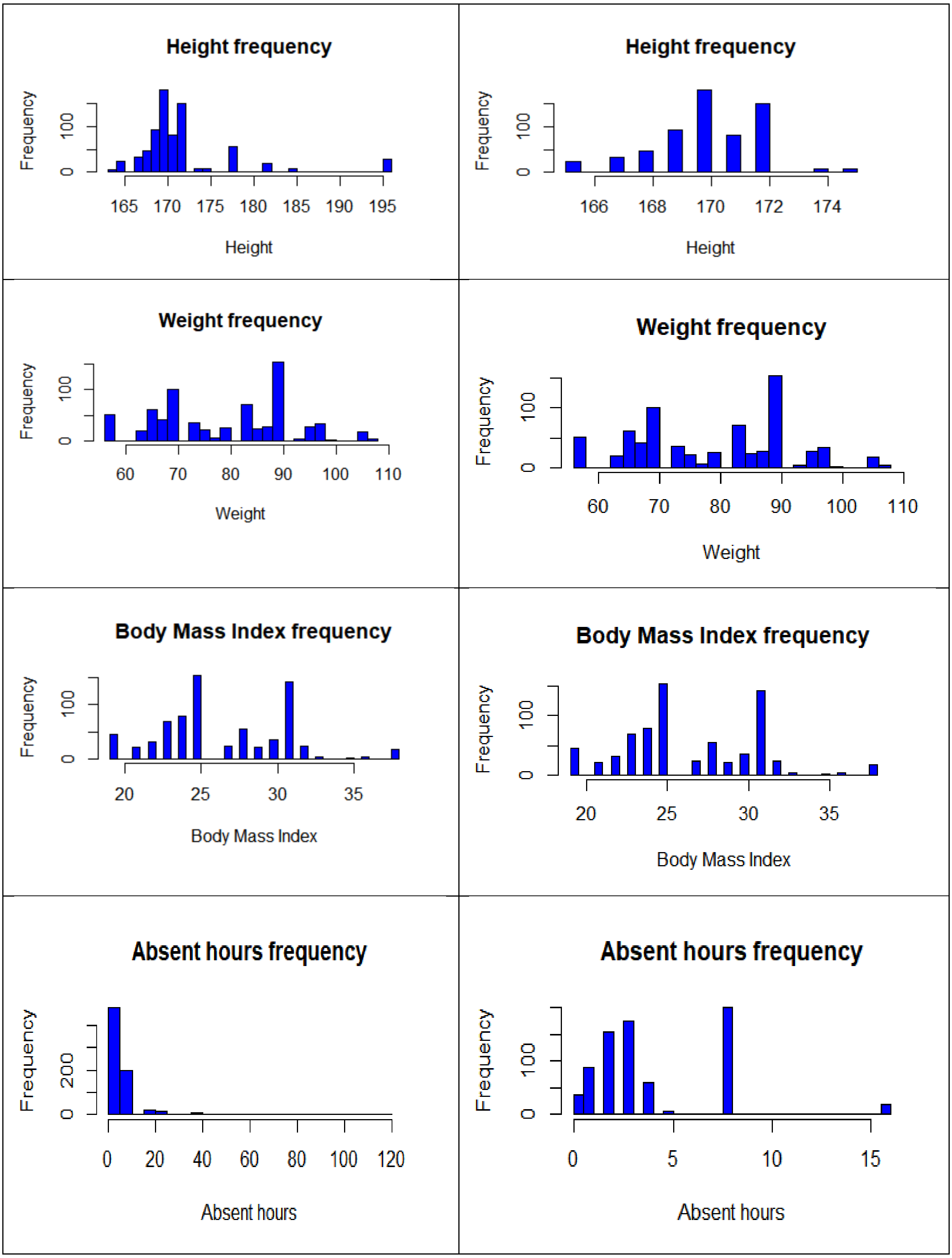
Before proceeding further with the analysis , we would like to do outlier analysis using boxplot method, which means that any data point that is less than 1.5*IQR(Inter Quartile range ) times the 25 the percentile and more than 1.5*IQR  the $75^{th}$ percentile ,is to be treated as an outlier .We replace these items with NaN in the dataset and then impute it with the median values.

Below is the histogram plot of the numerical features with and without outliers. We can see the range of the feature set has changed after imputing outliers with the median values.

Table 2.4 Frequency distribution of the numerical feature set with outliers and outliers imputed with median

| With outliers | Outliers imputed with median values |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |


Age frequency
Frequency
0 60 120
30 35 40 45 50 55
Age
Age frequency
Frequency
0 60 120
30 35 40 45 50
Age
Workload Average day frequency
Frequency
0 40 80
200000 250000 300000 350000
Workload Average day
Workload Average day frequency
Frequency
0 40 100
200000 240000 280000 320000
Workload Average day
Hit target frequency
Frequency
0 40 100
88 90 92 94 96 98 100
Hit target
Hit target frequency
Frequency
0 40 100
88 90 92 94 96 98 100
Hit target
Son frequency
Frequency
0 150
0 1 2 3 4
Son
Son frequency
Frequency
0 150
0 1 2 3 4
Son
Pet frequency
Frequency
0 200
0 2 4 6 8
Pet
Pet frequency
Frequency
0 200
0.0 0.5 1.0 1.5 2.0
Pet


Height frequency
Frequency
0
100
165 170 175 180 185 190 195
Height
Height frequency
Frequency
0
100
166 168 170 172 174
Height
Weight frequency
Frequency
0
100
60 70 80 90 100 110
Weight
Weight frequency
Frequency
0
100
60 70 80 90 100 110
Weight
Body Mass Index frequency
Frequency
0
100
20 25 30 35
Body Mass Index
Body Mass Index frequency
Frequency
0
100
20 25 30 35
Body Mass Index
Absent hours frequency
Frequency
0
200
0 20 40 60 80 100 120
Absent hours
Absent hours frequency
Frequency
0
100
0 5 10 15
Absent hours

## 2.4 Feature Selection :

One of the key task in any data science operation is to choose right set of predictors. This is because ,although more number of features implies more knowledge of our dataset but high dimension in the data set can also lead to higher variance which might fail to generalise on the test data leading to higher test MSE(Mean Square Error) .This is also known as the ***curse of dimensionality***. Apart from this , higher dimensional data in our model can also be computationally expensive. Thus we need to perform feature selection before supplying predictors to our model.

We plot correlation plot of our numerical data set :

Fig 2.4 Correlation plot of numerical dataset

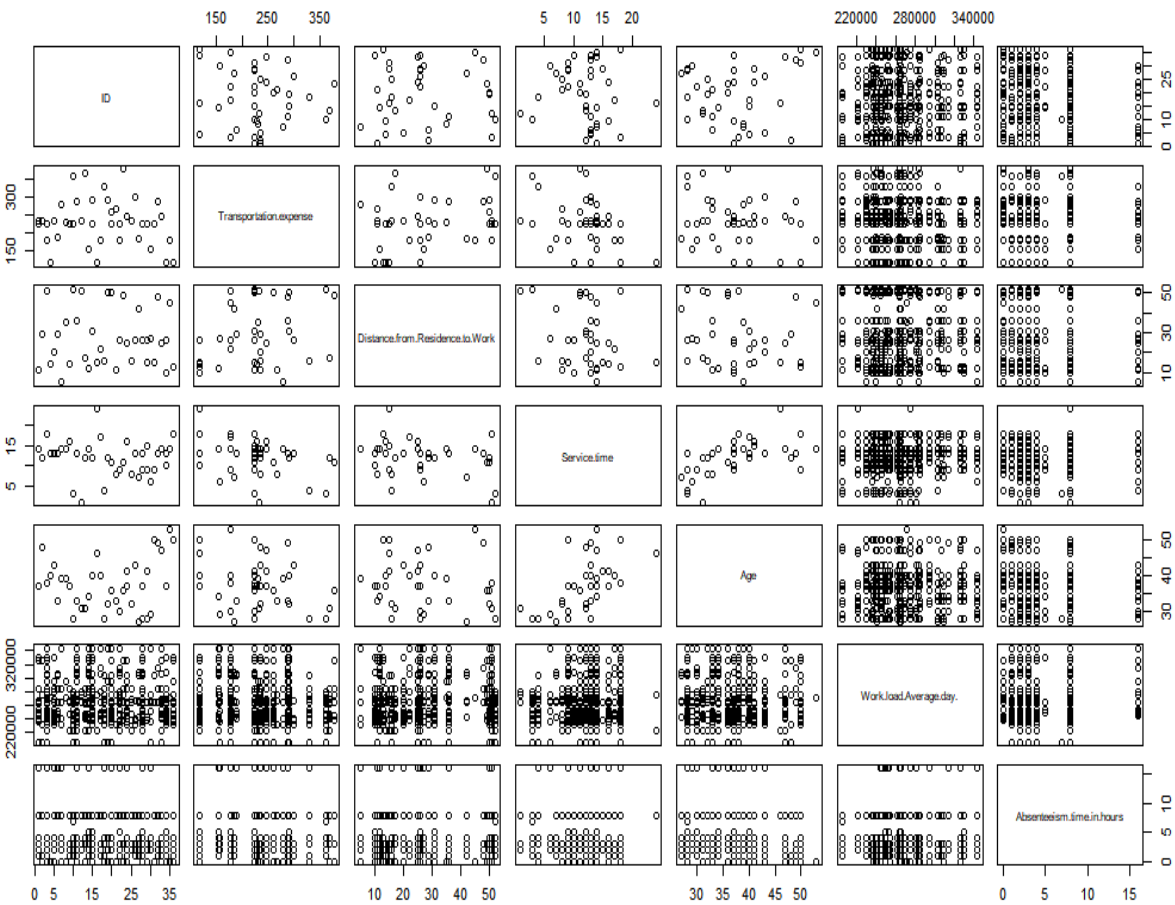
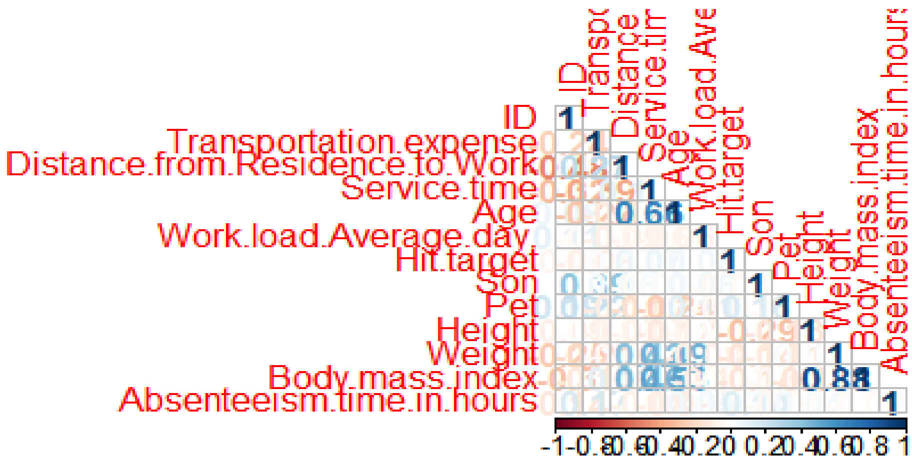



Fig 2.5 Correlation matrix of the numerical feature set




Multicollinearity can be checked through VIF(Variance Inflation Factor) values. We obtain following VIF values for our data set.

Table 2.5 VIF of numerical feature set

| S.No. | Variables | VIF |
|---|---|---|
| 1. | ID | 2.555528 |
| 2. | Transportation.expense | 2.199472 |
| 3. | Distance.from.Residence.to.Work | 1.593952 |

| | | |
|---|---|---|
| 4. | Service.time | 3.443374 |
| 5. | Age | 3.501619 |
| 6. | Work.load.Average.day. | 1.042375 |
| 7. | Hit.target | 1.024523 |
| 8. | Son | 1.532156 |
| 9. | Pet | 1.458590 |
| 10. | Height | 1.328574 |
| 11. | Weight | 6.039502 |
| 12. | Body.mass.index | 7.227522 |

Any feature set having more than 0.80 correlation will be removed .

Also, feature set with VIF > 5 ,will be removed. Thus, we remove one of the features out of weight and Body Mass Index .We chose to remove one of the variables that is Weight .

**Feature selection on categorical data set :**

As our target variable is continuous data ,we select anova test for performing feature selection on categorical data set .

Code :
```
> for(i in categorical_set){
+   print(i)
+   aov_summary = summary(aov(Absenteeism.time.in.hours~absenteeism_data[,
i],data = absenteeism_data))
+   print(aov_summary)
+ }
```

Result :

```
[1] "Reason.for.absence"
                        Df Sum Sq Mean Sq F value Pr(>F)    
absenteeism_data[, i]   26   2847  109.51   14.94 <2e-16 ***
Residuals              713   5225    7.33                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Month.of.absence"
                        Df Sum Sq Mean Sq F value Pr(>F)  
absenteeism_data[, i]   11    246   22.32   2.076 0.0199 *
Residuals              728   7827   10.75                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Day.of.the.week"
                        Df Sum Sq Mean Sq F value Pr(>F)
absenteeism_data[, i]    4     66   16.62   1.526  0.193
Residuals              735   8006   10.89               
[1] "Seasons"
                        Df Sum Sq Mean Sq F value Pr(>F)
absenteeism_data[, i]    3     38   12.72   1.166  0.322
Residuals              736   8034   10.92               
[1] "Disciplinary.failure"
```

```
                          Df Sum Sq Mean Sq F value  Pr(>F)    
absenteeism_data[, i]    1    441   441.3   42.68 1.2e-10 ***
Residuals              738   7631    10.3                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Education"
                          Df Sum Sq Mean Sq F value Pr(>F)
absenteeism_data[, i]    3     41   13.52   1.239  0.295
Residuals              736   8032   10.91               
[1] "Social.drinker"
                          Df Sum Sq Mean Sq F value Pr(>F)
absenteeism_data[, i]    1     29   29.19   2.678  0.102
Residuals              738   8043   10.90               
[1] "Social.smoker"
                          Df Sum Sq Mean Sq F value Pr(>F)
absenteeism_data[, i]    1     16   15.68   1.436  0.231
Residuals              738   8056   10.92               
```

Taking 95% as our confidence interval,we would select only those features whose p value is less than 0.05 i.e "Reason.for.absence","Month.of.absence","Disciplinary.failure"

Thus ,we reduce our overall dimension of 21 predictors to 15 predictors.

## 2.5 Modeling :

**Sampling :**

We choose ***stratified sampling*** to divide our dataset into test and train set stratified based on Reason of absence feature.We chose 75% of the data as train data and 25% data as test data

Code :

```
> train = createDataPartition(absenteeism_data$Reason.for.absence,times = 
1,p = 0.75,list = F)
> test = -(train)
```

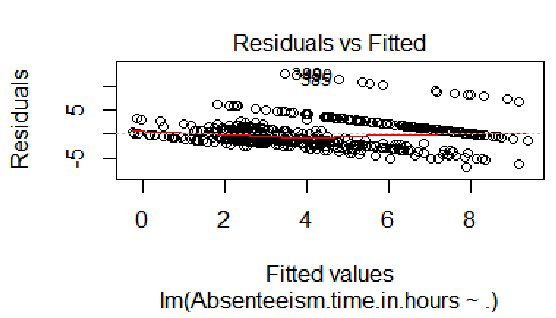
After feature selection we can now start using different regression models to predict .Let us start with the simplest model and then move towards more complex models if needed.

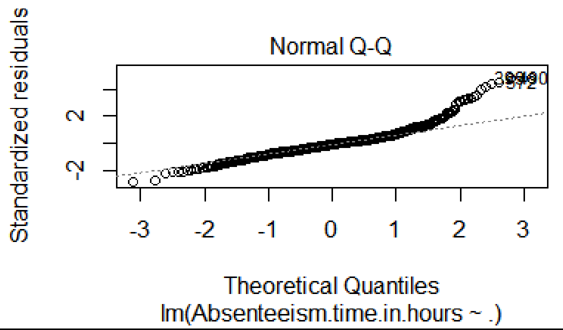
**2.5.1 Linear Regression :**

```
> modelLR = lm(Absenteeism.time.in.hours~.,data = absenteeism_data[train,]
)
> summary(modelLR)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8829 -1.4651 -0.1907  0.9266 12.5329 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      1.037e+01  1.393e+01   0.744  0.45701    
```

```
ID                              -3.172e-02  1.710e-02  -1.855  0.06415 .
Reason.for.absence2             -4.519e+00  2.806e+00  -1.610  0.10795
Reason.for.absence3             -5.461e+00  1.216e+00  -4.490 8.78e-06 ***
Reason.for.absence4             -1.550e+00  2.081e+00  -0.745  0.45666
Reason.for.absence5             -2.538e-01  1.746e+00  -0.145  0.88450
Reason.for.absence6             -2.064e+00  1.344e+00  -1.536  0.12525
Reason.for.absence7             -2.617e+00  1.098e+00  -2.384  0.01748 *
Reason.for.absence8             -1.114e+00  1.449e+00  -0.769  0.44244
Reason.for.absence9              2.499e+00  1.756e+00   1.423  0.15527
Reason.for.absence10            -1.382e+00  1.003e+00  -1.378  0.16874
Reason.for.absence11            -2.047e+00  9.876e-01  -2.072  0.03874 *
Reason.for.absence12            -1.837e+00  1.370e+00  -1.341  0.18046
Reason.for.absence13            -2.444e+00  8.929e-01  -2.737  0.00641 **
Reason.for.absence14            -2.225e+00  1.058e+00  -2.103  0.03591 *
Reason.for.absence15             8.042e-01  2.073e+00   0.388  0.69816
Reason.for.absence16            -5.272e+00  1.767e+00  -2.984  0.00298 **
Reason.for.absence17            -4.304e-01  2.804e+00  -0.153  0.87809
Reason.for.absence18            -4.685e-01  1.059e+00  -0.442  0.65838
Reason.for.absence19            -1.022e-01  9.394e-01  -0.109  0.91339
Reason.for.absence21            -7.660e-01  1.446e+00  -0.530  0.59647
Reason.for.absence22            -4.280e-02  9.550e-01  -0.045  0.96427
Reason.for.absence23            -4.336e+00  8.342e-01  -5.198 2.91e-07 ***
Reason.for.absence24             1.557e-01  1.767e+00   0.088  0.92981
Reason.for.absence25            -3.879e+00  9.650e-01  -4.019 6.70e-05 ***
Reason.for.absence26            -3.992e-01  9.639e-01  -0.414  0.67893
Reason.for.absence27            -4.924e+00  9.288e-01  -5.301 1.71e-07 ***
Reason.for.absence28            -4.212e+00  8.524e-01  -4.941 1.05e-06 ***
Month.of.absence2                7.824e-02  6.253e-01   0.125  0.90048
Month.of.absence3                3.982e-01  6.208e-01   0.641  0.52148
Month.of.absence4               -1.096e-01  7.014e-01  -0.156  0.87591
Month.of.absence5               -3.305e-01  7.239e-01  -0.457  0.64815
Month.of.absence6               -3.490e-01  7.080e-01  -0.493  0.62223
Month.of.absence7               -1.022e-01  7.035e-01  -0.145  0.88455
Month.of.absence8                1.068e-02  7.739e-01   0.014  0.98900
Month.of.absence9               -3.889e-01  7.644e-01  -0.509  0.61118
Month.of.absence10              -1.938e-01  7.271e-01  -0.267  0.78988
Month.of.absence11              -6.140e-01  6.751e-01  -0.909  0.36353
Month.of.absence12              -5.357e-01  7.053e-01  -0.760  0.44790
Transportation.expense           3.381e-03  2.651e-03   1.275  0.20284
Distance.from.Residence.to.Work -1.911e-02  9.980e-03  -1.915  0.05604 .
Service.time                     1.521e-02  5.132e-02   0.296  0.76710
Age                             -4.787e-02  3.660e-02  -1.308  0.19150
Work.load.Average.day.           4.113e-06  4.568e-06   0.900  0.36835
Hit.target                      -4.987e-03  5.373e-02  -0.093  0.92607
Disciplinary.failure1           -7.619e-01  9.230e-01  -0.825  0.40950
Son                              2.862e-01  1.374e-01   2.083  0.03770 *
Pet                             -2.813e-01  2.024e-01  -1.390  0.16512
Height                          -1.675e-02  7.513e-02  -0.223  0.82371
Body.mass.index                  4.036e-02  3.999e-02   1.009  0.31336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 516 degrees of freedom
Multiple R-squared:  0.4235,   Adjusted R-squared:  0.3687
F-statistic: 7.735 on 49 and 516 DF,  p-value: < 2.2e-16


> plot(modelLR)
```

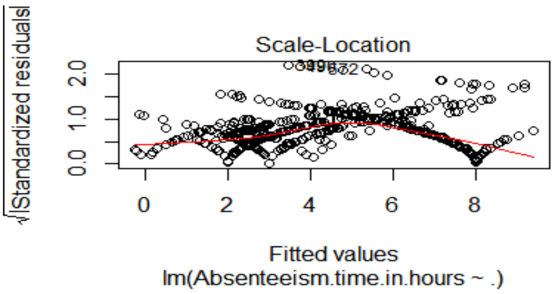
Fig 2.6 Linear Regression model summary

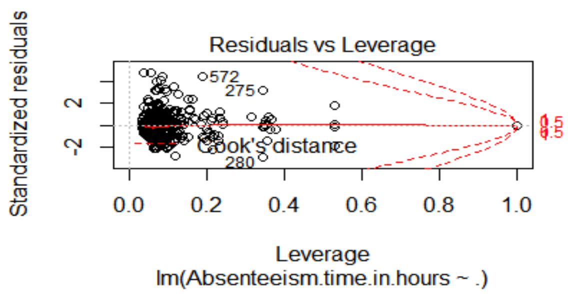



The above plot shows that the scatter plot between residual errors and fitted values(predicted values).We can observe that the above plot is not non linear and free from heteroskedacity.Thus non linear transformations are not required in our linear model

The Q-Q or the quantile quantile plot is the scatter plot .The above plot shows the presence of normality as the points almost passes through the straight line diagonal.However,some deviation can be seen only at the extreme end.




Scale Location plot is similar to the residual plot.The only difference is ,it uses square root of standardised residual errors.As no particular pattern can be seen in the plot,we can conclude that there is no heteroskedacity i.e.variance is equal

The above plot is also known as Cook's distance plot.It is used to identify if some predictors' point influence our prediction errors more than the others.The dotted red line shows the Cooks distance and any point beyond these points can be considered as high leverage points(extreme in X).Here we do not have any such points

From the above summary of linear model ,we observe that the only following features have higher importance :

{Reason.for.absence,Distance.from.Residence.to.Work,Son.}

Thus ,we try to remodel our linear regression model ,with only these 3 predictors.

```
> modelLR = lm(Absenteeism.time.in.hours~Reason.for.absence+Distance.from.
Residence.to.Work+Son,data = absenteeism_data[train,])
> summary(modelLR)


Call:
lm(formula = Absenteeism.time.in.hours ~ Reason.for.absence + 
    Distance.from.Residence.to.Work + Son, data = absenteeism_data[train, 
    ])

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7739 -1.5735 -0.3402  0.8554 13.6364 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                     7.068196   0.791186   8.934  < 2e-16 ***
Reason.for.absence2            -4.126588   2.770605  -1.489  0.13696    
Reason.for.absence3            -6.447535   0.888364  -7.258 1.39e-12 ***
Reason.for.absence4            -2.701403   2.038310  -1.325  0.18563    
Reason.for.absence5            -0.933981   1.718411  -0.544  0.58700    
Reason.for.absence6            -2.040058   1.330950  -1.533  0.12592    
Reason.for.absence7            -2.644877   1.086255  -2.435  0.01522 *  
Reason.for.absence8            -1.540481   1.416209  -1.088  0.27719    
Reason.for.absence9             1.858814   1.719776   1.081  0.28025    
Reason.for.absence10           -1.531283   0.980981  -1.561  0.11912    
Reason.for.absence11           -2.110977   0.975538  -2.164  0.03091 *  
Reason.for.absence12           -2.113442   1.331757  -1.587  0.11311    
Reason.for.absence13           -2.568175   0.873992  -2.938  0.00344 ** 
Reason.for.absence14           -2.677975   1.031242  -2.597  0.00967 ** 
Reason.for.absence15           -0.188753   2.050994  -0.092  0.92671    
Reason.for.absence16           -5.577575   1.718083  -3.246  0.00124 ** 
Reason.for.absence17            0.196931   2.771993   0.071  0.94339    
Reason.for.absence18           -0.521323   1.018265  -0.512  0.60888    
Reason.for.absence19           -0.346843   0.917041  -0.378  0.70542    
Reason.for.absence21           -0.910083   1.418687  -0.641  0.52147    
Reason.for.absence22            0.035584   0.929715   0.038  0.96948    
Reason.for.absence23           -4.762998   0.810217  -5.879 7.27e-09 ***
Reason.for.absence24            0.145838   1.726371   0.084  0.93271    
Reason.for.absence25           -3.990484   0.941357  -4.239 2.64e-05 ***
Reason.for.absence26           -0.320812   0.939538  -0.341  0.73289    
Reason.for.absence27           -4.914162   0.869337  -5.653 2.57e-08 ***
Reason.for.absence28           -4.635629   0.828857  -5.593 3.56e-08 ***
Distance.from.Residence.to.Work  0.003650   0.008155   0.448  0.65469    
Son                             0.327292   0.115303   2.839  0.00470 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.66 on 537 degrees of freedom
Multiple R-squared:  0.3952,	Adjusted R-squared:  0.3637 
F-statistic: 12.53 on 28 and 537 DF,  p-value: < 2.2e-16
```

After remodelling ,we do not see any significant difference in Residual Standard Errors and Adjusted R squared .

Let us now move to other models .

### 2.5.2 Decision Trees:

Code :

```
> DTmodel = tree(Absenteeism.time.in.hours~.,absenteeism_data,subset = tra
in)
> summary(DTmodel)
>plot(modelDT)
>text(modelDT,pretty = 0)
```

Result :

```
Regression tree:
tree(formula = Absenteeism.time.in.hours ~ ., data = absenteeism_data, 
    subset = train)
Variables actually used in tree construction:
```

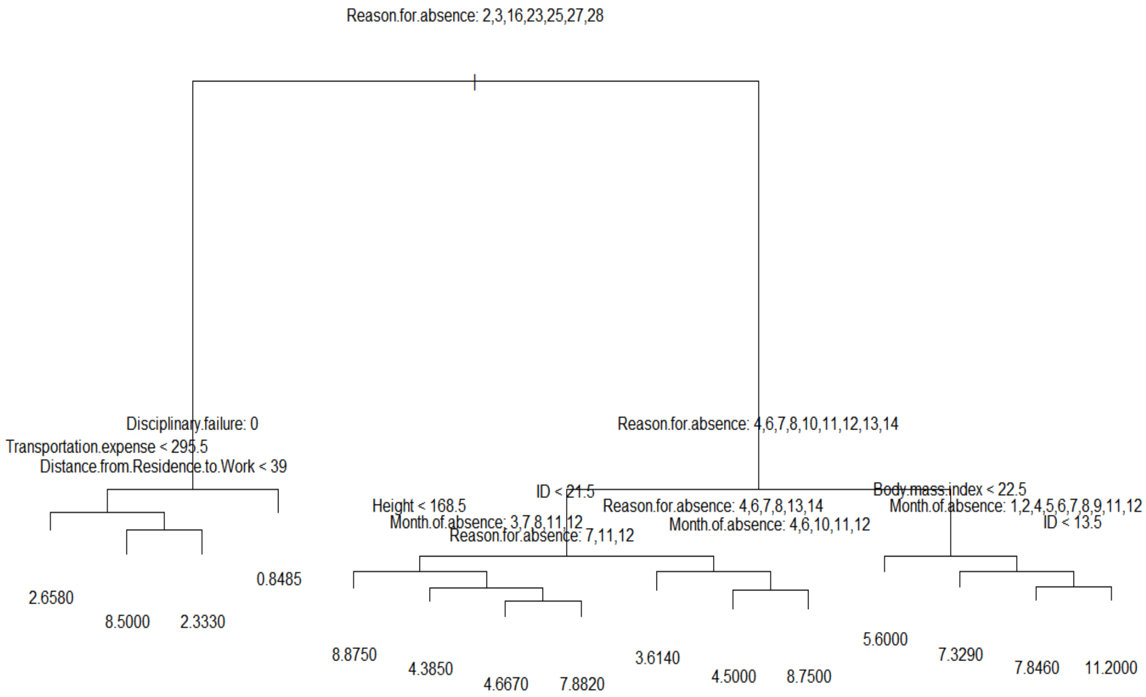
```
[1] "Reason.for.absence"              "Disciplinary.failure"          "T
ransportation.expense"
[4] "Distance.from.Residence.to.Work" "ID" "Height"
[7] "Month.of.absence"                "Body.mass.index"
Number of terminal nodes:  15
Residual mean deviance:  5.489 = 3024 / 551
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-6.3290 -0.8750 -0.6136  0.0000  0.6706 13.3400
```

We see the in the above summary that decision trees has chosen only 8 variables for prediction .They are :

```
{"Reason.for.absence","Disciplinary.failure","Transportation.expense"
,"Distance.from.Residence.to.Work","ID","Height","Month.of.absence",
"Body.mass.index"}
```

Residual mean deviance is equivalent to sum of squared errors for the tree which is around **5.849.**

Fig . 2.7 Decision tree structure




## 2.5.3 Random Forest :

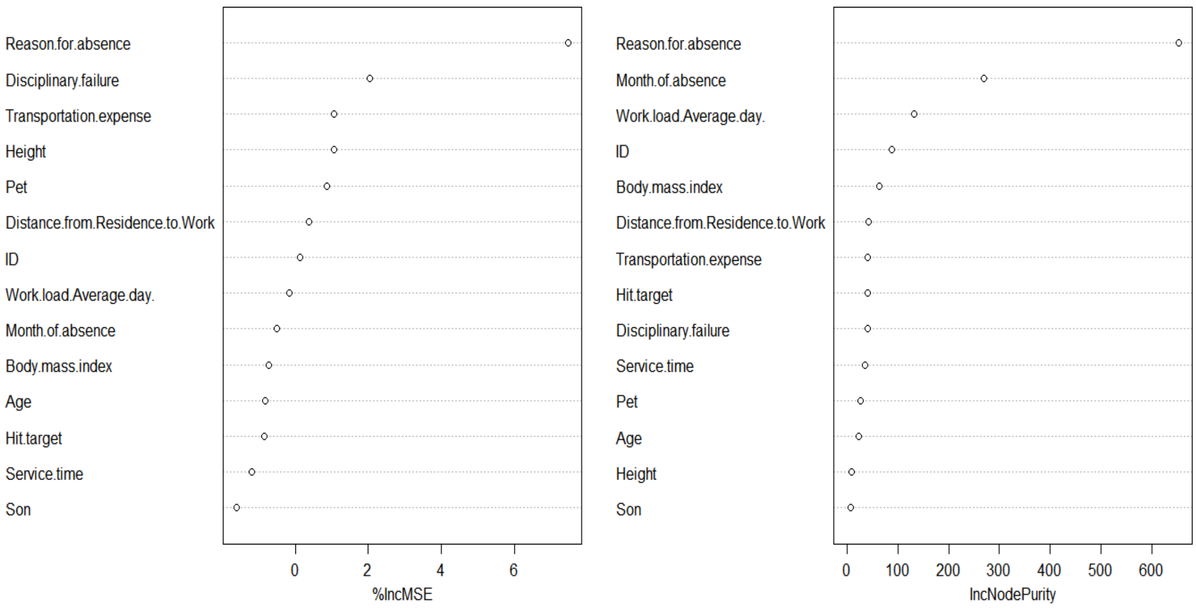
Code :

```
> RFmodel = randomForest(Absenteeism.time.in.hours~.,data = absenteeism_da
ta,subset = test,mtry = 10,ntree=10,importance = TRUE)
> varImpPlot(RFmodel)
> importance(RFmodel)
```

Result :

```
                                %IncMSE  IncNodePurity
ID                            0.1373691      87.602827
Reason.for.absence            7.4864674     653.928852
Month.of.absence             -0.5060437     268.936118
Transportation.expense        1.0610161      40.782444
Distance.from.Residence.to.Work 0.3666920    41.562755
Service.time                 -1.1934954      35.643946
Age                          -0.8264176      21.957876
Work.load.Average.day.       -0.1515439     132.028946
Hit.target                   -0.8510682      40.161114
Disciplinary.failure          2.0522754      39.748351
Son                          -1.6033152       6.762360
Pet                           0.8673430      25.444008
Height                        1.0603624       9.048111
Body.mass.index              -0.7150441      62.926948
```

Fig 2.8 Figure showing importance of variable on the basis of %inc MSE(left) and Node Purity (Right)




From Random Forest ,we see that when we chose 10 predictors ( mtry = 10) we get the above results. Thus we can select top 10 features from our %incMSE table and drop all others. Thus, also reducing feature set from 21 to finally 10 features. The first figure implies the mean decrease of accuracy if that particular feature is excluded from the model. The right hand side plot of IncNodePurity , implies the total decrease in node impurity that results from splits over that variable.

# Chapter 3

# Conclusion

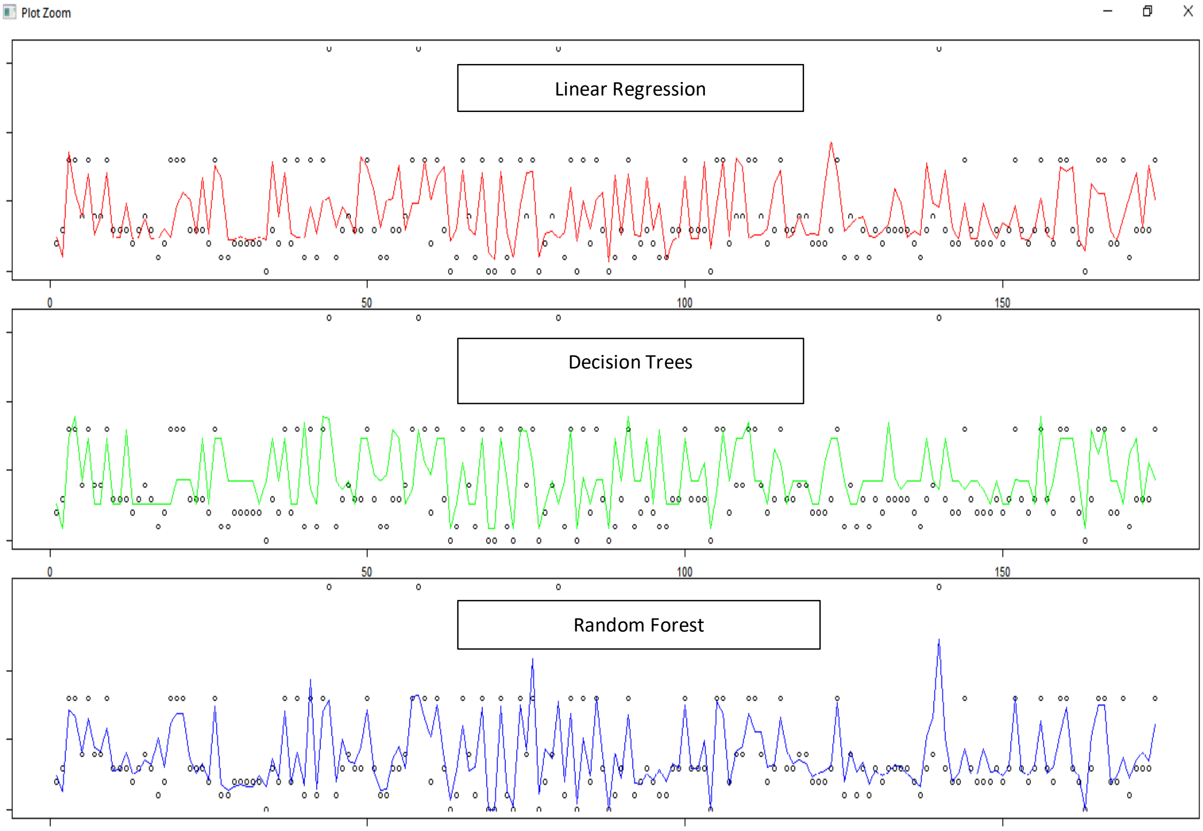
## 3.1 Model Evaluation

In the earlier section ,we had used following algorithms for modelling our dataset

i. Linear Regression
ii. Decision Trees
iii. Random Forest

We now compare the result by plotting the actual vs predicted plot for the test set for all the three models used.

Fig 2.9 Actual vs Predicted scatter plot. Actual values are shown as dots and predicted values are plotted as line




<u>RMSE for Linear Regression :</u>

```
> sqrt(mean((predictLR-absenteeism_data$Absenteeism.time.in.hours[test])^2
))
[1] 2.65533
```

<u>RMSE for Decision Trees :</u>

```
> sqrt(mean((predictDT-absenteeism_data$Absenteeism.time.in.hours[test])^2
))
[1] 2.924705
```

RMSE for Random Forest (parameters mtry = 10,ntree = 10)

```
> sqrt(mean((predictRF-absenteeism_data$Absenteeism.time.in.hours[test])^2
))
[1] 1.564675
```

Table 2.6 RMSE of Linear Model,Decision Trees and Random Forest

| Model | RMSE(in R ) | RMSE(in Python) |
|---|---|---|
| Linear Model | 2.94 | 3.06 |
| Decision Trees | 3.06 | 3.60 |
| Random Forest (mtry = 10,ntree = 10) | 1.45 | 3.04 |

Thus ,we can select Random Forest as the best model out of these 3 models.

## 3.2 Model Inference

Now that we have analysed our data set and selected or predictive model.We can see that the most important predictor in our absenteeism prediction is Reason for absence (fig no. 2.8). From the barplot and boxplot chart (fig no.2.3) of Reason for absence we see maximum frequency of absent reason is reason no. 23(medical consultation) and 28 (dental consultation).Thus , it is evident that maximum absenteeism is due to health related reasons.

On summarising the Count, Sum of Absenteeism hours and Mean of absenteeism hours Reason wise , we see that medical consultation(Reason no. 23) and dental consultation(Reason no. 28) is common cause of absenteeism .Hence the company can arrange for free regular medical consultation and den tal consultation in coalition with some hospitals and other promotion camps in its office.

Table 2.7 Table summarising frequency of Absent Reason and Sum and Mean Values of Absenteeism hours for each reasons

| S.No. | Reason No. | Frequency of Reason | Sum of Absent Hours | Mean of Absent Hours |
|---|---|---|---|---|
| 1 | **1** | **15** | **113** | **7.5333333** |
| 2 | 2 | 1 | 3 | 3 |
| 3 | 3 | 48 | 47 | 0.9791667 |
| 4 | 4 | 2 | 9 | 4.5 |
| 5 | 5 | 3 | 19 | 6.3333333 |
| 6 | 6 | 8 | 49 | 6.125 |
| 7 | 7 | 15 | 71 | 4.7333333 |
| 8 | 8 | 6 | 32 | 5.3333333 |
| 9 | 9 | 4 | 30 | 7.5 |
| 10 | 10 | 25 | 150 | 6 |
| 11 | 11 | 26 | 143 | 5.5 |
| 12 | 12 | 8 | 36 | 4.5 |

| | | | | |
|---|---|---|---|---|
| 13 | 13 | 55 | 305 | 5.5454545 |
| 14 | 14 | 19 | 96 | 5.0526316 |
| 15 | 15 | 2 | 16 | 8 |
| 16 | 16 | 3 | 6 | 2 |
| 17 | 17 | 1 | 8 | 8 |
| 18 | 18 | 21 | 140 | 6.6666667 |
| 19 | 19 | 40 | 263 | 6.575 |
| 20 | 21 | 6 | 35 | 5.8333333 |
| 21 | 22 | 37 | 265 | 7.1621622 |
| **22** | **23** | **149** | **426** | **2.8590604** |
| 23 | 24 | 3 | 24 | 8 |
| 24 | 25 | 31 | 108 | 3.483871 |
| 25 | 26 | 33 | 235 | 7.1212121 |
| 26 | 27 | 69 | 157 | 2.2753623 |
| **27** | **28** | **110** | **310** | **2.8181818** |

Our second problem at hand is to predict monthly work loss for the company is the same trend continues.We calculate work loss as :

Work Loss = (Work.load.Average.day/Service.time)*Absenteeism.time.in.hours

Work loss for 2011 ,considering the same trend in the absenteeism pattern is :
Code:

```
> loss_data$WorkLoss = round((loss_data$Work.load.Average.day./loss_data$S
ervice.time)*loss_data$Absenteeism.time.in.hours)
> View(loss_data)
> monthly_loss = aggregate(loss_data$WorkLoss,by = list(Category = loss_da
ta$Month.of.absence),FUN = sum)
> names(monthly_loss) = c("Month","WorkLoss")
> monthly_loss
```

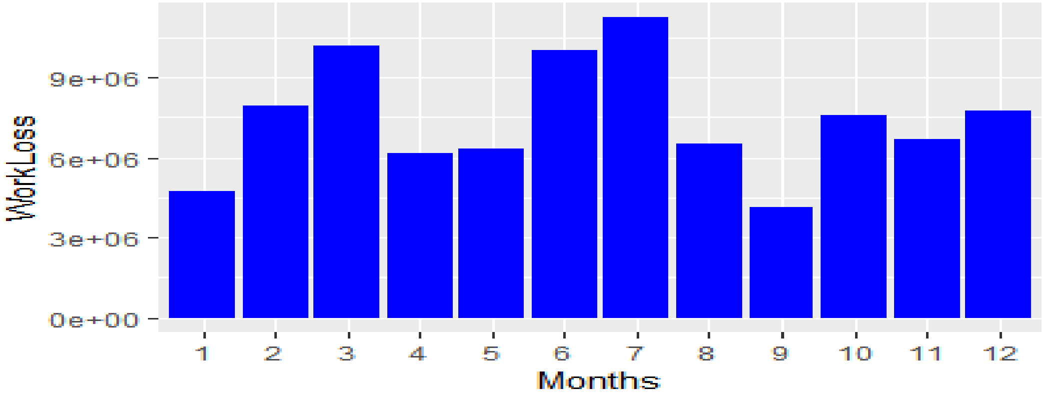
Result :

```
   Month WorkLoss
1      1  4730333
2      2  7938031
3      3 10195439
4      4  6140416
5      5  6341450
6      6 10033176
7      7 11256879
8      8  6520187
9      9  4159294
10    10  7598176
11    11  6674416
12    12  7742547
```

Code :

```
> ggplot(monthly_loss,aes(monthly_loss$Month,monthly_loss$WorkLoss))+geom_
bar(stat = "identity",fill = "blue")+labs(y="WorkLoss",x="Months")
```

Fig 2.8 Barplot of Monthly work loss



## References :


i.) https://edwisor.com/

ii.) https://www.analyticsvidhya.com

iii.) https://towardsdatascience.com/


*Note: Figures and References are made from R code outputs*