Employee Absenteeism

Harish Ravi

25 January 2019

Contents

Chapter 1

Introduction

1.1 Problem Statement

XYZ is a courier company. The company is passing through genuine issue of Absenteeism.
The company has shared it dataset and requested to have an answer on the following areas.

1. What changes company should bring to reduce the number of absenteeism?
2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

1.2 Data

Our task is to build a model which will project every month losses in 2011 based on previous data. Given below is a sample of the data set that we are using to project the 2011 losses.

Table 1.1: Employee Absenteeism Sample Data (Columns: 1-9)

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age
11	26	7	3	1	289	36	13	33
36	0	7	3	1	118	13	18	50
3	23	7	4	1	179	51	18	38
7	7	7	5	1	279	5	14	39
11	23	7	5	1	289	36	13	33
3	23	7	6	1	179	51	18	38

Table 1.2: Employee Absenteeism Sample Data (Columns: 10-21)

Work load Average/day	Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
239,554	97	0	1	2	1	0	1	90	172	30	4
239,554	97	1	1	1	1	0	0	98	178	31	0
239,554	97	0	1	0	1	0	0	89	170	31	2
239,554	97	0	1	2	1	1	0	68	168	24	4
239,554	97	0	1	2	1	0	1	90	172	30	2
239,554	97	0	1	0	1	0	0	89	170	31	5

The Explanation of variables is as follows

1. Individual identification (ID)
2. Reason for absence (ICD).
Absences attested by the International Code of Diseases (ICD) 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 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)

Chapter 2

Methodology

2.1 Pre Processing

Any predictive modeling requires the data before we start modeling. However, in data mining terms *looking at data* refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process we will first look at any missing values in our data set.

2.1.1 Missing Value Analysis

In Figure 2.1.1 we have plotted the missing values plot. So we can see from the plot that there are missing values in our data.

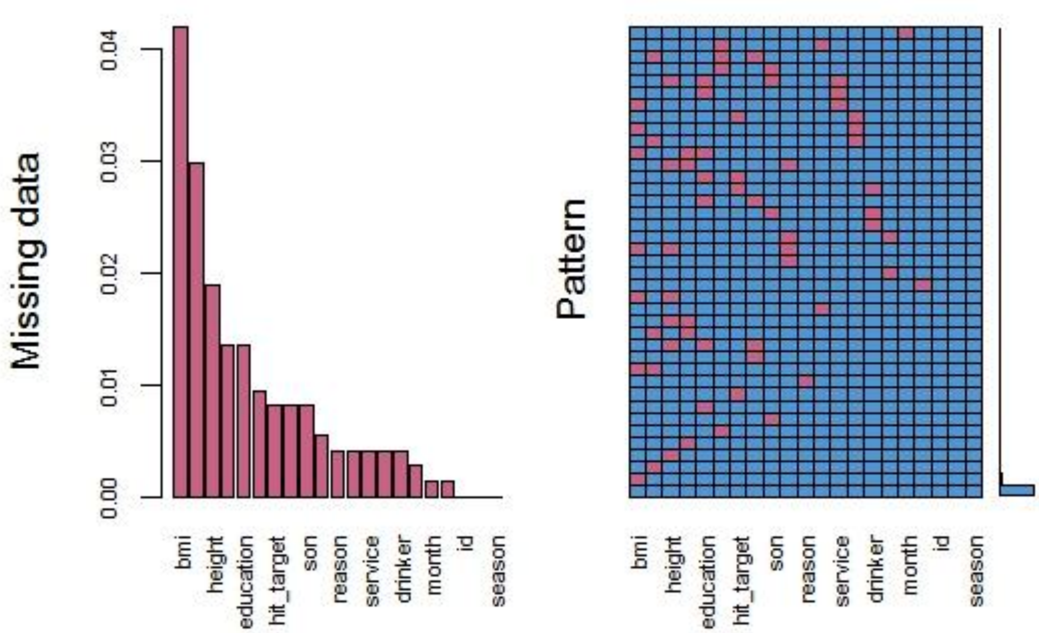


Fig 2.1.1

Missing data can occur because of no response or no information is provided for one or more items. Missing values leads to Biased Model. We can deal missing values using MICE package.

The MICE package in R, helps you imputing missing values with plausible data values. These plausible values are drawn from a distribution specifically designed for each missing data point.

Using MICE Package we have imputed the missing Values in our data.

2.1.2 Data Analysis

Data analysis is a process of inspecting and analyzing the data with the goal of discovering useful information and supporting decision making.

Let’s analyze our data by using plots and Visualizations to understand more from our data.

Plotting independent variables to understand their impact on our target variable

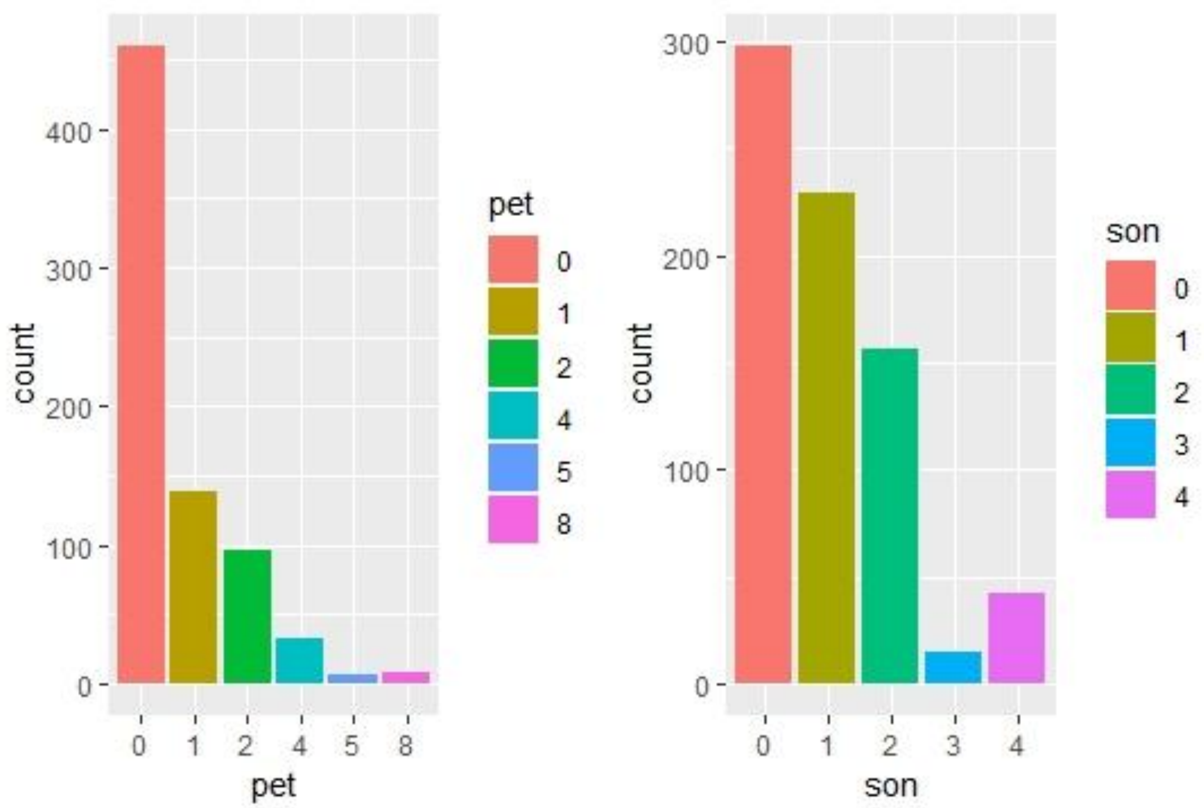


Fig 2.1.2.1

From above plots we can understand

- The people without any child have more absent time.
- The people without any pet or with only 1 pet have more absent time.

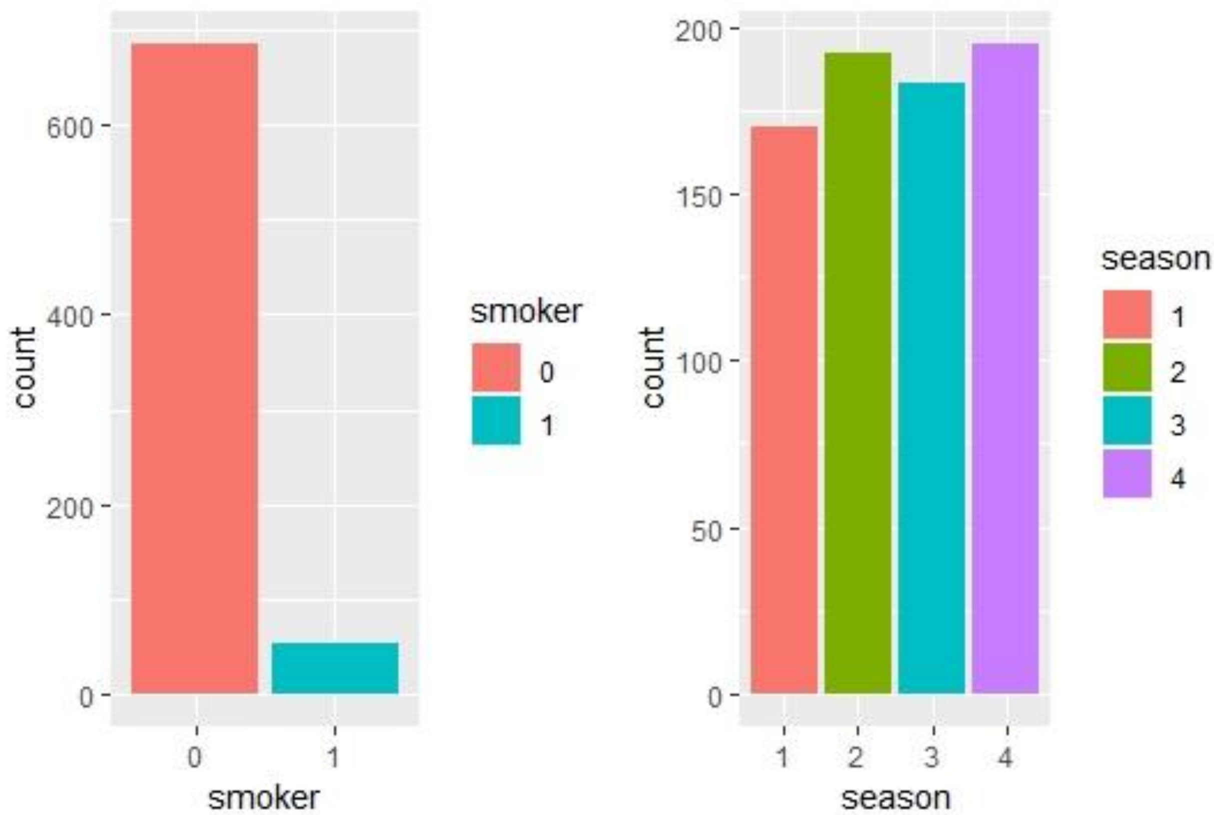


Fig 2.1.2.2

From the above plots we can understand that

- The employees with more absent time are non smokers
- Every season has same number of absent time but little more in seasons 2(autumn) and 4(spring).

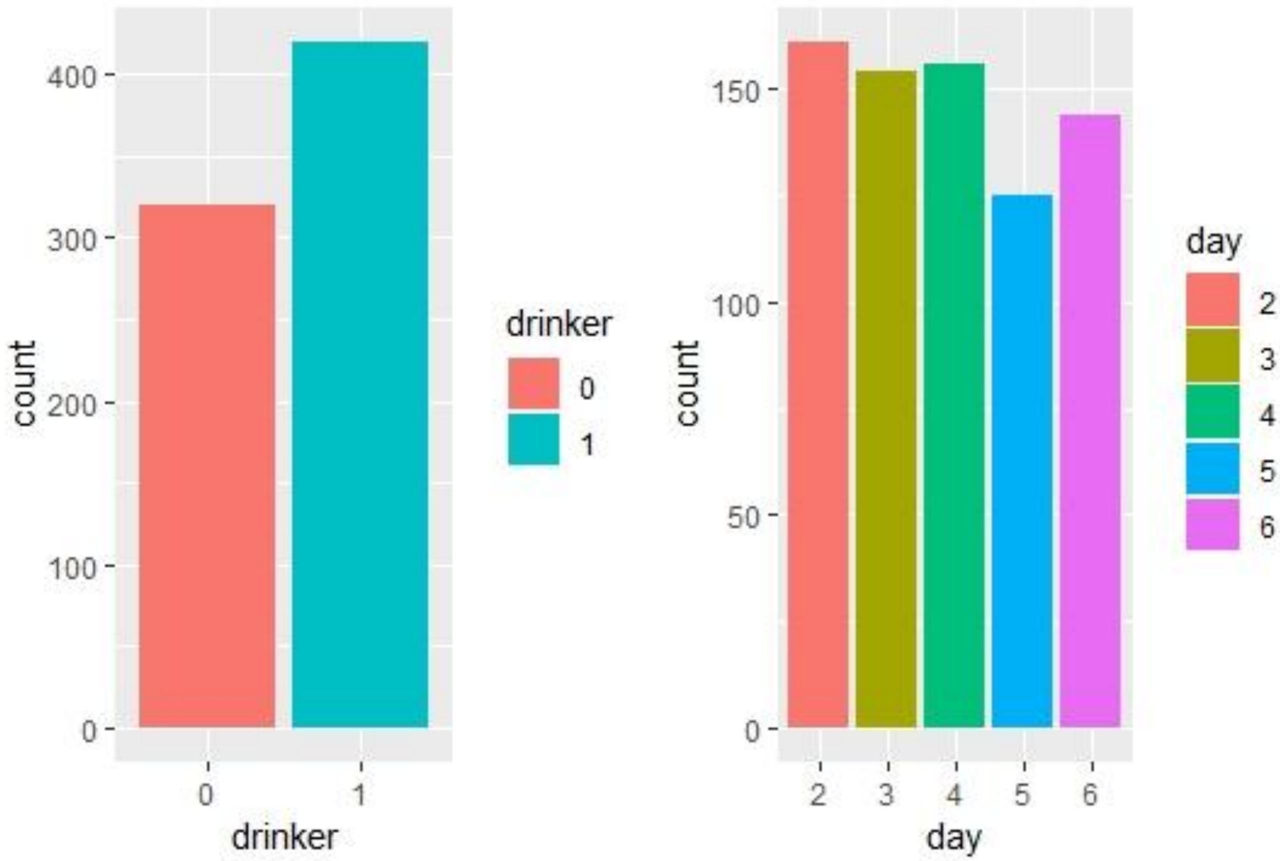


Fig 2.1.2.3

From the above plots we can observe that

- Employees with more absent time are social drinkers.
- More number of employees are absent in starting day and ending day of the week.

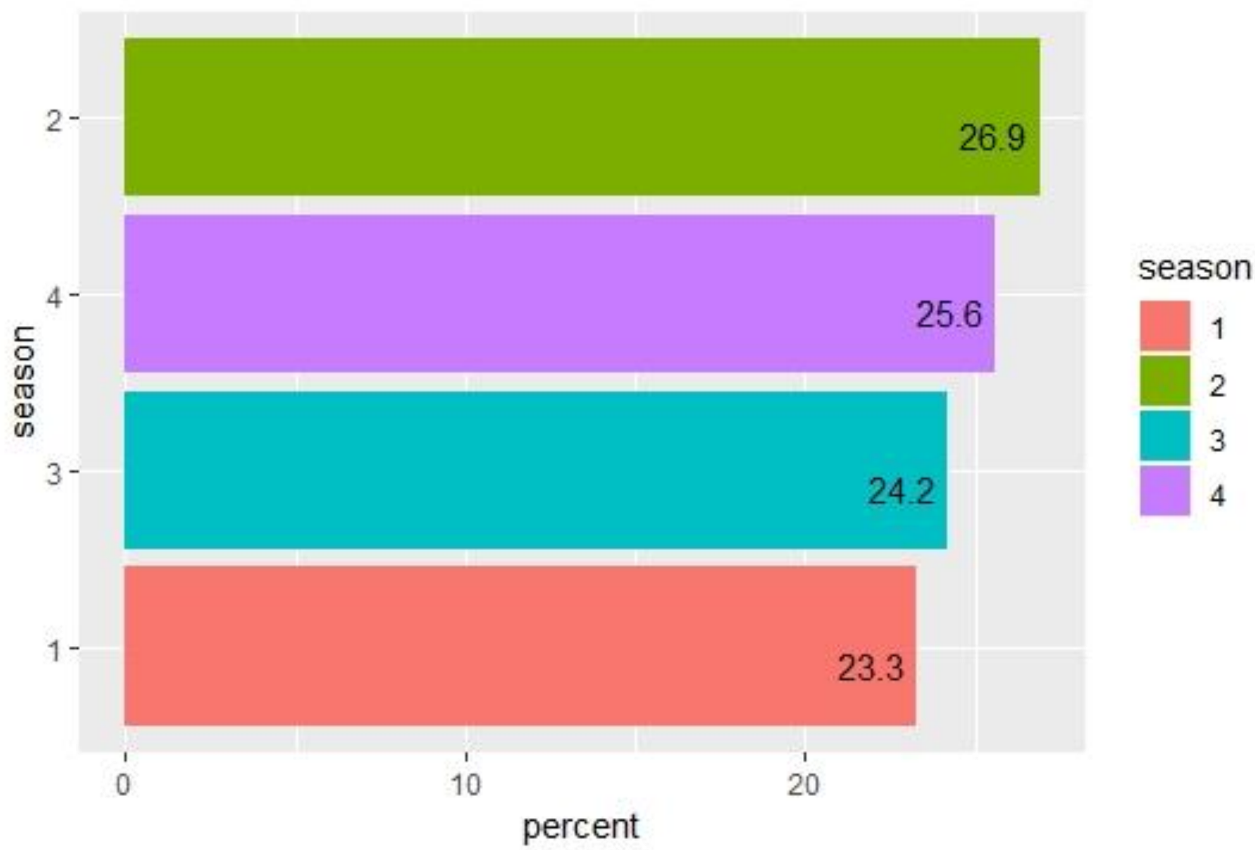


Fig 2.1.2.4

From the above plot we can understand the percentage of absent time in each season

1. Season 1 (Summer) shares 23.3 %
2. Season 2 (Autumn) shares 26.9 %
3. Season 3 (Winter) shares 24.2 %
4. Season 4 (Spring) shares 25.6 %

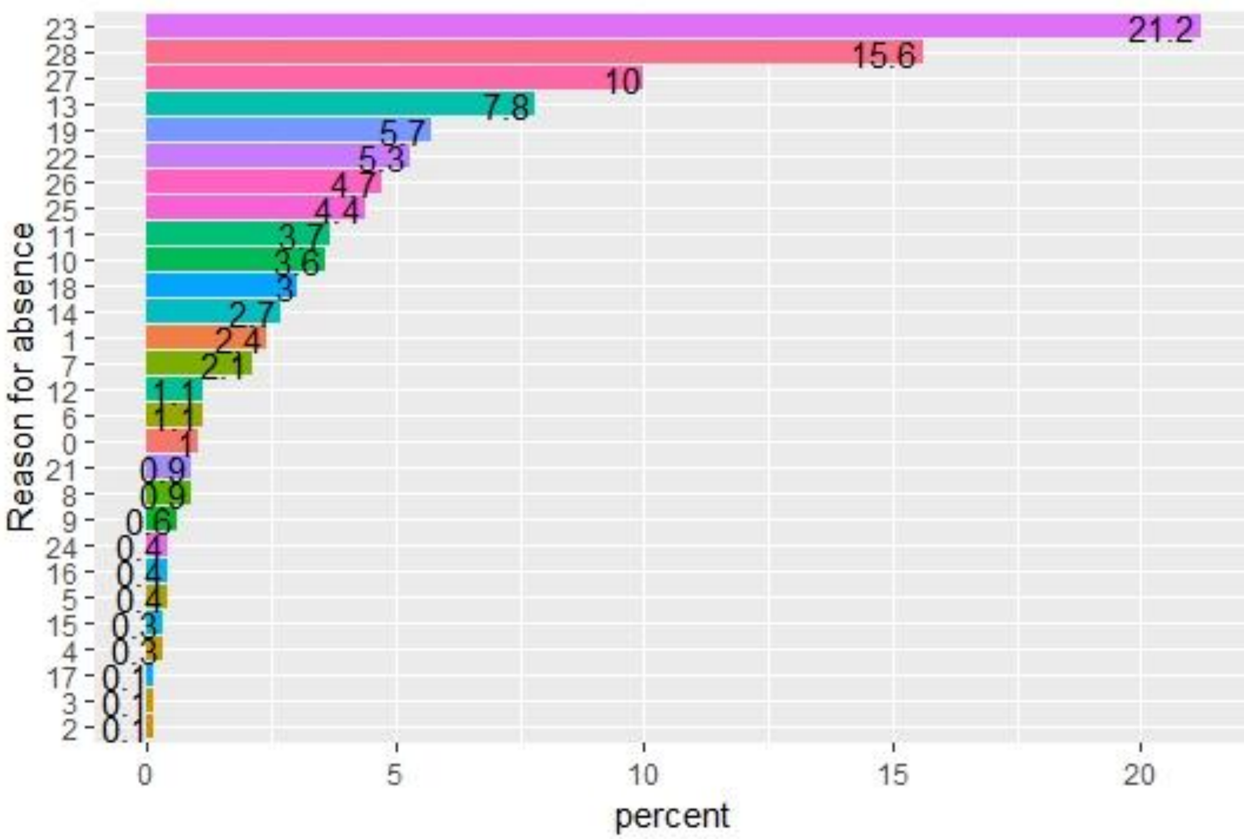


Fig 2.1.2.5

From the above plot we can observe the top medical reasons for absent time

The top reasons for more absent time are 23, 28 and 27.

- 23 : Medical Consultation and its share is 21.2 %
- 28 : Physiotherapy and its share is 15.6%
- 27 : Dental Consultation and its share is 10%

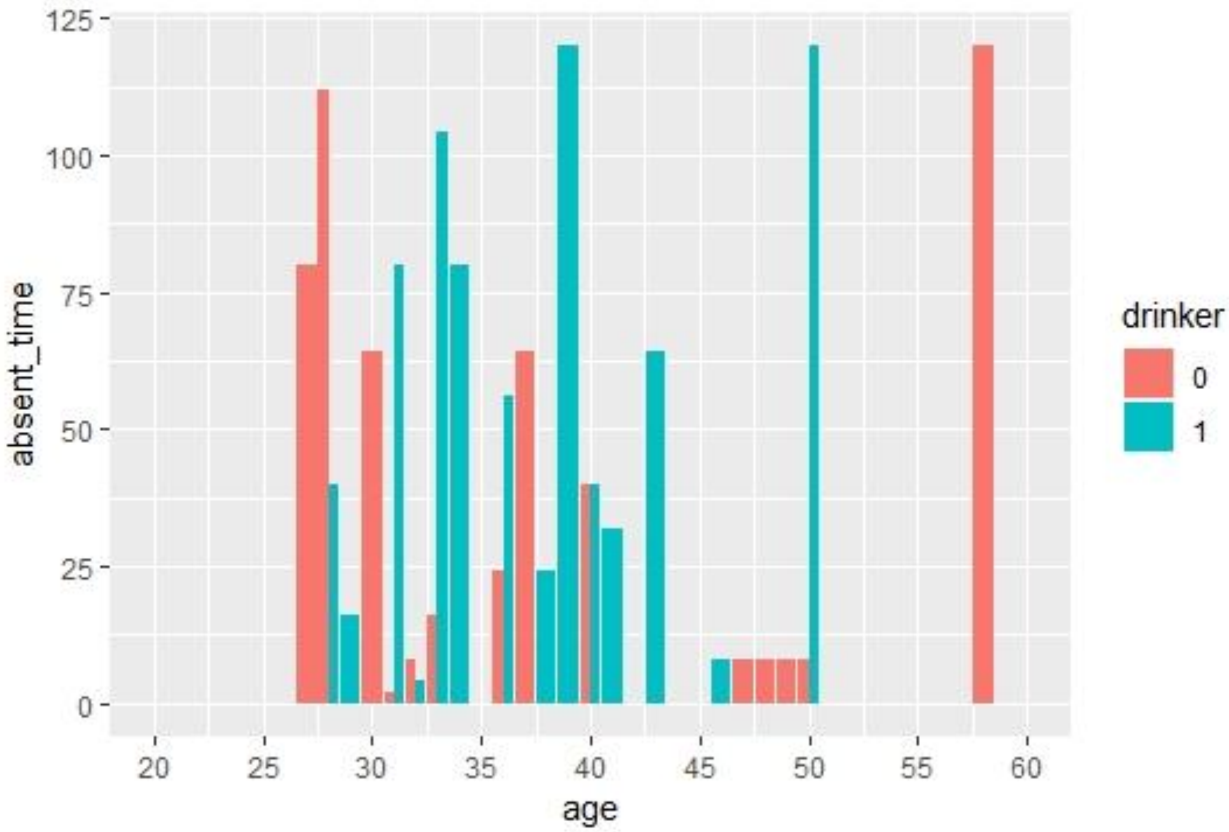


Fig 2.1.2.6

From the above plot we can observe that

- More number of employees who drink are between age group of 30 and 50
- Too young and too old employees are less absent and are not social drinkers.

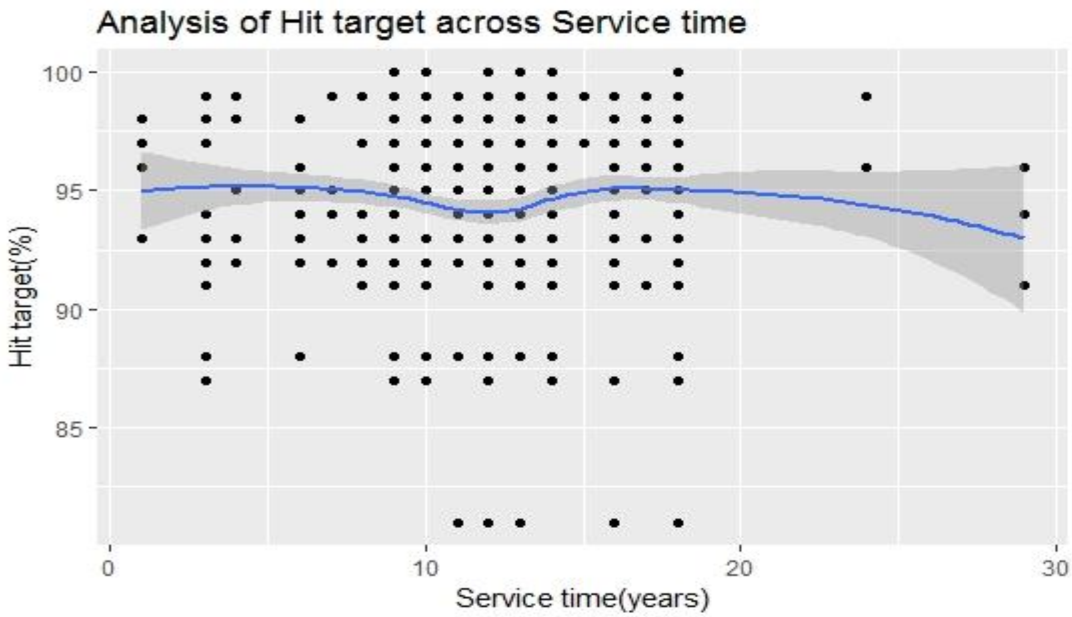


Fig 2.1.2.7

The hit target is good with employees of less service time and more service. There is observable decline in hit target of middle aged employees.

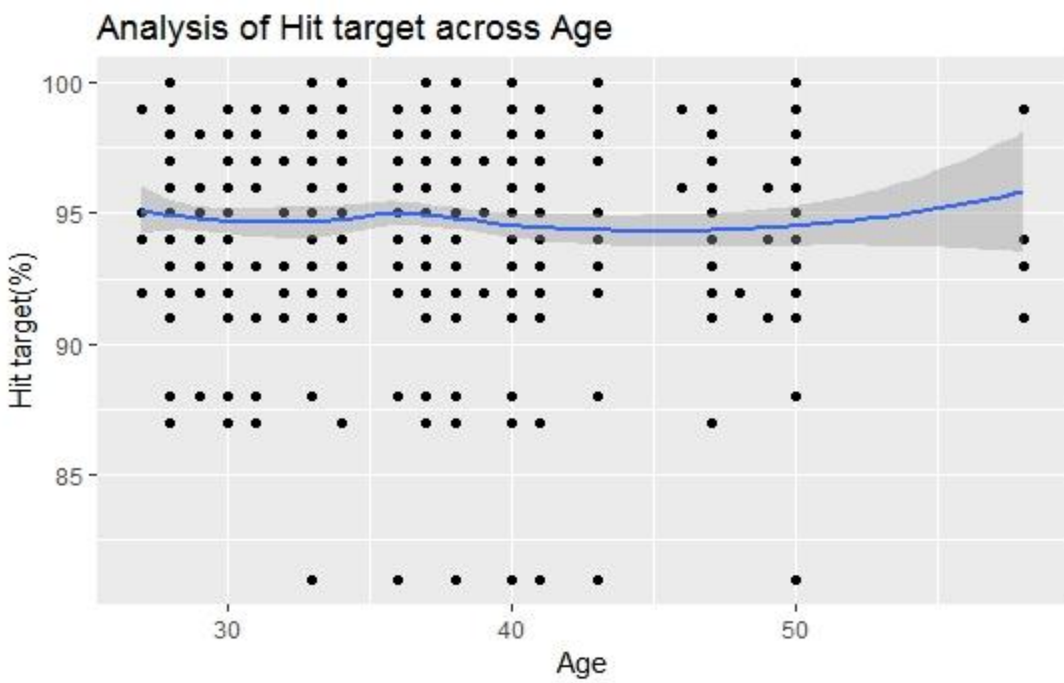


Fig 2.1.2.8

The hit target is constant across all age groups

We can observe that the Service time and age of employees are following same pattern.

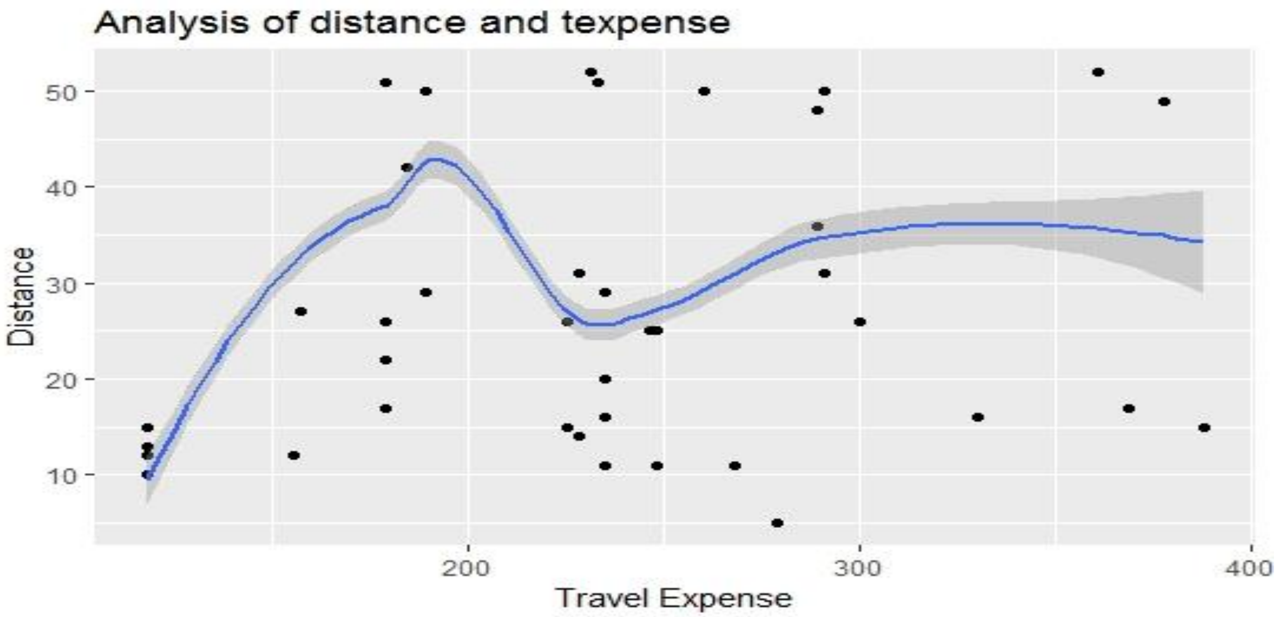


Fig 2.1.2.9

Travel Expense of employee is increasing with distance of home from work place.

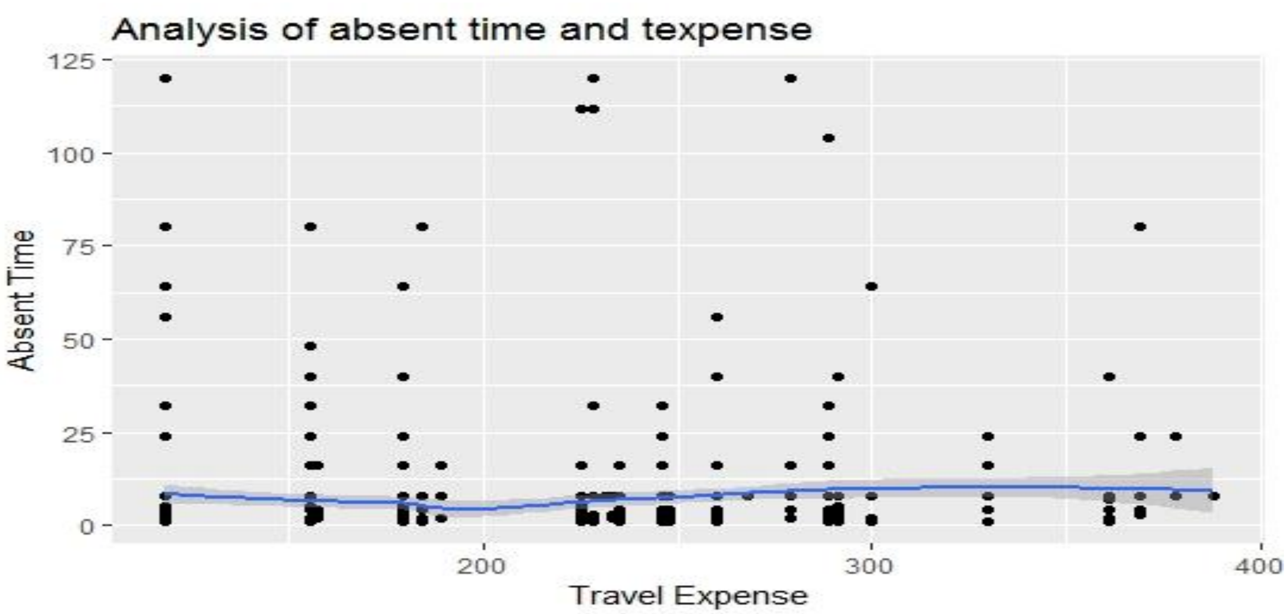


Fig 2.1.2.10

The absent time is constant irrespective of travel expense.

From the above two plots we can observe that travel distance and travel expense is not affecting the target variable much.

Summary from overall Analysis

- The people without any child have more absent time.
- The people without any pet or with only 1 pet have more absent time.
- The employees with more absent time are non smokers
- Every season has same number of absent time but little more in seasons 2(autumn) and 4(spring).
- Employees with more absent time are social drinkers.
- More number of employees are absent in starting day and ending day of the week.
- The top reasons for more absent time are 23, 28 and 27.
    - 23 : Medical Consultation and its share is 21.2 %
    - 28 : Physiotherapy and its share is 15.6%
    - 27 : Dental Consultation and its share is 10%
- More number of employees who drink are between age group of 30 and 50
- Too young and too old employees are less absent and are not social drinkers
- The hit target is good with employees of less service time and more service
- The hit target is constant across all age groups
- Travel Expense of employee is increasing with distance of home from work place.
- The absent time is constant irrespective of travel expense.

From all the above plots we can observe that each variable is affecting the target variable in some ratio. Let's reduce the dimensions or variables of our data using Principal Component Analysis.

2.1.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. It is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible

Eigen value conceptually represents that amount of variance accounted for by a factor.

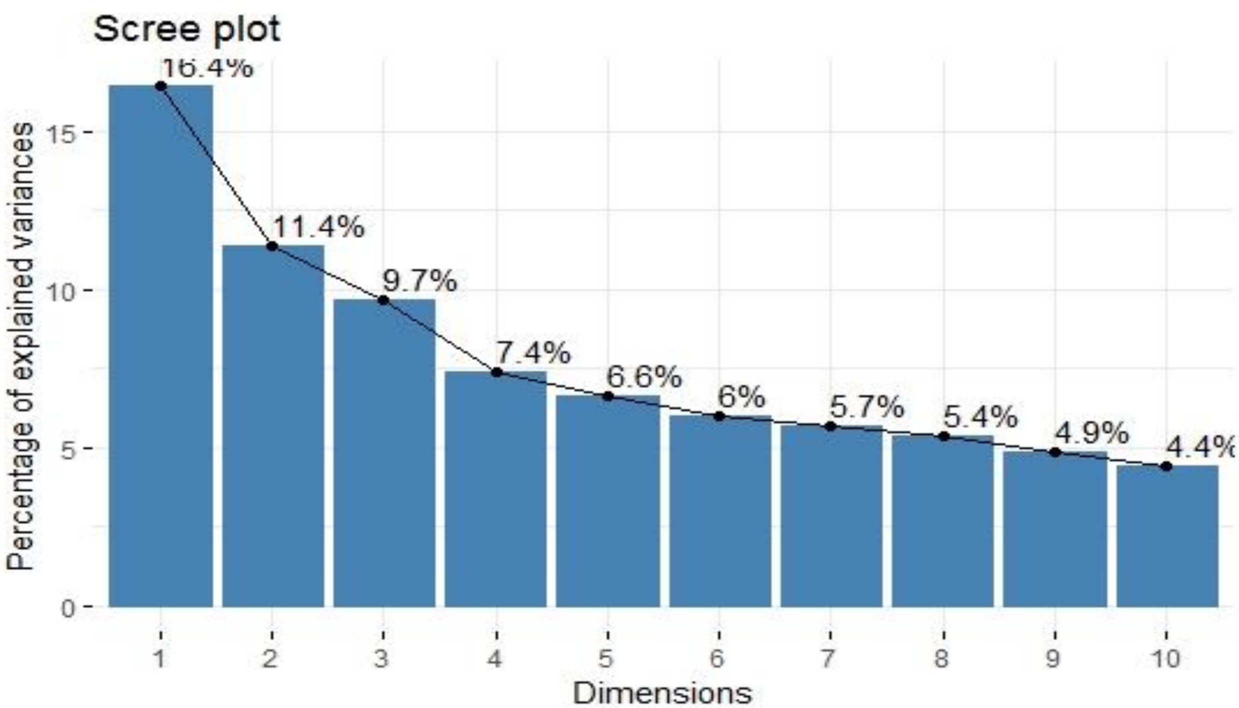


Fig 2.1.3.1

The above Plot shows the percentage of explained variance across Dimensions.

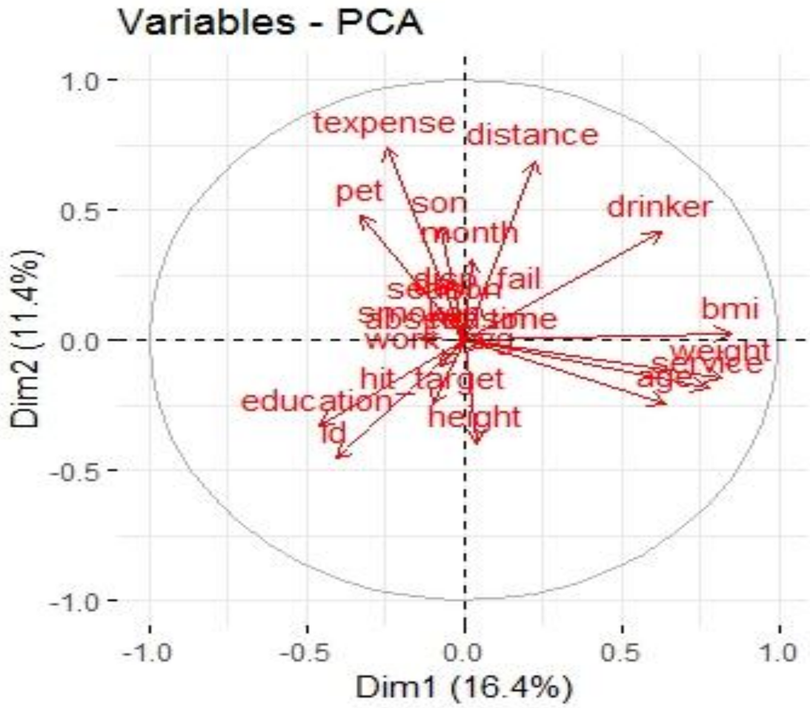


Fig 2.1.3.2

The above plot shows the magnitude and variance of each variable against dimension 1 to dimension 2.

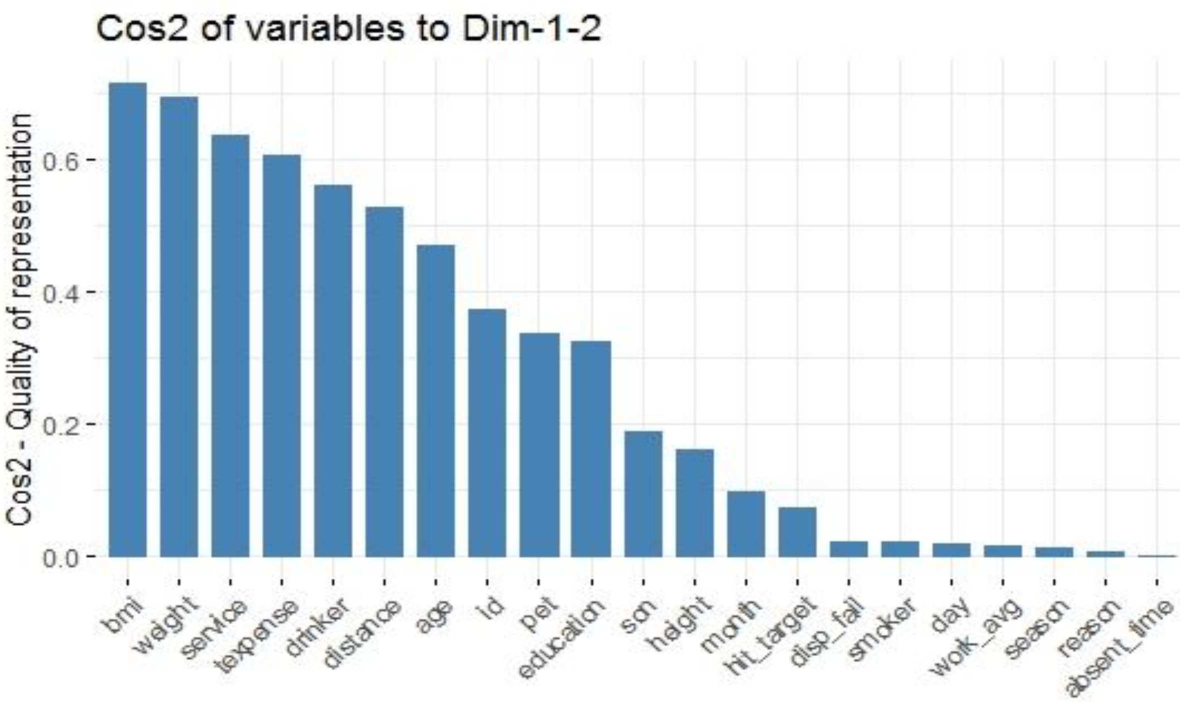


Fig 2.1.3.3

The above plot shows the quality of representation of each variable to dimension 1 to dimension 2.

2.1.4 Feature Selection

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem.

We will select only the following variables which are playing an important role while predicting based on PCA analysis.

1. Individual identification (ID)
2. Reason for absence (ICD).
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. Body mass index
15. Absenteeism time in hours (target)

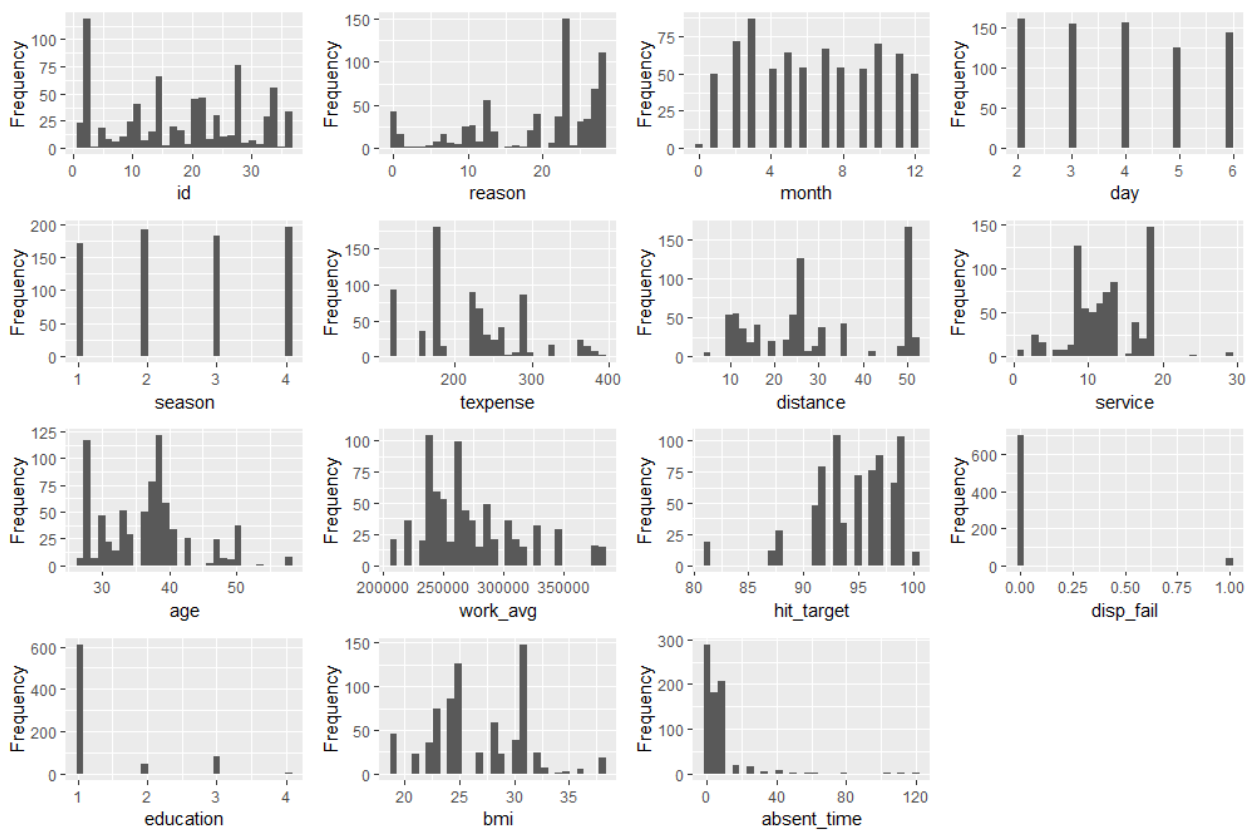


Fig 2.1.4.1

Histogram of all independent variables against target variable.

2.1.5 Correlation Analysis

Let's look for highly correlated variables in the data before feeding to the model. A very simple way of looking at correlations in the data is plotting Heat map between variables.

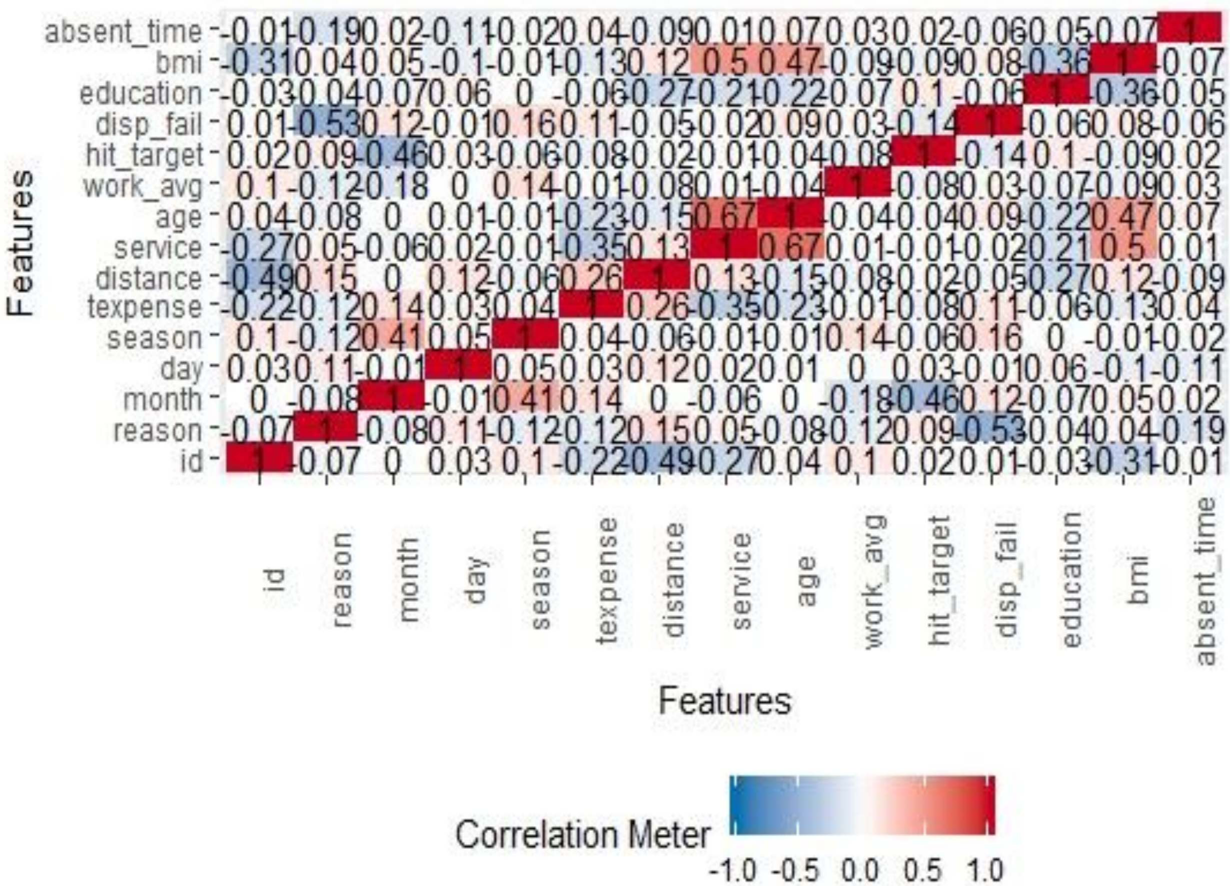


Fig 2.1.5.1

From the above heat map we can observe that there are no highly correlated variables in our data.

2.2 Modeling

2.2.1 Model Selection

The dependent variable is a Class Variable, the analysis that we can perform is **Classification and Clustering**.

You always start your model building from the simplest to the complex.

2.2.2 K-Nearest Neighbors Algorithm

K-Nearest Neighbors Model to predict the absenteeism time in hours

K= 2

knn_pred = knn(train_data, test_data, cl = train_data$absent_time, 2)

KNN_CM = confusionMatrix(test_data$absent_time, knn_pred)

KNN_CM$overall

Accuracy	Kappa	Accuracy Lower	Accuracy Upper	Accuracy Null
0.351351351	0.204479283	0.274761749	0.434024290	0.256756757

Accuracy PValue	McnemarPValue
0.006732354	NaN

K = 4

knn_pred2 = knn(train_data, test_data, cl = train_data$absent_time, 4)
KNN_CM2 = confusionMatrix(test_data$absent_time, knn_pred2)
KNN_CM2$overall

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull
0.391891892	0.240592930	0.312763089	0.475439894	0.277027027

AccuracyPValue	McnemarPValue
0.001636139	NaN

K = 6

knn_pred3 = knn(train_data, test_data, cl = train_data$absent_time, 6)
KNN_CM3 = confusionMatrix(test_data$absent_time, knn_pred3)
KNN_CM3$overall

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull
0.3243243	0.1467282	0.2497592	0.4060862	0.3378378

AccuracyPValue	McnemarPValue
0.6651666	NaN

With K = 4 is giving better accuracy than any in KNN.

2.2.3 Decision Tree Classifier Algorithm

Decision Tree Classifier Model to predict the absenteeism time in hours

```
DTree = rpart(absent_time~., data = train_data)

DTree_pred = predict(DTree, test_data, type = "class")
```

DTree_CM = confusionMatrix(test_data$absent_time, DTree_pred)
DTree_CM$overall

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull
0.547297297	0.404372898	0.463474967	0.629188113	0.425675676

AccuracyPValue	McnemarPValue
0.001919656	NaN

```
Decision tree model is giving good accuracy than KNN.
```

2.2.4 K-Means Model

Now we will try and use K-Means Clustering Model

```
kmeadata = data.frame(scale(meadata[-15]))
kmeans_model = kmeans(kmeadata,5,nstart = 25)

kmeans_model

K-means clustering with 5 clusters of sizes 278, 197, 89, 40, 136

Cluster means:
```

	id	reason	month	day	season	texpense	distance
1	0.07836930	0.150688601	0.1213899	0.12060851	0.02136181	0.7509725	0.3161803
2	0.78138834	-0.007220749	-0.1334335	-0.19005389	-0.07435092	-0.7181815	-0.9129215
3	-0.07805507	-0.033953947	-0.1081321	0.17843355	0.04579025	-0.2773862	-0.6439226
4	0.01882114	-2.316218094	0.4499965	-0.06321057	0.63445377	0.4956162	-0.2356610
5	-1.24651552	0.405894716	-0.1164420	-0.06941700	-0.15253630	-0.4590137	1.1667864

	service	age	work_avg	hit_target	disp_fail	education	bmi
1	-0.696576102	-0.6714894	0.05559112	-0.03967859	-0.2388841	-0.3213608	-0.40217691
2	0.356783275	0.7995283	0.16031179	0.06305475	-0.2388841	-0.3204658	0.04952998
3	-0.397186493	-0.5205993	-0.23961006	0.20323190	-0.2388841	2.4870850	-0.89350588
4	-0.002773146	0.4354665	0.12707172	-0.61459733	4.1804726	-0.2478943	0.30504743
5	1.167811579	0.4270710	-0.22642153	0.03753763	-0.2388841	-0.4335642	1.24535222

Within cluster sum of squares by cluster:
[1] 2450.4473 1885.3590 743.5843 482.3116 1029.2242
 (between_SS / total_SS = 36.3 %)

Chapter 3

Conclusion

3.1 Model Evaluation

Now that we have a few models for classifying the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Predictive Performance
2. Interpretability
3. Computational Efficiency

In our case of Employee Absenteeism problem. We will use *Predictive performance* as the criteria to compare and evaluate models.

3.1.1 Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

```
KNN_CM3 = confusionMatrix(test_data$absent_time, knn_pred3)
KNN_CM3$overall

Accuracy        Kappa      AccuracyLower  AccuracyUpper   AccuracyNull
0.3243243      0.1467282      0.2497592      0.4060862      0.3378378

AccuracyPValue    McnemarPValue
0.6651666              NaN
```

```
DTree_CM = confusionMatrix(test_data$absent_time, DTree_pred)
DTree_CM$overall

Accuracy         Kappa          AccuracyLower  AccuracyUpper   AccuracyNull
0.547297297    0.404372898    0.463474967    0.629188113    0.425675676

AccuracyPValue   McnemarPValue
0.001919656         NaN
```

3.2 Model Selection

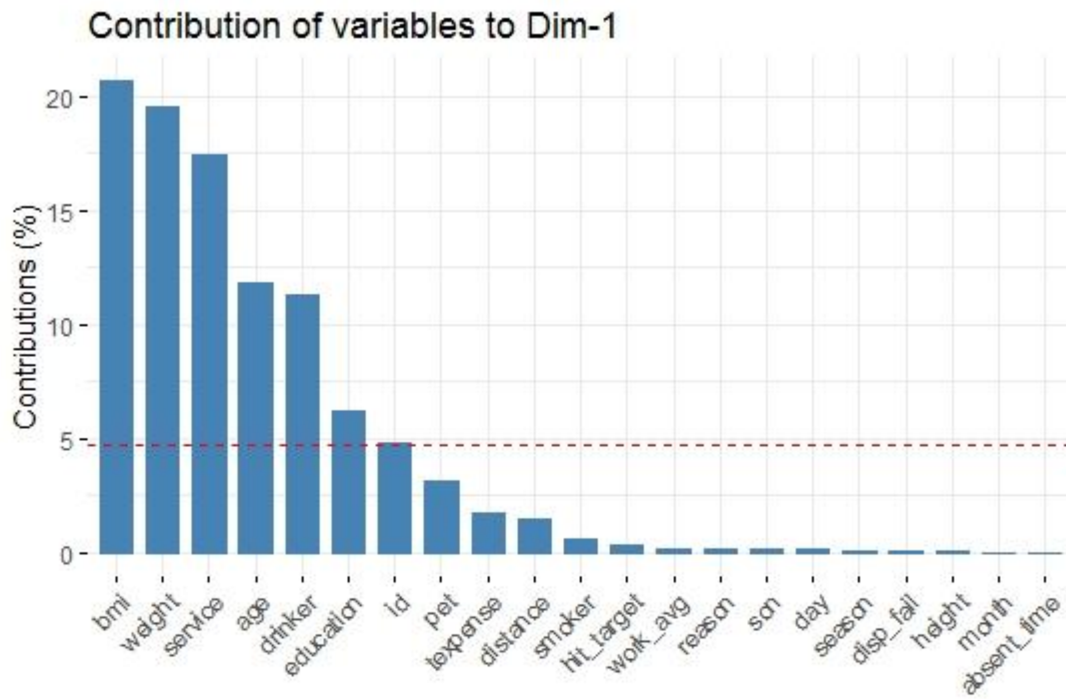
However we must understand that clustering algorithms are used to gain more insight about the data and find structure/patterns within the data points if there is any.

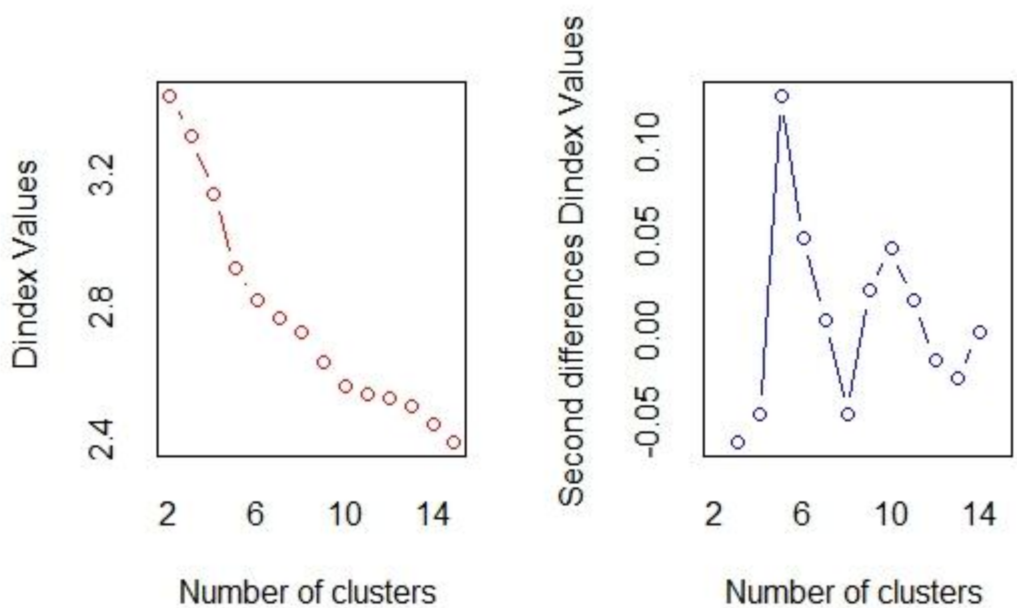
Clustering methods are not used for prediction or classification purposes and hence there is no meaning in evaluating the performance of such models in tasks such as prediction of hours of absenteeism. Even if we were to come up with a way to evaluate the models, their accuracies would be much greater than the accuracy of any other classification algorithm as there is a significant decrease in the number of classes while clustering.

We have to select the Model depending upon Problem statement.

Decision Tree Model is giving more accuracy than any classification model.

Appendix A - Extra Figures





Appendix B - R Code

Geometric Plot of Variables (Fig: 2.1.1)

```
p  = ggplot(feadata, aes(x = pet, fill = pet)) + geom_bar()
s  = ggplot(feadata, aes(x = son, fill = son)) + geom_bar()
ss = ggplot(feadata, aes(x = smoker, fill = smoker)) + geom_bar()
sd = ggplot(feadata, aes(x = drinker, fill = drinker)) + geom_bar()
d  = ggplot(feadata, aes(x = day, fill = day)) + geom_bar()
se = ggplot(feadata, aes(x = season,fill = season)) + geom_bar()

grid.arrange(p,s, nrow = 1)
grid.arrange(ss,se, nrow = 1)
grid.arrange(sd, d, nrow = 1)
```

Eigen Values (Fig 3.1.1)

```
#eigen values
egval = get_eigenvalue(pcaex)
fviz_eig(pcaex,addlabels=T)

#correlation of variables with PCA components
fviz_pca_var(pcaex,col.var='red')
```

Correlation Map (Fig 3.3.1)

```
plot_correlation(meadata)
```

K-Means Clusters Value Plot

```
nbclust_result = NbClust(kmeadata,min.nc = 2,max.nc = 15,method = "kmeans")

barplot(table(nbclust_result$Best.n[1,]),
      xlab = "No of Clusters",
      ylab = "No of Criteria",
      main = "No of Clusters Choosen")
```

Complete R File

```
#Clearing RAM
rm(list = ls())

#importing all the required libraries

library(plyr)
library(dplyr)
library(tibble)
library(readxl)
library(ggplot2)
library(rpart)
library(DataExplorer)
library(ggthemes)
library(grid)
library(gridExtra)
library(factoextra)
library(FactoMineR)
library(forcats)
library(mice)
library(VIM)
library(caret)
library(caTools)
library(class)
library(MASS)
library(NbClust)
library(fossil)

#Knowing the working directory
getwd()

#setting the working directory
setwd("C:/Users/Harish/Desktop/Projects")

#Importing the data
eadata  = read_excel("Absenteeism_at_work_Project.xls")

#Understanding the data or the summary of the day data

head(eadata,5)
summary(eadata)
View(eadata)

#Changing the variable or attribute names to a simple name

names(eadata) = c("id", "reason","month","day","season","texpense",
                "distance","service","age","work_avg","hit_target",
                "disp_fail","education","son","drinker","smoker","pet",
                "weight","height","bmi","absent_time")

#Knowing the data type of the variables
str(eadata)
```

```
#converting the data type of required variables to categorical form

eadata$reason     =as.factor(eadata$reason)
eadata$month      =as.factor(eadata$month)
eadata$day        =as.factor(eadata$day)
eadata$season     =as.factor(eadata$season)
eadata$disp_fail  =as.factor(eadata$disp_fail)
eadata$education  =as.factor(eadata$education)
eadata$son        =as.factor(eadata$son)
eadata$drinker    =as.factor(eadata$drinker)
eadata$smoker     =as.factor(eadata$smoker)
eadata$pet        =as.factor(eadata$pet)


#Cheking for any missing values in the dataset
sum(is.na(eadata))

#Finding the no of missing values for each variable

sapply(eadata, function(x) sum(is.na(x)))

#plotting missing values on a single plot using VIM package

miss_plot = aggr(eadata, col=mdc(1:2),
           numbers=TRUE, sortVars=TRUE,
           labels=names(eadata), cex.axis=.7,
           gap=3, ylab=c("Missing data","Pattern"))


#we have missing values in many columns with total of 135 values

#Imputing Missing Values using MICE package with max iterations of 10

mice_imputes = mice(eadata , m=5, maxit = 10)

#knowing the methods used by MICE to impute the values of each column

mice_imputes$method

#Selecting the best imputed data sets by MICE
#Lets take the 5th data set

feadata = complete(mice_imputes,5)

#checking for completeness of the imputed data

sum(is.na(feadata))

#Now we have data with all values imputed
#Data Exploration
#plotting data for better understanding

p  = ggplot(feadata, aes(x = pet, fill = pet)) + geom_bar()
s  = ggplot(feadata, aes(x = son, fill = son)) + geom_bar()
ss = ggplot(feadata, aes(x = smoker, fill = smoker)) + geom_bar()
sd = ggplot(feadata, aes(x = drinker, fill = drinker)) + geom_bar()
d  = ggplot(feadata, aes(x = day, fill = day)) + geom_bar()
se = ggplot(feadata, aes(x = season,fill = season)) + geom_bar()
```

```
grid.arrange(p,s, nrow = 1)
grid.arrange(ss,se, nrow = 1)
grid.arrange(sd, d, nrow = 1)

# creating a new variable 'absent' to filter absentees

absent = as.data.frame( feadata %>% filter(absent_time > 0))

#Plotting 'absent' againest all variables to understand each variable imapct on absent time
#Plotting againest seasons

season1 = as.data.frame(absent %>% group_by(season) %>% summarise(count= n(), percent =
round(count*100/nrow(absent),1))%>% arrange(desc(count)))
ggplot(season1,aes(x= reorder(season,percent), y= percent, fill = season)) + geom_bar(stat='identity') +
coord_flip() +
  geom_text(aes(label = percent), vjust = 1.1, hjust = 1.2) + xlab('season')

#From the plot it seems every season have around same absent_time but a little
#more in autumn and spring

#Plotting againest disciplinary failure

disciplinary = as.data.frame(absent %>% group_by(disp_fail) %>% summarise(count= n(), percent =
round(count*100/nrow(absent),1))%>% arrange(desc(count)))
ggplot(disciplinary,aes(x= reorder(disp_fail,percent), y= percent, fill = disp_fail)) +
geom_bar(stat='identity') + coord_flip() +
  geom_text(aes(label = percent), vjust = 1.1, hjust = 1.2) + xlab('Disciplinary failure')

#No dispilinary failures for absentees

#plotting againest each medical reason

med_reason = as.data.frame(absent %>% group_by(reason) %>% summarise(count= n(), percent =
round(count*100/nrow(absent),1))%>% arrange(desc(count)))
ggplot(med_reason,aes(x = reorder(reason,percent), y= percent, fill= reason)) + geom_bar(stat = 'identity')
+ coord_flip() + theme(legend.position='none') +
  geom_text(aes(label = percent), vjust = 0.5, hjust = 1.1) + xlab('Reason for absence')

#Top medical reasons are 23,28 and 27

#Plotting againest Social drinker

ggplot(absent,aes(x= age,y= absent_time,fill= drinker)) + geom_bar(stat='identity',position=
position_dodge()) +
  scale_x_continuous(breaks =c(seq(20,60,5)),limits=c(20,60))

#Middle age group people are drinking and with high absent time

#Plotting Hit target across service time

ggplot(absent,aes(x= service,y= hit_target)) + geom_point()+ geom_smooth(method = 'loess') +
  ggtitle('Analysis of Hit target across Service time') + xlab('Service time(years)') + ylab('Hit target(%)')

#service time is showing same trend as age

#Plotting Hit target across age

ggplot(absent,aes(x= age,y= hit_target)) + geom_point()+ geom_smooth(method = 'loess') +
  ggtitle('Analysis of Hit target across Age') + xlab('Age') + ylab('Hit target(%)')
```

```
#Analysis of travel expense across distance

ggplot(absent,aes(x= texpense,y= distance)) + geom_point()+ geom_smooth(method = 'loess') +
  ggtitle('Analysis of distance and texpense') + xlab('Travel Expense') + ylab('Distance')

#The travel expense is more above 35 range of distance
#analysis of travel expense across absenteeism time

ggplot(absent,aes(x= texpense,y= absent_time)) + geom_point()+ geom_smooth(method = 'loess') +
  ggtitle('Analysis of absent time and texpense') + xlab('Travel Expense') + ylab('Absent Time')

#Travel expense is not showing much variations in absent_time

#analysis of distance across absenteeism time
ggplot(absent,aes(x= distance,y= absent_time)) + geom_point()+ geom_smooth(method = 'loess') +
  ggtitle('Analysis of absent time and distance') + xlab('Distance') + ylab('Absent Time')

#above 35 range distance have more absent hours

#we can see from the plots every variable is contributing to absenteeism
#Lets reduce the dimensionality of data set by selecting important variables only
#We use PCA for Dimensionality Reduction
#coverting variables to numeric to carry out PCA

feadata$reason     =as.numeric(feadata$reason)
feadata$month      =as.numeric(feadata$month)
feadata$day        =as.numeric(feadata$day)
feadata$season     =as.numeric(feadata$season)
feadata$disp_fail  =as.numeric(feadata$disp_fail)
feadata$education  =as.numeric(feadata$education)
feadata$son        =as.numeric(feadata$son)
feadata$drinker    =as.numeric(feadata$drinker)
feadata$smoker     =as.numeric(feadata$smoker)
feadata$pet        =as.numeric(feadata$pet)

#Scaling the data for PCA

peadata = feadata

peadata = scale(peadata)


pcaex = PCA(peadata,graph = F)

#eigen values
egval = get_eigenvalue(pcaex)
fviz_eig(pcaex,addlabels=T)

#correlation of variables with PCA components
fviz_pca_var(pcaex,col.var='red')

#quality of presentation of variables in correlogram
fviz_cos2(pcaex,choice='var',axes=1:2)

#contribution of variables to the respective principal components
fviz_contrib(pcaex,choice='var',axes=1)
```

```
#Feature selection of dataset based on PCA analysis
#Selecting only impacting variables
meadata = subset(feadata, select=c("id", "reason","month","day","season",
                                   "texpense","distance","service","age",
                                   "work_avg","hit_target","disp_fail",
                                   "education","bmi","absent_time"))

#plotting histogram to see the impacts of variables

plot_histogram(meadata)

#checking for correlation among selected variables

plot_correlation(meadata)

#There is no much correlated variables in our selected data
#Converting required variables to categorical type

meadata$reason     =as.factor(meadata$reason)
meadata$month      =as.factor(meadata$month)
meadata$day        =as.factor(meadata$day)
meadata$season     =as.factor(meadata$season)
meadata$disp_fail  =as.factor(meadata$disp_fail)
meadata$education  =as.factor(meadata$education)
meadata$absent_time=as.factor(meadata$absent_time)

#Now the data is ready for feeding to the model
#Model building

#Starting with KNN

#set seed to ensure you always have same random numbers generated
set.seed(53)

#splitting the data into train and test sets

sample = sample.split(meadata,SplitRatio = 0.80)

train_data = subset(meadata,sample ==TRUE)

test_data = subset(meadata, sample==FALSE)

#Building the model with k_value =2

knn_pred = knn(train_data, test_data, cl = train_data$absent_time, 2)

#evaluating the model

KNN_CM = confusionMatrix(test_data$absent_time, knn_pred)

KNN_CM$overall

#Accuracy =33 with K=2
```

#Now with K value 4

knn_pred2 = knn(train_data, test_data, cl = train_data$absent_time, 4)

KNN_CM2 = confusionMatrix(test_data$absent_time, knn_pred2)

KNN_CM2$overall

#Accuracy =32 with k=4

#Now with K value 6

knn_pred3 = knn(train_data, test_data, cl = train_data$absent_time, 6)

KNN_CM3 = confusionMatrix(test_data$absent_time, knn_pred3)

KNN_CM3$overall

#Accuracy =37 with k = 6

#It seems K=6 is giving better accuracy in KNN

#Building Model with Decision Tree
#Multi Class Classification with Decision Tree Classifier

DTree = rpart(absent_time~., data = train_data)

DTree_pred = predict(DTree, test_data, type = "class")

#Evaluating the Model

DTree_CM = confusionMatrix(test_data$absent_time, DTree_pred)

DTree_CM$overall

#Accuracy = 54

#Building Clusttering Model with K means
#K_Means Clusttering

#Converting the variable type
meadata$reason   =as.numeric(meadata$reason)
meadata$month   =as.numeric(meadata$month)
meadata$day   =as.numeric(meadata$day)
meadata$season   =as.numeric(meadata$season)
meadata$education =as.numeric(meadata$education)
meadata$disp_fail =as.numeric(meadata$disp_fail)
meadata$absent_time=as.numeric(meadata$absent_time)

```
#Scaling the data
kmeadata = data.frame(scale(meadata[-15]))

#knowing the optimum K_value
nbclust_result = NbClust(kmeadata,min.nc = 2,max.nc = 15,method = "kmeans")

barplot(table(nbclust_result$Best.n[1,]),
	xlab = "No of Clusters",
	ylab = "No of Criteria",
	main = "No of Clusters Choosen")

#Building the model with K value

kmeans_model = kmeans(kmeadata,5,nstart = 25)

#Summary of the model
kmeans_model

#Evaluating the model

kcluster_accuracy = table(meadata$absent_time,kmeans_model$cluster)

rand.index(meadata$absent_time,kmeans_model$cluster)

#Accuracy =67

#Generally Clusttering Models give more accuracy than Classification
#Among Classification models, Decision tree is giving more accuracy than any of its type
```

References

An Introduction to Statistical Learning in application with R. Vol. 7. Springer.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer Science & Business Media.

Machine Learning with Python *by* Json Brownlee.