Project Report
Employee Absenteeism

Krishna R

19 JULY 2019

Contents

1. Introduction

2. Methodology

3. Conclusion

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 Data

There are 21 variables in our data in which 20 are independent variables and 1 (Absenteeism time in hours) is dependent variable. There are 13 continuous variables and 8 categorical variables. Since our target variable is continuous in nature, this is a regression problem.

Variables 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 behavioral disorders

VI. Diseases of the nervous system

VII. Diseases of the eye and adnexa

VIII. Diseases of the ear and mastoid process

IX. Diseases of the circulatory system

X. Diseases of the respiratory system

XI. Diseases of the digestive system

XII. Diseases of the skin and subcutaneous tissue

XIII. Diseases of the musculoskeletal system and connective tissue

XIV. Diseases of the genitourinary system

XV. Pregnancy, childbirth and the puerperium

XVI. Certain conditions originating in the perinatal period

XVII. Congenital malformations, deformations and chromosomal abnormalities

XVIII. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX. Injury, poisoning and certain other consequences of external causes

XX. External causes of morbidity and mortality

XXI. Factors influencing health status and contact with health services

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

3. Month of absence

4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons (summer (1), autumn (2), winter (3), spring (4))

6. Transportation expense

7. Distance from Residence to Work (kilometers)

8. Service time

9. Age

10. Work load Average/day

11. Hit target

12. Disciplinary failure (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

14. Son (number of children)

15. Social drinker (yes=1; no=0)

16. Social smoker (yes=1; no=0)

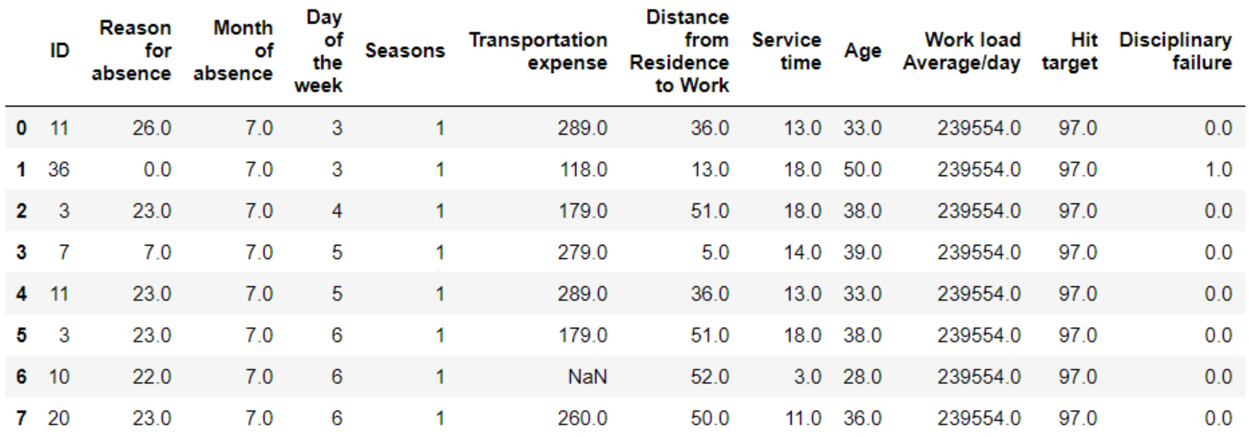
17. Pet (number of pet)

18. Weight

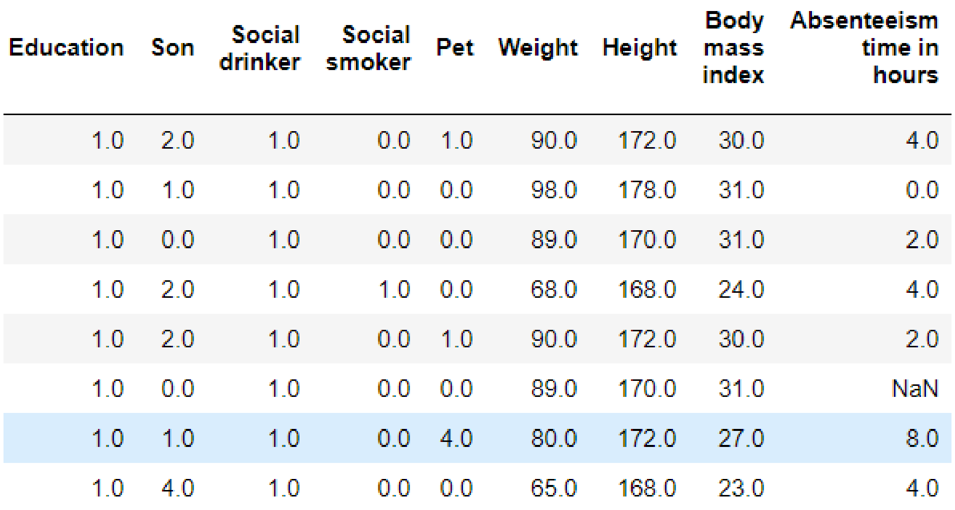
19. Height

20. Body mass index

21. Absenteeism time in hours (target)



	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure
0	11	26.0	7.0	3	1	289.0	36.0	13.0	33.0	239554.0	97.0	0.0
1	36	0.0	7.0	3	1	118.0	13.0	18.0	50.0	239554.0	97.0	1.0
2	3	23.0	7.0	4	1	179.0	51.0	18.0	38.0	239554.0	97.0	0.0
3	7	7.0	7.0	5	1	279.0	5.0	14.0	39.0	239554.0	97.0	0.0
4	11	23.0	7.0	5	1	289.0	36.0	13.0	33.0	239554.0	97.0	0.0
5	3	23.0	7.0	6	1	179.0	51.0	18.0	38.0	239554.0	97.0	0.0
6	10	22.0	7.0	6	1	NaN	52.0	3.0	28.0	239554.0	97.0	0.0
7	20	23.0	7.0	6	1	260.0	50.0	11.0	36.0	239554.0	97.0	0.0



Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
1.0	2.0	1.0	0.0	1.0	90.0	172.0	30.0	4.0
1.0	1.0	1.0	0.0	0.0	98.0	178.0	31.0	0.0
1.0	0.0	1.0	0.0	0.0	89.0	170.0	31.0	2.0
1.0	2.0	1.0	1.0	0.0	68.0	168.0	24.0	4.0
1.0	2.0	1.0	0.0	1.0	90.0	172.0	30.0	2.0
1.0	0.0	1.0	0.0	0.0	89.0	170.0	31.0	NaN
1.0	1.0	1.0	0.0	4.0	80.0	172.0	27.0	8.0
1.0	4.0	1.0	0.0	0.0	65.0	168.0	23.0	4.0

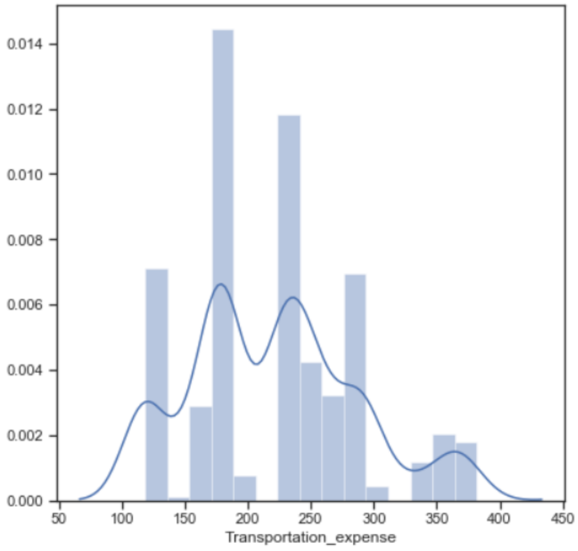
Chapter 2

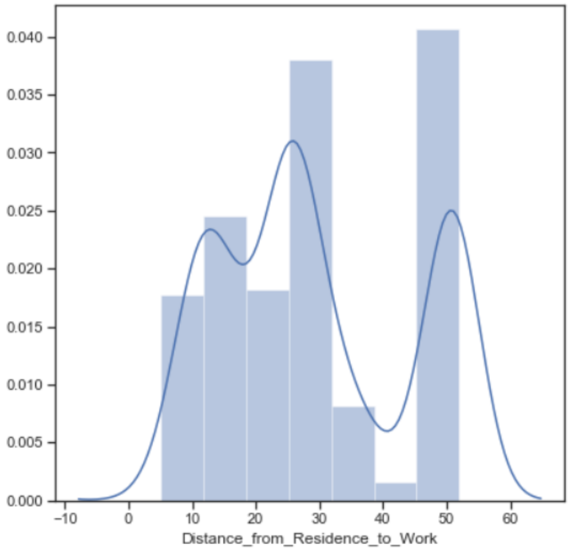
Methodology

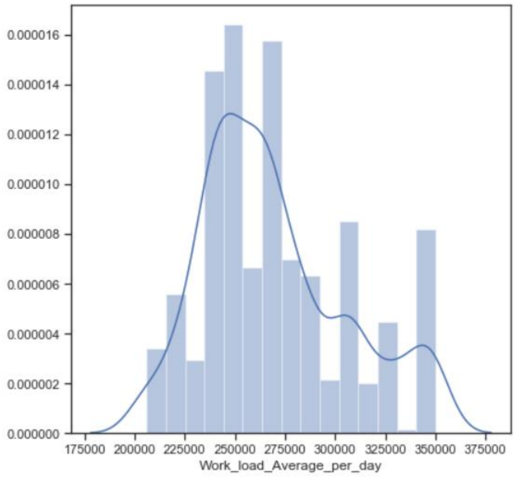
Before feeding the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project we spend almost 80% of time in it.

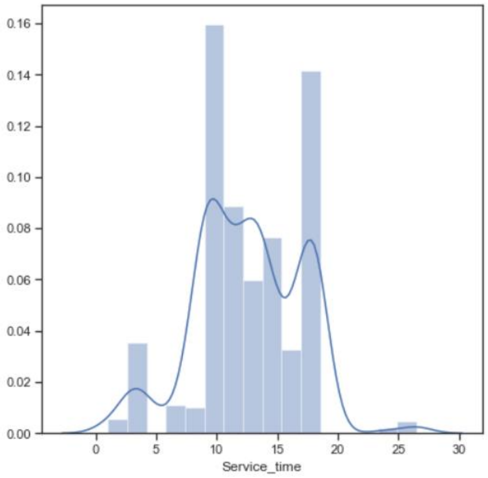
2.1 Pre Processing the data

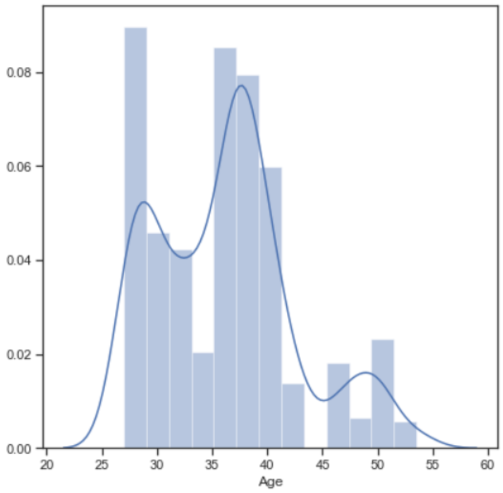
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process we will first try and look at all the probability distributions of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable

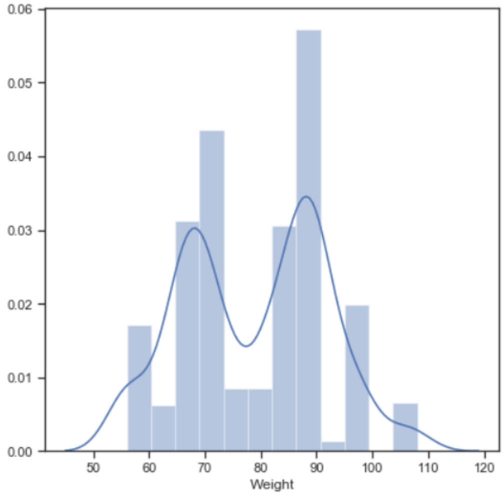


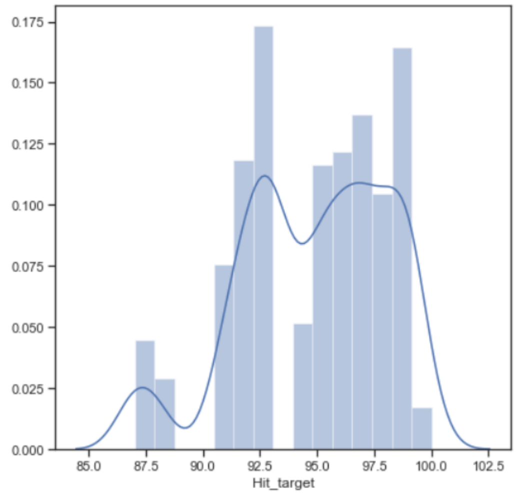


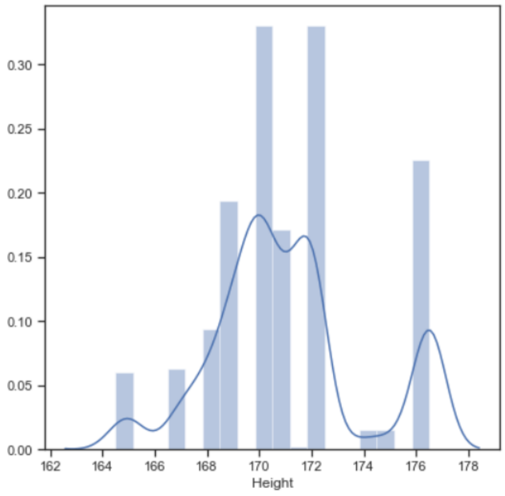

0.000016
0.000014
0.000012
0.000010
0.000008
0.000006
0.000004
0.000002
0.000000
175000 200000 225000 250000 275000 300000 325000 350000 375000
Work_load_Average_per_day

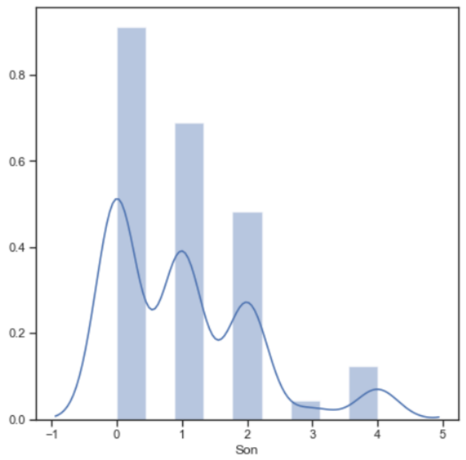

0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0 5 10 15 20 25 30
Service_time

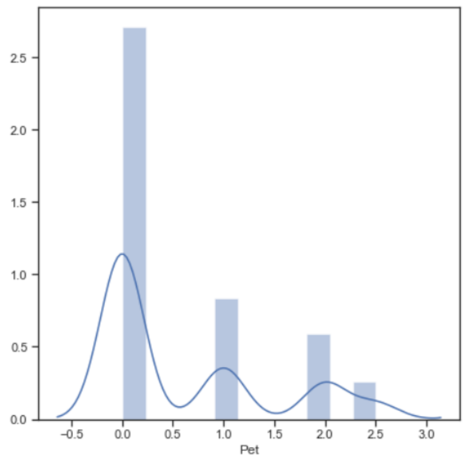

0.08
0.06
0.04
0.02
0.00
20 25 30 35 40 45 50 55 60
Age

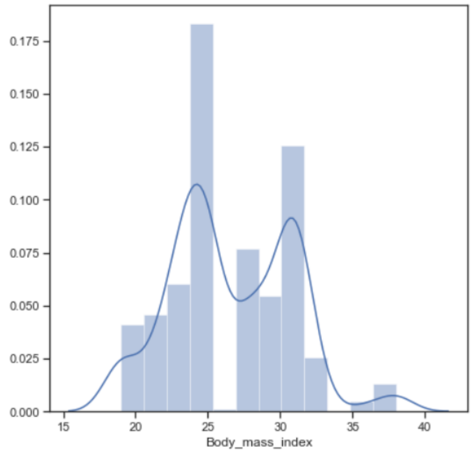

0.06
0.05
0.04
0.03
0.02
0.01
0.00
50 60 70 80 90 100 110 120
Weight

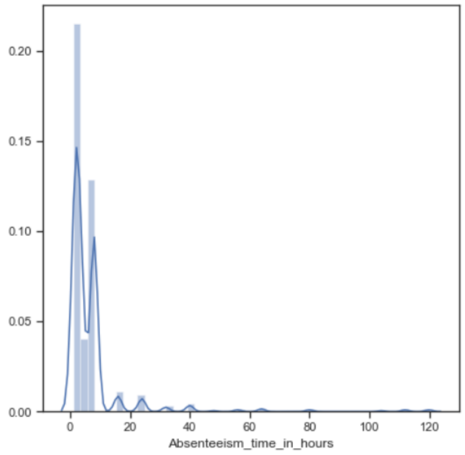

0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
85.0 87.5 90.0 92.5 95.0 97.5 100.0 102.5
Hit_target


0.30
0.25
0.20
0.15
0.10
0.05
0.00
162 164 166 168 170 172 174 176 178
Height


0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5
Son

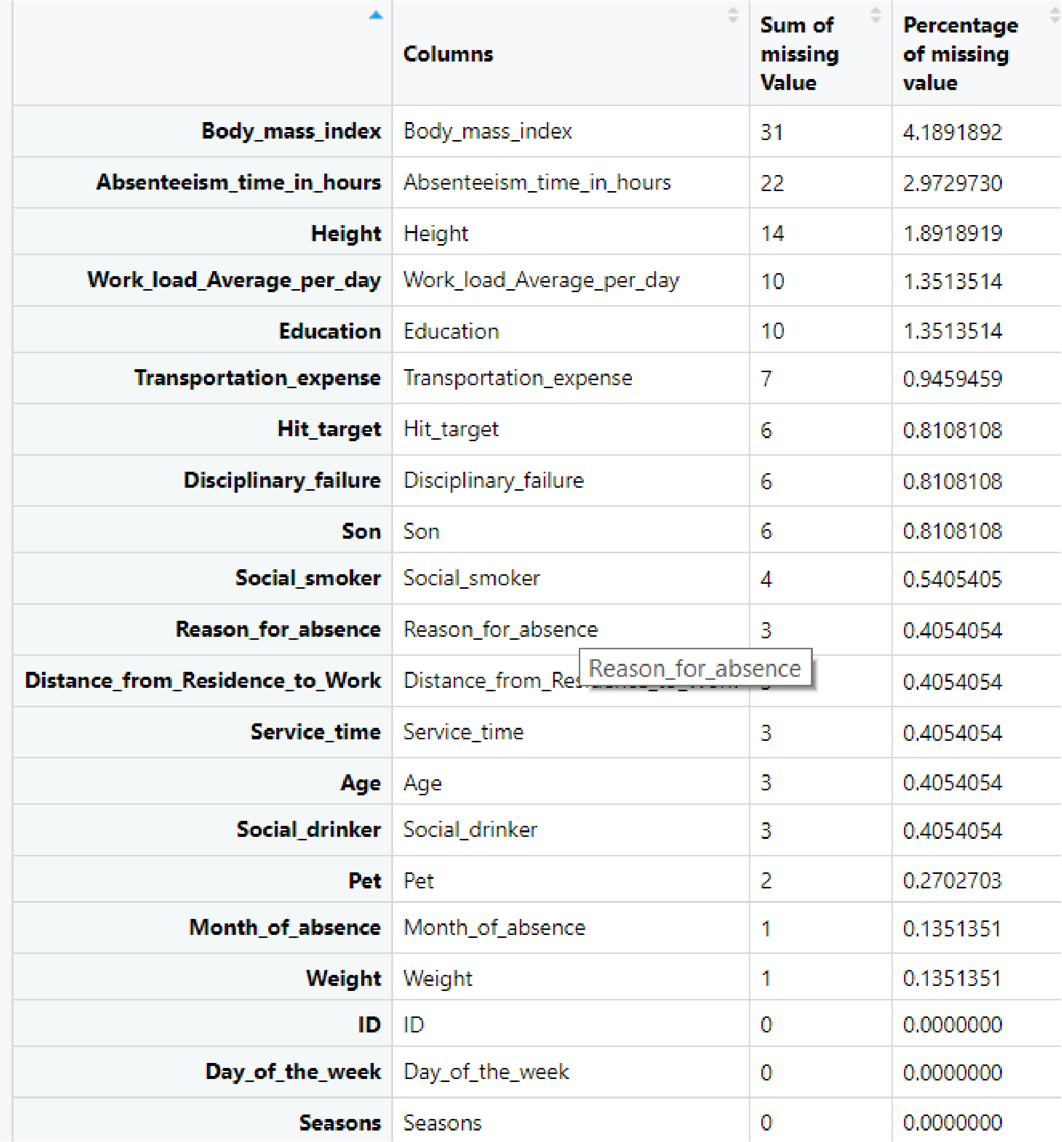

2.5
2.0
1.5
1.0
0.5
0.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Pet


0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
15
20
25
30
35
40
Body_mass_index


0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
120
Absenteeism_time_in_hours

2.2.1 Missing Value Analysis

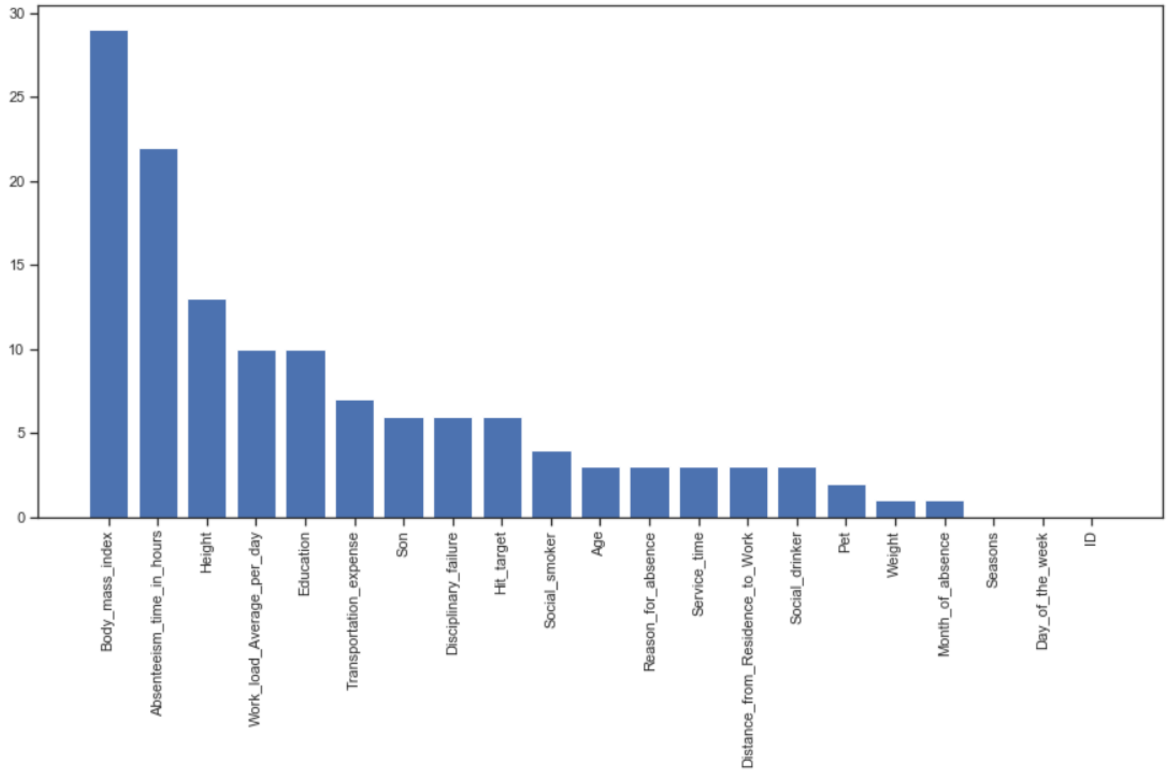
In statistics, *missing data*, or *missing values*, occur when no *data value* is stored for the variable in an observation. *Missing data* are a common occurrence and can have a significant effect on the conclusions that can be drawn from the *data*. If a columns has more than 30% of data as missing value either we ignore the entire column or we ignore those observations. In the given data the maximum percentage of missing value is 4.189% for **body mass index** column. To impute missing values we can use mean, median, mode or KNN method but here I chose to impute manually. Because with help of ID columns we can impute the data points as the IDs are repetitive in the data.



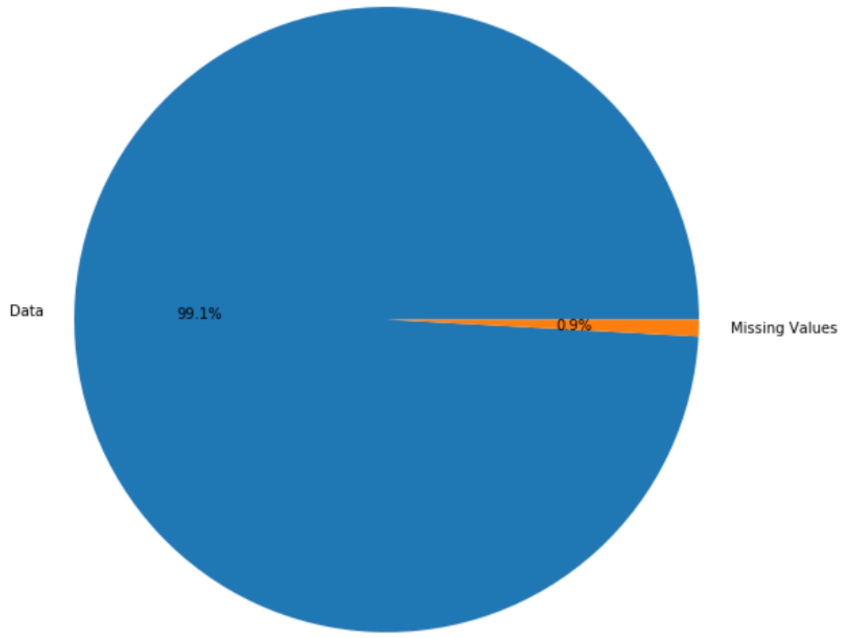
	Columns	Sum of missing Value	Percentage of missing value
Body_mass_index	Body_mass_index	31	4.1891892
Absenteeism_time_in_hours	Absenteeism_time_in_hours	22	2.9729730
Height	Height	14	1.8918919
Work_load_Average_per_day	Work_load_Average_per_day	10	1.3513514
Education	Education	10	1.3513514
Transportation_expense	Transportation_expense	7	0.9459459
Hit_target	Hit_target	6	0.8108108
Disciplinary_failure	Disciplinary_failure	6	0.8108108
Son	Son	6	0.8108108
Social_smoker	Social_smoker	4	0.5405405
Reason_for_absence	Reason_for_absence	3	0.4054054
Distance_from_Residence_to_Work	Distance_from_Res	3	0.4054054
Service_time	Service_time	3	0.4054054
Age	Age	3	0.4054054
Social_drinker	Social_drinker	3	0.4054054
Pet	Pet	2	0.2702703
Month_of_absence	Month_of_absence	1	0.1351351
Weight	Weight	1	0.1351351
ID	ID	0	0.0000000
Day_of_the_week	Day_of_the_week	0	0.0000000
Seasons	Seasons	0	0.0000000

There are total of 135 missing values in the data

Missing values in each columns:



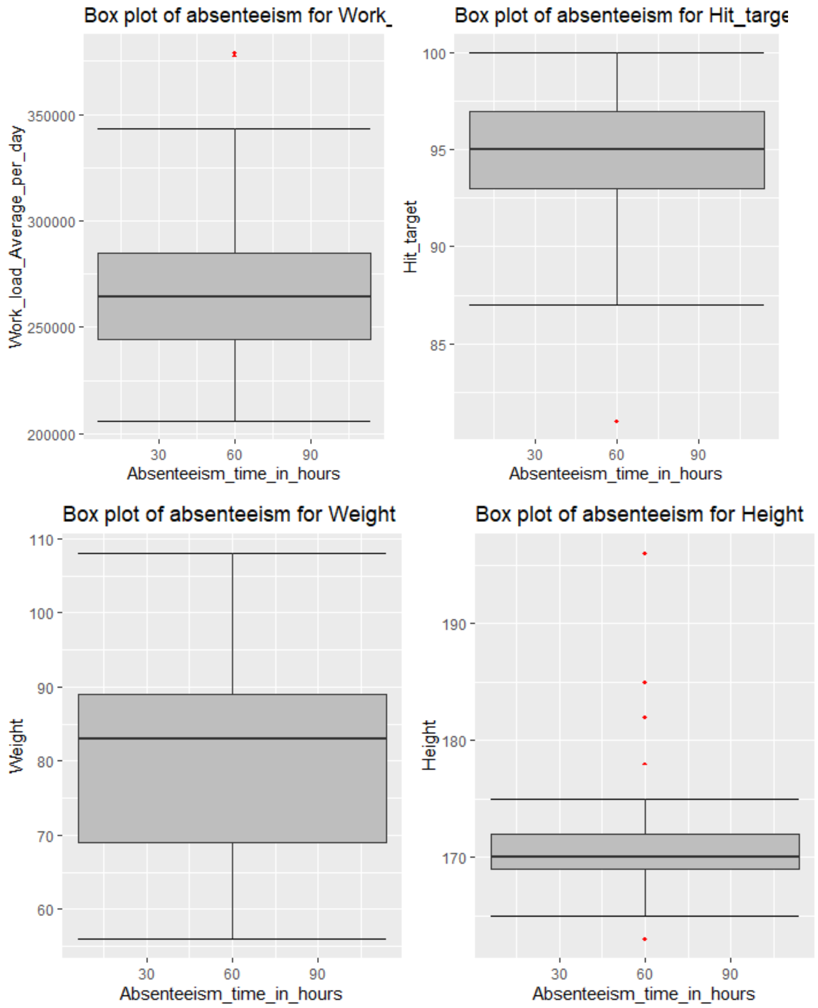
0.9% of the dataset are missing values:

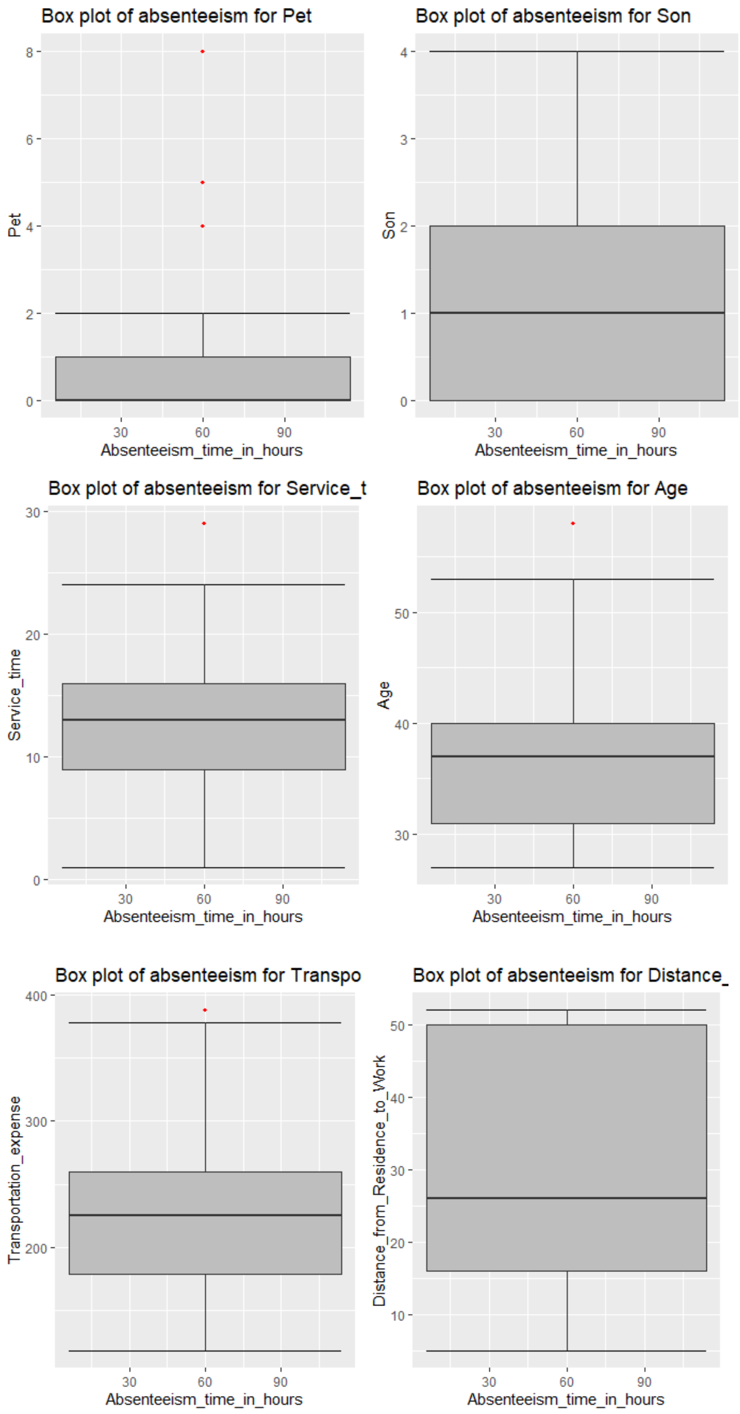


2.1.2 Outlier Analysis

We can clearly observe from these probability distributions that most of the variables are skewed. The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data. One of the other steps of pre-processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.

In figure we have plotted the boxplots of the 11 predictor variables with respect to **Absenteeism time in hour**. A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set.



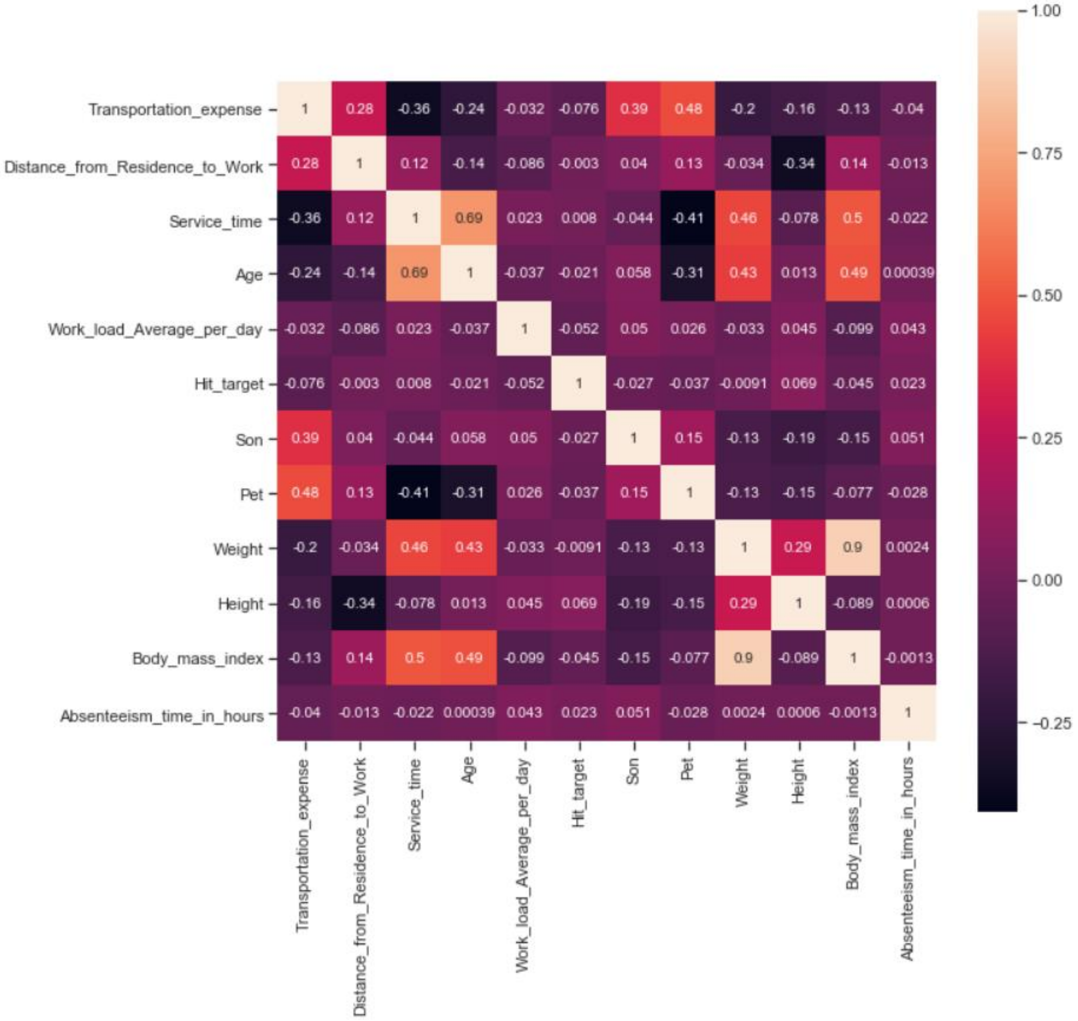


From the boxplot almost all the variables **except "Distance from residence to work", "Weight" and "Body mass index"** consists of outliers. We have converted the outliers (data beyond minimum with minimum value and beyond maximum with maximum value.

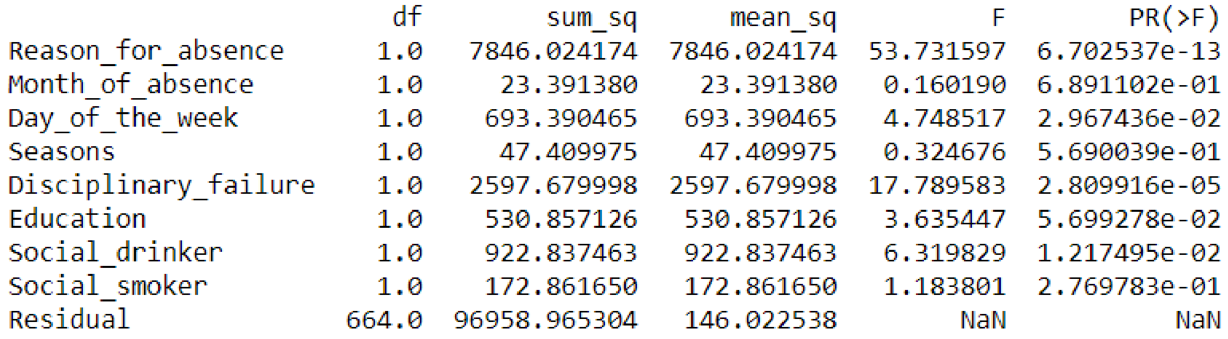
2.1.3 Feature Selection

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi collinearity. In this project we have selected **Correlation Analysis** for numerical variable and **ANOVA** (Analysis of variance) for categorical variable.

Correlation Heatmap for Continuous variable



Anova table for categorical variable



```
                          df        sum_sq       mean_sq          F        PR(>F)
Reason_for_absence       1.0   7846.024174   7846.024174  53.731597  6.702537e-13
Month_of_absence         1.0     23.391380     23.391380   0.160190  6.891102e-01
Day_of_the_week          1.0    693.390465    693.390465   4.748517  2.967436e-02
Seasons                  1.0     47.409975     47.409975   0.324676  5.690039e-01
Disciplinary_failure     1.0   2597.679998   2597.679998  17.789583  2.809916e-05
Education                1.0    530.857126    530.857126   3.635447  5.699278e-02
Social_drinker           1.0    922.837463    922.837463   6.319829  1.217495e-02
Social_smoker            1.0    172.861650    172.861650   1.183801  2.769783e-01
Residual               664.0  96958.965304    146.022538        NaN           NaN
```

From correlation analysis we have found that **Weight** and **Body mass index** has high correlation , so we have excluded the **Weight** column.

2.2.4 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

Normalization Formula:

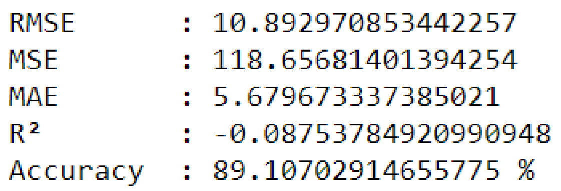
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2.2 Modeling

After a thorough preprocessing we will be using some regression models on our processed data to predict the target variable. Following are the models which we have built –

2.2.1 Decision Tree

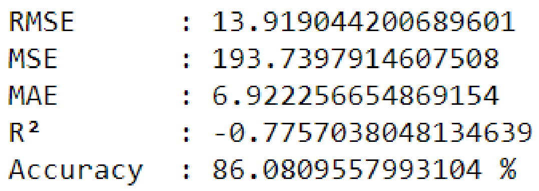
A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each branch connects nodes with "and" and multiple branches are connected by "or". It can be used for classification and regression. It is a supervised machine learning algorithm. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users. Split of decision tree is seen in the below tree. The RMSE value and R^2 value for our project in R and Python are –



```
RMSE       : 10.892970853442257
MSE        : 118.65681401394254
MAE        : 5.679673337385021
R²         : -0.08753784920990948
Accuracy   : 89.10702914655775 %
```

2.2.2 Random Forest

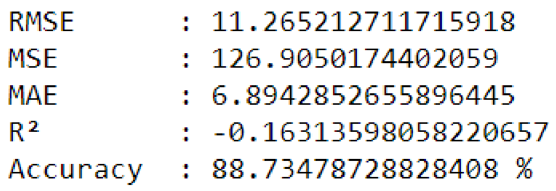
Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data. The RMSE value and R^2 value for our project in R and Python are –



```
RMSE       : 13.919044200689601
MSE        : 193.7397914607508
MAE        : 6.922256654869154
R²         : -0.7757038048134639
Accuracy   : 86.0809557993104 %
```

2.2.3 Liner Regression

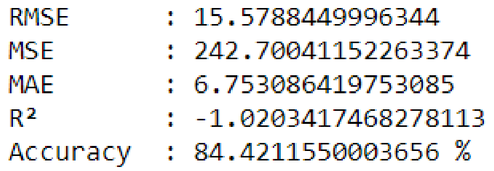
Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. To build any model we have some assumptions to put on data and model. Here are the assumptions to the linear regression model.



```
RMSE       : 11.265212711715918
MSE        : 126.9050174402059
MAE        : 6.8942852655896445
R²         : -0.16313598058220657
Accuracy   : 88.73478728828408 %
```

2.2.4 KNN Algorithm

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.



```
RMSE       : 15.5788449996344
MSE        : 242.70041152263374
MAE        : 6.753086419753085
R²         : -1.0203417468278113
Accuracy   : 84.4211550003656 %
```

Chapter 3

Conclusion

In this chapter we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

3.1 Model Evaluation

In the previous chapter we have seen the **Root Mean Square Error** (RMSE) and **R-Squared** Value of different models. **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, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Whereas **R-squared** is a relative measure of fit, **RMSE** is an absolute measure of fit. As the square root of a variance, **RMSE** can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of **RMSE** and higher value of **R-Squared Value** indicate better fit.

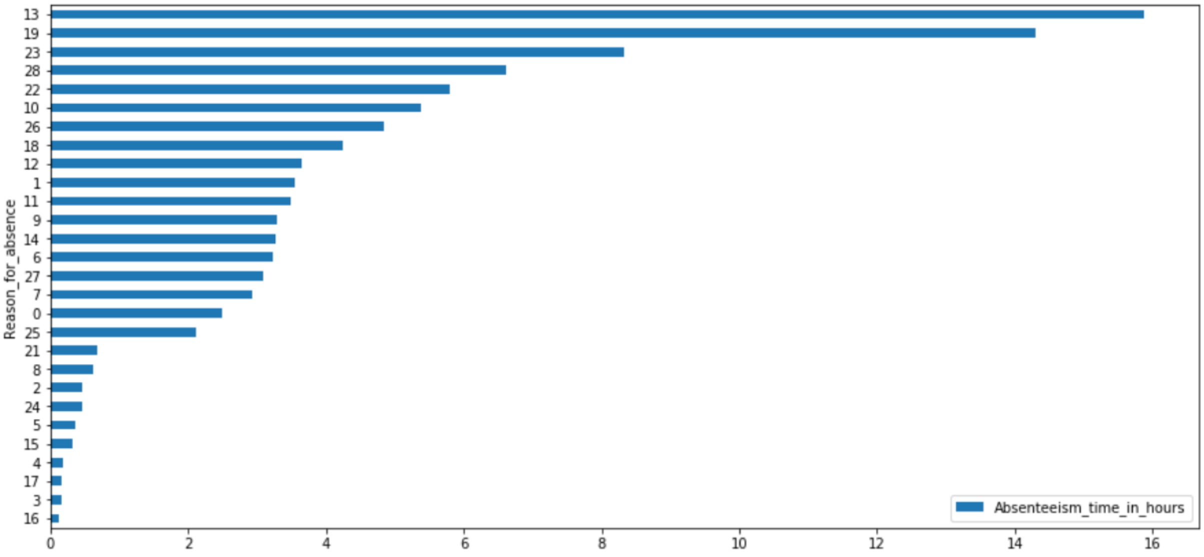
3.2 Model Selection

From the observation of all **RMSE Value** and **R-Squared** Value we have concluded that **Decision Tree** has minimum value of RMSE.

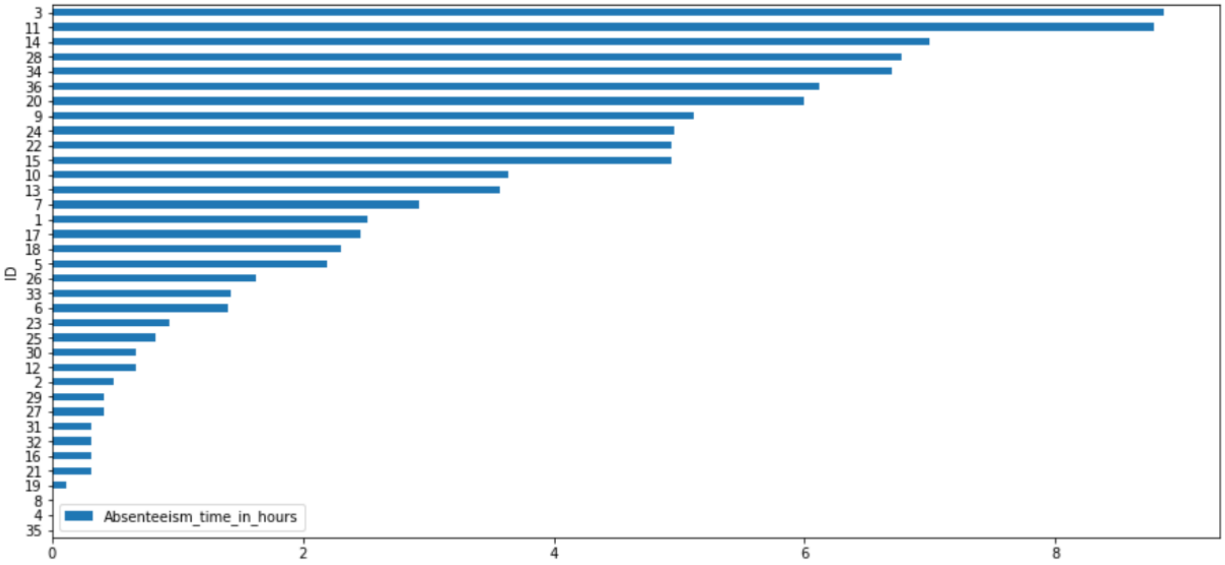
Answer to the questions

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

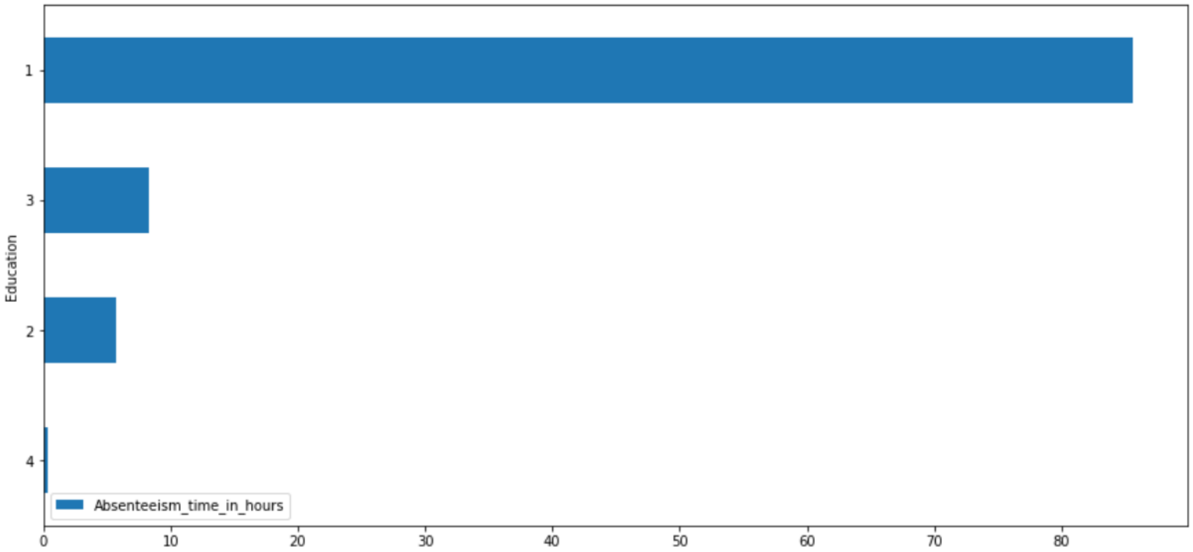
- The top 3 reasons employees are absent are due to musculoskeletal system and connective tissue, Physiotherapy and Disease of circulatory system. Therefore company need to take care of these employees



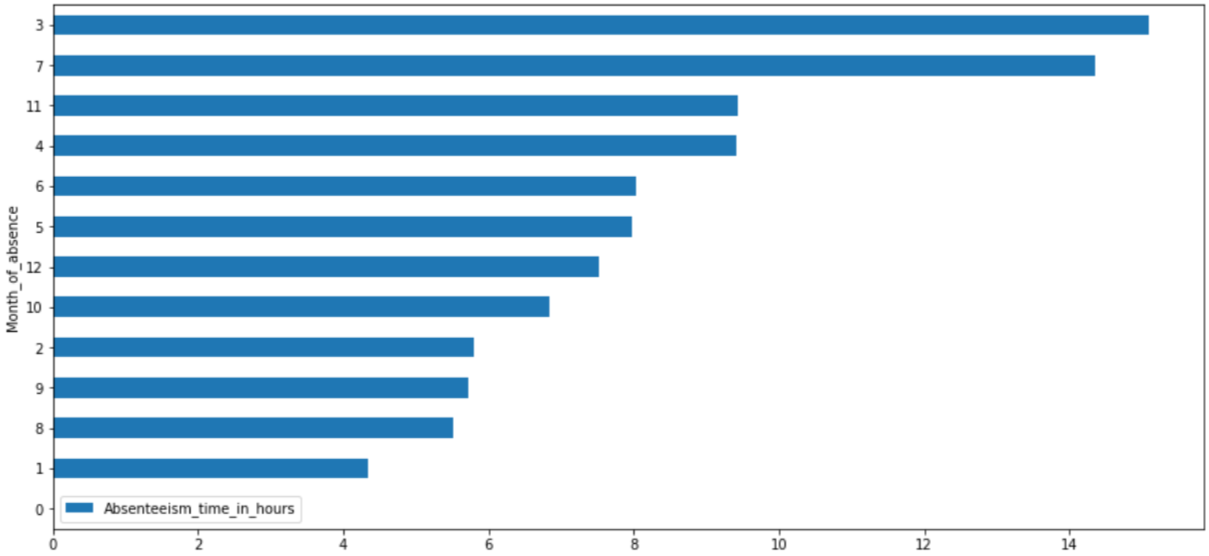
- Employee with ID 3 and 11 have the highest absent hours



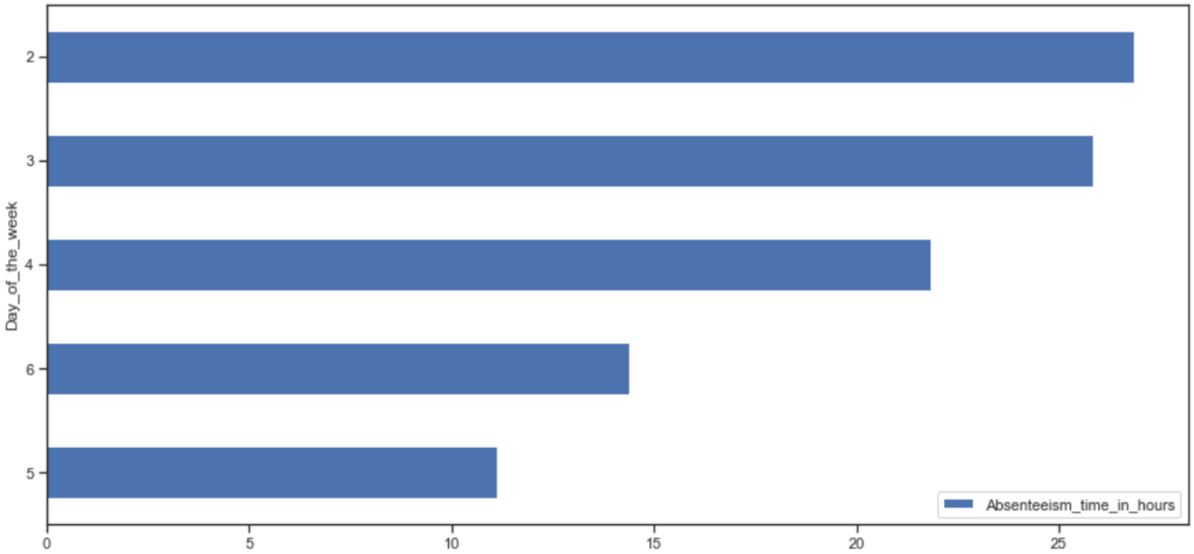
- Employees who studied only high school have most absent hours, Maybe because there are more employees in the dataset who only studied high school



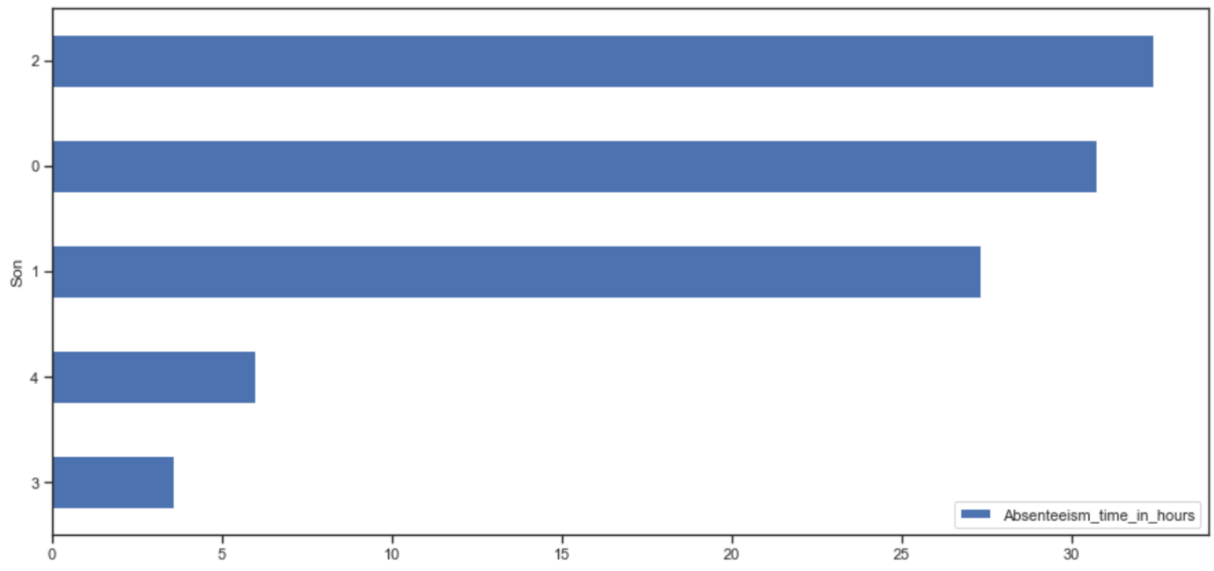
- From the plot we can tell that most of the employees are being absent in the month March



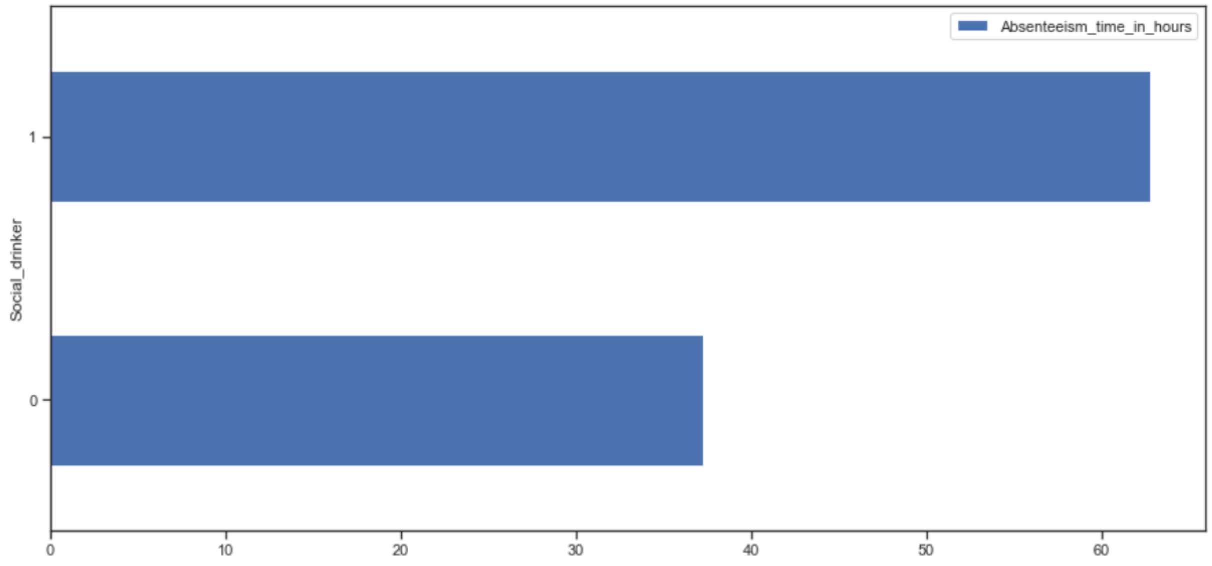
- Day with highest absent hours is Tuesday and Wednesday



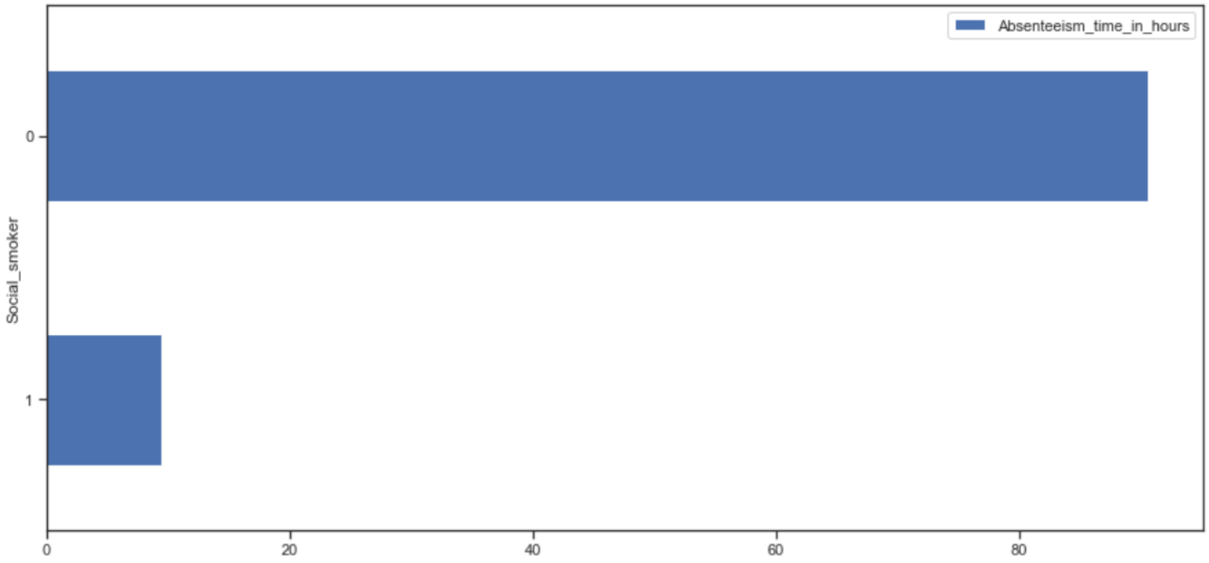
- Employees who has zero, one, or two kids have more absent hours



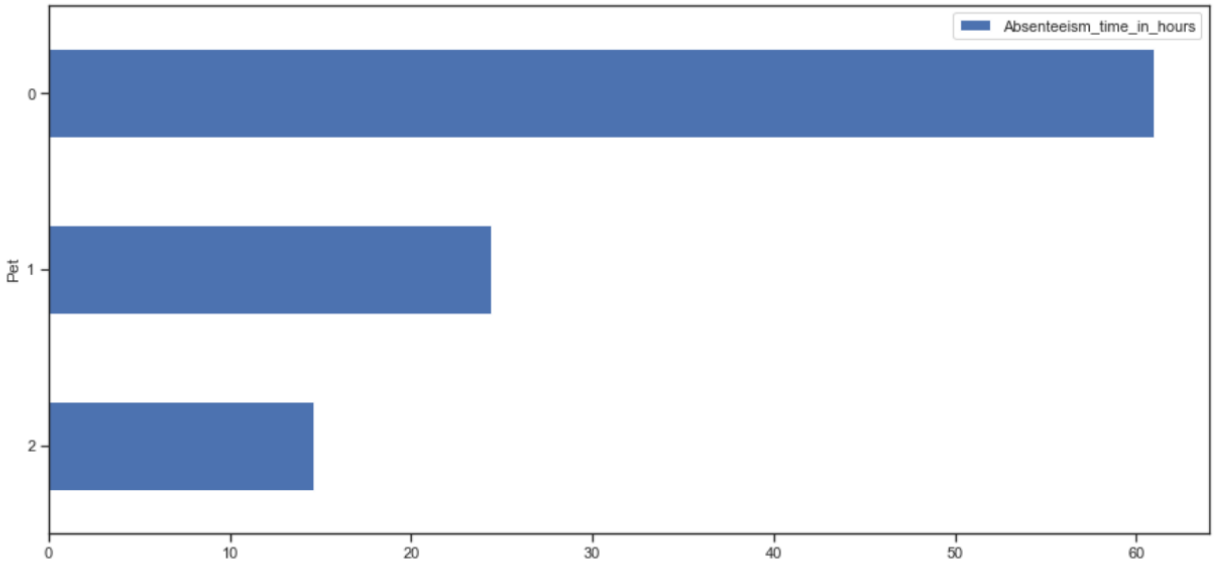
- Employees who drinks have more absent hours



- Employees who don’t smoke have more absence hours



- Employees with zero pets have more absent hours



Conclusion for first question:

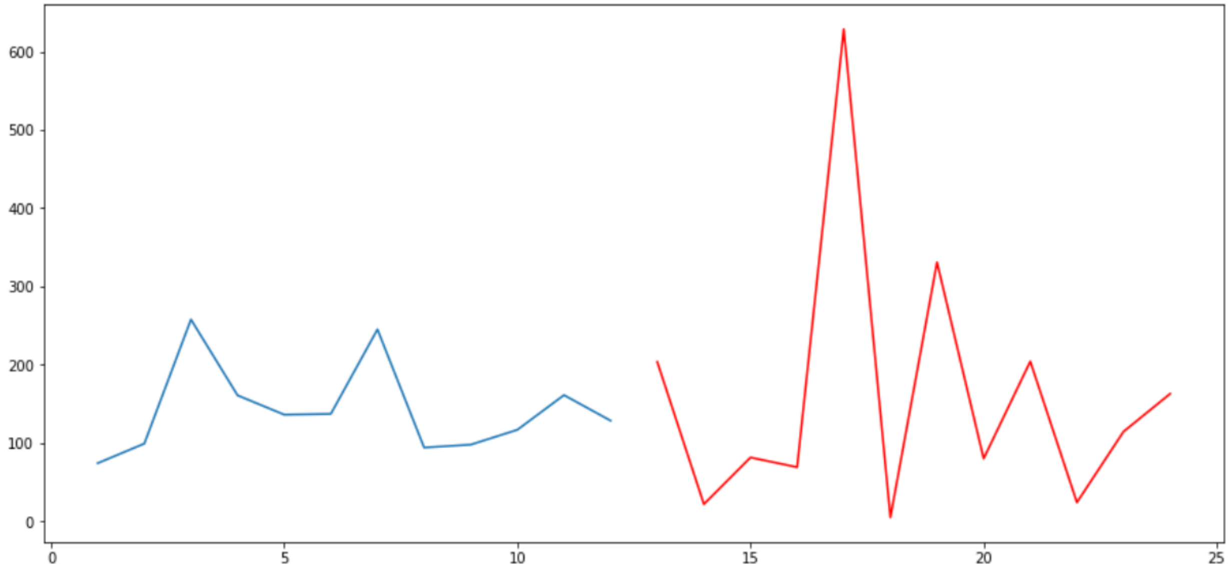
Some employees on the company have a series health condition which should be taken careoff. Company management should focus on the employee with no child, as there might be chance of high health risk. Moreover, employee ID 3 and 11 should be taken into consideration as these employees have taken the maximum absentee and also having some serious health issue which leads to the high absenteeism.

"Diseases of the musculoskeletal system and connective tissue" and "Injury, poisoning and certain other consequences of external causes" are not a common disease, there might be an environmental issue. There might a chance that due to some environmental factor, maximum employees are affecting due to this disease. These are the serious issue which should be taken into consideration as this can also affect other employees of an organisation. Around 60% of the workforce is affected by this disease.

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

To forecast Absenteeism I use the most popular model for Time Series – ARIMA which stands for Auto Regressive Integrated Moving Average.

The plot shows the monthly forecasting, If the trend of absenteeism continued then company will face an average of 160.36234 absence hours in every month in 2011


600
500
400
300
200
100
0
0
5
10
15
20
25