PORTFOLIO PROJECT

PROJECT REPORT
ON
EMPLOYEE ABSENTEEISM AT WORK

ABHIMANYU BHATIA

15 December 2018

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background

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 would like our help regarding the following:

1) Solutions on how they can reduce the number of absenteeism.
2) Projected Monthly Losses in 2011 if same trend of absenteeism continues.

1.2. Exploratory Data Analysis

The first objective is to analyse this dataset and find relations & patterns between the various predictor variables, so that we can provide our insights on how to reduce the number of absenteeism.
The second objective is to apply the machine learning algorithms and build models on this dataset in order to predict the employee absenteeism trend for the year 2011.
Given below is a sample of the data set that shared to us by the XYZ courier company:

Table 1: Absenteeism at work sample data (Columns: 1-7)

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work
11	26	7	3	1	289	36
36	0	7	3	1	118	13
3	23	7	4	1	179	51
7	7	7	5	1	279	5

Table 2: Absenteeism at work sample data (Columns: 8-14)

Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Education	Son
13	33	239,554	97	0	1	2
18	50	239,554	97	1	1	1
18	38	239,554	97	0	1	0
14	39	239,554	97	0	1	2

Table 3: Absenteeism at work sample data (Columns: 15-21)

Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
1	0	1	90	172	30	4
1	0	0	98	178	31	0
1	0	0	89	170	31	2
1	1	0	68	168	24	4

By observing the Month of absence column of the dataset we can infer that this dataset is from July 2007 to July 2010.

Thus, we have created a new column called 'year' which shows the year of absence of the employee

As you can see in the table 4 we have the following 21 predictor variables, using which we have to predict the trend of absenteeism in the year 2011:

Table 4: Predictor Variables of our Dataset

S. No.	PREDICTOR
1.	ID
2.	Reason for absence
3.	Month of absence
4.	year
5.	Day of the week
6.	Seasons
7.	Transportation expense
8.	Distance from Residence to Work
9.	Service time
10.	Age
11.	Work load Average/day
12.	Hit target
13.	Disciplinary failure
14.	Education
15.	Son
16.	Social drinker
17.	Social smoker
18.	Pet
19.	Weight
20.	Height
21.	Body mass index

Using the describe function in python we have obtained a brief description of our dataset

Table 5: Brief Description of the Dataset

	count	mean	std	min	25%	50%	75%	max
ID	740	18.02	11.02	1	9	18	28	36
Reason for absence	737	19.19	8.44	0	13	23	26	28
Month of absence	739	6.32	3.44	0	3	6	9	12
Year	740	2008.6	1.01	2007	2008	2009	2009	2010
Day of the week	740	3.91	1.42	2	3	4	5	6
Seasons	740	2.54	1.11	1	2	3	4	4
Transportation expense	733	221.04	66.95	118	179	225	260	388
Distance from Residence to Work	737	29.67	14.85	5	16	26	50	52
Service time	737	12.57	4.39	1	9	13	16	29
Age	737	36.45	6.48	27	31	37	40	58
Work load Average/day	730	271188.86	38981.88	205917	244387	264249	284853	378884
Hit target	734	94.59	3.79	81	93	95	97	100
Disciplinary failure	734	0.05	0.22	0	0	0	0	1
Education	730	1.30	0.68	1	1	1	1	4
Son	734	1.02	1.09	0	0	1	2	4
Social drinker	737	0.57	0.50	0	0	1	1	1
Social smoker	736	0.07	0.26	0	0	0	0	1
Pet	738	0.75	1.32	0	0	0	1	8
Weight	739	79.06	12.87	56	69	83	89	108
Height	726	172.15	6.08	163	169	170	172	196
Body mass index	709	26.68	4.29	19	24	25	31	38
Absenteeism time in hours	718	6.98	13.48	0	2	3	8	120

We can infer that out of the 22 variables in our dataset 12 are categorical and 10 are continuous variables.

- **CONTINUOUS VARIABLES:** Distance from Residence to Work, Service time, Age, Work load Average/day, Hit target, Transportation Expense, Weight, Height, Body mass index, Absenteeism time in hours

- **CATEGORICAL VARIABLES:** ID, Reason for absence, Month of absence, Day of the week, Seasons, Disciplinary failure, Education, Son, Social drinker, Social smoker, Pet, year

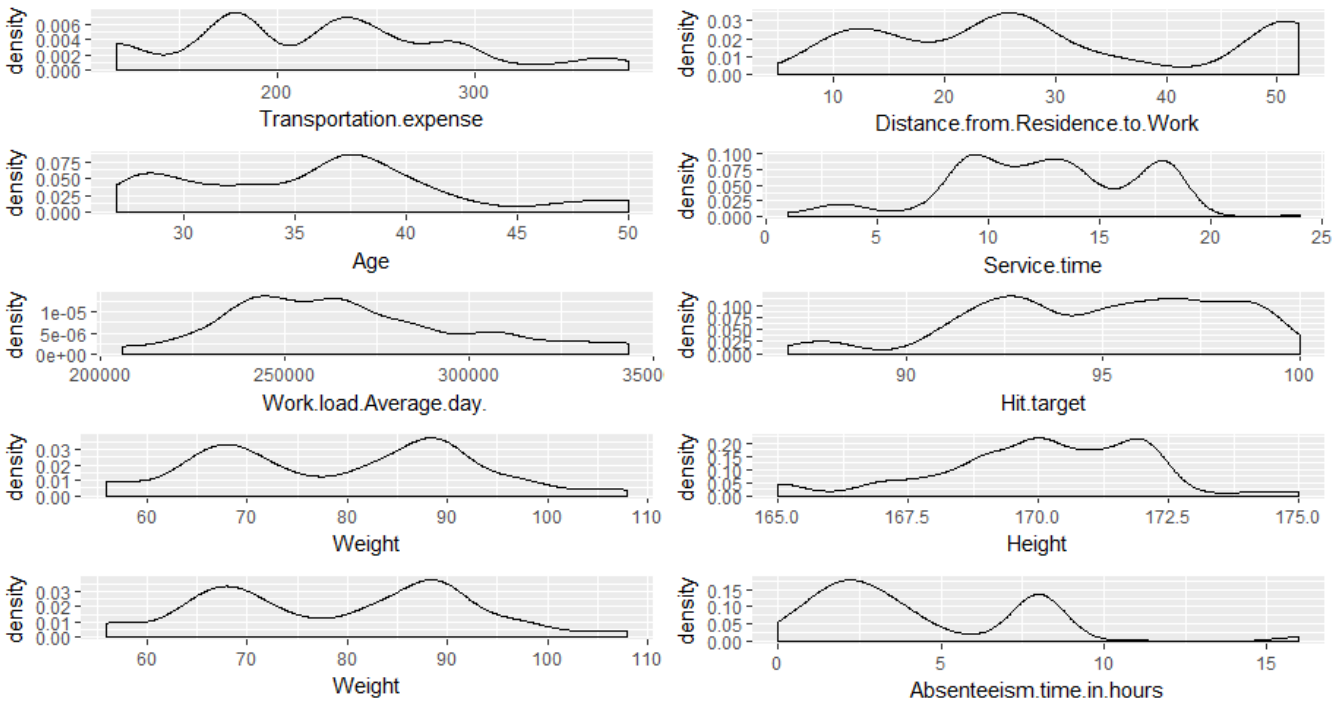

Fig.1: Density Plots of Continuous variables

- From the density plots above we can observe that the continuous variables in our dataset are skewed.
- This, can be attributed to the fact there are outliers in the dataset.

2. DATA PRE-PROCESSING

2.1. Missing Value Analysis

- 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 the missing data is less than 30% of the total data its value can be imputed using various statistical techniques.
- Table 6 shows the missing values in our dataset.

Table 6: Missing Values in Absenteeism at work dataset

Feature Variables	No. of missing values	Percentage of Missing Values (%)
ID	0	0.00
Reason for absence	3	0.41
Month of absence	1	0.14
year	0	0.00
Day of the week	0	0.00
Seasons	0	0.00
Transportation expense	7	0.95
Distance from Residence to Work	3	0.41
Service time	3	0.41
Age	3	0.41
Work load Average/day	10	1.35
Hit target	6	0.81
Disciplinary failure	6	0.81
Education	10	1.35
Son	6	0.81
Social drinker	3	0.41
Social smoker	4	0.54
Pet	2	0.27
Weight	1	0.14
Height	14	1.89
Body mass index	31	4.19
Absenteeism time in hours	22	2.97

- From the table 6 it can be clearly seen that the missing count of all the variables is in the range of 0 -5 %.
- Thus, we can impute the missing values for all the variables.

2.2. Outlier Analysis

- An outlier is an observation point that is distant from other observations.
- An outlier may be due to variability in the measurement or it may indicate experimental error.
- An outlier can cause serious problems in statistical analyses
- Box plots are non-parametric i.e. they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution.
- Thus, Box Plot technique for detecting Outliers can be used for any type of data distribution.
- To perform Outlier Analysis box plots for all the continuous variables are plotted and shown below.

Fig.2: Box Plot of variable “Distance from Residence to Work”

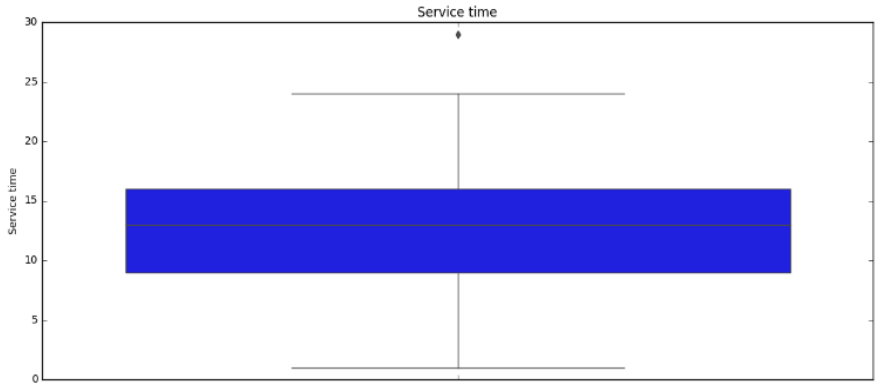

Fig.3: Box Plot of variable “Service time”

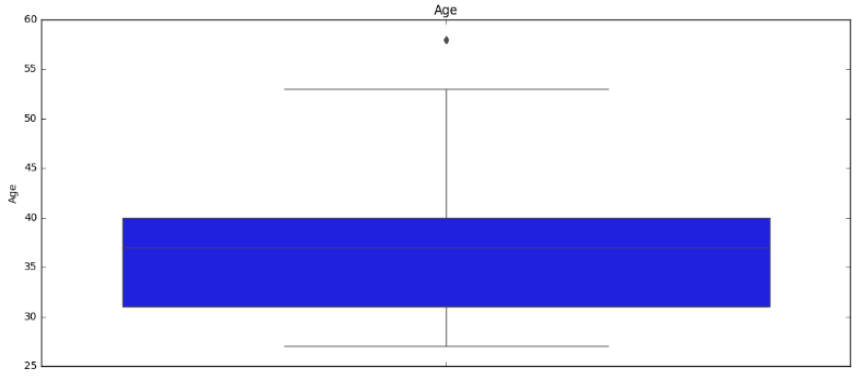

Fig.4: Box Plot of variable “Age”

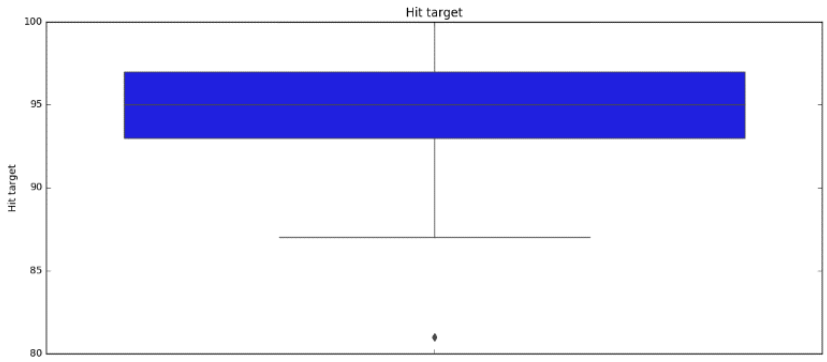

Fig.5: Box Plot of variable “Hit target”

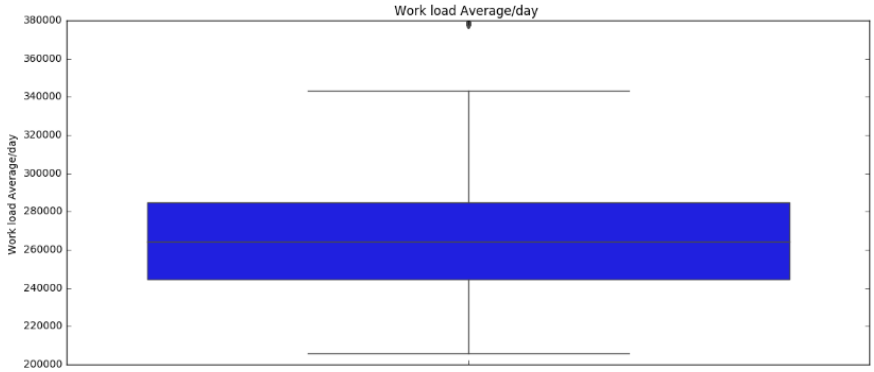

Fig.6: Box Plot of variable “Work load Average/day”

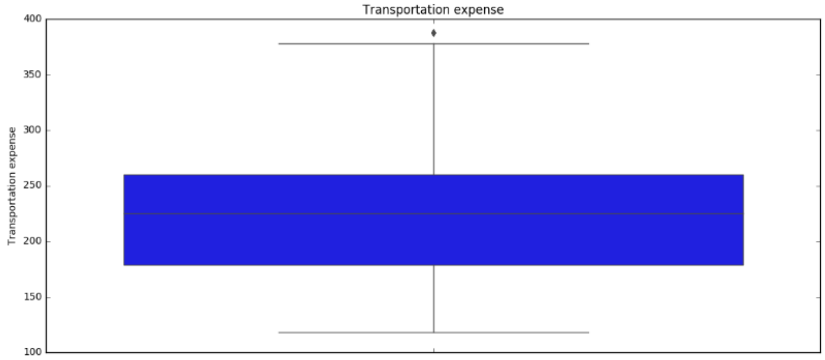

Fig.7: Box Plot of variable “Transportation expense”

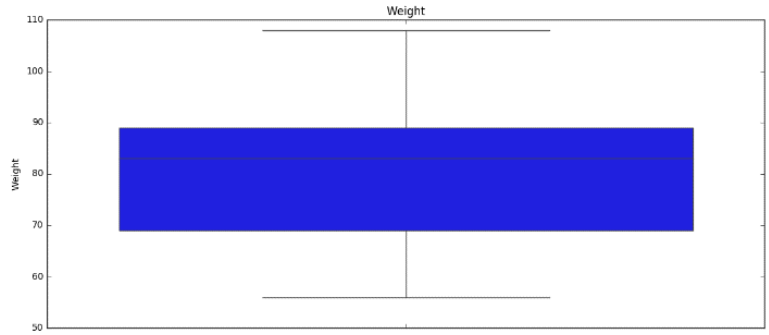

Fig.8: Box Plot of variable “Weight”

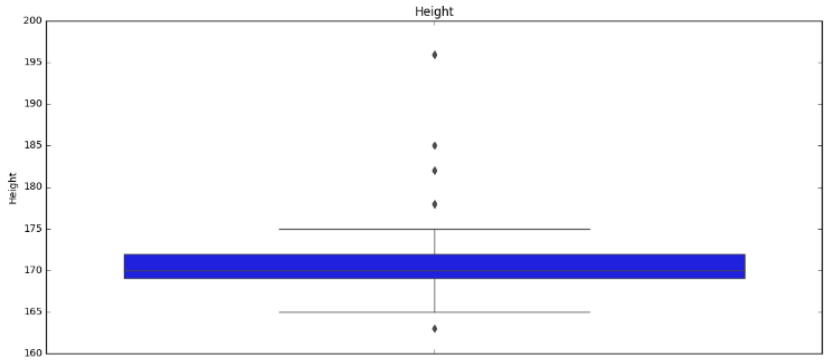

Fig.9: Box Plot of variable “Height”

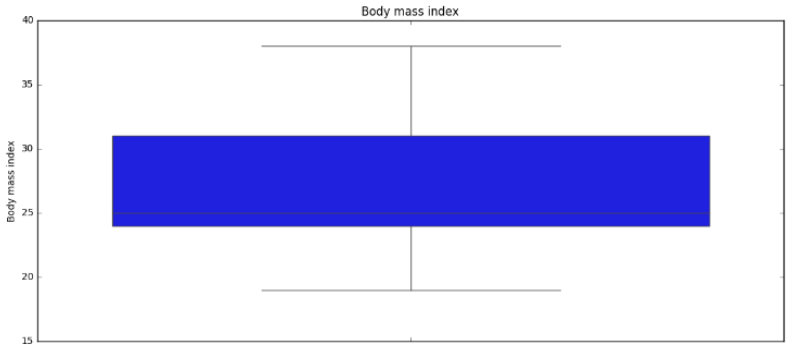

Fig.10: Box Plot of variable “Body mass index”

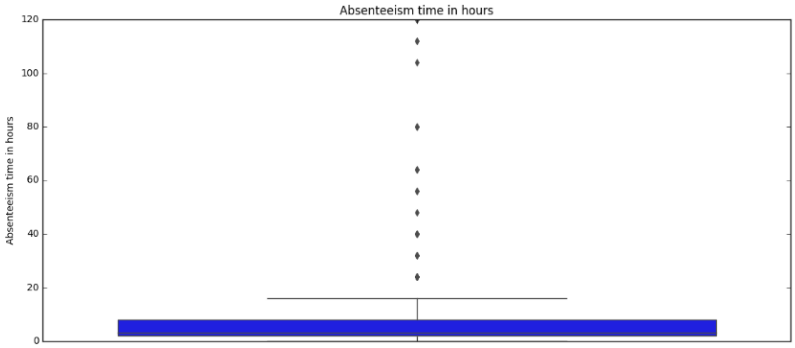

Fig.11: Box Plot of variable “Absenteeism time in hours”

From the above box plots we can see that the target variable “Absenteeism time in hours” has the most no. of outliers.
This is because for any working day the Absenteeism time cannot be more than 24 hours.

2.3. Missing Values & Outliers Imputation

- Since, both the missing value and outliers are very less in comparison to the rest of the observations they can be easily imputed without causing much variance to the model.
- For imputation three methods were tested: Mean, Median & KNN.
- KNN imputation was found to be the most accurate.
- KNN imputation (k=3) is used for imputation of Missing values & Outliers.

2.4. Feature Selection

- In large datasets the number of predictor variables are sometimes very high.
- When these variables have high correlation with each other the problem of multi-collinearty arises.
- By using techniques of Correlation plots and ANOVA we can check the variables that contribute towards multi-collinearity.
- Thus, some of these variables can be dropped.
- It then becomes easier to analyse the dataset.

2.4.1 Correlation

The correlation plot of all the continuous variables is shown below:

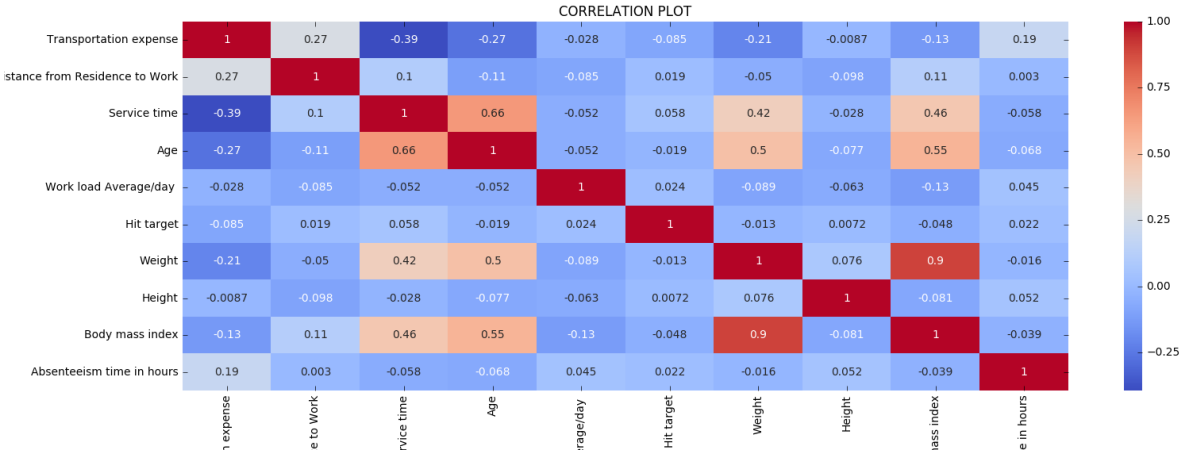

Fig.12: Correlation Plot

From the above plot following observations can be made:

- None of the predictor variables are highly correlated with the target variable
- Body Mass Index and Weight are highly correlated (factor of 0.9).
- Thus, we have dropped the Body mass index variable.

2.4.2 ANOVA

We have used the one way ANOVA test for categorical variables.

The P value of ID is 5.336353051624431e-172
The P value of Reason for absence is 3.652821281380938e-277
The P value of Month of absence is 4.032282401363404e-27
The P value of year is 0.0
The P value of Day of the week is 0.0005186559389246773
The P value of Seasons is 2.6074274571308503e-41
The P value of Disciplinary failure is 4.2953834629897344e-190
The P value of Education is 1.553542440364956e-108
The P value of Son is 6.035424443360698e-119
The P value of Social drinker is 7.859879891800084e-155
The P value of Social smoker is 1.5186474452639762e-188
The P value of Pet is 1.1700454953402614e-131

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

- From the above ANOVA test we observed that the p values for all the categorical variables was less than 5%
- Thus, we conclude that all the categorical variables are important and none of them can be dropped from our model.

3. DATA VISUALIZATION

3.1. Count Plots

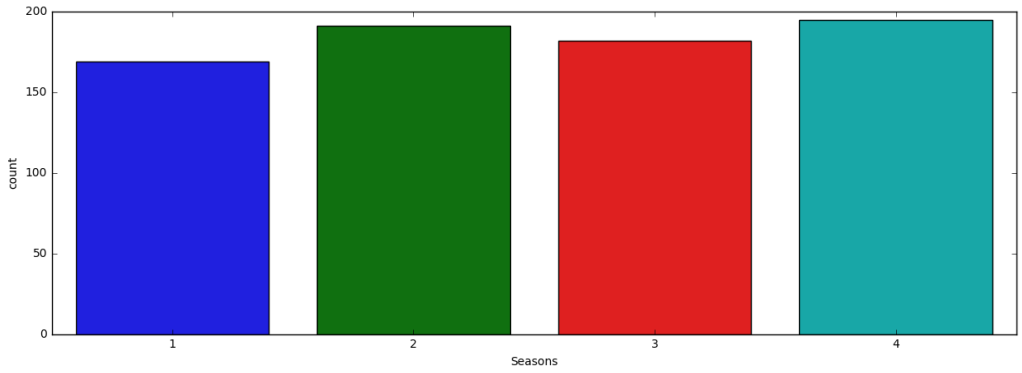

Fig.13: Count Plot of variable “Seasons”

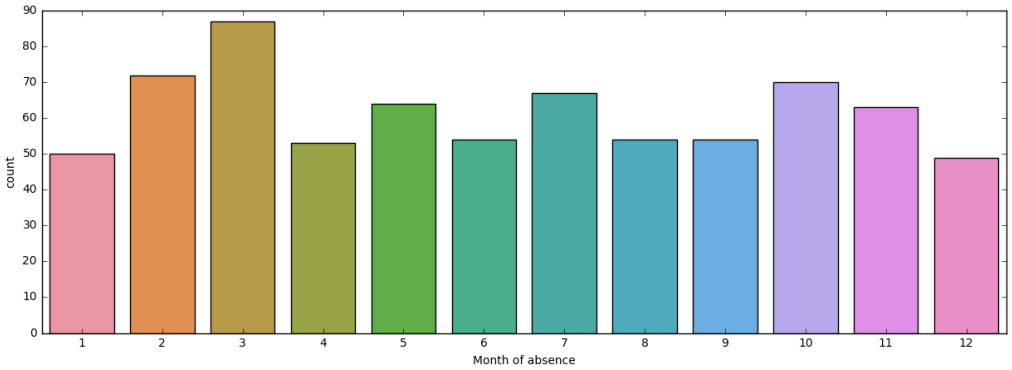

Fig.14: Count Plot of variable “Month of absence”

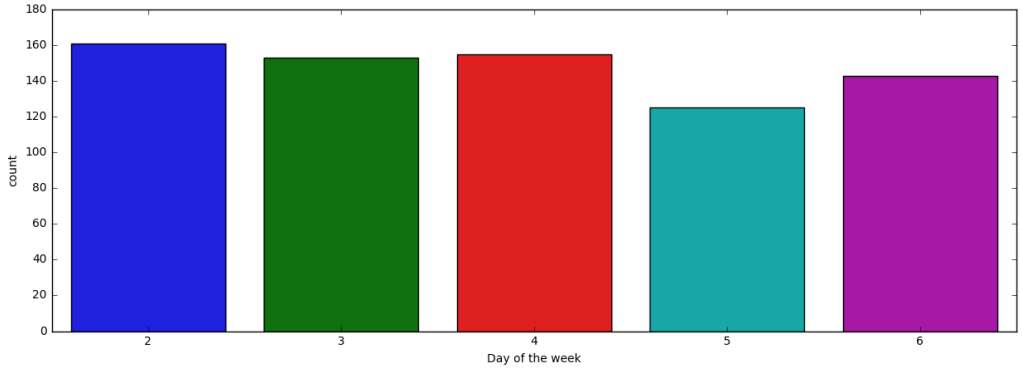

Fig.15: Count Plot of variable “Day of the week”

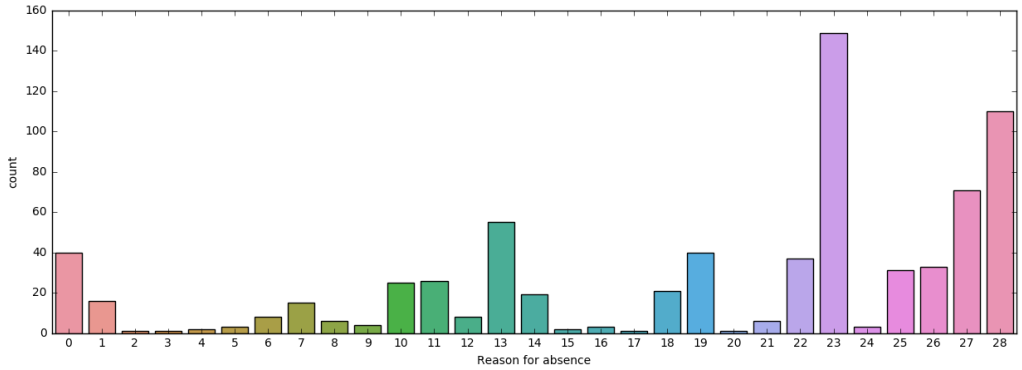

Fig.16: Count Plot of variable “Reason for absence”

INFERENCES BASED ON COUNT PLOTS:

From the count plot of “Reason for absence” the reason of absence it can be clearly observed that the most common reasons of absence are 23 followed by 28 and 27.
23: Medical Consultation. 28: Dental consultation 27: Physiotherapy.
These 3 reasons do not require attestation by the International Code of Diseases (ICD)

3.2. Bar Plots

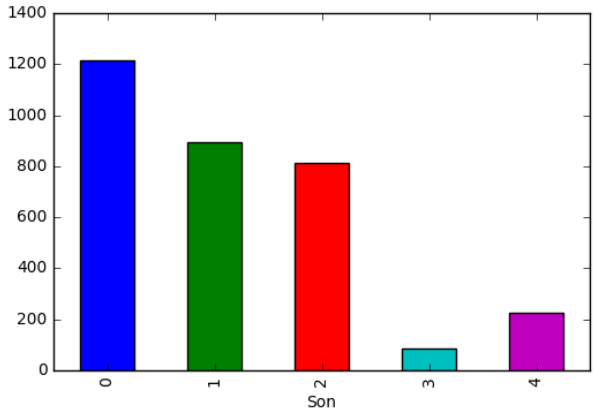

Fig.17: Bar Plot of variable “Son”

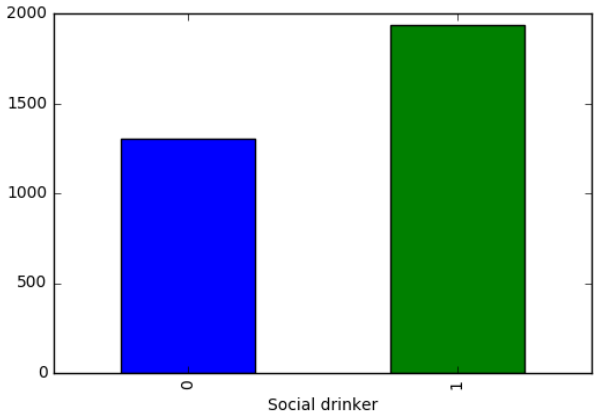

Fig.18: Bar Plot of variable “Social drinker”

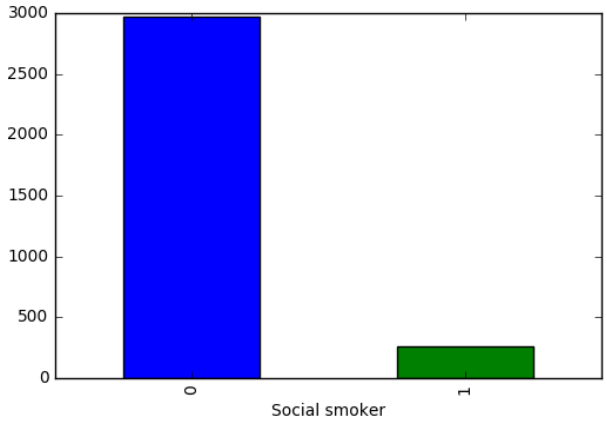

Fig.19: Bar Plot of variable “Social smoker”

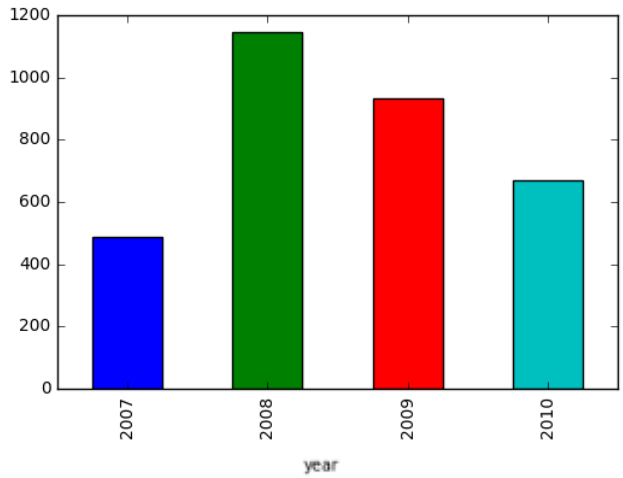

Fig.20: Bar Plot of variable “year”

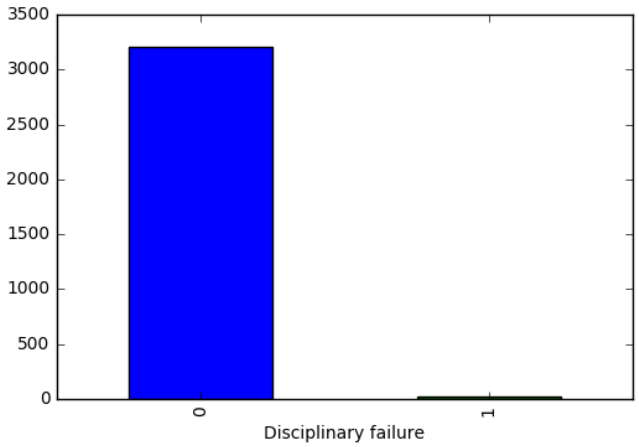

Fig.21: Bar Plot of variable “Disciplinary failure”

INFERENCES BASED ON BAR PLOTS:

From the bar plot of “Son” it can be seen that the people that don’t have any children have higher absent hours.
From the bar plot of “Disciplinary failure” it is inferred after a disciplinary action is taken against an employee the absenteeism hours decrease significantly.

3.3. Line Plots

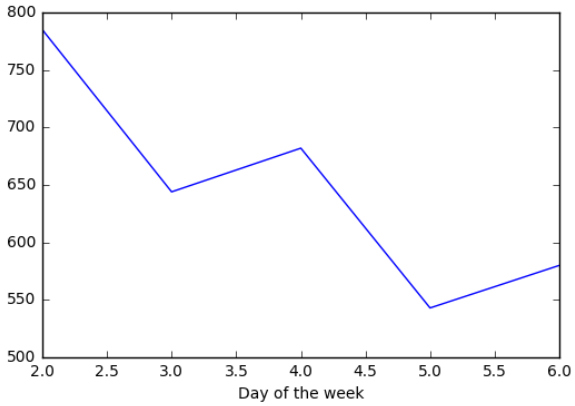

Fig.22: Line Plot of variable “Day of the week” with total absenteeism hours

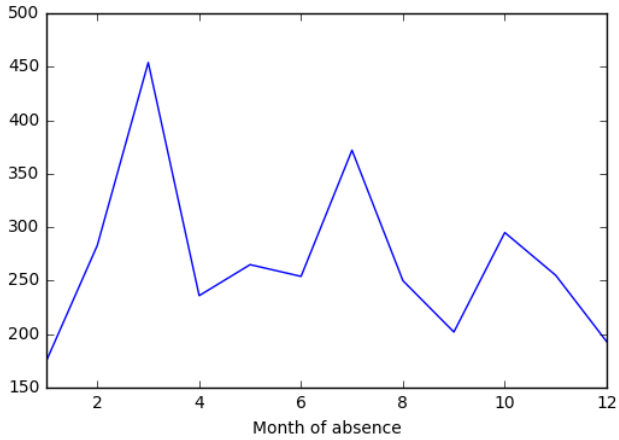

Fig.23: Line Plot of variable “Month of absence” with total absenteeism hours

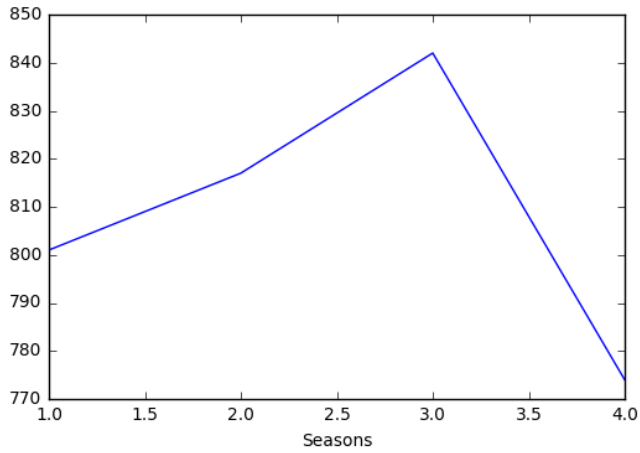

Fig.24: Line Plot of variable “Seasons” with total absenteeism hours

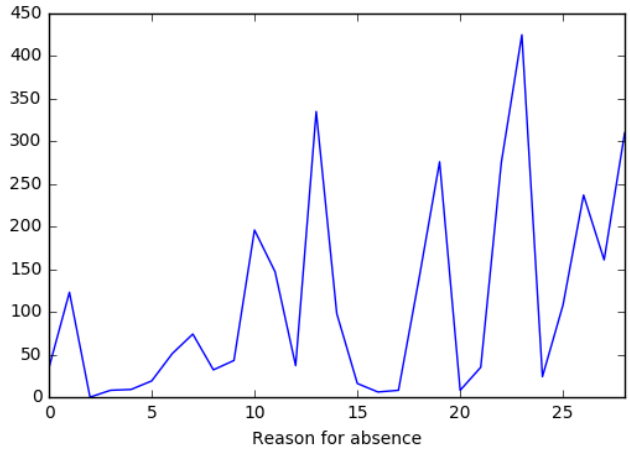

Fig.25: Line Plot of variable “Reason of absence” with total absenteeism hours

INFERENCES BASED ON LINE PLOTS:

From the line plot of “Month of absence” it can be seen that March and July have the highest number of absenteeism hours.
From the line plot of “Day of the week” it can be seen that Monday has the highest number of absenteeism hours.
From the line plot of “Seasons” it can be seen that winter season has the highest number of absenteeism hours.

4. INSIGHTS AND SOLUTIONS

4.1. Insights

- The most popular reasons of absence given by people are from category 23, 28 and 27. These categories are:
  23: Medical Consultation. 28: Dental consultation 27: Physiotherapy
- Employees are using these reasons as these reasons don’t require any medical certification from the doctor.
- Winter season has the highest number of absenteeism hours which seems likely as the climatic conditions are most difficult in winters.
- The months of March and July have the highest number of absenteeism hours which makes sense since, spring break and summer breaks are most common in these months.
- Monday has the highest number of absenteeism hours which seems likely as it is the first day of the job after the weekend.
- The employees who don’t have any children have the highest number of absent hours.
- When a disciplinary action is taken against the employee, they tend to take their job much more seriously and the absenteeism hours reduce significantly.

4.2. Solutions

- Since the most common reasons of absence are consultation, the company can organize health check-ups on a quarterly basis to ensure that all the employees stay in good health. This will also keep up the morale of employees and also help in reducing absenteeism hours.
- The employees tend to take disciplinary actions very seriously, so the company can they can implement strict actions against frequent defaulters thereby keeping their absentee hours in check.
- Since most of the employees are absent on Mondays, the company can decide to keep a lighter schedule for Mondays and Tuesdays and increase the workload slightly

on Fridays. This will help the employees ease into transition when they resume office after the weekend and thus, reduce company's losses.

- After observing the dataset we found that there were very few employees who had almost zero absentee hours. The company can decide to reward these employees which in turn might also inspire other employees to take off fewer hours from work.
- Winter season has the highest number of absenteeism hours. The company can organize an annual trip in winters which will give its employees some relaxation time and also keep their motivation levels high.

5. MACHINE LEARNING MODELS

To predict the absenteeism hours for the year of 2011 we trained the data using the following supervised machine learning algorithms:

- Linear Regression
- Decision Tree
- Random Forest

After treating the missing values and outliers the data was split into train data (80% of total data) and test data (20% of total data).
All the three algorithms were applied on the training data and then the values of the target variable of test dataset were predicted.
In the test dataset the actual values of target variables were compared with the calculated values of target variable and the Mean Absolute Error (MAE), Mean Squared Error (MSE) & Root Mean Squared Error (RMSE).

The least error was found to be with Linear Regression.

5.1. Linear Regression

The error metrics for linear regression were obtained as follows:

- MAE: 2.297175677996193
- MSE: 10.261481800256993
- RMSE: 3.2033547727744724

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

5.2. Decision Tree

The error metrics for decision tree were obtained as follows:

- MAE: 2.6283783783783785
- MSE: 17.972972972972972
- RMSE: 4.239454324907036

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

5.3. Random Forest

The number of estimators were taken as 300.
The error metrics for random forest were obtained as follows:

- MAE: 2.3110797940797936
- MSE: 10.697015416576441
- RMSE: 3.270629208054070

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

6. PREDICTIONS USING LINEAR REGRESSION

- Linear Regression model was trained on the complete employee absenteeism dataset.
- The values of target variables were groups by year and months. Thus, the dataset ranged from Month 1 i.e. July 2007 to Month 37 i.e. July 2010
- Using the Linear Regression model that was trained on this dataset the values of absenteeism hours were predicted for the time Period of August 2010 to December 2011 i.e. from Month 38 to Month 54.
- The following values were predicted by our model:

Table 7: Predicted Values of Target Variable
"Absenteeism time in hours" using Linear Regression

Month	Predicted Absenteeism Hours
38	86.51801802
39	86.47131342
40	86.42460882
41	86.37790422
42	86.33119962
43	86.28449502
44	86.23779042
45	86.19108582
46	86.14438122
47	86.09767662
48	86.05097202
49	86.00426743
50	85.95756283
51	85.91085823
52	85.86415363
53	85.81744903
54	85.77074443

- Thus, from our predicted values we can estimate the total absenteeism hours for the year 2011.
- It would be the sum of Absenteeism hours from Month 43 to Month 54. (i.e. from January 2011 to December 2011).
- Therefore, the total estimated absenteeism hours for 2011 are **1032 hours**.

NOTE: *For above results please refer the Jupyter Notebook file (python code).*

CONCLUSION:
The total loss of the company in the year 2011 will be 1032 man hours or 129 man days.