# ***DEBARUN BANERJEE***

***8th April 2019***

Employee Absenteeism – Project2

Contents

[**Chapter 1: Introduction** 2](#_Toc5499231)

[**1.1 Problem Statement** 2](#_Toc5499232)

[1. What changes company should bring to reduce the number of absenteeism? 2](#_Toc5499233)

[2. How much losses every month can we project in 2011 if same trend of absenteeism continues? 2](#_Toc5499234)

[**1.2 Dataset observations:** 2](#_Toc5499235)

[**1.3 Variables** 2](#_Toc5499236)

[**1.4 Sample Data** 4](#_Toc5499237)

[**1.5 Number of Unique Values** 5](#_Toc5499238)

[**Chapter 2: Methodology** 6](#_Toc5499239)

[**2.1 Pre – Processing** 6](#_Toc5499240)

[**2.2 Transforming Datatypes** 6](#_Toc5499241)

[ **In Python -** 7](#_Toc5499242)

[ **In R –** 10](#_Toc5499243)

[**2.3 Distribution of Categorical Variables:** 11](#_Toc5499244)

[**2.4 Distribution of Continues variables:** 14](#_Toc5499245)

[2.5 Standardization of categorical variables: 15](#_Toc5499246)

[**2.6 Outlier Analysis:** 15](#_Toc5499247)

[**2.7 Feature Selection:** 17](#_Toc5499248)

[**2.7.1 Correlation Plot:** 17](#_Toc5499249)

[**2.7.2 Chi-Square Test:** 18](#_Toc5499250)

[**2.8 Creating Dummy Variables for Categorical columns:** 18](#_Toc5499251)

[**2.9 Dimension Reduction using PCA:** 19](#_Toc5499252)

[Chapter 3: Modelling 19](#_Toc5499253)

[**3.1 Model Selection:** 19](#_Toc5499254)

[**3.2 Decision Tree:** 19](#_Toc5499255)

[**3.3 Random Forest:** 20](#_Toc5499256)

[**3.4 Linear Regression:** 20](#_Toc5499257)

[**Chapter 4: Conclusion:** 20](#_Toc5499258)

[**4.1 Model Evaluation:** 20](#_Toc5499259)

[**4.1.1 Root Mean Square Error (RMSE):** 20](#_Toc5499260)

[**4.1.2 R^2 value:** 20](#_Toc5499261)

# **Chapter 1: Introduction**

## **1.1 Problem Statement**

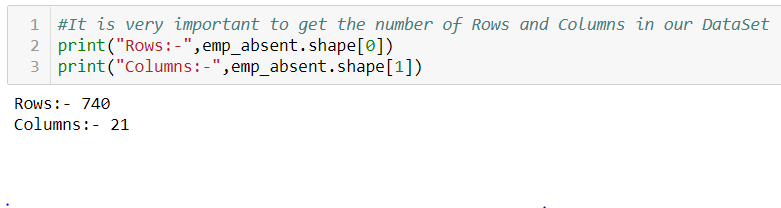
XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The company has shared it dataset and requested to have an answer on the following areas:

### 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?

## **1.2 Dataset observations:**

Before going to predict any model, it is very important to get the overview of the data that how many observations we have along with number of variables.



We can see, we have 740 observations with 21 variables in our dataset.

## **1.3 Variables**

There are 21 variables in our data in which 20 are independent variables and 1 (Absenteeism time in hours) is dependent variable. Since the type of target variable is continuous, this is a regression problem.

Variable 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

XXII medical consultation

XXIII blood donation

XXIV laboratory examination

XXV unjustified absence

XXVI physiotherapy

XXVII dental consultation

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 (KMs)

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

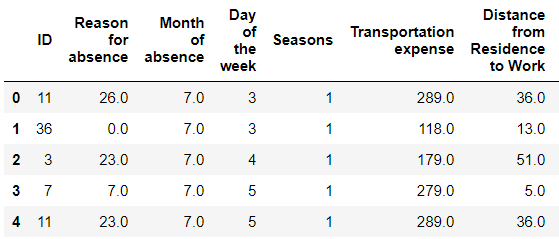
19. Height

20. Body mass index

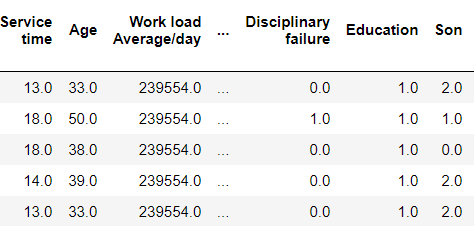
21. Absenteeism time in hours (target)

## **1.4 Sample Data**

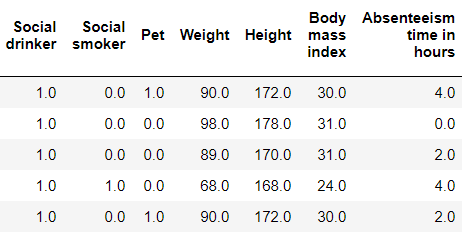
Variables (1-7)



**Variables (8-13)**

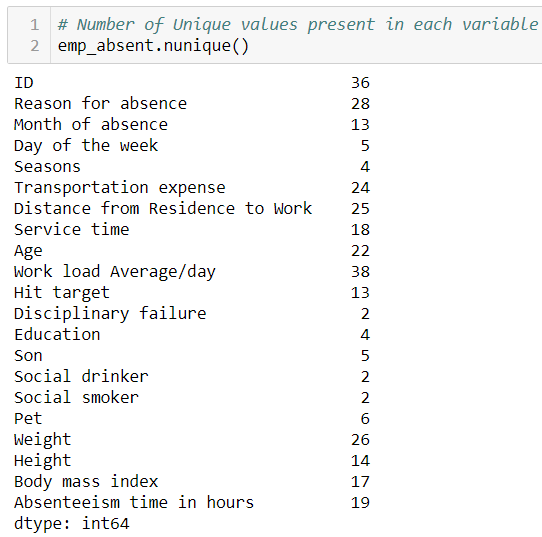


**Variables (14-21)**



## **1.5 Number of Unique Values**

In the below chart we can see the number of unique values in each variable.



# **Chapter 2: Methodology**

## **2.1 Pre – Processing**

In data mining, looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This process is known as Exploratory Data Analysis. In this project we will analyse missing values and the outliers present in the data. We will also look at the distribution of categorical variables and continuous variables.

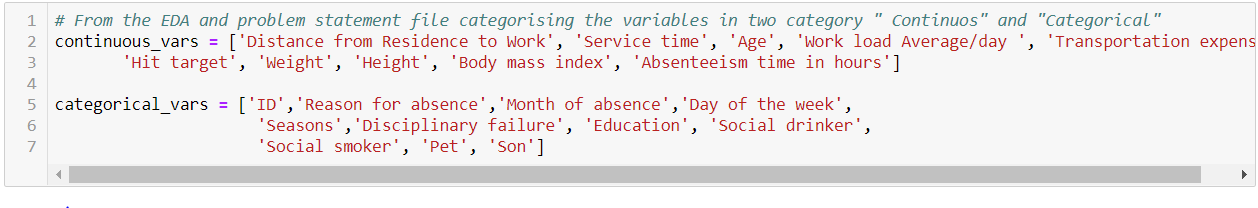
## **2.2 Transforming Datatypes**

Before entering missing value analysis, we need to check the datatypes of the dataset. Then we will transform the variables as required. Given below is the figure where we can see the datatypes of our dataset. So, we need to perform some operations to change these datatypes accordingly.



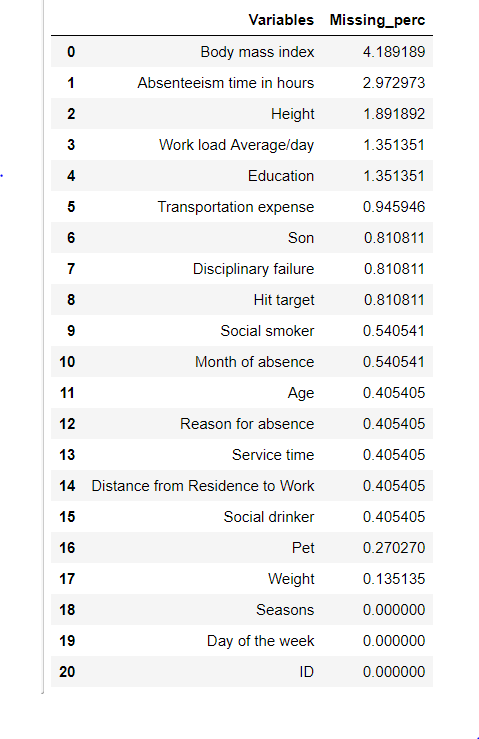


From the exploratory data analysis and problem statement given to us, we need to categorize the columns in “Continuous” and “Categorical” variables. Hereby in the below figure we are dividing columns into categorical and continuous variables.

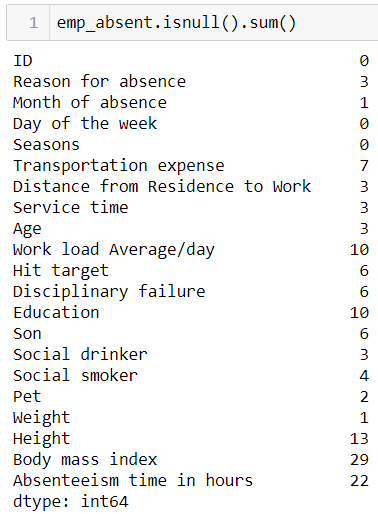
**2.3 Missing Value Analysis:**

* **In Python -**

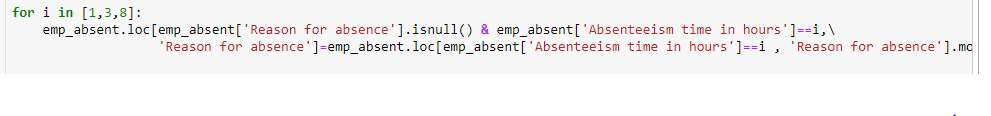
In statistics, missing data or missing values occur when no data value is stored for the variable in an observation. Missing values are a common occurrence in data analysis. These values can have a significant impact on the results or conclusions that would be drawn from these data. If a variable has more than 30% of its values missing, then those values can be ignored, or the column itself is ignored. In our case, none of the columns have a high percentage of missing values. The maximum missing percentage is 4.18% i.e., Body Mass Index column. Given below is the percentage of missing values in our variables.



Now we will investigate the number of missing values present in our data.



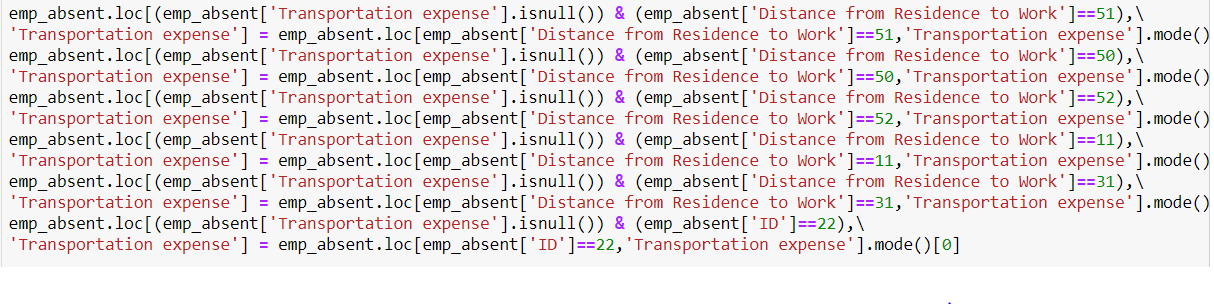
* Reason for absence is correlated with Absenteeism time in hours. Hence, we can see in our dataset that missing values are for which particular value of absent hours and then put their mode to fill the missing values.



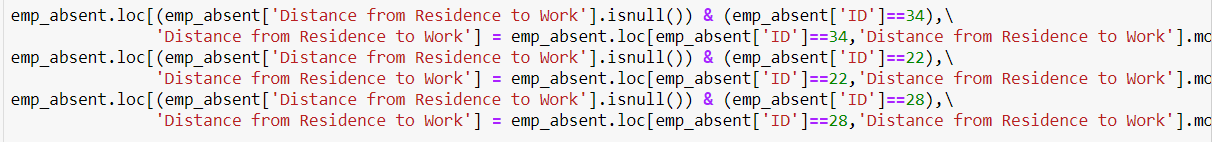
* For Month of absence we have taken the mode of this variable.



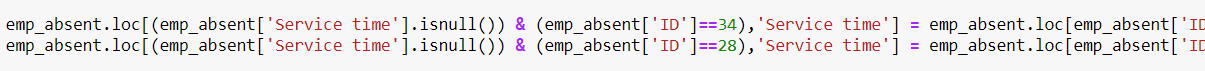
* Transportation expense is highly related with Distance from residence to wo work variable. Hence, we have taken the respective mode of those Distance from residence to work values and filled the missing values. For one Transportation expense missing value we have taken the mode of corresponding values for that id.



* We have seen the pattern of our dataset and found each id’s are having same Distance from residence to work. Hence, we have taken the mode for corresponding Distance from residence of each ids to impute missing values.



* We have seen that each id is having same Service time. Hence, we have imputed the mode of those corresponding ids.



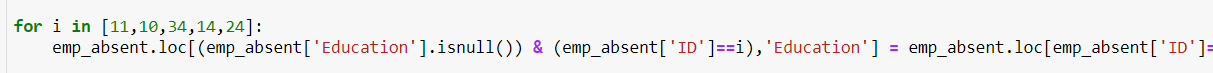
* Though each id is having same age, we have imputed the mode from each id to fill the missing values of Age column.



* For the categorical column “Disciplinary failure” we have imputed the mode of this variable.



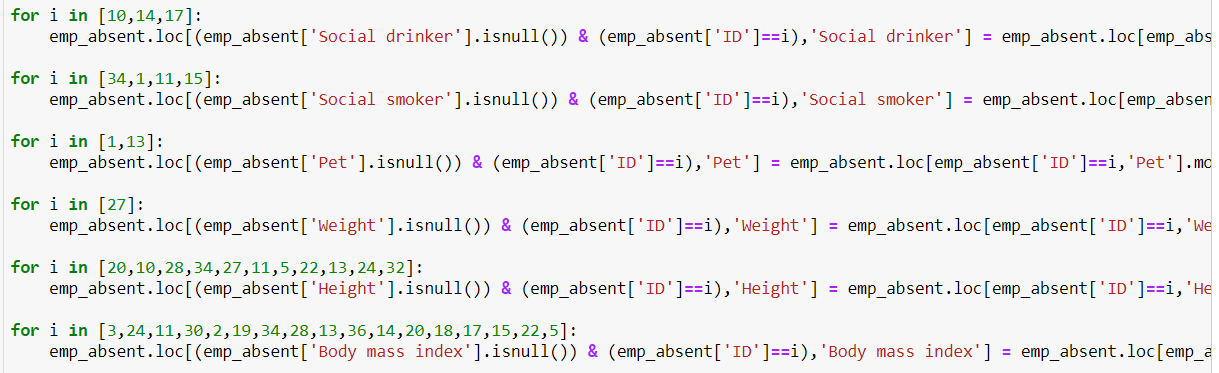
* As education is same for each id we have imputed the value from corresponding id to fill the missing values.



* Again, son of a id is same, hence we have imputed the mode of those ids to fill the missing values.



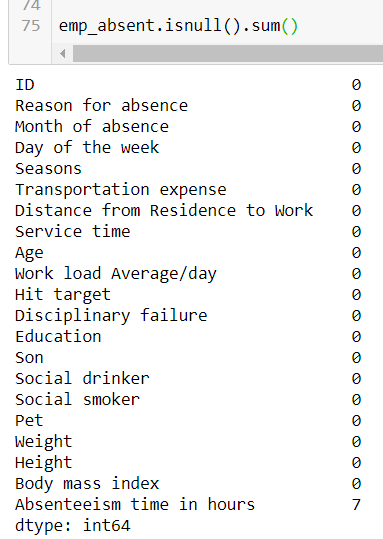
* For an id the value for Social Drinker, Social smoker, Pet, Height, Weight and Body mass index is same, hence to impute the missing values for those variables we used the mode of corresponding id.



* “Absenteeism time in hours” is highly related to the “Reason for absence”, hence we have imputed the mode of those corresponding “Reason for absence” values.



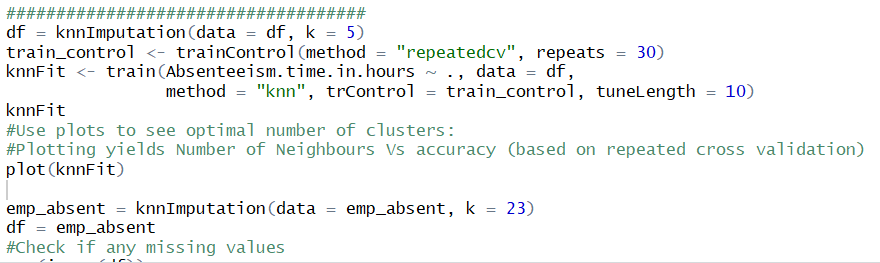
* Now we need to check the missing values for each column.



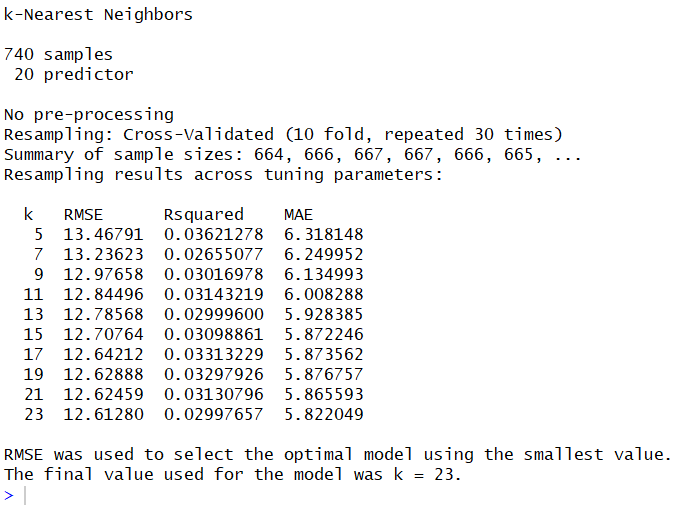
* **In R –**

In R, we have done kNN Imputation to impute the missing values. We have done separate processes so that we can compare the metrics in all possible ways.

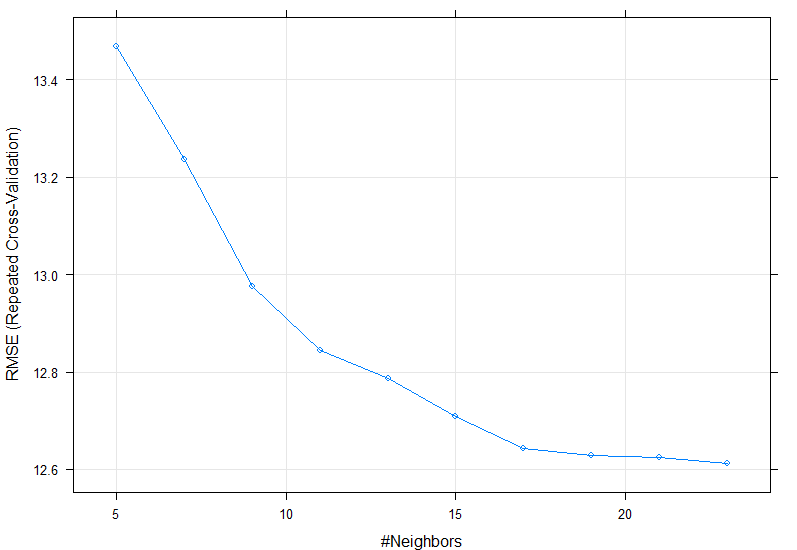
Given below is the process we followed to get the value of K. We have applied repeated cross validation process with 30 repetition and prepared 10 fold of our dataset.



We have used our copied dataset to get the value of K and used in our dataset “emp\_absent” to impute the missing values. Given below is the result of the knnFit.



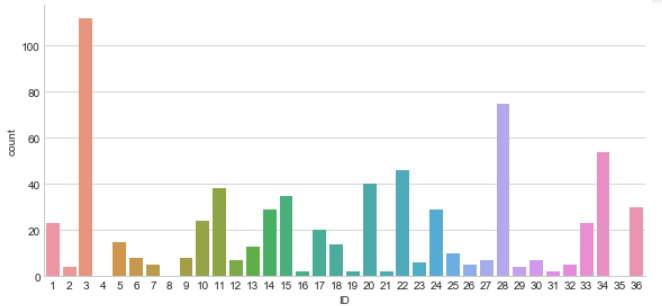
We can see that we will get the minimum Root Mean Square Error in K value of 23. Given below is the plot where we can visualize the value of K where it is minimum.



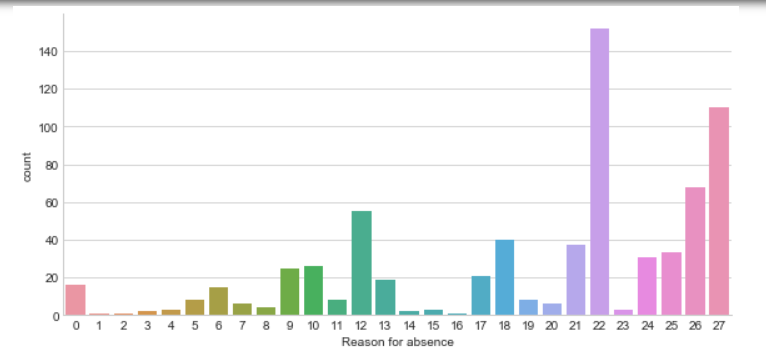
## **2.3 Distribution of Categorical Variables:**

Given below is the plots for categorical variable which will give us the idea regarding counts for each category.

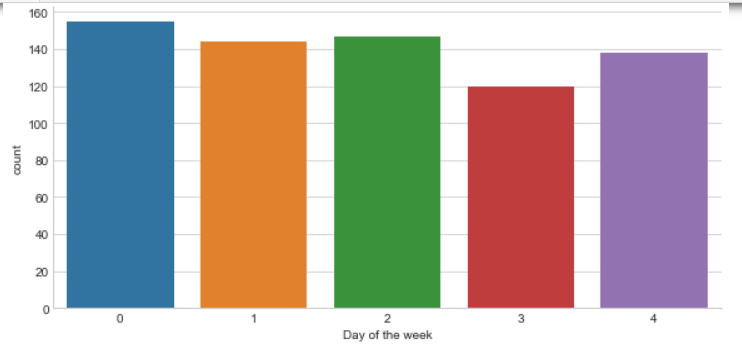
* ID - The company may act warn such employees to reduce being absent a lot or if repeated further, it can against them if necessary. ID 3 is most absent in our given period of dataset.



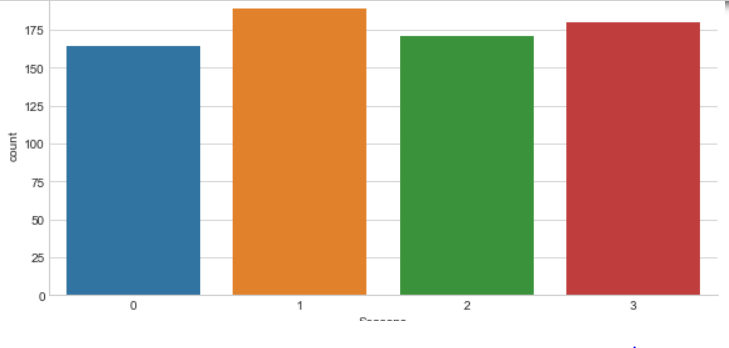
* Reason for absence - Reason 22 and 28 are the reasons employees give most for being absent. The company XYZ can help in informing employees on how to keep themselves healthier by having monthly campus consultations.



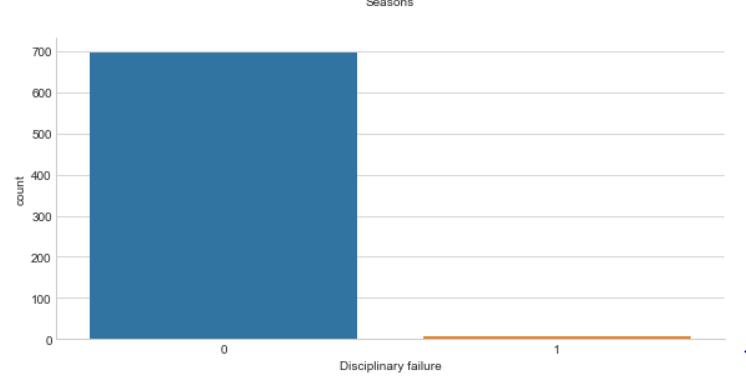
* Day of the week- Employees are absent most on Monday and least on Thursday.



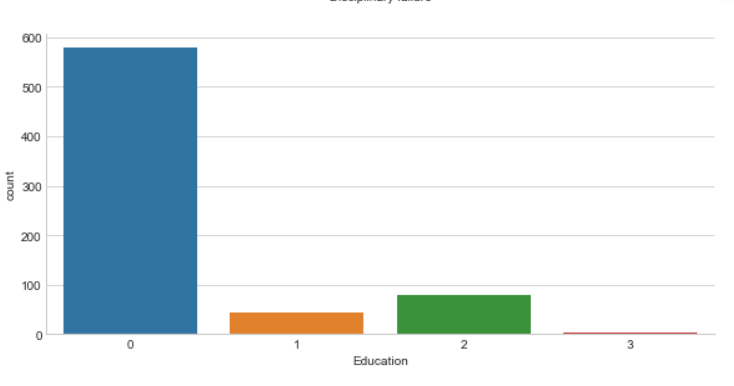
* Seasons- We can see from the below plot that employees are absent more on Autumn (1) and least in summer (0).



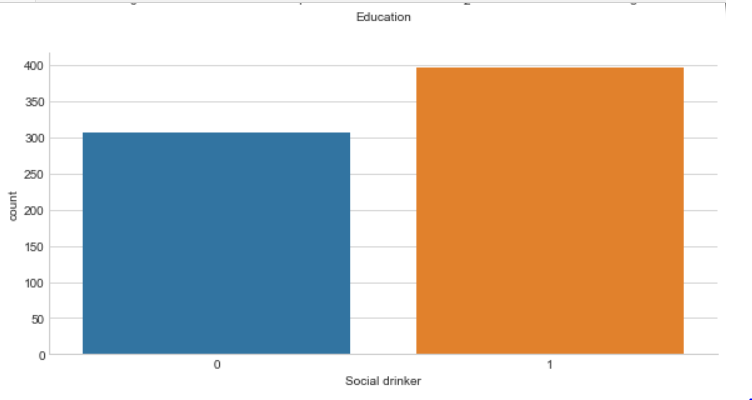
* Disciplinary failure- Employees with No disciplinary failure are more absent over employees with Disciplinary failure.



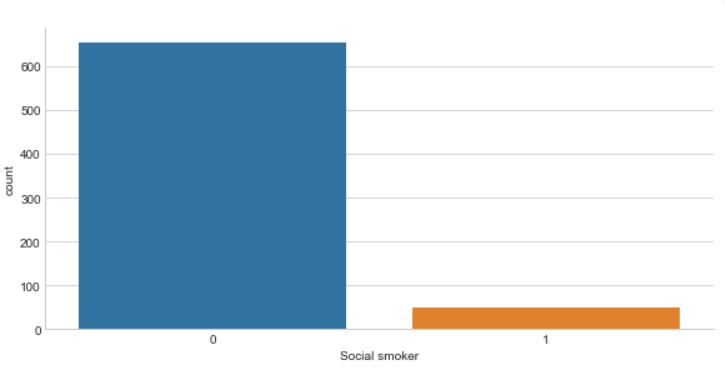
* Education – Employees with High school degree are more absent. So, the company can either hire employees who have at least graduated from college or inform those employees who have completed only their high school education to reduce the number of hours they are absent.



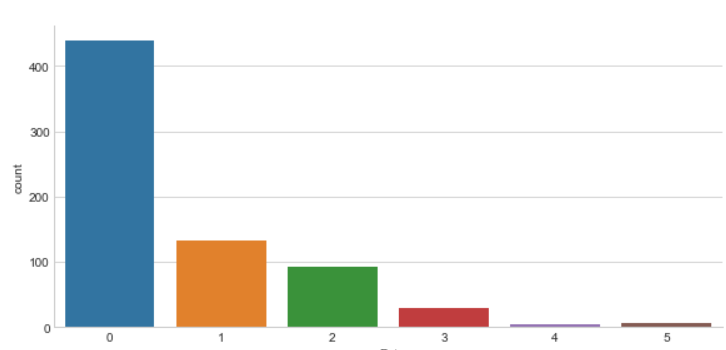
* Social Drinker: Employees who are social drinker are more absent than the people who are not social drinker. XYZ can keep a track of those people and inform those employees to reduce the intake of alcohol during working days.



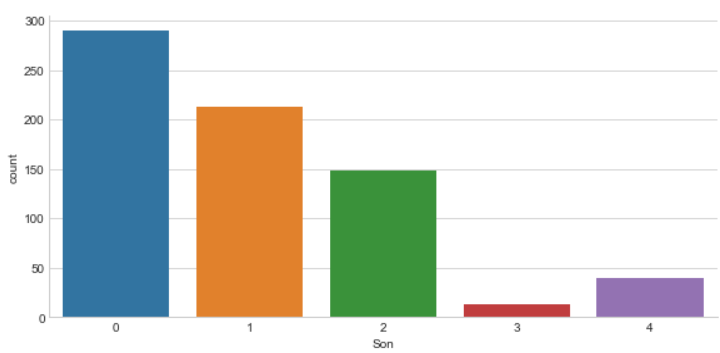
* Social Smoker: Employees who smoke are less absent then the employees who doesn’t smoke.



* Pet- Employees with no pet are most absent we can see from the below plot.

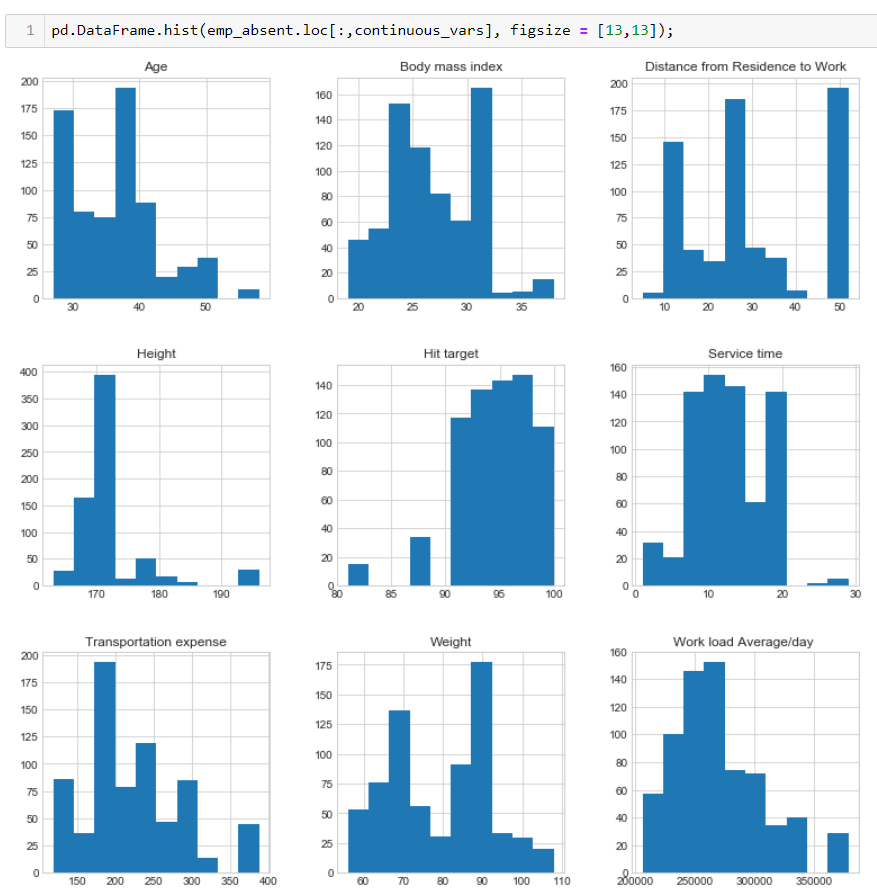


* Son- Given below is the plot where we can see that employees with no sons are more absent compare to the employees who have son.



## **2.4 Distribution of Continues variables:**

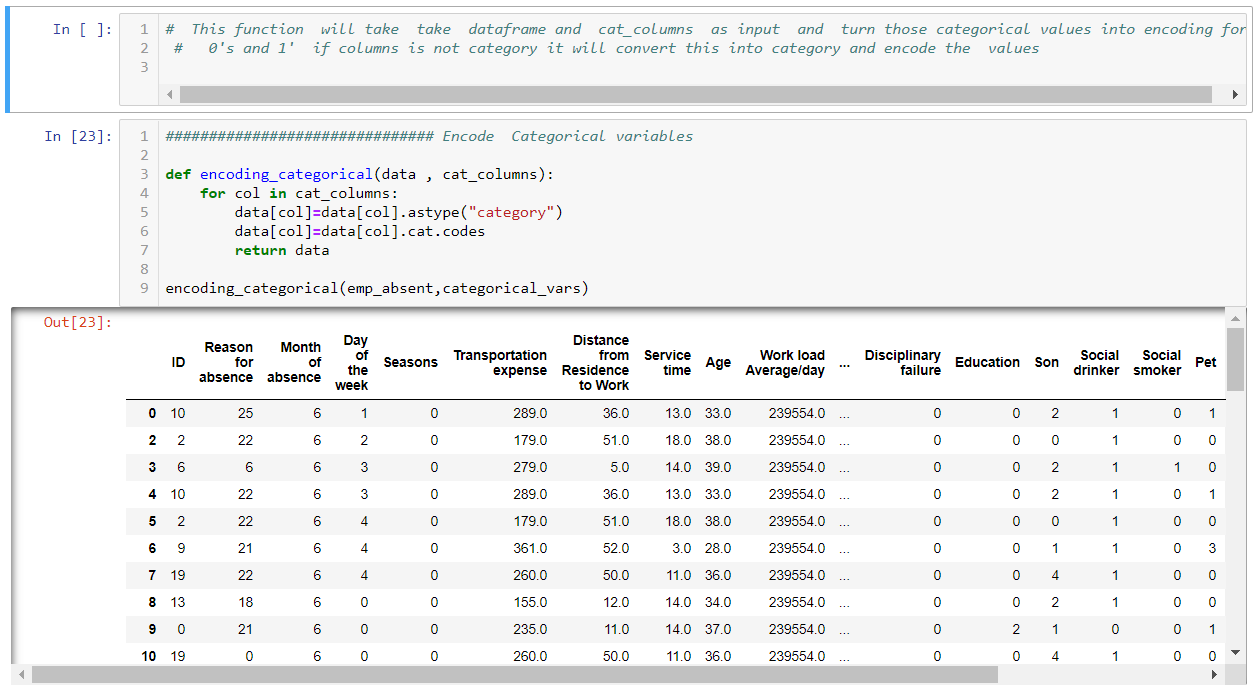
Given below are the plots of continues variables. We have plotted Histogram to check the normality of the variables.



It can be observed that none of the variables are normally distributed. So, we need to go for normalization for continues variables.

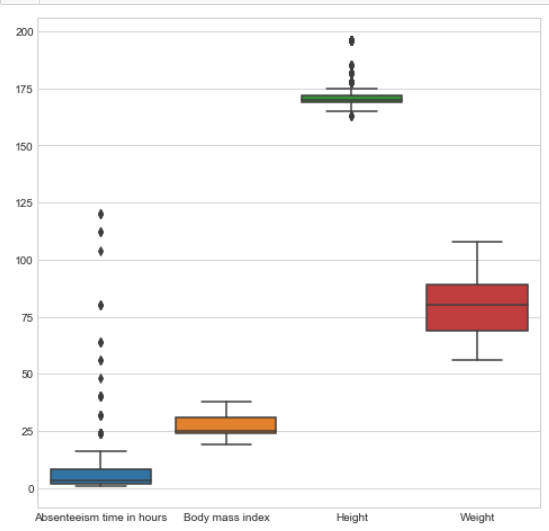
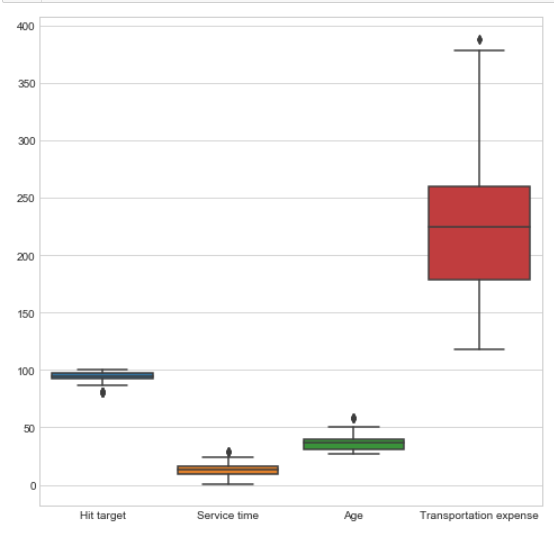
## 2.5 Standardization of categorical variables:

Before proceeding towards building any model, we need to standardize our categorical columns to numerical values to each factor. We have few factor columns in our dataset, though they are already input as number still we have done so that we can get default index for respective categories.

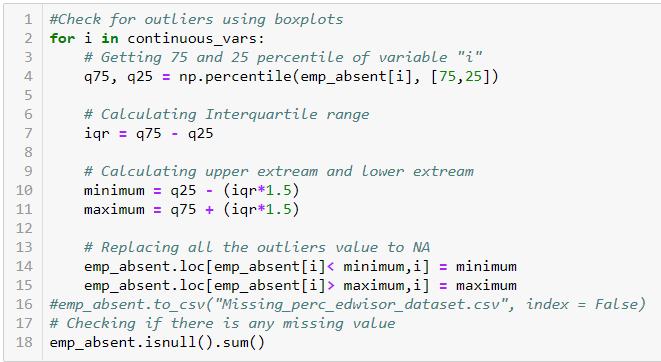


## **2.6 Outlier Analysis:**

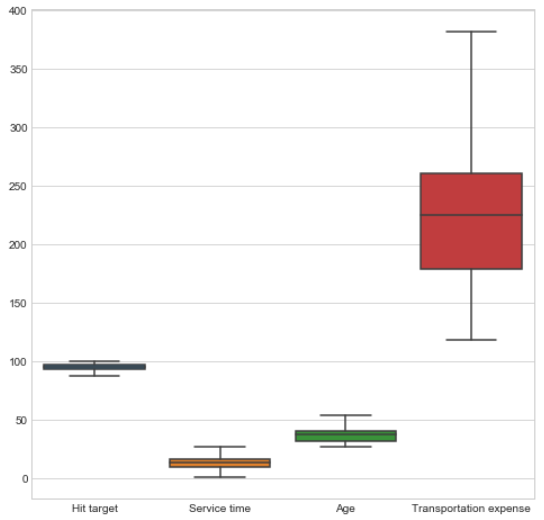
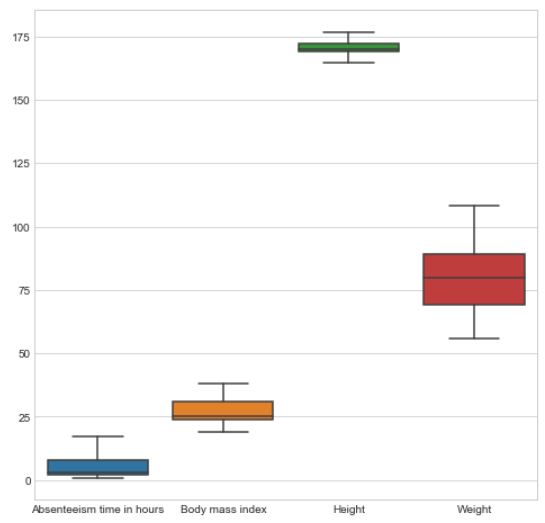
It can be observed from the distribution of variables that almost none of the variables are normally distributed. The skew in these distributions can be explained by the presence of outliers and extreme values in the data. One of the steps in pre-processing involves the detection and removal of such outliers. In this project, we use boxplot to visualize and remove outliers. Any value lying outside of the lower and upper whisker of the boxplot are outliers.

From the above plot, we can see that except “Body mass index”, “Weight” rest of the continuous variables have outliers, we can see points above or below any of the whisker. Now, we need to impute the outliers with predicted values. We have checked if the datapoint lies below the lower whisker then we replaced them with minimum value and if the datapoint lies above the upper whisker then we impute those outliers with maximum value.



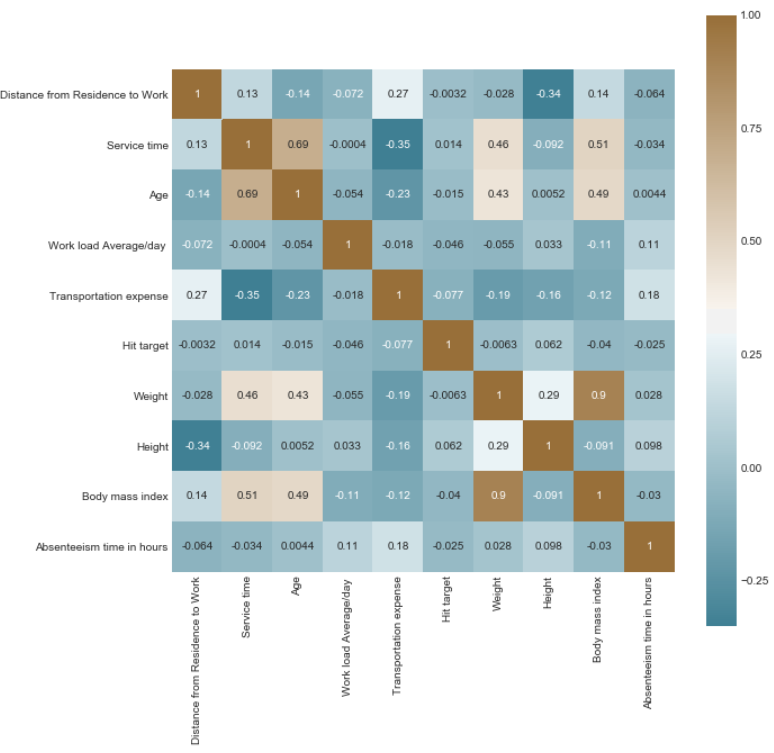
We have first calculated the 75th and 25th quartile and then we have calculated the inter quartile range which is nothing but the difference between the 75th and 25th quartile. Then, we got the value of minimum which is the difference between 25th quartile and 1.5 times of iqr value. Likewise, the maximum value is the addition of 75th quartile and 1.5 times of iqr value. After imputing the outliers, we have plotted the boxplots again can see there is no outlier detected anymore.



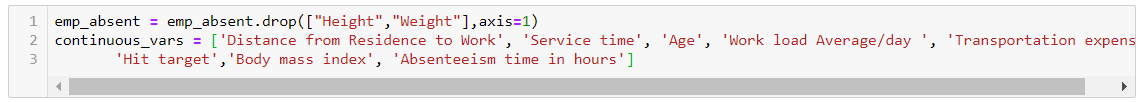
## **2.7 Feature Selection:**

We must select the possible best features to build our train data. We have plotted correlation plot for numeric columns and have performed Chi Square Test for categorical variables. Feature Selection reduces the complexity of a model and makes it easier to interpret. It also reduces overfitting. Features are selected based on their scores in various statistical tests for their correlation with the outcome variable. Correlation plot is used to find out if there is any multicollinearity between variables. The highly collinear variables are dropped and then the model is executed.

**2.7.1 Correlation Plot:** - The Correlation coefficient measures the linear relationship between two datasets. Pearson’s Correlation requires that each dataset to be normally distributed. Like other correlation coefficient it also has range between +1 to -1. 0 implies there is no correlation, 1 implies there is a linear relationship where y will also increase with increment of the value of x. -1 denotes there is a linear relationship but in opposite direction.



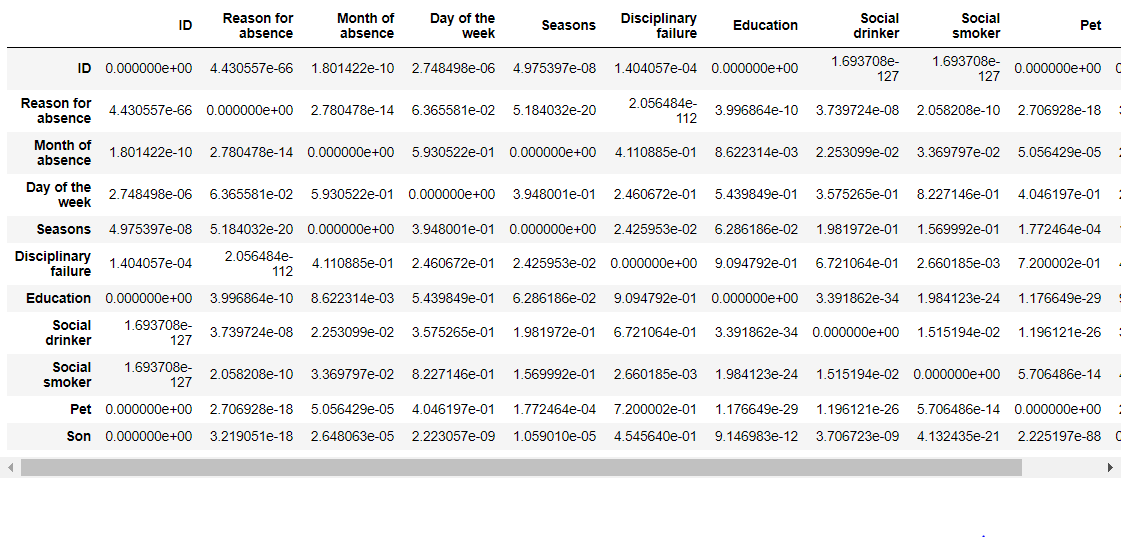
From our outcome of Correlation plot, we can see that “Height” and “Weight” is having highly correlated with “Body mass index”. Hence, we can drop “Height and Weight” and keep only the column “Body mass index”.

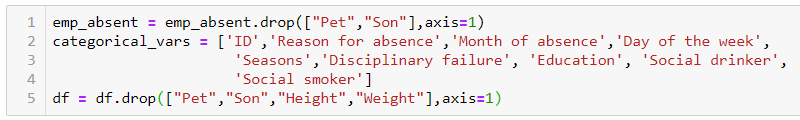


Hence, after the plot, we have simply removed those two variables from my dataset and updated the continues variable.

### **2.7.2 Chi-Square Test:**

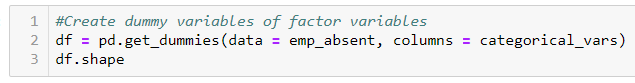
We often use the Chi Square test to view the dependencies between categorical variables before fitting the model to test and predict data. Pearson Chi square gives us the value of Chi square between each variable and we can eliminate the variables which are having less chi square value.

We can see from the above picture that we have two columns “Pet” and “Son” which is showing p-value of 0 with ID column. We can suspect that for an id, the value of Pet of Son for this given data is same. Hence, we can keep the id column only and can remove the “Pet” and “Son” column.



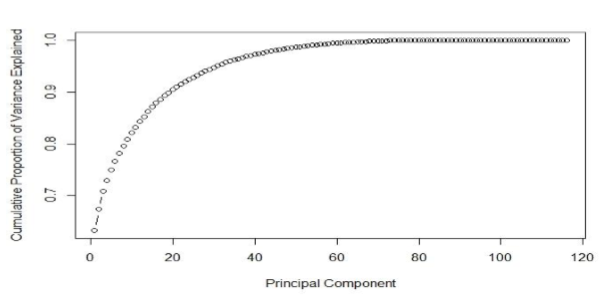
Hence, we have eliminated columns for both categorical and continuous columns and updated our dataset for further model building steps.

**2.8 Creating Dummy Variables for Categorical columns:**   
Before creating any model, we need to create dummy variables for our categorical columns with the number of categories from each variable. This process basically creates those many columns of each categories and put 1 for the corresponding observations and for the rest of the categories it creates a zero in respective places.



## **2.9 Dimension Reduction using PCA:**

Principal component analysis is a technique for *feature extraction.* It combines our input variables in a specific way, then we can drop the “least important” variables. *As an added benefit, each of the “new” variables after PCA are all independent of one another.* So, we can choose the number of Principal components that covers most of the features. We have plotted below PCA graph.



It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. After creating dummy variable of categorical variables, the data would have 116 columns and 740 observations. This high number of columns leads to bad accuracy.

# Chapter 3: Modelling

## **3.1 Model Selection:**

After cleaning our data with several pre processing technique, now we will create few regression models to predict our target variable. As our target variable “Absenteeism time in hours” is a continuous variable so we will build Decision Tree, Random Forest and Linear Regression. There are few regression metrics named Root Mean Square Error (RMSE), Co efficient of Determination R^2 we have checked to select the final model.

## **3.2 Decision Tree:**

Decision tree is a supervised machine learning algorithm which is generally used for both regression and classification problems. Here, each node represents a attribute and each branches denote a rule or decision and each leaf is a outcome which is in our case target variable.  
Basically we are going to create a decision tree model which will predict the target variable of our test data.

The RMSE values and R^2 values for the given project in R and Python are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature  Selection | RMSE | | R^2 | |
| R | Python | R | Python |
| Before | 1.164 | 4.652 | 0.388 | -0.019 |
| After | 0.444 | 0.053 | 0.97 | 1 |

## **3.3 Random Forest:**

Random forest is also a supervised machine learning model which is a collection of Decision trees and merge them together to get a more accurate and stable prediction. The method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner.

The RMSE values and R^2 values for the given project in R and Python are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature  Selection | RMSE | | R^2 | |
| R | Python | R | Python |
| Before | 0.163 | 3.579 | 0.388 | 0.293 |
| After | 0.48 | 0.001 | 0.97 | 0.999 |

## **3.4 Linear Regression:**

Linear Regression is a model to find a relationship between one or more feature (independent variables) and a continuous target variable (dependent variable). When there is only feature it is called Uni-variate Linear Regression and if there are multiple features, it is called Multiple Linear Regression. Here The line for which the error between the predicted values and the observed values is minimum is called the best fit line or the regression line. These errors are also called as residuals.

The RMSE values and R^2 values for the given project in R and Python are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature  Selection | RMSE | | R^2 | |
| R | Python | R | Python |
| Before | 0.174 | 3.729 | 0.344 | 0.232 |
| After | 0.003 | 5.972 | 0.003 | 1 |

# **Chapter 4: Conclusion:**

## **4.1 Model Evaluation:**

We have used the Root Mean Square Error (RMSE) and R^2 value to evaluate our predicted models.

### **4.1.1 Root Mean Square Error (RMSE):**

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are.

### **4.1.2 R^2 value:**

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.

R-squared is always between 0 and 100%:

* 0% indicates that the model explains none of the variability of the response data around its mean.
* 100% indicates that the model explains all the variability of the response data around its mean.

The higher the R-squared, the better the model fits our data. We use R square in case of precise precision.

## **4.2 Model Selection:**

## From the above understanding we can conclude that Linear Regression is having minimum RMSE and maximum R squared.

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

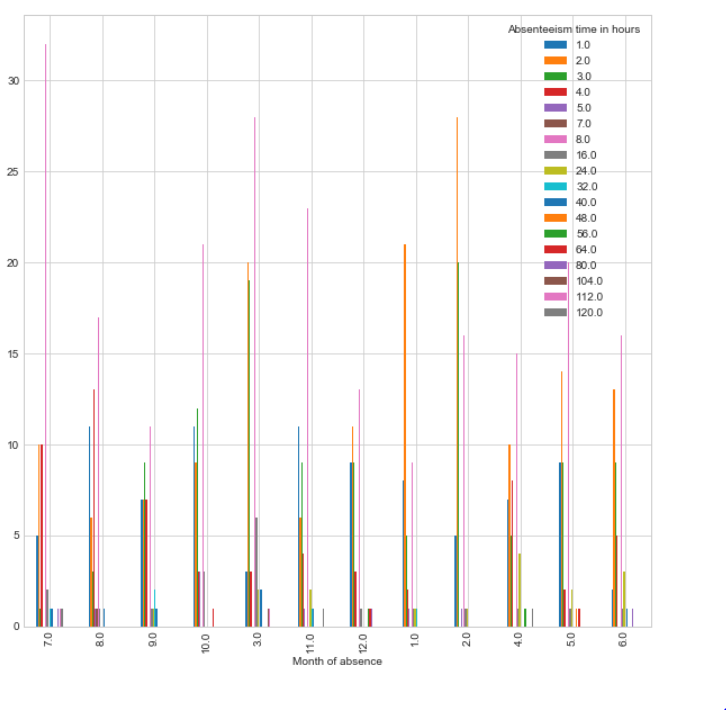
Please refer to the Distribution of Variables explained.

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

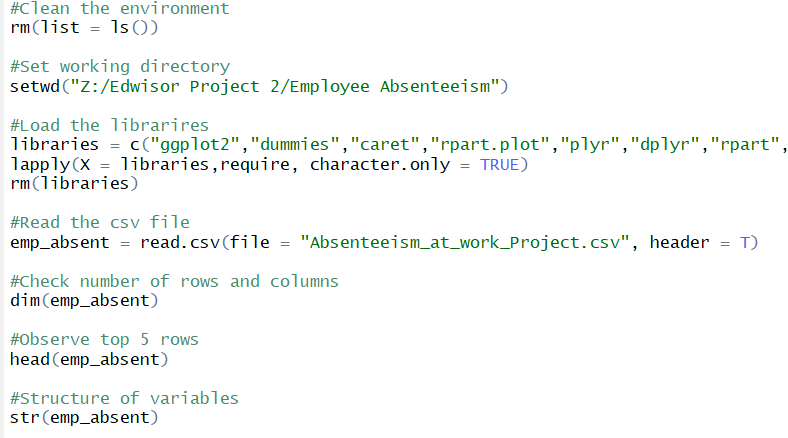
If the rate of absenteeism is same, then we can presume the amount of absenteeism for each month in the below picture.

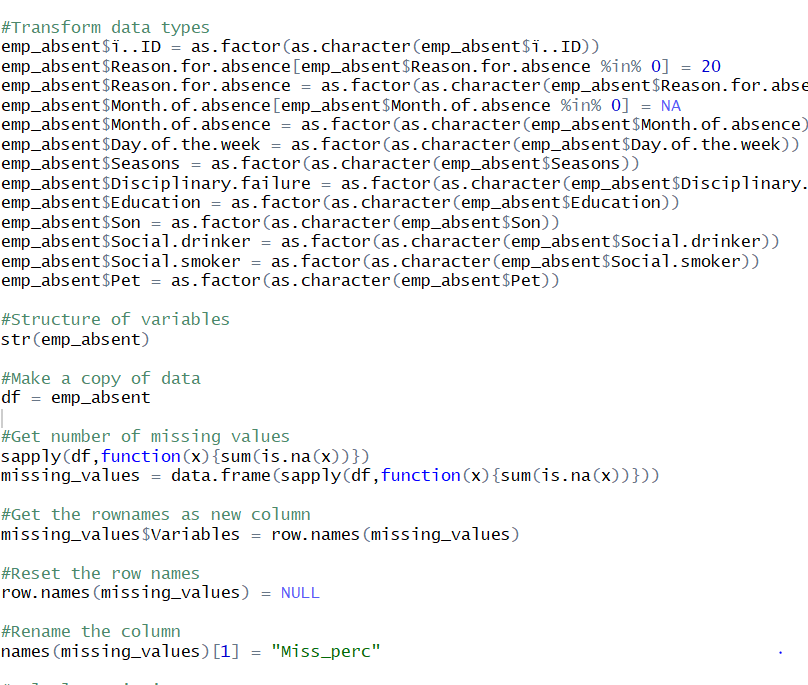
## 

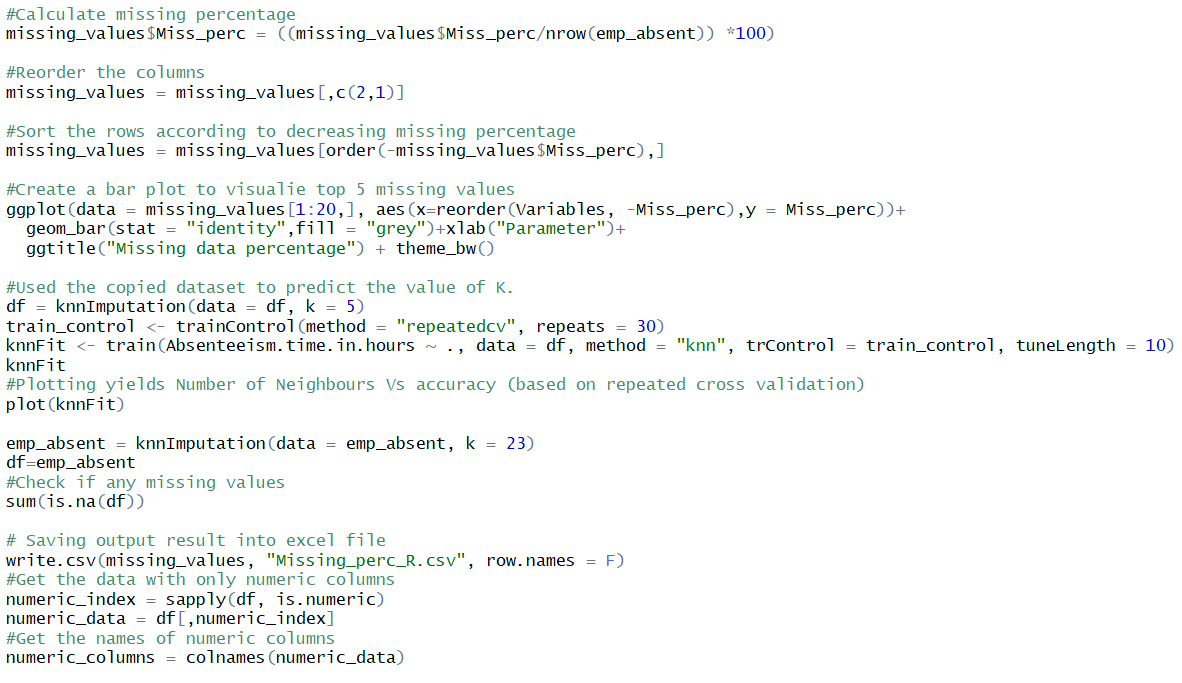
We can refer our detailed plots to see the absenteeism hour how it deviates between Month of Absence.



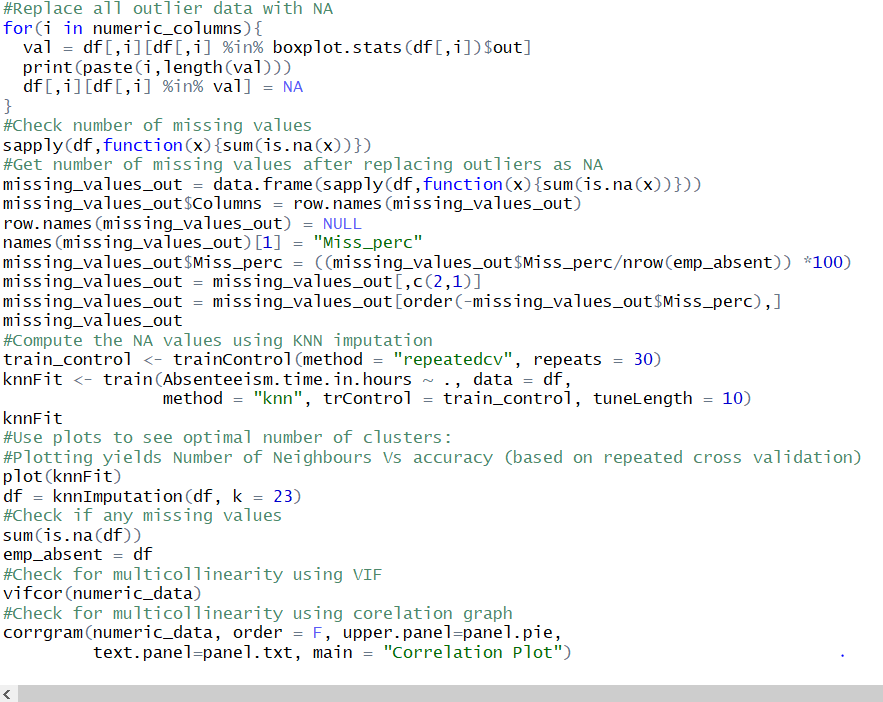
# **R Code:**

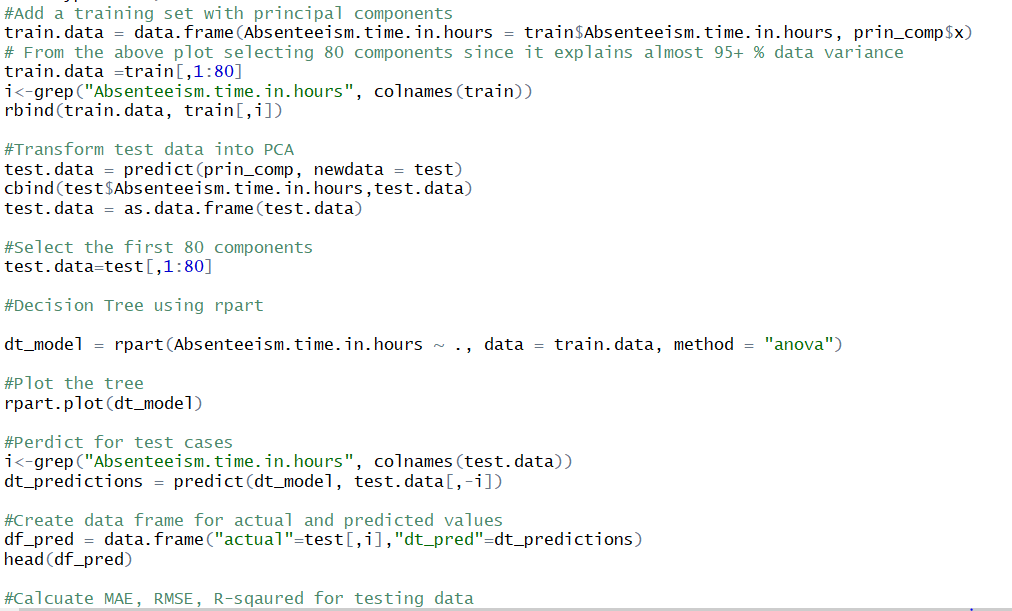


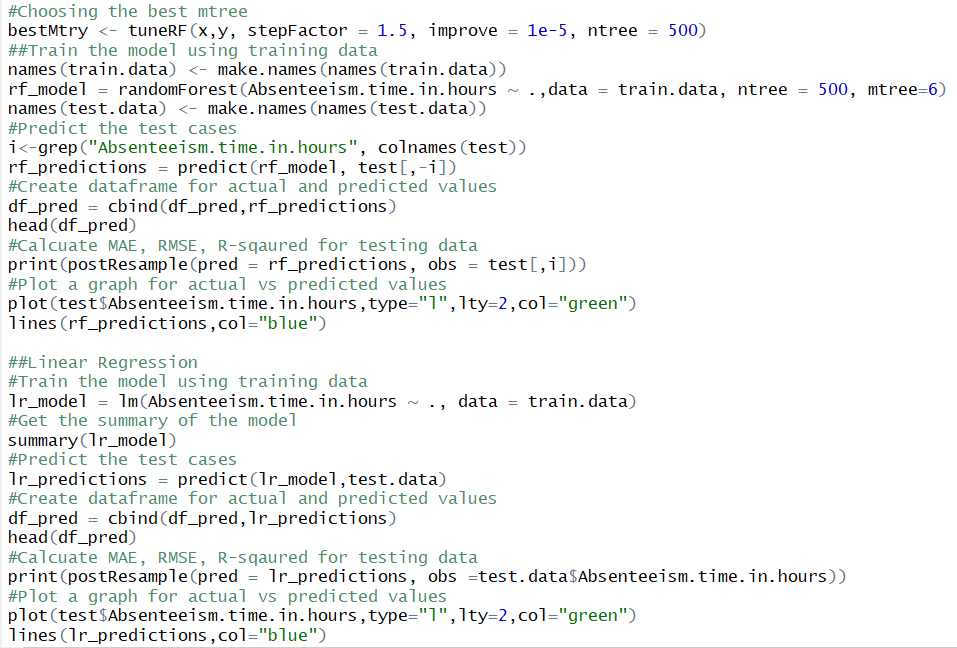












Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx Thank You xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx