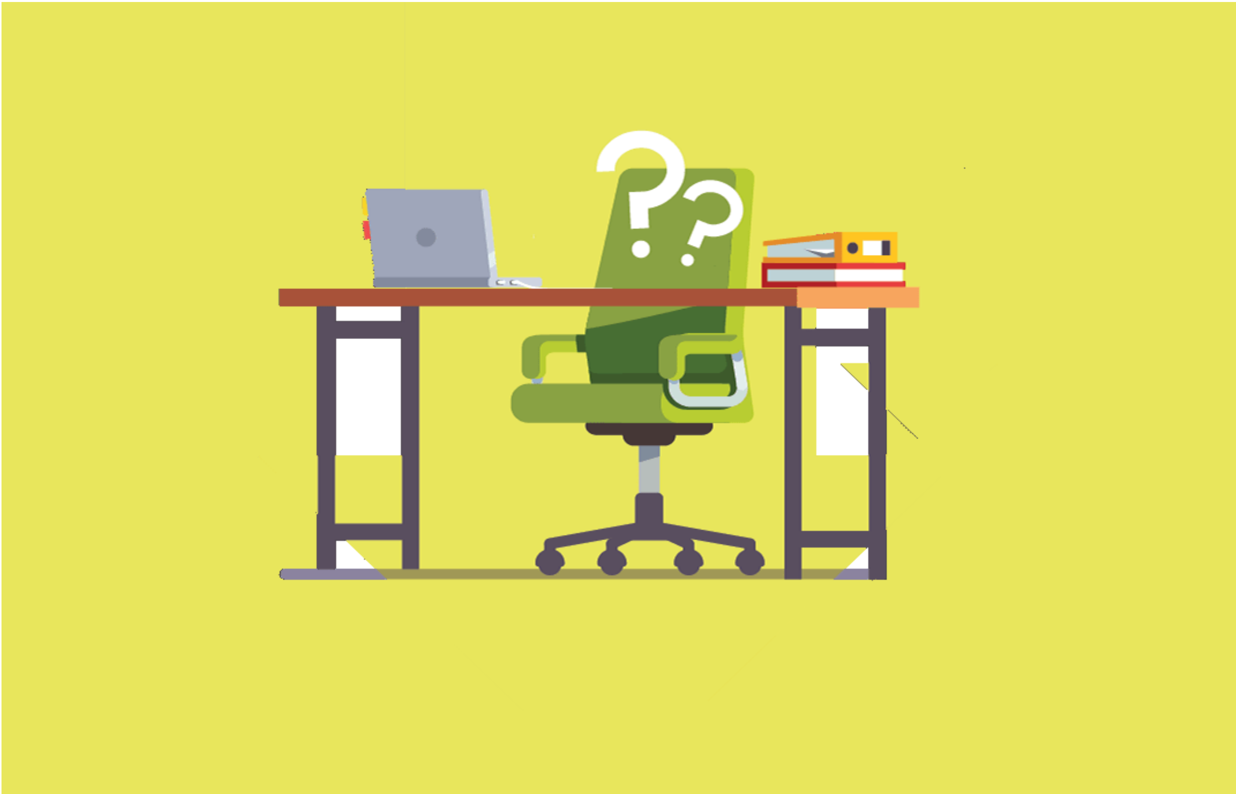


Employee Absenteeism

11-Oct-2019

Abhishek singh pushkar

Contents

Introduction

Problem Statement

The objective of this case is to define the changes for the company to reduce the number of Absenteeism. And to calculate how much losses will the company has if the trend remains the same. Also predicting future number of Absenteeism.

Data

The data given has 740 rows and 21 columns which includes

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
11	26.0	7.0	3	1	289.0	36.0	13.0	33.0	239554.0
36	0.0	7.0	3	1	118.0	13.0	18.0	50.0	239554.0
3	23.0	7.0	4	1	179.0	51.0	18.0	38.0	239554.0
7	7.0	7.0	5	1	279.0	5.0	14.0	39.0	239554.0
11	23.0	7.0	5	1	289.0	36.0	13.0	33.0	239554.0

Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
97.0	0.0	1.0	2.0	1.0	0.0	1.0	90.0	172.0	30.0	4.0
97.0	1.0	1.0	1.0	1.0	0.0	0.0	98.0	178.0	31.0	0.0
97.0	0.0	1.0	0.0	1.0	0.0	0.0	89.0	170.0	31.0	2.0
97.0	0.0	1.0	2.0	1.0	1.0	0.0	68.0	168.0	24.0	4.0
97.0	0.0	1.0	2.0	1.0	0.0	1.0	90.0	172.0	30.0	2.0

Here our target variable is Absenteeism time in hours. It is time in hours of employee absence. Reason for absence is the reason employee gave and the other variable represents the given name.

Methodology

Data preparation

We need to change the numerical value to the categorical value of the variable

Reason for absence (ICD):
1. Certain infectious and parasitic diseases
2. Neoplasms
3. Diseases of the blood and blood-forming organs and immune mechanism disorders
4. Endocrine, nutritional and metabolic diseases
5. Mental and behavioral disorders
6. Diseases of the nervous system
7. Diseases of the eye and adnexa
8. Diseases of the ear and mastoid process
9. Diseases of the circulatory system
10. Diseases of the respiratory system
11. Diseases of the digestive system
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Diseases of the genitourinary system
15. Pregnancy, childbirth and the puerperium
16. Certain conditions originating in the perinatal period
17. Congenital malformations, deformations, and chromosomal abnormalities
18. Symptoms, signs and abnormal clinical and laboratory findings, not classified
19. Injury, poisoning and certain other consequences of external causes
20. External causes of morbidity and mortality
21. Factors influencing health status and contact with health services.

Seasons:

1. Summer
2. autumn
3. Winter
4. spring

Education :

1. high school
2. Graduate
3. postgraduate
4. master and doctor

Exploratory data analysis

Distribution of a categorical variable with the target variable

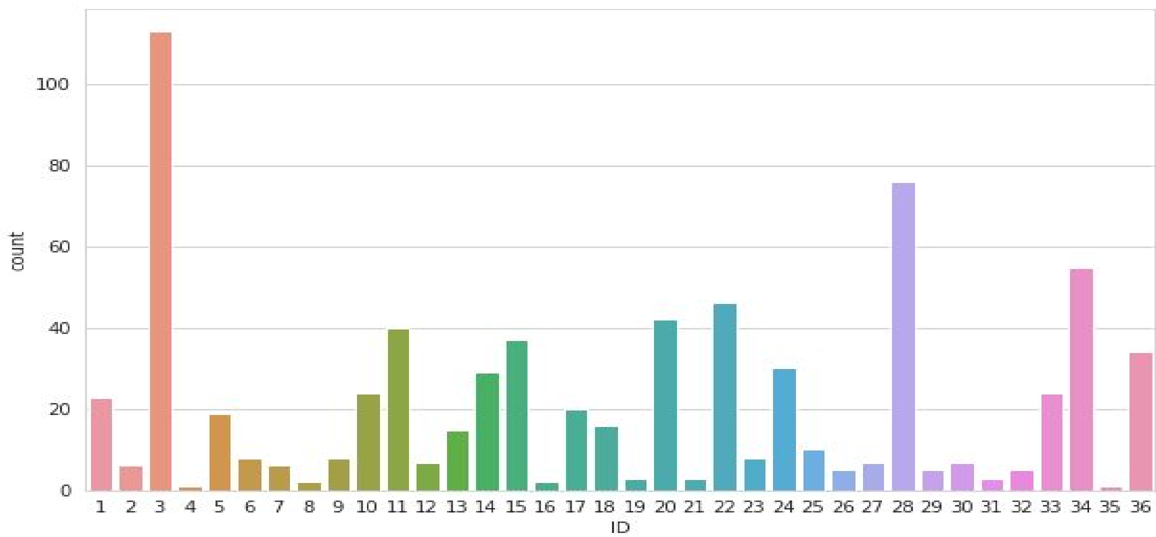


Fig. Bar graph showing the Frequency to reasons given by employees.

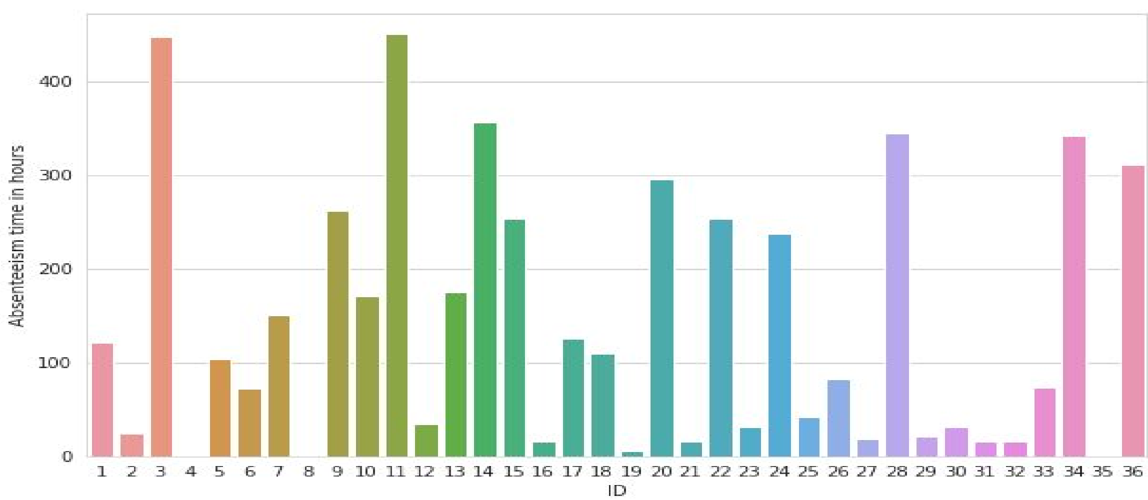


Fig. Bar graph showing total no. of absence hours reason wise

From the above graphs we can conclude that employees with id 3 ,11,14, 28 and 34 were mostly absent , from which employee 3 took frequent but 1 to 3 hours leave and employee 11 took less but ling hours leave

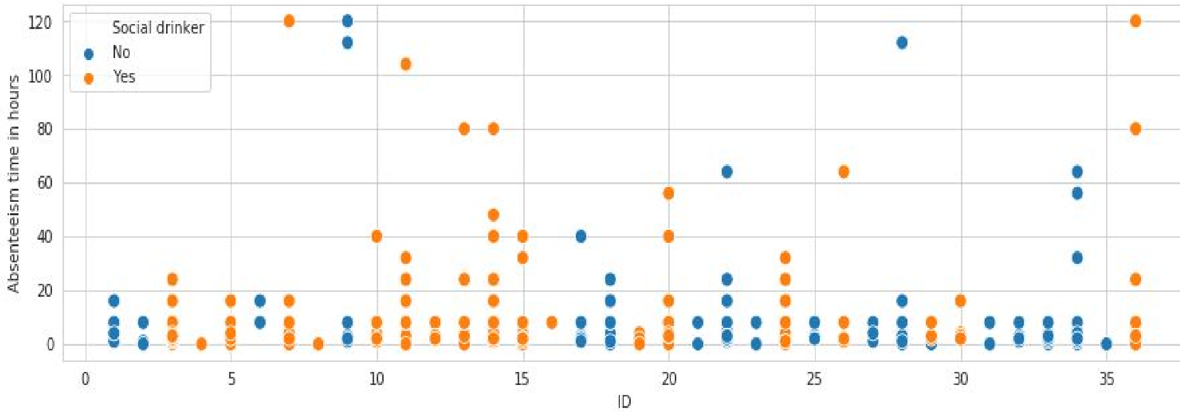


Fig. Scatter Plot showing absence of each employee with hours of absence and drinking habit

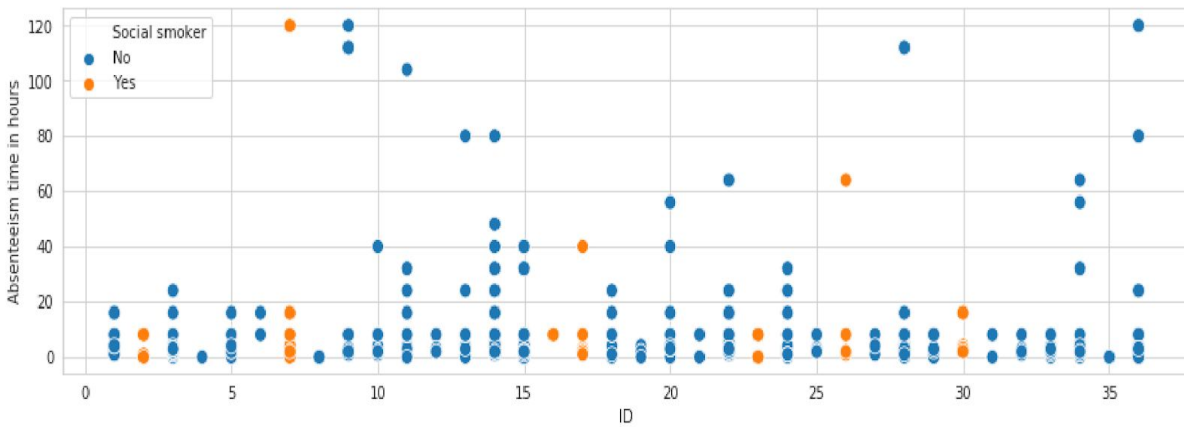


Fig.Scatter plot showing each leaves taken by each employee and their smoking habit

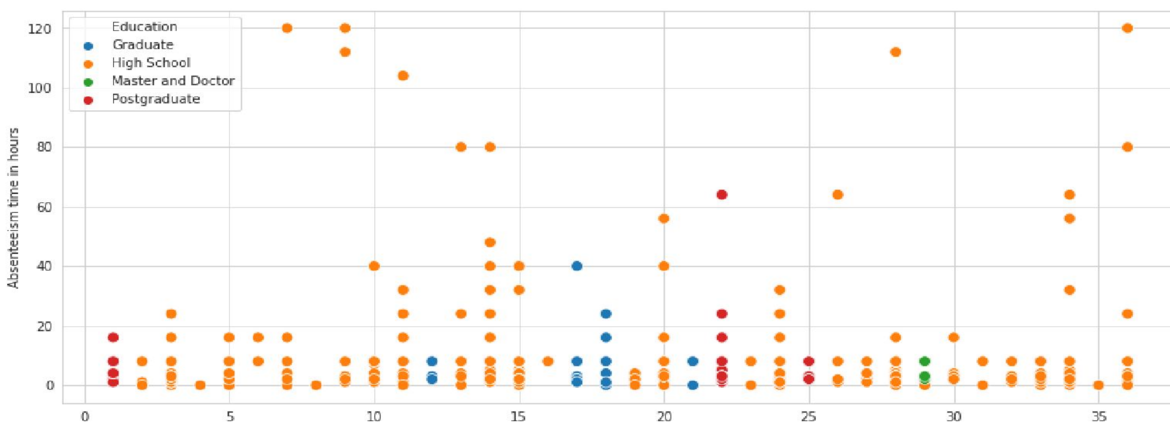


Fig. Scatter plot showing each leave took by each employee and their Educational qualification

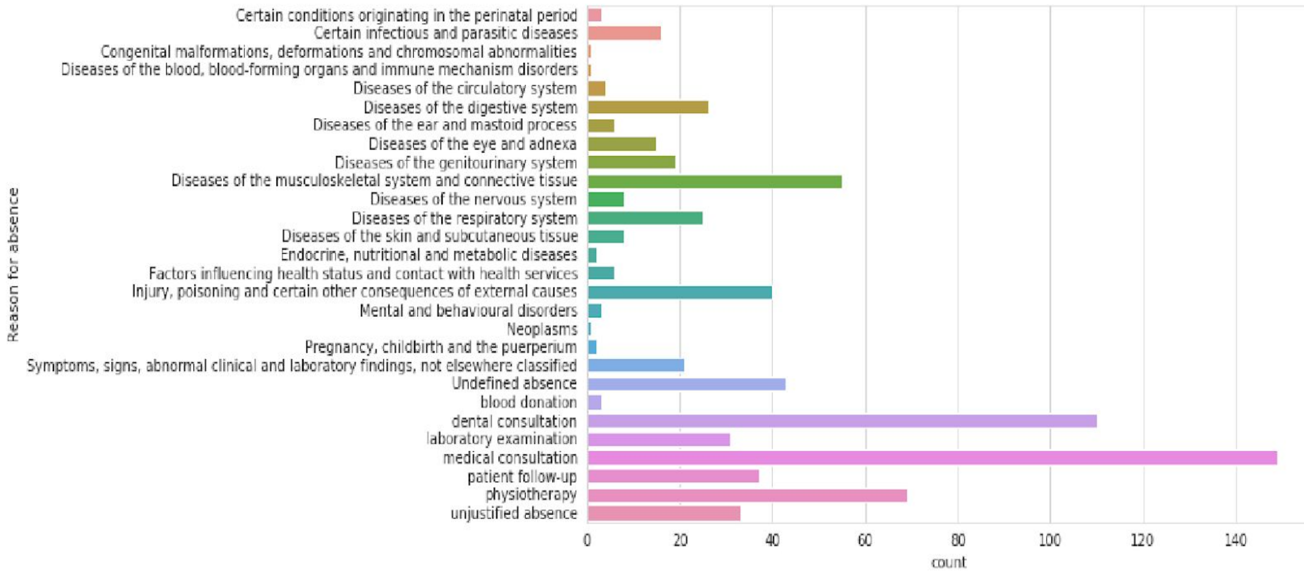


Fig. Bar graph showing the Frequency to reasons given by employees for absence

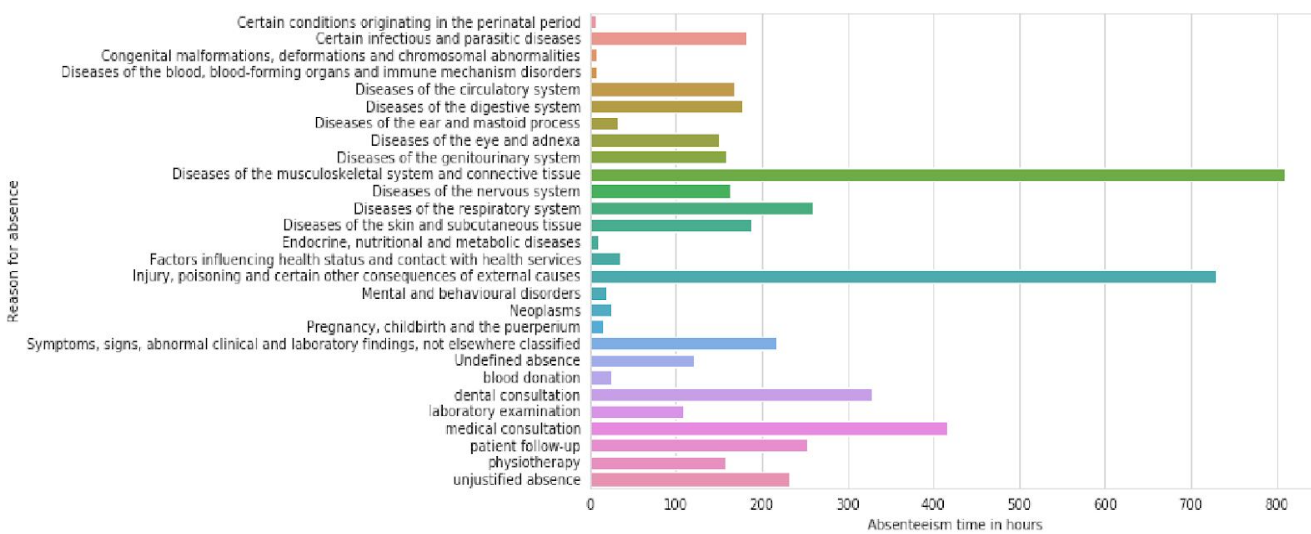


Fig. Bar graph showing the number of absence hours per reason

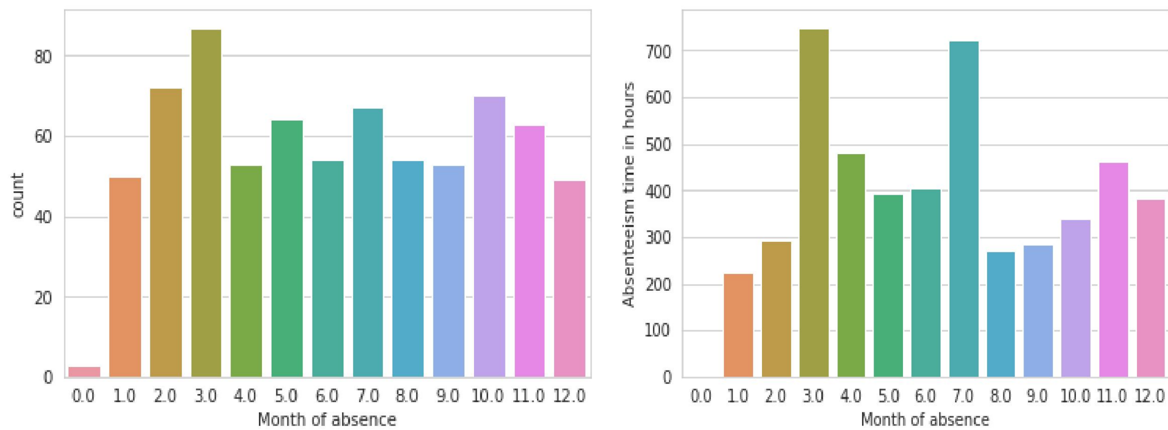


Fig. Bar graphs showing Month wise frequency of absence and total no. hours absence

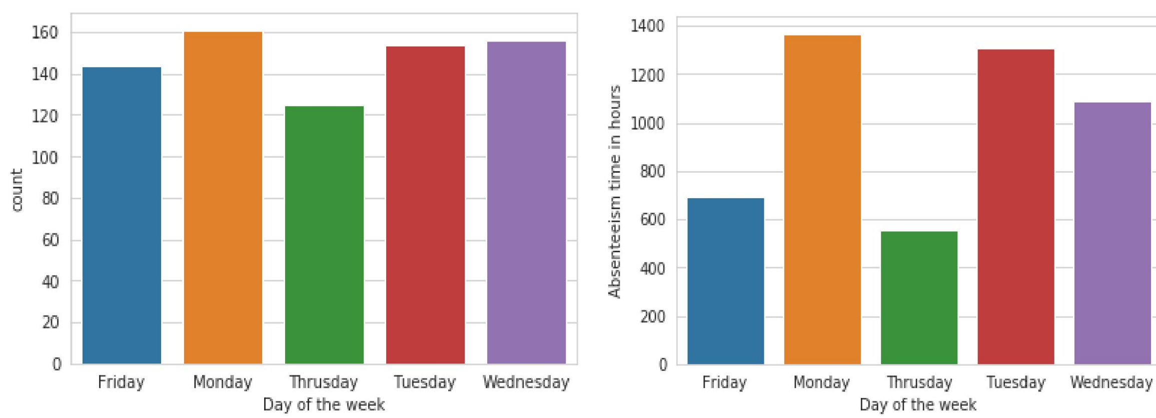


Fig.Bar graphs showing weekday frequency of absence and total no. hours absence

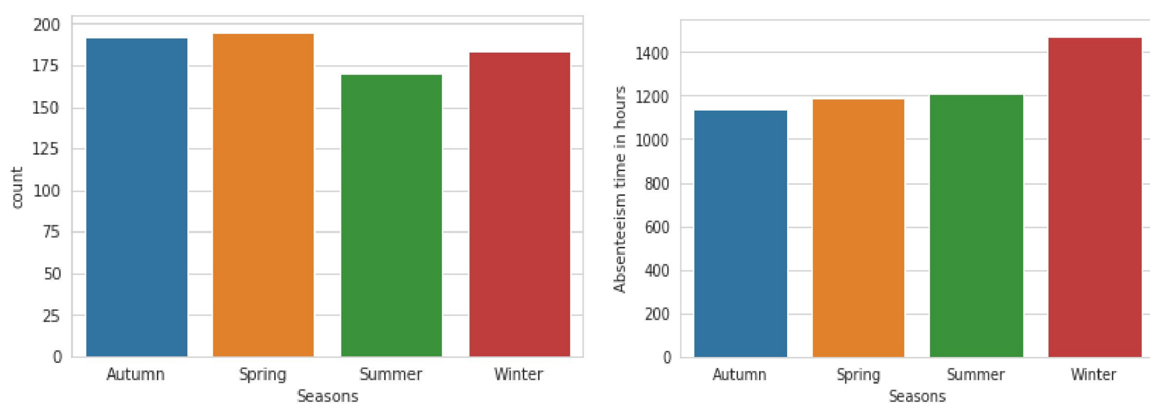


Fig. Bar graphs showing Season wise frequency of absence and total no. hours of absence

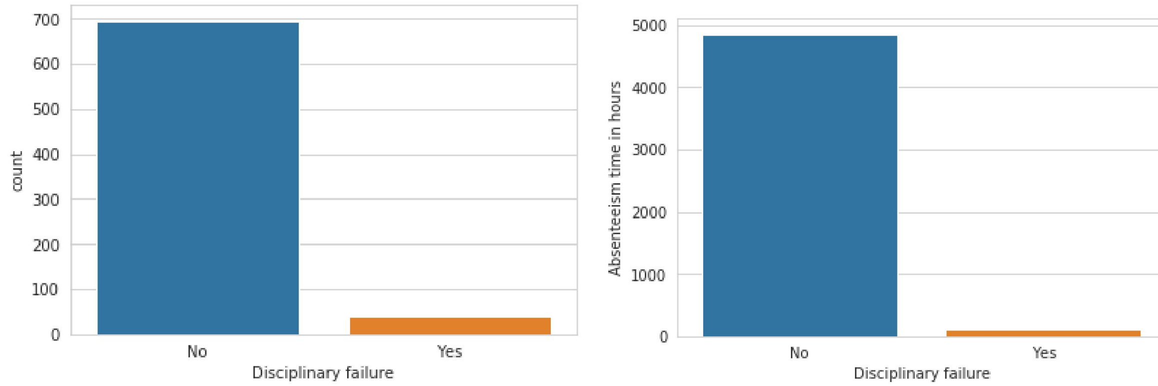


Fig. Bar graphs showing frequency of employee being absent based on their disciplinary status and total no. of hours absent by disciplinary category.

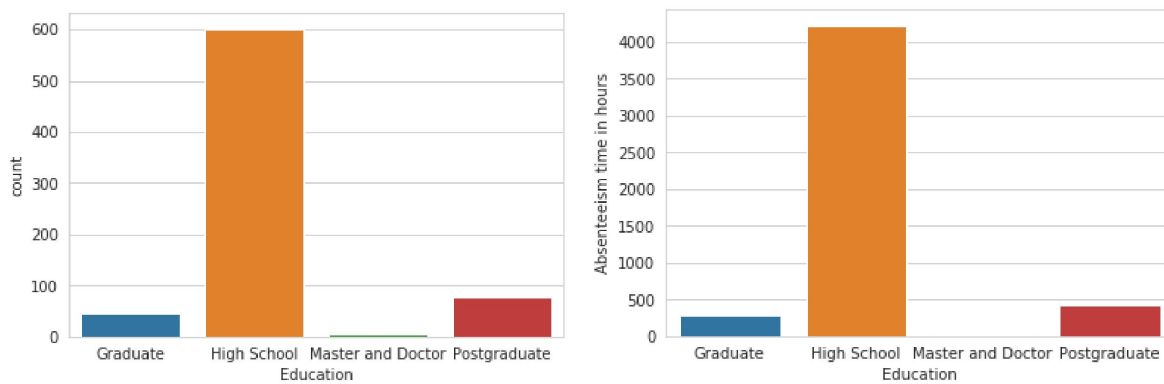


Fig. Bar graphs showing employees with what educational qualifications are mostly absent. It is clear employees with high school degree took significant no. of leaves.

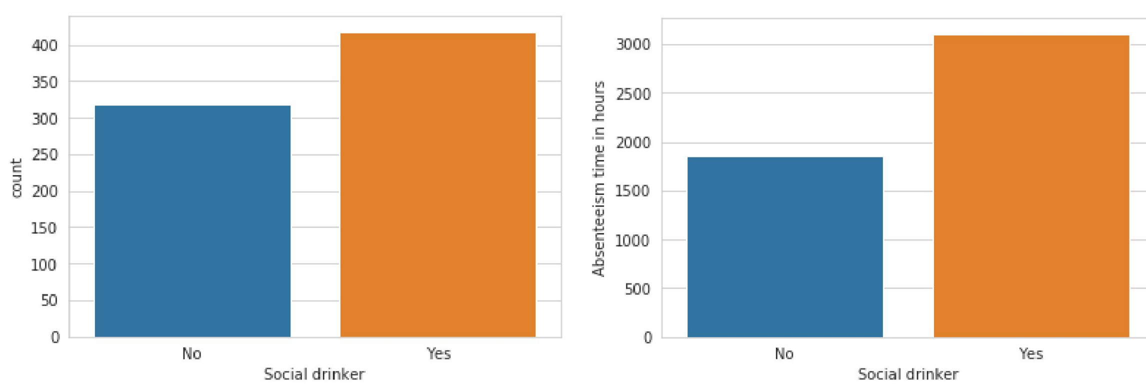


Fig. Bar graphs showing the frequency of leaves and total no of hours absent depending on the drinking habit of the employee

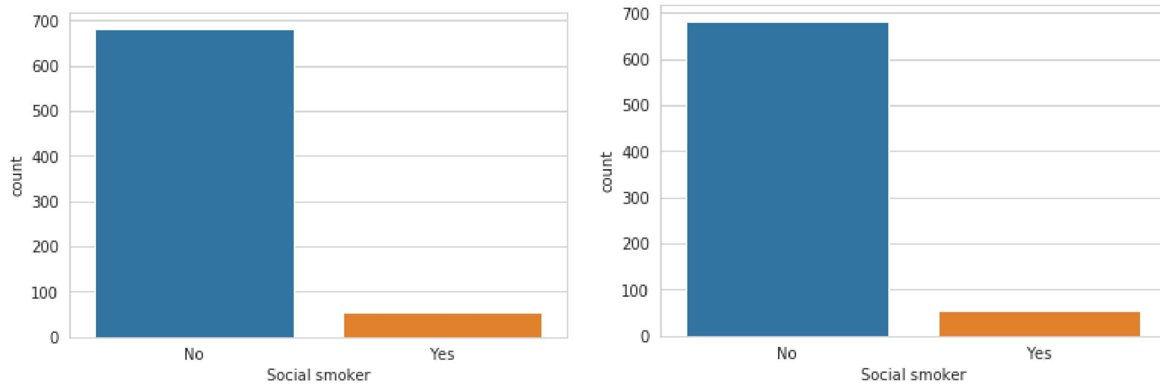


Fig. Bar graphs showing frequency of leaves and total no of hours absent depending on the Smoking habit of the employee. As we can see there is very less no. of people in the company who smokes.

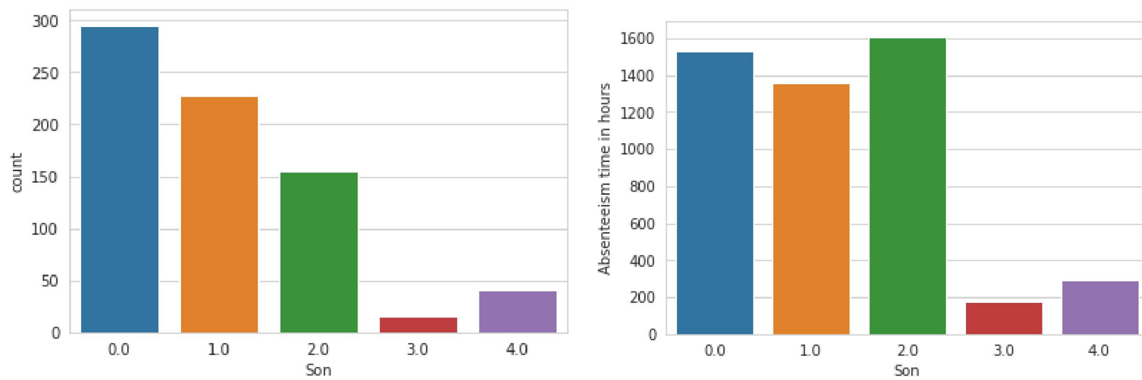


Fig. Bar graphs showing frequency of absence and total no. hours absent depending on the number of son employees have. People with 1 son took consistent hours of leave but employee with two son took long hours leave but not frequent.

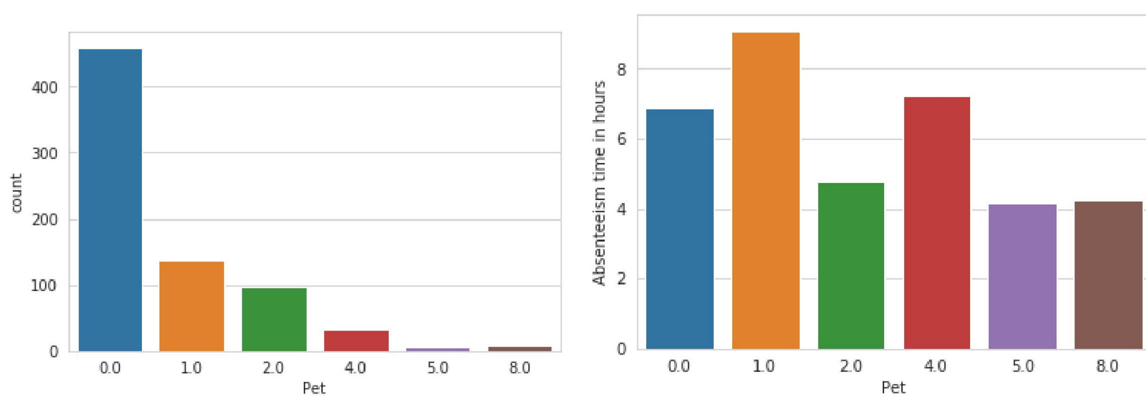


Fig. Bar graphs showing frequency and total no. hours absent depending on. Number pet employees have.

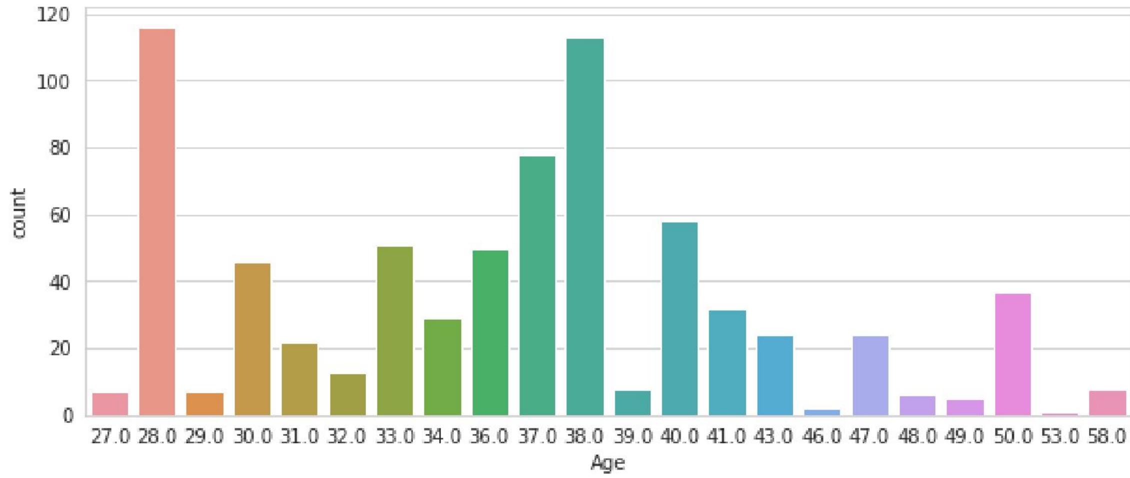


Fig. Bar graphs showing frequency of leaves applied age wise.Mostly employees of age 28 and 38 took more number of leaves.

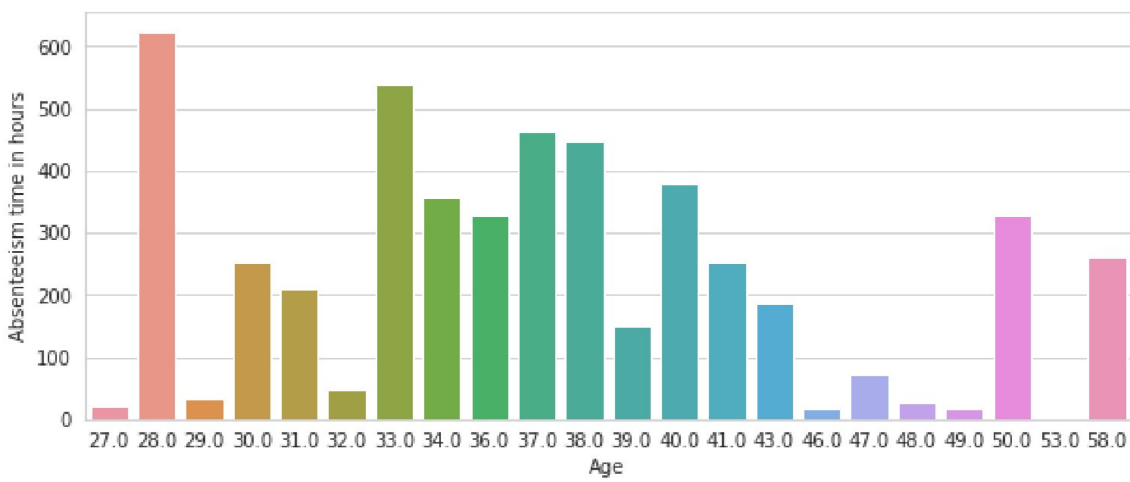
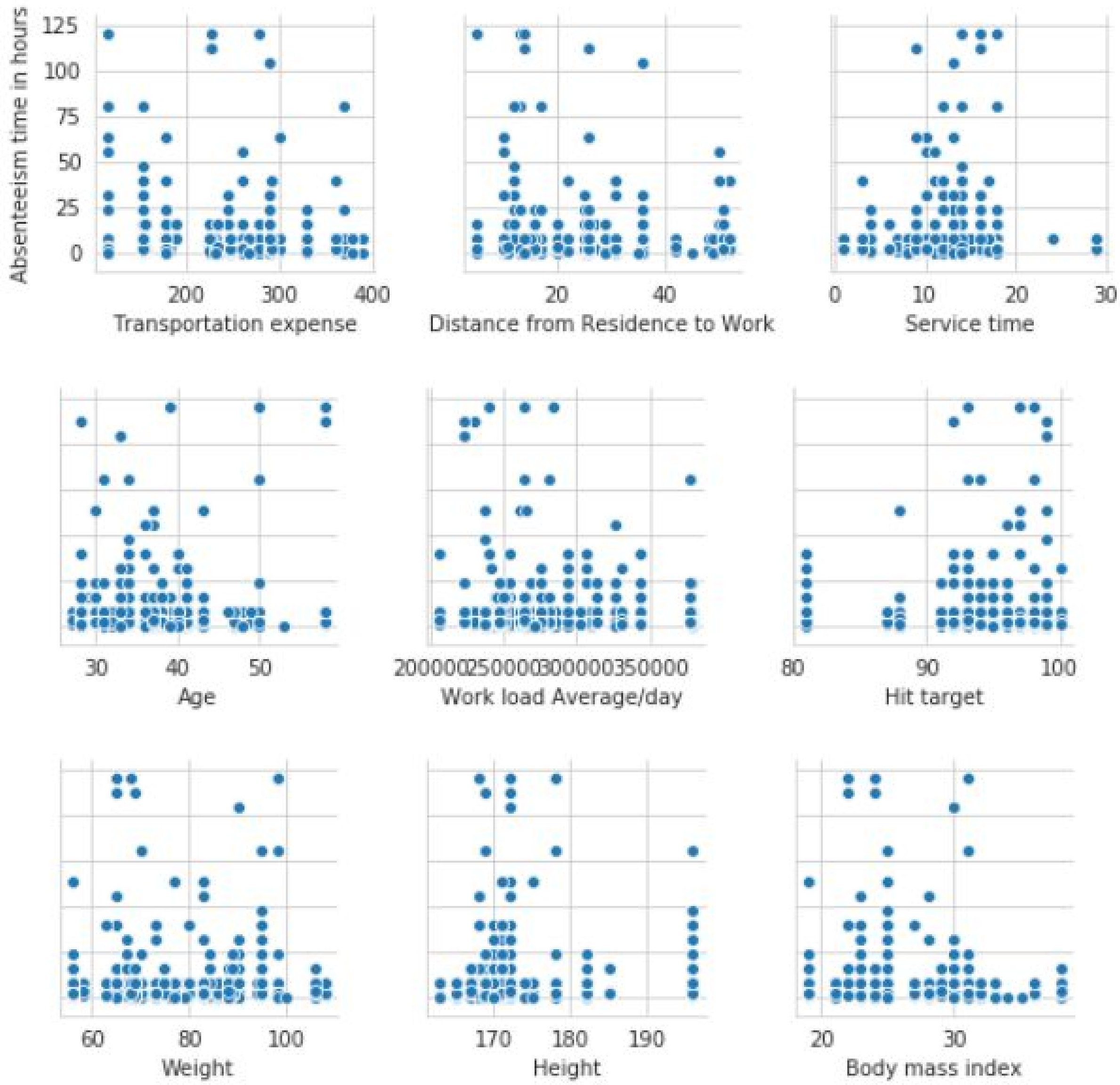
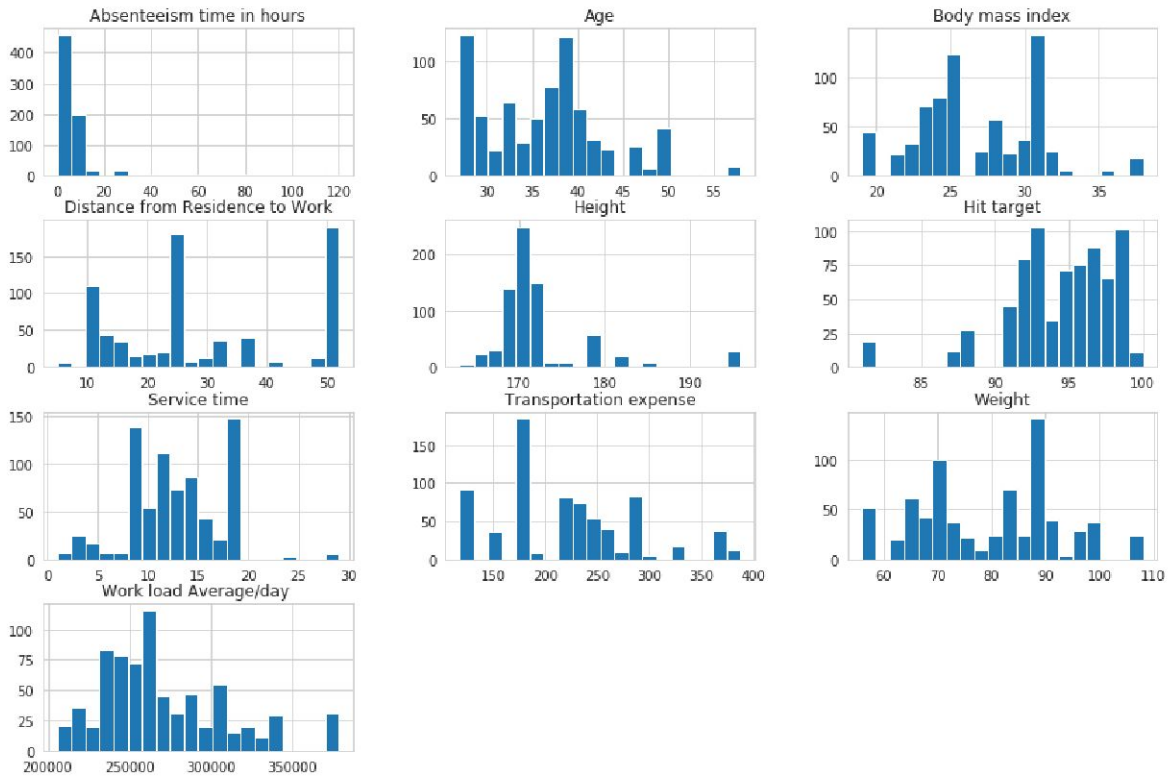


Fig. Bar graphs showing total no. of leaves applied age wise.Employees of age 28 and 33 took frequent leaves but for small hours but age 37,38,40 and 50 took long hours leave.

Distribution of Continuous variable with target variable



Distribution of continuous variables

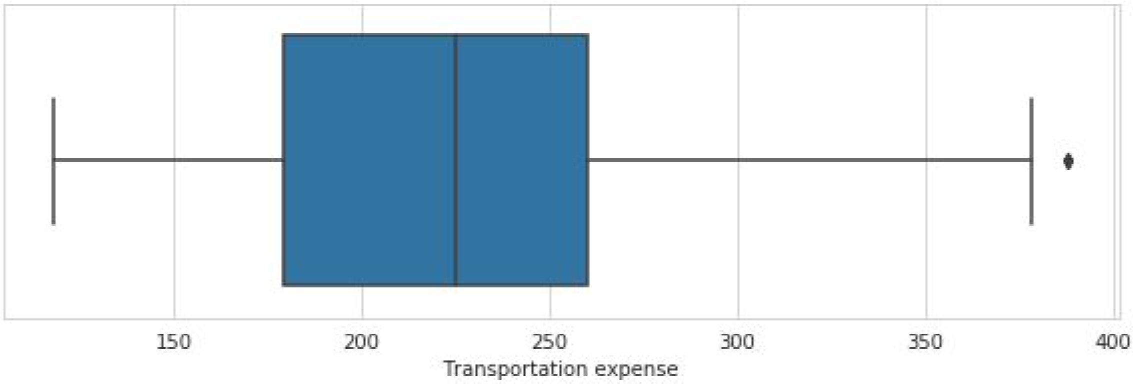


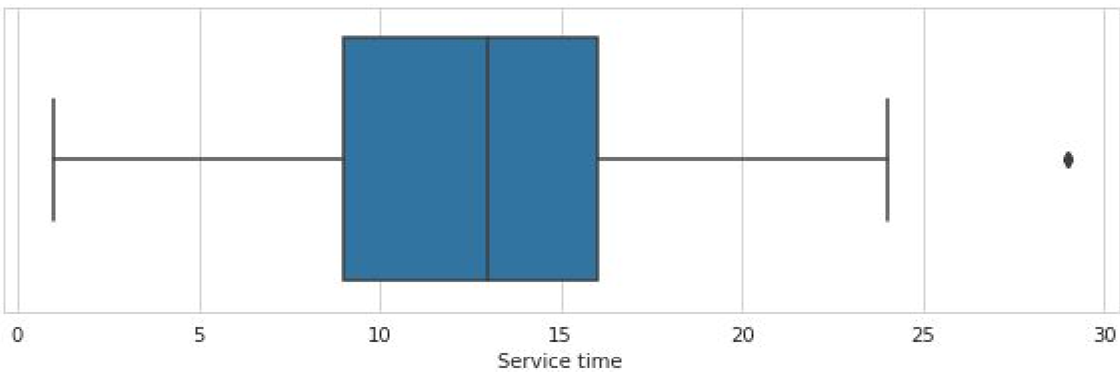
In the above image we can see that every variable is non uniformly distributed with significant numbers of outliers in Target variable.

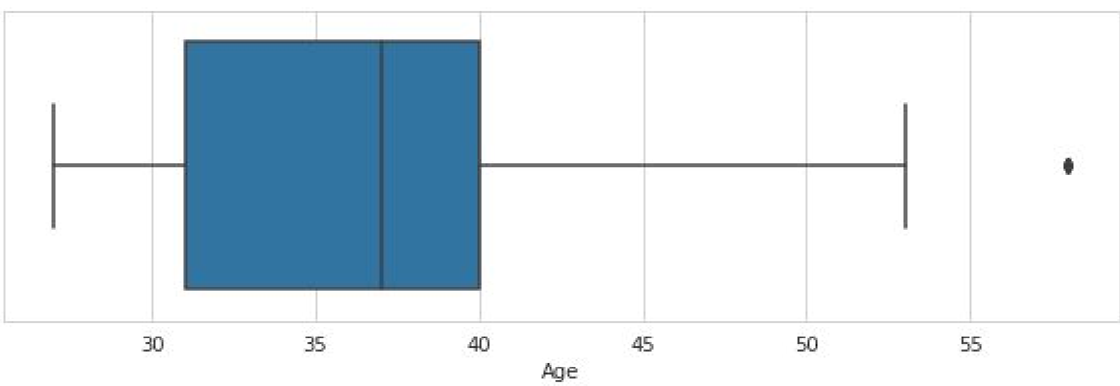
Missing Value Analysis

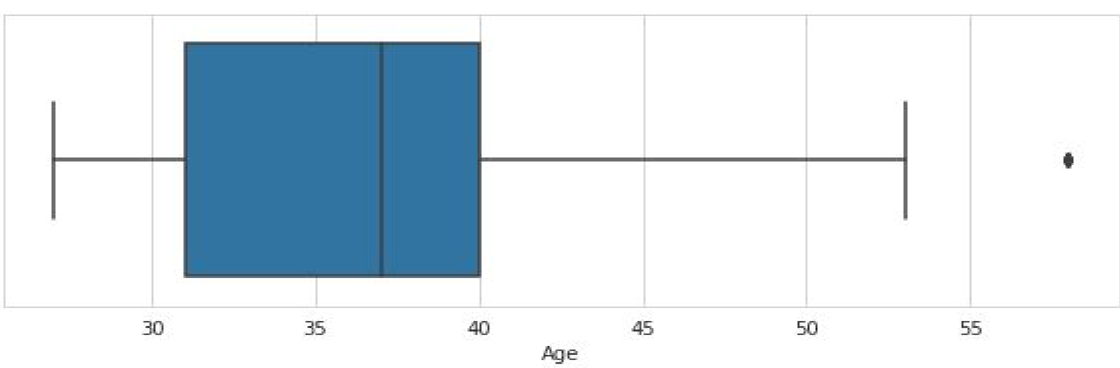
The missing values are imputed using mode for categorical variables and mean for continuous variables. Also Column 'Reason for absence', 'Month of absence' and 'Absenteeism time in hours' contains 0, but these variables cannot be 0. So these zeros are replaced with NA and then imputed using mode and median

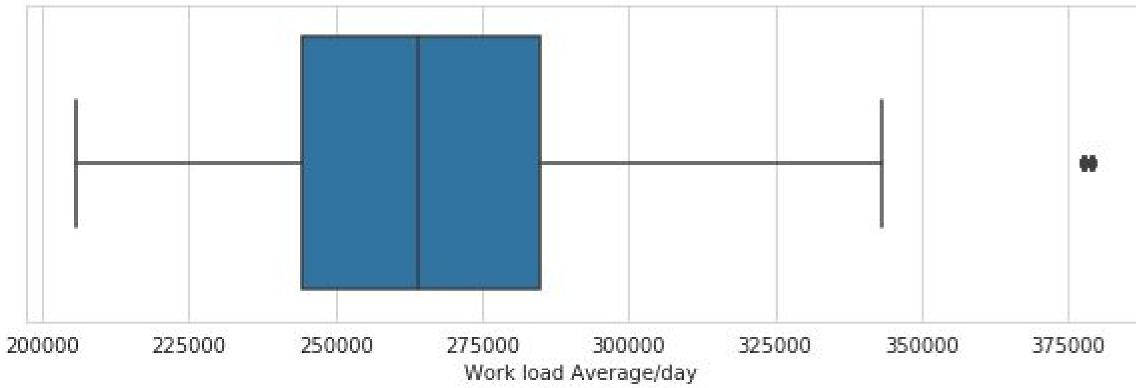
Outlier detection

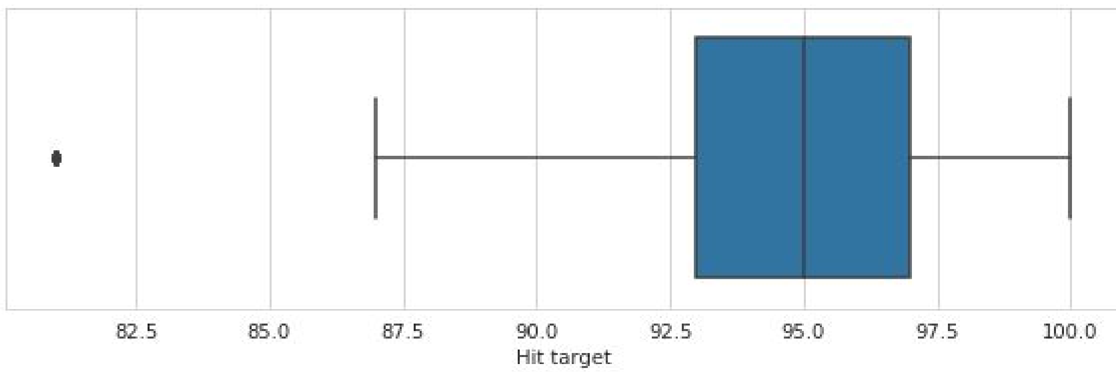


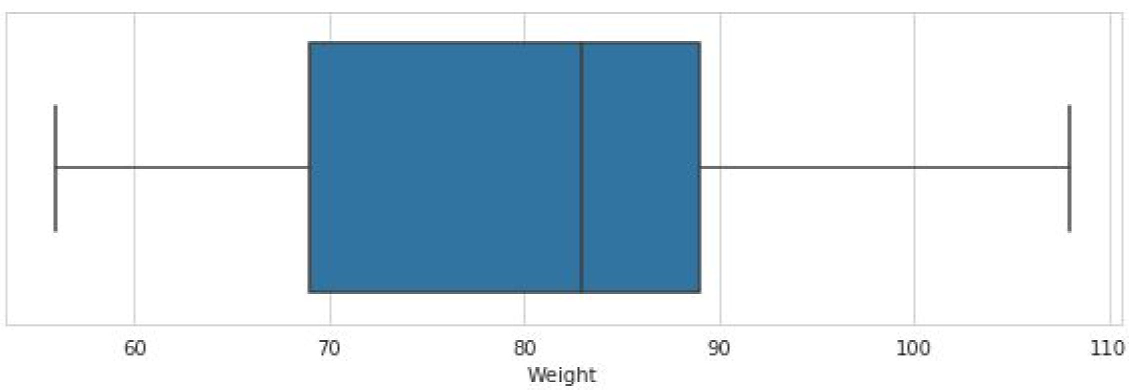


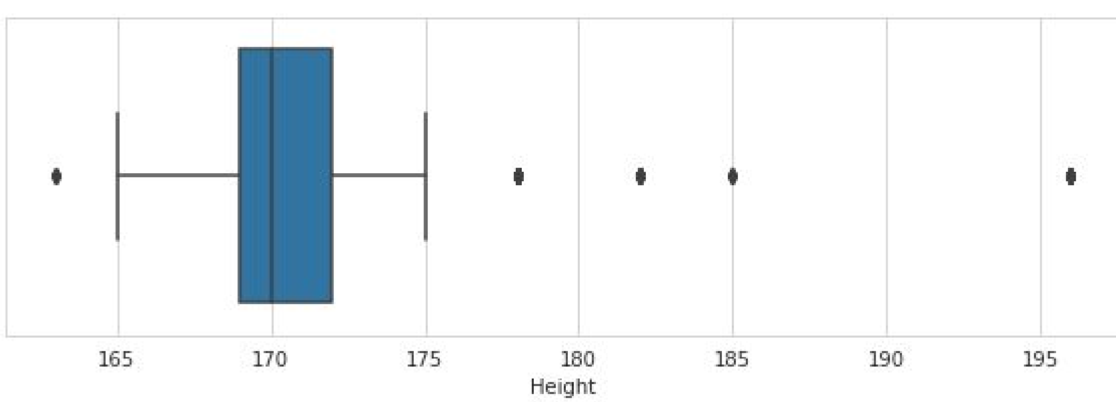


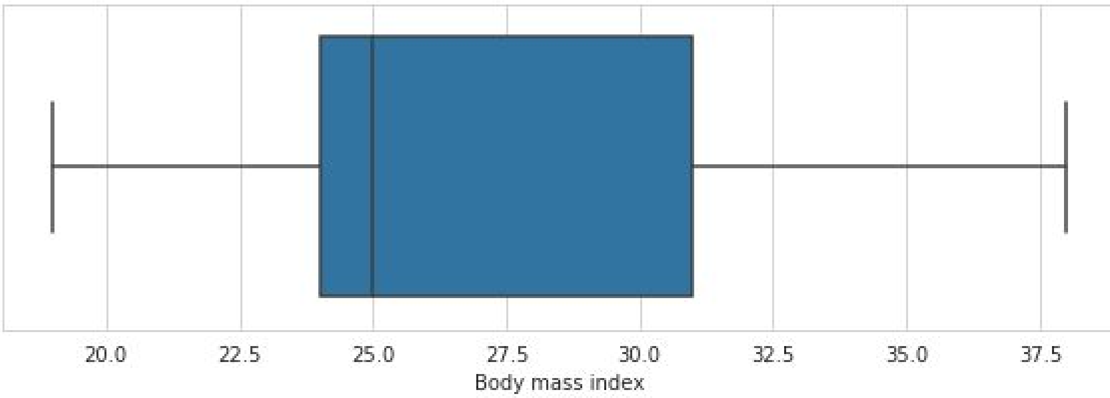


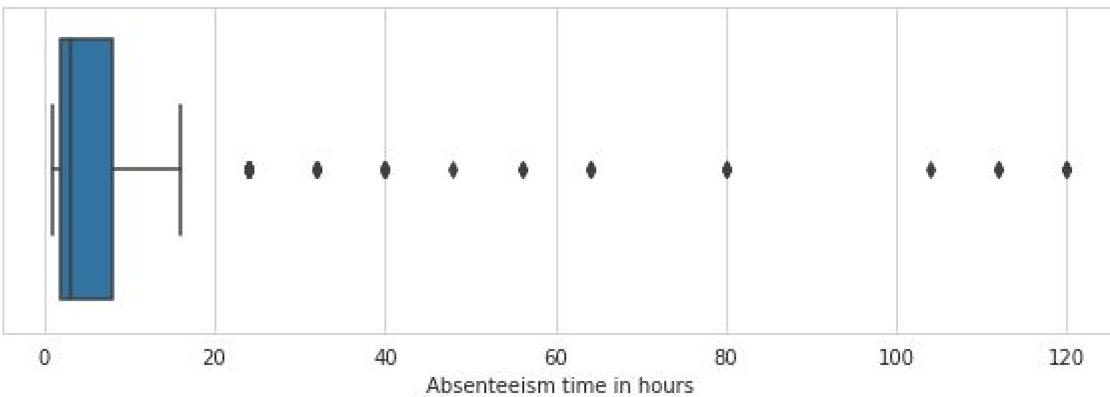

200000
225000
250000
275000
300000
325000
350000
375000
Work load Average/day


82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
Hit target


60
70
80
90
100
110
Weight


165
170
175
180
185
190
195
Height



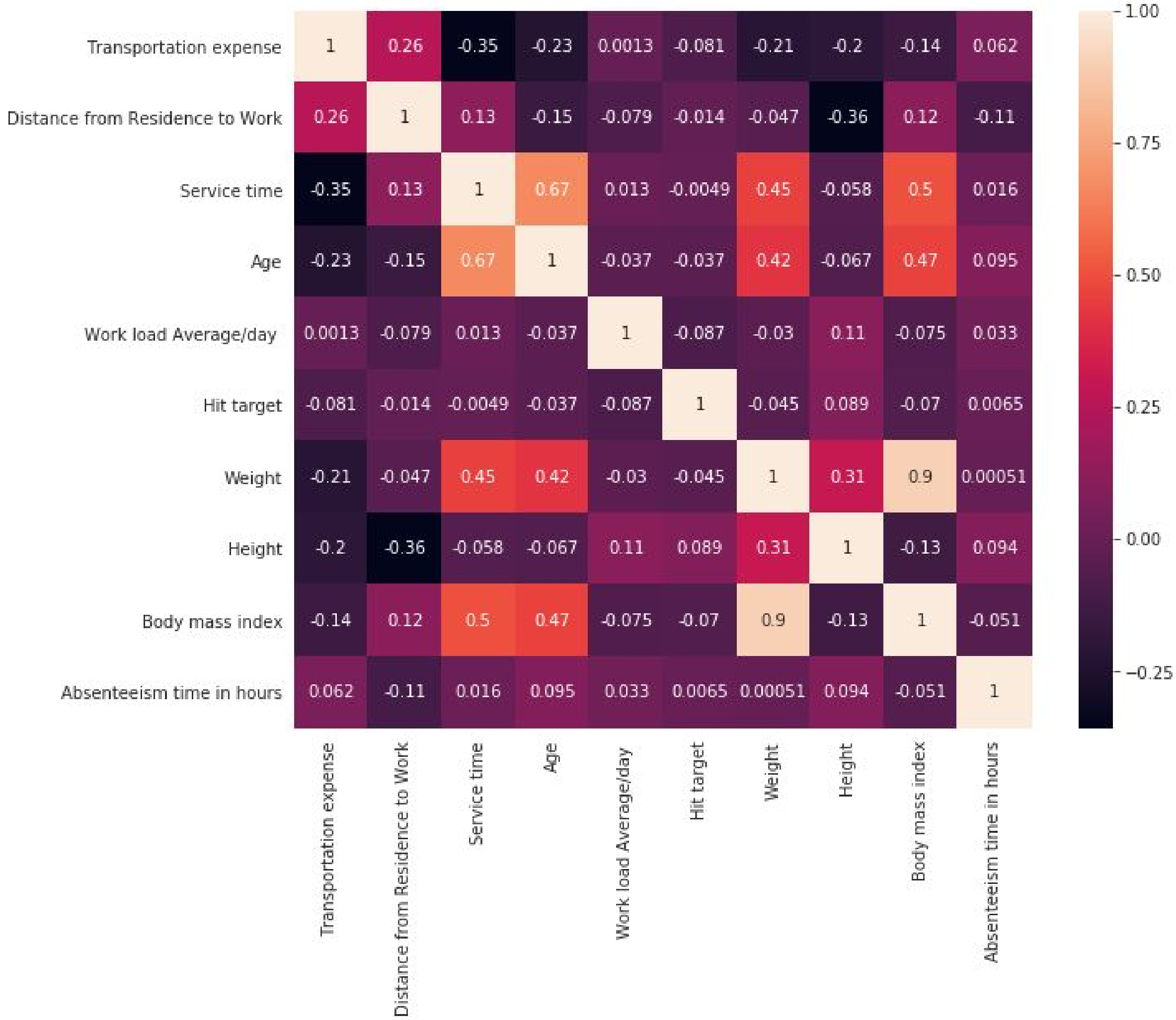


There is a small no. of outlier in every variable and significant in target variable Absenteeism time in hours.

Outlier Removal

Outliers can be removed Inter Quarentile Range. IQR is calculated with min and max value of a variable. Any value outside the min and max is considered an outlier.

Feature Selection



Here we need to select the feature which is valuable for the prediction task. Any variable with high collinearity should be discarded. From the above image, we can see that Weight and Body mass index are highly correlated so dropping Weight .

Feature Scaling

Normalizing the continuous variable.

Sampling

Split the dataset into 80 per cent training data and 20 per cent test data.

Modeling (Regression)

The target variable is continuous therefore the models should be regression type.

Tested with three algorithms:

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor

Evaluation

For score evaluation below are calculated

1. Mean Absolute Error
2. Mean Squared Error
3. Mean Absolute Percentage Error
4. R squared value

And the value of these scores are:

1. Mean Absolute Error
    a. LR : 2.17
    b. DTR : 1.971
    c. RFR : 1.908

2. Mean Squared Error
    a. LR : 9.8
    b. DTR : 10.3
    c. RFR : 8.46

3. Mean Absolute Percentage Error
    a. LR : 22.05
    b. DTR : 23.09
    c. RFR : 20.00

4. R Squared value
    a. LR : 0.23
    b. DTR : 0.19
    c. RFR : 0.34

Maximum RSquared value calculated was .52 after taking the cube root of the target variable.

Modeling and Evaluation (Classification)

As the data is non uniformly distributed it's hard to fit the data in any regression algorithm. Therefore converting the target variable into categorical.

Python: RandomForestClassifier gives nearly 80% accuracy when dividing into 2 classes.

R : RandomForestClassifier gives nearly 77% accuracy when dividing into 5 classes.

Conclusion

Calculation of the different scores tells us how often an algorithm can predict future cases. These are the definition of how these scores evaluate the data provided.

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It’s the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

MSE is a quadratic scoring rule that also measures the average magnitude of the error.

MAPE is a measure of prediction accuracy It usually expresses accuracy as a percentage and is defined by the formula.

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

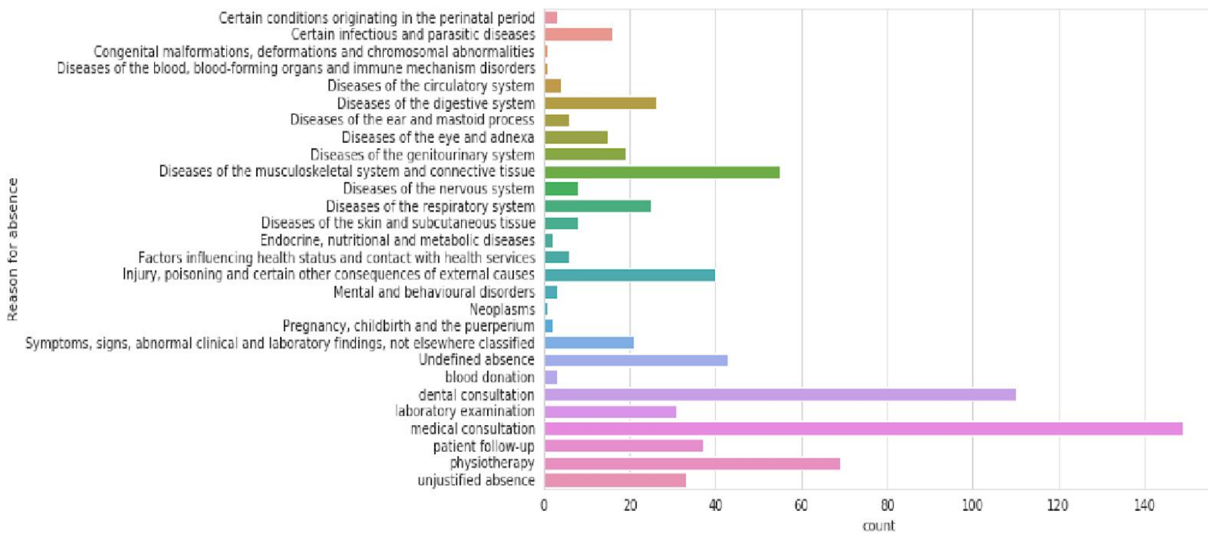
So considering these different scores of evaluation and comparing them among different algorithms it shows that due to nonuniform distribution of variables regression analysis cannot predict properly, so the target variable needs to be changed to categorical.

After evaluating with different number of categories the maximum accuracy obtained is 80 percent.

Solution of the problem statement

Q. What changes the company should bring to reduce the number of absenteeism?

a. It is observed from reasons of absence, employees mostly were absent because of the medical consultation reason. So, the company might add a full body check program for the well being of the employees.



b. Based on Educational qualification employees with high school degrees were mostly absent. The company can recruit graduate employees.

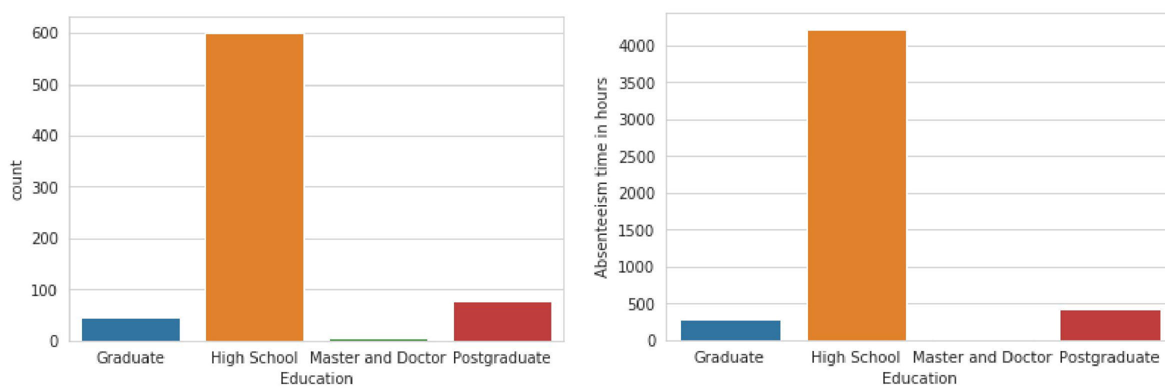


Fig. Bar graphs showing employees with what educational qualifications are mostly absent. It is clear employees with high school degree took significant no. of leaves.

c. Employees who are social drinker tends to be more absent. The company could call these employees and tell them not to drink in-office hours.

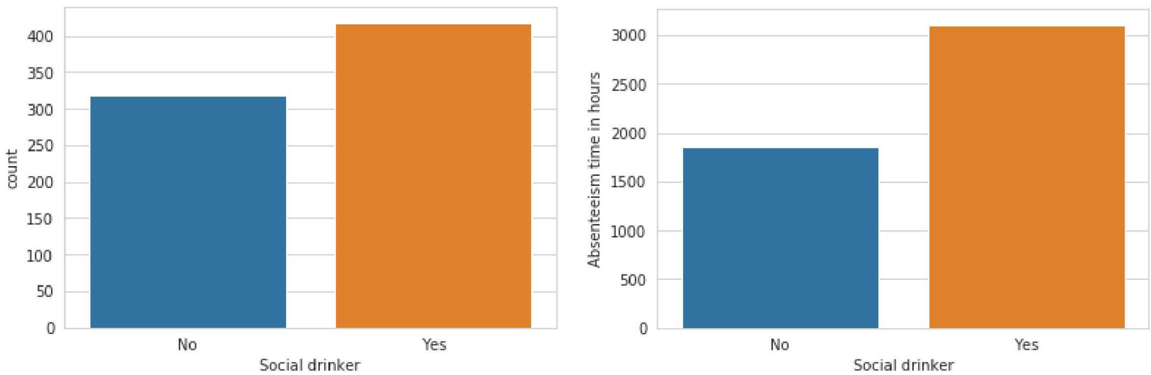


Fig. Bar graphs showing the frequency of leaves and total no of hours absent depending on the drinking habit of the employee.

d. Employees took long hours leave in the winter season. The company could provide them medical facility and transport facility or work from home. So that they won't miss their work.

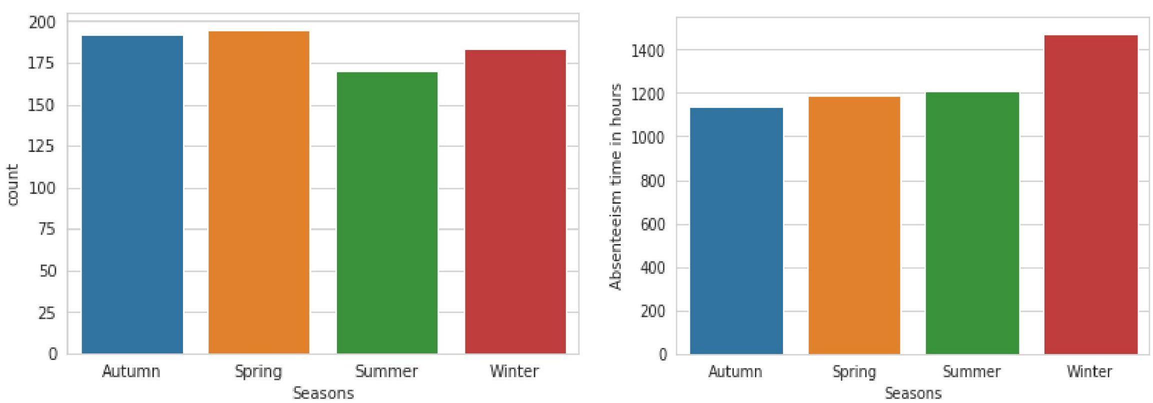


Fig. Bar graphs showing Season wise frequency of absence and total no. hours of absence

Q How much losses every month can we project in 2011 if same trend of absenteeism continues?

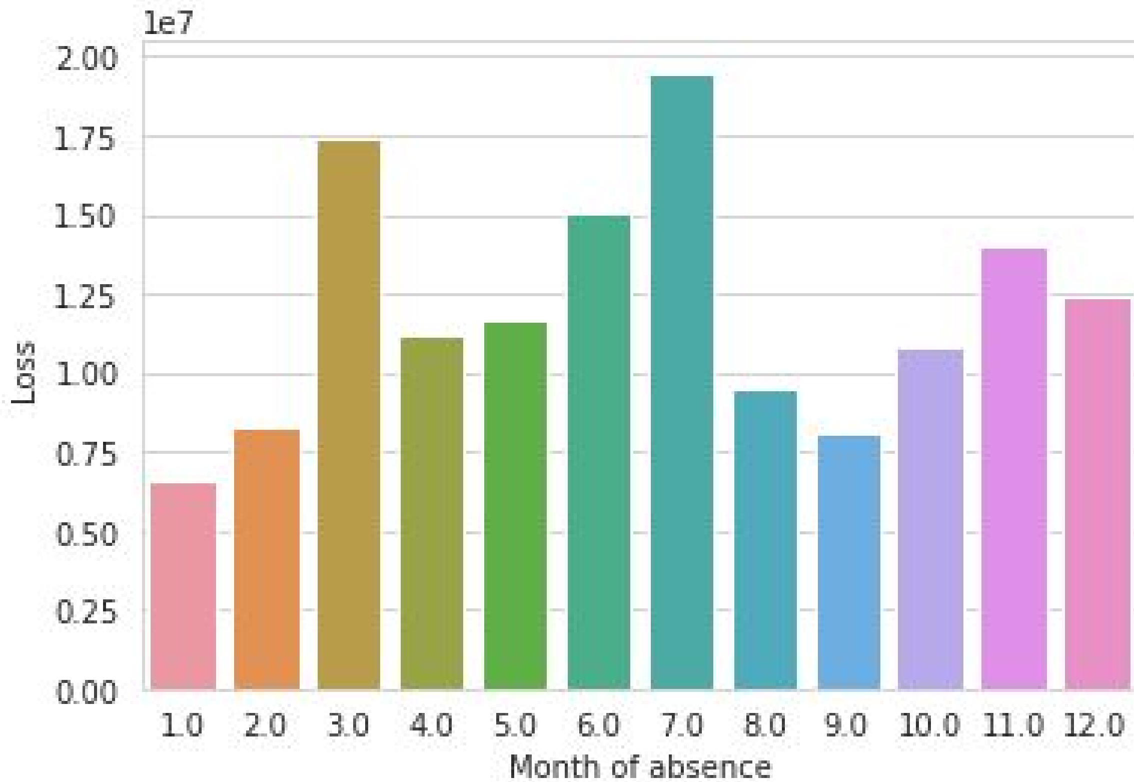


Fig. Bar graphs showing month wise total loss due to the absence of employees.

Appendix

Figures

Distribution of a categorical variable with the target variable

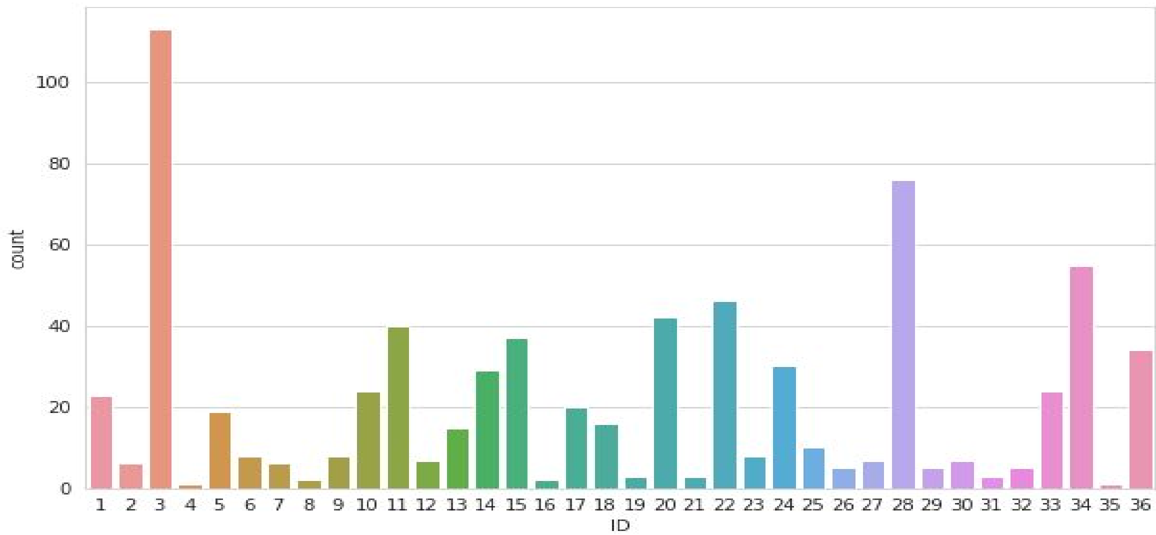


Fig. Bar graph showing the Frequency to reasons given by employees.

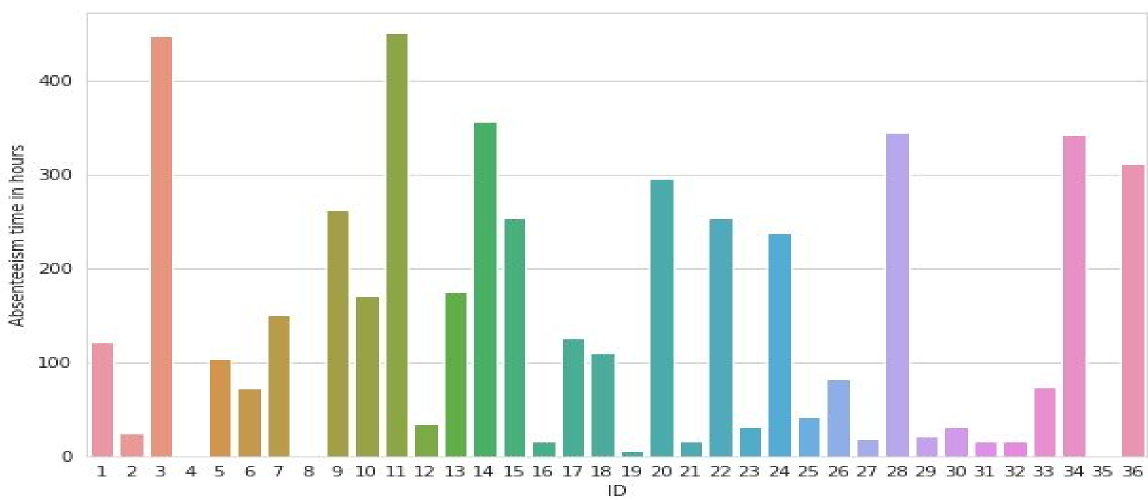


Fig. Bar graph showing total no. of absence hours reason wise

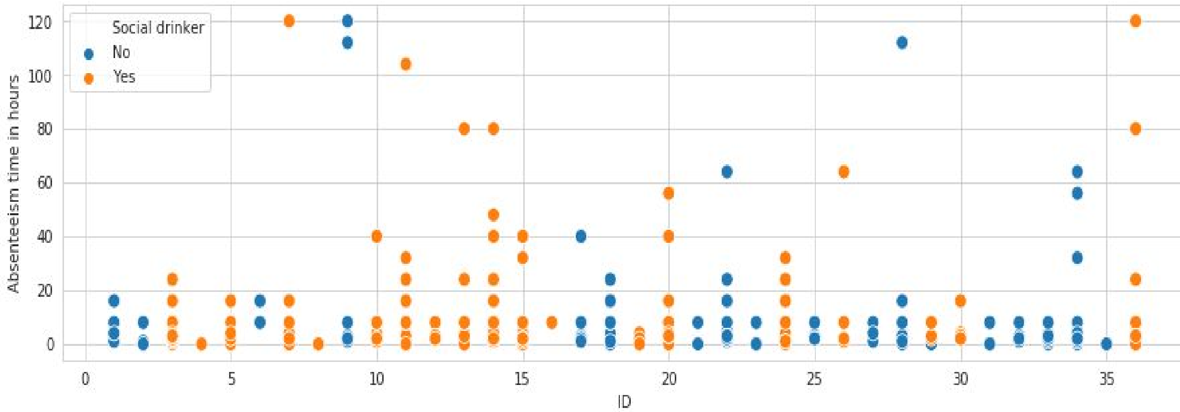


Fig. Scatter Plot showing absence of each employee with hours of absence and drinking habit

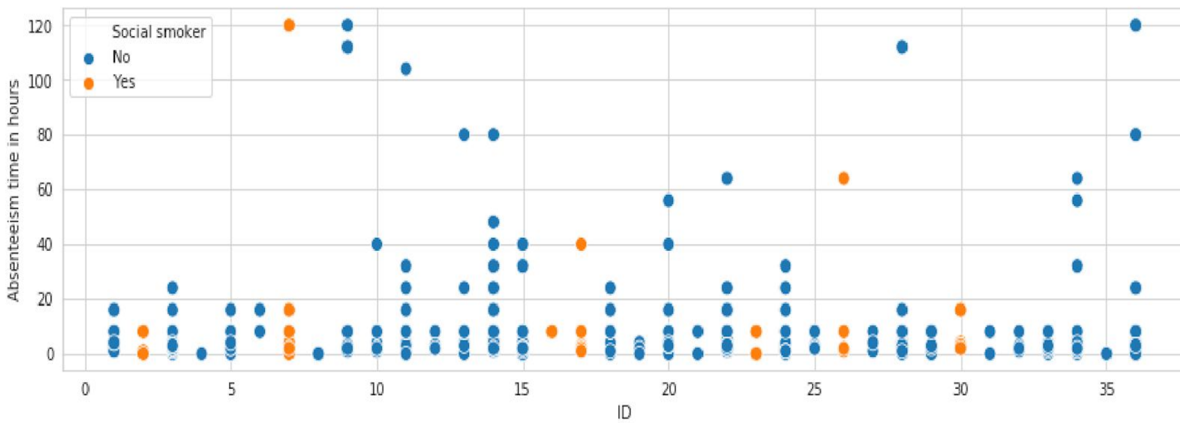


Fig.Scatter plot showing each leaves taken by each employee and their smoking habit

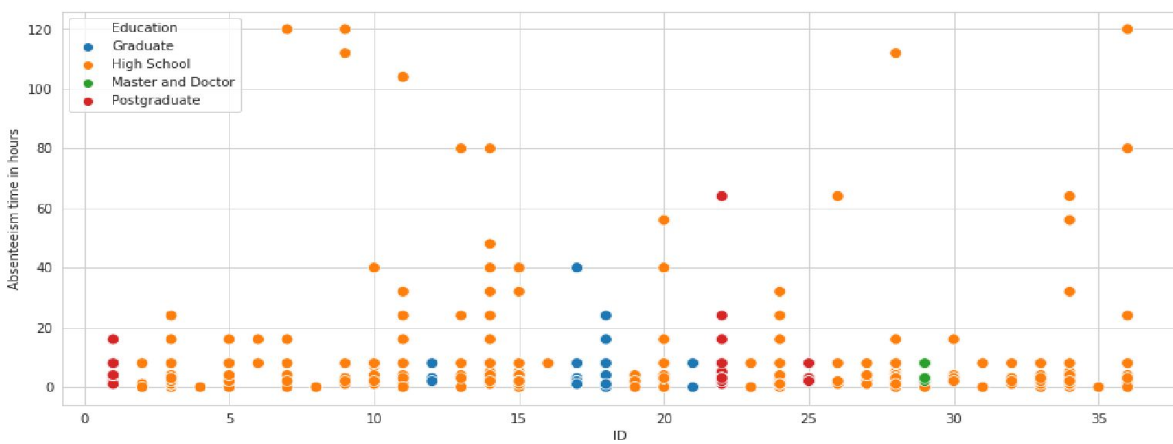


Fig. Scatter plot showing each leave took by each employee and their Educational qualification

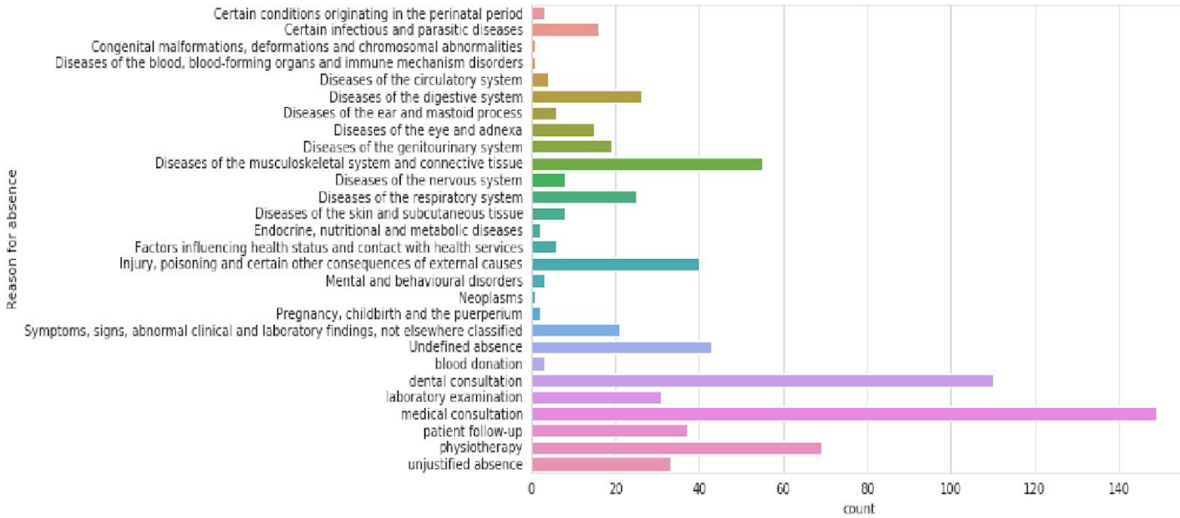


Fig. Bar graph showing the Frequency to reasons given by employees for absence

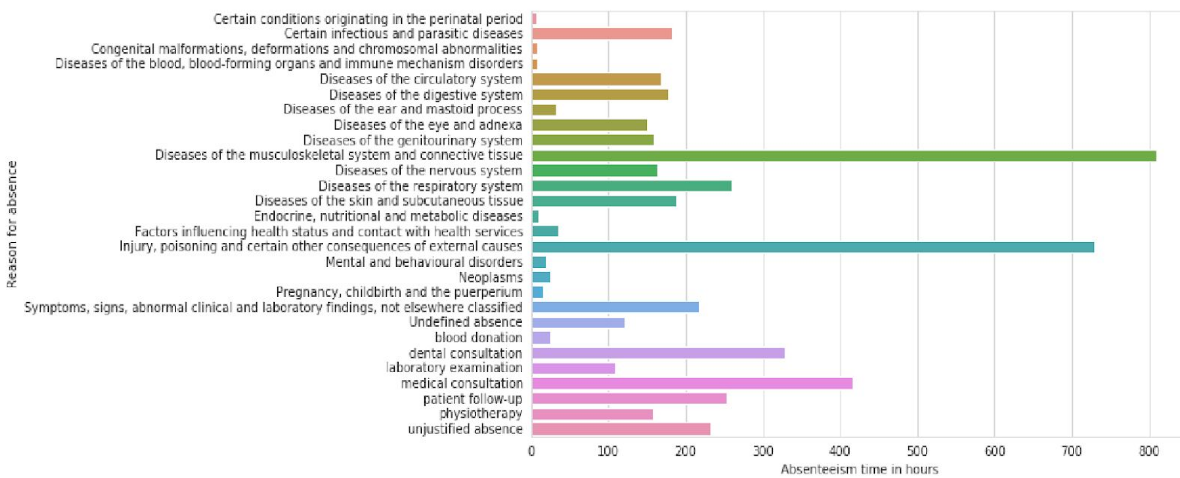


Fig. Bar graph showing the number of absence hours per reason

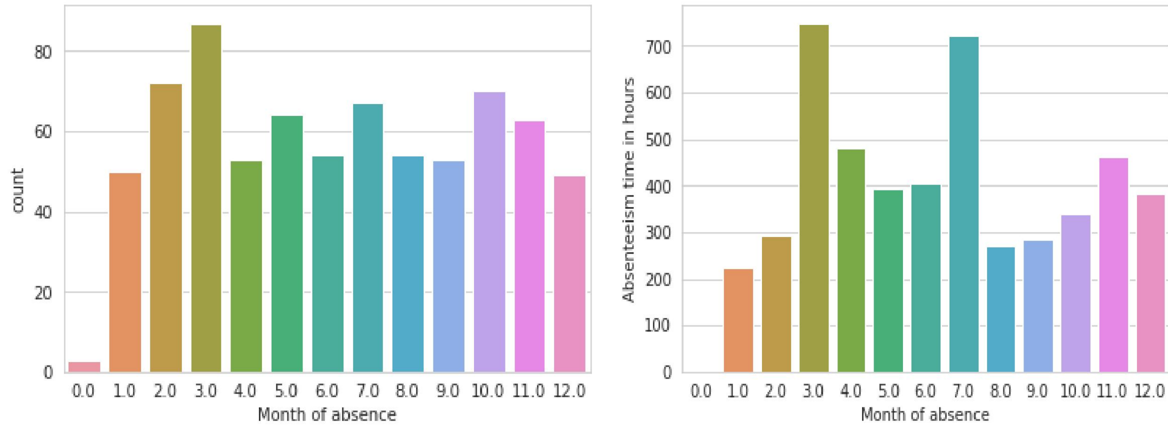


Fig. Bar graphs showing Month wise frequency of absence and total no. hours absence

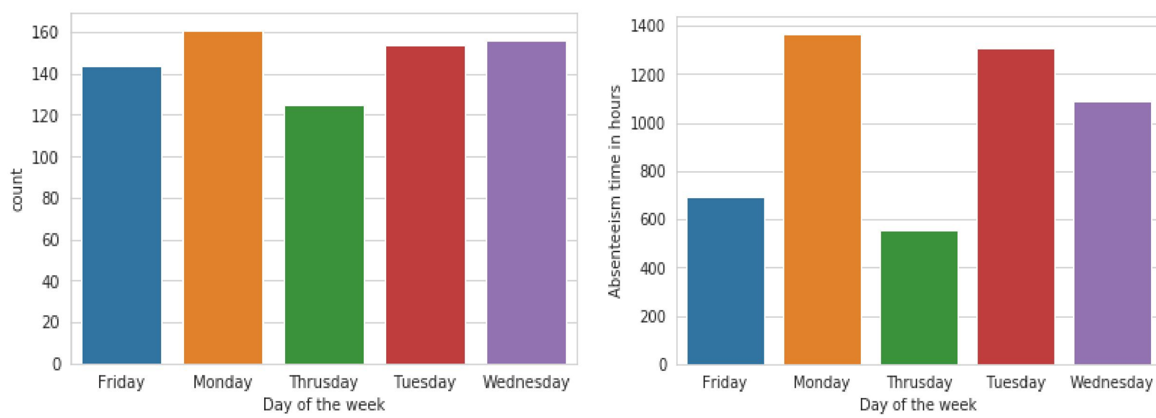


Fig.Bar graphs showing weekday frequency of absence and total no. hours absence

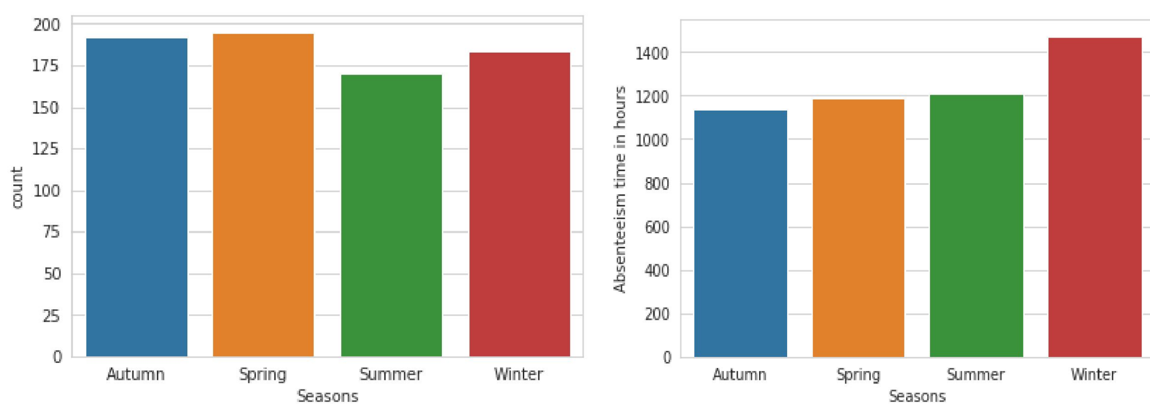


Fig. Bar graphs showing Season wise frequency of absence and total no. hours of absence

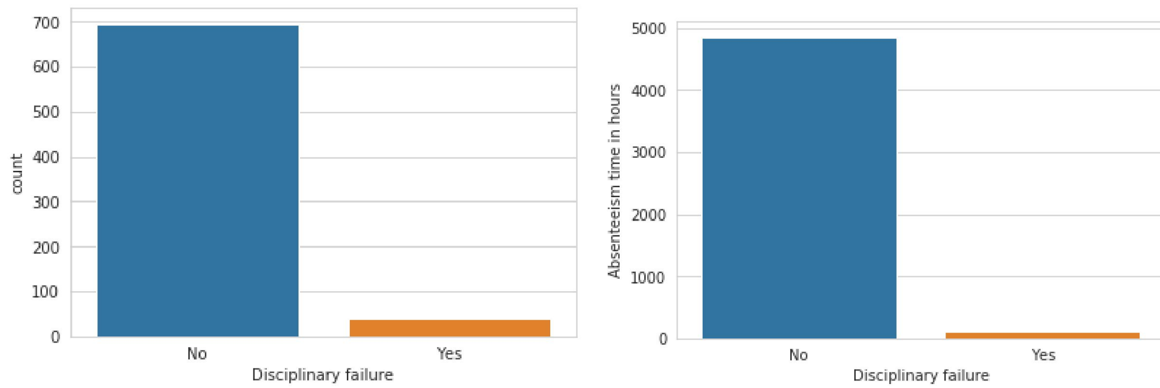


Fig. Bar graphs showing frequency of employee being absent based on their disciplinary status and total no. of hours absent by disciplinary category.

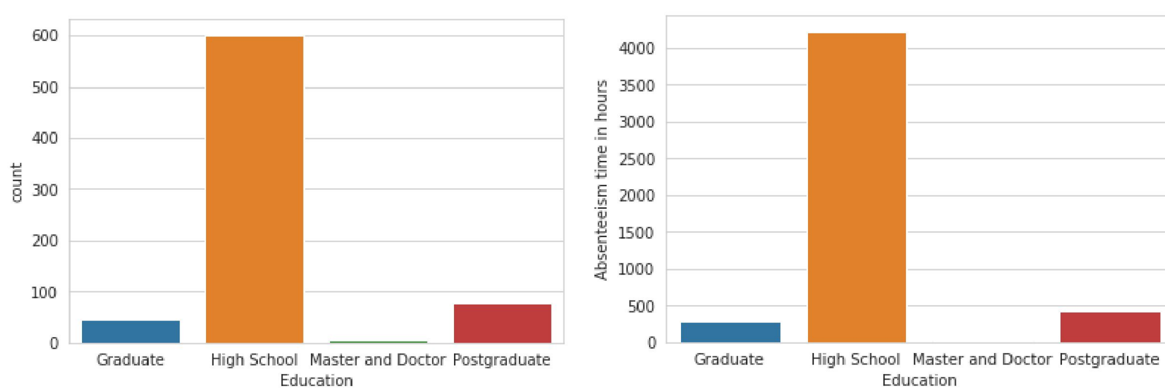


Fig. Bar graphs showing employees with what educational qualifications are mostly absent. It is clear employees with high school degree took significant no. of leaves.

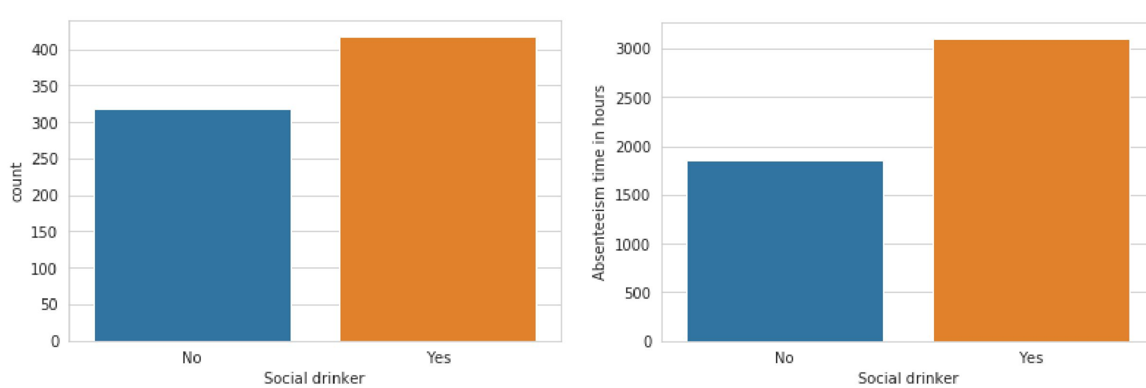


Fig. Bar graphs showing the frequency of leaves and total no of hours absent depending on the drinking habit of the employee

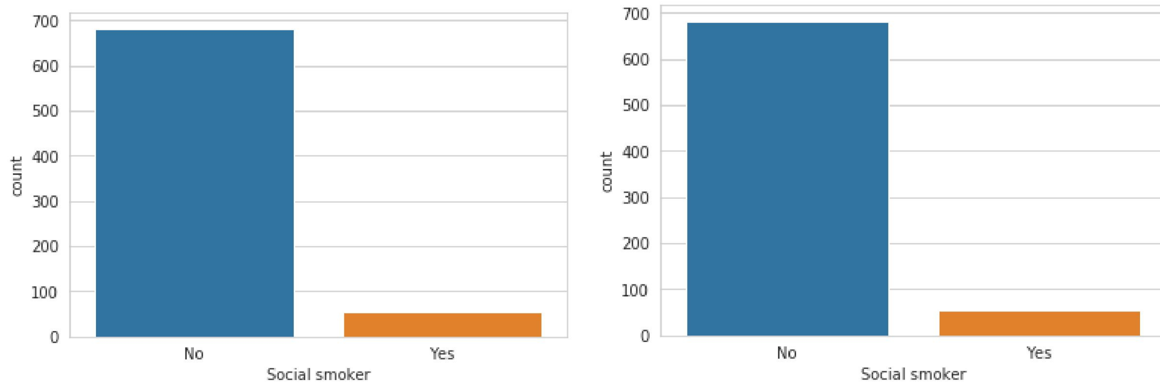


Fig. Bar graphs showing frequency of leaves and total no of hours absent depending on the Smoking habit of the employee. As we can see there is very less no. of people in the company who smokes.

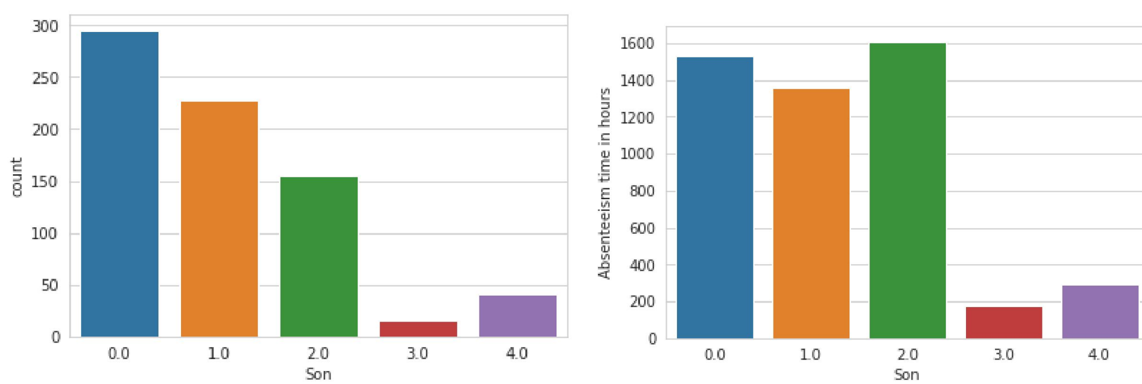


Fig. Bar graphs showing frequency of absence and total no. hours absent depending on the number of son employees have. People with 1 son took consistent hours of leave but employee with two son took long hours leave but not frequent.

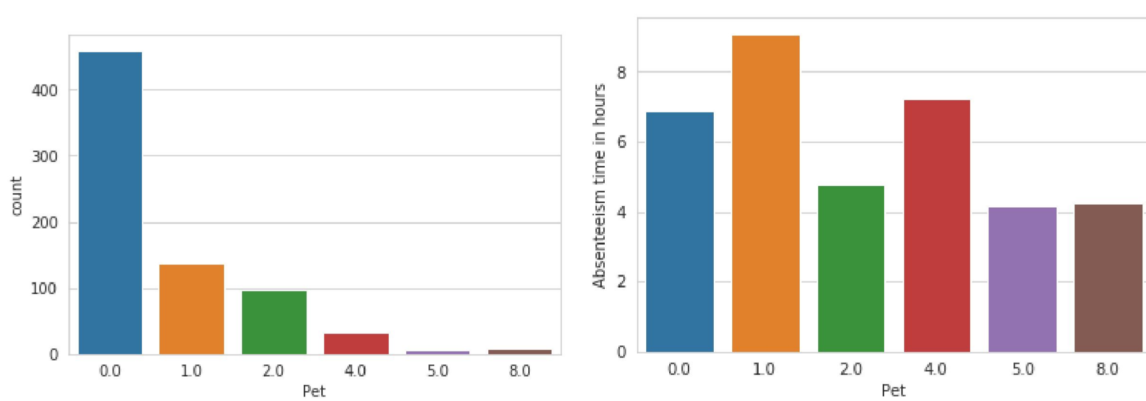


Fig. Bar graphs showing frequency and total no. hours absent depending on. Number pet employees have.

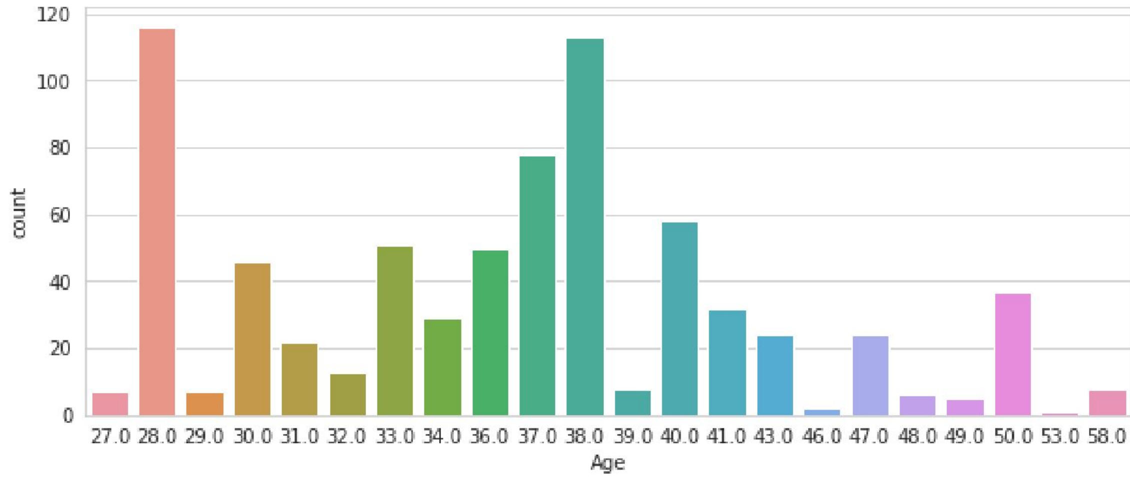


Fig. Bar graphs showing frequency of leaves applied age wise.Mostly employees of age 28 and 38 took more number of leaves.

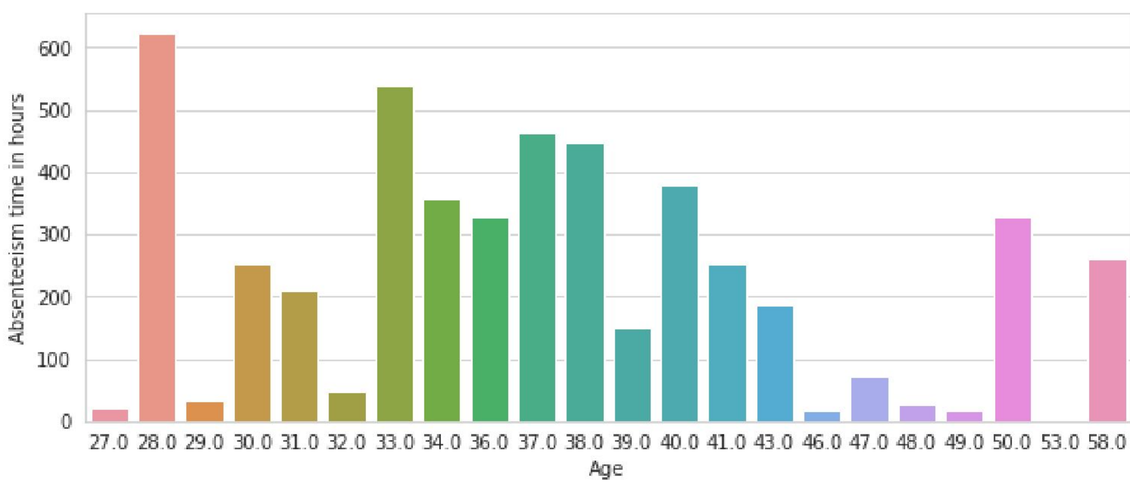
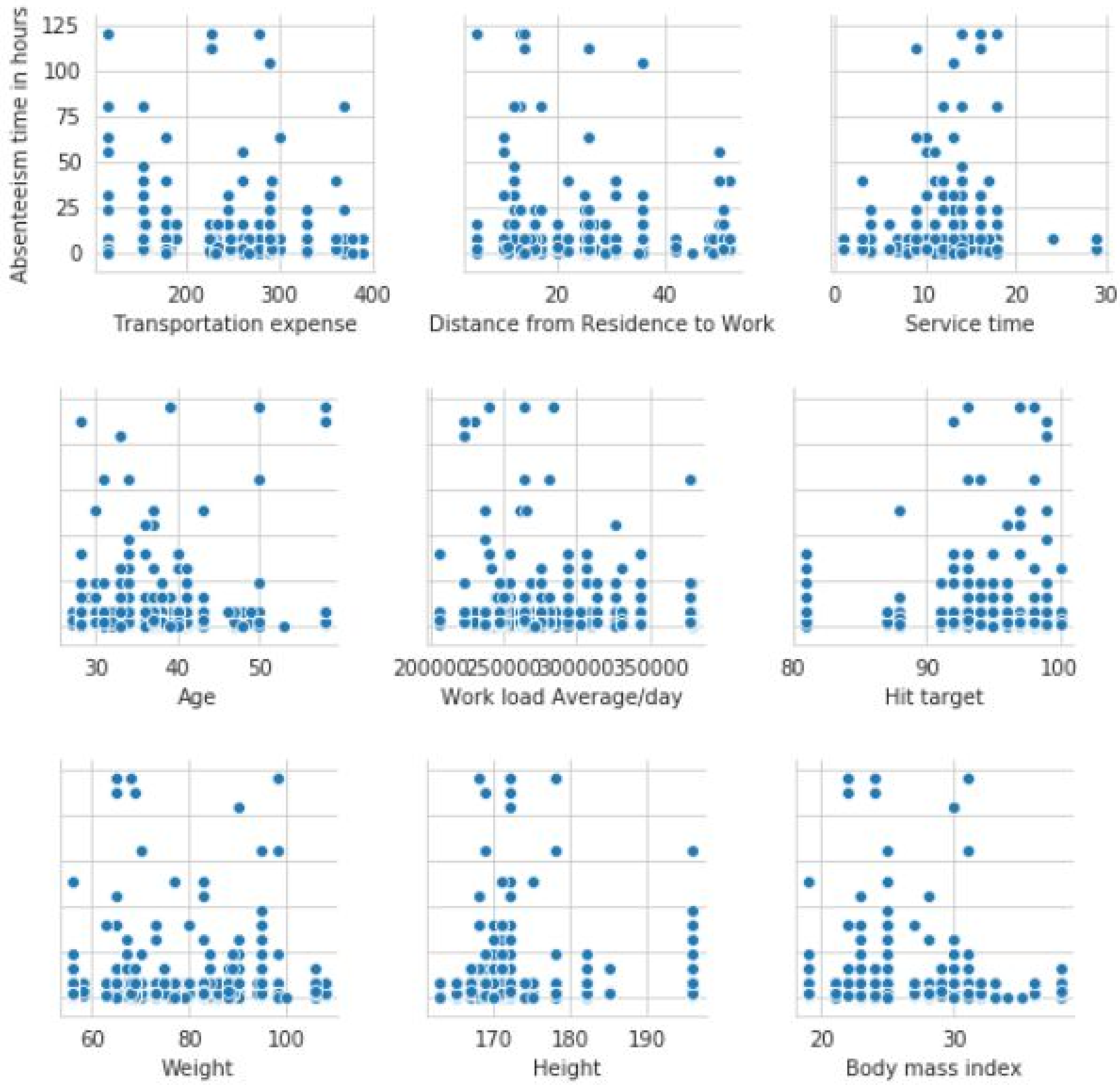
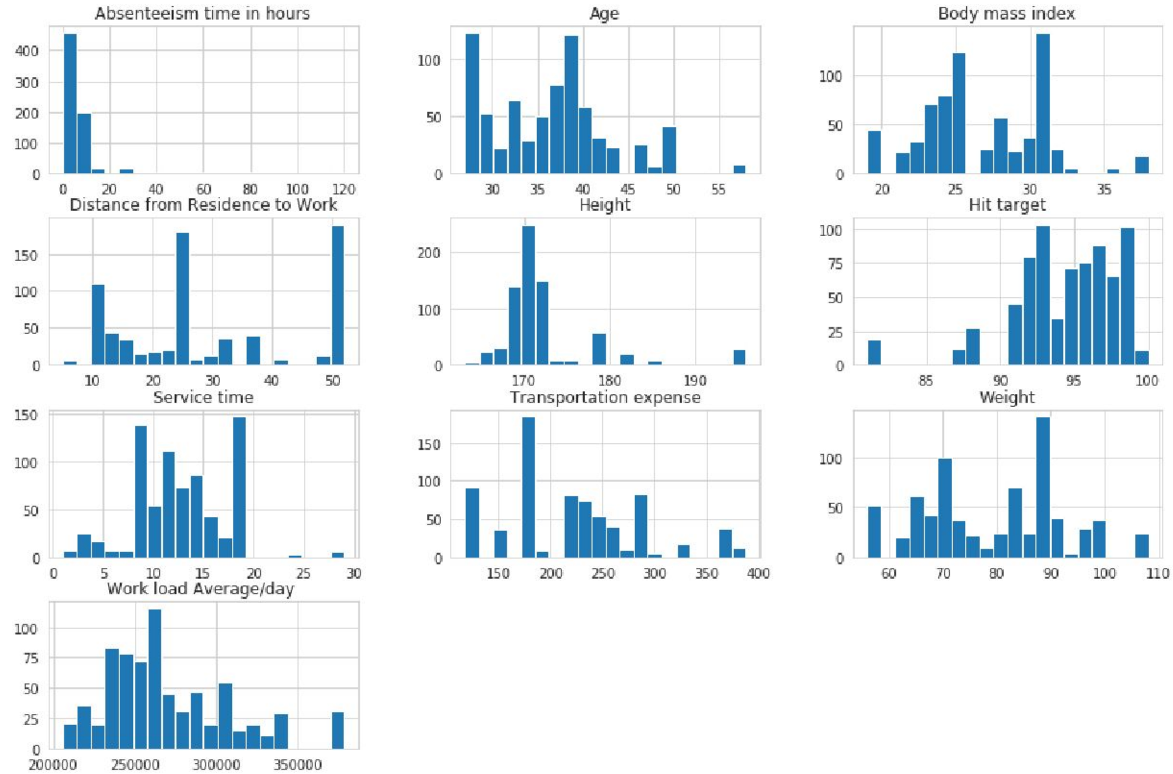


Fig. Bar graphs showing total no. of leaves applied age wise.Employees of age 28 and 33 took frequent leaves but for small hours but age 37,38,40 and 50 took long hours leave.

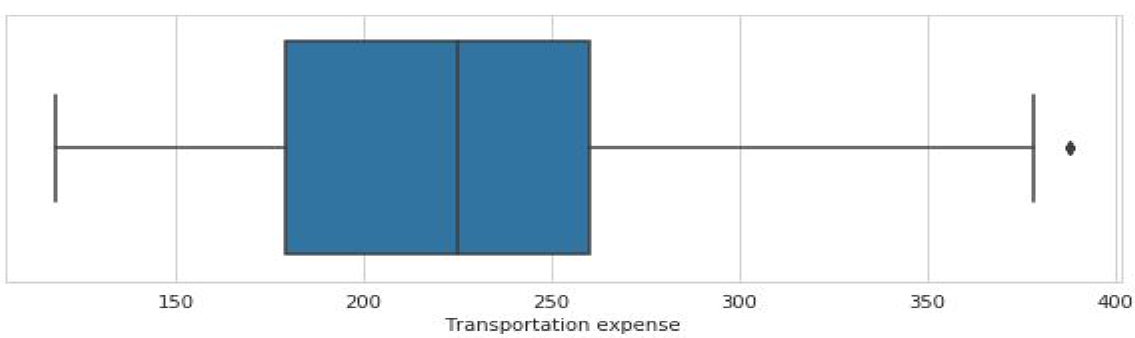
Distribution of Continuous variable with target variable

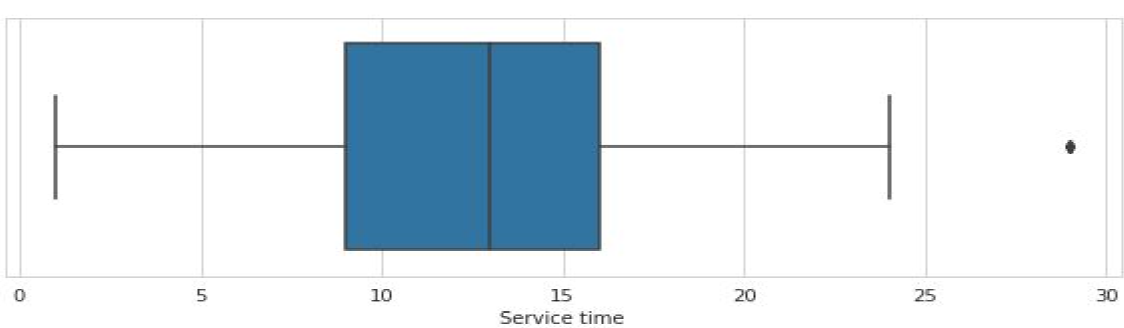


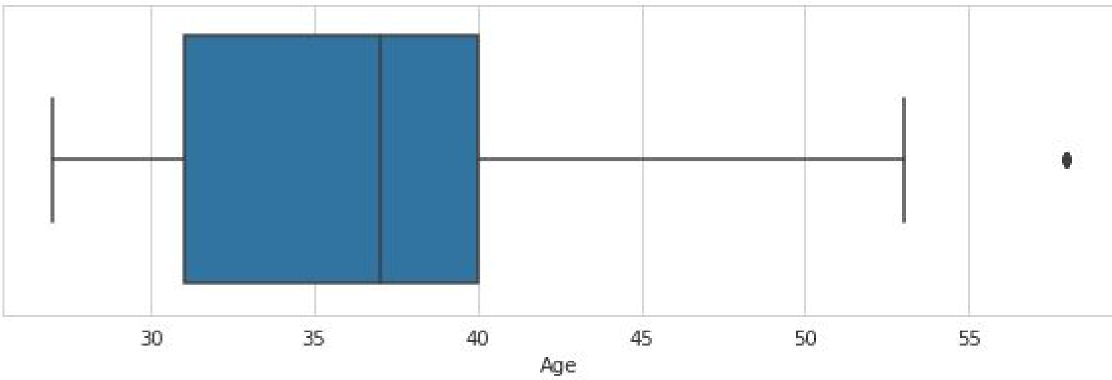
Distribution of continuous variables

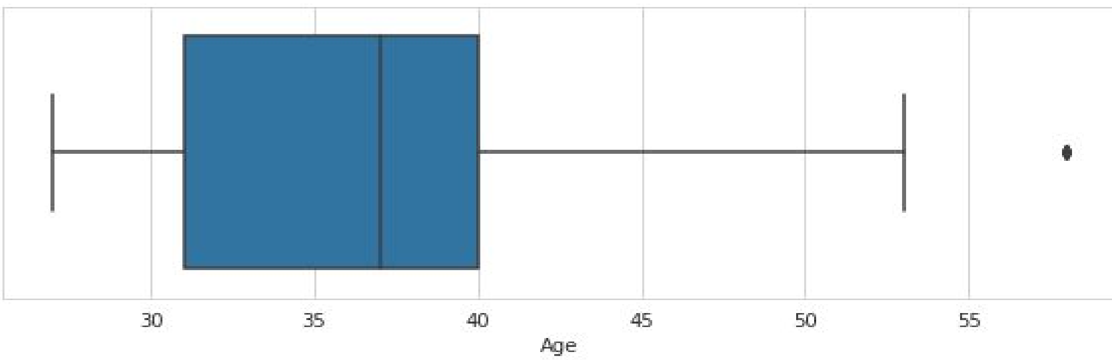


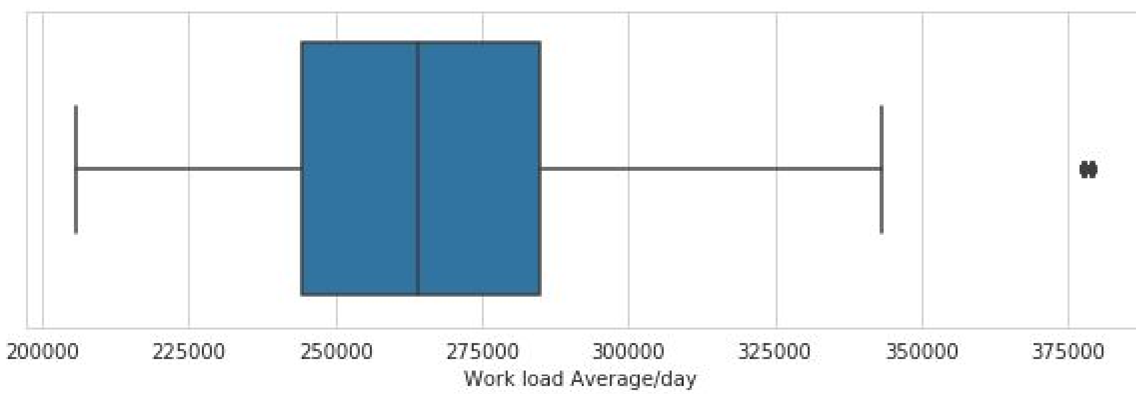
Outlier detection

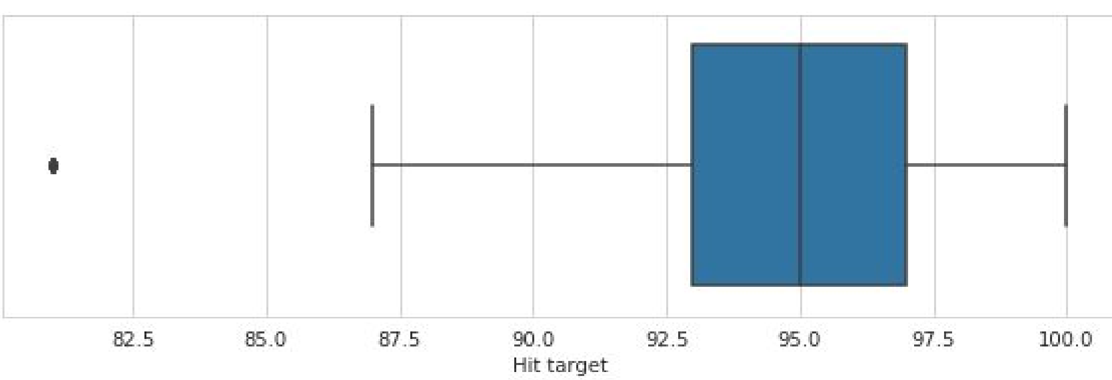


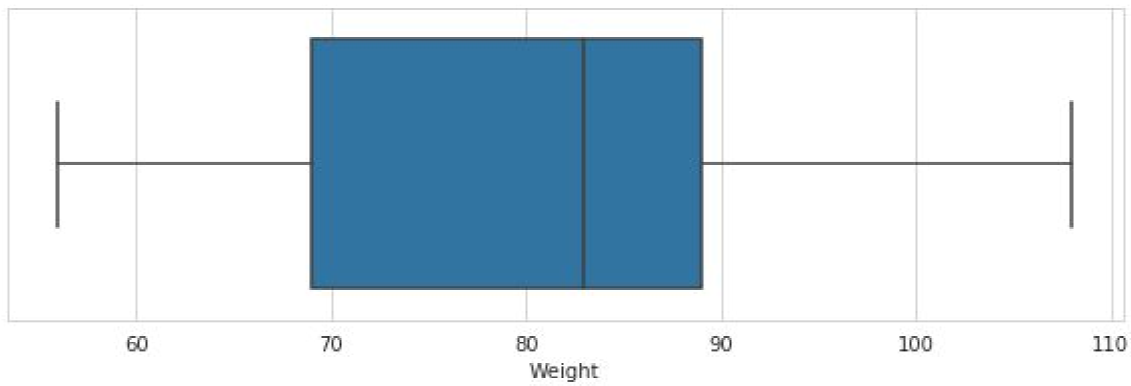


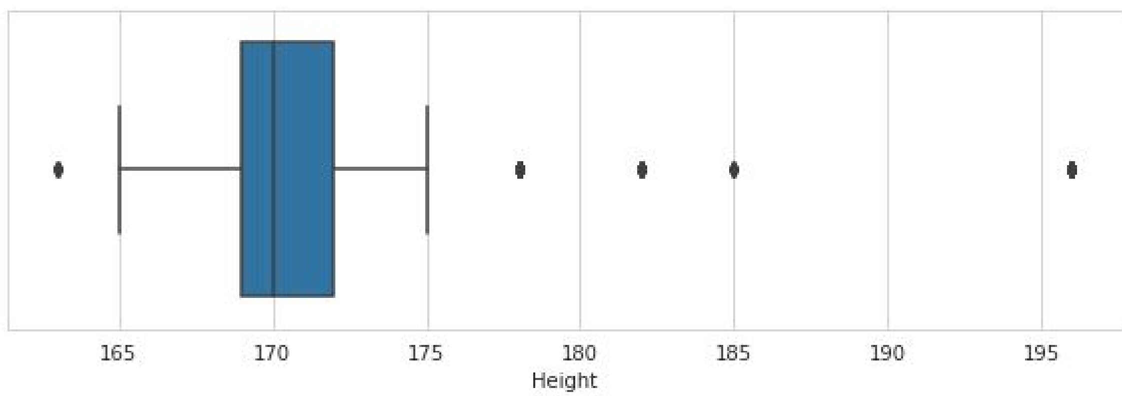

30
35
40
45
50
55
Age

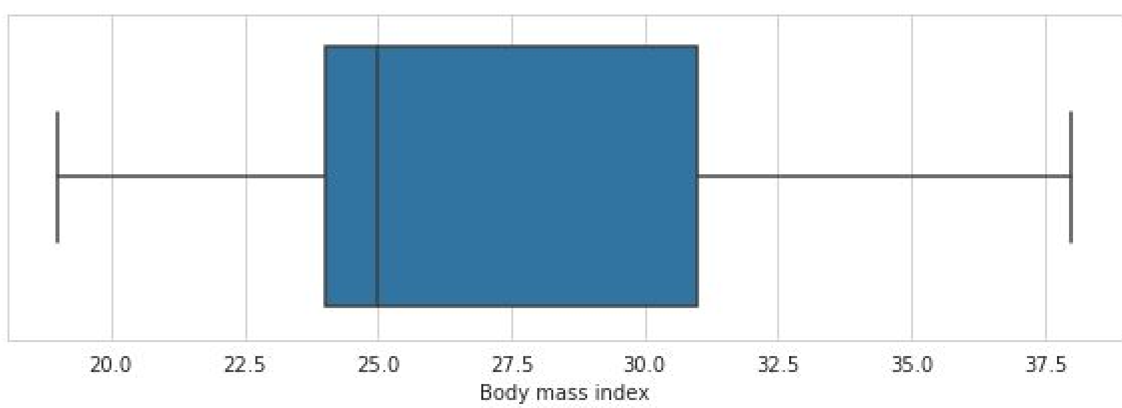

30
35
40
45
50
55
Age

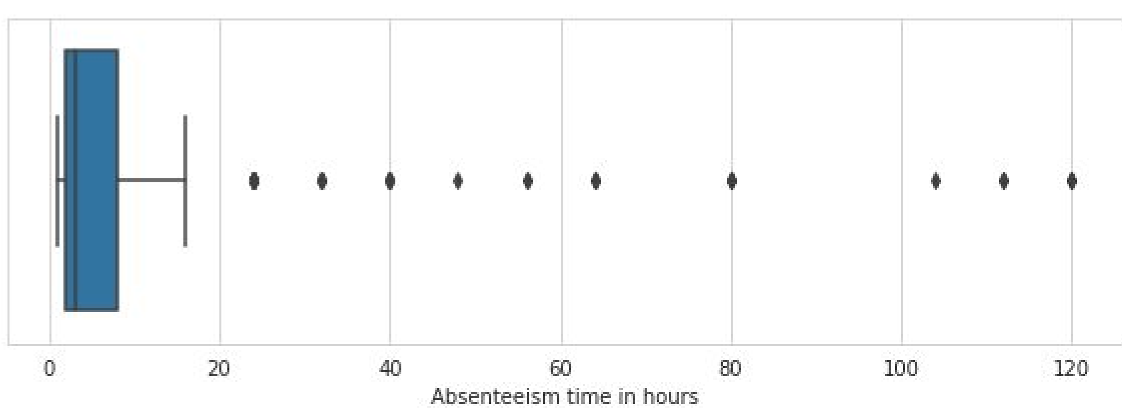

200000
225000
250000
275000
300000
325000
350000
375000
Work load Average/day


82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
Hit target

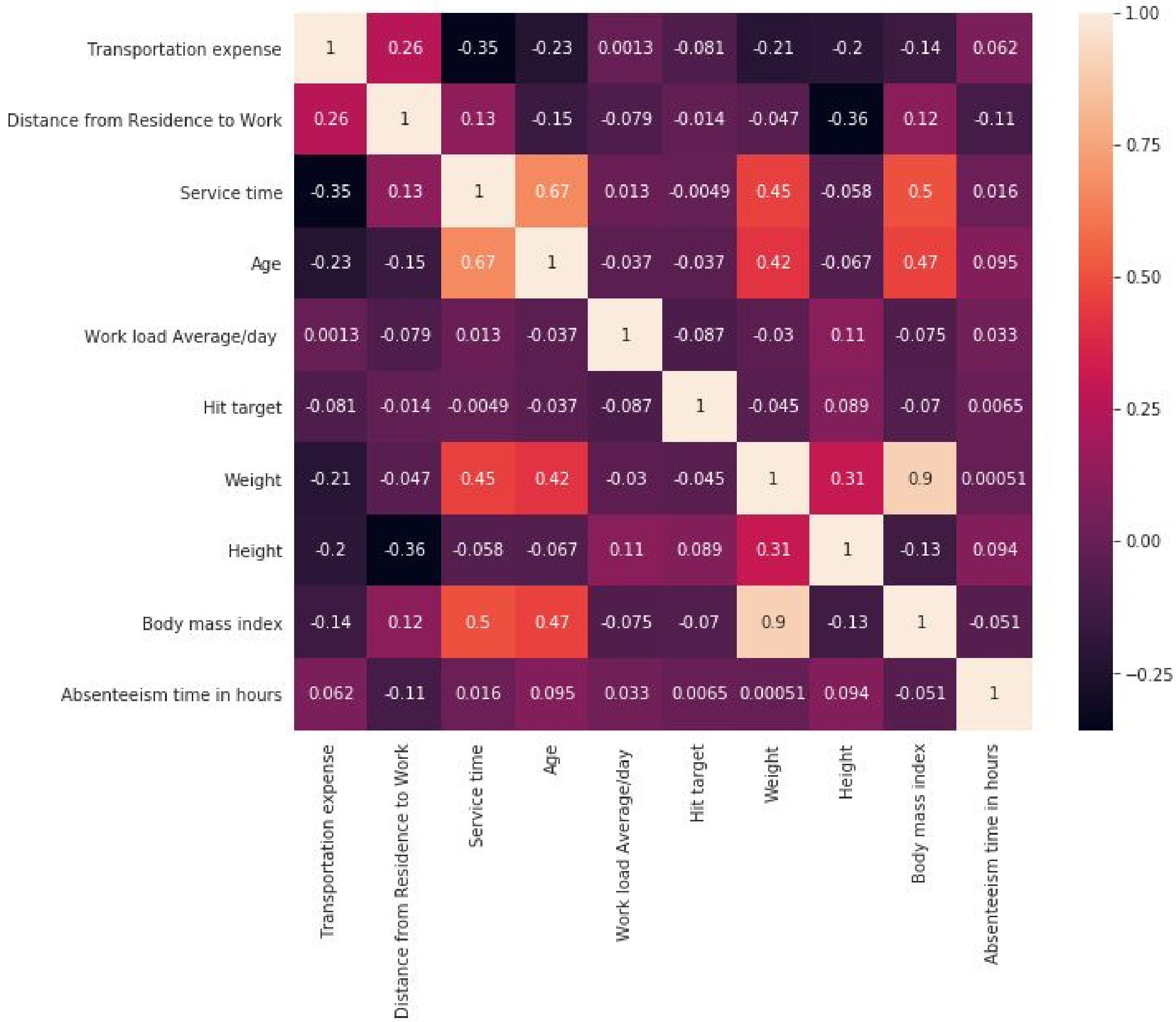

60
70
80
90
100
110
Weight


165
170
175
180
185
190
195
Height


20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
Body mass index


0
20
40
60
80
100
120
Absenteeism time in hours

Coorelation Matrix



R Code

```
#INITIALIZATION--------------------------------------------------------------------------------

#cleanup enviorment
rm(list = ls())

#installing required pacakages
install.packages("caret")
install.packages("Hmisc")
install.packages('corrplot')
install.packages('PerformanceAnalytics')
install.packages('caTools')
install.packages('randomForest')
install.packages('e1071')
install.packages('readxl')
install.packages('aod')

#read dataset
library("readxl")
FilePath =
"https://s3-ap-southeast-1.amazonaws.com/edwisor-india-bucket/projects/data/DataN0
101/Absenteeism_at_work_Project.xls"
File = download.file(FilePath,"EmpAb.xls")
EmpAb =  read_excel(path = 'EmpAb.xls')

#dimensions of dataset: 731 Rows, 16 columns
dim(EmpAb)
```

```
#getting datatypes and structure of columns
str(EmpAb)

#getting first five rows
head(EmpAb)

#getting statistical figures of columns of dataset
library(Hmisc)
describe(EmpAb)

#getting column names
names(EmpAb)

#DATA
PREPARATION---------------------------------------------------------------------------------------------------------
-----------------------------------

Categorical = c('ID','Reason for absence','Month of absence','Day of the week',
              'Seasons','Son','Pet','Disciplinary failure','Education',
              'Social drinker','Social smoker')

Continuous = c('Transportation expense','Distance from Residence to Work',
             'Service time','Age','Work load Average/day','Hit target','Weight',
             'Height','Body mass index','Absenteeism time in hours')

#creating new dataset for EXPLORTORY DATA ANALYSIS with proper categories names
data = EmpAb

data$'ID' = factor(data$'ID')
```

```
data$'Reason for absence'= factor(data$'Reason for absence',
                          levels = c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28),
                          labels = c('Undefined absence',
                                     'Certain infectious and parasitic diseases',
                                     'Neoplasms',
                                     'Diseases of the blood, blood-forming organs and immune mechanism disorders',
                                     'Endocrine, nutritional and metabolic diseases',
                                     'Mental and behavioural disorders',
                                     'Diseases of the nervous system',
                                     'Diseases of the eye and adnexa',
                                     'Diseases of the ear and mastoid process',
                                     'Diseases of the circulatory system',
                                     'Diseases of the respiratory system',
                                     'Diseases of the digestive system',
                                     'Diseases of the skin and subcutaneous tissue',
                                     'Diseases of the musculoskeletal system and connective tissue',
                                     'Diseases of the genitourinary system',
                                     'Pregnancy, childbirth and the puerperium',
                                     'Certain conditions originating in the perinatal period',
                                     'Congenital malformations, deformations and chromosomal abnormalities',
                                     'Symptoms, signs, abnormal clinical and laboratory findings, not elsewhere classified',
                                     'Injury, poisoning and certain other consequences of external causes',
                                     'External causes of morbidity and mortality',
                                     'Factors influencing health status and contact with health services',
                                     'patient follow-up',
```

```
                    'medical consultation',
                    'blood donation',
                    'laboratory examination',
                    'unjustified absence',
                    'physiotherapy',
                    'dental consultation'))

data$'Month of absence'= factor(data$'Month of absence')

data$'Day of the week'= factor(data$'Day of the week',levels = c(2,3,4,5,6),labels =
c('Monday',
                                            'Tuesday',
                                            'Wednesday',
                                            'Thrusday',
                                            'Friday'))

data$'Seasons'= factor( data$'Seasons',levels = c(1,2,3,4),labels = c('Summer',
                              'Autumn',
                              'Winter',
                              'Spring'))

data$'Disciplinary failure'= factor(data$'Disciplinary failure',levels = c(0,1),labels =
c('No','Yes'))

data$'Education'= factor(data$'Education',levels = c(1,2,3,4),labels = c('High School',
                               'Graduate',
                               'Postgraduate',
                               'Master and Doctor'))

data$'Social drinker'= factor(data$'Social drinker',levels = c(0,1),labels = c('No','Yes'))
```

```
data$'Social smoker'= factor(data$'Social smoker',levels = c(0,1),labels = c('No','Yes'))

data$'Son'= factor(data$'Son')

data$'Pet'= factor(data$'Pet')

data$'Transportation expense' = as.numeric(data$'Transportation expense')


sapply(data,class)

#EXPLORTORY DATA ANALYSIS-----------------------------------------------------------------------------------------------------------------------------

#Checking distribution of target variable
hist(data$'Absenteeism time in hours',breaks =  50)
#it seems target variable is nearly normally distributed

#plotting categorical variable vs target variable 'Absenteeism time in hours'
library(ggplot2)
c1 = ggplot(data, aes(y=data$'Absenteeism time in hours',x = data$'Reason for absence')) +  geom_bar(stat = 'identity')
c2 = ggplot(data, aes(y=data$'Absenteeism time in hours',x = data$'Month of absence')) + geom_bar(stat = 'identity')
c3 = ggplot(data, aes(x=data$'Day of the week',y =data$'Absenteeism time in hours')) + geom_bar(stat = 'identity')
c4 = ggplot(data, aes(x=data$'Seasons',y =data$ 'Absenteeism time in hours')) + geom_bar(stat = 'identity')
```

```
c5 = ggplot(data, aes(x=data$'Disciplinary failure',y =data$ 'Absenteeism time in hours'))
+  geom_bar(stat = 'identity')

c6 = ggplot(data, aes(x=data$'Education',y =data$ 'Absenteeism time in hours')) +
geom_bar(stat = 'identity')

c7 = ggplot(data, aes(x=data$'Social drinker',y =data$ 'Absenteeism time in hours')) +
geom_bar(stat = 'identity')

c8 = ggplot(data, aes(x=data$'Social smoker',y =data$ 'Absenteeism time in hours')) +
geom_bar(stat = 'identity')

c9 = ggplot(data, aes(x=data$'Son',y =data$ 'Absenteeism time in hours')) +
geom_bar(stat = 'identity')

c10= ggplot(data, aes(x=data$'Pet',y =data$ 'Absenteeism time in hours')) +
geom_bar(stat = 'identity')


gridExtra::grid.arrange(c1,c2,c3,c4,c5,c6,c7,c8,c9,c10,ncol=5)


#plotting continuous variable vs target variable 'Absenteeism time in hours'

c11 = ggplot(data, aes(x=data$'Transportation expense',y =data$ 'Absenteeism time in
hours')) +  geom_point(color = 'maroon')

c12 = ggplot(data, aes(x=data$'Distance from Residence to Work',y =data$ 'Absenteeism
time in hours')) +  geom_point()

c13 = ggplot(data, aes(x=data$'Service time',y =data$ 'Absenteeism time in hours')) +
geom_point()

c14 = ggplot(data, aes(x=data$'Age',y =data$ 'Absenteeism time in hours')) +
geom_point()

c15 = ggplot(data, aes(x=data$'Work load Average/day',y =data$ 'Absenteeism time in
hours')) +  geom_point()

c16 = ggplot(data, aes(x=data$'Hit target',y =data$ 'Absenteeism time in hours')) +
geom_point()

c17 = ggplot(data, aes(x=data$'Weight',y =data$ 'Absenteeism time in hours')) +
geom_point()

c18 = ggplot(data, aes(x=data$'Height',y =data$ 'Absenteeism time in hours')) +
geom_point()

c19 = ggplot(data, aes(x=data$'Body mass index',y =data$ 'Absenteeism time in hours')) +
geom_point()
```

```
gridExtra::grid.arrange(c11,c12,c13,c14,c15,c16,c17,c18,c19,ncol=3)

#plotting distribution of continuous variable
c20 = ggplot(data, aes(x=data$'Transportation expense')) +  geom_histogram(bins = 50)
c21 = ggplot(data, aes(x=data$'Distance from Residence to Work')) +
geom_histogram(bins = 50)
c22 = ggplot(data, aes(x=data$'Service time')) +  geom_histogram(bins = 50)
c23 = ggplot(data, aes(x=data$'Age')) +  geom_histogram(bins = 50)
c24 = ggplot(data, aes(x=data$'Work load Average/day')) +  geom_histogram(bins = 50)
c25 = ggplot(data, aes(x=data$'Hit target')) +  geom_histogram(bins = 50)
c26 = ggplot(data, aes(x=data$'Weight')) +  geom_histogram(bins = 50)
c27 = ggplot(data, aes(x=data$'Height')) +  geom_histogram(bins = 50)
c28 = ggplot(data, aes(x=data$'Body mass index')) +  geom_histogram(bins = 50)

gridExtra::grid.arrange(c20,c21,c22,c23,c24,c25,c26,c27,c28,ncol=3)

#MISSING VALUE
ANALYSIS------------------------------------------------------------------------------------------------------------------
----------------------

#Checking no. of missing values
sapply(EmpAb,function(x){sum(is.na(x))})

#below variables cannot contain 0 value so replacing it with NA
length(EmpAb$'Reason for absence'[EmpAb$'Reason for absence'==0])
length(EmpAb$'Month of absence'[EmpAb$'Month of absence'==0])
length(EmpAb$'Absenteeism time in hours'[EmpAb$'Absenteeism time in hours'==0])

EmpAb$'Reason for absence'[EmpAb$'Reason for absence'==0] = NA
```

```
EmpAb$'Month of absence'[EmpAb$'Month of absence'==0] = NA
EmpAb$'Absenteeism time in hours'[EmpAb$'Absenteeism time in hours'==0] = NA

Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

#Imputing categorical with mode and continuous with mean
for (cat in Categorical) {EmpAb[is.na(EmpAb[[cat]]),cat] = Mode(EmpAb[[cat]])}
for (con in Continuous) {EmpAb[is.na(EmpAb[[con]]),con] = mean(EmpAb[[con]], na.rm =
TRUE)}

#checking any missing value left
sum(is.na(EmpAb))

#changing categorical variabel dataytpe
for (i in Categorical) {factor(EmpAb[[i]])}

#OUTLIER
DETECTION------------------------------------------------------------------------------------------------------------
------------------------------
#Boxplots to detect outliers

ggplot(EmpAb, aes(y=EmpAb$'Transportation expense')) +  geom_boxplot()
ggplot(EmpAb, aes(y=EmpAb$'Distance from Residence to Work')) +  geom_boxplot()
ggplot(EmpAb, aes(y=EmpAb$'Service time')) +  geom_boxplot()
ggplot(EmpAb, aes(y=EmpAb$'Age')) +  geom_boxplot()
ggplot(EmpAb, aes(y=EmpAb$'Work load Average/day')) +  geom_boxplot()
ggplot(EmpAb, aes(y=EmpAb$'Hit target')) +  geom_boxplot()
```

```
ggplot(EmpAb, aes(y=EmpAb$'Weight')) +  geom_boxplot()
ggplot(EmpAb, aes(y=EmpAb$'Height')) +  geom_boxplot()
ggplot(EmpAb, aes(y=EmpAb$'Body mass index')) +  geom_boxplot()
ggplot(EmpAb, aes(y=EmpAb$'Absenteeism time in hours')) +  geom_boxplot()

#OUTLIER REMOVAL-------------------------------------------------------------------------------------------------------------------------------------------

#creating extra dataset with ourliers for furtur use
EmpAbWithOutliers = EmpAb

#sing quantile methos to remove outliers
OutlierRemoval = function(var){
  qnt = quantile(var, probs=c(.25, .75), na.rm = T)
  caps = quantile(var, probs=c(.05, .95), na.rm = T)
  H = 1.5 * IQR(var, na.rm = T)
  var[var < (qnt[1] - H)] <- caps[1]
  var[var > (qnt[2] + H)] <- caps[2]
  return (var)}

for (i in Continuous){EmpAb[[i]] = OutlierRemoval(EmpAb[[i]])}

sum(is.na(EmpAb))

#FEATURE SELECTION-------------------------------------------------------------------------------------------------------------------------------------------

#checking correlation between variable
```

```
library("PerformanceAnalytics")
chart.Correlation(EmpAb[Continuous], histogram=TRUE)

#bosy mass index and weight are highly correlated so dropping weight variable
EmpAb = EmpAb[, !colnames(EmpAb) %in% c('Weight'), drop = FALSE]
EmpAbWithOutliers = EmpAbWithOutliers[, !colnames(EmpAbWithOutliers) %in%
c('Weight'), drop = FALSE]



#FEATURE
SCALING-------------------------------------------------------------------------------------------------------------
----------------------------


#normalizing continuous values
for(i in Continuous){
  if (i == 'Weight' | i == 'Absenteeism time in hours')  {
    next}
  else
    EmpAb[i] = (EmpAb[i] - min(EmpAb[i]))/(max(EmpAb[i] - min(EmpAb[i])))}

#SAMPLING----------------------------------------------------------------------------------------------------------
---------------------------------------
library(caTools)

#divided dataset into 80% training set and 20% test set
sample = sample.split(EmpAb,SplitRatio = 0.8)
train =subset(EmpAb,sample ==TRUE)
test=subset(EmpAb, sample==FALSE)
```

```
#MODELLING AND
EVALUATION-----------------------------------------------------------------------------------------------------------
-------------------------

#evaluation (error calculation functions)
MAPE = function(actual,predicted){mean((abs(actual-predicted))/actual)*100}
MAE  = function(actual,predicted){mean((abs(actual-predicted)))}
RMSE = function(actual,predicted){sqrt(mean(((abs(actual-predicted)))^2))}
RSQ  = function(actual,predicted){1 - (sum((predicted - actual) ^ 2))/(sum((actual -
mean(actual)) ^ 2))}

#Linear Regression
#MAPE = 105.24%
#MAE  = 73.53
#RMSE = 5.02
#RSQ  = .14
LR =  lm(train[['Absenteeism time in hours']] ~.,
        data = train[, !colnames(train) %in% c('Absenteeism time in hours')])
LRpredicted = predict(LR,test[, !colnames(test) %in% c('Absenteeism time in hours')])
MAPE(test$'Absenteeism time in hours',LRpredicted)
MAE (test$'Absenteeism time in hours',LRpredicted)
RMSE(test$'Absenteeism time in hours',LRpredicted)
RSQ (test$'Absenteeism time in hours',LRpredicted)


#Decision Tree
#MAPE = 93.24%
#MAE  = 3.57
#RMSE = 5.02
#RSQ  = .11
```

```
library(rpart)
DT = rpart(train[['Absenteeism time in hours']] ~.,
      data = train[, !colnames(train) %in% c('Absenteeism time in hours')])
DTpredicted = predict(DT,test[, !colnames(test) %in% c('Absenteeism time in hours')])
MAPE(test$'Absenteeism time in hours',DTpredicted)
MAE (test$'Absenteeism time in hours',DTpredicted)
RMSE(test$'Absenteeism time in hours',DTpredicted)
RSQ (test$'Absenteeism time in hours',DTpredicted)


#Random Forest
#MAPE = 113.78%
#MAE  = 2.81
#RMSE = 4.34
#RSQ  = .18
DataRF = EmpAb
names(DataRF)<-str_replace_all(names(DataRF), c(" " = "." , "," = "" ,"/" = ""))
sample = sample.split(EmpAb,SplitRatio = 0.8)
train =subset(DataRF,sample ==TRUE)
test=subset(DataRF, sample==FALSE)
library(randomForest)
RF =  randomForest(train[['Absenteeism.time.in.hours']] ~.,
          data = train[, !colnames(train) %in% c('Absenteeism.time.in.hours')])
RFpredicted = predict(RF,test[, !colnames(test) %in% c('Absenteeism.time.in.hours')])
MAPE(test$'Absenteeism.time.in.hours',RFpredicted)
MAE (test$'Absenteeism.time.in.hours',RFpredicted)
RMSE(test$'Absenteeism.time.in.hours',RFpredicted)
RSQ (test$'Absenteeism.time.in.hours',RFpredicted)
```

```
#From above calcualtions RandomForest is the best fit for the dataset

#CONVERTING TARGET VARIABLE TO
CATEGORICAL-----------------------------------------------------------

library(caret)

DataCls = EmpAb
#DataCls = EmpAbWithOutliers

library(tidyverse)
names(DataCls)<-str_replace_all(names(DataCls), c(" " = "." , "," = "" ,"/" = ""))

DataCls$'Absenteeism.time.in.hours' = cut(DataCls$'Absenteeism.time.in.hours',
seq(0,30,5), right=FALSE, labels=c(1:6))
DataCls$'Absenteeism.time.in.hours' = factor(DataCls$'Absenteeism.time.in.hours')

sample = sample.split(DataCls,SplitRatio = 0.8)
train =subset(DataCls,sample ==TRUE)
test=subset(DataCls, sample==FALSE)



library(randomForest)
RFC =  randomForest(train$Absenteeism.time.in.hours ~.,
        data = train[, !colnames(train) %in% c('Absenteeism.time.in.hours')],
        family=binomial)
RFCpredicted = predict(RFC,test[, !colnames(test) %in% c('Absenteeism.time.in.hours')])
confusionMatrix(test$'Absenteeism.time.in.hours',RFCpredicted)
```

```
#After Converting target variable to categorical
#random forest provides 70% of accuracy with 5 classes
#MONTHLY LOSS FOR THE COMPANY-------------------------------------------------------------------

LossData = EmpAbWithOutliers[,c("Month of absence",
                                "Work load Average/day",
                                "Service time",
                                "Absenteeism time in hours")]

str(LossData)

LossData$WorkLoss = round((LossData$"Work load Average/day"/LossData$"Service time")
                          *LossData$"Absenteeism time in hours")

MonthlyLoss = aggregate(LossData$WorkLoss,by = list(Category = LossData$"Month of absence"),FUN = sum)

names(MonthlyLoss) = c("Month","WorkLoss")

ggplot(MonthlyLoss,aes(MonthlyLoss$Month,MonthlyLoss$WorkLoss))+geom_bar(stat = "identity",fill = "blue")+labs(y="WorkLoss",x="Months")
```