Employee Absenteeism

Riya Mittal

07 October 2018

Contents

Chapter 1

Introduction

1.1 Problem Statement

Productivity of an organization depends upon the productivity of its employees. However, it gets affected if the employee's absenteeism rate increases. This project aims at finding the prime causes of absenteeism in employees. It will help the organization reduce the number of absentees by taking appropriate measures. Also this project aims at determining the future trends of this issue if the same conditions persist.

1.2 Data

We would build a regression model here which will predict the absenteeism time in hours (target variable) for an employee based on multiple factors. Below is the sample of the dataset being used for this purpose:

Table 1.1: Sample Data (Columns: 1-6)

ID	***Reason for absence***	***Month of absence***	***Day of the week***	***Seasons***	***Transportation expense***
11	*26*	*7*	*3*	*1*	*289*
36	*0*	*7*	*3*	*1*	*118*
3	*23*	*7*	*4*	*1*	*179*
7	*7*	*7*	*5*	*1*	*279*
11	*23*	*7*	*5*	*1*	*289*

Table 1.2: Sample Data (Columns: 7-12)

Distance from Residence to Work	***Service time***	***Age***	***Work load Average/day***	***Hit target***	***Disciplinary failure***
36	*13*	*33*	*239554*	*97*	*0*
13	*18*	*50*	*239554*	*97*	*1*
51	*18*	*38*	*239554*	*97*	*0*
5	*14*	*39*	*239554*	*97*	*0*
36	*13*	*33*	*239554*	*97*	*0*

Table 1.3: Sample Data (Columns: 13-21)

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

This entire prediction model will be based on 740 X 21 dataset. Following is the bifurcation for predictor and target vaiables:

- ***Predictor Variables***: 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.

- ***Target Variable :*** Absenteeism time in hours

Also below is the list of categorical variables and continuous variables amongst them:

- ***Categorical Variables***: ID, Reason for absence, Month of absence, Day of the week, Seasons, Disciplinary failure, Education, Social drinker, Social smoker
- ***Continuous Variables:*** Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Hit target, Son, Pet, Weight, Height, Body Mass Index, Absenteeism time in hours

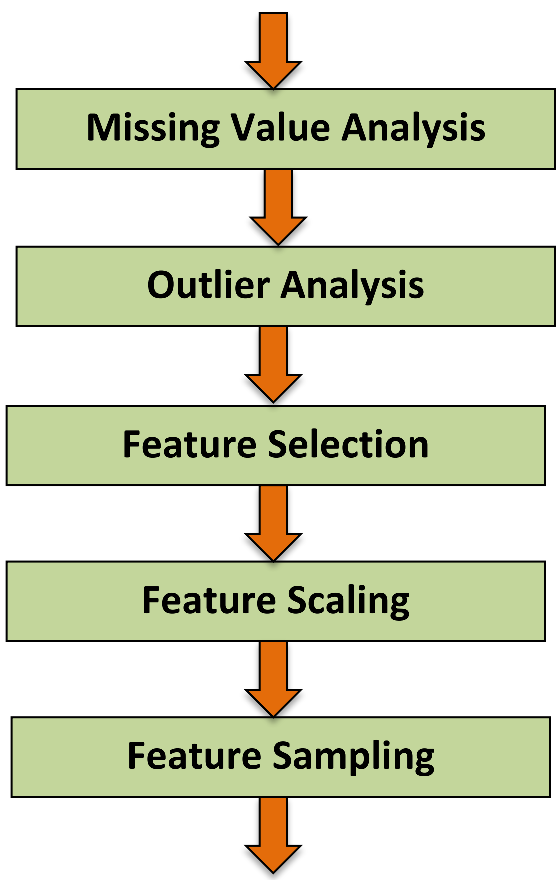
Chapter 2

Methodology

2.1 Pre Processing

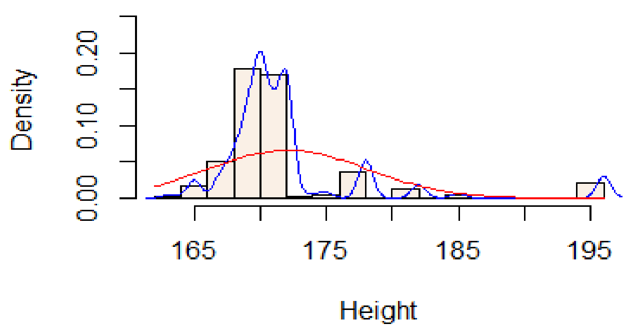
Pre-processing is a technique through which we make the data fit to be applied to any algorithm. Raw data undergoes a number of transformation before we feed it into an actual model. A data scientist roughly spends 80% of his time in pre-processing. It gives us an idea about how important this process is. Basic pre-processing steps involved before every model implementation is as shown in Figure 2.1 below:

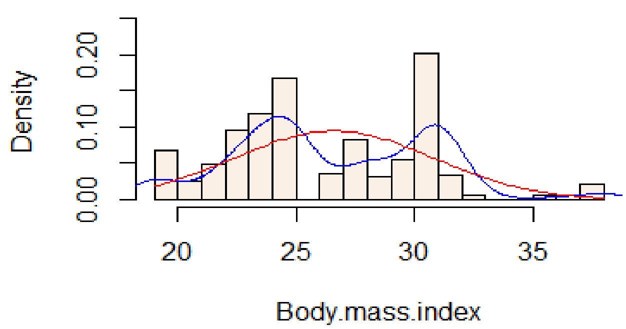
Figure 2.1: Pre-processing Steps

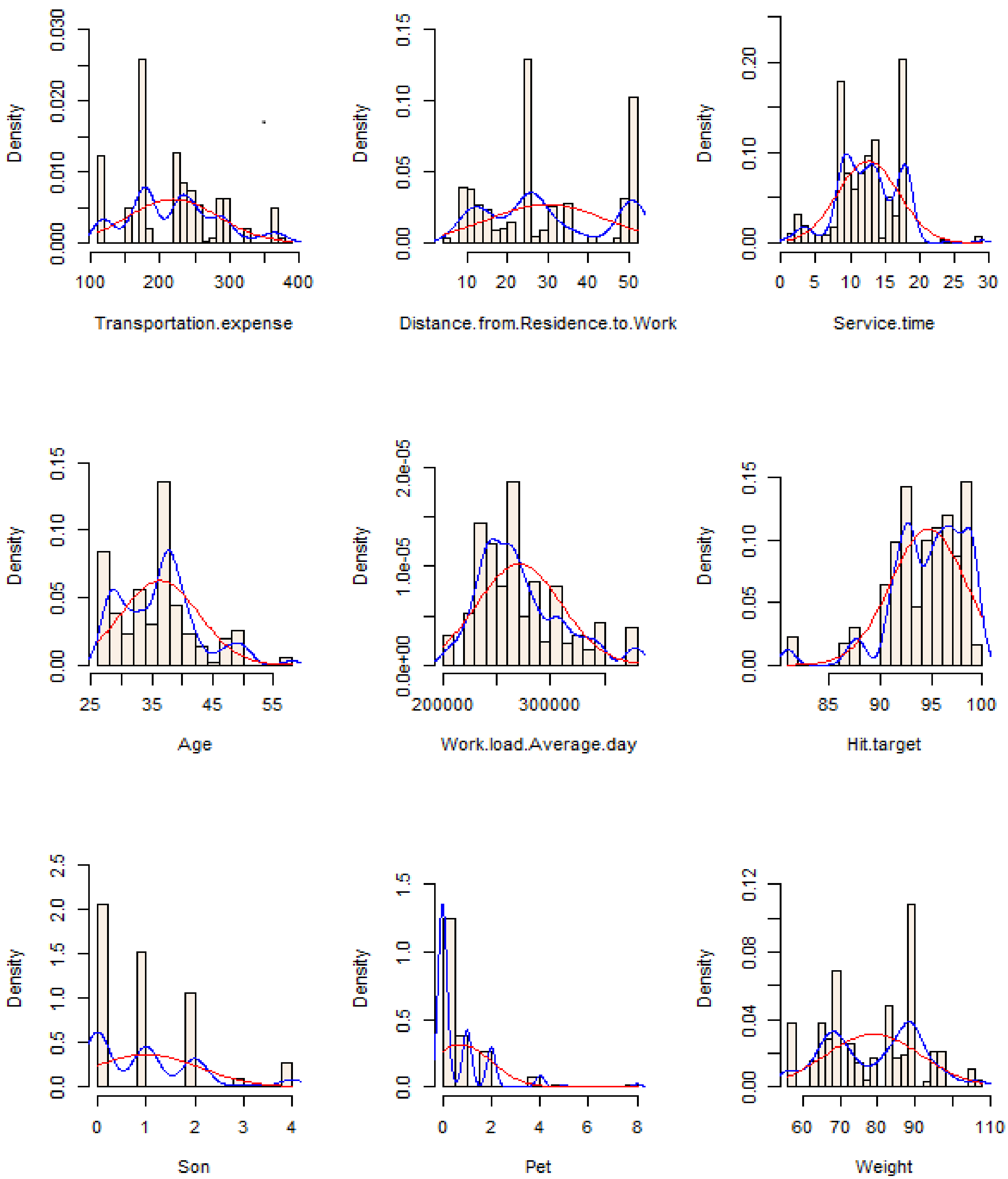


The Employee Absenteeism is a regression problem. This means we would have to predict a quantity. In this case the quantity is Absenteeism Time in Hours. Most of the regression problems analysis need normally distributed data. So we would have to look upon probability distributions or probability density functions of the continuous variables as shown in Figure 2.2. The blue lines indicate Kernel Density Estimations (KDE) of the variable. The red lines represent the normal distribution. The data is not normally distributed. This needs to be worked upon.

Figure 2.2: Probability Density Function of Employee Absenteeism Data






Density
Transportation.expense
Distance.from.Residence.to.Work
Service.time
Age
Work.load.Average.day
Hit.target
Son
Pet
Weight

2.1.1 Missing Value Analysis

Missing values are the most common problem encountered while dealing with the real world data. There could be multiple reasons for this like human error, skipped entry etc. To handle the missing values, we have a number of ways :

- Deleting the rows or columns
- Replacing with mean/median/mode
- Imputation techniques such as KNN-Imputation

In our case, we are going to delete the rows with null or zero value of target variable as it won't be of our use. We are interested in finding out why are the employees not showing up in office. For rest of the values we are dealing with KNN Imputation method. Table 2.1 describes the missing percentage of each variable.

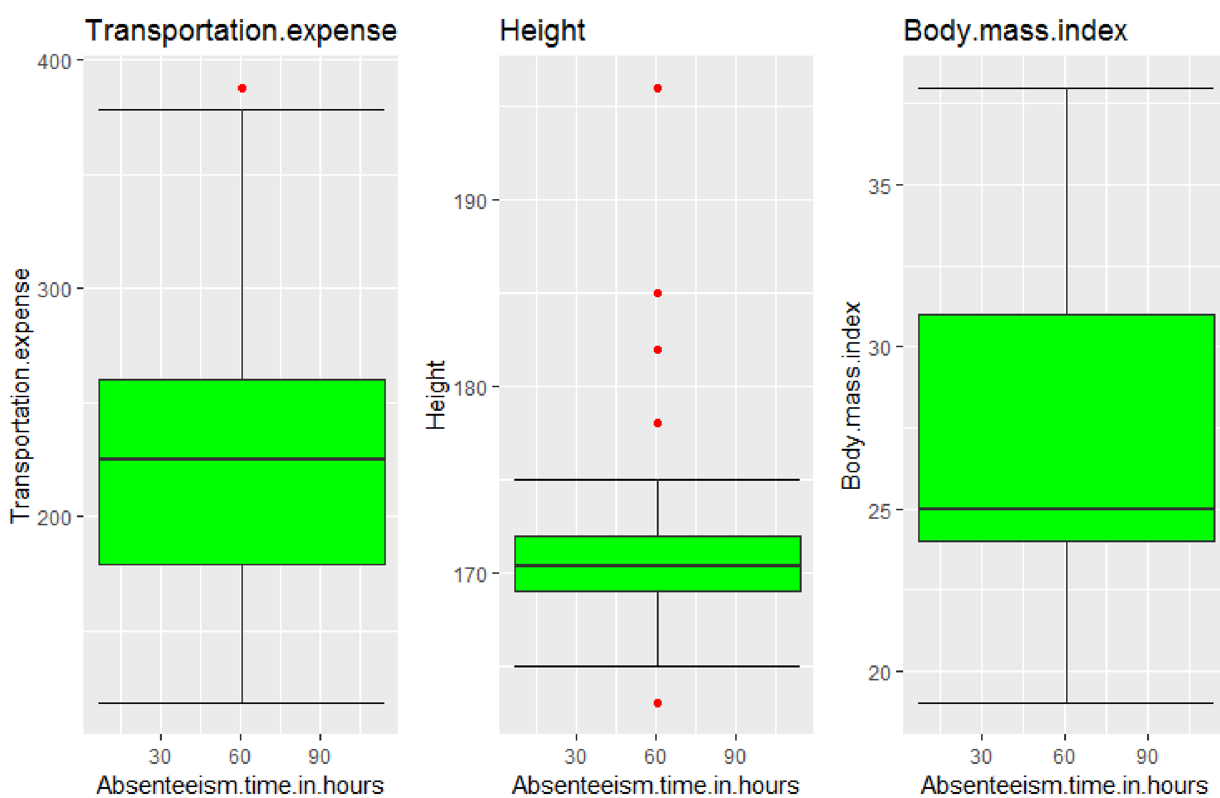
Table 2.1 Missing Percentage of each variable

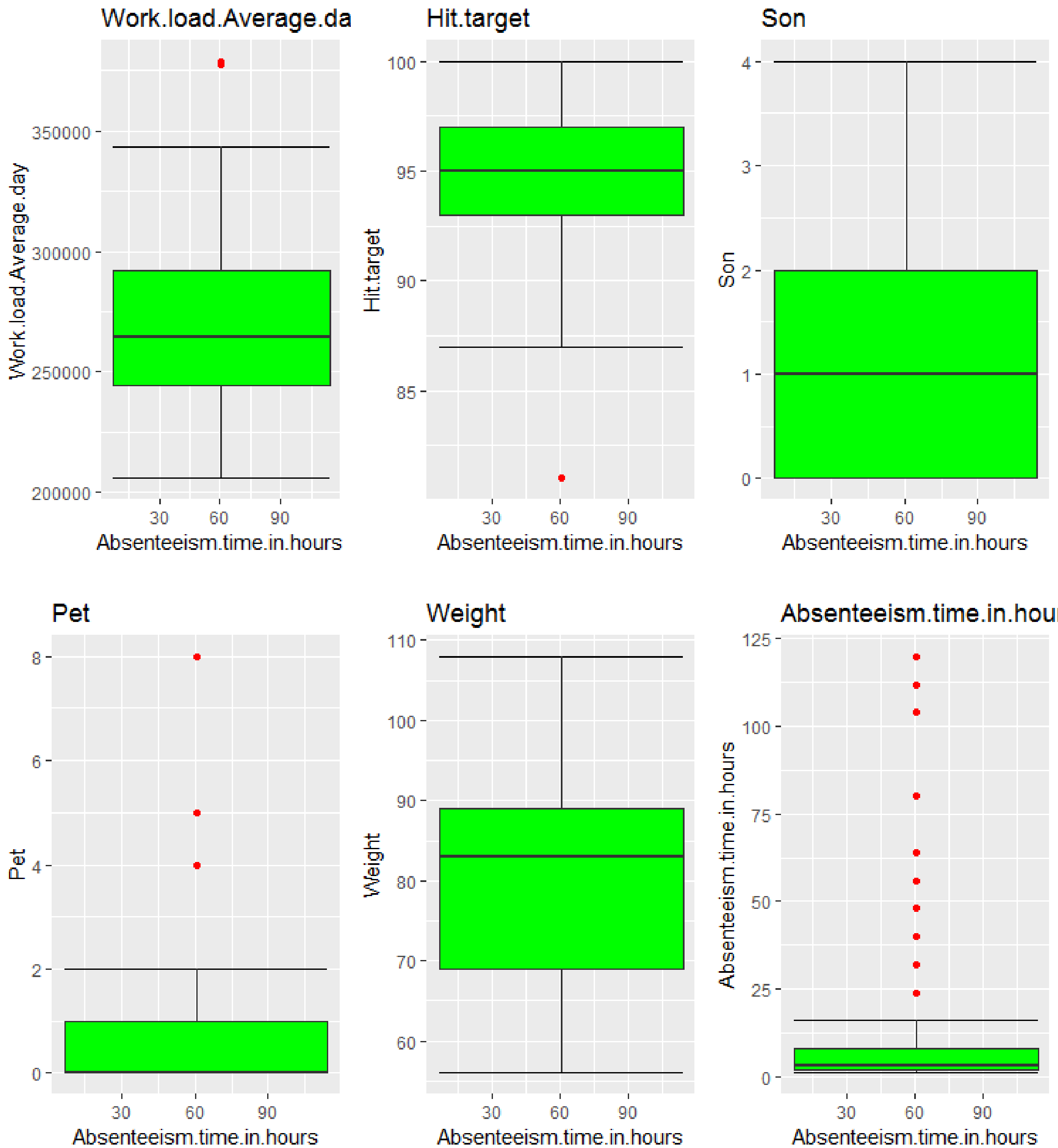
Columns	***Missing Percentage***
Body.mass.index	*4.189189189*
Absenteeism.time.in.hours	*2.972972973*
Height	*1.891891892*
Work.load.Average.day	*1.351351351*
Education	*1.351351351*
Transportation.expense	*0.945945946*
Hit.target	*0.810810811*
Disciplinary.failure	*0.810810811*
Son	*0.810810811*
Social.smoker	*0.540540541*
Reason.for.absence	*0.405405405*
Distance.from.Residence.to.Work	*0.405405405*
Service.time	*0.405405405*
Age	*0.405405405*
Social.drinker	*0.405405405*
Pet	*0.27027027*
Month.of.absence	*0.135135135*
Weight	*0.135135135*
ID/ Day.of.the.week/ Seasons	*0*

2.1.2 Outlier Analysis

Often we come across another common issue in the process of data exploration called Outliers. Outliers are observations which stand out from a normal range of a particular variable. It is of utter importance to analyze the reason of their deviation as they could many times lead to false predictions. It's sometimes observed that we may have valid and possible outliers. Outliers generally can be observed using Boxplots. Here we are performing multivariate analysis plotting each continuous variable against target variable as shown in Figure 2.3.

Figure 2.3 Boxplots for each Predictor





We can see a number of outliers in different predictors as well as target variables. But they seem to be valid outliers so we will proceed with two approaches i.e. with and without outliers.

2.1.3 Feature Dimension Reduction

Feature Selection is one of the most important step which decides the quality of our model. For this problem, our aim is to find out why employees are getting absent. This means we need to find out the prime features affecting the target variable. Two approaches have been used for features dimension reduction i.e. feature selection which involves correlation analysis and Principal Component Analysis. Let's first look upon correlation analysis.

2.1.3.1 Correlation Analysis

As we are interested in variable importance, correlation analysis can help a lot in reducing the dimension of features leaving us with the most important ones. As we have continuous target variable, we can go with heatmap to find out correlation between variables. Figure 2.4 shows the heatmap obtained.

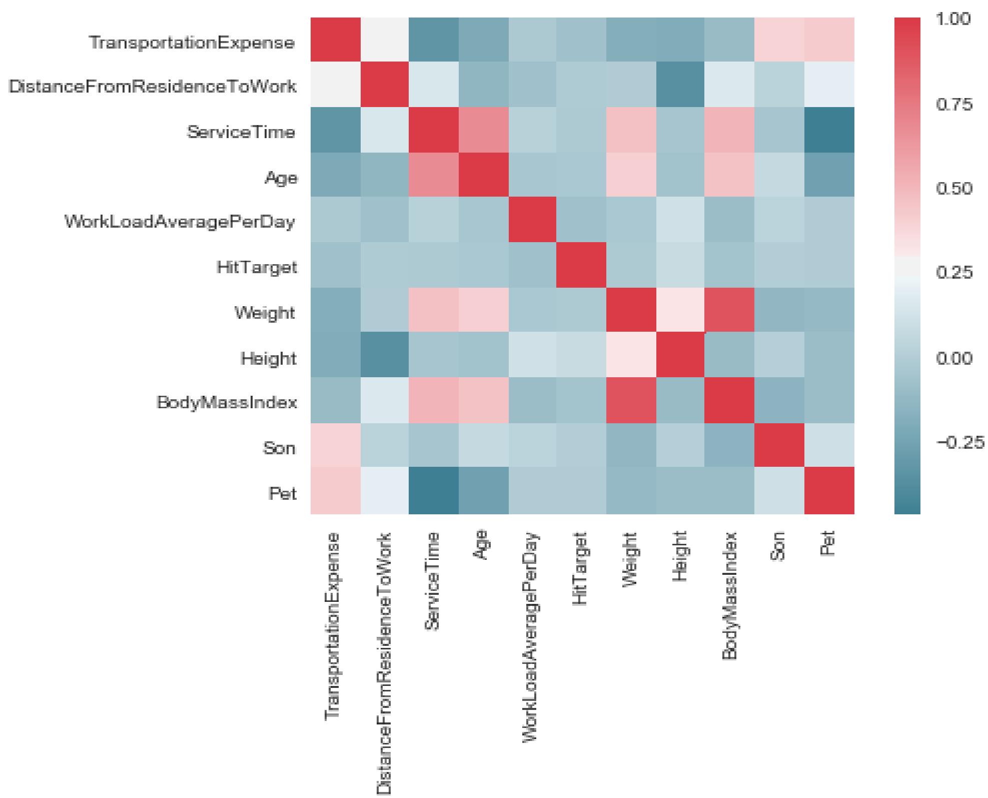


Figure 2.4 HeatMap between variables with outliers

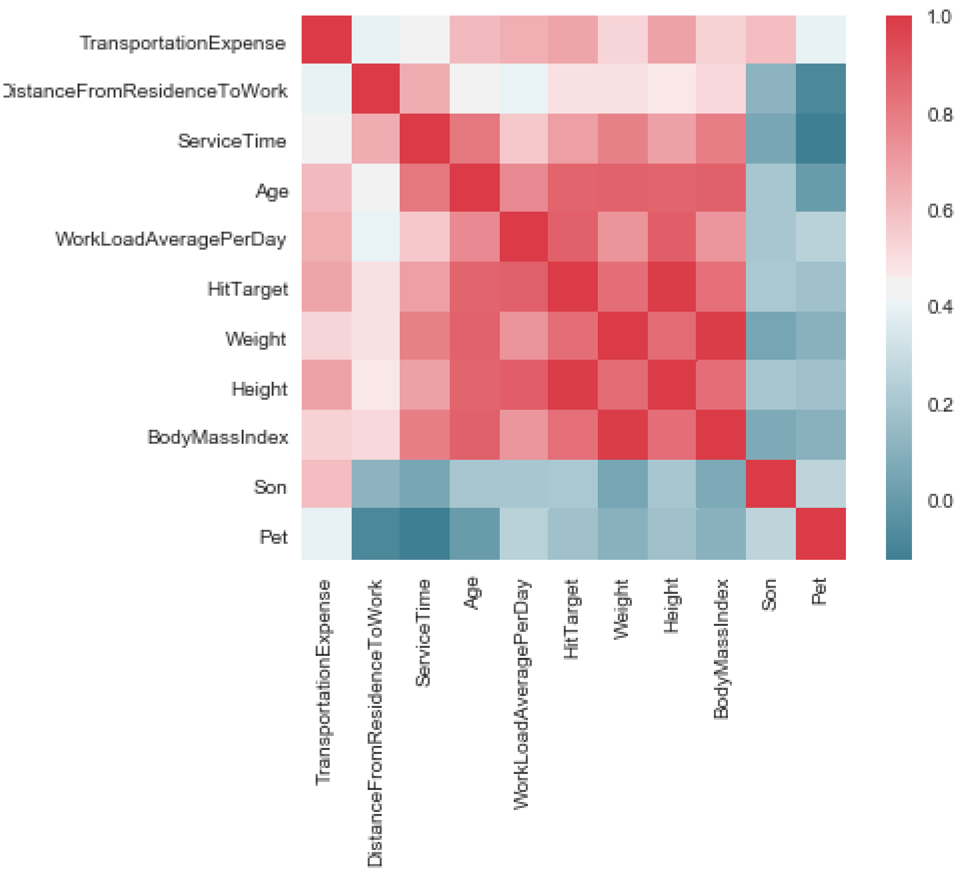


Figure 2.5 HeatMap between variables without outliers

Heatmap with outliers does not show any significant correlation amongst variables. But the other one without outliers shows many dark red parts indicating correlation.

We remove the variables Age, Weight and Height with high correlation for analysis without outliers from data for further analysis.

2.1.3.2 Principal Component Analysis(Calculated in Python)

Principal Component Analysis (PCA) is a dimensionality reduction technique. It works by combining the correlated features and creating the new which are decorrelated amongst each other. PCA assumes the data to be normally distributed. Hence, we need to apply standardization process before feeding the data to PCA.

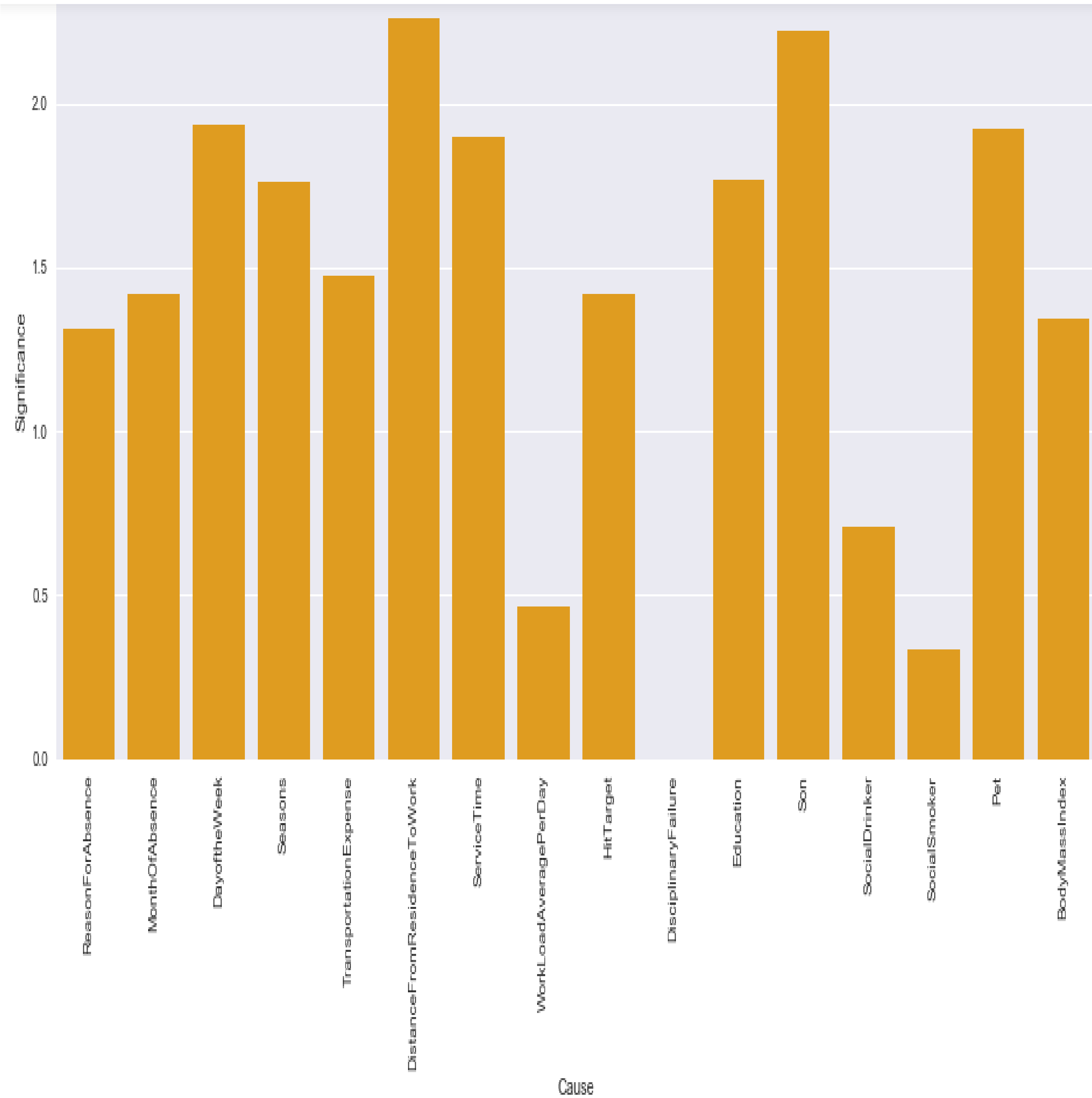


Figure 2.6 Most Significant Features Obtained through PCA(without outliers)

Figure 2.6 gives us very important insight about the factors responsible for Employee Absenteeism. It shows the net effect of old features involved in the new features space. We have now reduced features from 21 to 10. Our model can be built with these 10 features only. It shows Distance from Residence to work and Son are the most significant features in the new set.

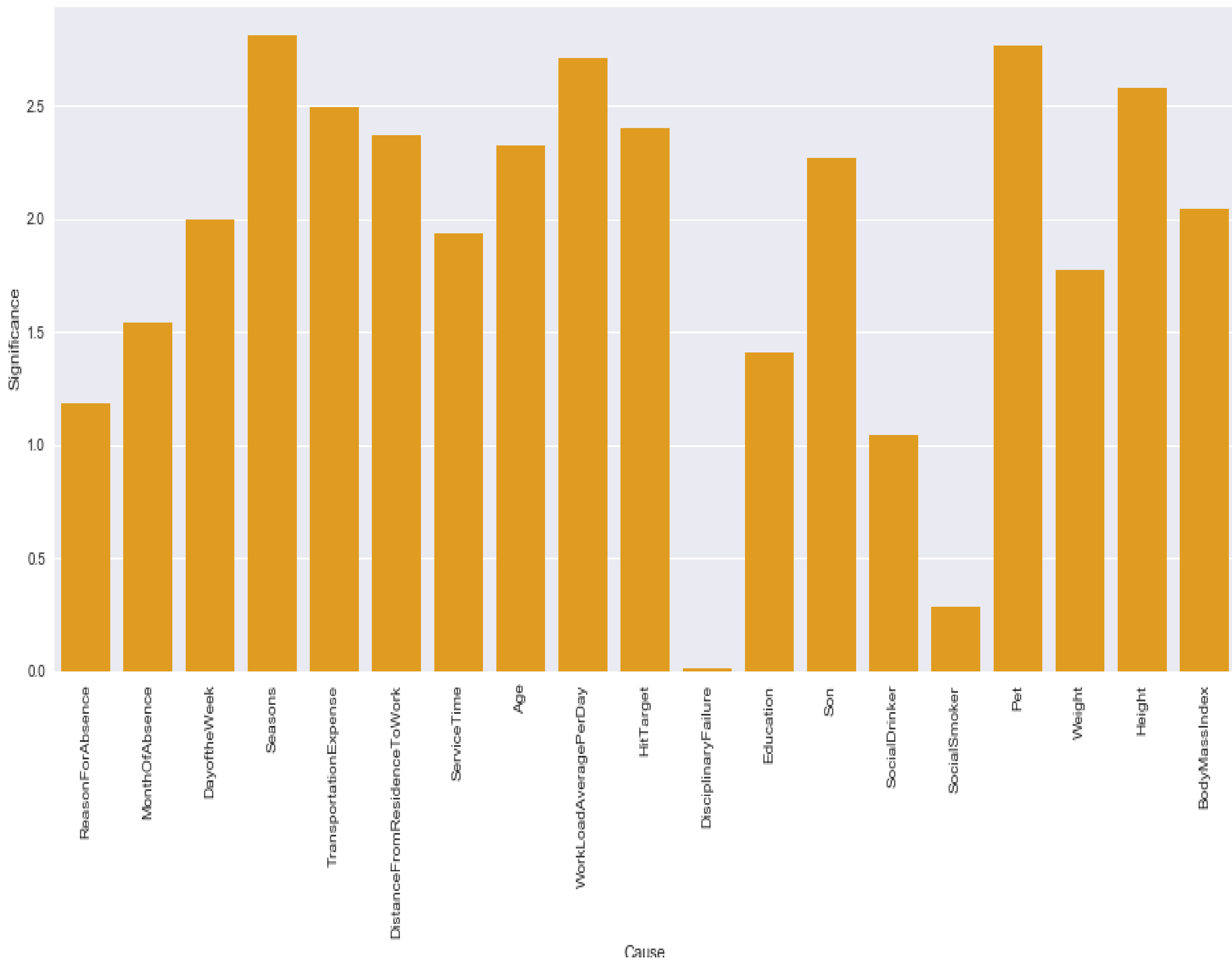


Figure 2.7 Most Significant Features Obtained through PCA(with outliers)

Figure 2.7 gives us insight about how much is the effect of each old variable in the new feature set is when we do the analysis on data with outliers. Here the predictor variables are reduced to 13 instead of 20. Above show the effect of Son, Wok Load Average and Pet is more in the new feature set.

2.2 Modeling

2.2.1 Model Selection

Now since we are ready with our features, we can apply models for predicting the absent time in hours. Here our dependent variable is a quantity hence we need to regression model. Let's first go with Multiple Linear Regression Model.

2.2.2 Multiple Linear Regression(R code without PCA)

- With Outliers

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

Residuals:
    Min      1Q  Median      3Q     Max 
-24.369  -4.981  -1.780   1.625  98.039 

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     24.35783    3.09573   7.868 2.05e-14 ***
Reason.for.absence              -0.52912    0.07938  -6.666 6.65e-11 ***
Month.of.absence                 0.04985    0.21554   0.231   0.8172    
Day.of.the.week                 -1.00247    0.39687  -2.526   0.0118 *  
Seasons                         -0.26252    0.57929  -0.453   0.6506    
Transportation.expense           0.58435    0.73833   0.791   0.4290    
Distance.from.Residence.to.Work -0.33196    0.90385  -0.367   0.7136    
Service.time                     0.05544    1.03017   0.054   0.9571    
Age                              1.51830    0.88428   1.717   0.0866 .  
Work.load.Average.day           -0.38420    0.61008  -0.630   0.5291    
Hit.target                       0.28462    0.66638   0.427   0.6695    
Disciplinary.failure                  NA         NA      NA       NA    
Education                       -1.79042    0.99548  -1.799   0.0727 .  
Son                              1.05733    0.64032   1.651   0.0993 .  
Social.drinker                   0.67363    1.81757   0.371   0.7111    
Social.smoker                   -1.80499    2.45854  -0.734   0.4632    
Pet                             -0.25405    0.74126  -0.343   0.7319    
Weight                           5.26166    6.15042   0.855   0.3927    
Height                          -1.67870    2.63169  -0.638   0.5238    
Body.mass.index                 -6.80952    5.84163  -1.166   0.2443    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.55 on 528 degrees of freedom
```

```
Multiple R-squared:  0.1474,   Adjusted R-squared:  0.1184
F-statistic: 5.073 on 18 and 528 DF,  p-value: 9.285e-11
```

Adjusted R square value shows that we can predict only 11.8 % of our data using this model. Although Reason of absence seems to be a significant factor.

```
Analysis of Variance Table

Response: Absenteeism.time.in.hours
                                 Df Sum Sq Mean Sq F value    Pr(>F)    
Reason.for.absence                1   9290  9290.2 58.9430 7.915e-14 ***
Month.of.absence                  1     10     9.9  0.0630  0.801935    
Day.of.the.week                   1    868   867.7  5.5052  0.019328 *  
Seasons                           1     66    66.4  0.4213  0.516595    
Transportation.expense            1    268   268.4  1.7028  0.192493    
Distance.from.Residence.to.Work   1    231   230.7  1.4635  0.226917    
Service.time                      1    296   296.0  1.8783  0.171113    
Age                               1    359   358.8  2.2763  0.131967    
Work.load.Average.day             1      0     0.2  0.0014  0.969775    
Hit.target                        1     56    55.6  0.3527  0.552842    
Education                         1    495   495.0  3.1409  0.076928 .  
Son                               1   1061  1060.9  6.7309  0.009739 ** 
Social.drinker                    1    165   164.9  1.0460  0.306903    
Social.smoker                     1     19    19.3  0.1222  0.726754    
Pet                               1     19    19.4  0.1231  0.725871    
Weight                            1    566   565.9  3.5901  0.058670 .  
Height                            1    408   407.9  2.5880  0.108277    
Body.mass.index                   1    214   214.2  1.3588  0.244267    
Residuals                       528  83219   157.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to this table we have Reason for absence, Day of the week and Son as significant factors.

```
LM_model2 = update(LM_model,. ~ . - Month.of.absence-Seasons-Work.load.Averag
e.day-Hit.target-Pet-Weight)
```

```

                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     23.52190    2.68859   8.749  < 2e-16 ***
Reason.for.absence              -0.52328    0.07792  -6.716  4.8e-11 ***
Day.of.the.week                 -1.00195    0.39047  -2.566   0.0106 *  
Transportation.expense           0.51130    0.70643   0.724   0.4695    
Distance.from.Residence.to.Work -0.46560    0.80060  -0.582   0.5611    
Service.time                     0.14840    0.87835   0.169   0.8659    
Age                              1.47122    0.83786   1.756   0.0797 .  
Disciplinary.failure                  NA         NA      NA       NA    
Education                       -1.60592    0.96986  -1.656   0.0983 .  
Son                              1.11312    0.63455   1.754   0.0800 .  
Social.drinker                   0.92436    1.70824   0.541   0.5886    
Social.smoker                   -2.02309    2.34972  -0.861   0.3896    
Height                           0.46021    0.69876   0.659   0.5104    
```

```
Body.mass.index                   -1.77844    0.73448  -2.421   0.0158 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.51 on 534 degrees of freedom
Multiple R-squared:  0.1445,   Adjusted R-squared:  0.1252
F-statistic: 7.514 on 12 and 534 DF,  p-value: 6.833e-13
```

Reason for absence shows maximum significance, Day of week, Body Mass Ind ex after that.

- Without Outliers

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

Residuals:
    Min      1Q  Median      3Q     Max
-7.1909 -1.8959 -0.4568  1.5329 13.1264

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      7.94777    0.72596  10.948  < 2e-16 ***
Reason.for.absence              -0.16780    0.01826  -9.188  < 2e-16 ***
Month.of.absence                 0.02118    0.04849   0.437 0.662431
Day.of.the.week                 -0.09224    0.09126  -1.011 0.312577
Seasons                         -0.08554    0.13427  -0.637 0.524350
Transportation.expense           0.67285    0.18075   3.723 0.000218 ***
Distance.from.Residence.to.Work -0.22467    0.15541  -1.446 0.148863
Service.time                    -0.13850    0.18028  -0.768 0.442660
Work.load.Average.day            0.19353    0.13688   1.414 0.157996
Hit.target                       0.02792    0.14508   0.192 0.847453
Disciplinary.failure            -1.75424    3.03210  -0.579 0.563135
Education                       -0.07757    0.22939  -0.338 0.735393
Son                              0.28497    0.15030   1.896 0.058498 .
Social.drinker                   1.07345    0.39818   2.696 0.007243 **
Social.smoker                    0.88746    0.54998   1.614 0.107205
Pet                             -0.32524    0.16469  -1.975 0.048807 *
Body.mass.index                 -0.02309    0.16875  -0.137 0.891229
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.943 on 531 degrees of freedom
Multiple R-squared:  0.2564,   Adjusted R-squared:  0.234
F-statistic: 11.44 on 16 and 531 DF,  p-value: < 2.2e-16
```

```
LM_model2 = update(LM_model,. ~ . - Month.of.absence-Seasons-Hit.target-Disci
plinary.failure-Education-Pet-Body.Mass.Index)
```

```
Call:
```

```
lm(formula = Absenteeism.time.in.hours ~ Reason.for.absence +
    Day.of.the.week + Transportation.expense + Distance.from.Residence.to.Wor
k +
    Service.time + Disciplinary.failure + Education + Son + Social.drinker +
    Social.smoker + Body.mass.index, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4312 -1.8761 -0.5427  1.5663 13.1265

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      7.614620   0.628912  12.108  < 2e-16 ***
Reason.for.absence              -0.171816   0.018115  -9.485  < 2e-16 ***
Day.of.the.week                 -0.084591   0.090715  -0.932  0.35150
Transportation.expense           0.541272   0.167925   3.223  0.00134 **
Distance.from.Residence.to.Work -0.258652   0.153929  -1.680  0.09347 .
Service.time                    -0.146823   0.179092  -0.820  0.41268
Disciplinary.failure            -1.735445   3.022377  -0.574  0.56607
Education                       -0.002493   0.219862  -0.011  0.99096
Son                              0.258075   0.149025   1.732  0.08389 .
Social.drinker                   1.440918   0.352682   4.086 5.07e-05 ***
Social.smoker                    1.044296   0.541728   1.928  0.05442 .
Body.mass.index                 -0.063557   0.166986  -0.381  0.70364
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.945 on 536 degrees of freedom
Multiple R-squared:  0.2482,   Adjusted R-squared:  0.2327
F-statistic: 16.08 on 11 and 536 DF,  p-value: < 2.2e-16
```

This model will be able to predict the results 23.4 % correctly. Significant vari ables being Reason for absence, Transportation Expense and Social Drinker.

Table 2.2 shows major Reason for Reasons for absence:

23: medical consultation	146
28: dental consultation	110
27: Physiotherapy	68

13: Diseases of the musculoskeletal system and connective tissue	54

2.2.3 Decision Tree(R code without PCA)

- Without outlier

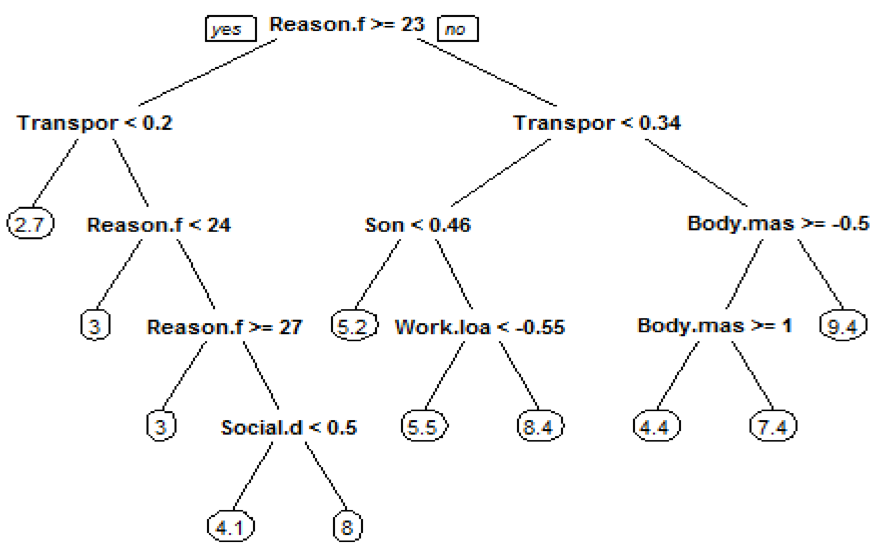


Figure 2.8 Decision tree for Employee Absenteeism

```
Variable Importance
Reason.for.absence            Transportation.expense            Service.time
   38                                15                              9
Body.mass.index                                  Son Distance.from.Residence.to.Work
        7                                         7                                6
Social.drinker              Work.load.Average.day                          Education
        4                             4                                    3
      Pet                     Day.of.the.week                     Social.smoker
      2                                 2                                2
                     Hit.target
                              1
```

- With Outliers

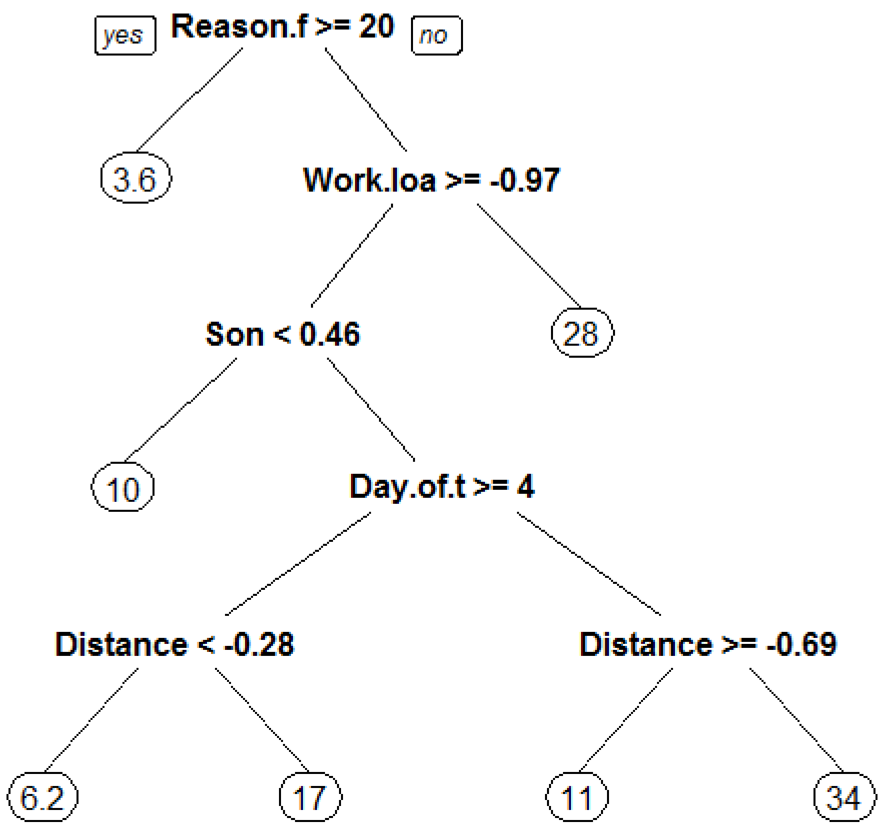


Figure 2.9 Decision tree for Employee Absenteeism (with outliers)

Variable Importance

```
Reason.for.absence Distance.from.Residence.to.Work Work.load.Average.day
32                                11                                10
Son           Transportation.expense                          Service.time
8                                 8                                7
Weight                      Day.of.the.week                           Age
6                                 5                                5
Height                            Education                           Pet
5                                 1                                1
Social.smoker
1
```

Chapter 3

Conclusion

3.1 Model Evaluation

Evaluation metrics help in explaining the performance of a model. The most p opular error metric to evaluate any regression model is Root Mean Square Err or (RMSE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Table 3.1 RMSE in case of R

	With Outlier	**Without Outlier**
Linear Regression	13.93	3
Decision Tree	14.06713	2.86
Random Forest	14.91	2.66

Table 3.2 RMSE in case of Python

	With PCA With Outlier	**With PCA without Outlier**
Linear Regression	7.75	2.98
Decision Tree	9.62	3.63
Random Forest	16	3.21
KNN	7.17	2.65
Naïve Baye's	35.9	2.95

3.2 Model Selection

We can see from the results that Linear Regression, KNN and Decision Tree models are somewhat better for this regression problem. This is when we have kept training and testing data ratio as 80:20.

Appendix B - R Code

Complete R File

```
#Load Libraries
x = c("ggplot2", "corrgram", "DMwR", "caret", "randomForest", "unbalanced", "C50", "du
mmies", "e1071", "Information","MASS", "rpart", "gbm", "ROSE", 'sampling', 'DataCombi
ne', 'inTrees',"Metrics","psych","party","rpart.plot")

lapply(x, require, character.only = TRUE)
rm(x)

## Read the data
Absenteeism_at_work = read.csv("Absenteeism_at_work_Project.csv", header = T, na.s
trings = c(" ", "", "NA"))

###Explore the data########
str(Absenteeism_at_work)

####Missing Values Analysis######################
missing_val = data.frame(apply(Absenteeism_at_work,2,function(x){sum(is.na(x))}))
missing_val$Columns = row.names(missing_val)
names(missing_val)[1] =  "Missing_percentage"
missing_val$Missing_percentage = (missing_val$Missing_percentage/nrow(Absenteeis
m_at_work)) * 100
missing_val = missing_val[order(-missing_val$Missing_percentage),]
row.names(missing_val) = NULL
missing_val = missing_val[,c(2,1)]
write.csv(missing_val, "Missing_perc.csv", row.names = F)
#Bar graph#
ggplot(data = missing_val[1:3,], aes(x=reorder(Columns, -Missing_percentage),y = Miss
ing_percentage))+
  geom_bar(stat = "identity",fill = "grey")+xlab("Parameter")+
  ggtitle("Missing data percentage (Train)") + theme_bw()

####Convert work load avg from factor to numeric######
```

```
#Relace comma by blank
Absenteeism_at_work$Work.load.Average.day=gsub(",","",Absenteeism_at_work$Work
.load.Average.day)
Absenteeism_at_work$Work.load.Average.day=as.numeric(as.character(Absenteeism_
at_work$Work.load.Average.day))

#check datatype
str(Absenteeism_at_work)

#Remove all 0s and NAs from target variable
str(Absenteeism_at_work)
Absenteeism_at_work=Absenteeism_at_work[!is.na(Absenteeism_at_work$Absenteeis
m.time.in.hours) & !(Absenteeism_at_work$Absenteeism.time.in.hours)==0,]


# kNN Imputation
Absenteeism_at_work = knnImputation(Absenteeism_at_work, k = 3)
sum(is.na(Absenteeism_at_work))

write.csv(Absenteeism_at_work, 'Absenteeism_at_work_missing.csv', row.names = F)
multi.hist(Absenteeism_at_work[,c(1:4)], main = NA, dcol = c("blue", "red"),
        dlty = c("solid", "solid"), bcol = "linen")
multi.hist(Absenteeism_at_work[,c(12:13,15:16)], main = NA, dcol = c("blue", "red"),
        dlty = c("solid", "solid"), bcol = "linen")
multi.hist(Absenteeism_at_work[,c(5,21)], main = NA, dcol = c("blue", "red"),
        dlty = c("solid", "solid"), bcol = "linen")


### BoxPlots - Distribution and Outlier Check
#print(colnames(Absenteeism_at_work))
cnames = colnames(Absenteeism_at_work[,-c(1:5,12:13,15:16)])
#print(cnames)
for (i in 1:length(cnames))
{
  assign(paste0("gn",i), ggplot(aes_string(y = (cnames[i]), x = "Absenteeism.time.in.hour
s"), data = subset(Absenteeism_at_work))+
        stat_boxplot(geom = "errorbar", width = 0.5) +
        geom_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=18,
                outlier.size=1, notch=FALSE) +
        theme(legend.position="bottom")+
        labs(y=cnames[i],x="Absenteeism.time.in.hours")+
        ggtitle(paste("Box plot of Absenteeism time in hours for",cnames[i])))
}

# ## Plotting plots together
gridExtra::grid.arrange(gn1,gn10,gn11,ncol=3)
```

```
gridExtra::grid.arrange(gn2,gn3,gn4,ncol=3)
gridExtra::grid.arrange(gn5,gn6,gn7,ncol=3)
gridExtra::grid.arrange(gn8,gn9,ncol=2)

# #Replace all outliers with NA and impute
# #create NA on "custAge
for(i in cnames){
  val = Absenteeism_at_work[,i][Absenteeism_at_work[,i] %in% boxplot.stats(Absenteei
sm_at_work[,i])$out]
  #print(length(val))
  Absenteeism_at_work[,i][Absenteeism_at_work[,i] %in% val] = NA
}

Absenteeism_at_work = knnImputation(Absenteeism_at_work, k = 3)

cor(Absenteeism_at_work[,-c(1:5,12:13,15:16,21)])

######Feature Selection##############################
## Correlation Plot
corrgram(Absenteeism_at_work[,-c(1:5,12:13,15:16,21)], order = F,
         upper.panel=panel.pie, text.panel=panel.txt, main = "Correlation Plot")

cnames=colnames(Absenteeism_at_work[,-c(1:5,12:13,15:16,21)])

# #Standardisation
for(i in cnames){
  print(i)
  Absenteeism_at_work[,i] = (Absenteeism_at_work[,i] - mean(Absenteeism_at_work[,i])
)/
    sd(Absenteeism_at_work[,i])
}

#Drop ID and correlated column
Absenteeism_at_work=Absenteeism_at_work[,-c(1,9,18,19)]
str(Absenteeism_at_work)

#Divide data into train and test using stratified sampling method
set.seed(123)
train.index = createDataPartition(Absenteeism_at_work$Absenteeism.time.in.hours, p =
.80, list = FALSE)
train = Absenteeism_at_work[ train.index,]
test  = Absenteeism_at_work[-train.index,]


##Decision tree for classification
#Develop Model on training data
```

```
C50_model = rpart(Absenteeism.time.in.hours ~., data=train)

#Summary of DT model
summary(C50_model)
prp(C50_model)
C50_Predictions = predict(C50_model, test[,-17])
rmse(test[,17],C50_Predictions)

###Random Forest
RF_model = randomForest(Absenteeism.time.in.hours ~., train, importance = TRUE, ntr
ee = 500)
#Predict test data using random forest model
RF_Predictions = predict(RF_model, test[,-17])
rmse(test[,17],RF_Predictions)

#Develop Linear Regression model
LM_model = lm(Absenteeism.time.in.hours ~., train)

#predict on test cases #raw
LM_Predictions = predict(LM_model, test[,1:17])
rmse(test[,17],LM_Predictions)
summary(LM_model)
anova(LM_model)
LM_model2 = update(LM_model,. ~ . - Month.of.absence-Seasons-Hit.target-Disciplinar
y.failure-Education-Pet-Body.Mass.Index)
summary(LM_model2)
f=as.data.frame(table(Absenteeism_at_work$Reason.for.absence))
```

References

Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.