Employee Absenteeism

Priyanka Thakare

Contents:

Chapter 1. Introduction

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?

Data:

ID	Reason absence	Month absence	Day of week	Seasons	Transportation expense	Distance	Service time	Age	Work load
11	26	7	3	1	289	36	13	33	2,39,554
36	0	7	3	1	118	13	18	50	2,39,554
3	23	7	4	1	179	51	18	38	2,39,554
7	7	7	5	1	279	5	14	39	2,39,554
11	23	7	5	1	289	36	13	33	2,39,554

Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index
97	0	1	2	1	0	1	90	172	30
97	1	1	1	1	0	0	98	178	31
97	0	1	0	1	0	0	89	170	31
97	0	1	2	1	1	0	68	168	24
97	0	1	2	1	0	1	90	172	30

Absenteeism time in hours
4
0
2
4
2

There are total 21 variables out of which Absenteeism time in hours is our dependent variable. Rest are predictor variables.

Chapter 2 Methodology:

Data Preprocessing:

Exploratory Data Analysis:

Let's us understand the data to answer our first question. Our data has 740 rows and 21 variables. Out of which 135 are missing observations.

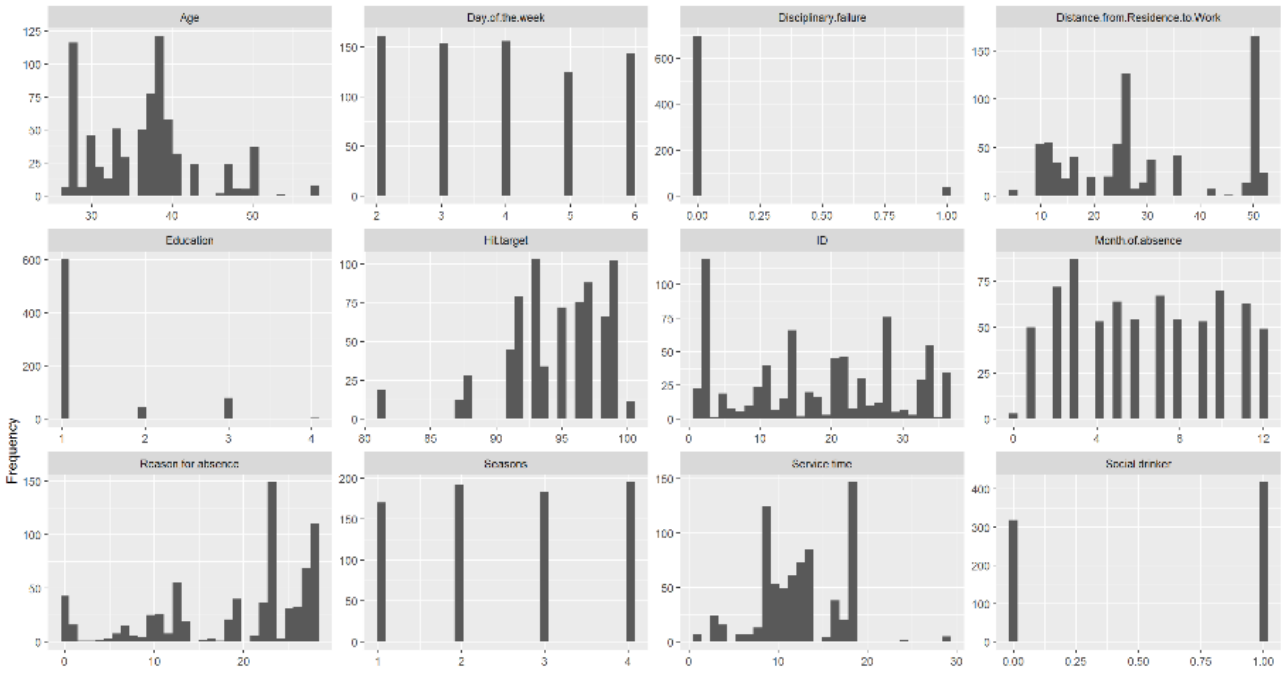
Basic Statistics

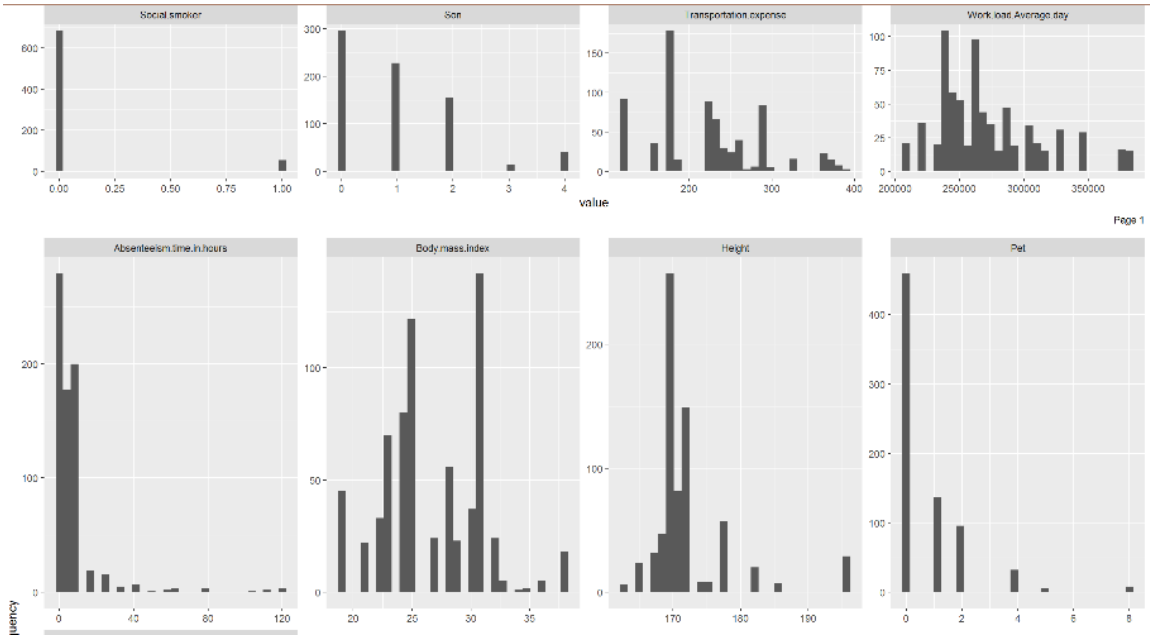
Raw Counts

Name	Value
Rows	740
Columns	21
Discrete columns	0
Continuous columns	21
All missing columns	0
Missing observations	135
Complete Rows	639
Total observations	15,540
Memory allocation	126.6 Kb

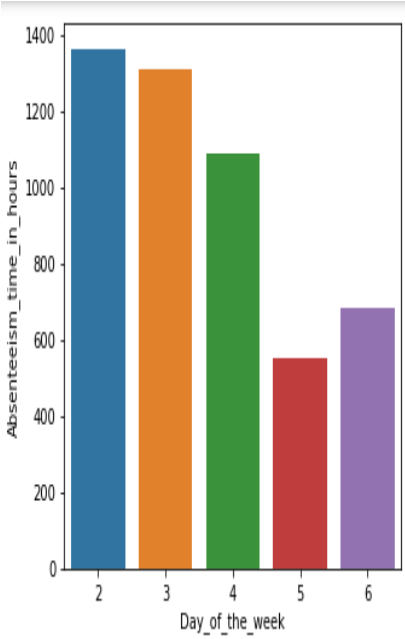
Univariate Distribution

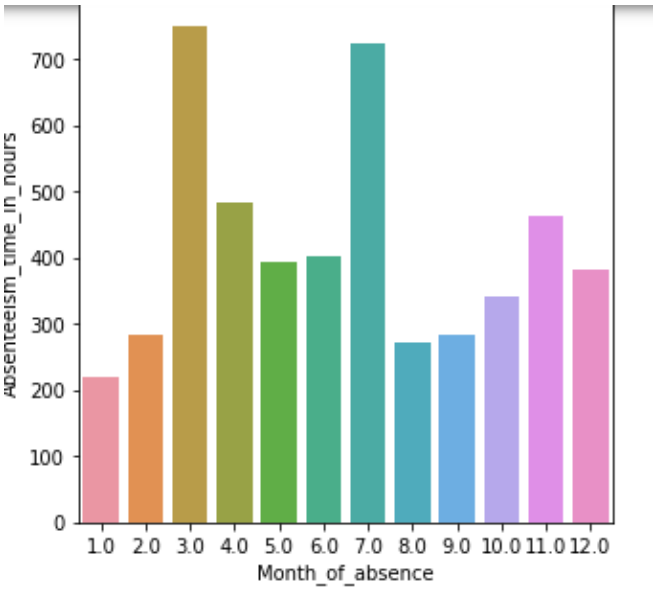
Histogram

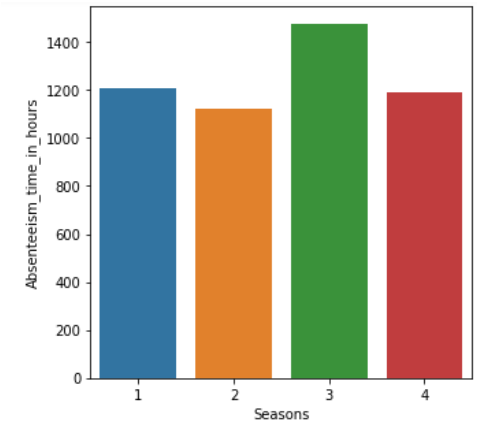


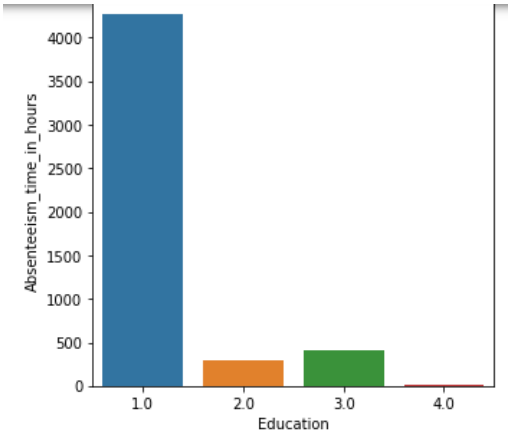


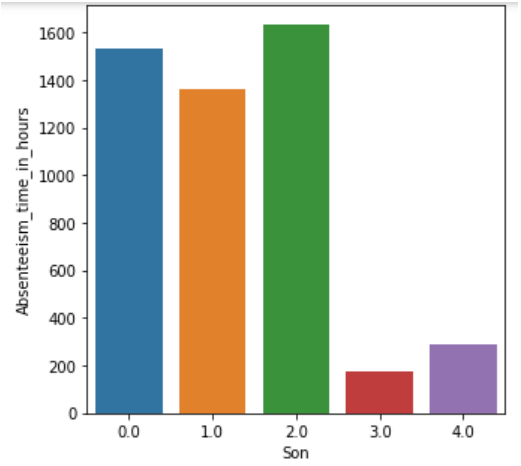
Let’s understand what is affecting the absenteeism hours.

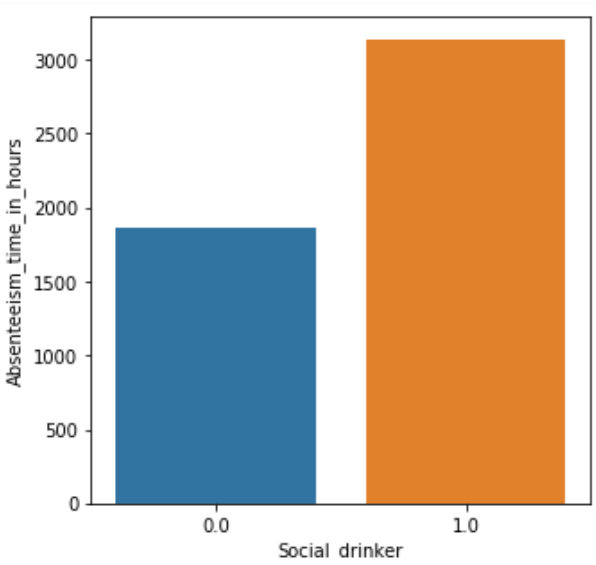


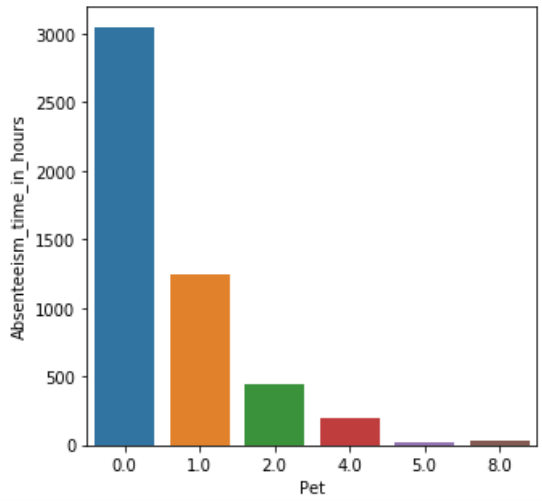


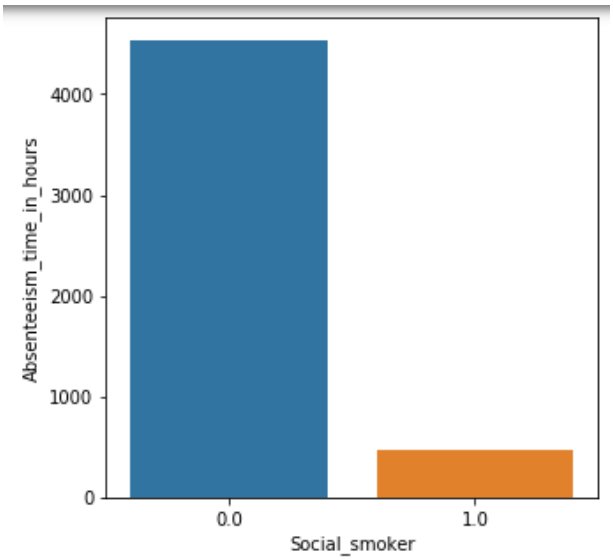












From the above graphs we can understand below thing

Absenteeism is more on

1) Mondays and Tuesday than rest of the days in the week.

2) Social drinker are frequently absent than non drinker

3) People with no pet have more absenteeism than with one or more pets

4) Months like March, April, June, Nov have more absenteeism

5) People with less no of children take more leaves than with ore children

6) During winter we can see more absenteeism which is justified because of cold and low temp might make people sick.

7) The important factor is reason of absence, from the reasons we can see

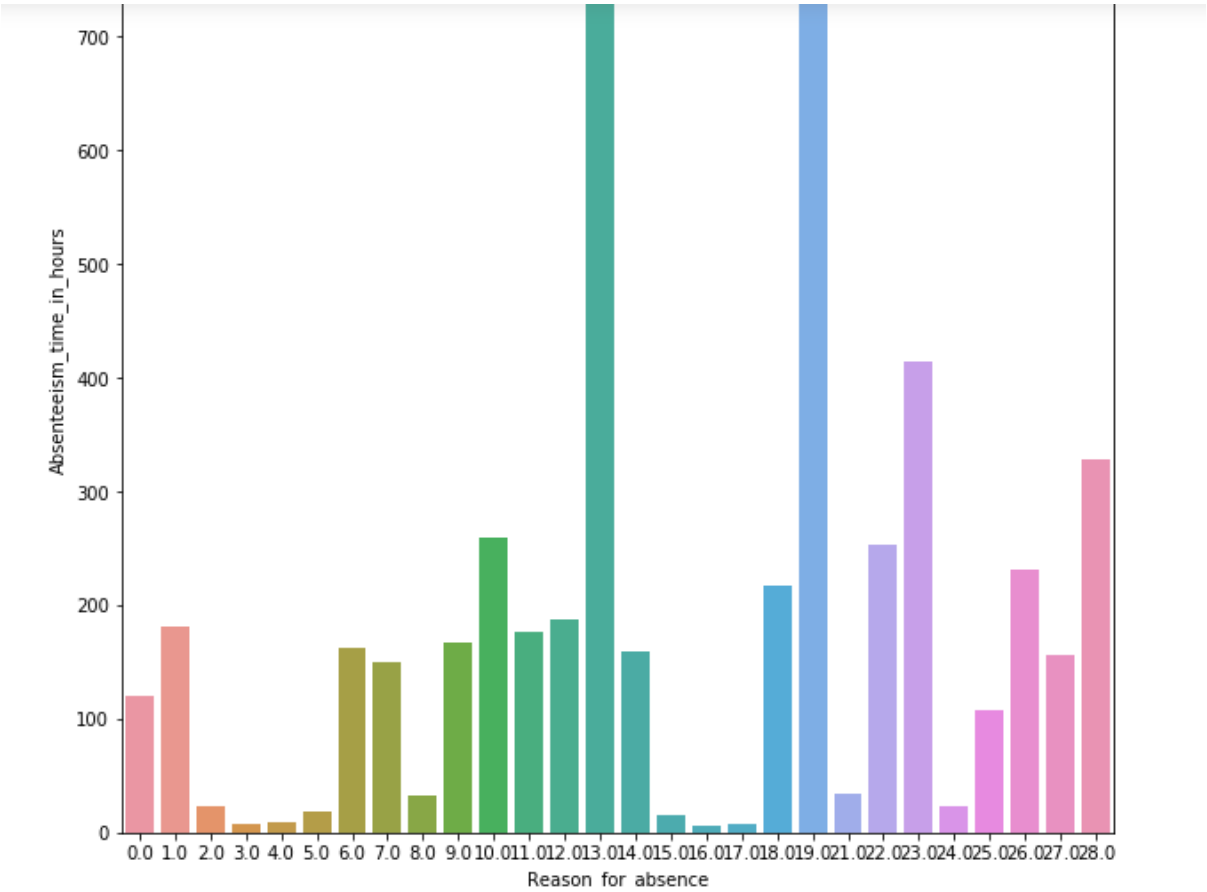
Reason 13, 19, 23 and 28 has more absenteeism,

13→ Diseases of the musculoskeletal system and connective tissue

19→ Injury, poisoning and certain other consequences of external causes

23→ medical consultation

28→ dental consultation



To answer first question which is,

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

Ans→ following are the ways to reduce absenteeism

1) Spread fitness awareness among employees, encourage employees to practise healthy life styles.

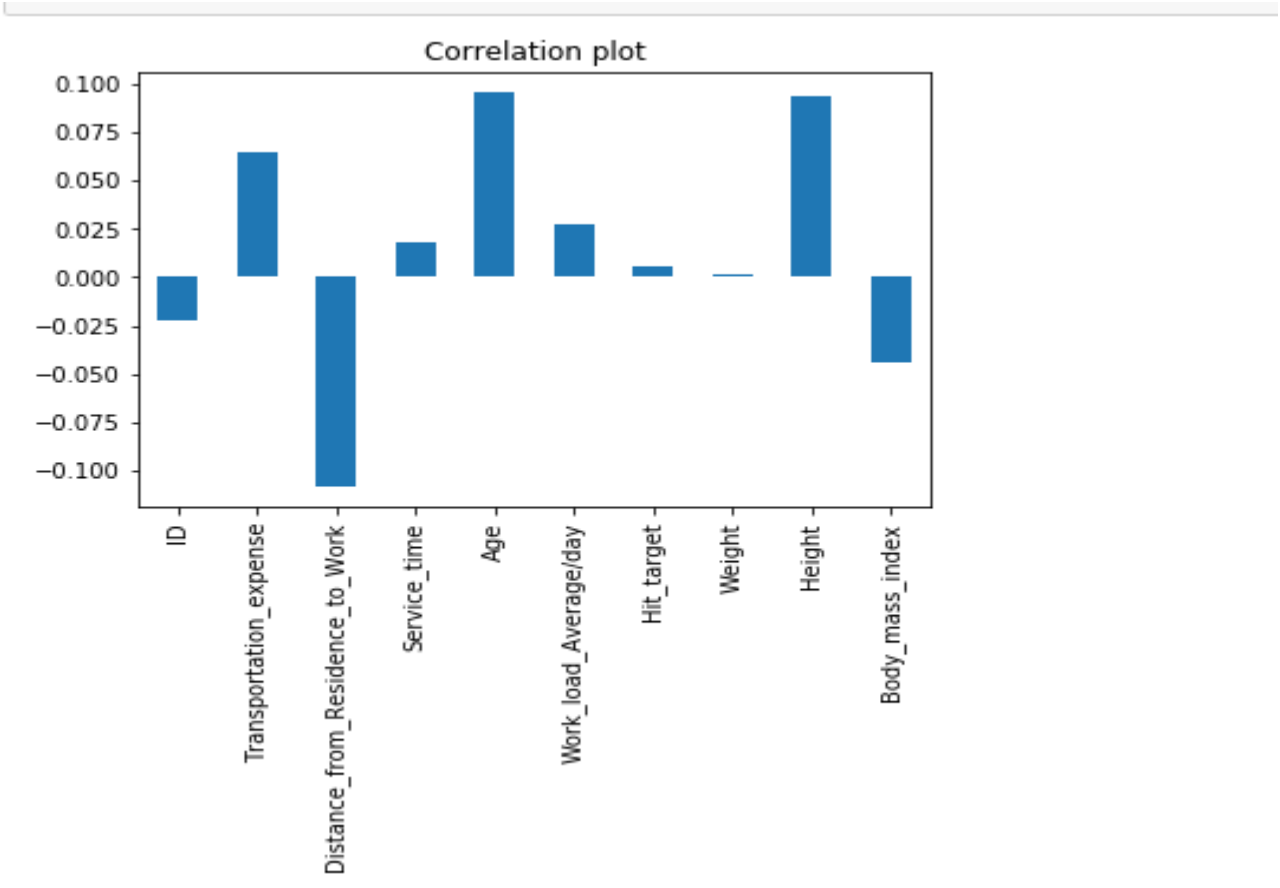
Have a hospital at workplace / have rest rooms where employees can take rest for few hours and work instead of taking whole day off.

2) As we can see more absenteeism on Monday and Tuesday, companies can encourage work from home this will potential reduce the leaves during holiday months like June/March/November also on Mondays. And during winter season also. Also people who are absent due to muscle pain and medical consultation can work from home instead.

3) As we can see drinkers take more leaves so company can promote rewards to non drinkers.

4) As we can see employee with higher education qualification takes less leaves than less qualified employee so, company can promote higher education.

Let's understand how continuous variables are affecting Absenteeism.



From the above graph it can be seen that variables like distance to work, Age, transportation expenses height are affecting Absenteeism. So Following can be done to reduce Absenteeism

5) Company can provide vehicles for commute which would provide less transportation expenses, and discourage absenteeism.

6) People with more height tend to take more leaves and also more the age more is absenteeism.

So to company can provide more comfortable chairs/furniture to more heighted/aged employees to reduce absentism.

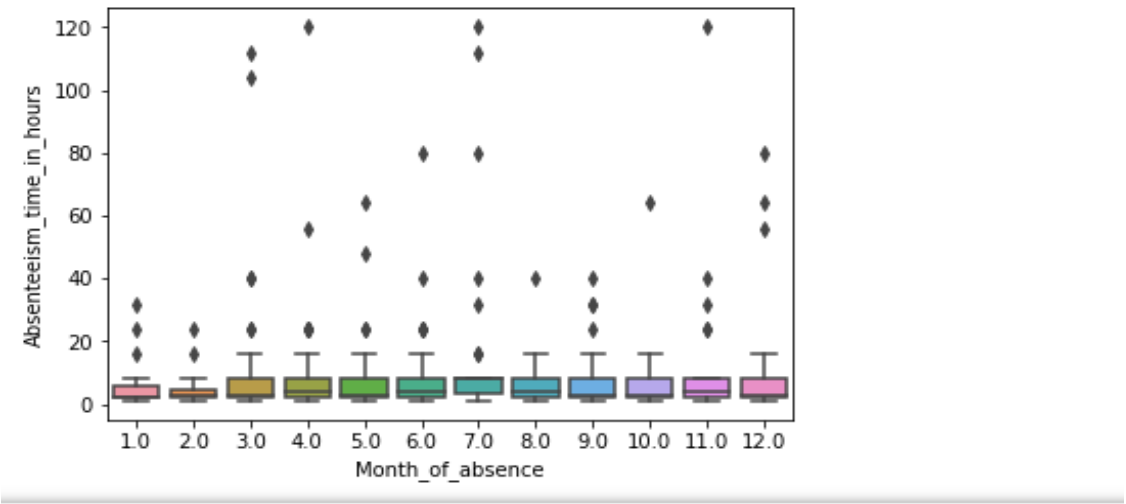
Missing value analysis:

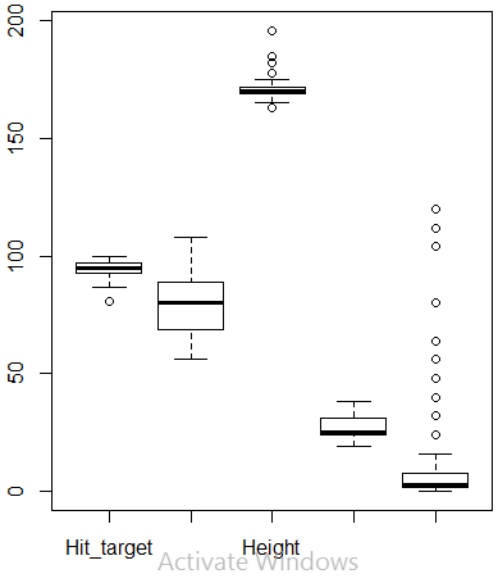
ID	0
Reason_for_absence	3
Month_of_absence	1
Day_of_the_week	0
Seasons	0
Transportation_expense	7
Distance_from_Residence_to_Work	3
Service_time	3
Age	3
Work_load_Average_day	10
Hit_target	6
Disciplinary_failure	6
Education	10
Son	6
Social_drinker	3
Social_smoker	4
Pet	2
Weight	1
Height	14
Body_mass_index	31
Absenteeism_time_in_hours	22

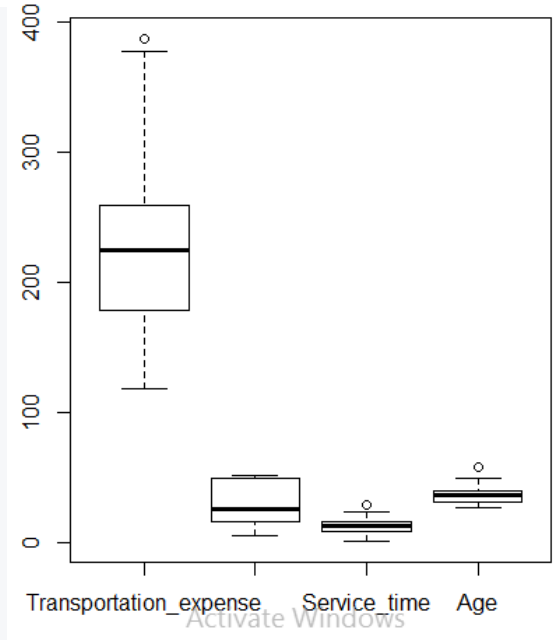
We can compute missing values for Age, BMI, Service time, son, smoker, education, weight, height,Drinker, smoker, distance from work, transportation expense for individual Ids. Because ID is unique for each employee so these characteristics. For work load we can consider month of absence and can take mode from the month to impute missing value. I have used mean method to impute missing values.

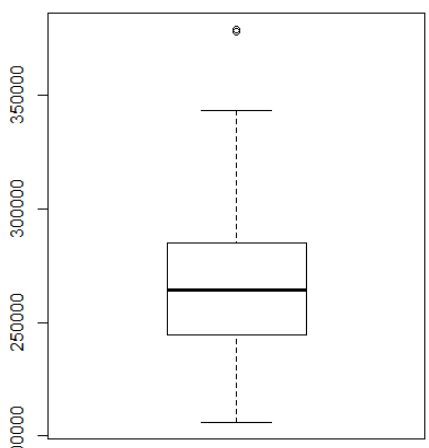
Outlier analysis:

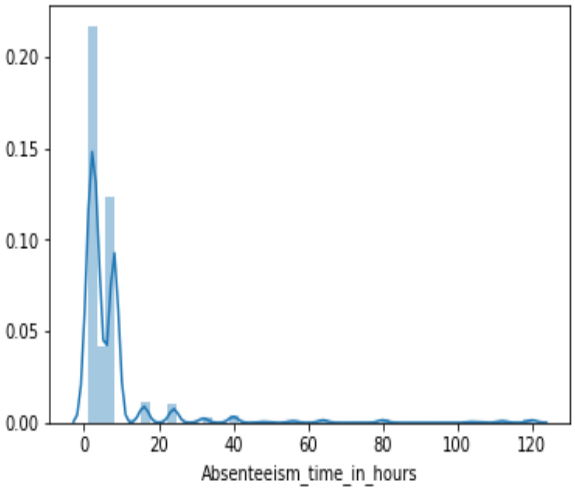
We can see our data has lot of outliers. lets plot box plot



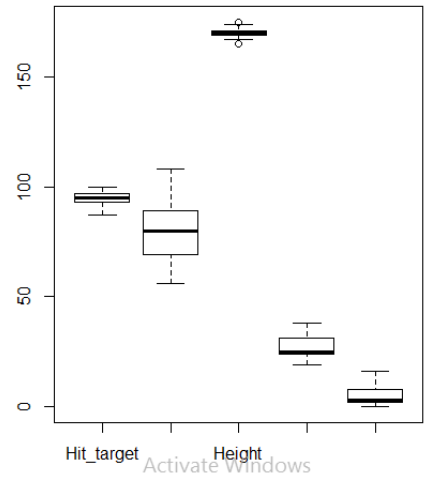


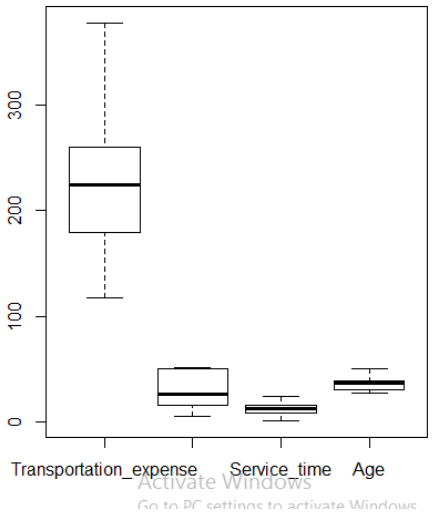


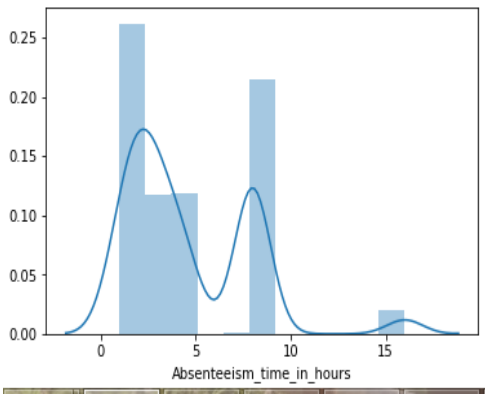


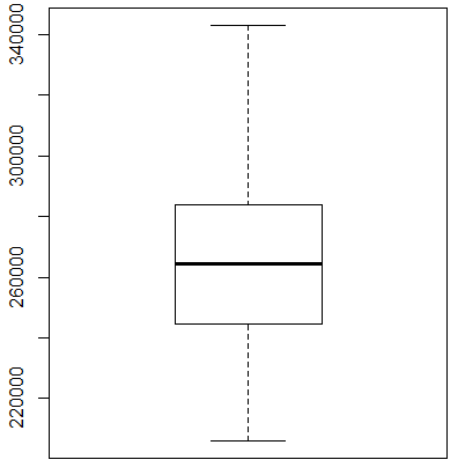


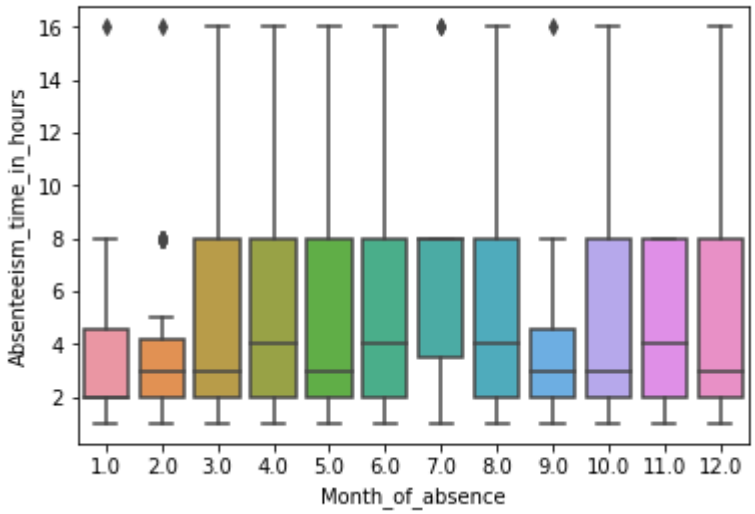
Due to presence of outliers dependent variable is not normally distributed. Outliers have been removed and computed by mean. After removal let's plot boxplot again.







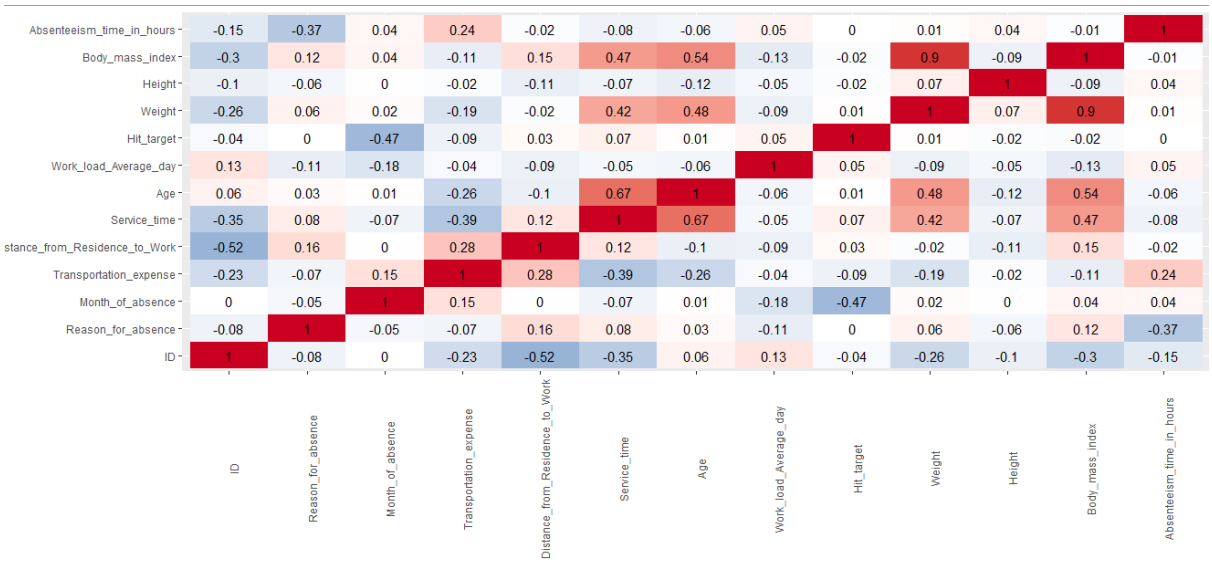




Its clear from the boxplot that outliers have been moved and we can see significant increase in the mean of the variables compare to presence with outliers however we still bimodal distribution of the dependent variable.

Feature extraction:

Lets see for correlation between independent variable to avoid multi-co linearity effect.



From the plot we can see weight and BMI has high correlation so will drop BMI. Also, To answer the monthly losses we don't actually need all the variables will just select necessary variables like, ID ,month, distance, service

time, age, work_load, height ,son, drinker, pet. Anyways we need to predict output by month so taking season/days won't make any sense.

And there are few invalid entries in reason so will just exclude rason.

Feature Scaling :

Looking at the final set of data, we need normalization because there work_load has range much higher than other variables this will impact performance of our model.

ID	Month_of_absence	Distance_from_Residence_to_Work	Service_time	Age	Work_load_Average/day	Height	Son	Social_drinker	Pet	Absenteeism_time_in_h
1	1.0	11.0	14.0	37.0	330061.0	172.0	1.0	0.0	1.0	
1	2.0	11.0	14.0	37.0	330061.0	172.0	1.0	0.0	1.0	
1	3.0	11.0	14.0	37.0	244387.0	172.0	1.0	0.0	1.0	
1	4.0	11.0	14.0	37.0	326452.0	172.0	1.0	0.0	1.0	
1	5.0	11.0	14.0	37.0	246074.0	172.0	1.0	0.0	1.0	

After normalization:

ID	Month_of_absence	Distance_from_Residence_to_Work	Service_time	Age	Work_load_Average/day	Height	Son	Social_drinker	Pet	Absenteeism_ti
1	1.0	0.12766	0.52381	0.454545	0.903944	0.7	1.0	0.0	1.0	
1	2.0	0.12766	0.52381	0.454545	0.903944	0.7	1.0	0.0	1.0	
1	3.0	0.12766	0.52381	0.454545	0.280116	0.7	1.0	0.0	1.0	
1	4.0	0.12766	0.52381	0.454545	0.877665	0.7	1.0	0.0	1.0	
1	5.0	0.12766	0.52381	0.454545	0.292400	0.7	1.0	0.0	1.0	

Model performance and selection.

For predicting the losses we need regression model so, we can go for linear regression , random forest ,decision trees and xg boost.But off course our data is not normally distributed so using linear regression would be a mistake. Will go with random forest and xgboost and decision tree.

	Model	R-Squared	MeanSquaredError	RootMeanSquaredError	MeanAbsoluteError
0	Decision tree	0.970873	0.000028	0.005274	0.003238
1	Random forest	0.870785	0.000014	0.003789	0.002628
2	XG Boost	0.943515	0.000020	0.004463	0.003068

As from the above metric we can see random forest is giving minimum error will choose random forest as our final model.

Monthly loss prediction:

As in 2011 same condition are going to be present so let's predict values . and calculate the monthly losses.

	Month_of_absence	absent_hrs	Predict_hr
0	1.0	7.0	20.715425
1	2.0	26.0	23.081331
2	3.0	40.0	30.963266
3	4.0	18.0	24.456989
4	5.0	37.0	39.143216
5	6.0	22.0	28.530617
6	7.0	19.0	24.155075
7	8.0	10.0	22.641775
8	9.0	20.0	18.737455
9	10.0	24.0	13.900121
10	11.0	18.0	17.883630
11	12.0	34.0	38.273423

And for loss calculation :

Total hrs= 22*8*36 =6336

Loss= (predicted hrs/total loss)*100

	Month_of_absence	absent_hrs	Predict_hr	loss
0	1.0	7.0	20.715425	0.326948
1	2.0	26.0	23.081331	0.364289
2	3.0	40.0	30.963266	0.488688
3	4.0	18.0	24.456989	0.386000
4	5.0	37.0	39.143216	0.617791
5	6.0	22.0	28.530617	0.450294
6	7.0	19.0	24.155075	0.381235
7	8.0	10.0	22.641775	0.357351
8	9.0	20.0	18.737455	0.295730
9	10.0	24.0	13.900121	0.219383
10	11.0	18.0	17.883630	0.282254
11	12.0	34.0	38.273423	0.604063

Chapter 3 Conclusion :

To reduce absenteeism company can provide work from home also provide transportation promote healthy lifestyle; provide medical consultation at work place along with comfortable desk. Using the RF model we can predict the loss percentage for 2011.