PROJECT ON EMPLOYEE ABSENTEEISM

-BY ABHISHEK VIGG

TABLE OF CONTENTS:

Project Description and Problem statement - XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism.

The company has shared it dataset and requested to have an answer on the following areas: 1. What changes company should bring to reduce the number of absenteeism? 2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

Build suitable model (both R and Python) to answer the above two questions with a proper report.

Dataset Details: Dataset Characteristics: Timeseries Multivariant

Number of Attributes: 21

Missing Values: Yes

Attribute Information: 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)

1)Introduction:

1.1) **Problem Statement**:

We are given a dataset for XYZ company that is facing a problem due to employee absenteeism. Our goal is to understand the data, model a machine learning algorithm around the given data and predict future loss to the company due to employee absenteeism.

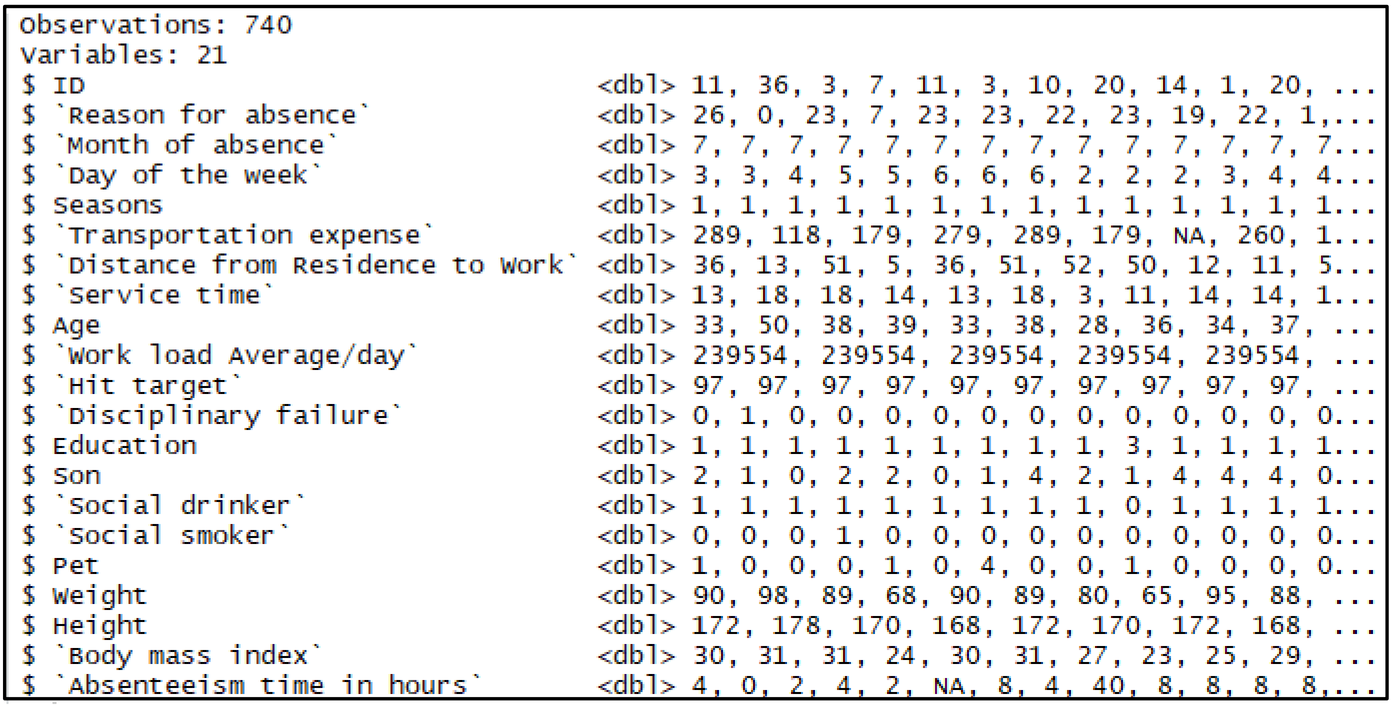
We import the dataset into RStudio to perform the exploratory data analysis followed by the model development in R. Midway in the project after exploratory data analysis, there will be insightful recommendations made to improve reduce the problem of employee absenteeism.

1.2) **Understanding Variables Row Values**:

We use the Tidyverse package in RStudio to perform the preliminary data analysis and ggplot2 to perform exploratory data analysis in the further stages.

After importing, we have a glance at our dataset to know the types of variables and a few values:

Going for a window type view into the data:



```
Observations: 740
Variables: 21
$ ID                                <dbl> 11, 36, 3, 7, 11, 3, 10, 20, 14, 1, 20, ...
$ `Reason for absence`              <dbl> 26, 0, 23, 7, 23, 23, 22, 23, 19, 22, 1,...
$ `Month of absence`                <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7...
$ `Day of the week`                 <dbl> 3, 3, 4, 5, 5, 6, 6, 6, 2, 2, 2, 3, 4, 4...
$ Seasons                           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ `Transportation expense`          <dbl> 289, 118, 179, 279, 289, 179, NA, 260, 1...
$ `Distance from Residence to Work` <dbl> 36, 13, 51, 5, 36, 51, 52, 50, 12, 11, 5...
$ `Service time`                    <dbl> 13, 18, 18, 14, 13, 18, 3, 11, 14, 14, 1...
$ Age                               <dbl> 33, 50, 38, 39, 33, 38, 28, 36, 34, 37, ...
$ `Work load Average/day`           <dbl> 239554, 239554, 239554, 239554, 239554, ...
$ `Hit target`                      <dbl> 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, ...
$ `Disciplinary failure`            <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ Education                         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1...
$ Son                               <dbl> 2, 1, 0, 2, 2, 0, 1, 4, 2, 1, 4, 4, 4, 0...
$ `Social drinker`                  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1...
$ `Social smoker`                   <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ Pet                               <dbl> 1, 0, 0, 0, 1, 0, 4, 0, 0, 1, 0, 0, 0, 0...
$ Weight                            <dbl> 90, 98, 89, 68, 90, 89, 80, 65, 95, 88, ...
$ Height                            <dbl> 172, 178, 170, 168, 172, 170, 172, 168, ...
$ `Body mass index`                 <dbl> 30, 31, 31, 24, 30, 31, 27, 23, 25, 29, ...
$ `Absenteeism time in hours`       <dbl> 4, 0, 2, 4, 2, NA, 8, 4, 40, 8, 8, 8, 8,...
```

The following view shows the glimpse of the data. As we can see, all the variables are coded as a double type of variable. They need to be recoded for us to perform data analysis successfully. We cannot work on double type of variables. We first change all the variables to a numeric type and then understand the data to change the categorical variables to factors and leave numeric data as numeric. We use the sapply function to recode all the variables at once.

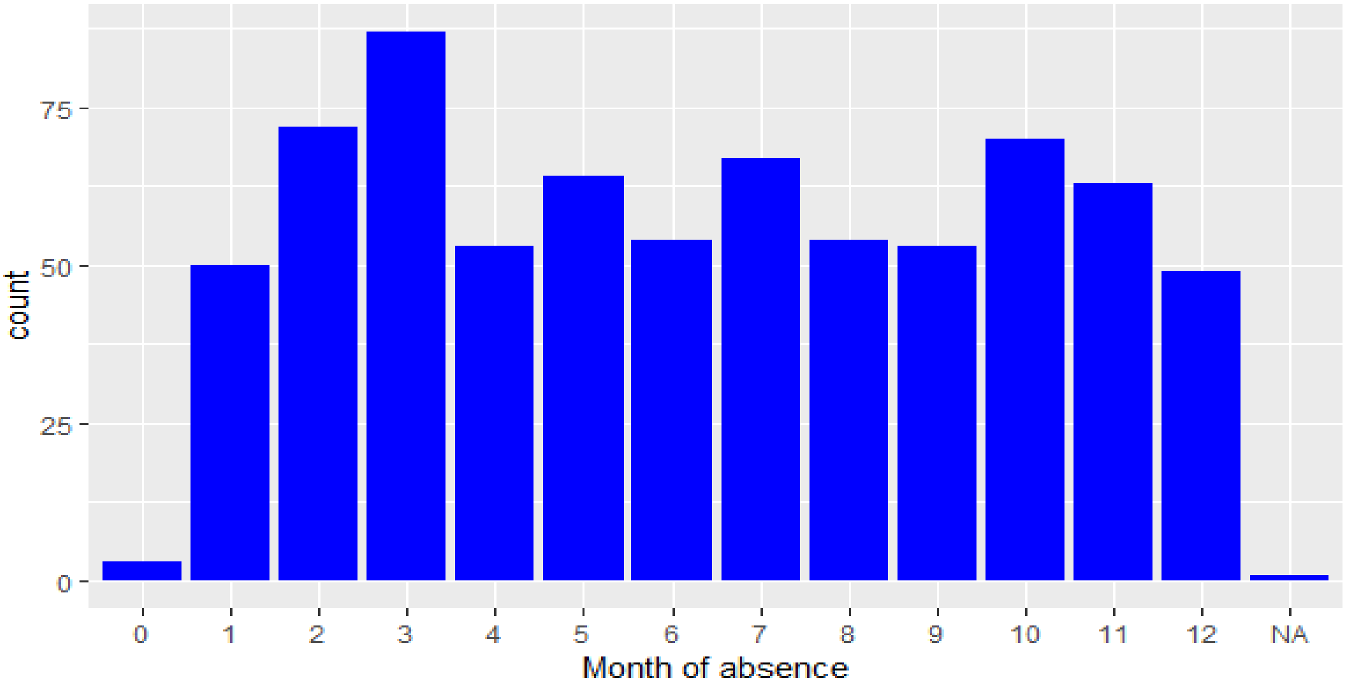
After applying the sapply function we have a new dataframe called df_data that has all the variables recoded as numeric variables.

2)Tidying the data for model development:

2.1) **Missing Value Analysis**:

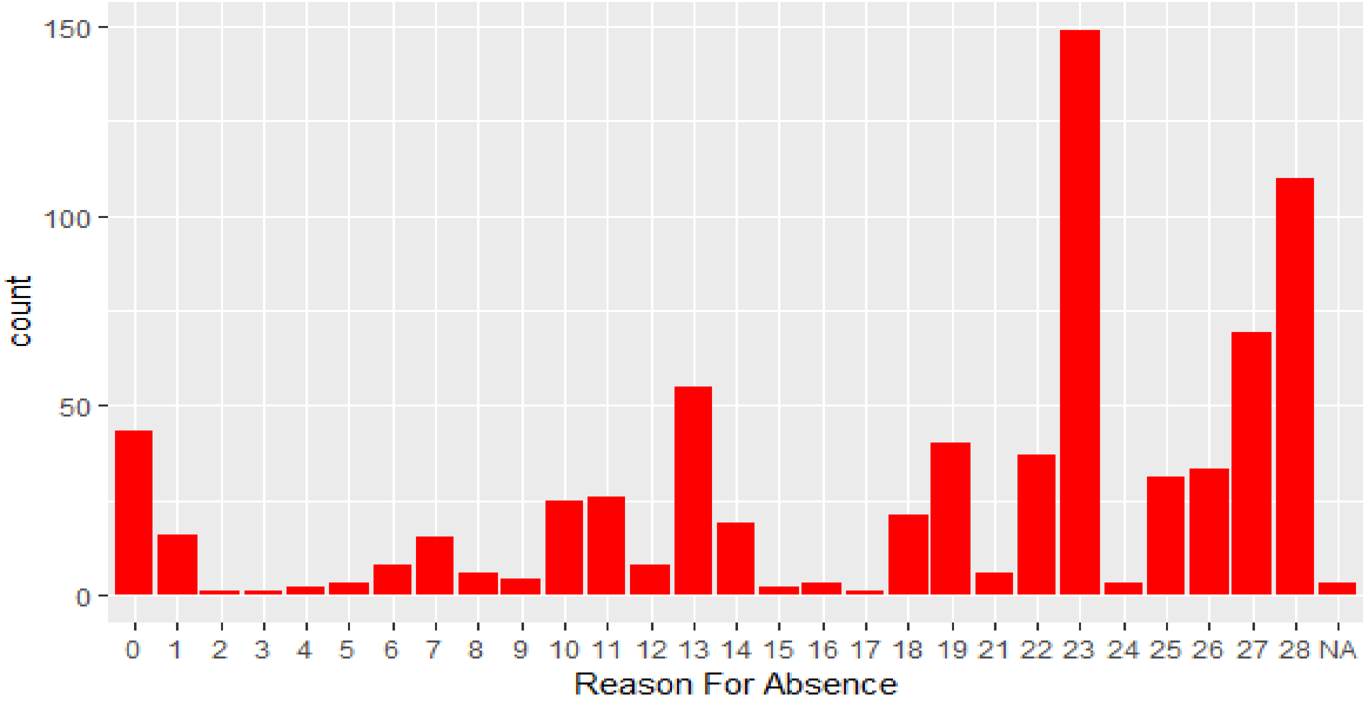
When we view the data in length, we find that NA is not the only missing value in the dataset. When we have such multiple types of missing values in the data, it is a good practice to convert them to NA as anything other than NA as a missing value does consume the RAM of the kernel.

Let's plot a few visualizations of variables to know how many of the variables have multiple missing values.



As we can see in the histogram, the normal observations without missing values take Month of Absence value between 1 to 12 denoting months January to December respectively. The observations with value 0 do not represent any month and should be replaced by NA. We use the following command in R to replace the 0s with NA.

Here we find the indexes of the values in the Month Of Absence with the value 0 and then pass it in the replace function to replace them with NA.
Plotting visualizations for Reason Of Absence:



As we can see the number 0 represents no reason for absence in the project description. We have to replace it with NA. Find index of 0s and replace the NA.

Using the summary function to calculate the number of NA values for each variable, we come to know that the following is the number of NA values in each variable:

Variable	Number Of Missing Values
Reason for Absence	46
Transportation Expense	7
Distance From Residence To Work	3
Age	3
Month Of Absence	4
Service Time	3
Work Load Average	10
Hit Target	6
Disciplinary Failure	6
Education	10
Son	6
Social Drinker	3
Social Smoker	4
Pet	2
Weight	1
Height	14
Body Mass Index	31
Absenteeism Time In Hours	22
Total	181

None of the columns have more than 20% missing data. Thus, we perform imputation.

Now that we have treated the data to have only one type of missing value(NA). We impute the missing value NA with an imputation method that suits the data. We have three types of Missing value treatments:

1) Mean
2) Median
3) K-NN Imputation

We experiment with each method to find the method that best suits our data.

Starting with mean, we first turn an actual observation that is not NA into NA and record the actual value.

>df_data[10,2]

The actual value is 22

Now we perform the mean method of imputation.

When we check the value of df_data[10,2] it is 20.38.Reloading the data and turning the value at df_data[10,2] to NA will give us the original dataset. Now we perform the median method of imputation.

Checking value at df_data[10,2] gives 23. Reloading the data and turning df_data[10,2] to NA. Applying the knn imputation

The following are values recorded:

1) Mean = 20.38 2) Median = 23 3) Knn =16.83

We can conclude that Median Imputation is the best method to fill NA values of our dataset.

Now that we have a dataset with full data we just have to round the data to integers as Median Imputation leaves a lot of decimal places.

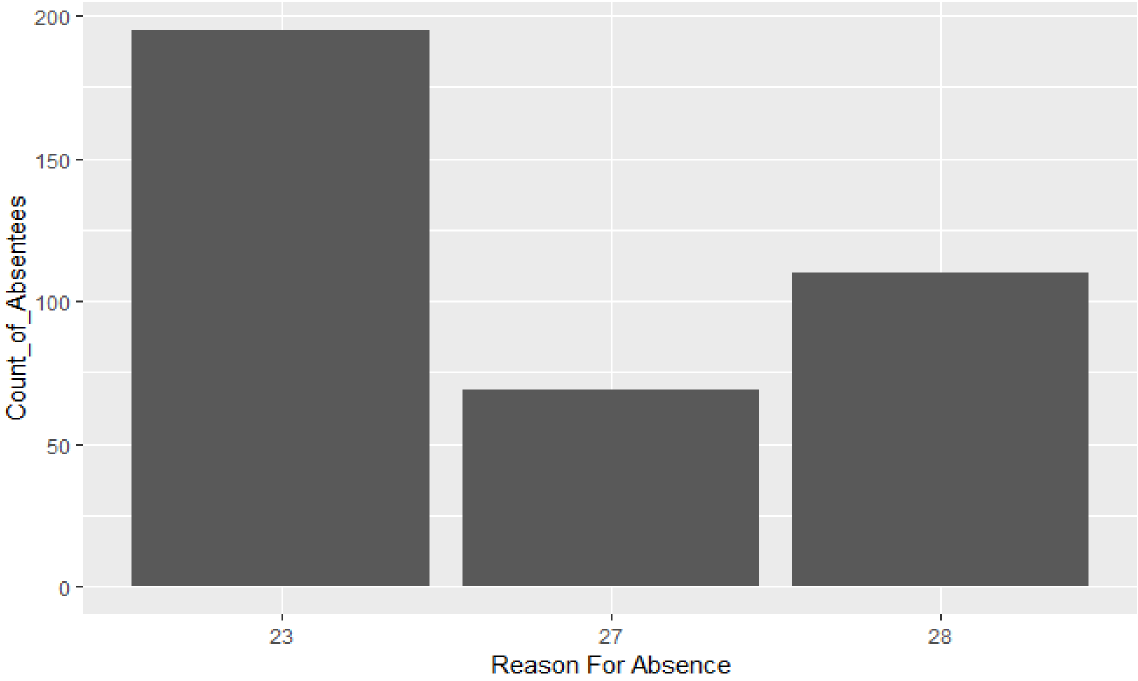
Now that we have a clean dataset without any missing values, we will perform exploratory data analysis to draw inferences from the data given to us. This is performed with an agenda of identifying patterns in the data.

3)Exploratory Data Analysis & Recommendations:

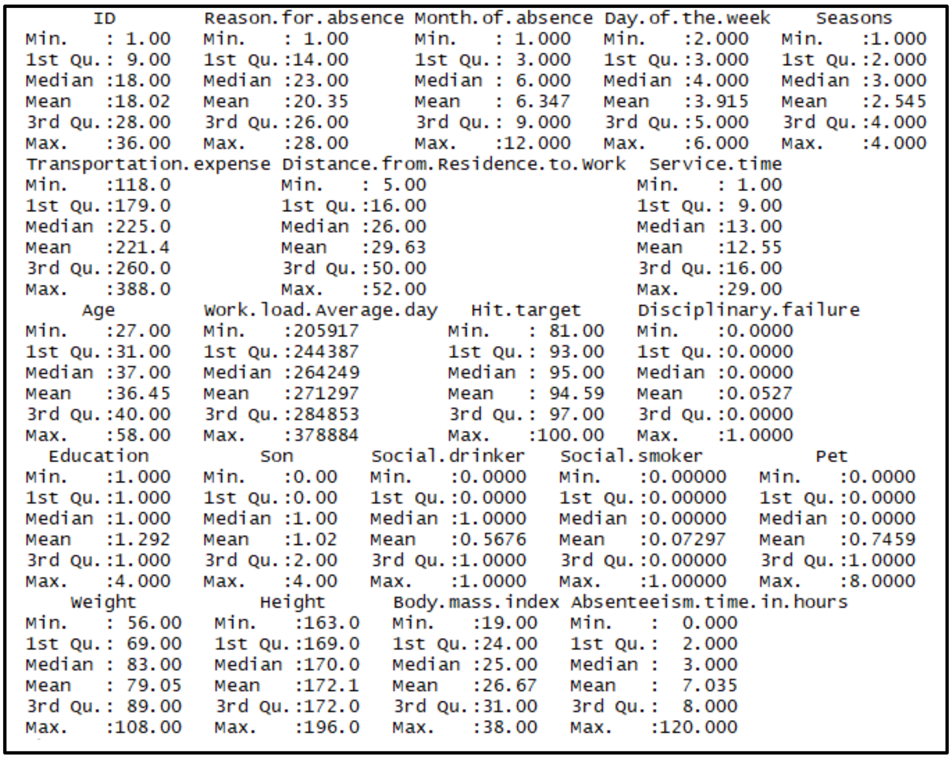
3.1)Exploratory Data Analysis:

Now we try to find the top five reasons for absence from our dataset. This will give us a clear idea of which disease or ailment is causing the absence.

Tabulating a data frame that list the count of absentees grouped by each reason of absence and visualizing them on a bar graph with x axis as the reason for absence and y axis as the count of absentees:

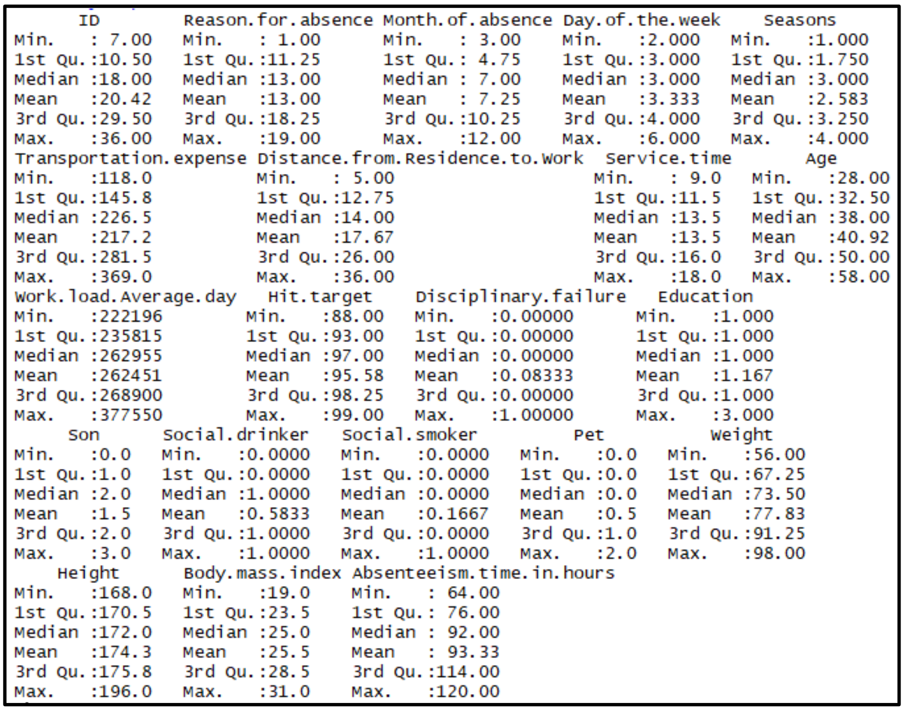


The above graph shows the top five reasons for people being absent from work.
When we take a look at the project description we can trace out what these numbers denote.
23-> Blood Donation
27-> Physiotherapy
28-> Dental Consultation
As we can see, the biggest reason for absence of employees is Blood Donation. This is not a sickness or an ailment that is preventing a person from being absent from work. The XYZ company can initiate blood donation drives at its own company to reduce this number drastically. The second biggest reason for absenteeism is Dental Consultation.
The XYZ company can hire an inhouse dentist to give the employees dental consultation at the workplace during off work hours. This will drastically reduce the number of people being absent.
Now we check the data to understand what are the parameters of employees who have taken the longest hours off as abse nt from work.
Let's understand the summary statistics of our entire data to make inferences later.



```
       ID        Reason.for.absence Month.of.absence Day.of.the.week    Seasons     
 Min.   : 1.00   Min.   : 1.00      Min.   : 1.000   Min.   :2.000   Min.   :1.000  
 1st Qu.: 9.00   1st Qu.:14.00      1st Qu.: 3.000   1st Qu.:3.000   1st Qu.:2.000  
 Median :18.00   Median :23.00      Median : 6.000   Median :4.000   Median :3.000  
 Mean   :18.02   Mean   :20.35      Mean   : 6.347   Mean   :3.915   Mean   :2.545  
 3rd Qu.:28.00   3rd Qu.:26.00      3rd Qu.: 9.000   3rd Qu.:5.000   3rd Qu.:4.000  
 Max.   :36.00   Max.   :28.00      Max.   :12.000   Max.   :6.000   Max.   :4.000  
 Transportation.expense Distance.from.Residence.to.Work  Service.time  
 Min.   :118.0          Min.   : 5.00                   Min.   : 1.00  
 1st Qu.:179.0          1st Qu.:16.00                   1st Qu.: 9.00  
 Median :225.0          Median :26.00                   Median :13.00  
 Mean   :221.4          Mean   :29.63                   Mean   :12.55  
 3rd Qu.:260.0          3rd Qu.:50.00                   3rd Qu.:16.00  
 Max.   :388.0          Max.   :52.00                   Max.   :29.00  
      Age        Work.load.Average.day   Hit.target     Disciplinary.failure
 Min.   :27.00   Min.   :205917        Min.   : 81.00   Min.   :0.0000      
 1st Qu.:31.00   1st Qu.:244387        1st Qu.: 93.00   1st Qu.:0.0000      
 Median :37.00   Median :264249        Median : 95.00   Median :0.0000      
 Mean   :36.45   Mean   :271297        Mean   : 94.59   Mean   :0.0527      
 3rd Qu.:40.00   3rd Qu.:284853        3rd Qu.: 97.00   3rd Qu.:0.0000      
 Max.   :58.00   Max.   :378884        Max.   :100.00   Max.   :1.0000      
   Education          Son       Social.drinker   Social.smoker          Pet        
 Min.   :1.000   Min.   :0.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
 1st Qu.:1.000   1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
 Median :1.000   Median :1.00   Median :1.0000   Median :0.00000   Median :0.0000  
 Mean   :1.292   Mean   :1.02   Mean   :0.5676   Mean   :0.07297   Mean   :0.7459  
 3rd Qu.:1.000   3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000  
 Max.   :4.000   Max.   :4.00   Max.   :1.0000   Max.   :1.00000   Max.   :8.0000  
     Weight           Height     Body.mass.index Absenteeism.time.in.hours
 Min.   : 56.00   Min.   :163.0   Min.   :19.00   Min.   :  0.000          
 1st Qu.: 69.00   1st Qu.:169.0   1st Qu.:24.00   1st Qu.:  2.000          
 Median : 83.00   Median :170.0   Median :25.00   Median :  3.000          
 Mean   : 79.05   Mean   :172.1   Mean   :26.67   Mean   :  7.035          
 3rd Qu.: 89.00   3rd Qu.:172.0   3rd Qu.:31.00   3rd Qu.:  8.000          
 Max.   :108.00   Max.   :196.0   Max.   :38.00   Max.   :120.000          
```

Let us find the summary statistics of top 10 absentees in our data.



```
       ID        Reason.for.absence Month.of.absence Day.of.the.week    Seasons     
 Min.   : 7.00   Min.   : 1.00      Min.   : 3.00    Min.   :2.000   Min.   :1.000  
 1st Qu.:10.50   1st Qu.:11.25      1st Qu.: 4.75    1st Qu.:3.000   1st Qu.:1.750  
 Median :18.00   Median :13.00      Median : 7.00    Median :3.000   Median :3.000  
 Mean   :20.42   Mean   :13.00      Mean   : 7.25    Mean   :3.333   Mean   :2.583  
 3rd Qu.:29.50   3rd Qu.:18.25      3rd Qu.:10.25    3rd Qu.:4.000   3rd Qu.:3.250  
 Max.   :36.00   Max.   :19.00      Max.   :12.00    Max.   :6.000   Max.   :4.000  
 Transportation.expense Distance.from.Residence.to.Work  Service.time       Age       
 Min.   :118.0          Min.   : 5.00                   Min.   : 9.0   Min.   :28.00  
 1st Qu.:145.8          1st Qu.:12.75                   1st Qu.:11.5   1st Qu.:32.50  
 Median :226.5          Median :14.00                   Median :13.5   Median :38.00  
 Mean   :217.2          Mean   :17.67                   Mean   :13.5   Mean   :40.92  
 3rd Qu.:281.5          3rd Qu.:26.00                   3rd Qu.:16.0   3rd Qu.:50.00  
 Max.   :369.0          Max.   :36.00                   Max.   :18.0   Max.   :58.00  
 Work.load.Average.day   Hit.target     Disciplinary.failure   Education    
 Min.   :222196        Min.   :88.00   Min.   :0.00000       Min.   :1.000  
 1st Qu.:235815        1st Qu.:93.00   1st Qu.:0.00000       1st Qu.:1.000  
 Median :262955        Median :97.00   Median :0.00000       Median :1.000  
 Mean   :262451        Mean   :95.58   Mean   :0.08333       Mean   :1.167  
 3rd Qu.:268900        3rd Qu.:98.25   3rd Qu.:0.00000       3rd Qu.:1.000  
 Max.   :377550        Max.   :99.00   Max.   :1.00000       Max.   :3.000  
      Son       Social.drinker    Social.smoker          Pet            Weight     
 Min.   :0.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0   Min.   :56.00  
 1st Qu.:1.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0   1st Qu.:67.25  
 Median :2.0   Median :1.0000   Median :0.0000   Median :0.0   Median :73.50  
 Mean   :1.5   Mean   :0.5833   Mean   :0.1667   Mean   :0.5   Mean   :77.83  
 3rd Qu.:2.0   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0   3rd Qu.:91.25  
 Max.   :3.0   Max.   :1.0000   Max.   :1.0000   Max.   :2.0   Max.   :98.00  
     Height      Body.mass.index Absenteeism.time.in.hours
 Min.   :168.0   Min.   :19.0    Min.   : 64.00           
 1st Qu.:170.5   1st Qu.:23.5    1st Qu.: 76.00           
 Median :172.0   Median :25.0    Median : 92.00           
 Mean   :174.3   Mean   :25.5    Mean   : 93.33           
 3rd Qu.:175.8   3rd Qu.:28.5    3rd Qu.:114.00           
 Max.   :196.0   Max.   :31.0    Max.   :120.00           
```

When we check the two summary statistics we come to know of very interesting insights. The top ten absentees had a median distance from residence to work of 14 whereas the median distance from residence to work of the entire dataset is 26. Maybe the employees who have their homes very near to the workplace go to their home for lunch or to take a nap during working hours. This could be a reason for the excessive hours of absenteeism.

Moreover, the top ten absentees are absent more often on a particular day of the week. That is 3. I think this denotes Wednesday. There is no conclusion that can be given from this insight.

There is another interesting finding here. The median weight of the top 10 employee absentees is 10 Kgs less than the median weight of the entire sample. Maybe the employees who are absent excessively are malnourished.

<u>3.2)Recommendations</u>:

The following recommendations provide key insights into the data. The recommendations can be listed as follows:

1)Blood Donation Drives and Dental Consultations should be arranged in the company to reduce the number of Absentees.

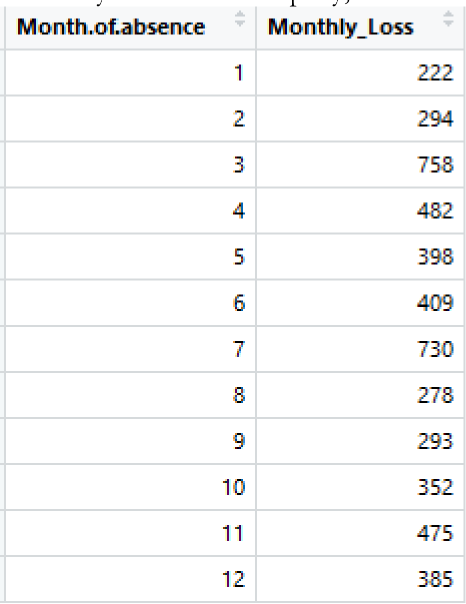
2)If the distance of residence of employees is very less from the workplace, they should not be going home to have a nap or eat lunch. They can have lunch with their respective teams.

3) If the employees absent for very long hours are malnourished they should be provided medical care to improve their health under some medical health plan.

Note: These recommendations will be reiterated with the conclusion of the model.

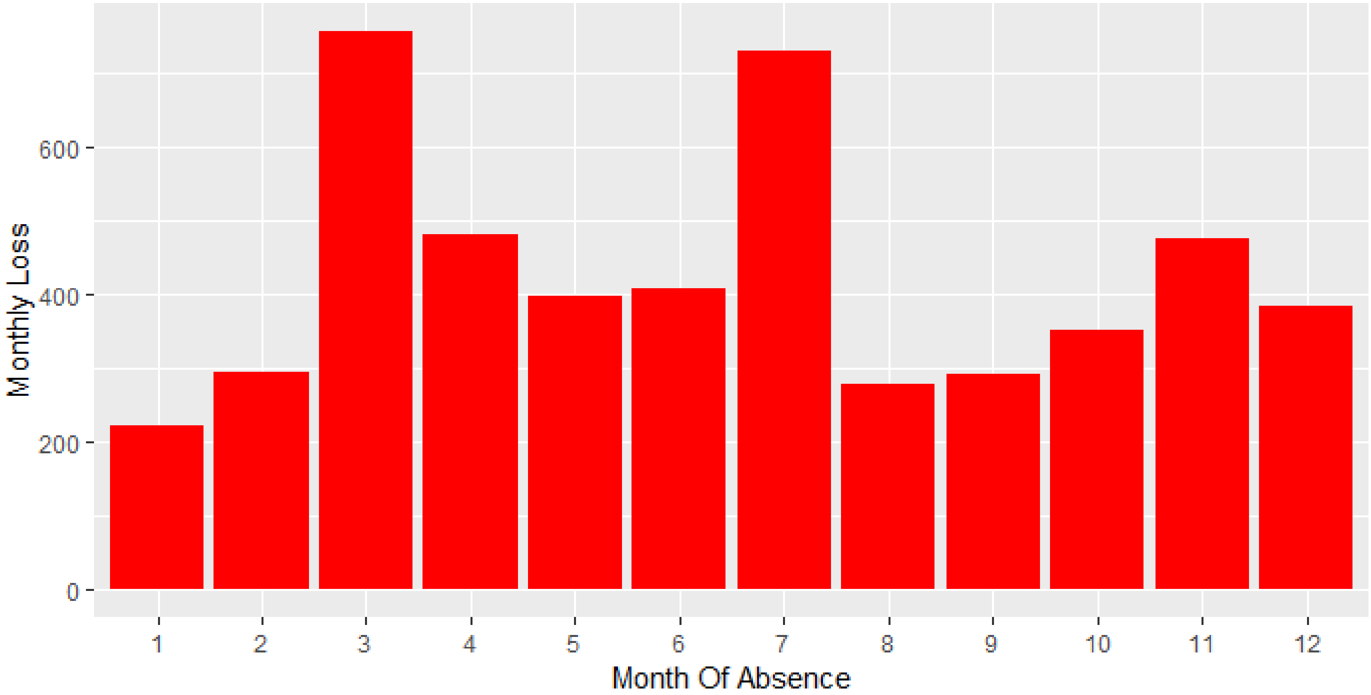
3.3) Monthly Loss To Company In Hours:

As the project requires us to calculate the monthly loss to the company, we have to calculate the total losses every month.



Month.of.absence	Monthly_Loss
1	222
2	294
3	758
4	482
5	398
6	409
7	730
8	278
9	293
10	352
11	475
12	385

Visualizing the monthly losses on a bar plot:

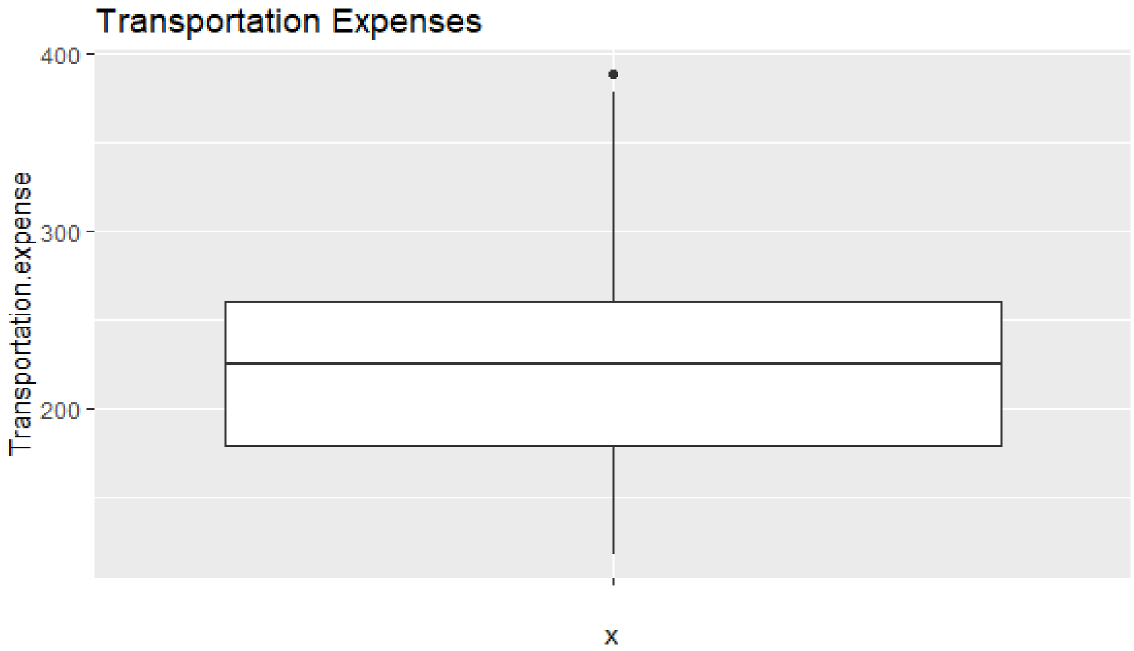


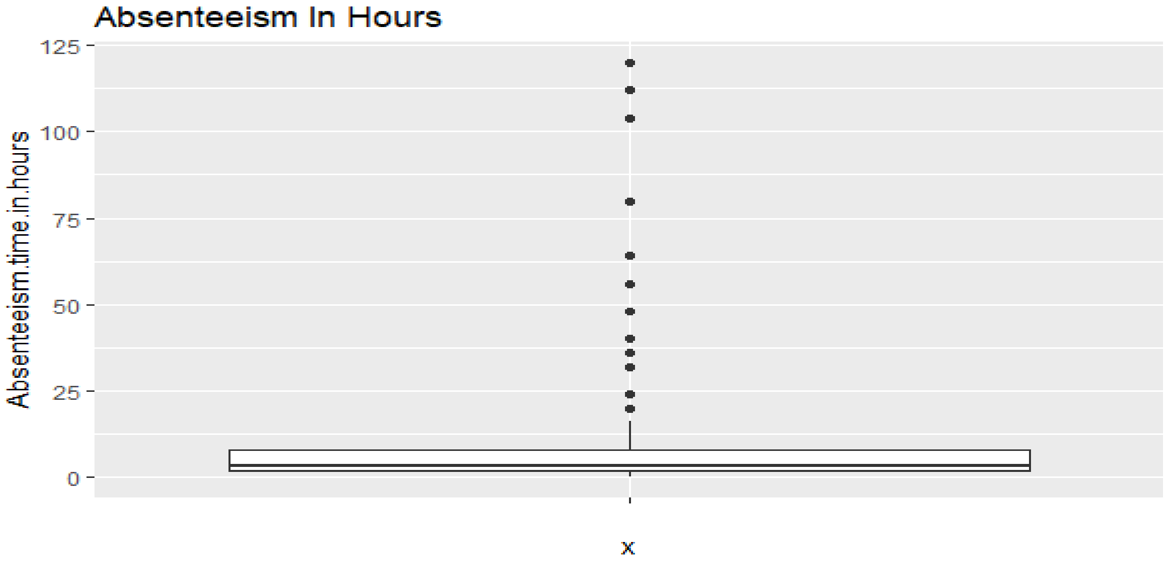
The losses seem to be highest in the month of March and July.

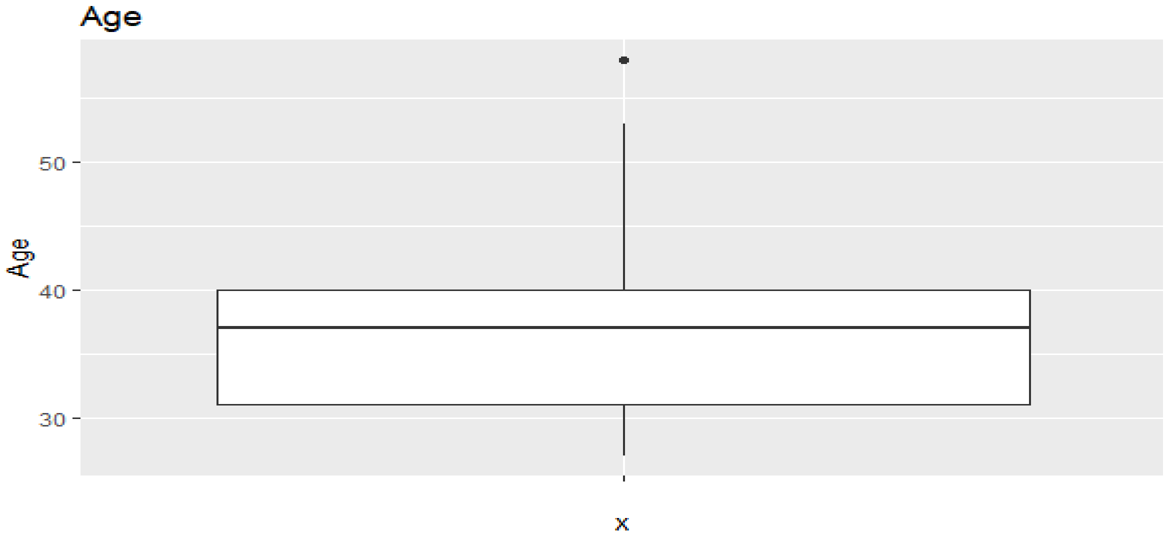
4)Model Development:

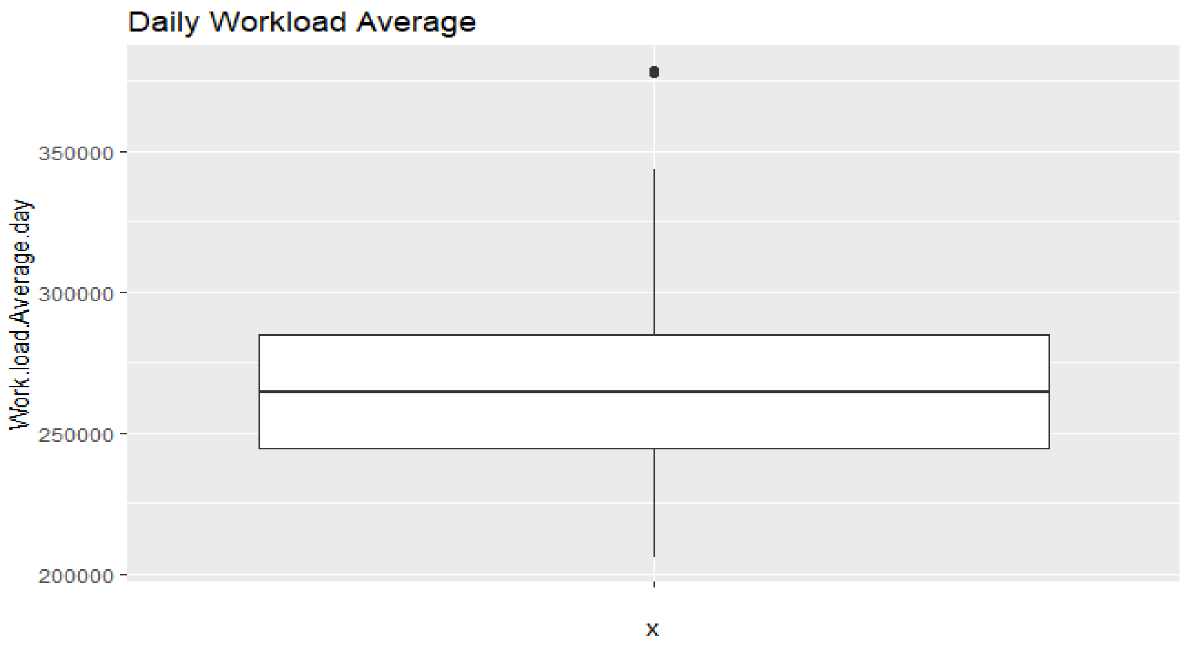
4.1) Outlier Analysis:

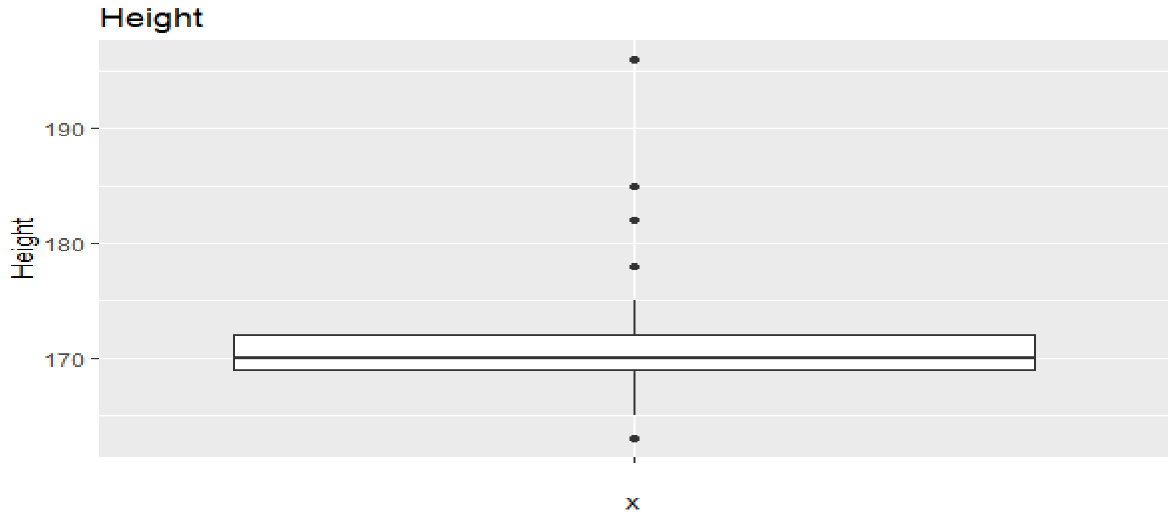
Finding the numeric data from the entire data to do outlier analysis.
We usually plot boxplots for numeric data to understand the number of outliers we have per variable.









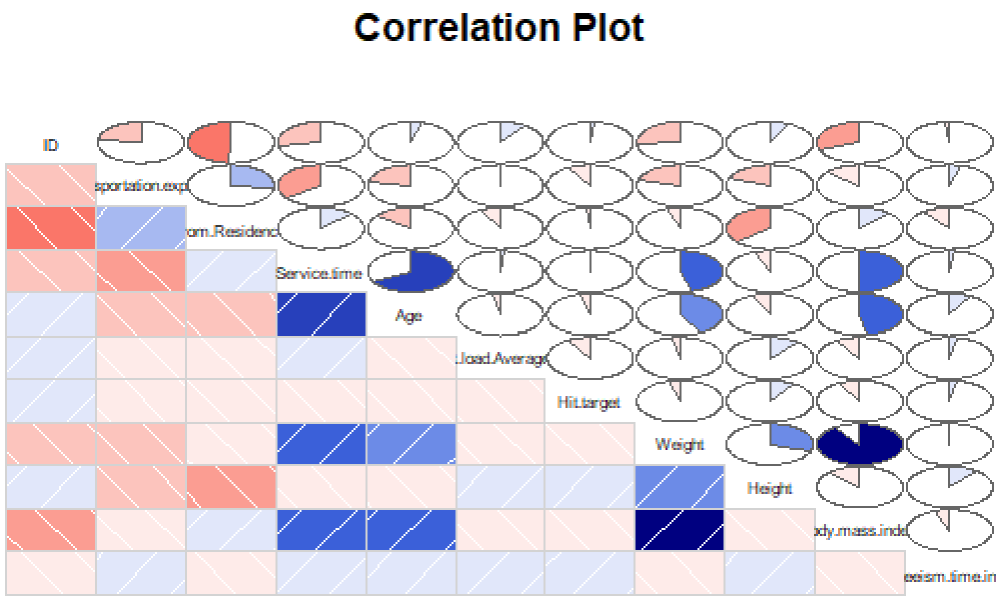


As we can see, there are a couple of variables that have outliers. These outliers should be eliminated.

<u>4.2) Feature Selection</u>:

Feature selection is a process of selecting the variables that are important for model development. We eliminate the variables that have high correlation between each other as they carry similar information. We do this separately for numeric variables and categorical variables.

We use correlation plots to check the correlation between numeric variables using the corrgram library:



As we can see the Weight and Body.mass.index variable has high correlation. Since Body.mass.index seems to explain the dependent variable better we will eliminate the independent variable Weight from our data. Performing Chi Square Tests to do feature selection on categorical variables.

Recoding the variables to factors before performing the Chi Square Tests and performing the chi square test on categorical variables, we get the following observations:

Variable Name	X-Squared	Degree Of Freedom	P-value
Reason.for.absence	1589.2	675	P < 2.2e-16
Month.of.absence	392.21	275	4.248e-06
Day.of.the.week	128.82	100	0.02769
Seasons	144.99	75	2.273e-06
Disciplinary.failure	580.75	25	P<2.2e-16
Education	44.149	75	0.9983
Son	188.51	100	2.148e-07
Social.smoker	32.442	25	0.1456
Pet	141.46	125	0.1491

From the above table we can check the p values. The p-values greater than 0.05 will be removed as these values statistically are unimportant for the prediction of the target variable. The following variables have a p value greater than 0.05: Pet,Social.smoker and Education.

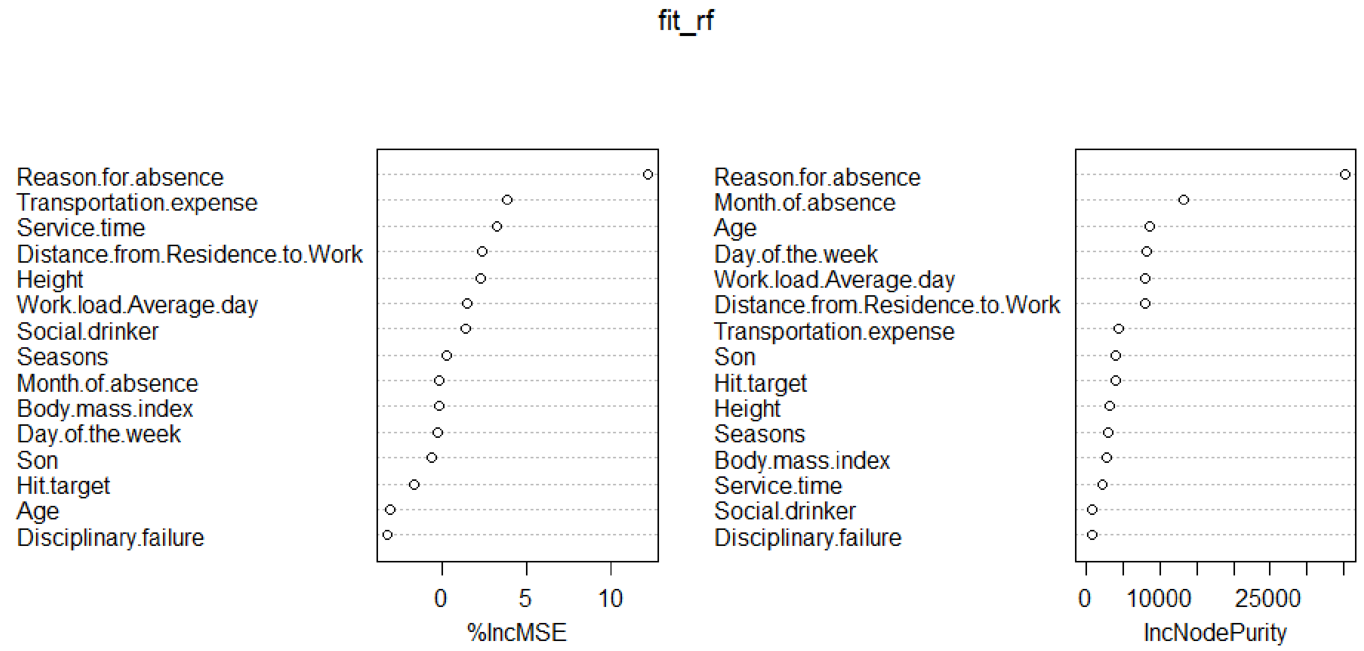
Now removing the variables that were to be eliminated in the correlation analysis and chi square tests along with the variable ID.

4.3)Feature importance:

We find the feature importance to understand which variables are more important for the prediction of the target variable. The importance of features can be estimated from data by building a model. Some methods like decision trees have a built-in mechanism to report on variable importance. But for this data we use the Random Forest Model along with the varImp function of Caret package to understand the variable importance.

Random Forests are based on decision trees and use bagging to come up with a model over the data. Random forests also have a feature importance methodology which uses 'gini index' to assign a score and rank the features.

We build a model called fit_rf and then apply the varImp function to it. This will give us the horizontal bar graph of the importance of each variable. Afterwards, we use the varImpPlot over the fit_rf model to visualize the importance of each variable.



Here we see the %IncMSE and IncNodePurity graphs. IncNodePurity relates to the loss function which by best splits are chosen. The loss function is mse for regression and gini-impurity for classification. More useful variables achieve higher increases in node purities, that is to find a split which has a high inter node 'variance' and a small intra node 'variance'. %IncMSE is a similar feature where greater the value of %IncMSE more is the variable important. As we can see,"Reason For Absence" is the most important variable followed by Transportation Expenses and Service Time.

4.4)Feature Scaling:

Performing feature scaling involves identifying whether the data is normally distributed or is skewed and then normalizing the values of the variables of our data. After seeing the Exploratory Data Analysis, we can say that our data is clearly skewed. So we use the normalization method of feature scaling to make sure the variables whose range is very large do not dominate the machine learning model's predictions.

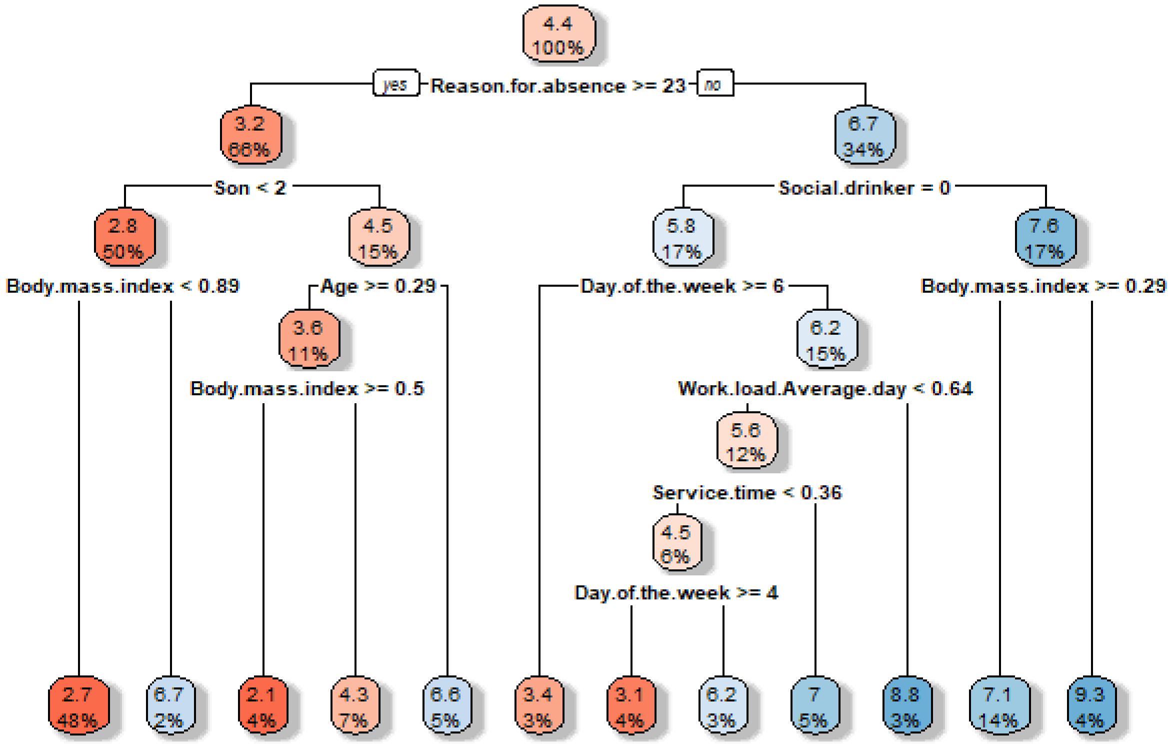
After this we have data that has all the numeric variables with a range between 0 to +1. This way none of the numeric variables dominate model's results.

4.5) Decision Tree Model:

Decision Tree Regression model is a basic model that decides certain business rules as part of the model development. These rules are developed on a train data and then applied on test data. We will derive the train and test data from the principle data itself by using sampling techniques.

The train_data is the data on which the decision tree model is built. Then it is applied on the test_data. The code to build the decision tree regression model using rpart library and rpart.plot is written in the R file.

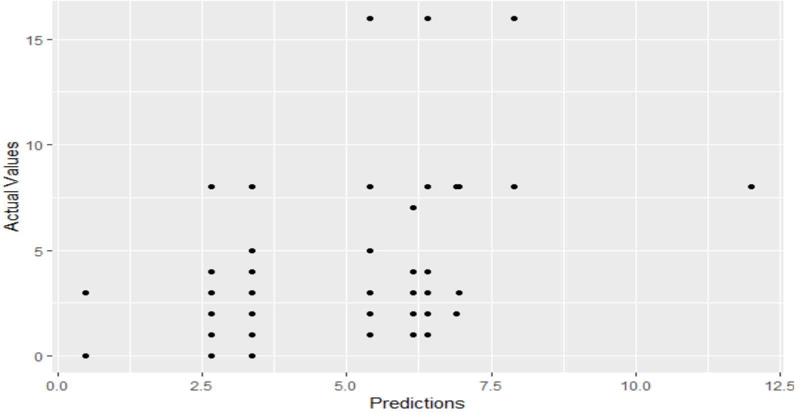
We can understand the business rules by using the rpart.plot library. If there is a need to summarize the model's business rules then we can use the summary function on the model developed(decision_tree_fit)
Here is the visualization of the business rules using rpart.plot:



The above business rules are a pictorial representation of the business rules in summary of the model. We use the parameters RMSE,R^2 and MAE to compare different models as these metrics are good to determine a model's efficiency relative to others.

We predict the model's output on test data and calculate Root Mean Squared Error. Once our model is ready, we input the test data to find the predicted or target values.

The RMSE for Decision tree model comes out to be **2.88**. The Mean Absolute Error of this model is **1.92**.
We try to calculate the correlation coefficient between the predicted test values and actual test values to calculate R^2 for decision tree algorithm because R^2 is basically the square of the Pearson Correlation Coefficient. After calculations, we get R^2 equal to **0.19.** The scatterplot of Predicted vs Actual values is as shown:



This is a decent RMSE,MAE and R^2 but we will have to develop more models to know if there can be improvement on t his.

4.6)Linear Regression Model:

Linear Regression is an algorithm that determines the coefficients of a numerical equation between the dependent variable and the independent variable. It is represented as:

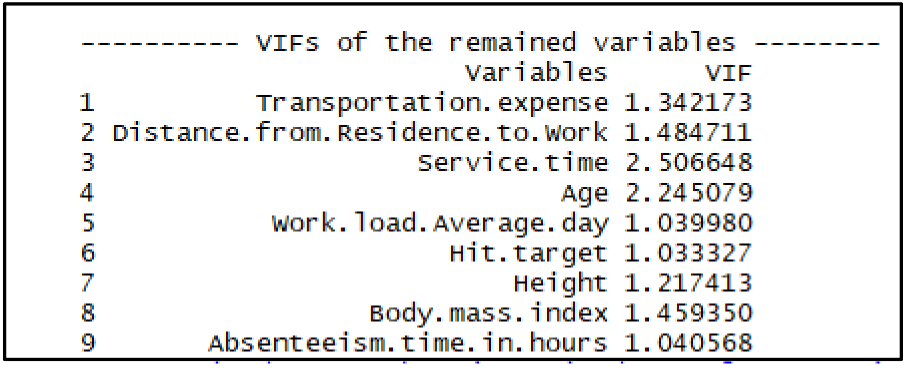
$$Y = A0 + A1*b1 + A2*b2 + A3*b3 + A4*b4 \ldots\ldots$$

Here Y is the dependent variable and A0,A1,A2,etc are the independent variables with b1,b2,b3 as the coefficients.

For developing a linear regression model we have to check if any variable has high collinearity between each other. This is calculated by checking the Variance Inflation Factor. We have the library usdm that helps us calculate the Variance Inflation Factor for numeric variables.

Since we have the numeric data saved in the data frame numeric_data, we use it to find the variance inflation factors.

The following is the results we get.



```
---------- VIFs of the remained variables --------
                        Variables      VIF
1          Transportation.expense 1.342173
2 Distance.from.Residence.to.Work 1.484711
3                    Service.time 2.506648
4                             Age 2.245079
5           Work.load.Average.day 1.039980
6                      Hit.target 1.033327
7                          Height 1.217413
8                 Body.mass.index 1.459350
9       Absenteeism.time.in.hours 1.040568
```

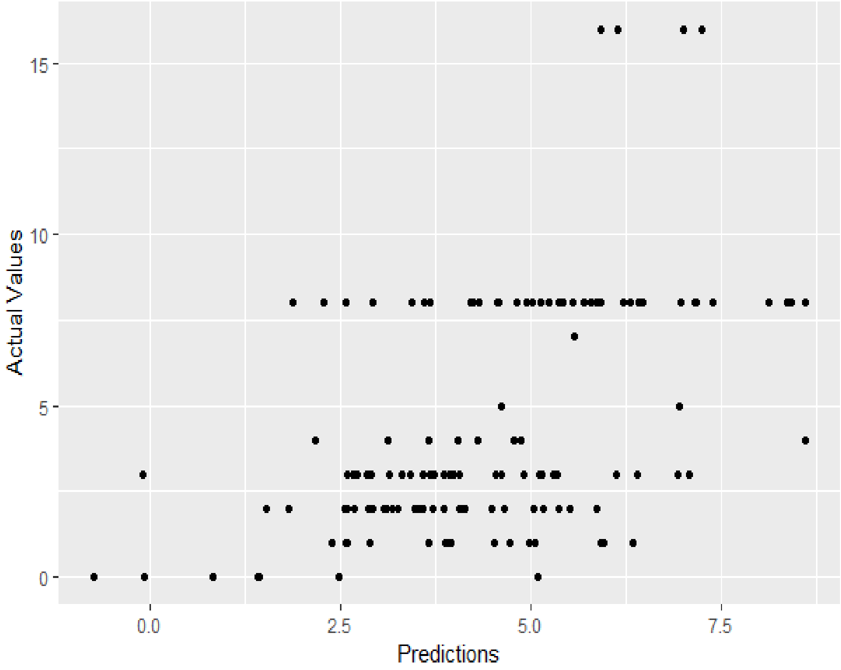
All the numeric variables in our data have a VIF of less than 5. So we can safely say that they are not collinear.

Now we develop the model on train data and then predict values on the independent variables of test data. A summary of the linear regression model gives the following result:

The RMSE comes out to be **2.76**. The Mean Absolute Error of this model is **1.97**.

As we read the summary of the linear regression model, we understand that the R^2 is equal to **0.18.** The coefficients show the intercepts on x axis of all the variables and standard error along with t value.

The value of R^2 represents the correlation coefficient between the original dependent variable values of the test data and the predicted dependent variable values of the test data.
Ideally this value should be as close to 1 as possible. But in our model it is far away from 1. Thus, we still try building another model.
The following is the scatterplot of predicted vs actual values.



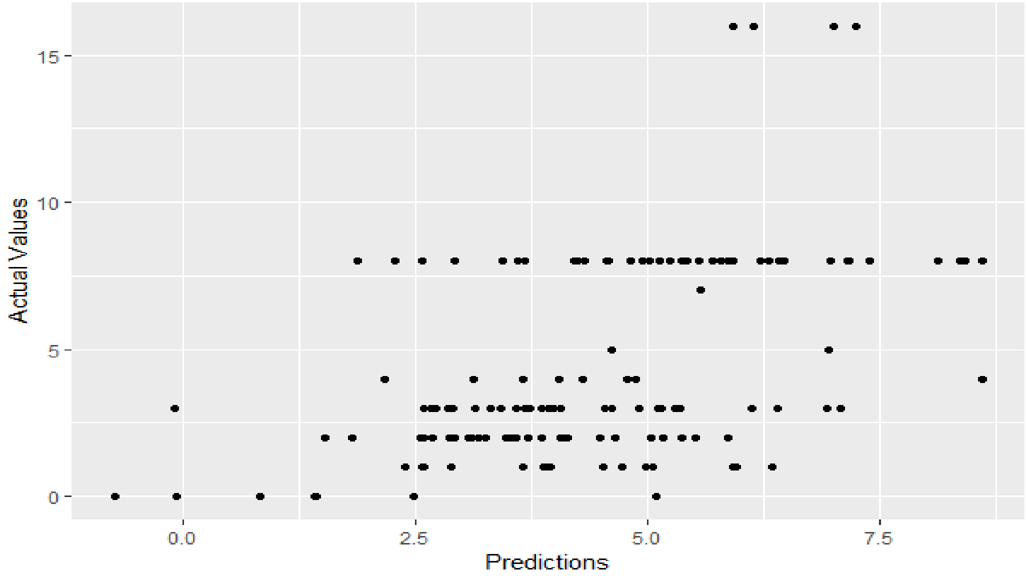
Let’s try out the last model development for our data

4.7)Random Forest Model:

Random Forest Algorithm is an ensemble method of developing multiple decision trees randomly creating a forest of decision trees that do not split at the nodes on the basis of variable importance but split in bags of variables. This is a widely used algorithm for both regression and classification.

We start by developing a random forest model by fitting it to the train data by importing the library randomForest and applying the function for model development. Later, we fit this model on the test data.

The RMSE comes out to be **2.69** and the MAE is calculated to be **1.84**. Going for the calculation of the R^2 for the random forest model using the correlation coefficient gives the R^2 value to be **0.25.**
The following shows a scatterplot between the actual values vs predicted values for the model.



After developing these three models, we conclude the project by deciding on a final model to be used. We tabulate the results of these models to compare them. From our previous analysis on feature importance, we can clearly say that "Reason For Absence" is one of the most important variable. According to the Random Forest variable importance indicators, if we can mitigate the reasons for absence then the problem of absenteeism might improve.

5)Conclusion:

Tabulating the various error metrics rounded to two decimal places of the above three models to finalize on one:

Model Name	RMSE	MAE	R Squared
Random Forest Model	2.69	1.84	0.25
Linear Regression Model	2.76	1.97	0.18
Decision Tree Model	2.88	1.92	0.19

As we can see, Random Forest Model seems to have the least Root Mean Squared Error and the highest R Squared. This means it best explains the variance in the data. So we should use the Random Forest Model for the following problem of Employee Absenteeism in XYZ company.

The recommendations for the company to save on the loss of man hours have been clearly given in the exploratory data analysis of this section.

6)References:

- Datacamp
- Stack Overflow
- Code Academy