EMPLOYEE ABSENTEEISM

-Sarthak Gupta

CONTENTS

Chapter 1:

Introduction

1.1 Problem Statement

XYZ is a courier company. It is going through the problem of employee absenteeism. The company has shared its dataset. Our task is to answer the following two questions :

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

1.2 Data

A sample of the data is given below :

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expens	Distance from Residence to Wor	Service time	Age	Work load Average/day	Hit target	Disciplinary failure
11	26	7	3	1	289	36	13	33	239,554	97	0
36	0	7	3	1	118	13	18	50	239,554	97	1
3	23	7	4	1	179	51	18	38	239,554	97	0
7	7	7	5	1	279	5	14	39	239,554	97	0
11	23	7	5	1	289	36	13	33	239,554	97	0
3	23	7	6	1	179	51	18	38	239,554	97	0
10	22	7	6	1		52	3	28	239,554	97	0
20	23	7	6	1	260	50	11	36	239,554	97	0
14	19	7	2	1	155	12	14	34	239,554	97	0

Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
1	2	1	0	1	90	172	30	4
1	1	1	0	0	98	178	31	0
1	0	1	0	0	89	170	31	2
1	2	1	1	0	68	168	24	4
1	2	1	0	1	90	172	30	2
1	0	1	0	0	89	170	31	
1	1	1	0	4	80	172	27	8
1	4	1	0	0	65	168	23	4
1	2	1	0	0	95	196	25	40

Chapter 2

Methodology

2.1 Exploratory Data Analysis

Any type of data analysis has to start with preliminary exploration about the data. The primary purpose is to look at the structure of data set, number of rows and coloumns, class of variables and get a overview of what the dataset is all about.

The R code is :

```
str(data2)

head(data2)

tail(data2)

summary(data2)
```

From this we observe that all the variables are numeric. However, from the problem statement and looking at the values of dataset it is clear that many of these variables are categorical in nature where different categories are instead represented by the numbers as dummies. Therefore it is imperative that these variables are converted into factor.

The variables of this nature are :

- ID
- Reason.for.absence.
- Month.of.absence
- Day.of.the.week
- Seasons
- Disciplinary.failure
- Education
- Social.drinker
- Social.Smoker

The R code is:

```
data3 = data2


factor_data =
c("ID","Reason.for.absence","Month.of.absence","Day.of.the.week","Seasons","Disciplinary.failure","Ed
ucation","Social.drinker","Social.smoker")


#converting variables which are of categorical nature into factor

for (i in factor_data) {

  data2[,i] = as.factor(data2[,i])

}
```

2.2 Missing Value Analysis

We need to check whether the dataset has missing values. The R code is :

```
table(is.na(data3))
```

From this it is clear that the dataset has multiple missing values and further analysis is required.

First we determine the count of missing values for each variable :

```
 #count of missing values

for(i in colnames(data3)){


  b=table(is.na(data3[,i]))

  print(i)

  print(b)

}
```

We need to measure the percentage of missing values for each variable. The percentage should be less than 30% to go for imputation. Otherwise if it is more than 30% then the variable needs to be deleted.

The R code :

```
#create dataframe with missing percentage
```

```
missing_val = data.frame(apply(data3,2,function(x){sum(is.na(x))}))

#convert row names into coloumn
missing_val$columns = row.names(missing_val)
row.names(missing_val)=NULL

#Rename the variable name
names(missing_val)[1] = "Missing_Percentage"

#Calculate Percentage (should be less than 30)
missing_val$Missing_Percentage = (missing_val$Missing_Percentage/nrow(data3))*100

#Arrange in descending order
missing_val = missing_val[order(-missing_val$Missing_Percentage),]
```

From the values it is clear that all variables have missing values much below 30%. Therefore, we need to go for imputation

Imputation for missing values

Imputation can be done through mean, median, mode or K NN. The way of choosing a method is to remove a value that exists and replace it with a missing value. Then we use each of these methods to impute the missing value. The method which gives value closest to missing value should be chosen.

The R code :

```
data4 = data3
#Actual Value data4[1,20] = 30
data4[1,20] = NA
#Mean Method
```

```
data4$Body.mass.index[is.na(data4$Body.mass.index)] = mean(data4$Body.mass.index,na.rm = T)
data4[1,20]
# Value Obtained = 26.68
#Median Method
data4 = data3 #reloading the data
data4[1,20] = NA
data4$Body.mass.index[is.na(data4$Body.mass.index)] = median(data4$Body.mass.index,na.rm = T)
data4[1,20]
# Value Obtained = 25
#KNN
data4 = data3
data4[1,20] = NA
library(DMwR)
data4 = knnImputation(data4,k=5)
data4[1,20]
#Value Obtained = 29.59  closest value to actual value
```

From this we get that KNN method gives value closest to actual value. Therefore, we use KNN to impute the missing values.

2.3 Outlier Analysis

Any variable might contain values that are inconsistent with the rest of the values i.e. they are either too big or too small. These values can potentially distort any kind of model and subsequent prediction. However, there might be cases where outliers itself might provide useful information. For example, in case of health industry outliers in human data might reveal the individuals with likelihood of diseases. In case of time series dataset it is possible that outliers might reveal seasonality , cyclicity . Therefore, we go for outlier analysis but at the same time preserve the dataset with outliers. We can later analyze which one would be more suitable for our needs.

The R code :

```
data5 = data4
#outlier analysis
numeric_index = sapply(data5,is.numeric) #selecting only numeric
numeric_data = data5[,numeric_index]
cnames = colnames(numeric_data)
#Delete the outliers using boxplot method
for(i in cnames){
  print(i)
  val = data5[,i][data5[,i] %in% boxplot.stats(data5[,i])$out]
  #print(length(val))
  data5 = data5[which(!data5[,i] %in% val),]
}
#Imputing the oultiers after replacing them with missing values
data6 = data4
numeric_index = sapply(data6,is.numeric) #selecting only numeric
numeric_data = data6[,numeric_index]
cnames = colnames(numeric_data)
#Replace outliers with NA
for(i in cnames){
  print(i)
  val = data6[,i][data6[,i] %in% boxplot.stats(data6[,i])$out]
  #print(length(val))
  data6[,i][data6[,i] %in% val] = NA
}
#Imputing for the newly created missing values
data7 = data6
data7 = knnImputation(data7, k = 3)
sum(is.na(data7))
```

data_deleted = data5

data_imputed = data7

Using boxplot method we identify the outliers. After this we can either delete the outliers or impute them using the same approach we used for missing values. The advantage with deleting is that we do not impute values on our own, but we lose a reasonable amount of data. With imputation we do not lose data but impute values which can potentially modify the nature of data. We observe that dataset with deleted outliers has 100 less observations compared to original dataset. As the total number of observations are 740, this is a huge loss of data. Therefore, we reject the use of data where outliers are deleted.

NOTE: Generally, after doing outlier analysis we go for feature selection where techniques like correlation, chi square test etc. to select which variables to chose for modelling. But, here our primary purpose is to search for the answer to the questions asked. Therefore, we will not go for feature selection here and look at it according to the relevant question.

3. MODELLING (Finding the Answer to the given Questions)

Through the process of modeling and statistical techniques we seek to answer two questions

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

Now, we come to searching the answer for the first question.

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

This problem is more about finding the causation factors of Absenteeism. Once we find them and the nature of relationship the company can suitably make changes to reduce Absenteeism.

Causation factors can be found by various metrics such as correlation analysis. As a factor that causes absenteeism will have high correlation with Absenteeism. However, there is bottleneck the reverse is not true. It is not necessary that a variable which has high correlation with the Absenteeism necessarily causes it.

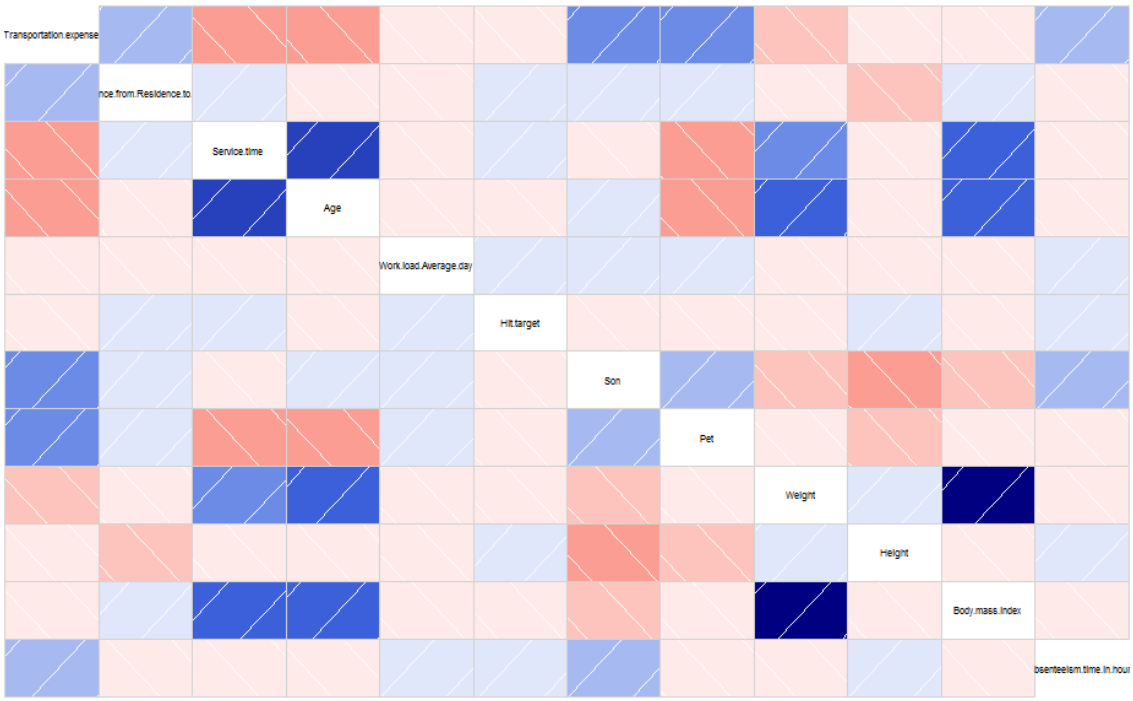
This becomes more clear by seeing an example of ice creams sold on a beach and the number of deaths due to drowning. Both are found to have a high correlation. But basic common sense would conclude that ice creams sold cannot be the causation factor. It happens so because more people come to beaches in summer, when more ice creams are sold. More people on beaches leads to more incidents of drowning and deaths. Therefore it is quite possible that a hidden factor/s which has/have high correlation with these variables are the real causation factor.

The same analysis is true for any other statistic metric as well such as p-value, chi-square etc. Therefore, purely by seeing the dataset and using mathematical techniques to find the relationship it is not possible to find causation factors. It is possible that all the given variables have high correlation with the target and none are causation factors. If we cannot find the causative factors it is not possible to tell as to how can we reduce absenteeism. Therefore, it is not possible

to answer this question. However, for answering the question we need to make certain assumptions :

- The variables given in the dataset are assumed to be exhaustive in nature, meaning there is no variable outside the dataset which is the cause of Absenteeism. It could be possible that any variable within the dataset is not the causation factor of Absenteeism
- If two or more variables are found to have high correlation among themselves then the one having higher correlation with target variable is assumed to be the causation factor.
- We observed during data exploration that there were 43 observations in Reasons for Absence variable with value =0. Now, value 0 has not been defined in the problem statement. Therefore, one method is to simply delete the observations. But this will lead to loss of data. Another would be to go for imputation after replacing them with NA, whereas it could be highly likely that the 0 would have a certain meaning but due to clerical error in the problem statement it was not mentioned. This would be putting values based on statistical measures where perhaps the meaning is something different. Therefore, we assume that the 0 signifies Miscellaneous reasons not covered in other reasons. We always have the option of replacing 0 with actual reason once we take the solution back to the company.

First we go for finding spearman correlation between numeric variables and absenteeism. We find that all numeric variables have low correlation with absenteeism.in.hours. (Maximum correlation = 0.2) .



The maximum three are :

Transportain Expense(0.2), Age(-0.09), Son(0.16)

Therefore, it would be possible to conclude that other variables do not contribute towards Absenteeism hours. They are not causative factors. Therefore, the company should not focus on these variables.

As far as these three variables are concerned the correlation is not very high. But by this we can conclude the following steps that can be taken :

- Provide reimbursement for transportation allowance
- Increase the age of workforce.
- As far as number of children are concerned, it's a personal choice and perhaps the company cannot do anything about it. Moreover, it can lead to ethical and legal issues.

Now, we convert categorical variables into numeric (dummies). Now, we find correlation.

maximum corelation is with ID, Reasons for absence, Disciplinary Failure

Random Forest

Random Forest can be used to see importance of variables in predicting the depe ndent or target variable which in this case is Absenteesis.in.hours

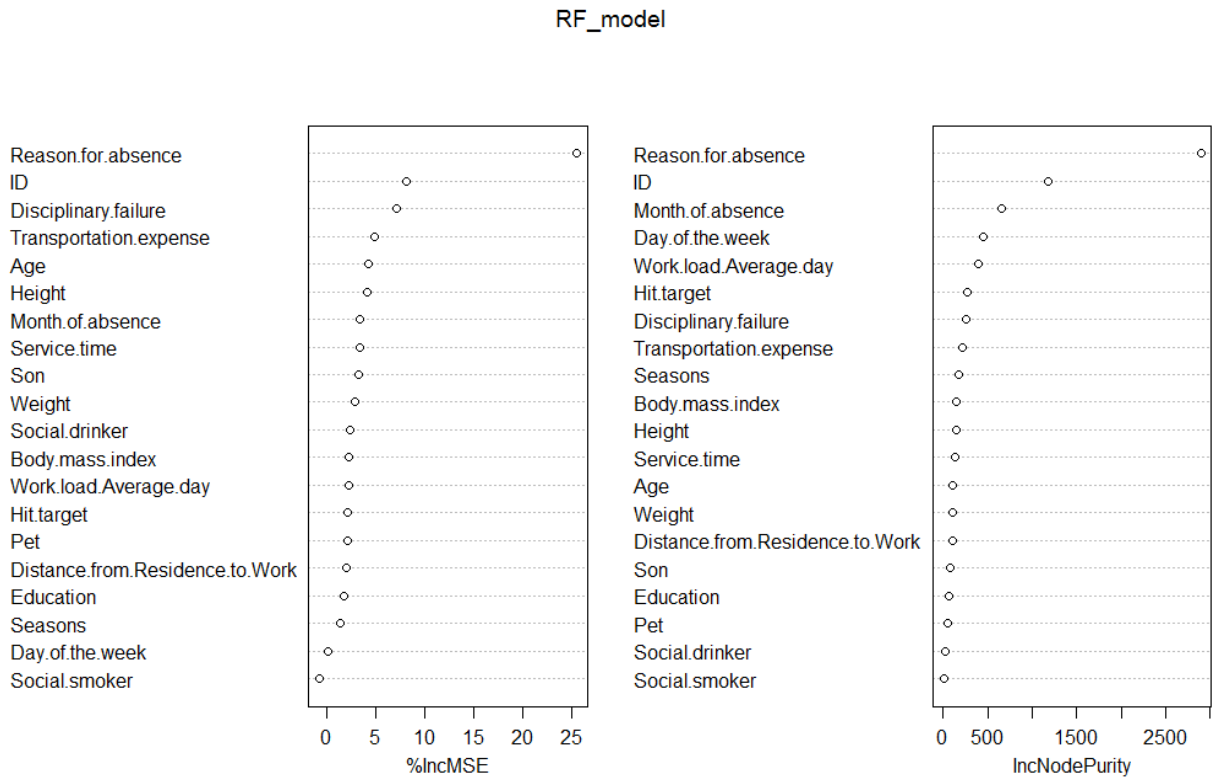
The R code is :

```
RF_model = randomForest(Absenteeism.time.in.hours~.,data4,importance = TRU
E,ntree = 100)

importance(RF_model, type = 1)

varImpPlot(RF_model)
```

```
                                 %IncMSE
ID                             8.1845059
Reason.for.absence            25.5540900
Month.of.absence               3.4884030
Day.of.the.week                0.1369741
Seasons                        1.4301611
Transportation.expense         4.9666494
Distance.from.Residence.to.Work 2.1108916
Service.time                   3.4829330
Age                            4.2780550
Work.load.Average.day          2.2940805
Hit.target                     2.1400808
Disciplinary.failure           7.1751112
Education                      1.7953722
Son                            3.2528300
Social.drinker                 2.4507319
Social.smoker                 -0.7207928
Pet                            2.1270347
Weight                         2.9037462
Height                         4.1173533
Body.mass.index                2.3007140
```



Reason for absence, ID, Disciplinary Failure, Transportation Expense are the most important variables. This matches with the correlation one also. Although, theoretically it is not necessary to.

Linear Regression

We try different combinations of variables based on analysis we have done till now. In the end we apply backward propagation and forward elimination as well.

The R code is :

lm_model = lm(Absenteeism.time.in.hours~.,data=data4)

summary(lm_model)

Adjusted R-squared: 0.4132. No variable is significant

lm_model2 =
lm(Absenteeism.time.in.hours~Reason.for.absence+Disciplinary.failure+Transportation.expense+Son+Age,data=data_imputed)

summary(lm_model2)

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.251429   1.602843   0.157 0.875397
Reason.for.absence1    6.461334   1.464390   4.412 1.18e-05 ***
Reason.for.absence2    1.683961   2.293848   0.734 0.463118
Reason.for.absence3    6.801765   2.957749   2.300 0.021759 *
Reason.for.absence4    3.634095   2.303862   1.577 0.115152
Reason.for.absence5    5.952505   2.021781   2.944 0.003344 **
Reason.for.absence6    5.552570   1.622989   3.421 0.000659 ***
Reason.for.absence7    4.238383   1.495870   2.833 0.004737 **
Reason.for.absence8    4.621632   1.709081   2.704 0.007012 **
Reason.for.absence9    7.115111   1.875506   3.794 0.000161 ***
Reason.for.absence10   5.993439   1.429060   4.194 3.09e-05 ***
Reason.for.absence11   5.039238   1.427148   3.531 0.000441 ***
Reason.for.absence12   4.410474   1.624629   2.715 0.006794 **
Reason.for.absence13   5.098655   1.373402   3.712 0.000221 ***
Reason.for.absence14   4.237518   1.458464   2.905 0.003781 **
Reason.for.absence15   6.523570   2.289324   2.850 0.004505 **
Reason.for.absence16   1.134027   2.017380   0.562 0.574206
Reason.for.absence17   7.146192   2.950075   2.422 0.015669 *
Reason.for.absence18   5.770127   1.444915   3.993 7.19e-05 ***
Reason.for.absence19   6.046915   1.389656   4.351 1.55e-05 ***
Reason.for.absence21   5.087560   1.708347   2.978 0.003000 **
Reason.for.absence22   6.016307   1.405854   4.279 2.13e-05 ***
Reason.for.absence23   1.968212   1.344755   1.464 0.143741
Reason.for.absence24   6.558122   2.023694   3.241 0.001248 **
Reason.for.absence25   2.575368   1.409559   1.827 0.068110 .
Reason.for.absence26   6.092000   1.401970   4.345 1.59e-05 ***
Reason.for.absence27   1.710199   1.373440   1.245 0.213473
Reason.for.absence28   1.917587   1.347995   1.423 0.155308
Disciplinary.failure1 -0.379550   1.390807  -0.273 0.785010
Transportation.expense 0.004509   0.001835   2.458 0.014210 *
Son                    0.160611   0.102015   1.574 0.115844
Age                   -0.013150   0.017884  -0.735 0.462408
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.638 on 708 degrees of freedom
Multiple R-squared:  0.4113,   Adjusted R-squared:  0.3855
F-statistic: 15.95 on 31 and 708 DF,  p-value: < 2.2e-16
```

Reasons for absence 1,3,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,21,22,24,26,28

Disciplinary Failure 1 and Transportation Expense are found to be significant.

But still Adjusted R-squared is low.

lm_model3 =
lm(Absenteeism.time.in.hours~Reason.for.absence,data=data_imputed)

summary(lm_model3)

```
Call:
lm(formula = Absenteeism.time.in.hours ~ Reason.for.absence,
    data = data_imputed)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4256 -1.3523 -0.2754  1.0181 13.1656

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.6867     0.4056   1.693 0.090857 .
Reason.for.absence1    6.7388     0.7466   9.026  < 2e-16 ***
Reason.for.absence2    2.4793     1.9238   1.289 0.197898
Reason.for.absence3    7.3133     2.6903   2.718 0.006719 **
Reason.for.absence4    3.8133     1.9238   1.982 0.047845 *
Reason.for.absence5    5.6466     1.5881   3.556 0.000402 ***
Reason.for.absence6    5.8205     1.0240   5.684 1.92e-08 ***
Reason.for.absence7    4.6217     0.7975   5.795 1.03e-08 ***
Reason.for.absence8    4.6466     1.1590   4.009 6.74e-05 ***
Reason.for.absence9    7.1780     1.3902   5.163 3.15e-07 ***
Reason.for.absence10   6.2952     0.6689   9.411  < 2e-16 ***
Reason.for.absence11   5.1322     0.6607   7.768 2.79e-14 ***
Reason.for.absence12   4.6110     1.0240   4.503 7.83e-06 ***
Reason.for.absence13   5.2767     0.5414   9.747  < 2e-16 ***
Reason.for.absence14   4.6109     0.7326   6.294 5.42e-10 ***
Reason.for.absence15   7.3133     1.9238   3.801 0.000156 ***
Reason.for.absence16   1.3133     1.5881   0.827 0.408550
Reason.for.absence17   7.3133     2.6903   2.718 0.006719 **
Reason.for.absence18   6.0142     0.7080   8.494  < 2e-16 ***
Reason.for.absence19   6.4324     0.5842  11.010  < 2e-16 ***
Reason.for.absence21   5.1466     1.1590   4.440 1.04e-05 ***
Reason.for.absence22   6.7729     0.5964  11.357  < 2e-16 ***
Reason.for.absence23   2.1849     0.4604   4.746 2.51e-06 ***
Reason.for.absence24   7.3133     1.5881   4.605 4.89e-06 ***
Reason.for.absence25   2.7972     0.6266   4.464 9.36e-06 ***
Reason.for.absence26   6.5060     0.6155  10.570  < 2e-16 ***
Reason.for.absence27   1.5886     0.5167   3.074 0.002189 **
Reason.for.absence28   2.1476     0.4783   4.490 8.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.66 on 712 degrees of freedom
Multiple R-squared:  0.3983,   Adjusted R-squared:  0.3755
F-statistic: 17.46 on 27 and 712 DF,  p-value: < 2.2e-16
```

Backward Propogation

```
model_final<-step(lm_model,direction = 'backward',trace = 0)

summary(model_final)
```

```
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.478252   1.014441   0.471 0.637469
Reason.for.absence1   6.667903   0.738524   9.029  < 2e-16 ***
Reason.for.absence2   1.700412   1.903028   0.894 0.371878
Reason.for.absence3   7.334135   2.648293   2.769 0.005764 **
Reason.for.absence4   4.223290   1.905281   2.217 0.026966 *
Reason.for.absence5   6.164636   1.571362   3.923 9.59e-05 ***
Reason.for.absence6   5.788596   1.016340   5.696 1.80e-08 ***
Reason.for.absence7   4.570694   0.794396   5.754 1.30e-08 ***
Reason.for.absence8   4.884222   1.148275   4.254 2.39e-05 ***
Reason.for.absence9   7.395763   1.372876   5.387 9.76e-08 ***
Reason.for.absence10  6.117024   0.664414   9.207  < 2e-16 ***
Reason.for.absence11  5.259954   0.664236   7.919 9.30e-15 ***
Reason.for.absence12  4.758758   1.015474   4.686 3.34e-06 ***
Reason.for.absence13  5.289052   0.542563   9.748  < 2e-16 ***
Reason.for.absence14  4.737826   0.724278   6.541 1.17e-10 ***
Reason.for.absence15  6.826760   1.898423   3.596 0.000346 ***
Reason.for.absence16  1.152651   1.567232   0.735 0.462298
Reason.for.absence17  7.127420   2.651974   2.688 0.007366 **
Reason.for.absence18  6.052601   0.703433   8.604  < 2e-16 ***
Reason.for.absence19  6.436655   0.581031  11.078  < 2e-16 ***
Reason.for.absence21  5.107116   1.151268   4.436 1.06e-05 ***
Reason.for.absence22  6.427122   0.601911  10.678  < 2e-16 ***
Reason.for.absence23  2.339562   0.464933   5.032 6.16e-07 ***
Reason.for.absence24  6.772789   1.571897   4.309 1.88e-05 ***
Reason.for.absence25  2.872973   0.628234   4.573 5.67e-06 ***
Reason.for.absence26  6.257563   0.607982  10.292  < 2e-16 ***
Reason.for.absence27  1.752590   0.546553   3.207 0.001404 **
Reason.for.absence28  2.142081   0.476669   4.494 8.17e-06 ***
Transportation.expense 0.006335  0.001885   3.360 0.000821 ***
Service.time          0.057912   0.033666   1.720 0.085833 .
Age                  -0.051044   0.022966  -2.223 0.026559 *
Son                   0.158798   0.101505   1.564 0.118161
Pet                  -0.467884   0.153430  -3.049 0.002378 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.615 on 707 degrees of freedom
Multiple R-squared:  0.4223,   Adjusted R-squared:  0.3962
F-statistic: 16.15 on 32 and 707 DF,  p-value: < 2.2e-16
```

Forward elimination

```
model_final2<-step(lm_model3,direction = 'forward',trace = 0)

summary(model_final2)
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.6867     0.4056   1.693 0.090857 .
Reason.for.absence1   6.7388     0.7466   9.026  < 2e-16 ***
Reason.for.absence2   2.4793     1.9238   1.289 0.197898
Reason.for.absence3   7.3133     2.6903   2.718 0.006719 **
Reason.for.absence4   3.8133     1.9238   1.982 0.047845 *
Reason.for.absence5   5.6466     1.5881   3.556 0.000402 ***
Reason.for.absence6   5.8205     1.0240   5.684 1.92e-08 ***
Reason.for.absence7   4.6217     0.7975   5.795 1.03e-08 ***
Reason.for.absence8   4.6466     1.1590   4.009 6.74e-05 ***
Reason.for.absence9   7.1780     1.3902   5.163 3.15e-07 ***
Reason.for.absence10  6.2952     0.6689   9.411  < 2e-16 ***
Reason.for.absence11  5.1322     0.6607   7.768 2.79e-14 ***
Reason.for.absence12  4.6110     1.0240   4.503 7.83e-06 ***
Reason.for.absence13  5.2767     0.5414   9.747  < 2e-16 ***
Reason.for.absence14  4.6109     0.7326   6.294 5.42e-10 ***
Reason.for.absence15  7.3133     1.9238   3.801 0.000156 ***
Reason.for.absence16  1.3133     1.5881   0.827 0.408550
Reason.for.absence17  7.3133     2.6903   2.718 0.006719 **
Reason.for.absence18  6.0142     0.7080   8.494  < 2e-16 ***
Reason.for.absence19  6.4324     0.5842  11.010  < 2e-16 ***
Reason.for.absence21  5.1466     1.1590   4.440 1.04e-05 ***
Reason.for.absence22  6.7729     0.5964  11.357  < 2e-16 ***
Reason.for.absence23  2.1849     0.4604   4.746 2.51e-06 ***
Reason.for.absence24  7.3133     1.5881   4.605 4.89e-06 ***
Reason.for.absence25  2.7972     0.6266   4.464 9.36e-06 ***
Reason.for.absence26  6.5060     0.6155  10.570  < 2e-16 ***
Reason.for.absence27  1.5886     0.5167   3.074 0.002189 **
Reason.for.absence28  2.1476     0.4783   4.490 8.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.66 on 712 degrees of freedom
Multiple R-squared:  0.3983,   Adjusted R-squared:  0.3755
F-statistic: 17.46 on 27 and 712 DF,  p-value: < 2.2e-16
```

The value of Adjusted R-squared is similar in all models except for the first one where all variables are taken and none are found to be significant. Therefore, we select the last model as it has least variable.

According to this, Reason for absence 1, 3, 4, 5 ,6 , 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28 are significant.

The final steps that can be taken by company :

- The first priority of the company should be to look at workers suffering from these diseases and provide incentives for preventive care of these diseases.
- Provide reimbursement for transportation allowance
- Increase the age of workforce. Replace younger ones by older ones. Hire experienced workers.
- As far as number of children are concerned, it's a personal choice and perhaps the company cannot do anything about it. Moreover, it can lead to ethical and legal issues.

We will tackle the **second** question now. We need to project losses every month in 2011. It is given that the same trend will continue. We need to make certain assumptions here :

- For the variable month.of.absence 1,2,3.. signify different months like January, February ,March…
- The given dataset is for the year 2010. This assumption has been made as any other assumption would require dates or atleast the information about year, which has not been provided.
- There are three observations in the variable Month.of.absence with value as 0. We can either choose to delete it or consider as values for December 2009. However, considering the fact that it is not as intuitive to assume it as December 2009 as was for months in the first assumption, it would be wise to delete these observations. Another fact helping our cause is that these are only 3 observations and deleting does not lead to significant loss of data.
- Absenteeism leads to losses. It is possible esp. in a overstaffed company that absenteeism might be reducing the losses. But we will reject this possibility. We will assume that absenteeism will lead to loss. Losses will be calculated in terms of number of hours of absenteeism

We need to project losses in 2011 on per month basis. Therefore, we group the data on the basis of months.

A condition which has been given in the question itself is that the same trend will continue. Moreover, by the analysis of correlation, regression in first question it is clear that the given variables do not explain the absenteeism to a great degree. Therefore, the problem reduces to a univariate timeseries and modelling for prediction on future. The other predictor variables have limited role to play.

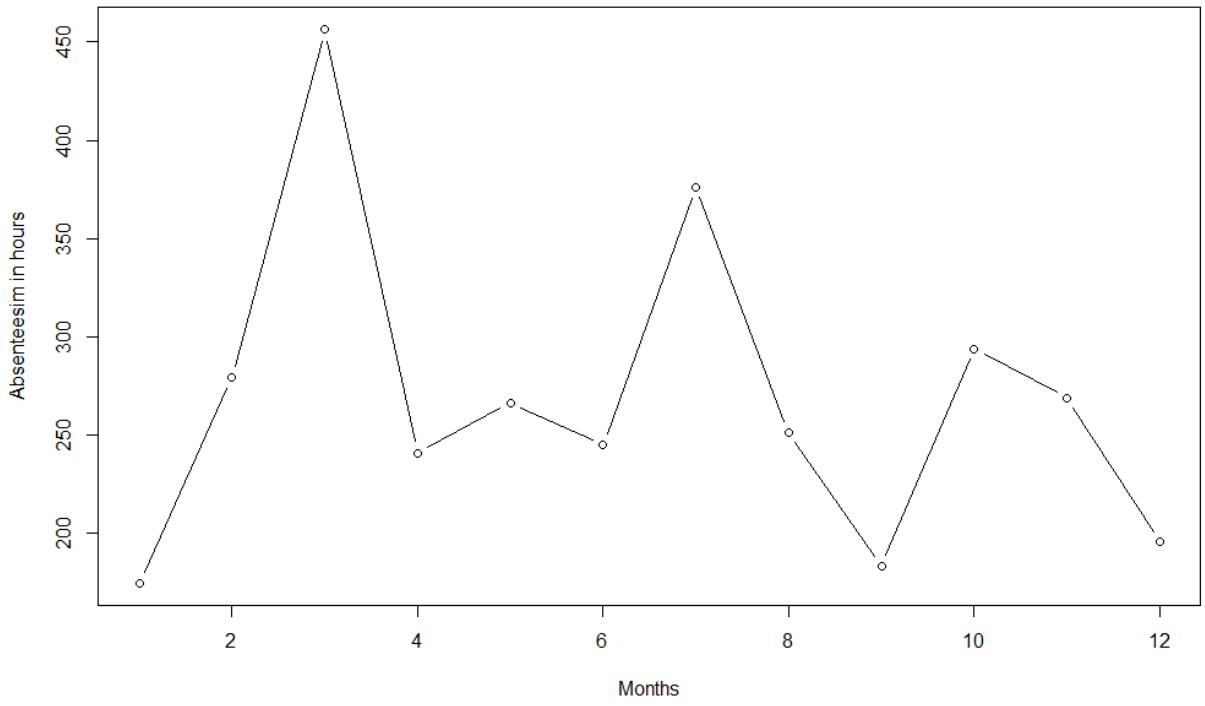
The R code:

```
data_ordered = data_imputed[,c(3,21)]


data_ordered_sum = aggregate(data_ordered$Absenteeism.time.in.hours,
by=list(Month=data_ordered$Month.of.absence), FUN=sum)


data_ts_sum = data_ordered_sum[-1,2]

View(data_ts_sum)
```



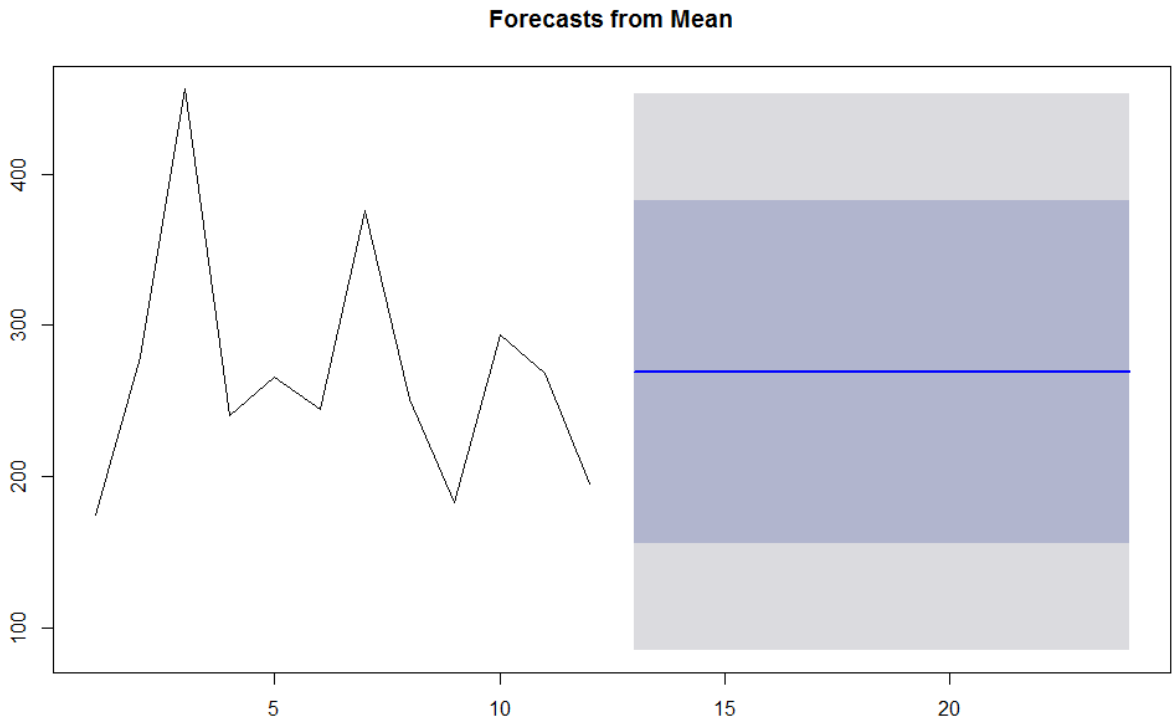
The above graph shows time series plot for the year 2010, Now, we need to project for the year 2011.

Here , it is important for us relook at one of the assumptions that is given in the question itself i.e. the same trend continues. The question arises as to what does trend mean. The first guess would be that it means the trend component of time series. However, we see that the time series plot above has no trend, therefore, this assumption would be absurd.

Another assumption for the meaning of trend as used in the question could be that the entire month wise pattern remains exactly the same in case the company takes no measures. In such a case the same graph would predict for 2011 as well without change .

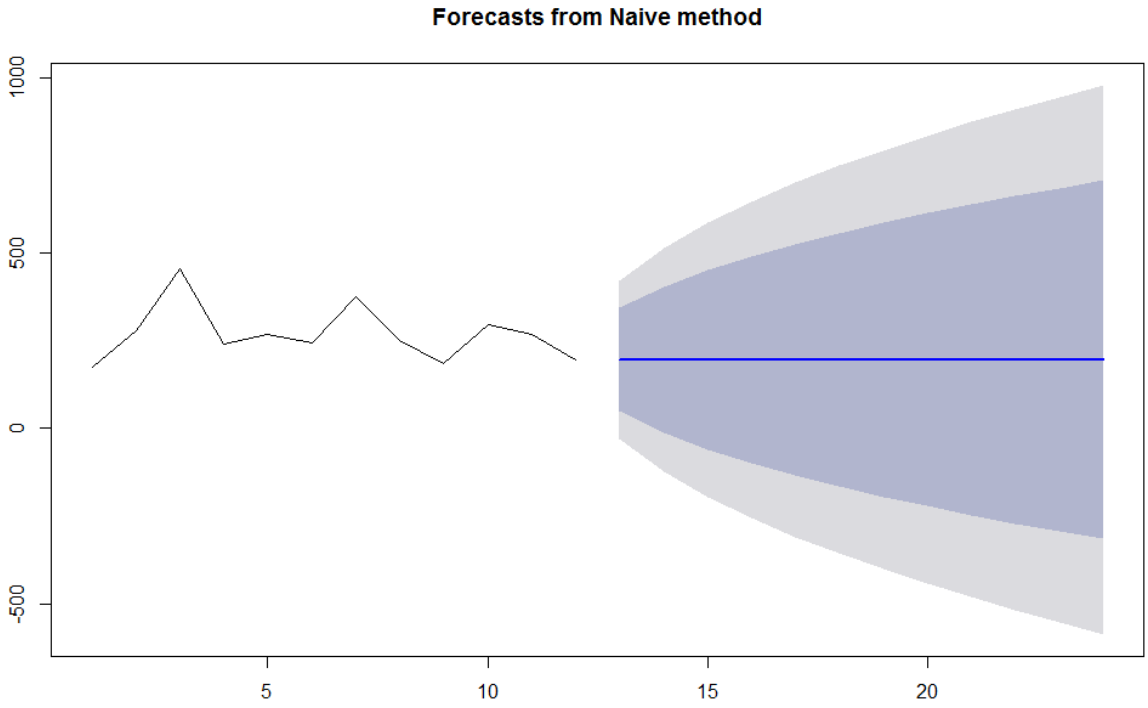
Average Method

Here, the forecasts of all future values are equal to mean of historical data.



Naïve Method

All forecasts are simply set to be the value of last observation.



Check for Stationarity

We use Augmented Dickey Fuller Test for this purpose.

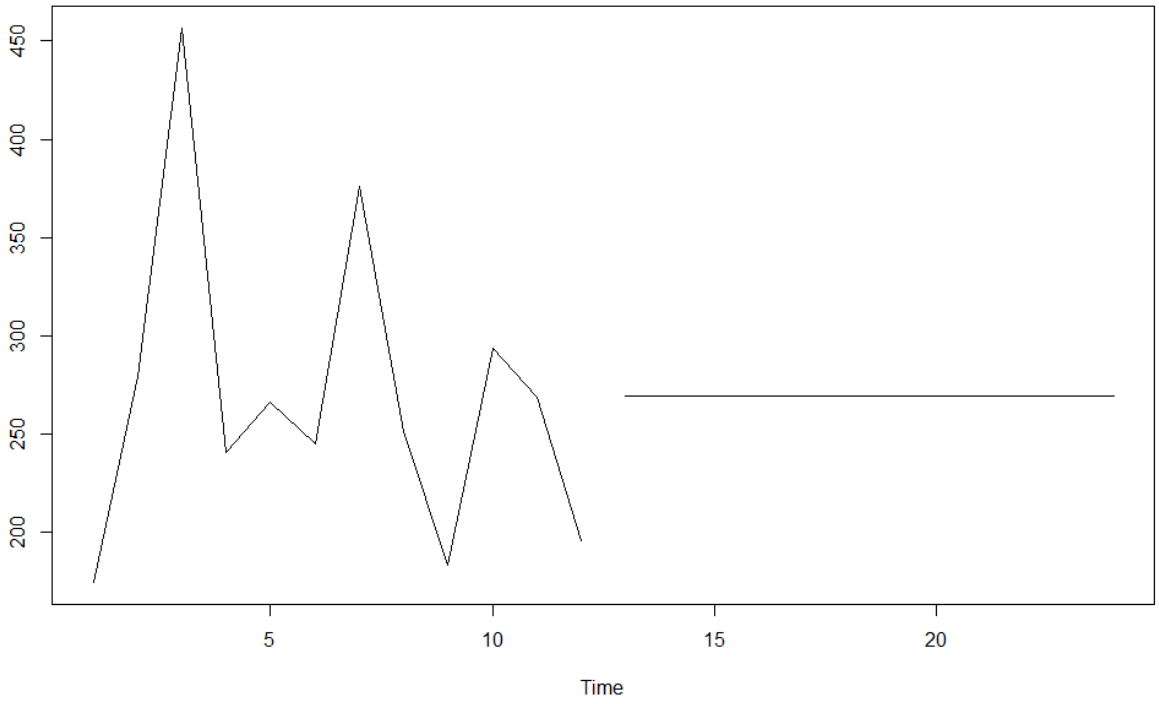
```
library(tseries)
adf.test(x,alternative = "stationary")
```

p<0.05, therefor, time series is stationary.

Now, we go for Arima modeling

```
library(forecast)
fit2<-auto.arima(x)
pred2<-predict(fit2,n.ahead = 12)
```

ts.plot(x,pred2$pred)



Simple Explonential Method

library(forecast)

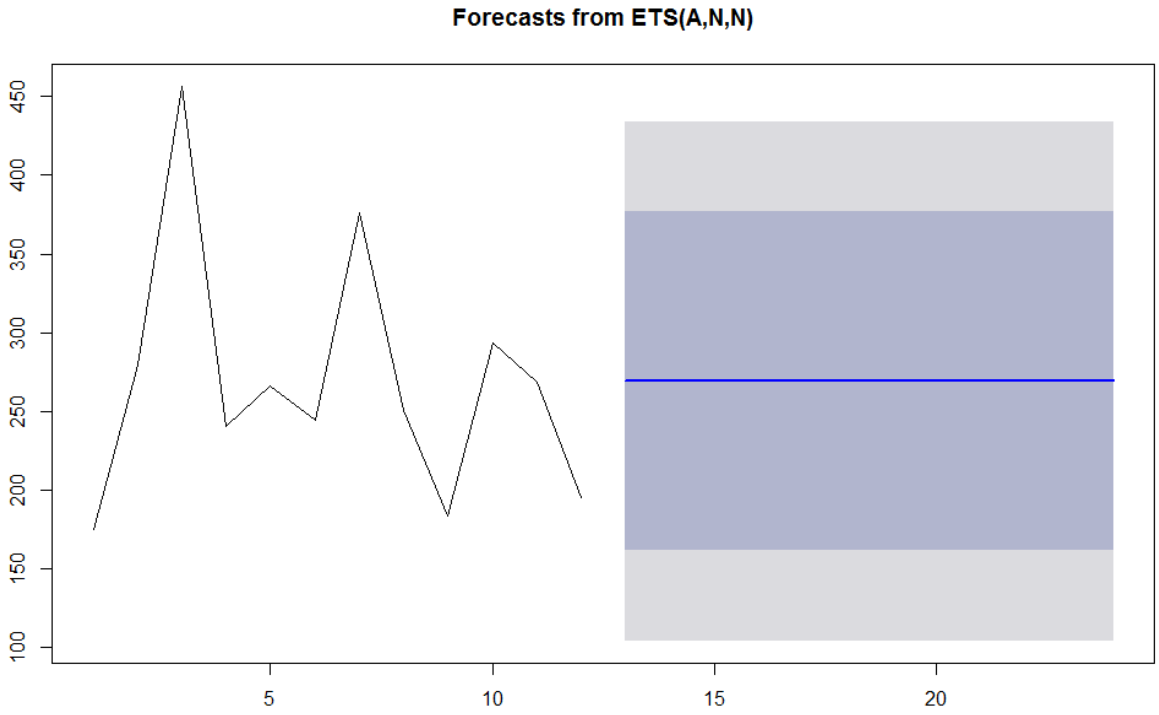
fit3<-ets(x,model = "ANN")

fit3

forecast(fit3,12)

plot(forecast(fit3,12))

accuracy(fit3)



We are getting a flat forecast with similar results(except the naïve method which has lower accuracy compared to others) in all the methods. The flat forecast shows that there is no trend, no seasonality, and insufficient temporal dynamics to allow the future observations to have different conditional means.

4.Conclusion

Through the use of various statistical techniques and modelling we conclude that the company should take the following measures :

- The first priority of the company should be to look at workers suffering from these diseases and provide incentives for preventive care of these diseases.
- Provide reimbursement for transportation allowance
- Increase the age of workforce. Replace younger ones by older ones. Hire experienced workers.
- As far as number of children are concerned, it's a personal choice and perhaps the company cannot do anything about it. Moreover, it can lead to ethical and legal issues.

For the second question we build different models for getting the losses. These have given similar results as depicted in the graphs.