Employee Absenteeism

Gaurav Malik

Contents

Introduction

1.1 Problem Statement

1.1.1What is Absenteeism?

Absenteeism is a regular pattern of absences from work or obligation without any good reason. This is a rather difficult problem to address as you cannot force your employee to show up to work on time. We could just understand the pattern and find out reasonable policies to reduce absenteeism. As to why it is necessary, the company in the end has to face loss because of this habit.

In this report we would be making various model prediction pick the best one and also answer two questions that are demanded by the client. Those questions are:

- 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 Interpretation

The data has 21 columns observed by the company and the target feature is the Absenteeism time in hours. As the target variable is continuous, regression would be better suited.

The data consists of

COLUMN 1-13

ID	eason for absenc	Month of absence	Day of the week	Seasons	nsportation exper	e from Residence	Service time	Age	rk load Average/d	Hit target	Disciplinary failure	Education
11	26	7	3	1	289	36	13	33	239554	97	0	1
36	0	7	3	1	118	13	18	50	239554	97	1	1
3	23	7	4	1	179	51	18	38	239554	97	0	1
7	7	7	5	1	279	5	14	39	239554	97	0	1
11	23	7	5	1	289	36	13	33	239554	97	0	1
3	23	7	6	1	179	51	18	38	239554	97	0	1

COLUMN 14-21

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

There are total 21 columns in the dataset consisting all the information of the customer and even whether the customer churned or not.

The description of the columns are provided below:

```
ID                                740 non-null int64
Reason for absence                737 non-null float64
Month of absence                  739 non-null float64
Day of the week                   740 non-null int64
Seasons                           740 non-null int64
Transportation expense            733 non-null float64
Distance from Residence to Work   737 non-null float64
Service time                      737 non-null float64
Age                               737 non-null float64
Work load Average/day             730 non-null float64
Hit target                        734 non-null float64
Disciplinary failure              734 non-null float64
Education                         730 non-null float64
Son                               734 non-null float64
Social drinker                    737 non-null float64
Social smoker                     736 non-null float64
Pet                               738 non-null float64
Weight                            739 non-null float64
Height                            726 non-null float64
Body mass index                   709 non-null float64
Absenteeism time in hours         718 non-null float64
```

Total 21 columns with missing data.

Methodology

2.1 Pre Processing

Data Pre Processing is just as important as creating a machine learning model. Before using the models of Machine Learning it is very essential that we first make the data which can be better suited to the algorithms. This process mostly includes cleaning the data, look for missing values, handling categorical data, looking for outliers. During our pre processing we also need to need to plot various bars and plots to look for outliers and the distribution of the data too.

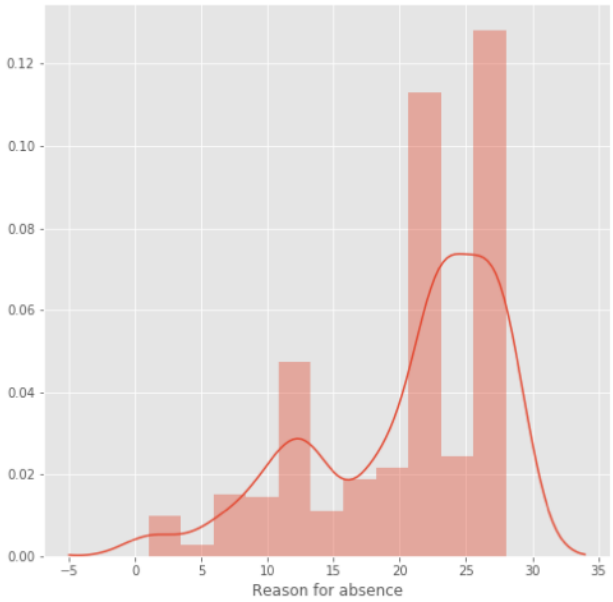
2.1.1 Imputing NA Values

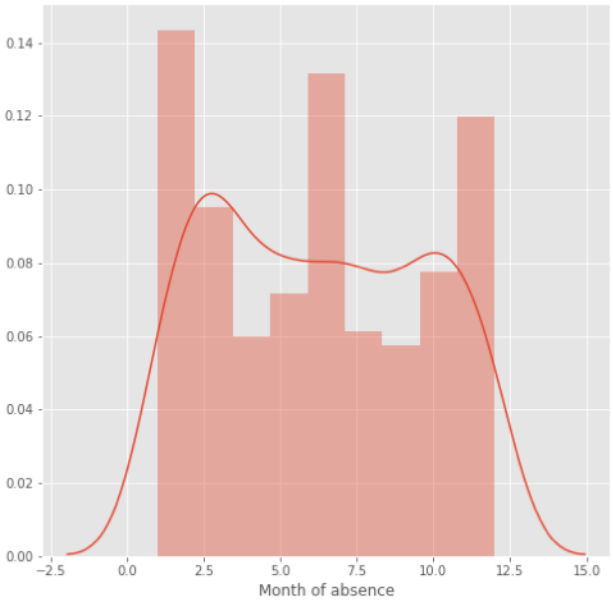
As the data contained a lot of null values, first remove all the null values which are in our target variable. After removing all the null target variable we can impute the missing values in our independent variables. This time KNN imputation was used as it is a very popular method. This is because KNN imputation uses other features of the dataset as well while computing rather than using mean or median strategies to fill the missing values.

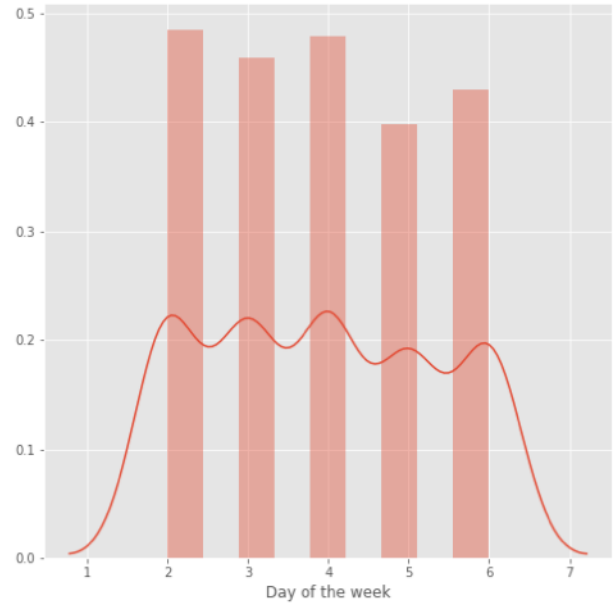
2.1.2 Distribution Histograms

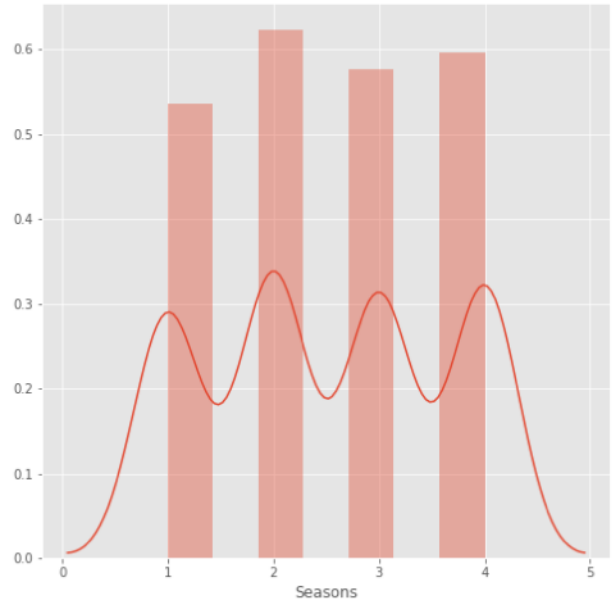
While we do our pre processing it becomes very important to observe the distribution of data. To at least have a sense of the data distribution in our dataset we must have an insight of the data distribution and we do this by plotting histograms of the data features.

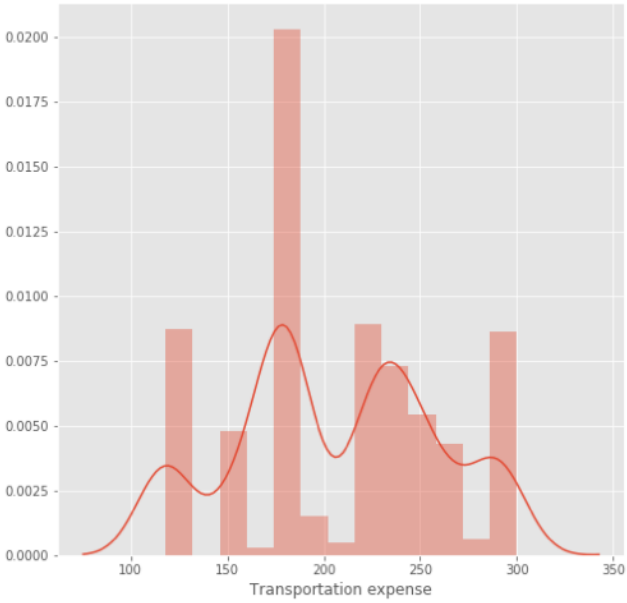
Following are the histograms of columns indicating their distribution

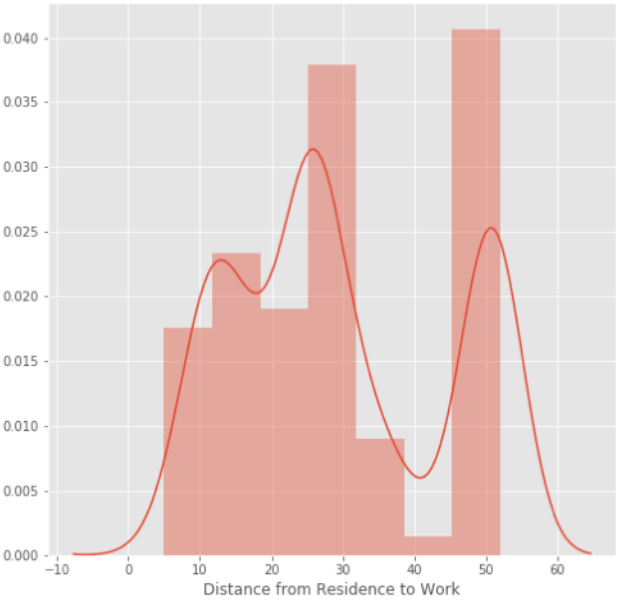

Reason for absence

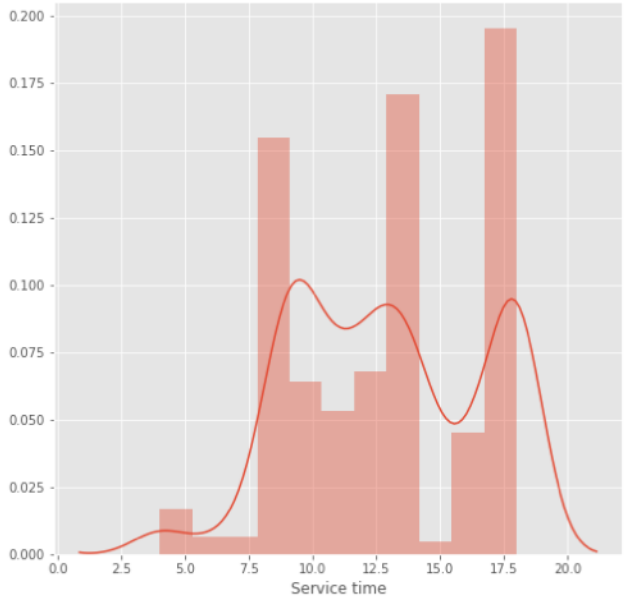

Month of absence

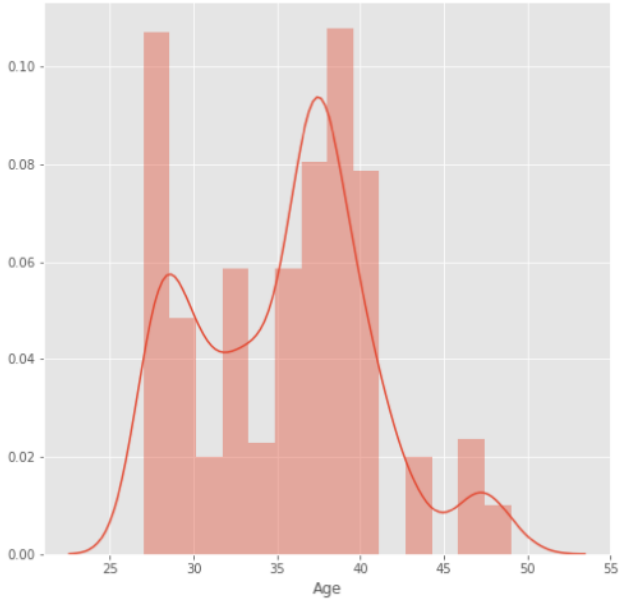

Day of the week

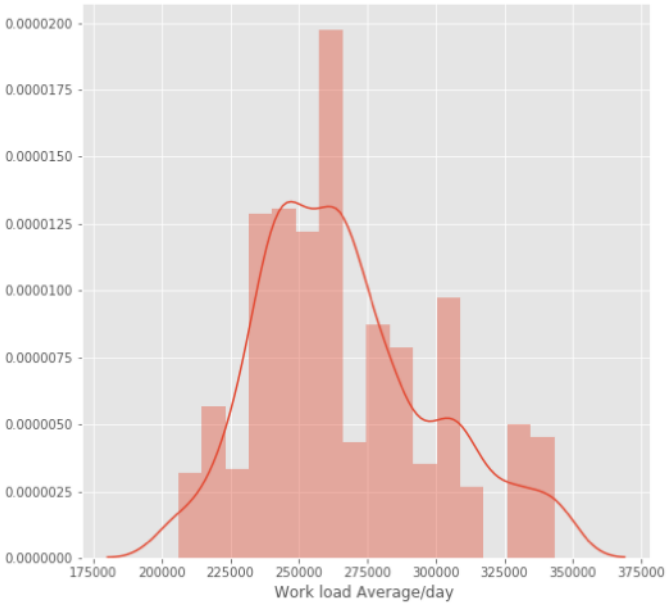

Seasons

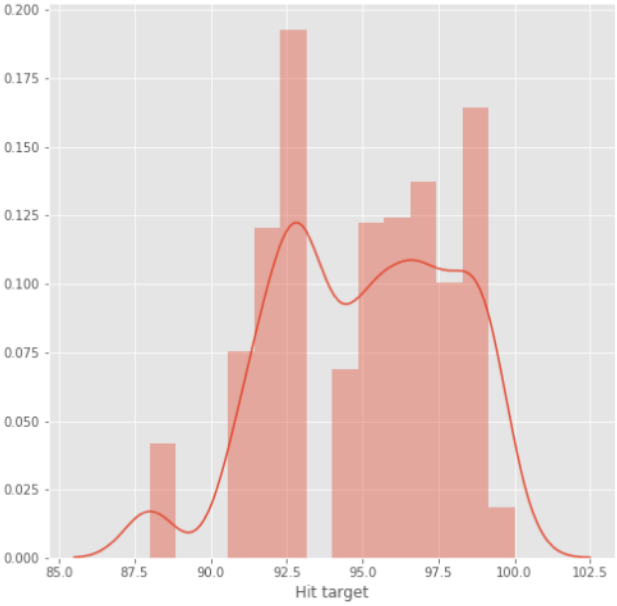

Transportation expense

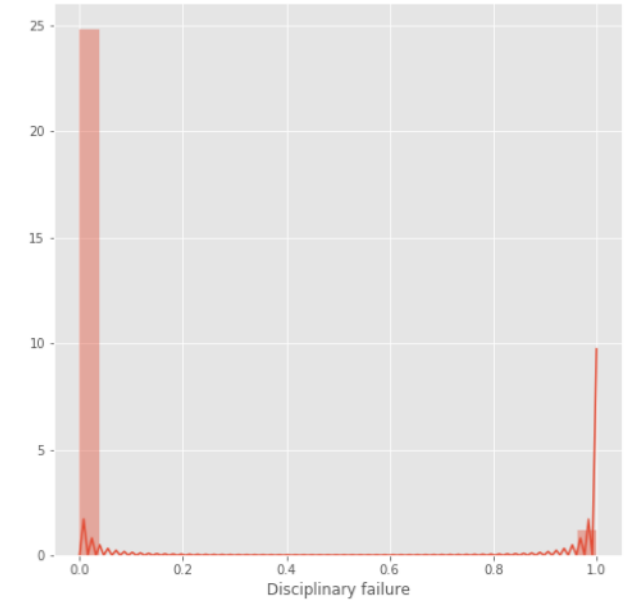

Distance from Residence to Work

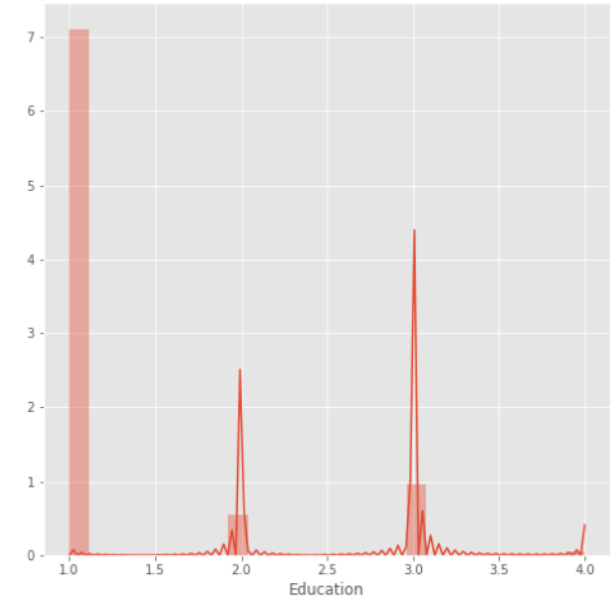

Service time

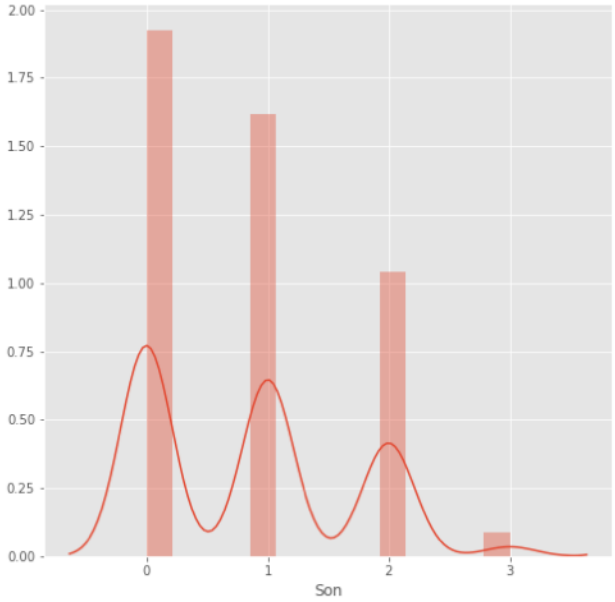

Age

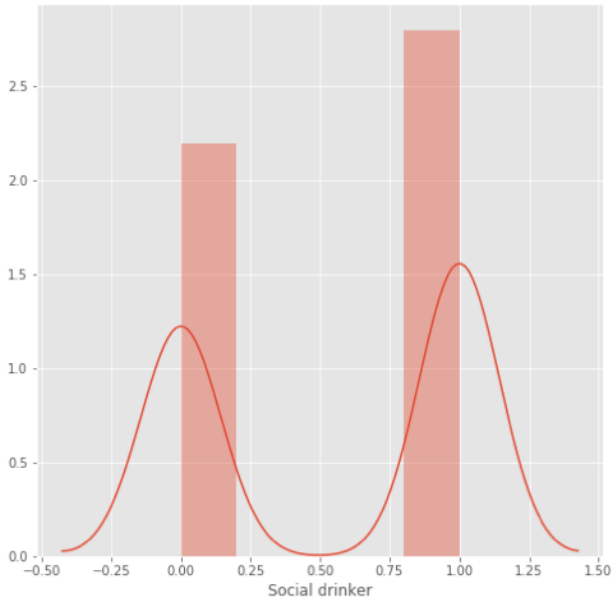

Work load Average/day

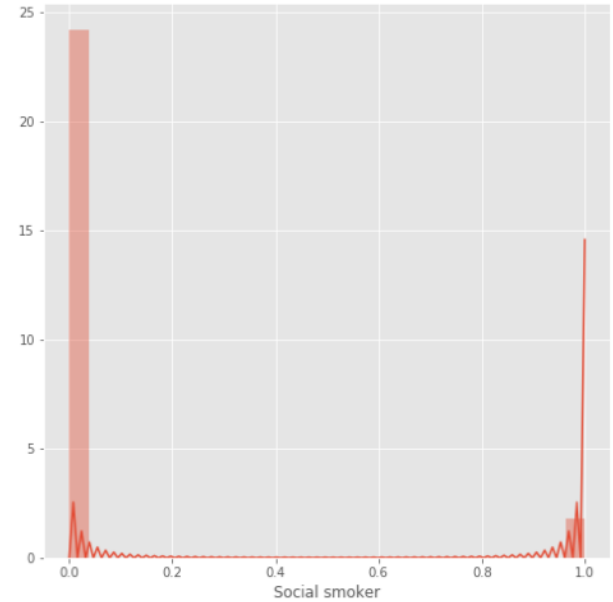

Hit target

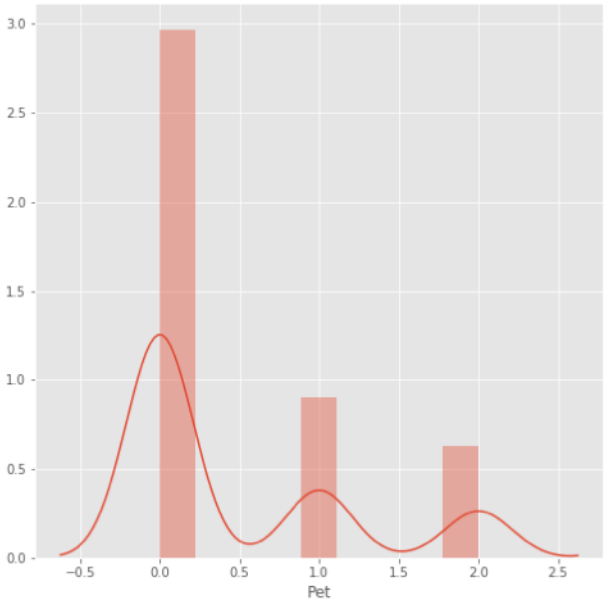

Disciplinary failure

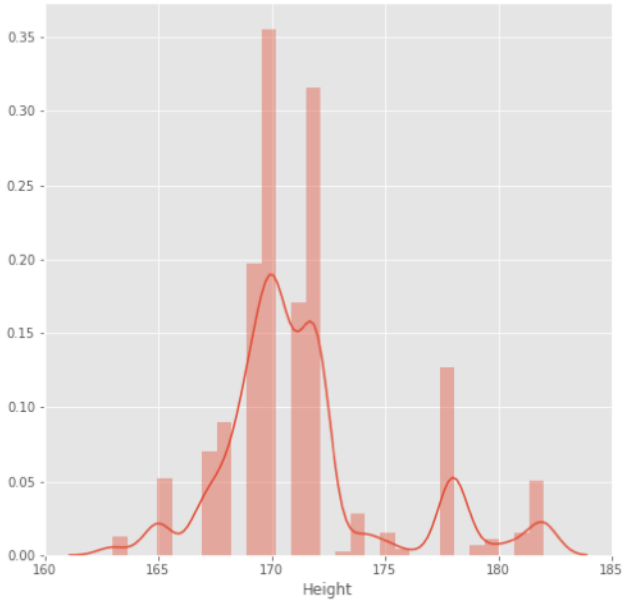

Education

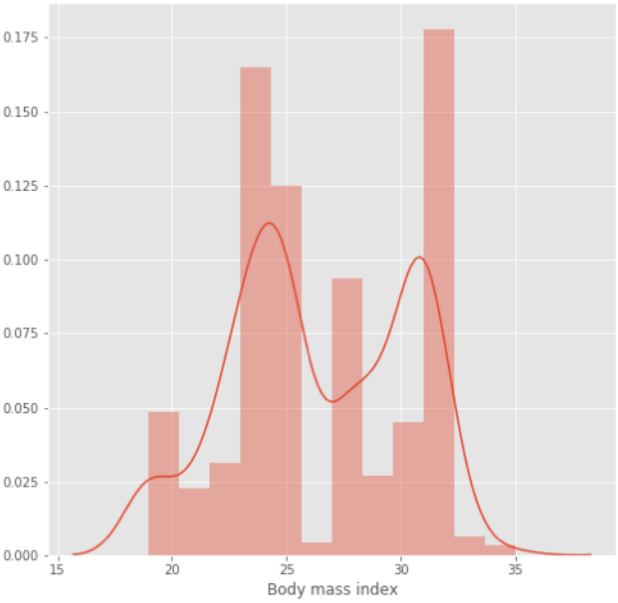

Son

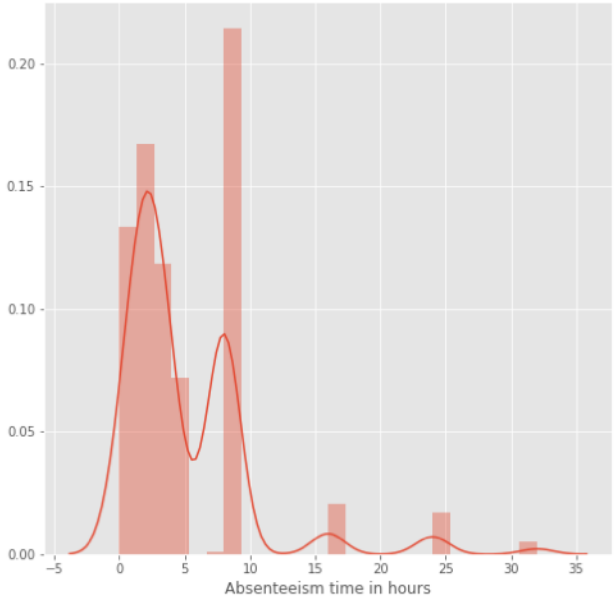

Social drinker


Social smoker


Pet


Height


Body mass index



From the above data distribution plots we can observe that in:

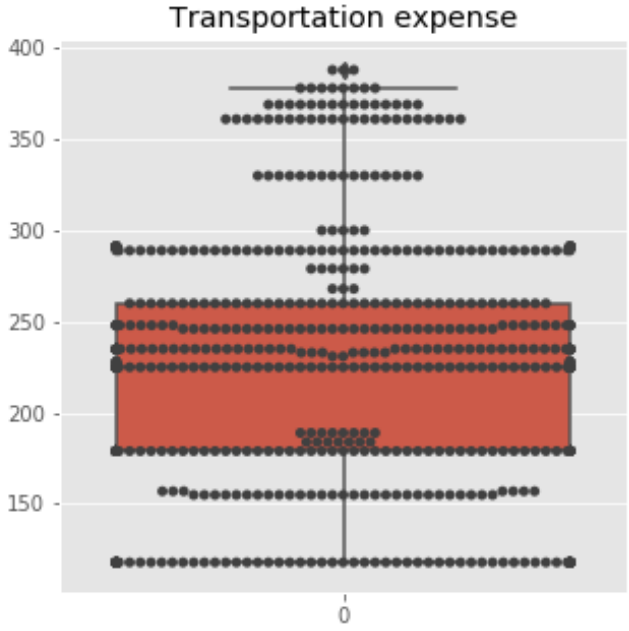
- The reason for absence is mostly due to medical purpose.
- Those whose Disciplinary Failure is False or we could say that those who are disciplined are more inclined to be absent from their work without any proper reason.
- Those who are less educated also tend to give less attention to their work.

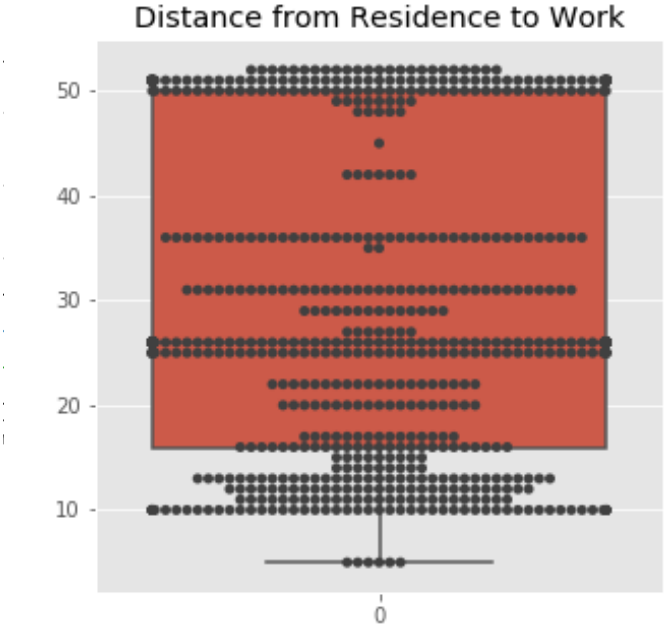
The data is almost normally distributed but there are some continuous variables like features including Son and Pet columns that are have some skew.

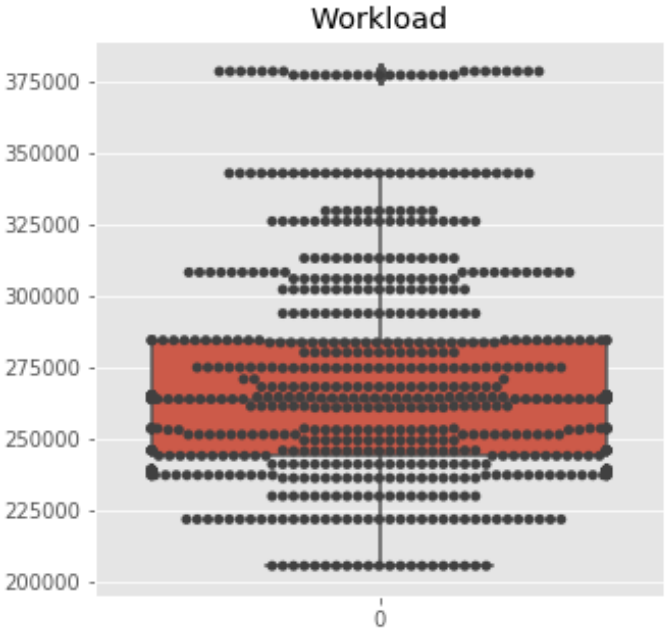
2.1.3 Outliers Analysis

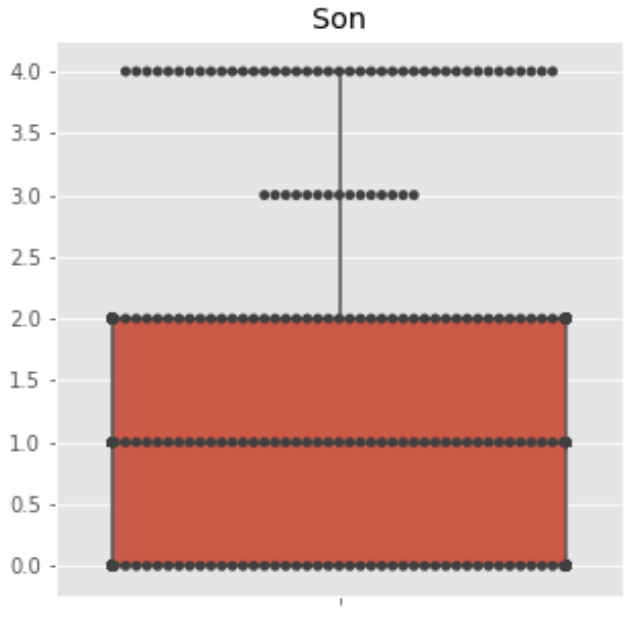
When we are cleaning our data, analysing outliers becomes among essential things to do. With the help of outliers box plot we can see whether the data which is too far from the rest of the data points will affect the prediction and remove them replacing with better values like mean of the data.

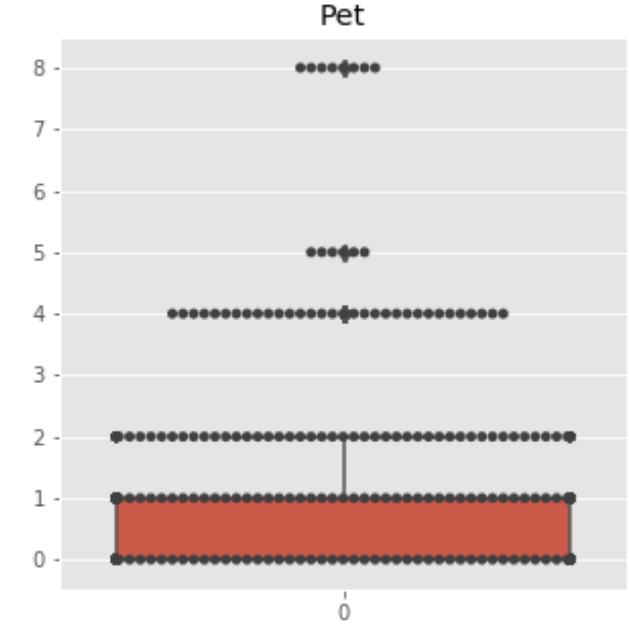
Following are the box plots for the data we have on customers:

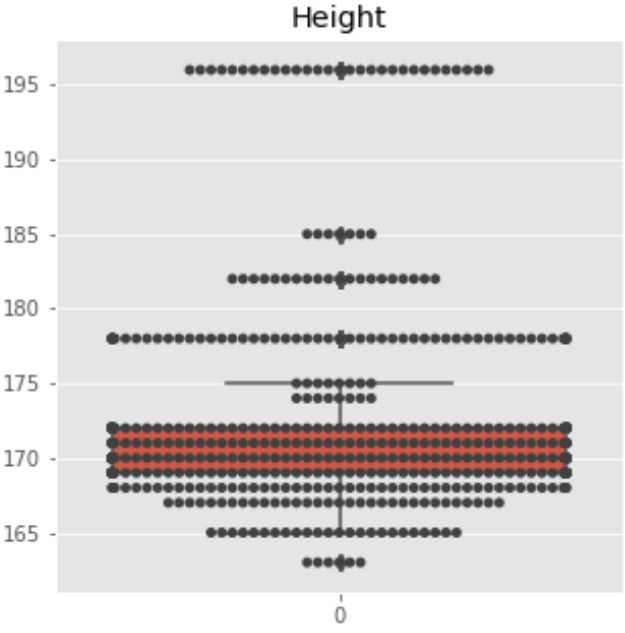

Transportation expense
400
350
300
250
200
150
0

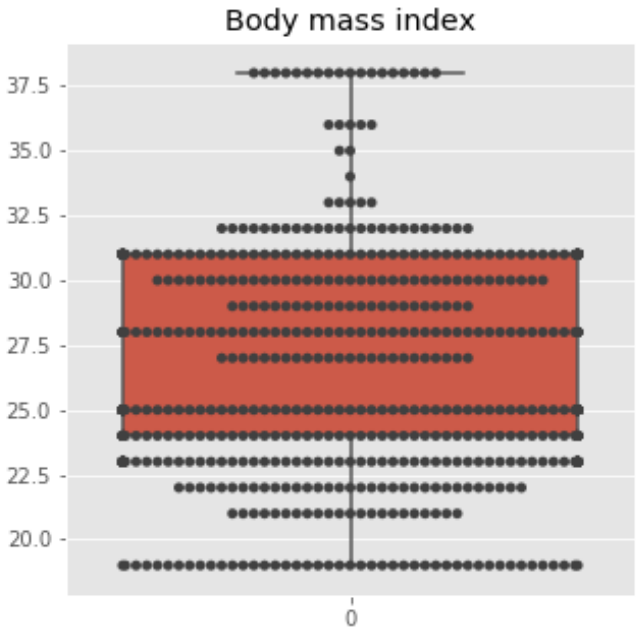

Distance from Residence to Work
50
40
30
20
10
0

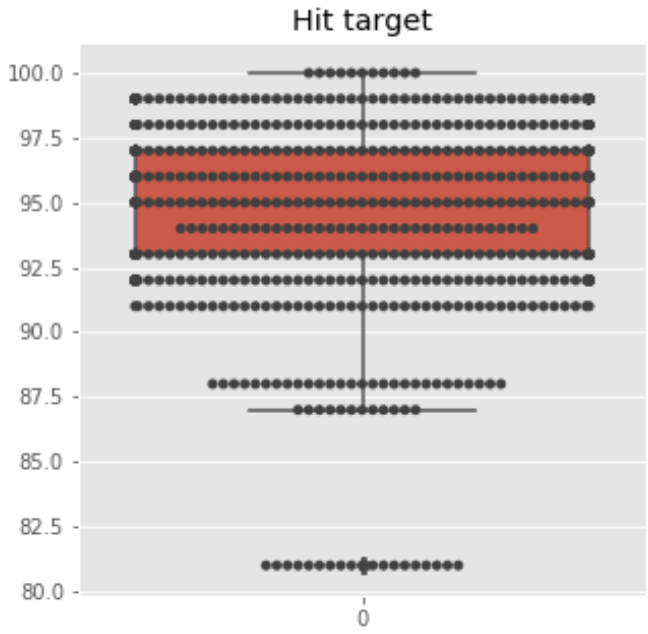

Workload
375000
350000
325000
300000
275000
250000
225000
200000
0

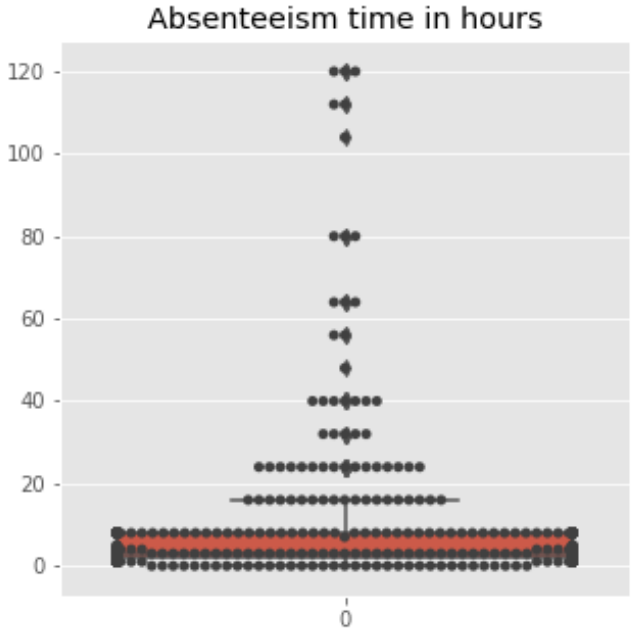

Son
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0


Pet
8
7
6
5
4
3
2
1
0
0


Height
195
190
185
180
175
170
165
0







As the box plot interprets that many of the numerical features contains outliers. We have to see which outliers are to be excluded and which not to exclude. I myself don’t generally exclude outliers as these are very special data points which even though are out of the box but still they give you an entire new perspective on the data.

Hit target is not one of the features that would not be included as the outlier is 0% which is one the required insight in the dataset so it won't be included for the removal of outliers. Rest of the features would be normalised to certain extent.

2.1.4 Feature Selection

We have looked upon our data, got a general hint of what it includes, how it is distributed. During data pre processing there comes a problem of whether to select all the features of the data or not. At times the features are all independent of each other but sometimes they are not. So, we have to filter out those highly correlated features.

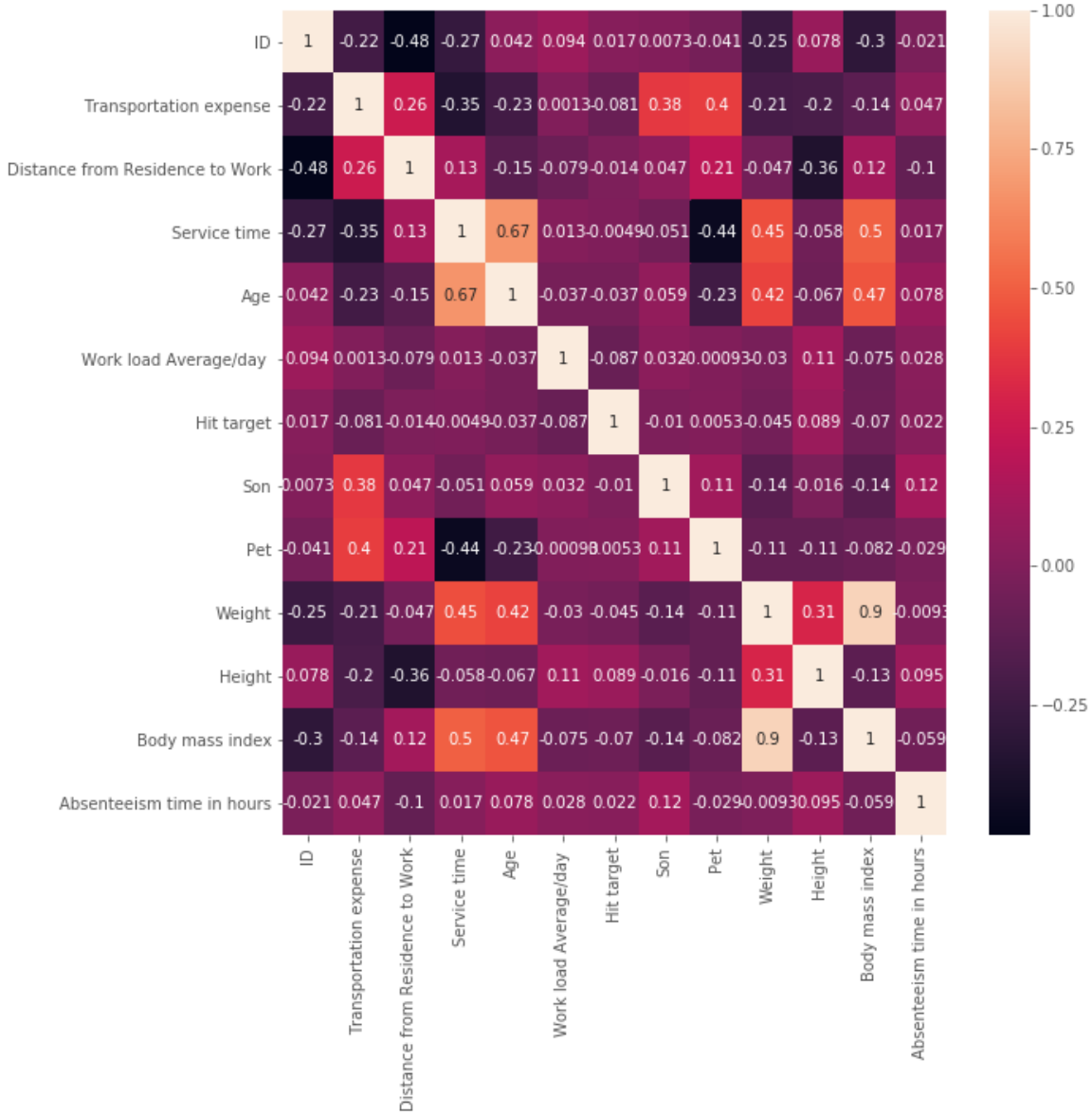
Correlated features in general do not improve the models. There are various benefits to removing the correlated features-

- Making the learning of algorithms faster
- Decreases the biasing in data
- Make the model much simpler and interpretable

For linear models like linear regression, logistic regression multicollinearity can make the prediction highly unstable. For models like Naive Bayes actually benefits from "positive" correlated data and for random forest regression the benefits are indirect as random forest is good at detecting interactions between different features.

So, removing these features can at times be necessary to speed up the learning. As the aim of the data scientist is to make the data interpretable it becomes essential to make the data simple.

Following shows the heat map of our data:



From the heat map we can observe that there are some highly correlated features which needs to be removed:

- Age is correlated to Service Time
- Service Time is correlated to Weight & Body Mass Index
- Weight is highly correlated to Body Mass Index

With this our dataset has features which are not highly dependent and this would make our model fitting to the data perfect and simple.

2.1.5 Encode Categorical Data

What is Categorical Data?

Categorical data are variables that contains labels as their value instead of having numbers.

Problem With Categorical Data

Many machine learning algorithms requires input and output variables to be numeric. They cannot work on categorical data directly, unlike decision tree which can work on categorical data without any problem.

In order to avoid these problems it is preferred to convert all categorical data to numeric form. In case the output variable is categorical variable too, we have to convert it back to categorical variable to present them or to use in applications.

Different Methods To Encode Categorical Data

We can encode categorical data using either of the two steps:

- Integer Encoding
- One-Hot Encoding

In integer encoding each unique categorical variable gets a unique number. This is also called label encoding.

For example in a data set we have three cities: New York, Mumbai, Moscow

In label encoding each of the variables get a number. Like, “New York” would be 1, “Mumbai” would be 2, and “Moscow” would be 3.

With label encoding there comes an uninvited guest called natural ordering in the data. This means that the machine automatically assumes that there is a natural order in the categories with one being higher than the other. To get rid of this problem we do One-Hot Encoding. One-Hot Encoding makes columns same as the number of categorical variables and that too of boolean type.

For example,

NEW YORK	MUMBAI	MOSCOW
1	0	0
0	1	0
0	0	1

The binary variables are often referred as dummy variables.

2.1.6 Feature Scaling

Why Scaling?

Most of the time the data in different features of dataset have high varying magnitude of data. This is a problem in most of the algorithms as they mostly use Eucledian distance to do their computation. This high magnitude features will weigh more than others features in the distance calculations with low magnitude.

To suppress this problem we need to scale the features at a same level for which we do Feature Scaling.

What Happens while Scaling?

When we use algorithms to scale, all the data is reduced to between -1 to 1 or 0 to 1 depending on algorithms and even the parameters.

When to Scale?

There are various algorithms where scaling makes a difference and some where it does not.

- k-Nearest Neighbours is very sensitive with magnitudes as they depend on Eucledian distance and scaling is needed for this algorithm.
- Models which are tree based like Decision Tree & Random Forest do not depend on distance due to which they can handle ranges of features which vary a lot.

Model Selection

The motive is to answer the questions by the customer who wants to know about the measures that they could take to minimise absenteeism and to have a guess on what should they expect in 2011 as their loss.

We need to find a model which better suits the data. It might not be enough for the model to well fitting the data as the data is not enough. For better prediction data is the key. The more the data, the better is the prediction. We will be using few regression models like:

- Multiple Linear Regression
- Decision Tree
- Random Forest
- SVR

Multiple Linear Regression

Multiple Linear Regression is the most common form of linear regression. This regression is used to explain relation between one continuous dependent variable and one or more than one independent variables which can be continuous or categorical.

MSE Score of Multiple Linear Regression: 23.93

R square Score of Multiple Linear Regression: 0.22

Decision Tree Regression

This regressor organises a series of test questions and conditions in a tree structure. In the decision tree, the roots and internal nodes contains different test conditions to separate records having different characteristic. Following are the results of the regression:

MSE Score of Decision Tree Regression: 24.29

R square Score of Decision Tree Regression: 0.20

Random Forest Regression

Random Forest algorithm creates the forest with a number of trees. In general, the more the number of trees the more robust the model. In random forest, as we increase the number of trees the accuracy also increases but too many trees can lead to overfitting. In our model, the following was the result produced:

MSE Score of Random Forest Regression: 26.65

R square Score of Random Forest Regression: 0.13

Support Vector Regression

Support Vector Machine is highly preferred by many as it produces very good accuracy with minimum computational power. The goal of SVR is to build a hyperplane in an N-dimensional data to classify the data points. For our data SVR produced the following results:

MSE Score of Random Forest Regression: 27.68

R square Score of Random Forest Regression: 0.09

Conclusion

The answer to the two questions that the customer wanted:

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

  It was observed that the most common reason for absenteeism was among medical reason. So better medical checkup and dispensaries in the office could be a great help.

  Also, less educated people tend to be absent from their work more than those who were educated so educational programs for the employee might also create an impact on reducing the absenteeism.

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

  Expected loss in 2011(if the data is considered for three years):

  Absenteeism time in hours: 1150.00

  Target Missed Total: 42401449.25

  Projection of loss in 2011(if the data is considered of only 1 year):

  Total Absenteeism in hours: 5161.4

  Total Workload missed: 10861797

References

- Medium
- www.towardsdatascience.com
- www.stackoverflow.com
- https://stats.stackexchange.com/
- Analytics Vidya
- Research Gate