Employee Absenteeism

Nivedha Radhakrishnan

20 June 2020

Contents

Chapter 1

Introduction

1.1 Problem Statement

Human capital plays an important role in the collection, transportation, and delivery of a courier company XYZ. The company is passing through a genuine issue of Absenteeism. The company has shared its dataset and requested to have an answer on the following areas:

1. Measures to reduce the number of absenteeism

2. Predict the trend of absenteeism in 2011 provided the trend remains the same

1.2 Data

It is crucial to understand the data before processing it. We perform initial investigations to unravel patterns and anomalies and test hypotheses. Doing so gives a rough outline of insights the data can provide.

Dataset Details:

Dataset Characteristics: Timeseries Multivariant

Number of Attributes: 21

Missing Values : Yes

The variables can fall into either of the categories:

1. Nominal
2. Ordinal
3. Interval
4. Ratio
5. Continuous

The attributes are numeric, however, we will split them into categorical and numeric variables logically. Ordinal variables will be considered categorical variables. ‘Absenteeism in hour’ is the target variable and it is continuous.

Attribute details:

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)

```
  ID Reason.for.absence Month.of.absence Day.of.the.week Seasons Transportation.expense
1 11                26                7               3       1                    289
2 36                 0                7               3       1                    118
3  3                23                7               4       1                    179
4  7                 7                7               5       1                    279
5 11                23                7               5       1                    289
6  3                23                7               6       1                    179
  Distance.from.Residence.to.Work Service.time Age Work.load.Average.day Hit.target
1                              36           13  33              2,39,554         97
2                              13           18  50              2,39,554         97
3                              51           18  38              2,39,554         97
4                               5           14  39              2,39,554         97
5                              36           13  33              2,39,554         97
6                              51           18  38              2,39,554         97
  Disciplinary.failure Education Son Social.drinker Social.smoker Pet Weight Height
1                    0         1   2              1             0   1     90    172
2                    1         1   1              1             0   0     98    178
3                    0         1   0              1             0   0     89    170
4                    0         1   2              1             1   0     68    168
5                    0         1   2              1             0   1     90    172
6                    0         1   0              1             0   0     89    170
  Body.mass.index Absenteeism.time.in.hours
1              30                         4
2              31                         0
3              31                         2
4              24                         4
5              30                         2
6              31                        NA
>
```

Figure 1.1 Sample dataset shows the first 6 rows and all columns

```
'data.frame':   740 obs. of  21 variables:
 $ ID                            : Factor w/ 36 levels "1","2","3","4",..: 11 36 3 7 11 3 10 20 14 1 ..
 $ Reason.for.absence            : Factor w/ 28 levels "0","1","2","3",..: 26 1 23 8 23 23 22 23 20 22
 ...
 $ Month.of.absence              : Factor w/ 13 levels "0","1","2","3",..: 8 8 8 8 8 8 8 8 8 8 ...
 $ Day.of.the.week               : Factor w/ 5 levels "2","3","4","5",..: 2 2 3 4 4 5 5 5 1 1 ...
 $ Seasons                       : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 $ Transportation.expense        : int  289 118 179 279 289 179 NA 260 155 235 ...
 $ Distance.from.Residence.to.Work: int  36 13 51 5 36 51 52 50 12 11 ...
 $ Service.time                  : int  13 18 18 14 13 18 3 11 14 14 ...
 $ Age                           : int  33 50 38 39 33 38 28 36 34 37 ...
 $ Work.load.Average.day         : chr  "2,39,554" "2,39,554" "2,39,554" "2,39,554" ...
 $ Hit.target                    : int  97 97 97 97 97 97 97 97 97 97 ...
 $ Disciplinary.failure          : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ Education                     : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 3 ...
 $ Son                           : Factor w/ 5 levels "0","1","2","3",..: 3 2 1 3 3 1 2 5 3 2 ...
 $ Social.drinker                : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 1 ...
 $ Social.smoker                 : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
 $ Pet                           : Factor w/ 6 levels "0","1","2","4",..: 2 1 1 1 2 1 4 1 1 2 ...
 $ Weight                        : int  90 98 89 68 90 89 80 65 95 88 ...
 $ Height                        : int  172 178 170 168 172 170 172 168 196 172 ...
 $ Body.mass.index               : int  30 31 31 24 30 31 27 23 25 29 ...
 $ Absenteeism.time.in.hours     : int  4 0 2 4 2 NA 8 4 40 8 ...
```

Figure 1.2 Summary of data after changing data type

Chapter 2

Methodology

2.1 Pre Processing

Pre-processing refers to exploring the data, cleaning the data, and visualizing the data through summary statistics and graphical representation. This is done with the help of Exploratory Data Analysis.

From the data we obtain that the target/dependent variable is continuous, hence we will deal with the model using regression. Regression analysis requires the data to be normally distributed. We use histograms and bar plots to visualize the probability distribution of the variables. In Figure 2.1 and Figure 2.2 we see a distribution plot of all the variables. The numeric variables are plotted by histogram and categorical data is plotted by barplot.EDA includes stages based on the requirement of the data. Let us explore each stage in detail in further sections.

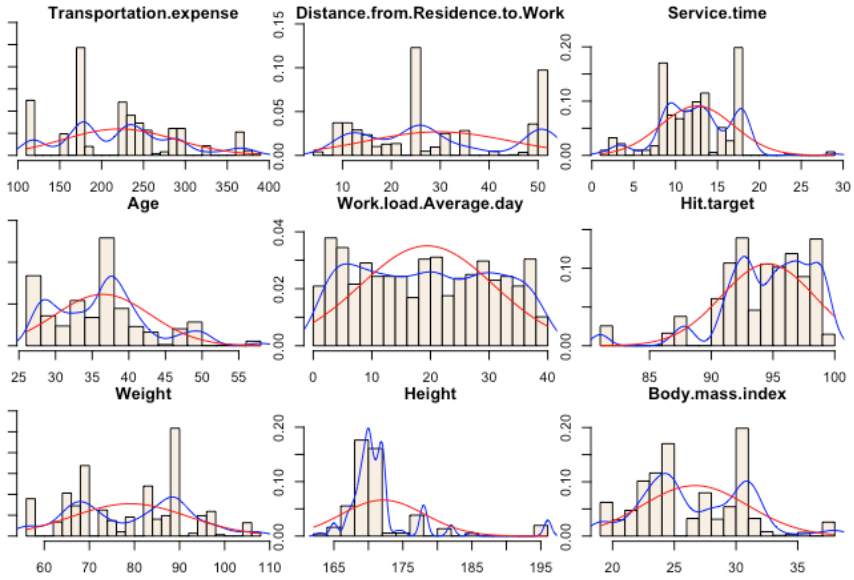

Figure 2.1: Histogram of numerical variables showing the distribution

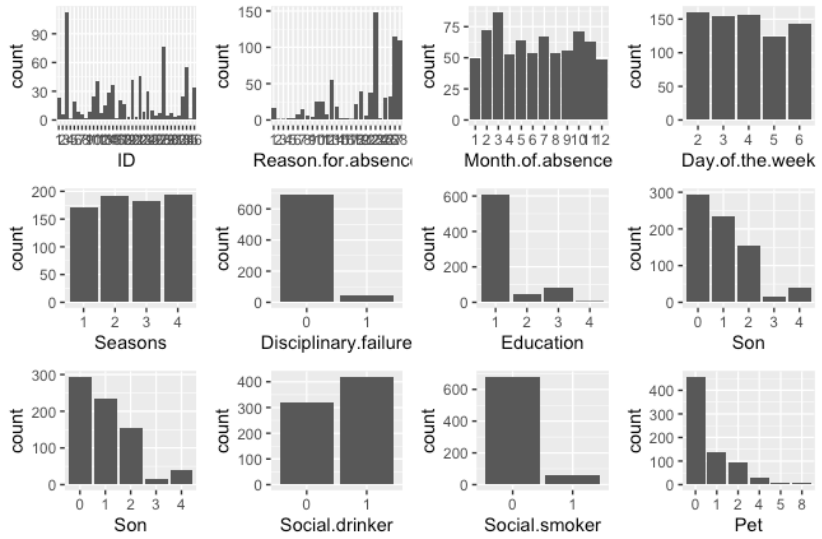

Figure 2.2: Barplot of numerical variables showing the distribution

2.1.1 Missing Value Analysis

The data includes null values in most of the columns. It also includes ‘0’s in columns like Reason for absence, Month of absence, Day of the week, Seasons, Education, ID, Age, Weight, Height, and Body mass index which is logically not possible. So we consider them as null values too. For numerical variables, we use mean and median to impute the values whereas for categorical variables we used mode. In Figure 2.3 and Fig 2.4 we have the list of several null values in each column before and after imputing 0s with NA.

	NA_count	NA_Percent
Body.mass.index	31	4.1891892
Absenteeism.time.in.hours	22	2.9729730
Height	14	1.8918919
Education	10	1.3513514
Transportation.expense	7	0.9459459
Hit.target	6	0.8108108
Disciplinary.failure	6	0.8108108
Son	6	0.8108108
Social.smoker	4	0.5405405
Reason.for.absence	3	0.4054054
Distance.from.Residence.to.Work	3	0.4054054
Service.time	3	0.4054054
Age	3	0.4054054
Social.drinker	3	0.4054054
Pet	2	0.2702703
Month.of.absence	1	0.1351351
Weight	1	0.1351351
ID	0	0.0000000
Day.of.the.week	0	0.0000000
Seasons	0	0.0000000
Work.load.Average.day	0	0.0000000

Figure 2.3 : Column wise missing value count before imputing 0s

	NA_count	NA_Percent
Reason.for.absence	46	6.2162162
Body.mass.index	31	4.1891892
Absenteeism.time.in.hours	22	2.9729730
Height	14	1.8918919
Education	10	1.3513514
Transportation.expense	7	0.9459459
Hit.target	6	0.8108108
Disciplinary.failure	6	0.8108108
Son	6	0.8108108
Month.of.absence	4	0.5405405
Social.smoker	4	0.5405405
Distance.from.Residence.to.Work	3	0.4054054
Service.time	3	0.4054054
Age	3	0.4054054
Social.drinker	3	0.4054054
Pet	2	0.2702703
Weight	1	0.1351351
ID	0	0.0000000
Day.of.the.week	0	0.0000000
Seasons	0	0.0000000
Work.load.Average.day	0	0.0000000

Figure 2.3 : Column wise missing value count after imputing 0s

2.1.2 Outlier Analysis

On observing the plots Figure 2.1 and Figure 2.2 most variables are found to be skewed. The skew in the distribution shows the presence of outliers and extreme values in the data.

Boxplots are used to visualize the outliers present in each variable. These outliers are then removed using Tukey's method.

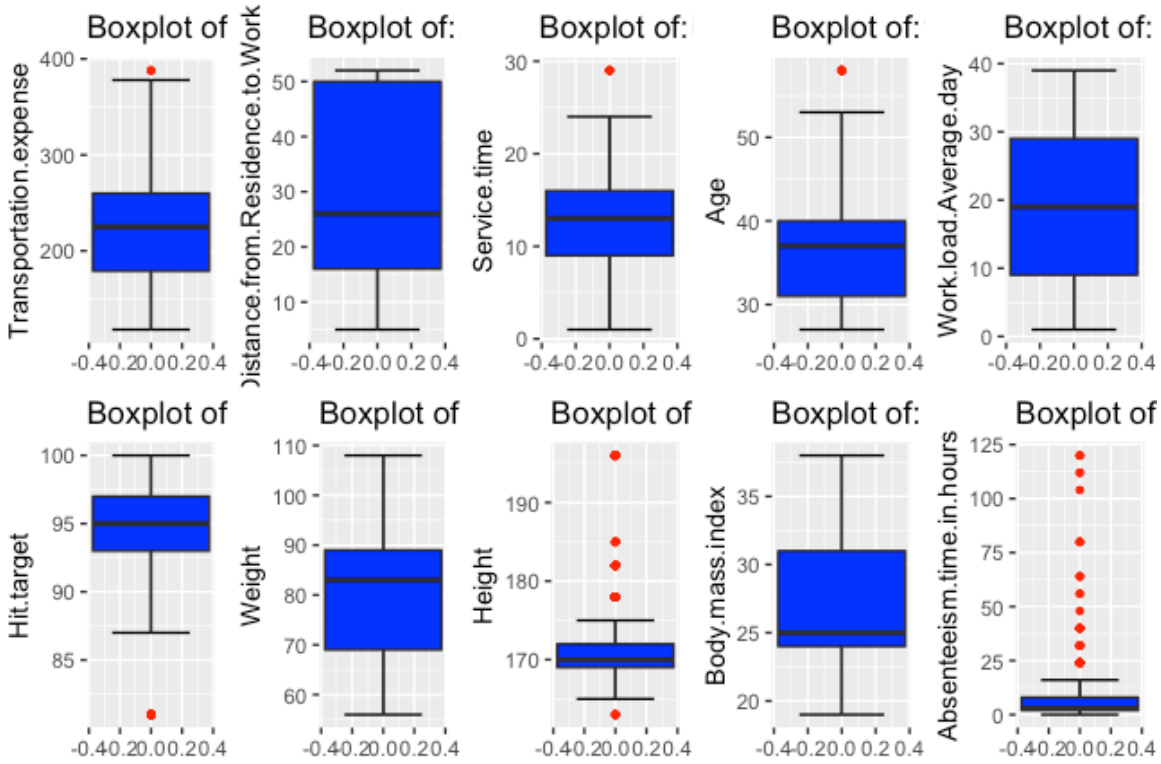

Figure 2.4 : Boxplot for outlier analysis

From observing the box plot (Figure 2.4) we see that variable height and the dependent variable absenteeism in hours have many outliers. They are detected and replaced with NA using Tukey's method. They are then imputed with the median.

2.1.3 Feature Selection

Before building the model we need to assess the significance of each predictor value and correlation between the independent variables. If multicollinearity is present, it has to be rectified. The methods we are using are as follows:

2.1.3.1.1 Correlation matrix and correlation plot

We use a correlation matrix and heat map (Figure 2.5) to find the correlation between numerical variables. Figure 2.6 shows the correlation plot of the numerical variables. From Figure 2.6 (can be viewed in the appendix) we derive that both weight and body mass index is above 5 and hence one of the predictors can be removed.

```
correlation_matrix = cor(emp_abs_data[,numeric_pred_variable_set])
corrplot(correlation_matrix,method = "number",type = 'lower')
```

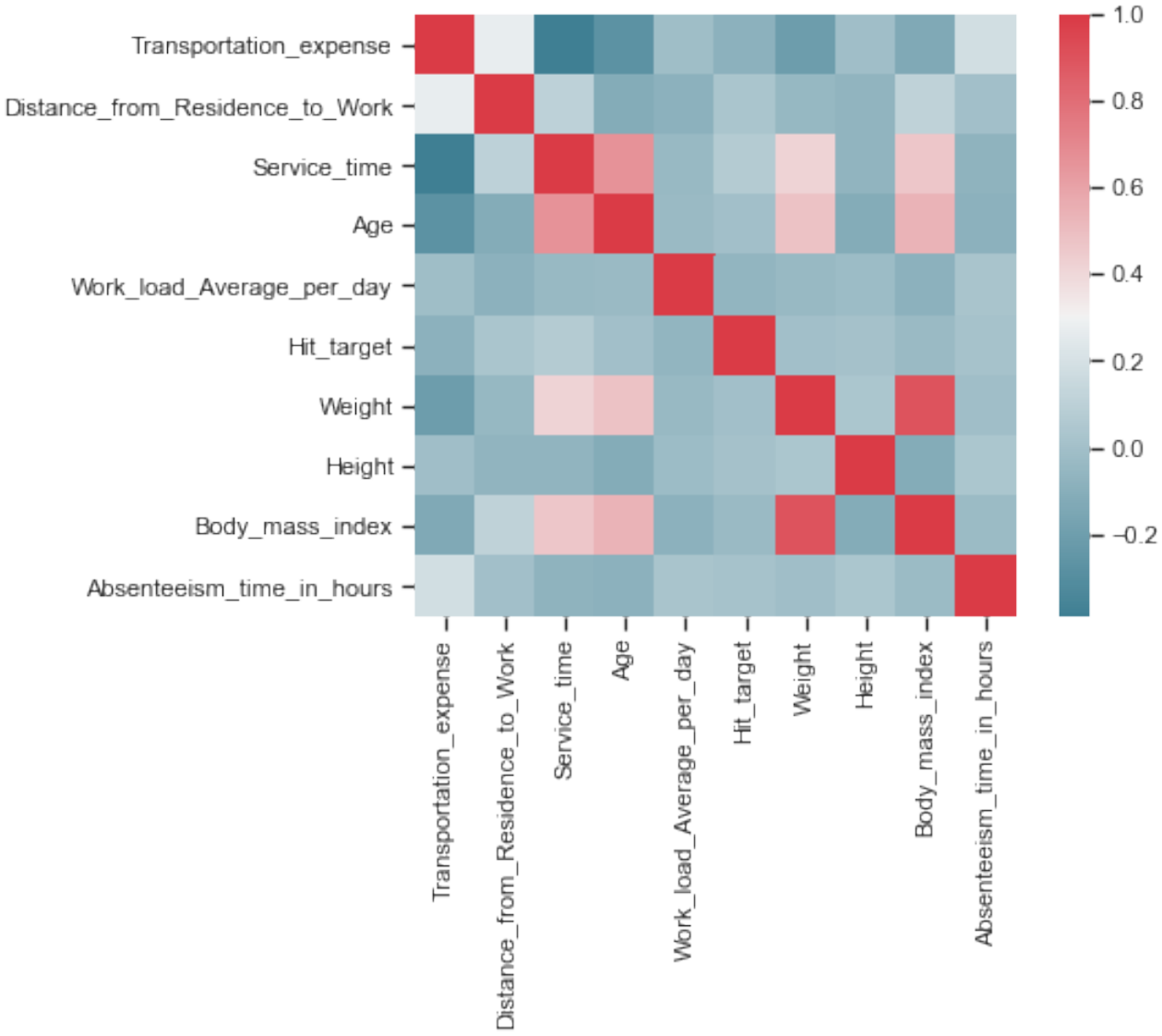

Figure 2.5 Heat map for correlation of variables

2.1.3.2 VIF

To detect multicollinearity, we use VIF for numeric variables. Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is the ratio of the overall model variance to the variance of a model that includes only that single independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model. From Fig 2.7(can be viewed in the appendix) we derive that both weight and body mass index is above 5 and hence one of the predictors can be removed.

```
vif(emp_abs_data[,numeric_pred_variable_set])
```

2.1.3.3 ANOVA

To detect multicollinearity, we use VIF for numeric variables. Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means. From Fig 2.8(can be viewed in the appendix) we derive that the variables Day of the week, Seasons, Education, Social smoker, Social drinker have p-value more than the level of significance (ie.,0.05) and hence can be omitted.

```
for(i in categorical_pred_variable_set){
  print(i)
  aov_summary = summary(aov(Absenteeism.time.in.hours~emp_abs_data[,i],data =
emp_abs_data))
  print(aov_summary)
}
```

2.1.3.4 Dimentionality Reduction

From the observations from the correlation plot, VIF and ANOVA we conclude that the columns Weight, Day of the week, Seasons, Education, Social smoker, Social drinker are not significant.

```
emp_abs_data = subset(emp_abs_data, select = -(which(names(emp_abs_data) %in%
c("Weight","Day.of.the.week","Seasons","Education","Social.smoker","Social.drinker"))))
```

2.2 Modeling

2.2.1 Model Selection

As discussed in the data section, the variables can be nominal, ordinal, interval, ratio, or continuous. If the dependent variable, in our case *Absenteeism in hours*, is Nominal the only predictive analysis that we can perform is Classification, and if the dependent variable is Interval or Ratio the normal method is to do a Regression analysis or classification after binning. If the dependent variable we are dealing with is *Ordinal*, for which both classification and regression can be done because even though the Quality variable has categories, these categories have an order associated with them, which is ranks. But our dependent variable is continuous, so we use **Regression Analysis**.

2.2.3 Sampling

We split the data into 4:1 ratio ie 80 % data would be used to train and the remaining 25% data will be used to test.

```
set.seed(123)
#Splitting the train and test set in 4:1 ratio (ie., 80% training data and 20% test data)
train_index = sample(1:nrow(emp_abs_data), 0.8*nrow(emp_abs_data))
train = emp_abs_data[train_index,]
test = emp_abs_data[-train_index,]
```

2.2.3 Multiple Linear Regression

```
regressor_mlr = lm(formula = Work.load.Average.day ~ .,data = train)
y_pred_mlr = predict(regressor_mlr, newdata = test)
summary(regressor_mlr)
```

2.2.4 Decision Tree

```
regressor_dt = rpart(Absenteeism.time.in.hours ~., data = train, method = "anova")
y_pred_dt = predict(regressor_dt, newdata = test)
summary(regressor_dt)
```

From Figure 2. , it is observed that the reason for absence from 22 to 28 contributes to 65% of absenteeism. Working more than 5 hours with a lesser service period has caused 13% of absenteeism. These variables are more significant compared to the others.

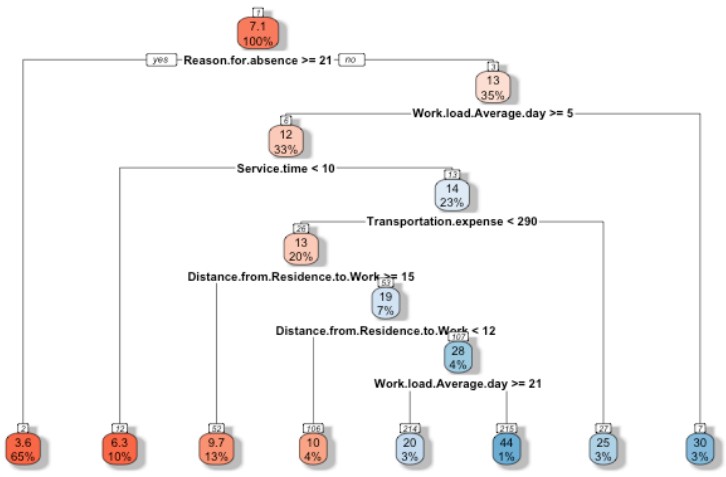

Figure 2.5 Dendogram for Decision tree

2.2.5 Random Forest

```
regressor_rf = randomForest(Absenteeism.time.in.hours~., data = train, ntrees = 500)
y_pred_rf = predict(regressor_rf,test)
summary(regressor_rf)
```

Chapter 3

Conclusion

3.1 Model Evaluation

In the previous section we obtained the results from the following models:

1. Multiple linear regression
2. Decision tree
3. Random forest

We evaluate the models based on the MAE, RMSE, and R-squared values.

Root Mean Square Error(RMSE) is the square root of the average of squared errors. **Mean Absolute Error(MAE)** of a model refers to the mean of the absolute values of each prediction error on all instances of the test data-set. **R squared** value is the percentage of the response variable variation that is explained by a linear model.

Table 3.1 shows the comparison of MAE

Language	Model	Value
Python	Multiple linear regression	
	Decision Tree	
	Random forest	
R	Multiple linear regression	14.8248566614
	Decision Tree	5.61130582
	Random forest	5.10285898

Table 3.2 shows the comparison of RMSE

Language	Model	Value
Python	Multiple linear regression	0.391861939284227
	Decision Tree	0.304944290275688
	Random forest	0.0272624878403135
R	Multiple linear regression	15.79761152
	Decision Tree	10.90598041
	Random forest	9.97901412

Table 3.3 show the comparison of R squared value

Language	Model	Value
Python	Multiple linear regression	0.359458047365649
	Decision Tree	0.612097637432515
	Random forest	0.996899638164159
R	Multiple linear regression	0.01206002
	Decision Tree	0.05805556
	Random forest	0.06205605

3.2 Model Selection

After comparing MAE, RMSE and R squared values, **the Random forest** model is chosen as the best fit as it has the minimum RMSE value and maximum R squared value.

3.3 Business Solution

3.3.1 Trend of absenteeism in 2011

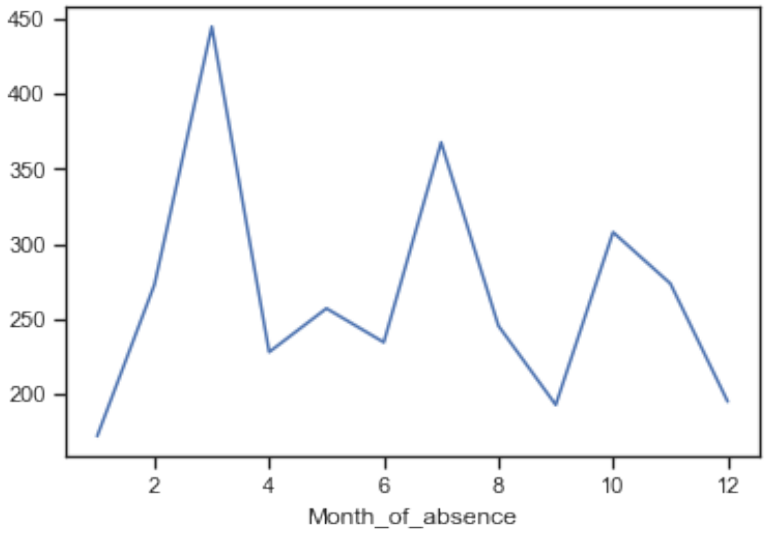

Figure 3.1 Trend of anbsenteeism in 2011

It is assumed that the trend will be the same in 2011 as now. By observing the plot it is evident that absenteeism is high in March followed by July and October. We can expect a higher work loss in March, July, and October than in the other months.

3.3.2 Measures to be taken to reduce hours of absenteeism and workload

From the previous section, we infer that the loss is higher in March and July. It could be due to the climatic conditions. According to Indian seasons, summer occurs in march and the rainy season occurs in July. The employees might face trouble traveling to the workplace. Let us plot few more bar charts to validate this assumption

3.3.2.1 Distance to work vs Absenteeism in hours and Transport expense

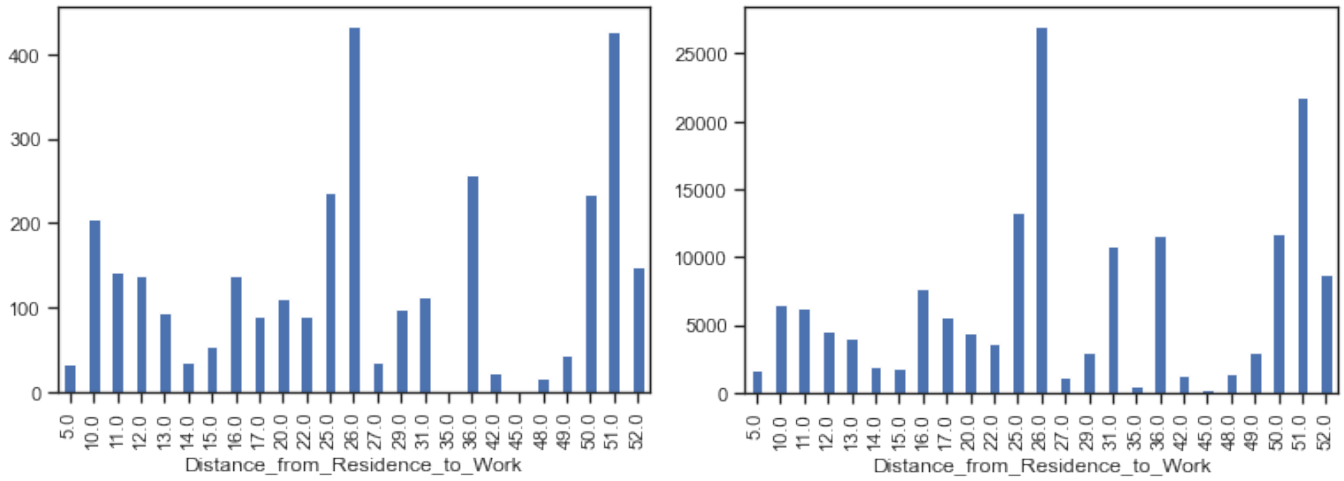

Figure 3.2 and Figure 3.3 (Distance vs Absenteeism , Distance vs Transport expense)

From the plot, we see that absenteeism is higher at 26 km. It is also observed that the transport expenses are higher in that category. Employees can be provided office vehicles to make travel convenient.

3.3.2.2 ID vs Absenteeism in hours

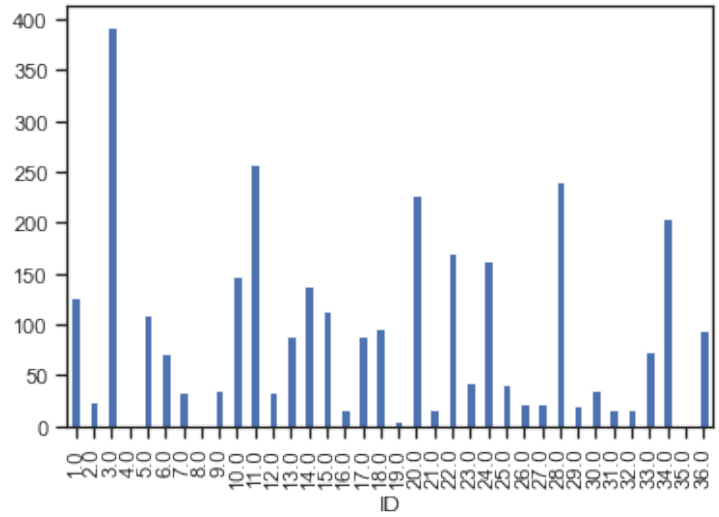

Figure 3.4 ID vs Absenteeism in hours

It is observed that id numbers 3,11,20,28 and 34 have long absent hours. Disciplinary actions can be taken if valid justification is not given.

3.3.2.2 Reason for absence vs Absenteeism in hours

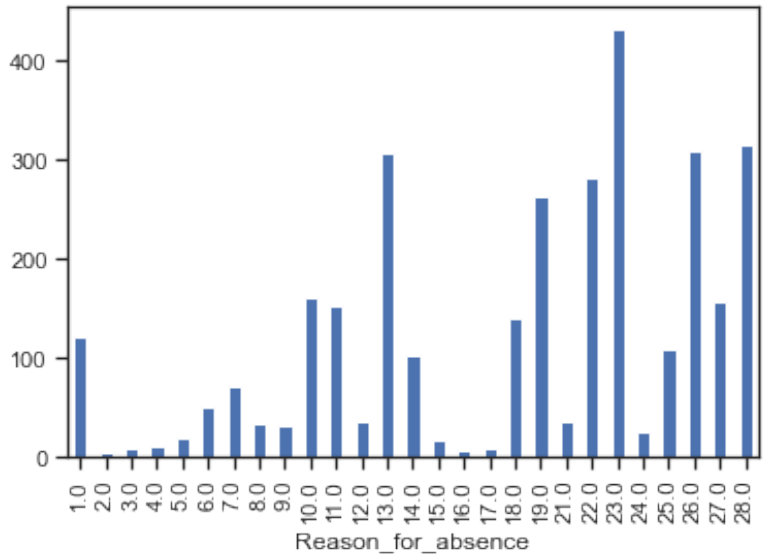

Figure 3.5 Reason of absence vs Absenteeism in hours

The primary reason for absenteeism seems to be categories 23 and 13 which are medical consultation and Diseases of the musculoskeletal system and connective tissue. This could be due

to unhealthy posture and work style. Employees can be given consultation on healthy practices while they are at work. Strategies to optimize workload among employee must be constructed.

3.3.2.2 Day of the week vs Absenteeism in hours

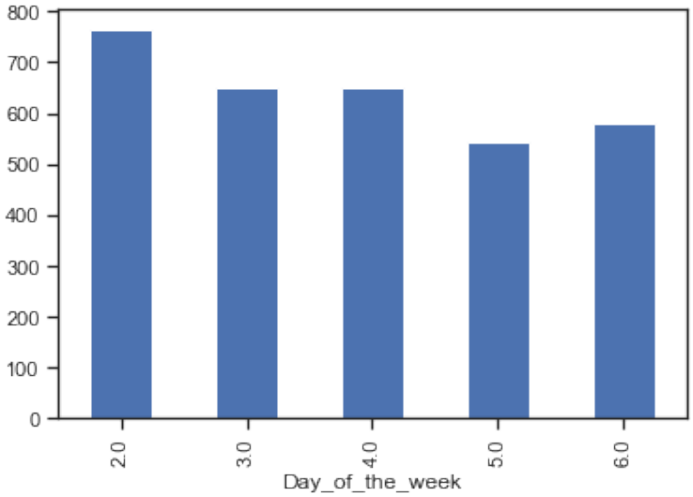

Figure 3.6 Day of week vs Absenteeism in hours

It is observed from the plot that absenteeism hours are greater on Monday which shows a lack of motivation. Employees can be encouraged with incentives and bonuses.

Appendix A - Extra Figures

Figure 2.6 Correlation plot

	Variables	VIF
1	Transportation.expense	1.374129
2	Distance.from.Residence.to.Work	1.426735
3	Service.time	2.288992
4	Age	2.299679
5	Work.load.Average.day	1.024001
6	Hit.target	1.028390
7	Weight	7.344420
8	Height	1.157580
9	Body.mass.index	8.237345

Figure 2.7 VIF

```
ID
F_onewayResult(statistic=1046.84059570791, pvalue=4.428940518145323e-174)
Reason_for_absence
F_onewayResult(statistic=3252.598691961359, pvalue=0.0)
Month_of_absence
F_onewayResult(statistic=130.60513133889617, pvalue=4.811469772568523e-29)
Day_of_the_week
F_onewayResult(statistic=8.790490943738725, pvalue=0.0030765757200355574)
Seasons
F_onewayResult(statistic=184.50092252932188, pvalue=1.0925911694626689e-39)
Disciplinary_failure
F_onewayResult(statistic=1182.937758420434, pvalue=6.045767744318007e-191)
Education
F_onewayResult(statistic=574.42081814584, pvalue=1.649394908824148e-107)
Son
F_onewayResult(statistic=641.7926847908078, pvalue=6.829638928484887e-118)
Social_drinker
F_onewayResult(statistic=898.7769490791635, pvalue=1.1525923573593214e-154)
Social_smoker
F_onewayResult(statistic=1169.878208775719, pvalue=2.300481866993278e-189)
Pet
F_onewayResult(statistic=721.9258282157014, pvalue=8.095704928790522e-130)
```

Figure 2.7 ANOVA

Appendix B - R Code

The complete R code is attached below.

```
##########################Employee Absenteeism R
Code################################
#Cleaning the environment
rm(list = ls())

#Setting the working directory
setwd("/Users/nivedharakigmail.com/Desktop/Edwisor/Project 1")

#Load libraries
libraries =
c("psych","ggpubr","corrplot","usdm","caret","rpart","randomForest","rpart.plot")
lapply(X = libraries,require, character.only = TRUE)
rm(libraries)

##Importing the csv file
#emp_abs_data_complete = read.csv(file = "Absenteeism_dataset.csv", header = T)
#Storing only predictor data
#emp_abs_data = emp_abs_data_complete[1:20]
emp_abs_data= read.csv(file = "Absenteeism_dataset.csv", header = T)

##########################Exploratory data
analysis################################
#Observing sample data
head(emp_abs_data)

#Observing structure of the data
str(emp_abs_data)
#dimnames(emp_abs_data)

##Sorting the variables into numerical and categorical
numeric_all_variable_set =
c("Transportation.expense","Distance.from.Residence.to.Work","Service.time","Age","Wo
rk.load.Average.day","Hit.target","Weight","Height","Body.mass.index","Absenteeism.tim
e.in.hours")
categorical_pred_variable_set =
c("ID","Reason.for.absence","Month.of.absence","Day.of.the.week","Seasons","Disciplin
ary.failure","Education","Son","Social.drinker","Social.smoker","Pet")
numeric_pred_variable_set =
c("Transportation.expense","Distance.from.Residence.to.Work","Service.time","Age","Wo
rk.load.Average.day","Hit.target","Weight","Height","Body.mass.index")

#Parsing the datatype of categorical variables and assigning levels
```

```
for(i in categorical_pred_variable_set){
  emp_abs_data[,i] = as.factor(emp_abs_data[,i])
}

#verifying the structure of the categorical variables
str(emp_abs_data)

#Storing the numeric and categorical data separately
#Get the data with only numeric columns
numeric_index = sapply(emp_abs_data, is.numeric)
numeric_data = emp_abs_data[,numeric_index]

#Get the data with only factor columns
factor_data = emp_abs_data[,!numeric_index]

##Copy the dataset for reference/comparison
#df = emp_abs_data

############################Missing value
Analysis#################################

###----------------------Observation--------------------------------------###
#Count of missing values in each variable
missing_values_df = data.frame(apply(emp_abs_data,2,function(x){sum(is.na(x))}))

#Creating a new column for missing value percentage
names(missing_values_df)[1] = "NA_count"
missing_values_df$NA_Percent = (missing_values_df$NA_count/
nrow(emp_abs_data))*100

#Sorting missing values in decreasing order
missing_values_df = missing_values_df[order(-missing_values_df$NA_Percent),]

###----------------------Imputation---------------------------------------###
##Imputing 0s with NA for variables
emp_abs_data$Reason.for.absence[emp_abs_data$Reason.for.absence==0] = NA
emp_abs_data$Month.of.absence[emp_abs_data$Month.of.absence==0] = NA
emp_abs_data$Day.of.the.week[emp_abs_data$Day.of.the.week==0] = NA
emp_abs_data$Seasons[emp_abs_data$Seasons==0] = NA
emp_abs_data$Education[emp_abs_data$Education==0] = NA
emp_abs_data$ID[emp_abs_data$ID==0] = NA
emp_abs_data$Age[emp_abs_data$Age==0] = NA
emp_abs_data$Weight[emp_abs_data$Weight==0] = NA
emp_abs_data$Height[emp_abs_data$Height==0] = NA
emp_abs_data$Body.mass.index[emp_abs_data$Body.mass.index==0] = NA
```

```
##Imputing missing values with default values
#Reason for absence
#emp_abs_data$Reason.for.absence[emp_abs_data$Reason.for.absence==0] = 26
emp_abs_data$Reason.for.absence[is.na(emp_abs_data$Reason.for.absence)] = 27

#Parsing the column Work load average day into categorical value
emp_abs_data$Work.load.Average.day  =
as.numeric(as.factor(as.character(emp_abs_data$Work.load.Average.day )))
##Imputing NA values with id wise mean of the column values for numeric columns
#Transportation expense
x = emp_abs_data$ID[is.na(emp_abs_data$Transportation.expense)]
for (i in x){
  emp_abs_data$Transportation.expense[is.na(emp_abs_data$Transportation.expense) &
emp_abs_data$ID==i] =
mean(emp_abs_data$Transportation.expense[emp_abs_data$ID==i],na.rm = T)
}

#Distance from Residence to Work
x = emp_abs_data$ID[is.na(emp_abs_data$Distance.from.Residence.to.Work)]
for (i in x){

emp_abs_data$Distance.from.Residence.to.Work[is.na(emp_abs_data$Distance.from.Res
idence.to.Work) & emp_abs_data$ID==i] =
mean(emp_abs_data$Distance.from.Residence.to.Work[emp_abs_data$ID==i],na.rm =
T)
}

#Service time
x = emp_abs_data$ID[is.na(emp_abs_data$Service.time)]
for (i in x){
  emp_abs_data$Service.time[is.na(emp_abs_data$Service.time) &
emp_abs_data$ID==i] =
mean(emp_abs_data$Service.time[emp_abs_data$ID==i],na.rm = T)
}

#Age
x = emp_abs_data$ID[is.na(emp_abs_data$Age)]
for (i in x){
  emp_abs_data$Age[is.na(emp_abs_data$Age) & emp_abs_data$ID==i] =
mean(emp_abs_data$Age[emp_abs_data$ID==i],na.rm = T)
}

#Weight
x = emp_abs_data$ID[is.na(emp_abs_data$Weight)]
```

```
for (i in x){
  emp_abs_data$Weight[is.na(emp_abs_data$Weight) & emp_abs_data$ID==i] =
mean(emp_abs_data$Weight[emp_abs_data$ID==i],na.rm = T)
}

#Height
x = emp_abs_data$ID[is.na(emp_abs_data$Height)]
for (i in x){
  emp_abs_data$Height[is.na(emp_abs_data$Height) & emp_abs_data$ID==i] =
mean(emp_abs_data$Height[emp_abs_data$ID==i],na.rm = T)
}

##Imputing NA values with id wise mode of the column values for categorical columns
#Defining mode function
getmode = function(x){
  unique_x = unique(x)
  mode_val = which.max(tabulate(match(x,unique(x))))
}

#Reason for absence
x = emp_abs_data$ID[is.na(emp_abs_data$Reason.for.absence)]
for (i in x){
  emp_abs_data$Reason.for.absence[is.na(emp_abs_data$Reason.for.absence) &
emp_abs_data$ID==i] =
getmode(emp_abs_data$Reason.for.absence[emp_abs_data$ID==i])
}

#Education
x = emp_abs_data$ID[is.na(emp_abs_data$Education)]
for (i in x){
  emp_abs_data$Education[is.na(emp_abs_data$Education) & emp_abs_data$ID==i] =
getmode(emp_abs_data$Education[emp_abs_data$ID==i])
}

#Son
x = emp_abs_data$ID[is.na(emp_abs_data$Son)]
for (i in x){
  emp_abs_data$Son[is.na(emp_abs_data$Son) & emp_abs_data$ID==i] =
getmode(emp_abs_data$Son[emp_abs_data$ID==i])
}

#Social drinker
x = emp_abs_data$ID[is.na(emp_abs_data$Social.drinker)]
for (i in x){
```

```
  emp_abs_data$Social.drinker[is.na(emp_abs_data$Social.drinker) &
emp_abs_data$ID==i] =
getmode(emp_abs_data$Social.drinker[emp_abs_data$ID==i])
}

#Social smoker
x = emp_abs_data$ID[is.na(emp_abs_data$Social.smoker)]
for (i in x){
  emp_abs_data$Social.smoker[is.na(emp_abs_data$Social.smoker) &
emp_abs_data$ID==i] =
getmode(emp_abs_data$Social.smoker[emp_abs_data$ID==i])
}

#Pet
x = emp_abs_data$ID[is.na(emp_abs_data$Pet)]
for (i in x){
  emp_abs_data$Pet[is.na(emp_abs_data$Pet) & emp_abs_data$ID==i] =
getmode(emp_abs_data$Pet[emp_abs_data$ID==i])
}

##Imputing NA BMI values using height and weight
#Body mass index
x = emp_abs_data$ID[is.na(emp_abs_data$Body.mass.index)]
for (i in x){
  emp_abs_data$Body.mass.index[is.na(emp_abs_data$Body.mass.index) &
emp_abs_data$ID==i] =
round(((emp_abs_data$Weight[emp_abs_data$ID==i])*10000)/
((emp_abs_data$Height[emp_abs_data$ID==i])**2))
}

##Imputing remaining NA values with median for numerical data and mode for
categorical data
#Numerical
#Absenteeism.time.in.hours
x = emp_abs_data$ID[is.na(emp_abs_data$Absenteeism.time.in.hours)]
for (i in x){

emp_abs_data$Absenteeism.time.in.hours[is.na(emp_abs_data$Absenteeism.time.in.hour
s) & emp_abs_data$ID==i] = median(emp_abs_data$Absenteeism.time.in.hours,na.rm
= T)
}
#Hit.target
x = emp_abs_data$ID[is.na(emp_abs_data$Hit.target)]
for (i in x){
```

```
  emp_abs_data$Hit.target[is.na(emp_abs_data$Hit.target) & emp_abs_data$ID==i] =
median(emp_abs_data$Hit.target,na.rm = T)
}

#Categorical
#Disciplinary.failure
x = emp_abs_data$ID[is.na(emp_abs_data$Disciplinary.failure)]
for (i in x){
  emp_abs_data$Disciplinary.failure[is.na(emp_abs_data$Disciplinary.failure) &
emp_abs_data$ID==i] = getmode(emp_abs_data$Disciplinary.failure)
}

#Reason.for.absence
x = emp_abs_data$ID[is.na(emp_abs_data$Reason.for.absence)]
for (i in x){
  emp_abs_data$Reason.for.absence[is.na(emp_abs_data$Reason.for.absence) &
emp_abs_data$ID==i] = getmode(emp_abs_data$Reason.for.absence)
}

#Month.of.absence
x = emp_abs_data$ID[is.na(emp_abs_data$Month.of.absence)]
for (i in x){
  emp_abs_data$Month.of.absence[is.na(emp_abs_data$Month.of.absence) &
emp_abs_data$ID==i] = getmode(emp_abs_data$Month.of.absence)
}

#Checking for missing values
sum(is.na(emp_abs_data))

############################Analysis of data distribution using histogram#########
#Storing the numeric and categorical data separately
#Get the data with only numeric columns
numeric_index = sapply(emp_abs_data, is.numeric)
numeric_data = emp_abs_data[,numeric_index]

#Get the data with only factor columns
factor_data = emp_abs_data[,!numeric_index]
#Multiple histograms for numerical predictors
emp_data_plot_hist = numeric_data[1:9]
multi.hist(emp_data_plot_hist,dcol= c("blue","red"),dlty=c("solid",
"solid"),bcol="linen")

#Multiple barplot
#Bar plot for categorically predictors
emp_data_plot_bar = factor_data[1:11]
```

```
gplot1 = ggplot(emp_data_plot_bar, aes(x = ID ) )+ geom_bar()
gplot2 = ggplot(emp_data_plot_bar, aes(x = Reason.for.absence ) )+ geom_bar()
gplot3 = ggplot(emp_data_plot_bar, aes(x = Month.of.absence ) )+ geom_bar()
gplot4 = ggplot(emp_data_plot_bar, aes(x = Day.of.the.week ) )+ geom_bar()
gplot5 = ggplot(emp_data_plot_bar, aes(x = Seasons ) )+ geom_bar()
gplot6 = ggplot(emp_data_plot_bar, aes(x = Disciplinary.failure ) )+ geom_bar()
gplot7 = ggplot(emp_data_plot_bar, aes(x = Education ) )+ geom_bar()
gplot8 = ggplot(emp_data_plot_bar, aes(x = Son ) )+ geom_bar()
gplot9 = ggplot(emp_data_plot_bar, aes(x = Social.drinker ) )+ geom_bar()
gplot10 = ggplot(emp_data_plot_bar, aes(x = Social.smoker ) )+ geom_bar()
gplot11 = ggplot(emp_data_plot_bar, aes(x = Pet ) )+ geom_bar()
ggarrange(gplot1,gplot2,gplot3,gplot4,gplot5,gplot6,gplot7,gplot8,gplot8,gplot9,gplot10,
gplot11)

############################Outlier
Analysis#########################################
#Outlier boxplot
for(i in 1:ncol(numeric_data)) {
  assign(paste0("box_plot",i), ggplot(data = emp_abs_data, aes_string(y =
numeric_data[,i])) +
         stat_boxplot(geom = "errorbar", width = 0.5) +
         geom_boxplot(outlier.colour = "red", fill = "blue", outlier.size = 1) +
         labs(y = colnames(numeric_data[i])) +
         ggtitle(paste("Boxplot of: ",colnames(numeric_data[i]))))
}

#Arrange the plots in grids
gridExtra::grid.arrange(box_plot1,box_plot2,box_plot3,box_plot4,box_plot5,
                        box_plot6,box_plot7,box_plot8,box_plot9,box_plot10,ncol=5)

#Replacing all outliers with NA
for(i in numeric_pred_variable_set){
  val1 = emp_abs_data[,i][emp_abs_data[,i] %in% boxplot.stats(emp_abs_data[,i])$out]
  print(paste(i,length(val1)))
  emp_abs_data[,i][emp_abs_data[,i] %in% val1] = NA
}

#Checking for missing values
sum(is.na(emp_abs_data))

#Imputing data in columns with NA by median method
impute_median_mode = function(data_set){
  for(i in colnames(data_set)){
    if(sum(is.na(data_set[,i]))!=0){
      if(is.numeric(data_set[,i])){
```

```
      data_set[is.na(data_set[,i]),i] = median(data_set[,i],na.rm = TRUE)
    }


    else if(is.factor(data_set[,i])){

      data_set[is.na(data_set[,i]),i] = getmode(data_set[,i])
    }

  }
 }
 #print(data_set)
 return(data_set)
}
emp_abs_data = impute_median_mode(emp_abs_data)
#Checking for missing values
sum(is.na(emp_abs_data))

############################Feature
Selection#########################################
correlation_matrix = cor(emp_abs_data[,numeric_pred_variable_set])
# dev.off()
corrplot(correlation_matrix,method = "number",type = 'lower')

#Checking multi-collinearity
#Using VIF technique for numerical data
vif(emp_abs_data[,numeric_pred_variable_set])
#Using ANOVA technique for categorical data
for(i in categorical_pred_variable_set){
 print(i)
 aov_summary = summary(aov(Absenteeism.time.in.hours~emp_abs_data[,i],data =
emp_abs_data))
 print(aov_summary)
}

#Dimentionality reduction
emp_abs_data_all_columns = emp_abs_data
#From the observations obtained from from VIF and ANOVA test,we derive the following
inference
#From VIF both weight and body mass index are above 5. Either of the predictor can be
removed
#From ANOVA the following columns had p-value greater than the level of
significance(ie.,0.05) and hence can be removed
```

```
emp_abs_data = subset(emp_abs_data, select = -(which(names(emp_abs_data) %in%
c("Weight","Day.of.the.week","Seasons","Education","Social.smoker","Social.drinker")))
)

##############################Feature
Sampling#########################################
variable_set =
c("Transportation.expense","Distance.from.Residence.to.Work","Service.time","Age","Wo
rk.load.Average.day","Hit.target","Height","Body.mass.index","Absenteeism.time.in.hour
s","ID","Reason.for.absence","Month.of.absence","Disciplinary.failure","Son","Pet")
#Parsing all the columns to numeric value
for(i in variable_set){
  emp_abs_data[,i] = as.numeric(emp_abs_data[,i])
}
#Separating dataset into test and train set
set.seed(123)
#Splitting the train and test set in 4:1 ratio (ie., 80% training data and 20% test data)
train_index = sample(1:nrow(emp_abs_data), 0.8*nrow(emp_abs_data))
train = emp_abs_data[train_index,]
test = emp_abs_data[-train_index,]

################################## Model Development
##################################

###----------------Multiple linear regression-------------------------------###
#Train the model using training data
# Fitting Multiple Linear Regression to the Training set
regressor_mlr = lm(formula = Work.load.Average.day ~ .,data = train)
# Predicting the Test set results
y_pred_mlr = predict(regressor_mlr, newdata = test)
#Get the summary of the model
summary(regressor_mlr)
#Create dataframe for actual and predicted values
model_pred = data.frame("actual"=test, "model_pred"=y_pred_mlr)
head(model_pred)
#Calcuate MAE, RMSE, R-squared for testing data
print(postResample(pred = y_pred_mlr, obs =test$Absenteeism.time.in.hours))
#Plot a graph for actual vs predicted values
plot(test$Absenteeism.time.in.hours,type="l",lty=2,col="red")
lines(y_pred_mlr,col="blue")

###-------------------Decision Tree---------------------------------------###
#Build decsion tree using rpart
regressor_dt = rpart(Absenteeism.time.in.hours ~., data = train, method = "anova")
```

```
#Predict the test cases
y_pred_dt = predict(regressor_dt, newdata = test)

#Create data frame for actual and predicted values
model_pred = cbind(model_pred,y_pred_dt)
head(model_pred)

#Get the summary of the model
summary(regressor_dt)

#Calcuate MAE, RMSE, R-squared for testing data
print(postResample(pred = y_pred_dt, obs = test$Absenteeism.time.in.hours))

#Plot a graph for actual vs predicted values
plot(test$Absenteeism.time.in.hours,type="l",lty=2,col="red")
lines(y_pred_dt,col="blue")

# Visualize the decision tree with rpart.plot
rpart.plot(regressor_dt, box.palette="RdBu", shadow.col="gray", nn=TRUE)

###----------Random forest----------###
#Train the model using training data
regressor_rf = randomForest(Absenteeism.time.in.hours~., data = train, ntrees = 500)

#Predict the test cases
y_pred_rf = predict(regressor_rf,test)

#Create dataframe for actual and predicted values
model_pred = cbind(model_pred,y_pred_rf)
head(model_pred)

#Get the summary of the model
summary(regressor_rf)

#Calcuate MAE, RMSE, R-squared for testing data
print(postResample(pred = y_pred_rf, obs = test$Absenteeism.time.in.hours))

#Plot a graph for actual vs predicted values
plot(test$Absenteeism.time.in.hours,type="l",lty=2,col="red")
lines(y_pred_rf,col="blue")

################################## Model Inference
##################################
#The trend in absenteesim in hours is assumed to be same as this year in 2011
#Plotting the Absenteeism trend
```

References

[1] https://rdrr.io

[2] https://www.rdocumentation.org

[3] https://edwisor.com/home

[4] https://www.udemy.com

[5] https://docs.python.org/3.9/