Employee Absenteeism
Case Study
Asheesh Aggarwal

Contents

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

Sample Dataset-

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work
11	26.0	7.0	3	1	289.0	36.0
36	0.0	7.0	3	1	118.0	13.0
3	23.0	7.0	4	1	179.0	51.0
7	7.0	7.0	5	1	279.0	5.0
11	23.0	7.0	5	1	289.0	36.0
3	23.0	7.0	6	1	179.0	51.0

Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Education	Son
13.0	33.0	239554.0	97.0	0.0	1.0	2.0
18.0	50.0	239554.0	97.0	1.0	1.0	1.0
18.0	38.0	239554.0	97.0	0.0	1.0	0.0
14.0	39.0	239554.0	97.0	0.0	1.0	2.0
13.0	33.0	239554.0	97.0	0.0	1.0	2.0
18.0	38.0	239554.0	97.0	0.0	1.0	0.0

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

Dataset has 21 variables in which 20 variables are independent and 1 (Absenteeism time in hours) is dependent variable. Since target variable is continuous in nature, this is a regression problem.

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),

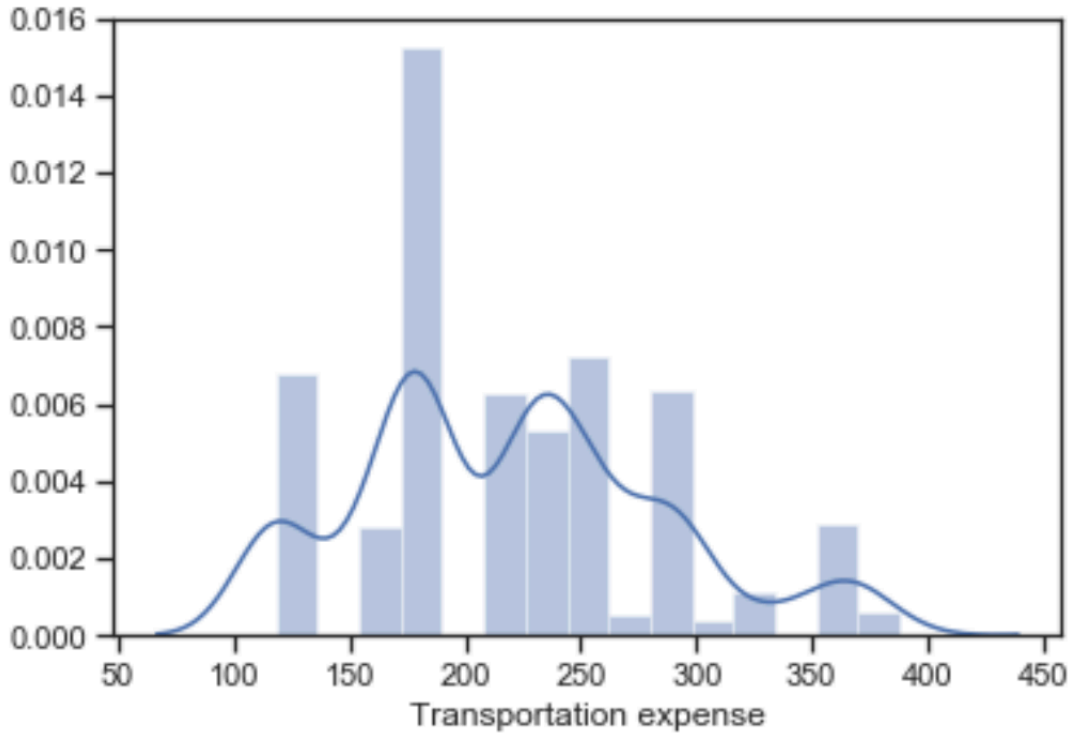
dental consultation (28).
3. Month of absence
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
5. Seasons (summer (1), autumn (2), winter (3), spring (4))
6. Transportation expense
7. Distance from Residence to Work (kilo meters)
8. Service time
9. Age
10. Work load 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)
4
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)

Methodology

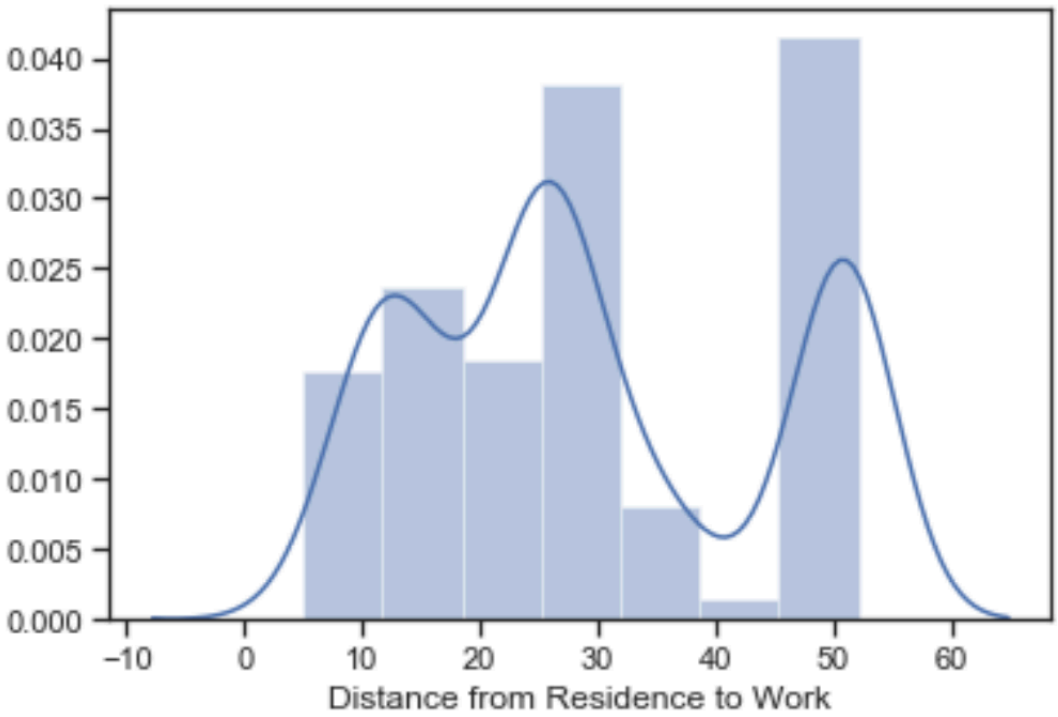
2.1 Pre Processing

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.

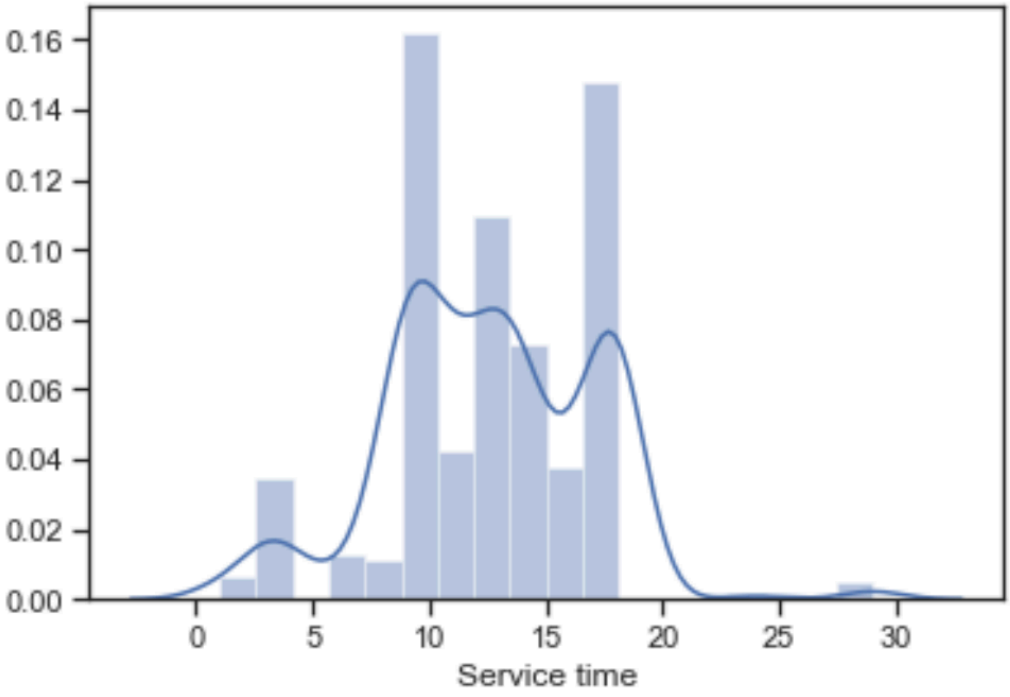
1. Transportation expense



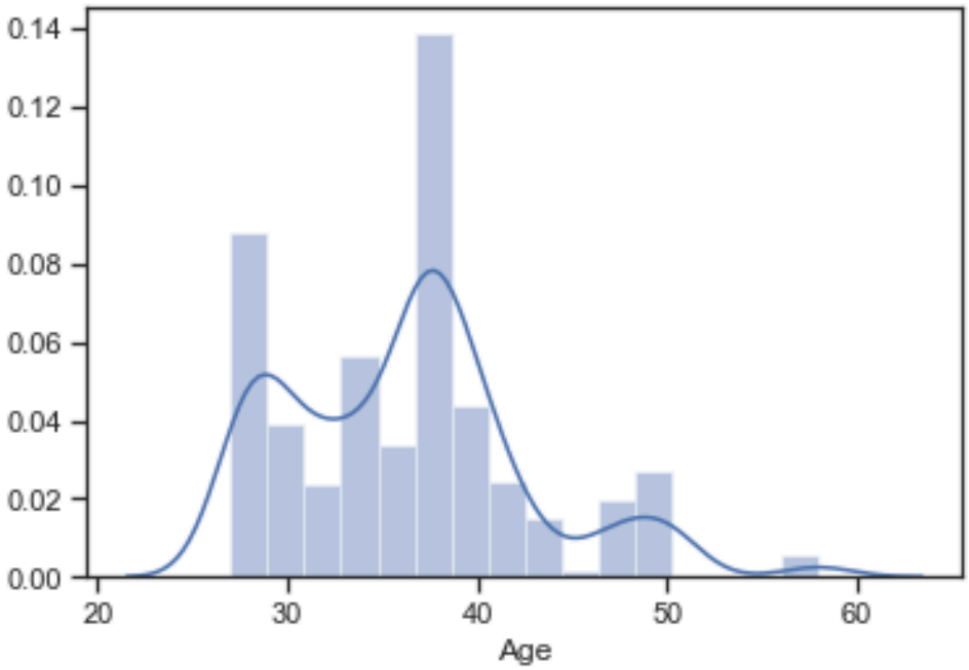
2. Distance from Residence to Work



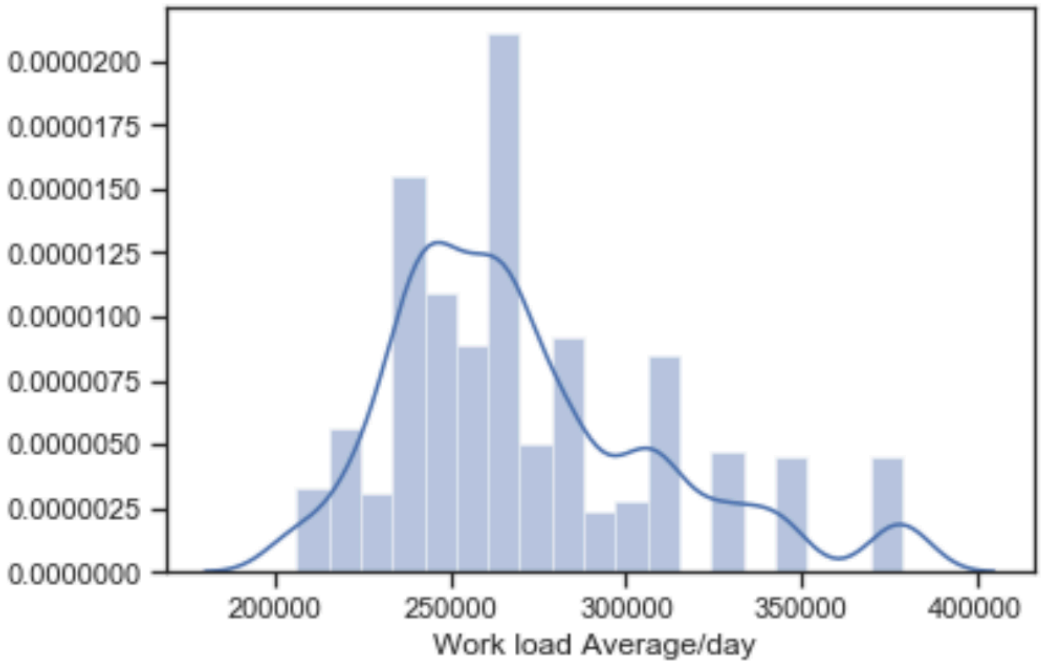
3. Service time



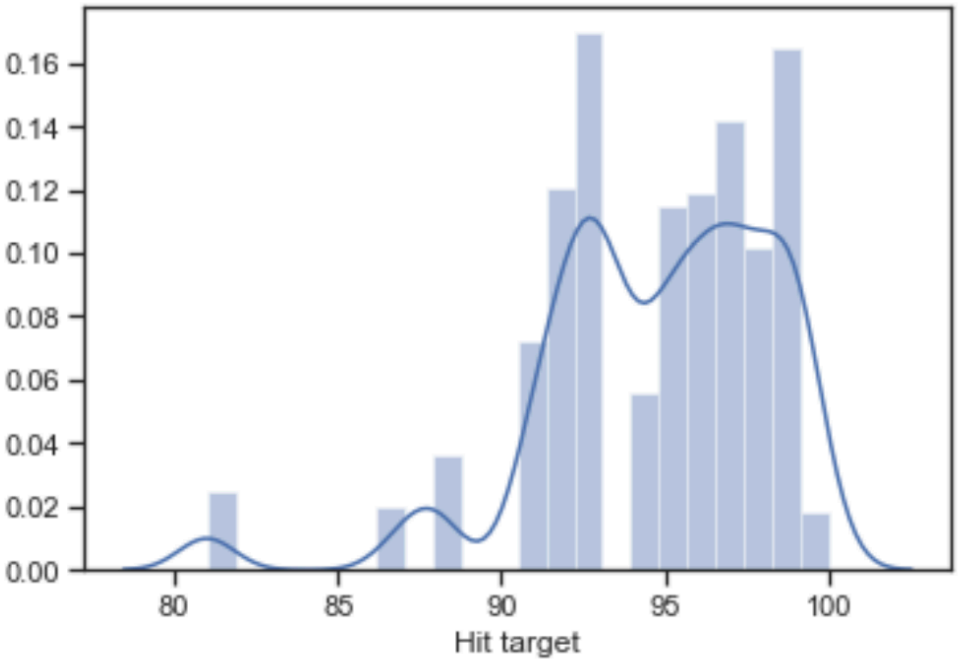
4. Age



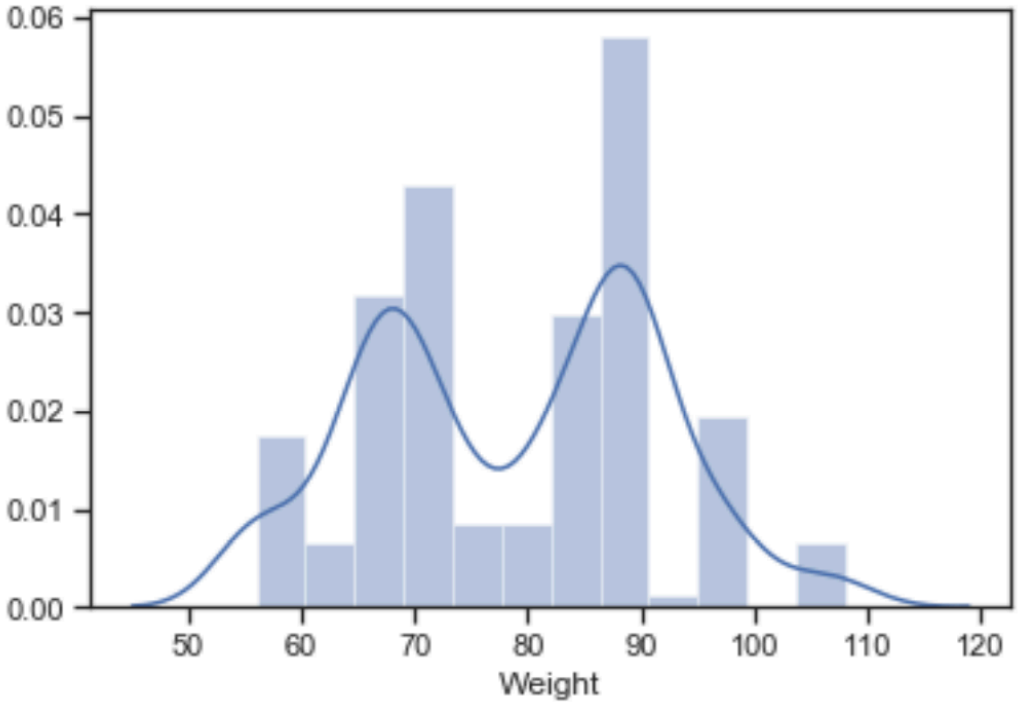
5. Work load Average/day



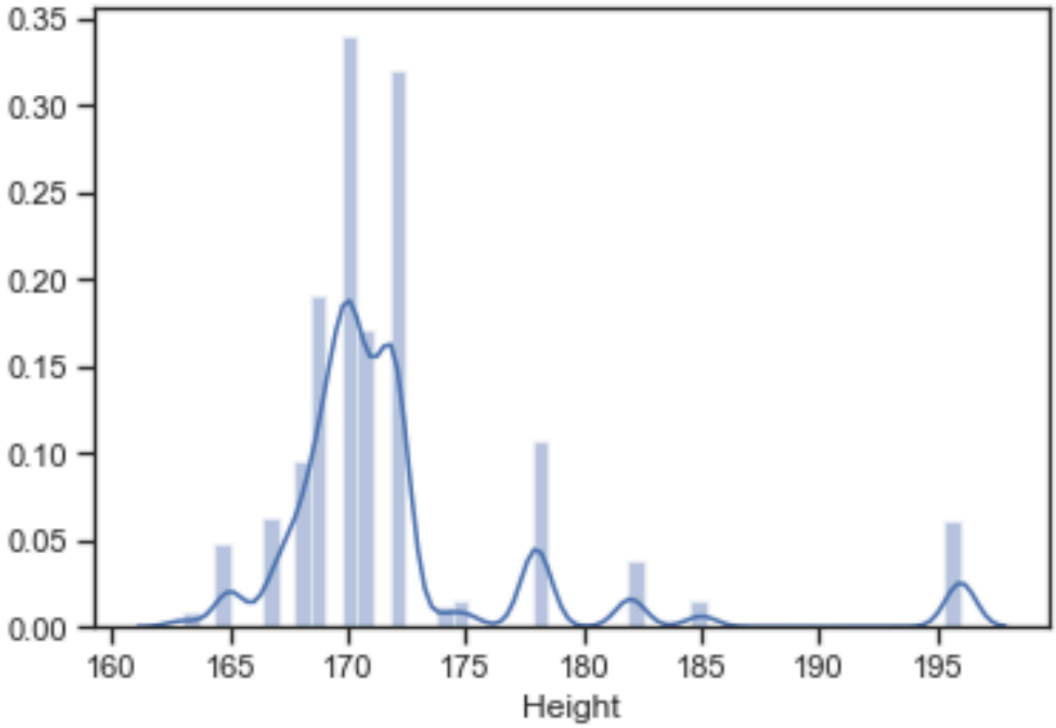
6. Hit target



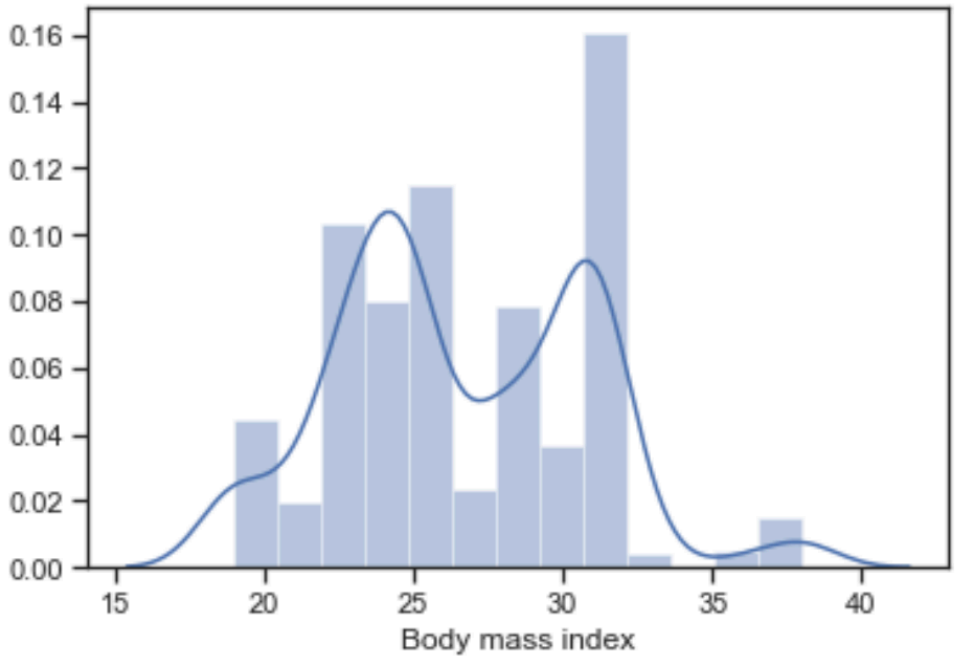
7. Weight



8. Height



9. Body mass index



Looking at the plots we can conclude that all continuous variables have skewed distributions, hence when we come to the feature scaling part we can go ahead with the normalization technique.

2.1.1 Missing Value Analysis

Missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

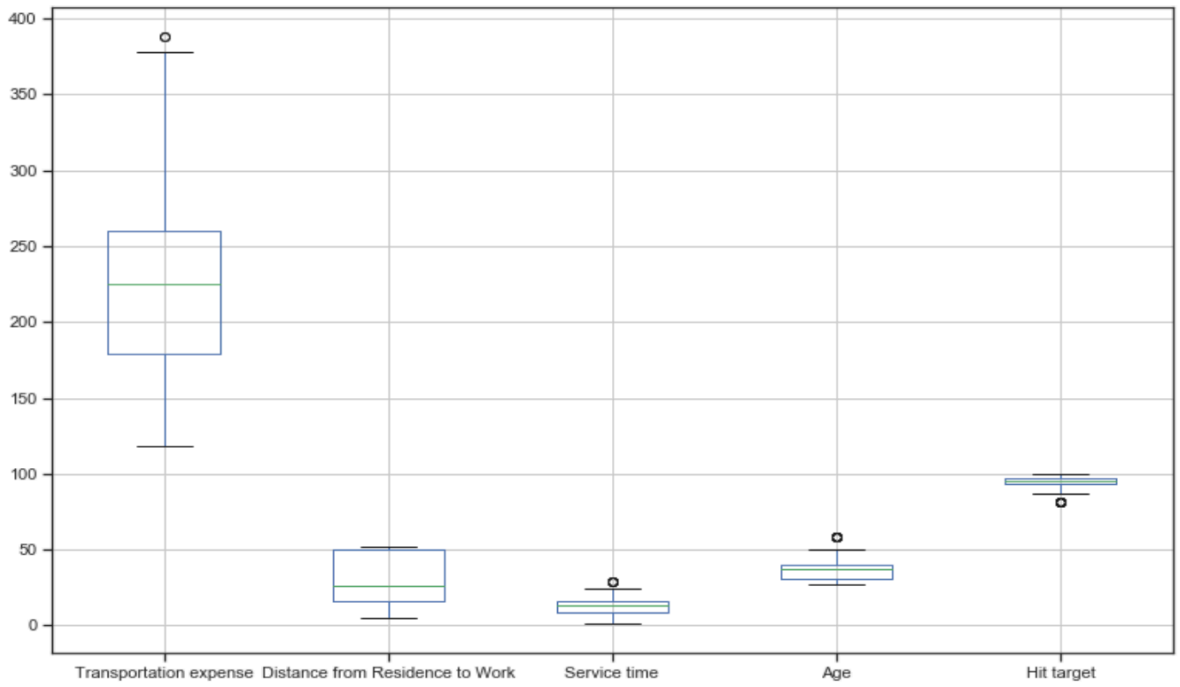
Looking at the data we drew the below conclusions which were used to impute missing values.

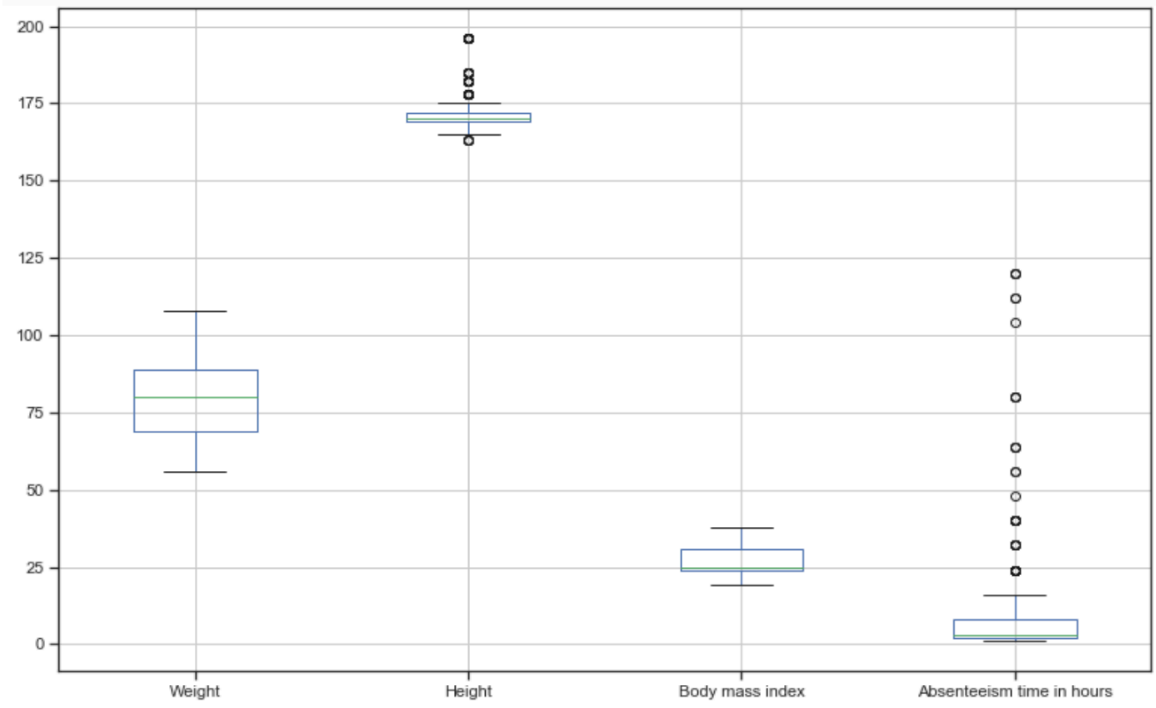
- The variable ID can be used to uniquely identify the following variables:
    - Transportation expense
    - Distance from Residence to Work
    - Service time
    - Age
    - Education
    - Son
    - Social drinker
    - Social smoker
    - Pet
    - Weight
    - Height
    - Body mass index
- The variable Work load Average/day can be uniquely identified using a combination of Month of absence and the Hit target variables.
- The variable Hit target can be uniquely identified using a combination of Month of absence and the Work load Average/day variables.
- The variable Month of absence can be uniquely identified using a Work load Average/day variable.
- For the variable Reason for absence, the median value for rows with the same value for Absenteeism time in hours was used to impute as it was more accurate than KNN and mean values.
- Disciplinary failure value is 1 for the rows where the reason for absence is 0, otherwise it is 0. The same rule was used for the imputation.

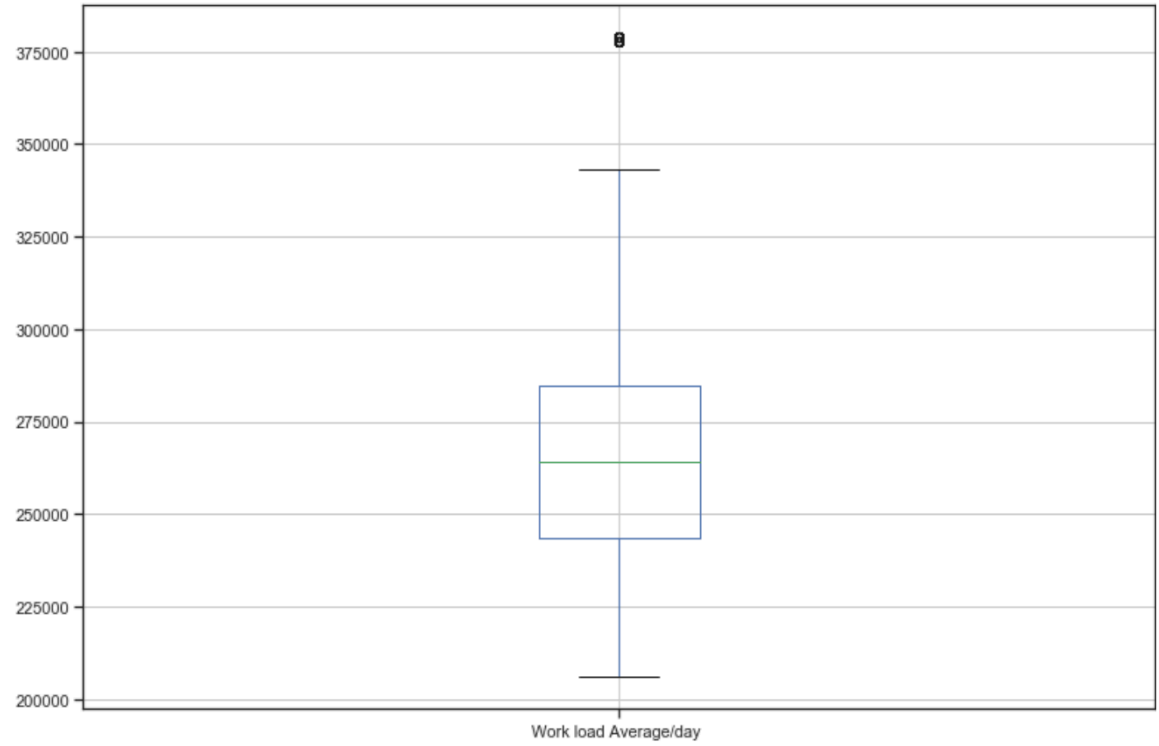
- For the variable Absenteeism time in hours, the median value for rows with the same value for variable Reason for absence was used to impute as it was more accurate than KNN and mean values.

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. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.
In below figures we have plotted the boxplots of the predictor variables with respect to target variable Absenteeism time in hour, and detect the outliers by visualization.





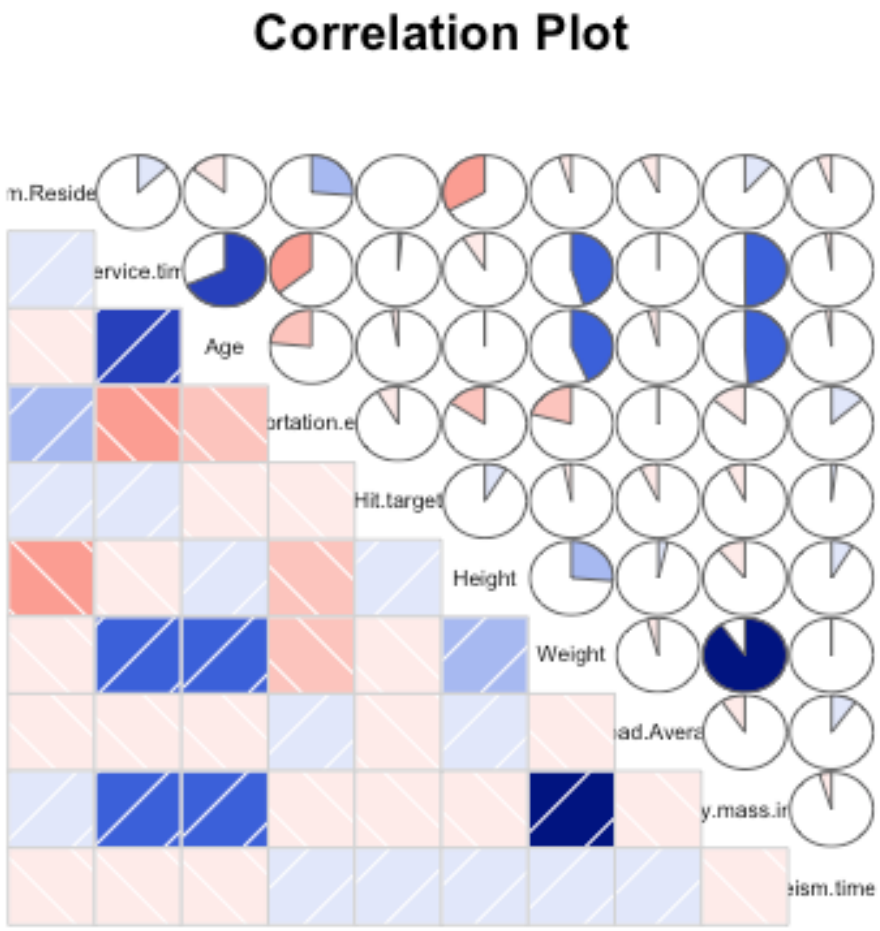


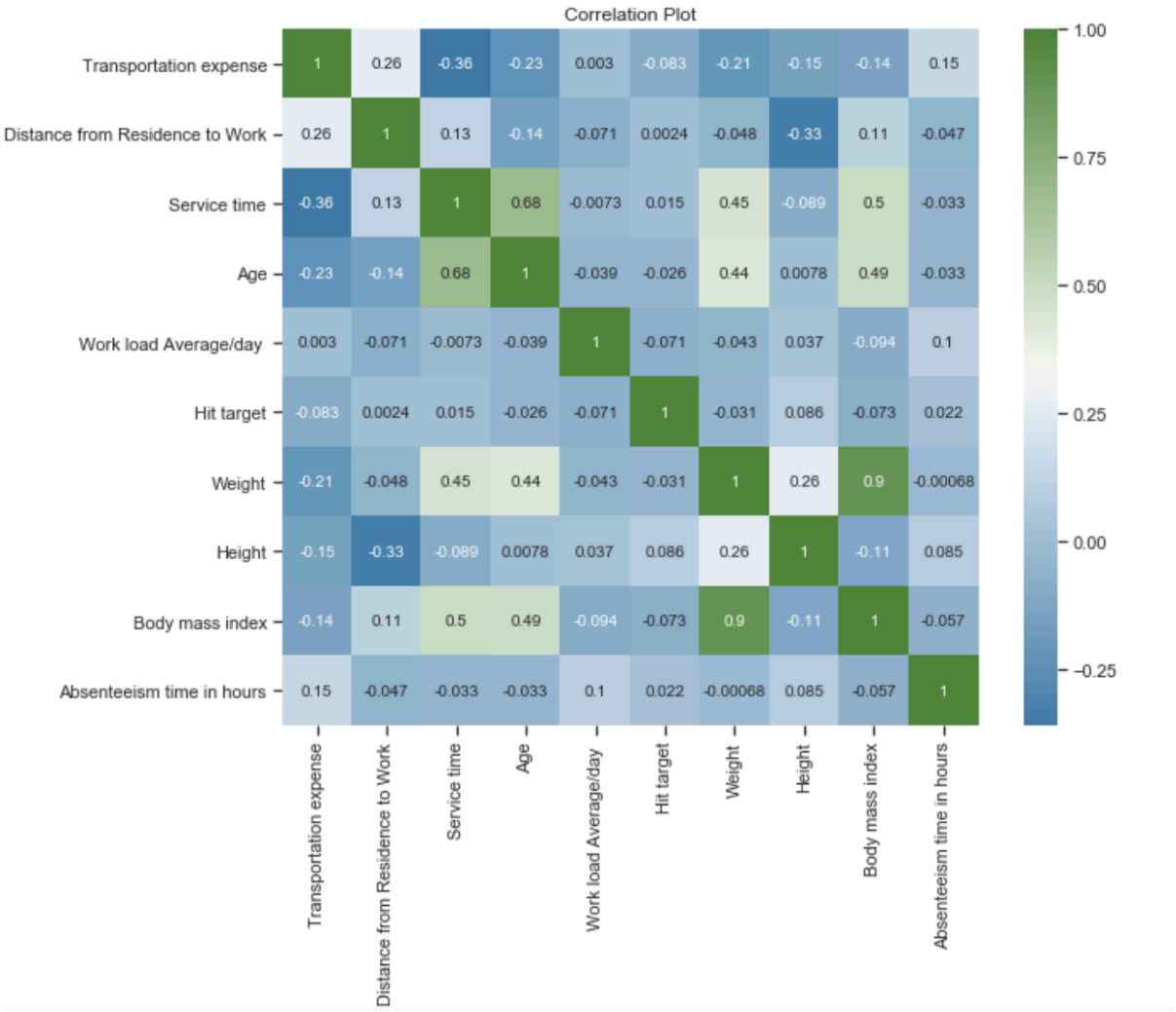
Outliers were capped to the maximum and the minimum values as per the boxplots i.e. values above upper fence set to maximum and values below the lower fence set to minimum.

2.1.3 Feature Selection

We cannot use all the features becausesome 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 chi square test for categorical variables.

Correlation Analysis plot for continuous variables-





P values of categorical variables from chi square test :

ID - 2.2365273988343456e-18
Reason for absence - 7.286648099847831e-164
Month of absence - 4.2684498955085155e-14
Day of the week - 0.022796924191761825
Seasons - 5.362095696190738e-10
Disciplinary failure - 3.9490277004803864e-134
Education - 0.31031968497259027
Son - 1.374139582757932e-10
Social drinker - 0.0002987130412010769
Social smoker - 0.19661000773680085
Pet - 1.2450025241126808e-05

From correlation analysis we have found that Weight and Body mass index has high correlation (>0.7), so we have excluded the Weight column, and from chi

square analysis we have found that in categorical variables Social smoker, Education have the p value(>0.05), so we excluded them.

Remaining Values :
Continuous values -> ['Transportation expense', 'Distance from Residence to Work', 'Service time', 'Age','Work load Average/day ', 'Hit target','Height', 'Body mass index','Absenteeism time in hours']
categorical values -> ['ID', 'Reason for absence','Month of absence', 'Day of the week','Seasons','Disciplinary failure', 'Son', 'Social drinker','Pet',]

2.1.4 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.

Data Sample after preprocessing :

	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	Son	Social drinker	Pet	Height	Body mass index	Absenteeism time in hours
0	11	26.0	7.0	3	1	0.648956	0.659574	0.470588	0.226415	0.206305	0.769231	0.0	2.0	1.0	1.0	0.625000	0.578947	0.235294
1	36	0.0	7.0	3	1	0.000000	0.170213	0.666667	0.867925	0.206305	0.769231	1.0	1.0	1.0	0.0	1.000000	0.631579	0.000000
2	3	23.0	7.0	4	1	0.231499	0.978723	0.666667	0.415094	0.206305	0.769231	0.0	0.0	1.0	0.0	0.458333	0.631579	0.117647
3	7	7.0	7.0	5	1	0.611006	0.000000	0.509804	0.452830	0.206305	0.769231	0.0	2.0	1.0	0.0	0.291667	0.263158	0.235294
4	11	23.0	7.0	5	1	0.648956	0.659574	0.470588	0.226415	0.206305	0.769231	0.0	2.0	1.0	1.0	0.625000	0.578947	0.117647
5	3	23.0	7.0	6	1	0.231499	0.978723	0.666667	0.415094	0.206305	0.769231	0.0	0.0	1.0	0.0	0.458333	0.631579	0.117647
6	10	22.0	7.0	6	1	0.922201	1.000000	0.078431	0.037736	0.206305	0.769231	0.0	1.0	1.0	4.0	0.625000	0.421053	0.470588
7	20	23.0	7.0	6	1	0.538899	0.957447	0.392157	0.339623	0.206305	0.769231	0.0	4.0	1.0	0.0	0.291667	0.210526	0.235294

2.2 Model Development

After Data pre-processing the next step is to develop a model using a train or historical data which can perform to predict accurate result on test data or new data. Here we have tried with different models and will choose the model which will provide the most accurate values.

2.2.1 Decision Tree

Decision Tree is a supervised machine learning algorithm, which is used to predict the data for classification and regression. It accepts both continuous and categorical variables. 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”.
We have prepared a model by using decision tree algorithm and calculate RMSE value and R^2 value for our project in R and Python are -

Decision Tree	R	PYTHON
RMSE Test	0.181170	0.25044771158101437
R^2 Test	0.2446165	0.05693086

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. It means to build each decision tree on random forest we are not going to use the same data. The higher no of trees in the random forest will give higher no of accuracy, so in random forest we can go for multiple trees. It can handle large no of independent variables without variable deletion and it will give the estimates that what variables are important. The RMSE value and R^2 value for our project in R and Python are -

Decision Tree	R	PYTHON
RMSE Test	0.1913823	0.196181310561
R^2 Test	0.4537444	0.3514735259

2.2.3 Liner Regression

Linear Regression is one of the statistical method of prediction. It is most common predictive analysis algorithm. It uses only for regression, means if the target variable is continuous than we can use linear regression machine learning algorithm. The RMSE value and R^2 value for our project in R and Python are -

Decision Tree	R	PYTHON
RMSE Test	0.241316	0.2025154298
R^2 Test	0.1394887	0.3089194302

3. Conclusion

In methodology we have done data cleaning and then applied different-different machine learning algorithms on the data set to check the performance of each model, now in conclusion we will finalize the model of Employee Absenteeism dataset.

3.1 Model Evaluation

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). In simple words R-squared tells how much variance of dependent variable explained by the independent variable. It is a measure if goodness of fit in regression line. Value of R-squared between 0-1, where 0 means independent variable unable to explain the target variable and 1 means target variable is completely explained by the independent variable. So, Lower values of RMSE and higher value of R-Squared Value indicate better fit of model.

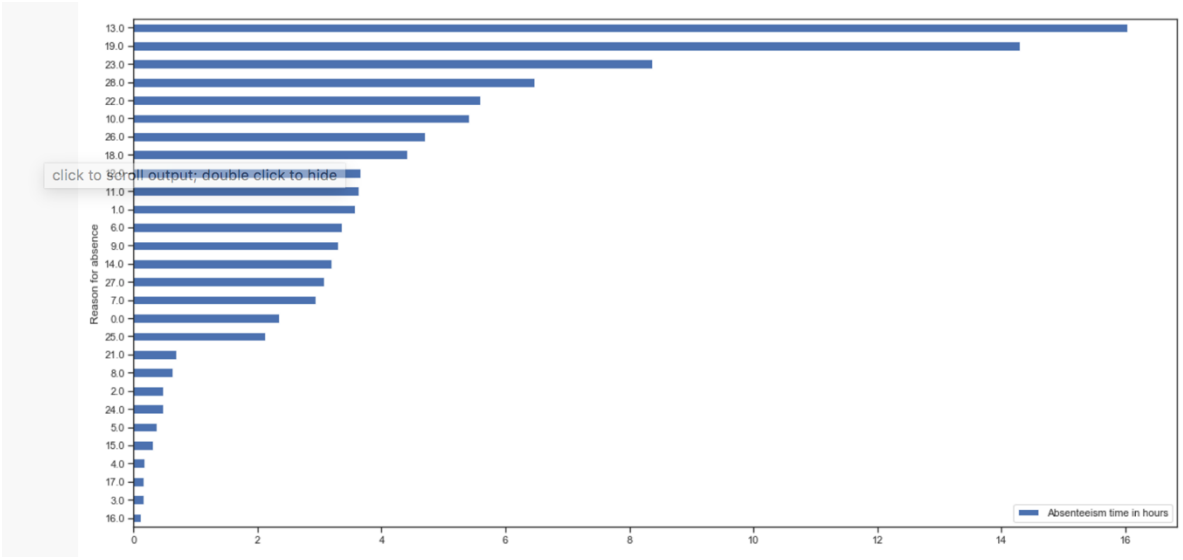
3.2 Model Selection

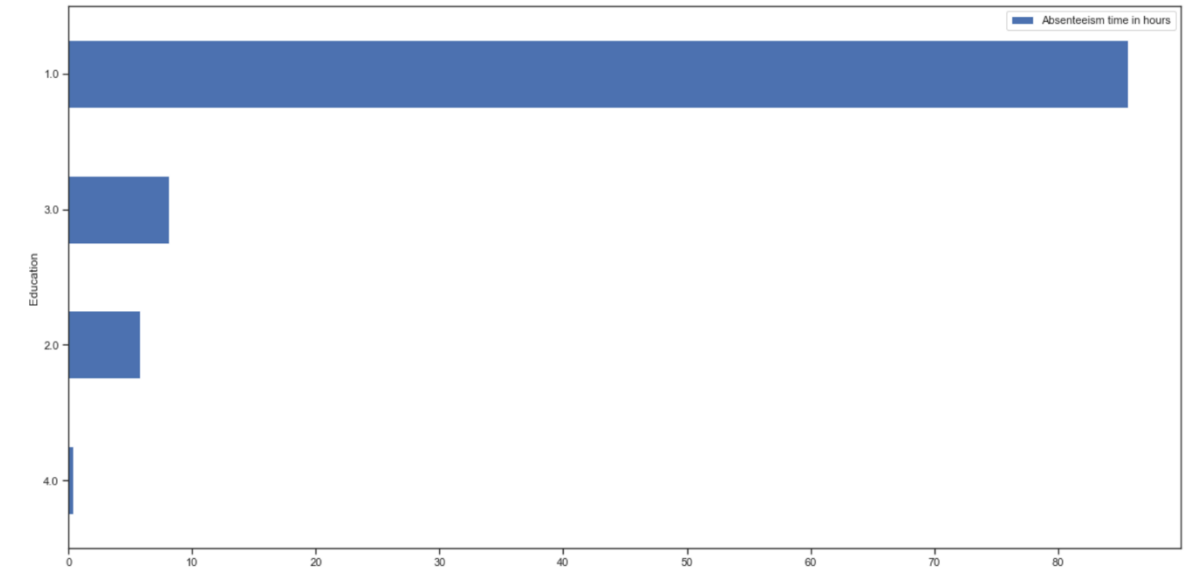
From the observation of all RMSE Value and R-Squared Value we have concluded that Random Forest has minimum value of RMSE (0.1913823) and it's R-Squared Value is also maximum (0.4537444). By Random forest algorithm predictor are explain 45% to the target variable on the test data.

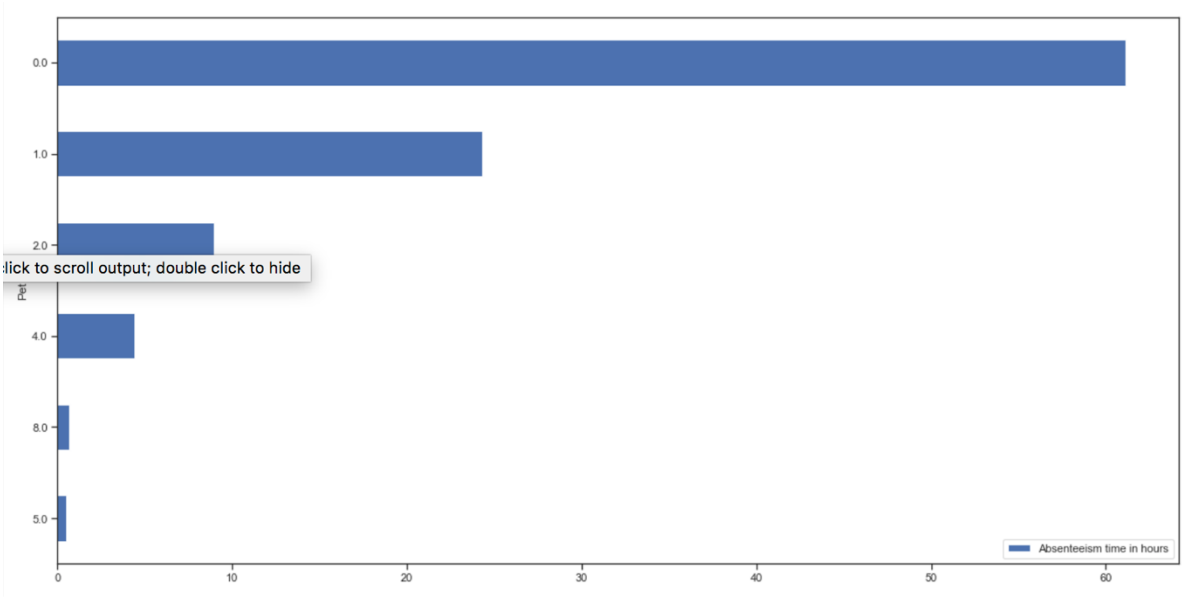
3.3 Answers of asked questions

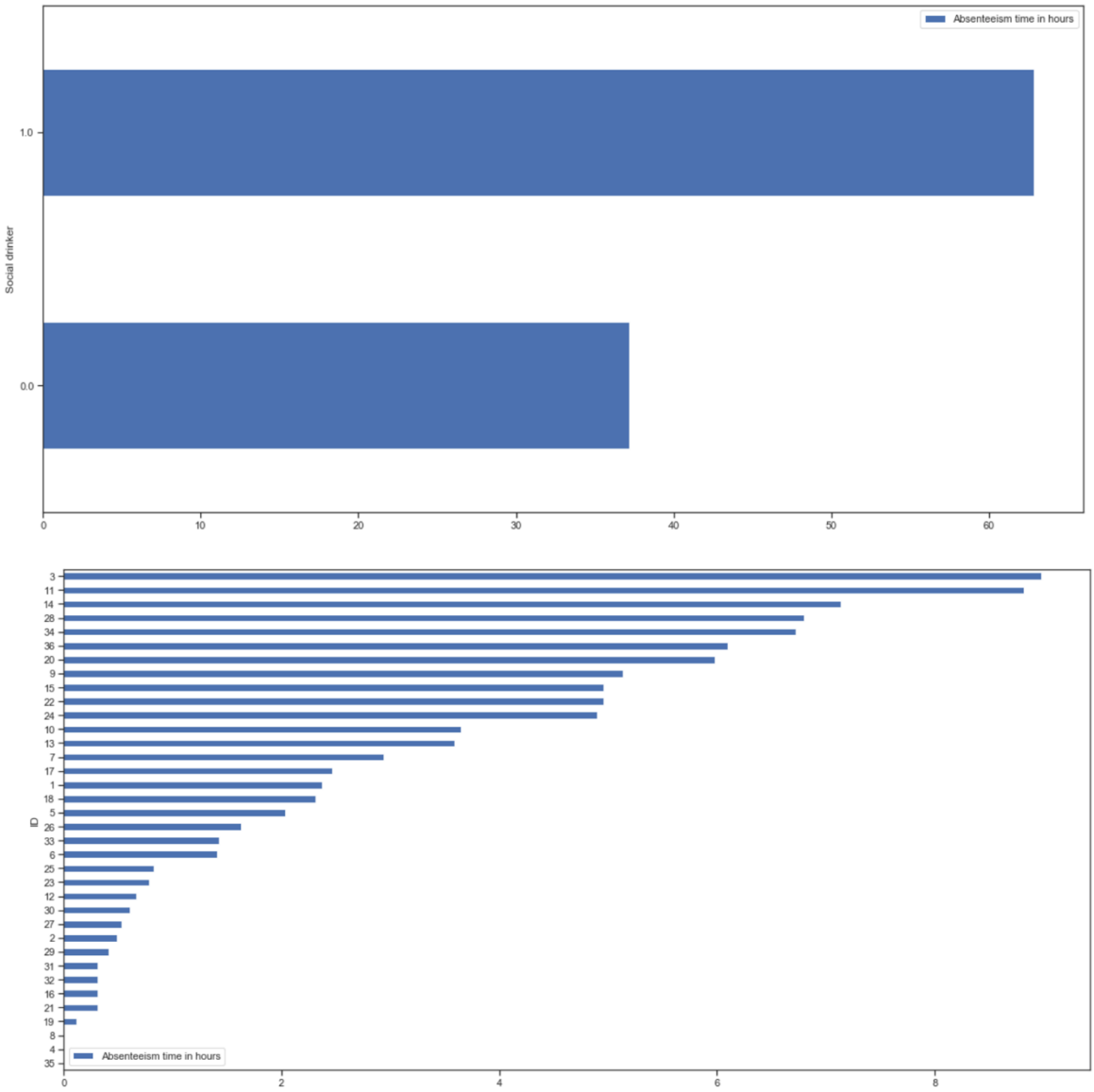
1. What changes company should bring to reduce the number of absenteeism?

For the above query, we analyzed the data properly and plotted some visualizations to find the answer to this question as below.


Reason for absence
13.0
19.0
23.0
28.0
22.0
10.0
26.0
18.0
11.0
1.0
6.0
9.0
14.0
27.0
7.0
0.0
25.0
21.0
8.0
2.0
24.0
5.0
15.0
4.0
17.0
3.0
16.0
0
2
4
6
8
10
12
14
16
Absenteeism time in hours


Education
1.0
3.0
2.0
4.0
0
10
20
30
40
50
60
70
80
Absenteeism time in hours


0.0
1.0
2.0
4.0
8.0
5.0
0
10
20
30
40
50
60
Absenteeism time in hours



- From the above plots we found that the maximum tend of absenteeism people as education wise are those who have only high school degree. 82.69 % of absenteeism time is contributed by people having high school education. The company could take care of this during hiring.
- Top 3 categories in order of Absenteeism time are:
    - Category 13 : Diseases of the musculoskeletal system and connective tissue
    - Category 23 : medical consultation
    - Category 19 : Injury, poisoning and certain other consequences of external causes

      The company can take steps to curb these.
- Most of the time of absenteeism is taken up by people with no pets.

- Social drinkers contribute to more than 64% of the absence. The company could take care of this during hiring new employees or pushing employees to curb drinking.
- The top 3 employees that contribute to absenteeism are employees with IDs - 3,11 and 14. The company can take steps to reduce this by understanding the concerns of these employees and issuing warnings if necessary

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

As we don't have the data for 2011, we shall use the given data to calculate the loss in 2011assuming the trend remains same. For this, we shall calculate the loss of work in time (hrs) which would be the average of absenteeism time in hours for each month respectively. Summing up and averaging the absenteeism time for separate years(2007,2008,2009 and 2010) shows us that there is no specific pattern for the average monthly time in absenteeism.

Hence, to best predict the pattern for 2011 we can find the average absenteeism time by month considering the data from all the years. This can be done easily using the pivot table in excel.

Once we have the total absenteeism data for all the months, we will divide the data by 3 and the data for the 7th month by 4 as we have the data for 3 years for all months except July.

Doing this we get the below values, these are the average absenteeism times for each month.

Sum of Absenteeism time in hours		
Month of absence	Total	
0	0	
1	222	74
2	294	98
3	749	249.6666667
4	482	160.6666667
5	392	130.6666667
6	403	134.3333333
7	724	181
8	272	90.66666667
9	284	94.66666667
10	340	113.3333333
11	463	154.3333333
12	382	127.3333333
(blank)	3	1
Grand Total	5010	