Employee Absenteeism

Gautam Pai

24 December 2019

Contents

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:

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

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

1.2 Data

Our task is to find out features that are contributing to more number of absentees as well as finding the trend in the number of absentees every month

Table 1.1: Employee Absenteeism Sample Data (Columns: 1-6)

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense

4	19	6	2	3	246
10	22	6	2	3	361
11	10	6	3	3	246
10	23	6	5	3	291
13	10	6	6	3	246
12	11	6	2	3	179

Table 1.2: Employee Absenteeism Sample Data (Columns: 7-12)

Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure
25	16		377,550	94	0
52	3	28	377,550	94	0
25	16	41	377,550	94	0
31	12	40	377,550	94	0
25	16		377,550	94	0
51	18	38	377,550	94	0

Table 1.3: Employee Absenteeism Sample Data (Columns: 13-18)

Education	Son	Social drinker	Social smoker	Pet	Weight
1	0	1	0	0	67
1	1	1	0	4	80
1	0	1	0	0	67
1	1	1	0	1	73
1	0	1	0	0	67
1	0	1	0	0	89

Table 1.4: Employee Absenteeism Sample Data (Columns: 19-21)

Height	Body mass index	Absenteeism time in hours
170		8
172	27	8
170	23	24
171	25	4
170	23	
170	31	8

As you can see in the table below we have the following 20 predictor variables

Table 1.5: Predictor Variables

S.No.	Predictor

1	Individual identification (ID)
2	Reason for absence (ICD).
3	Month of absence
4	Day of the week
5	Seasons
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
13	Education
14	Son
15	Social drinker
16	Social smoker
17	Pet
18	Weight
19	Height
20	Body mass index

Chapter 2

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 the data types and then 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.

In Figure2.1 we have plotted the histograms of all the variables we have available in the data.

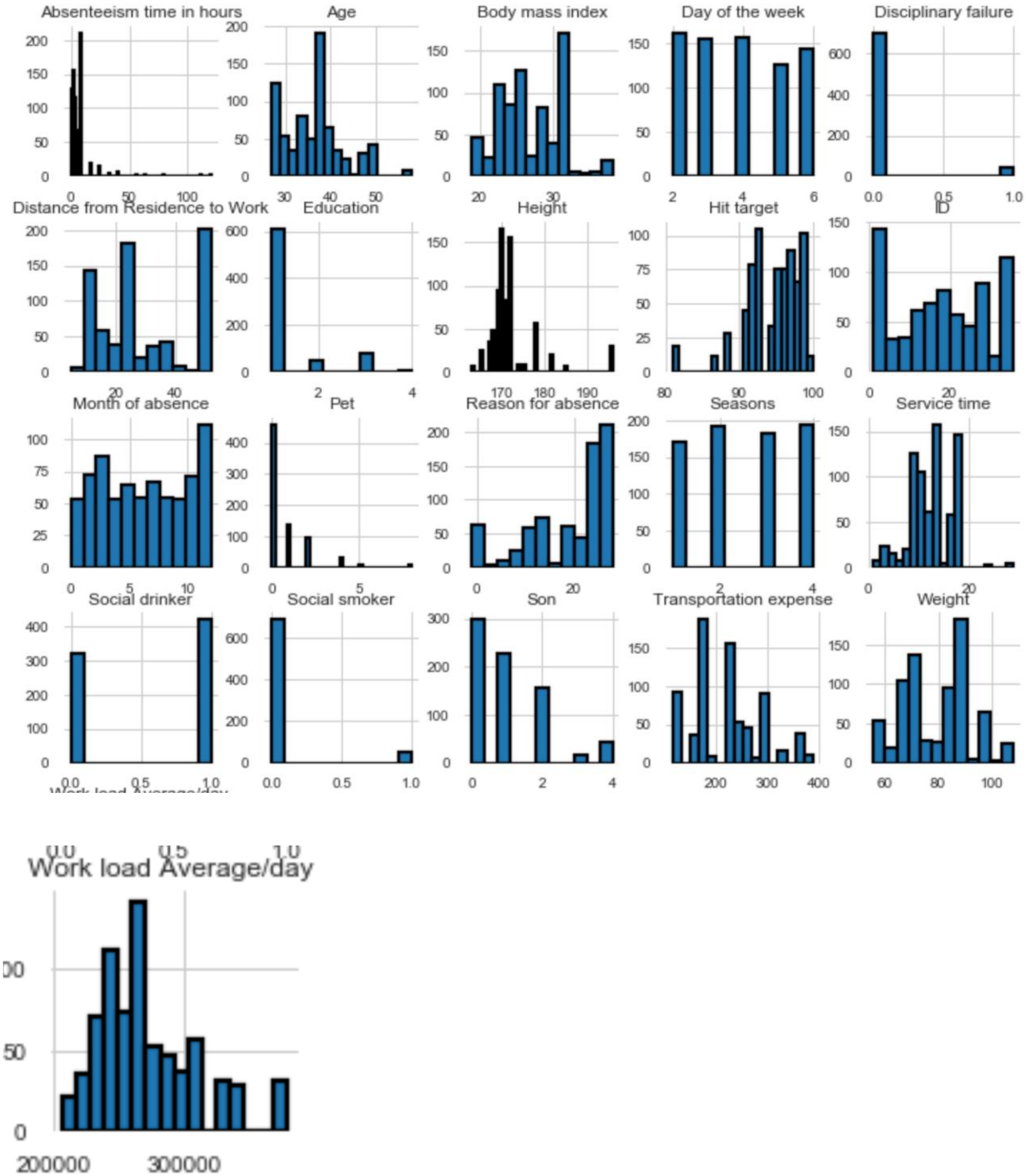


Figure 2.1: Histograms of all the variables in the dataset

From the above plots it looks like all the variables are discreate in nature. Also from the exploratory analysis, there were invalid data within ‘Reason for absence’ and ‘Month of absence’. Since the variables were discrete in nature we tried to print the unique values of each variables.

2.1.1 Missing Value Analysis

In this stage we identify the missing values in each and every feature and gather the count of all the variables. Missing values can arise due to many reasons like business unable to capture the data due to privacy or error in capturing the data. Missing values are present as NA/NAN or empty values. The rule of thumb is if the missing values are more than 30%, it is advisable to drop that variable. If it is below to some percentage based on business scope, we can either drop those samples or else impute those values with either mean, median, knnimputation, random sampling imputation. In Fig 2.2, shows the missing value statistics for every variable.

There are 2 groups of variables present in the data. Ones related to employee and the others related to company.

We see that Height, Weight, Body mass index, Education, Transportation Expense, Son, Social Smoker, Age, Service Time are related to employee attributes and the other attributes like Month of absence, Workload average/day, Hit target, absentee hours are related to workplace attributes. The strategy is to group the similar looking observations coming under above 2 groups and try to impute the missing values using the median statistics. The missing values are imputed in the order of highest missing value features.

index	percentage
Body mass index	4.189189
Absenteeism time in hours	2.972973
Height	1.891892
Work load Average/day	1.351351
Education	1.351351
Transportation expense	0.945946
Son	0.810811
Disciplinary failure	0.810811
Hit target	0.810811
Social smoker	0.540541
Age	0.405405
Reason for absence	0.405405
Service time	0.405405
Distance from Residence to Work	0.405405
Social drinker	0.405405
Pet	0.270270
Weight	0.135135
Month of absence	0.135135
Seasons	0.000000

Figure 2.2: Missing value percentages of all the variables in the dataset

2.1.2 Feature Selection

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 significant enough to explain the variance in the target. Since we have all categorical variables we decided to go with chi square analysis. Once we get significant lower p-values for a variable we will perform a frequency counts for that variable and consider only the values which are in higher frequency (90% quantile). After one iteration we drop that feature from the dataset along with the features that are not significant (higher p-values). We get to a reduced set of variables that will enable the major reasons for employee absenteeism.

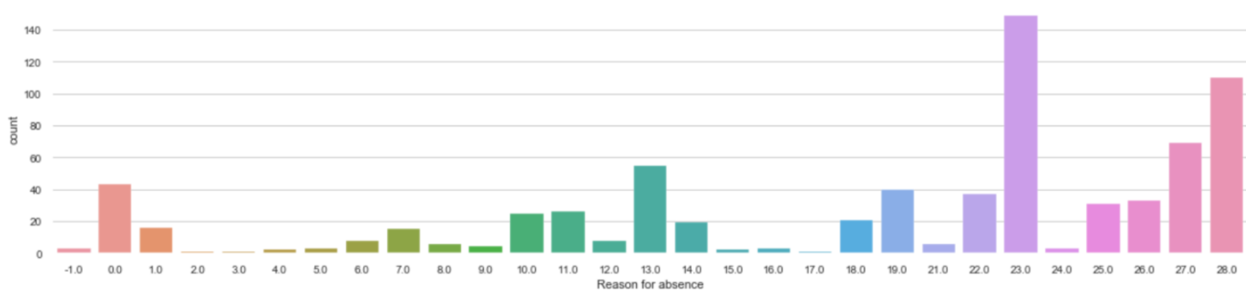
In Chi-Square, we select 2 hypothesis

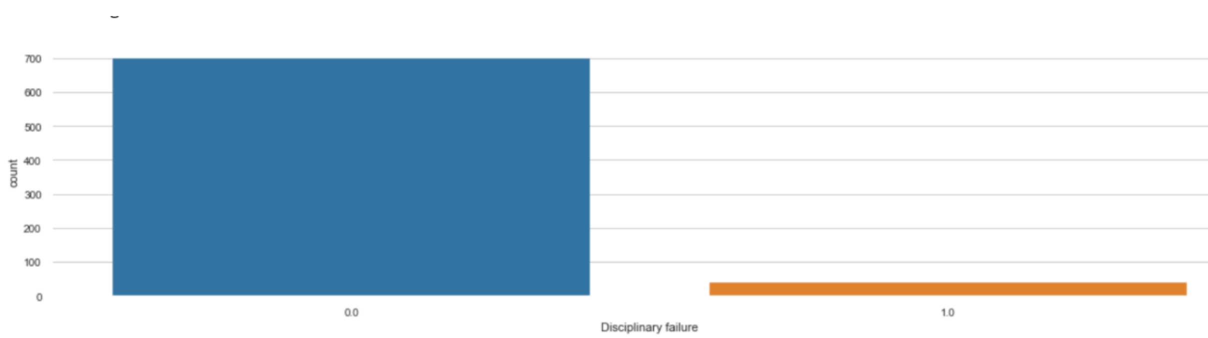
i) H0: Variables are independent of each other
ii) H1: Variables are not independent of each other

If the p-value statistics between our target variable and the predictor variable are <0.05, we reject the null hypothesis, claiming that the variance in the target is contributed by the relevant predictor variable.

1st Iteration

Fig 2.3 shows the chi square tests for all the variables wrt the target variable. We see that the p-values are less for 'Reason for absence' and 'Disciplinary failure' variables. We get the frequency statistics of these variables and then plot as shown below. We see that the major frequencies contributing to absentees are reasons 23, 27, 28 i.e. medical consultation, physiotheraphy and dental consultation.




700
600
500
400
300
200
100
0
count
0.0
1.0
Disciplinary failure

```
Chi Sq test for features........... ID
889.2453819016475
3.815031774677936e-11
Chi Sq test for features........... Reason for absence
1656.032873284991
1.1805219347568775e-122
Chi Sq test for features........... Month of absence
359.71824497828754
2.8944754330752625e-09
Chi Sq test for features........... Day of the week
103.87669508720488
0.008287792011308939
Chi Sq test for features........... Seasons
122.5974320111169
3.0053076994296445e-07
Chi Sq test for features........... Transportation expense
697.0521607127151
9.406544679138365e-17
Chi Sq test for features........... Distance from Residence to Work
721.8301626052225
6.626733002732717e-17
Chi Sq test for features........... Service time
415.04671629021135
3.1920877098617734e-05
Chi Sq test for features........... Age
592.5256190584283
1.0231171439551832e-11
Chi Sq test for features........... Work load Average/day
960.8794673589287
4.611084344149994e-13
Chi Sq test for features........... Hit target
311.5771325961067
2.1957618620549616e-05
Chi Sq test for features........... Disciplinary failure
531.3246434435844
2.670951153834707e-101
Chi Sq test for features........... Education
37.16679804314187
0.9609122922050121
Chi Sq test for features........... Son
161.97098572180235
7.040447791252274e-09
Chi Sq test for features........... Social drinker
40.02404101447068
0.0020715847836446314
Chi Sq test for features........... Social smoker
```

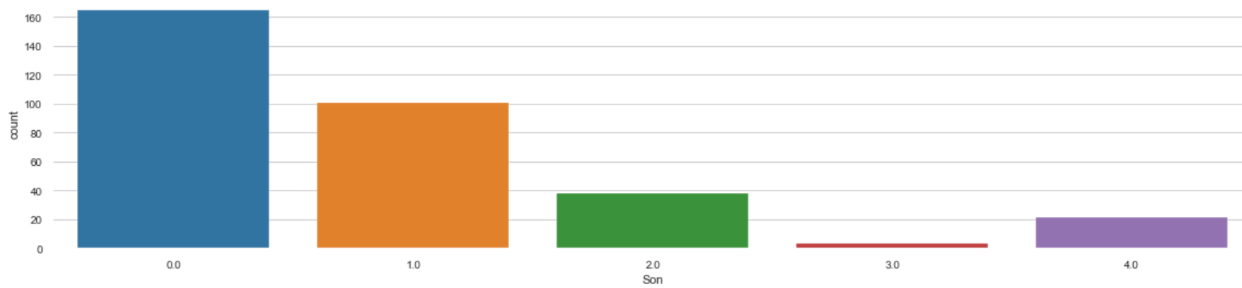
Figure 2.3: Chi Square analysis of 1st iteration

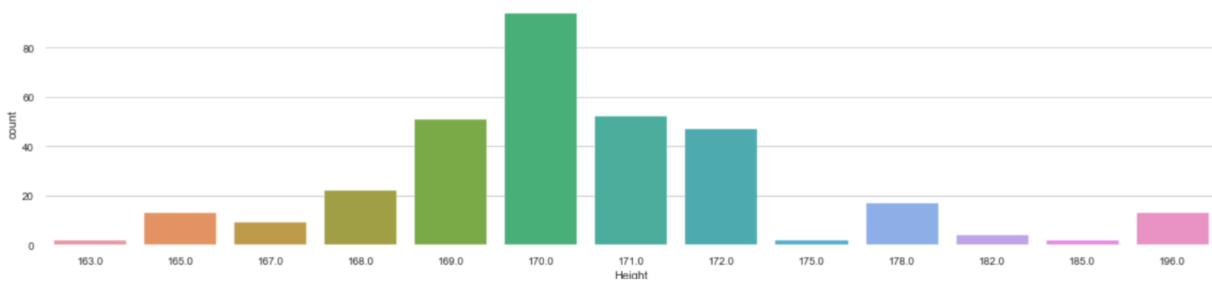
2nd Iteration

```
Chi Sq test for features........... ID
332.12499465433973
6.269475251322483e-07
Chi Sq test for features........... Month of absence
142.29129722332132
0.00022175251549340464
Chi Sq test for features........... Day of the week
34.704861371063146
0.3402020386876126
Chi Sq test for features........... Seasons
64.8447952055207
1.2826957190186773e-05
Chi Sq test for features........... Transportation expense
247.69397736079668
1.6038659403946965e-08
Chi Sq test for features........... Distance from Residence to Work
265.13558001976423
3.389668904893843e-07
Chi Sq test for features........... Service time
218.2666068694327
1.1492810417805194e-06
Chi Sq test for features........... Age
242.81609988283606
3.992498695754827e-06
Chi Sq test for features........... Work load Average/day
398.42892972840764
4.070080102567428e-06
Chi Sq test for features........... Hit target
168.86593903122395
6.381792655333278e-06
Chi Sq test for features........... Son
124.47161226449576
7.628603607064547e-13
Chi Sq test for features........... Social drinker
18.027111754333465
0.021024141460881326
Chi Sq test for features........... Pet
32.544233040452454
0.4399935549519931
Chi Sq test for features........... Weight
264.71097198879613
1.719527161186957e-05
Chi Sq test for features........... Height
207.87315521620843
3.103467557585875e-10
Chi Sq test for features........... Body mass index
134.629689302162
0.07151888246125382
```

Figure 2.4: Chi Square analysis of 2nd iteration

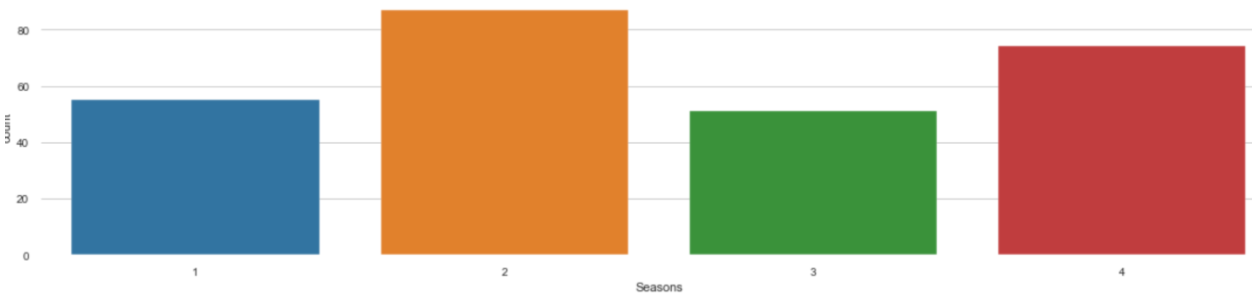
Fig 2.4, shows the chi square statistics of 2nd iteration with more significant and less significant variables dropped from 1st iteration and with a reduced dataset. From this analysis we find that the more significant variables are Son and Height. While plotting the frequency distribution of both these variables we find that candidates with number of Son equal to 0 or 1 are contributing to more number of absentees. Also the Height values in 170 or 171 are contributing to more number of absentees.





3rd Iteration

Fig 2.5, shows the chi square statistics for variables after removing the more/least significant variables from the earlier 2 iterations. We found that 'Work load Average/day' and 'Seasons' has very low p-values compared to other variables. We pick those variables and find their frequencies as below.




```
Chi Sq test for features........... ID
160.14696733964365
0.07702939692007128
Chi Sq test for features........... Month of absence
129.95303996966217
0.0024469776330491608
Chi Sq test for features........... Seasons
62.58652529673983
2.7310607157995167e-05
Chi Sq test for features........... Transportation expense
94.05365786096962
0.13481098400624203
Chi Sq test for features........... Distance from Residence to Work
135.867906613664
0.15270320074873878
Chi Sq test for features........... Service time
114.43414501918697
0.22768907060147142
Chi Sq test for features........... Age
104.9306895868834
0.2503704055396815
Chi Sq test for features........... Work load Average/day
392.1338705699634
2.4346970652230627e-06
Chi Sq test for features........... Hit target
151.54678143747674
0.000260848793349091
Chi Sq test for features........... Social drinker
12.76251183205612
0.12028627152459102
Chi Sq test for features........... Weight
143.59727302810728
0.06999475993823973
```

Figure 2.5: Chi Square analysis of 3rd iteration

We find the following values for 'Work load Average/day' had more frequencies than other values

308593.0 264249.0 222196.0 251818.0

Also Season 2 (Autumn) contributed to more number of absentees than other seasons

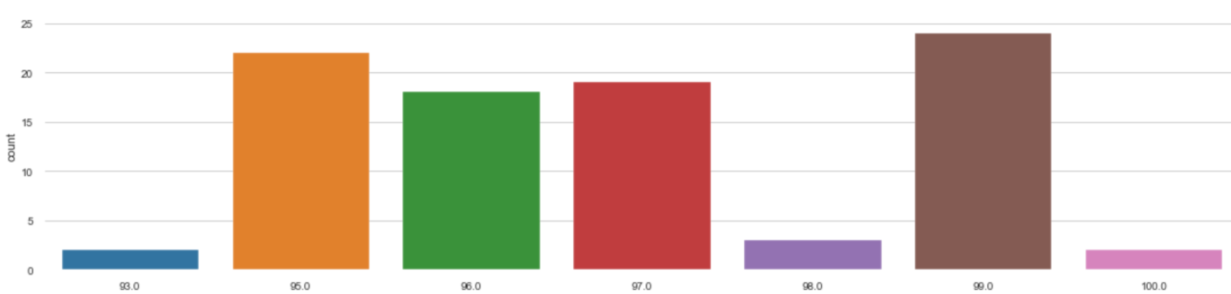
4th Iteration

Fig 2.6, shows the chi-square statistics of the last iteration, we had filtered many variables in the earlier iteration and this iteration we had only least significant variables within the dataset. We found in this iteration that 'Hit target' has a least p-value compared to other predictors.

```
Chi Sq test for features........... ID
75.46066636851522
0.08607643619703208
Chi Sq test for features........... Month of absence
18.547700161130894
0.42015756388824266
Chi Sq test for features........... Hit target
110.0194055944056
2.0035204983634793e-09
Chi Sq test for features........... Weight
75.4606663685152
0.08607643619703223
```

Figure 2.6: Chi Square analysis of 4th iteration

We then plot the frequency plot for Hit target as below and the values within Hit target of 95,96,97,99 contributed to more number of absentees.



Chapter 3

Conclusion

3.1 Solutions of Problem Statement

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

Solution:

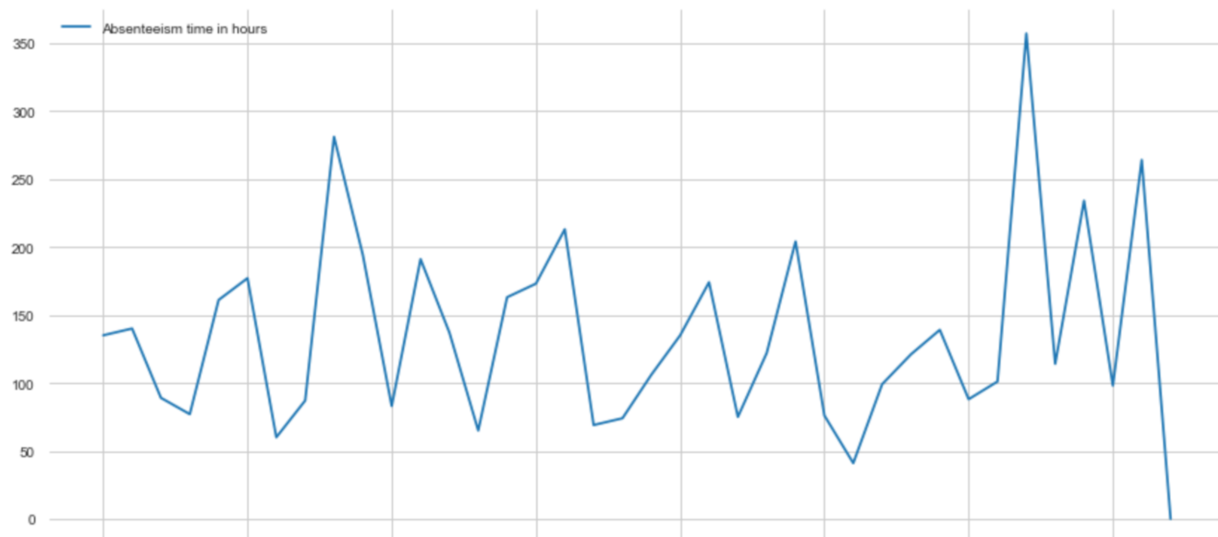
In our case of Employee Absenteeism Data, we found that the significant reasons used by employees are related to consultation and physiotheraphy. The company can arrange for inhouse regular health checkup to their employees.

Employees with less than 2 Son's are having more absentees than others, which gives indication that employees with no children will not have to deal with school days of their children and they will go on leave as they like. For employees with 1 Son considering Son of small age, employee could have availed more leaves to look after their children. It would be better if company could provide day care facility to their children nearby. It was observed that Autumn season contributed to more number of absentees. The company could warn their employees to not take more leaves during autumn. Also company could reduce the Hit targets to less than 95, to avoid more number of absences.

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

Solution:

The year 2011 does not explain anything within the problem statement or the data. If the same trend of absenteeism continues, then the total losses for a profile consisting of month, workload average, Hit target is as shown in the graph below. Employees are absent the most in the month of March with Work load of 222196 and Hit target of 99, with total Absenteeism hours equal to 357 hours. Employees are absent the least in the month of Septemeber with WorkLoad 261756 and Hit target of 87, with total Absenteeism hours equal to 41.



References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 6. Springer.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer Science & Business Media.