Introduction

What is the reason for workers being absent?

According to Forbes, Absenteeism is an employee's intentional or habitual absence from work where excessive absences can equate to decreased productivity and can have a major effect on company finances, morale and other factors.

A list of reasons that can be a cause of an employee being absent can be -

1. Childcare
2. Illness
3. Harassment at workplace
4. Stress

There can be no number of reasons for absenteeism depending upon various individuals. This can have adverse consequences like

1. Decreased productivity
2. Poor morale among co- workers
3. Increased labor costs
4. Poor customer service
5. High administration cost

Data Exploration

The Dataset consists of total 740 rows having 21 features each. These can be further classified into following types:

1. In Categorical Features
    - Reason for absence
    - Day of the week
    - Disciplinary failure
    - Social drinker
    - Month of absence
    - Seasons
    - Education
    - Social smoker
2. In Numerical Features
    - ID
    - Distance from residence to work
    - Age
    - Transportation expense
    - Service time
    - Work load Average/day

- Hit target
- Pet
- Height
- Son
- Weight
- Body mass index

Expected output: Absenteeism time in hours

Data Pre-Processing

To perform the data analysis and build a good model we should have a good knowledge of the data. In the given data set we can see:

1. We can see that the “month of absence” column consists of some invalid data i.e. month data 0. And from the given column definition we know that the number of months can be from 1-12. So, we can remove those rows.
   After removing those rows 737 rows are left.

2. If we assume that the disciplinary failure = 1 means that the person is necessarily absent due to some discipline issue. The count of such rows is 40.
   And if we count the no of rows where 'Absenteeism time in hours' is zero, it comes out to 41. So, we can consider that the value of that 1 row is inconsistent and is not correct. So, we will replace the value of that row with the mean of the 'Absenteeism time’ of that id.

3. We assumed that disciplinary failure = 1 means that the employees were asked to be absent. So, in such cases (40 rows) we can replace the value 0 with 8.

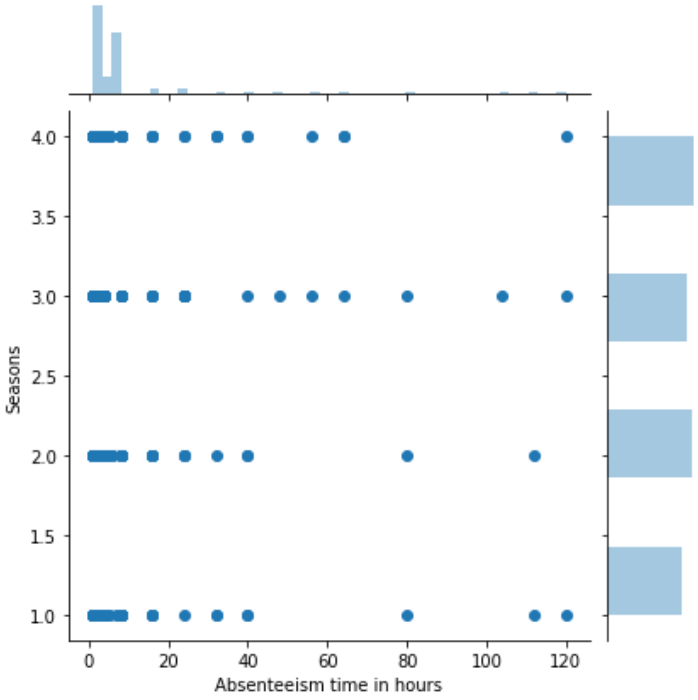
4. We know that the BMI is calculated on weight and height and since this is a derived column, We can remove weight and height columns since they information is already captured in BMI

After doing the above pre-processing we have filtered our data and replaced all the inconsistent values. We are now ready for the for our data exploration and model building.

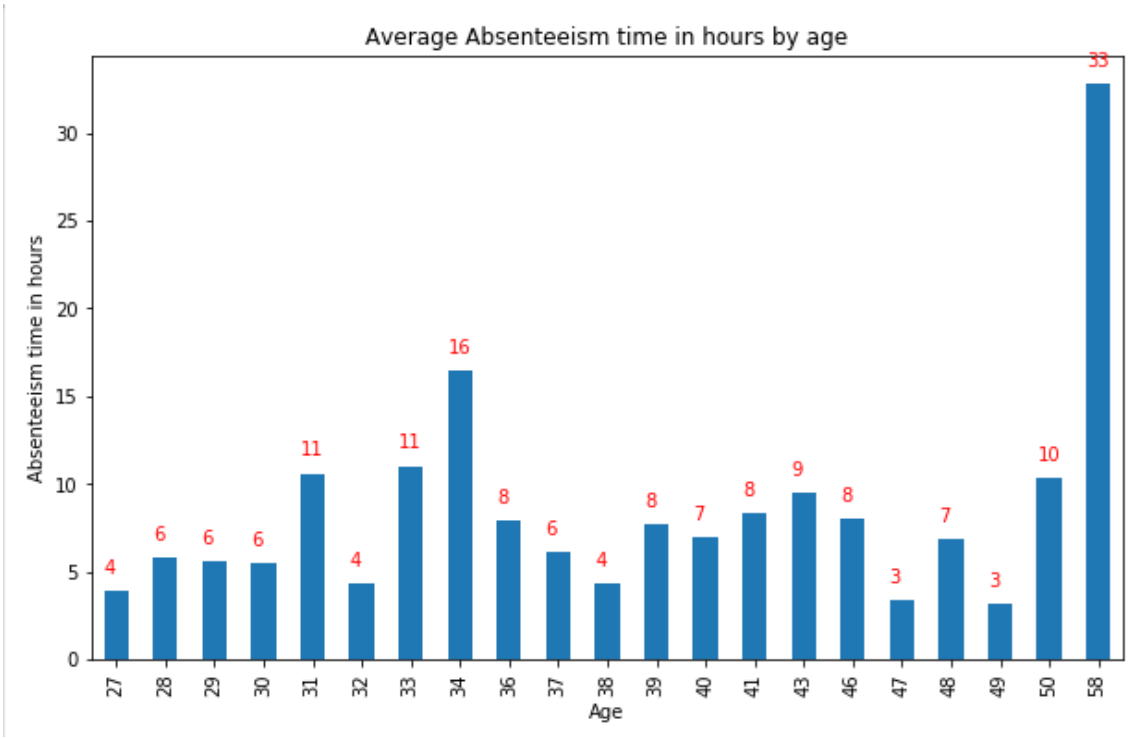
Data Analysis

Let’s have a closer look for a few columns:

1. **Seasons:**
   All the employee data is present over 4 seasons. If we plot a joint-plot between 'Absenteeism time in hours' and ‘Seasons’, we can see that the data is evenly distributed over the all four seasons. So, we can’t conclude any results from this.

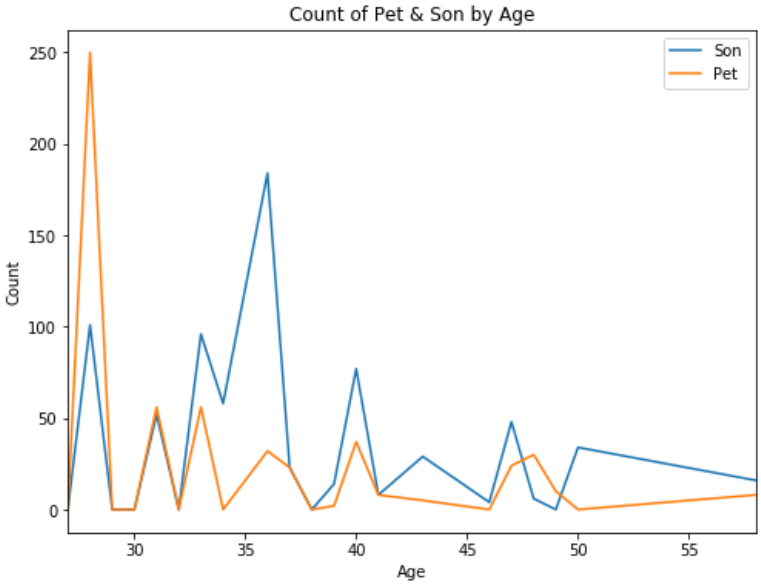


2. **Age**
   For this column, I am trying to gather the type of employees who work at this company, is there a younger generation or is there a range?



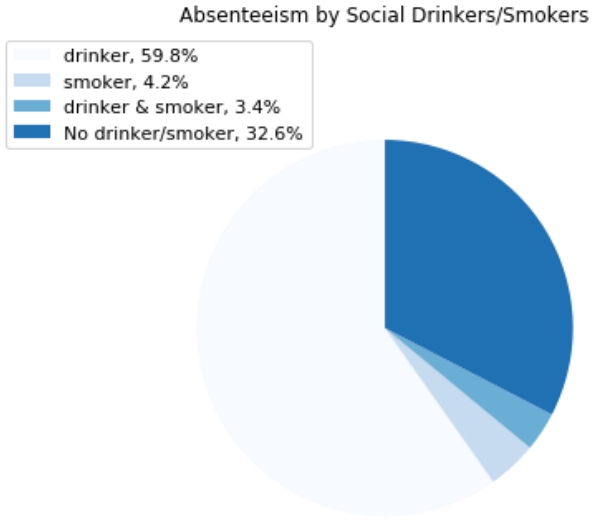
From the plot we can see that most workers are in between their late 30's and early 60's.

3. **Son and Pet**



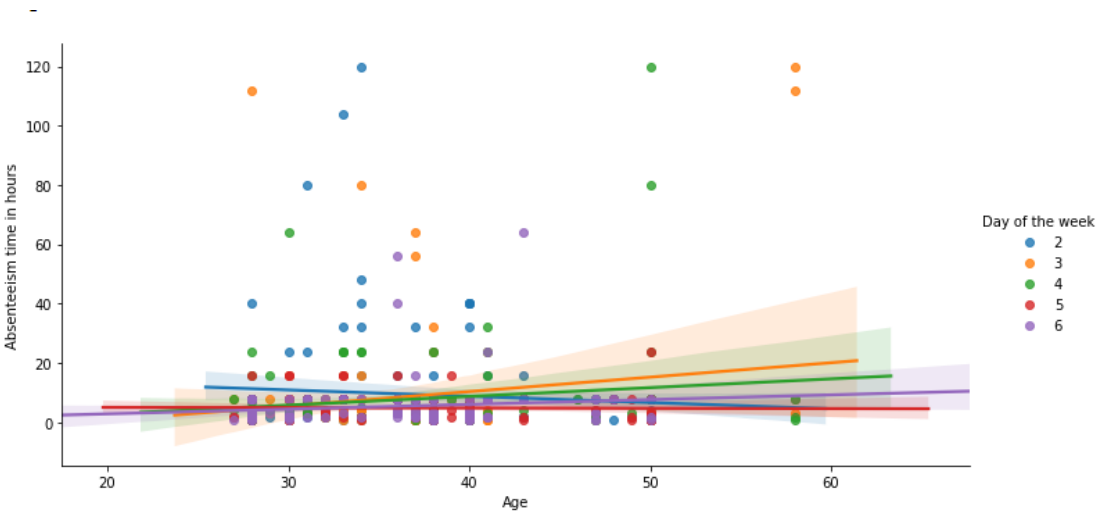
When we plotted a chart between the count of 'Son and Pet' vs the age. We get an interesting analysis that most workers till the age of 40 have a higher pet and son count.

4. Social drinking and smoking



If we plot a chart for the social drinking and smoking. We see that the percentage of workers who drink more are more likely to be absent.

5. Day of the week



From the above plot we can see that most of the workers are absent on the second day of the week and are in the age group of 30-40.

6. Reason for absence

Another column of importance is Reason for absence with its 28 levels where each number represents a reason given by an employee for their absence. The column originally had categorical input that was converted into numerical input.
Most common reasons are the categories outside of the ICD (International Code of Diseases) meaning that the most reasons were non serious diseases such as:

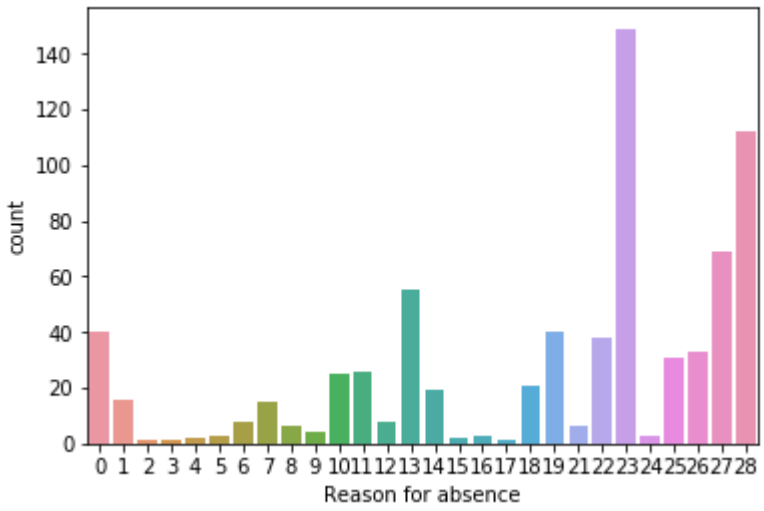
- Medical consultation (23) - 149
- Dental consultation (28) - 112
- Physiotherapy (27) - 69

These 7 categories are used by **45% of employees**.

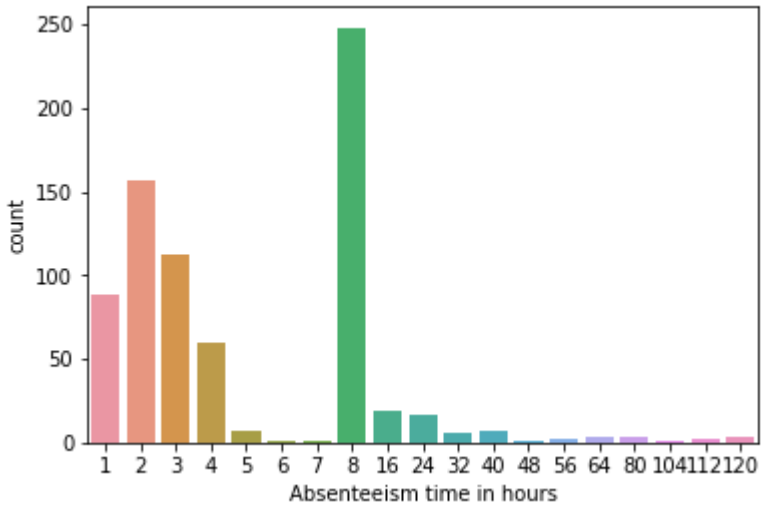
For the ICD diseases or serious disease, the main 2 categories with relatively high numbers Are

- Diseases of the musculoskeletal system and connective tissue (13) – 55
- Injury, poisoning and certain other consequences of external causes (19) – 40

These two seem like they could be related to injuries one gets from this type of industry. For example, someone who is always doing deliveries for a long period of time is more likely to have musculoskeletal issues. It is surprising how only a few absences are related to birthing or children related.

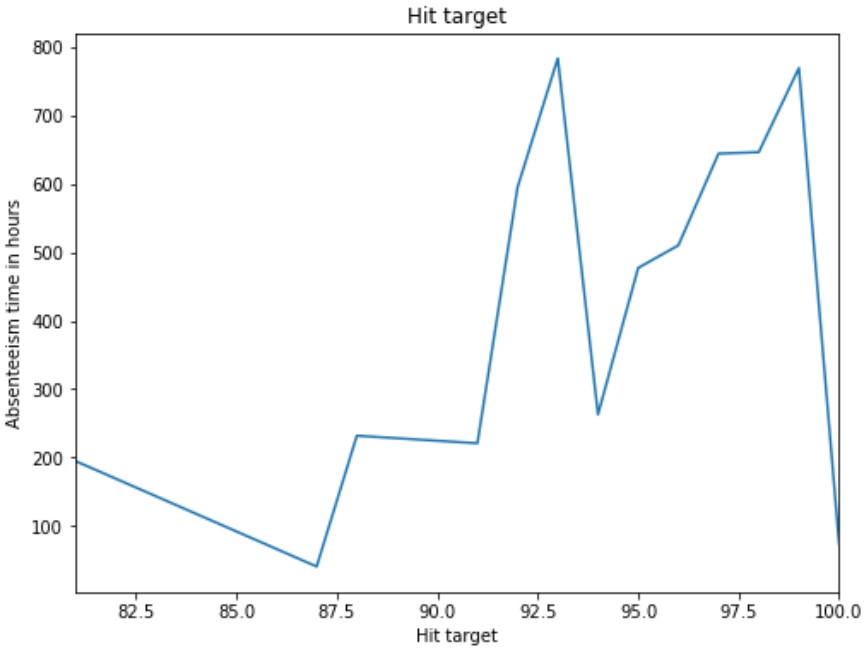


7. Absenteeism time in hours



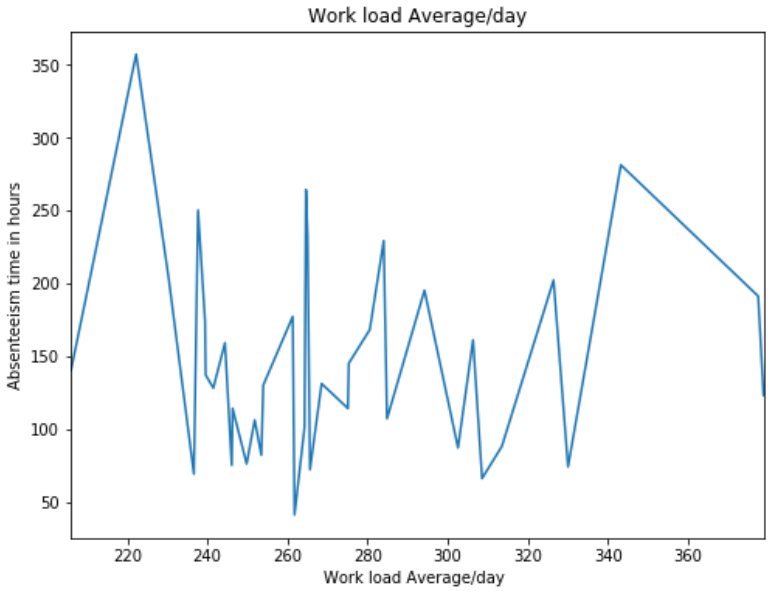
From the above chart we can clearly see that most of the employees have taken only one day leave. Few of them are also the case of disciplinary failure.
Other than these workers, most frequently absent in the bin of 1-4 hrs. and these cannot be considered as absenteeism.

8. Hit Target



Above plot indicates Higher the hit target higher the absenteeism hours.

9. Work Load Average



Above plot indicates higher the work load average higher the absenteeism time in hours.

10. Co-relation between columns

From the above correlation analysis, we cannot find any pattern which can lead to absenteeism.

The above analysis shows how both Age, hit target and Work load Average/day have high absenteeism time in hours.

Model Building and Evaluation

1. We need to predict the value of absenteeism in hrs. We first tried solving this problem using using regression.

   We split the dataset into two parts:
   1. Training Dataset - Containing 70 % of total records.
   2. Test Dataset – Containing 30 % of total records.

We used 2 regression models:

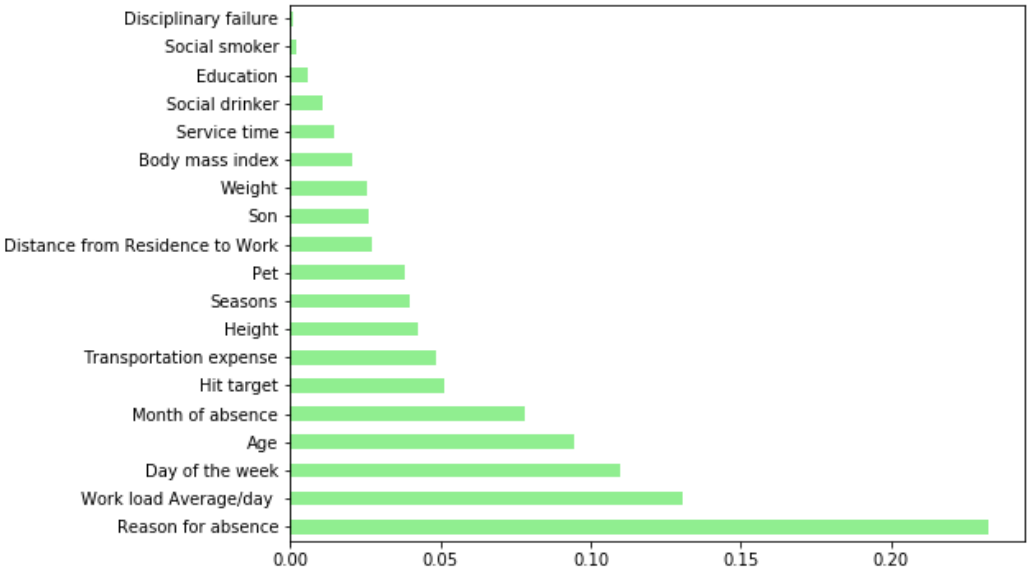
a) RandomForestRegressor
b) Linear Regression

We first used RandomForestRegressor to predict the values of Absenteeism.

When we verified the output of the test data following was the output for various methods:

Test data R-2 score: 0.125
Test set of RMSE: 10.89

Accuracy for the predicted data is not up to the mark. And on checking the feature importance of each feature we got to know that it is giving more importance to “Reason of absence” than any other feature. This behavior is not correct as it is giving higher priority to one and neglecting all other features.



Next we tried linear regression.
Using box plot we found that absenteeism hours > 24 hours were outliers. We removed these records and found that the RMSE value improved significantly. However R-2 score is still pretty low.

Test R-2 score: 0.125
Test RMSE : 3.5000

We also scaled the all the attributes 'StandardScaler' library
We also converted the categorical attributes into encoded attributes using one hot encoding however that did not improve the score further.

2. Since overall score with regression turned out to low even after using various techniques to optimize the model we decided to convert this problem into a classification one.

   First, we divided the absenteeism time into bins. For that we will create 4 bins:

   Bin 0: <= 4 hrs
   Bin 1: > 4 and <=8 hrs
   Bin 2: >8 and <=24 hrs
   Bin 3: > 24 hrs

   We have used this partition because:

- we found that, majority of the hours are less than 24.
- Also for business perspective absenteeism can be meaningfully classified as half a day or less, full day, 2-3 days and more than 3 days.

   We tried the following 3 models for classification:

a) RandomForestClassifier
b) Naive Bayes
c) Decision tree

   Following is a comparison of the 3 classification models in terms of accuracy and f1-score

	RandomForestClassifier	Naive Bayes	Decision Tree
Overall accuracy	0.73	0.61	0.77
<= 4 hrs f1-score	0.83	0.79	0.85
> 4 and <=8 hrs f1-score	0.64	0.29	0.72
>8 and <=24 hrs f1-score	0.15	0.22	0.00
> 24 hrs f1-score	0.40	0.15	0.00

The best accuracy we have achieved is using decision tree classifier (77%). Hence we recommend decision tree classifier model for predicting absenteeism of employees.

Conclusion

This analysis was done to sort out a method to reduce workers' absenteeism time.

According to the observed pattern, it was noticed that workload, hit target and doctor's consultations were the main reasons of remaining absent.

To ensure higher presence rate the following steps can be taken-

- **Flexible schedule**.
  Since medical consultations are high, giving an option for a flexible schedule where the few hours lost can be compensated by either coming in early or leaving late, depending on the employee's role.

- **Employee Wellness program**.
  For example, the amount of Work load can lead to absenteeism due to stress. Relaxing activities during lunch breaks targeting specific muscles for a delivery employee not only reduce stress, potential health problems but also increase employee morale.