Project Report on Employee Absenteeism

(Chandradeep Pokhariya)

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:

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?

1.2 DATA

Data includes one dataset. Dataset contains 21 variables and around 740 observations. It also has missing values, which we have to fill by imputing method.

Dataset Characteristics: Time Series Multivariate
Number of Attributes: 21
Missing Values : Yes

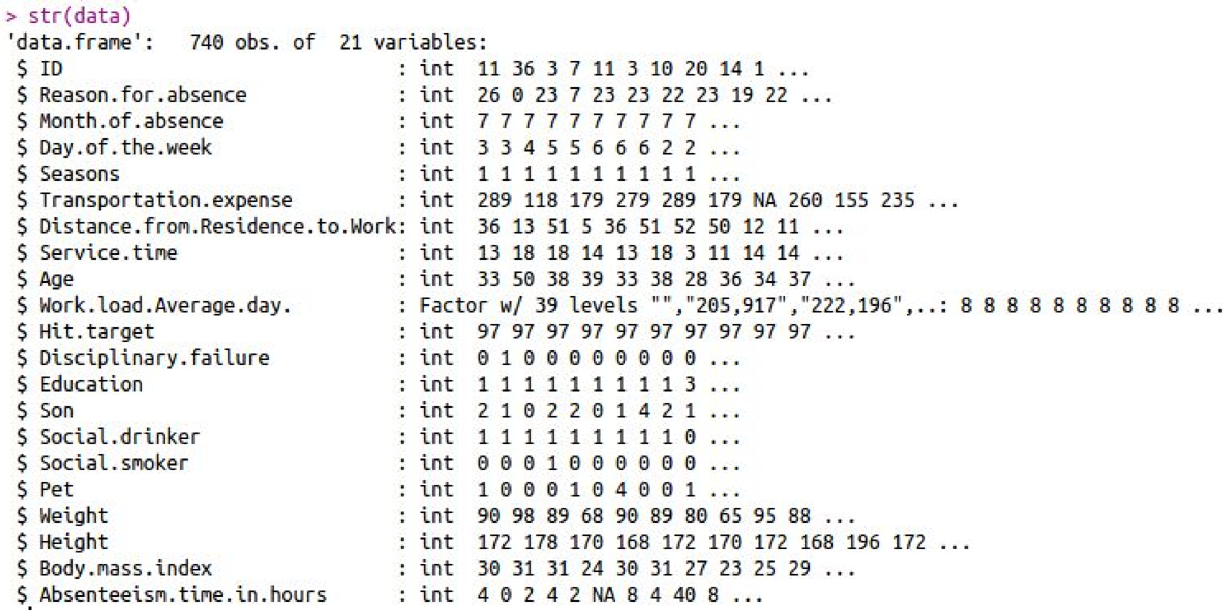
The variables are:

1. Individual identification (ID)
2. Reason for absence (ICD)
3. Month of absence
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
5. Seasons (summer (1), autumn (2), winter (3), spring (4))
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 (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
14. Son (number of children)
15. Social drinker (yes=1; no=0)
16. Social smoker (yes=1; no=0)
17. Pet (number of pet)
18. Weight
19. Height
20. Body mass index
21. Absenteeism time in hours (target)

In these variables, **Absenteeism time in hours** is target variable while all other 20 variables are combination of continuous and categorical variables.

The dataset is not cleaned properly, so we have to do some exploratory data analysis to meet our requirement.



```
> str(data)
'data.frame':	740 obs. of  21 variables:
 $ ID                             : int  11 36 3 7 11 3 10 20 14 1 ...
 $ Reason.for.absence             : int  26 0 23 7 23 23 22 23 19 22 ...
 $ Month.of.absence               : int  7 7 7 7 7 7 7 7 7 7 ...
 $ Day.of.the.week                : int  3 3 4 5 5 6 6 6 2 2 ...
 $ Seasons                        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Transportation.expense         : int  289 118 179 279 289 179 NA 260 155 235 ...
 $ Distance.from.Residence.to.Work: int  36 13 51 5 36 51 52 50 12 11 ...
 $ Service.time                   : int  13 18 18 14 13 18 3 11 14 14 ...
 $ Age                            : int  33 50 38 39 33 38 28 36 34 37 ...
 $ Work.load.Average.day.         : Factor w/ 39 levels "","205,917","222,196",..: 8 8 8 8 8 8 8 8 8 8 ...
 $ Hit.target                     : int  97 97 97 97 97 97 97 97 97 97 ...
 $ Disciplinary.failure           : int  0 1 0 0 0 0 0 0 0 0 ...
 $ Education                      : int  1 1 1 1 1 1 1 1 1 3 ...
 $ Son                            : int  2 1 0 2 2 0 1 4 2 1 ...
 $ Social.drinker                 : int  1 1 1 1 1 1 1 1 1 0 ...
 $ Social.smoker                  : int  0 0 0 1 0 0 0 0 0 0 ...
 $ Pet                            : int  1 0 0 0 1 0 4 0 0 1 ...
 $ Weight                         : int  90 98 89 68 90 89 80 65 95 88 ...
 $ Height                         : int  172 178 170 168 172 170 172 168 196 172 ...
 $ Body.mass.index                : int  30 31 31 24 30 31 27 23 25 29 ...
 $ Absenteeism.time.in.hours      : int  4 0 2 4 2 NA 8 4 40 8 ...
```

CHAPTER 2 METHODOLOGY

2.1 EXPLORATORY DATA ANALYSIS

As the dataset may contain missing values and outliers, we need to first clean the data before feeding it to the algorithm. So at first we will proceed with missing value analysis and then outlier analysis and followed by feature selection.

First we clean the data by removing commas and extra spaces from the dataset, then after we make variables name according to our need.

We even remove extra variable like ‘ID’ which do not have any need in the dataset to get insights from the data.

Then we change the datatype of the variable. For Example, some variables are not in the factor type so they are converted to factor type from numeric data type.

2.2 MISSING VALUE ANALYSIS

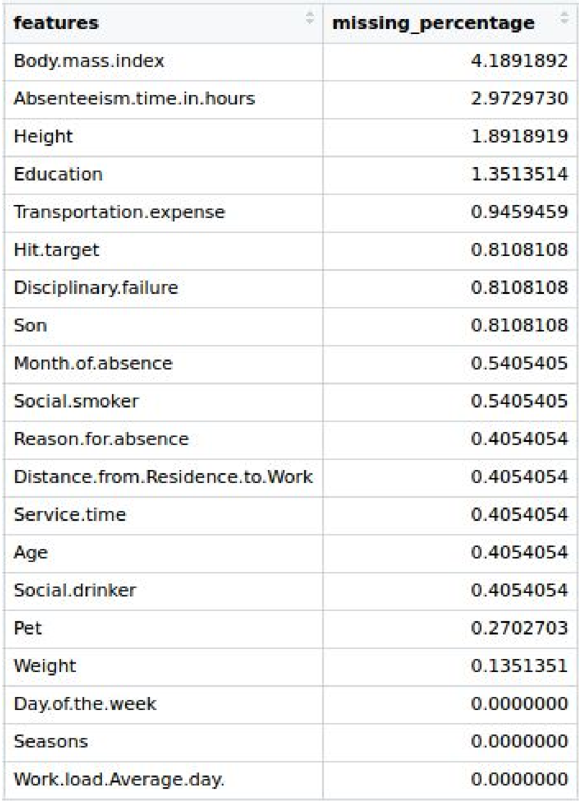
Missing values can be due to various reasons. It may be due to some human error, refuse to answer while surveying or optional box in the questionnaire. We will first find the missing values in the dataset and then impute them with different methods of imputation.

Since this dataset is a hybrid type of dataset. It contains both categorical and continuous type of variables. So firstly we will do imputation of categorical variable and thereafter imputation of continuous variable.

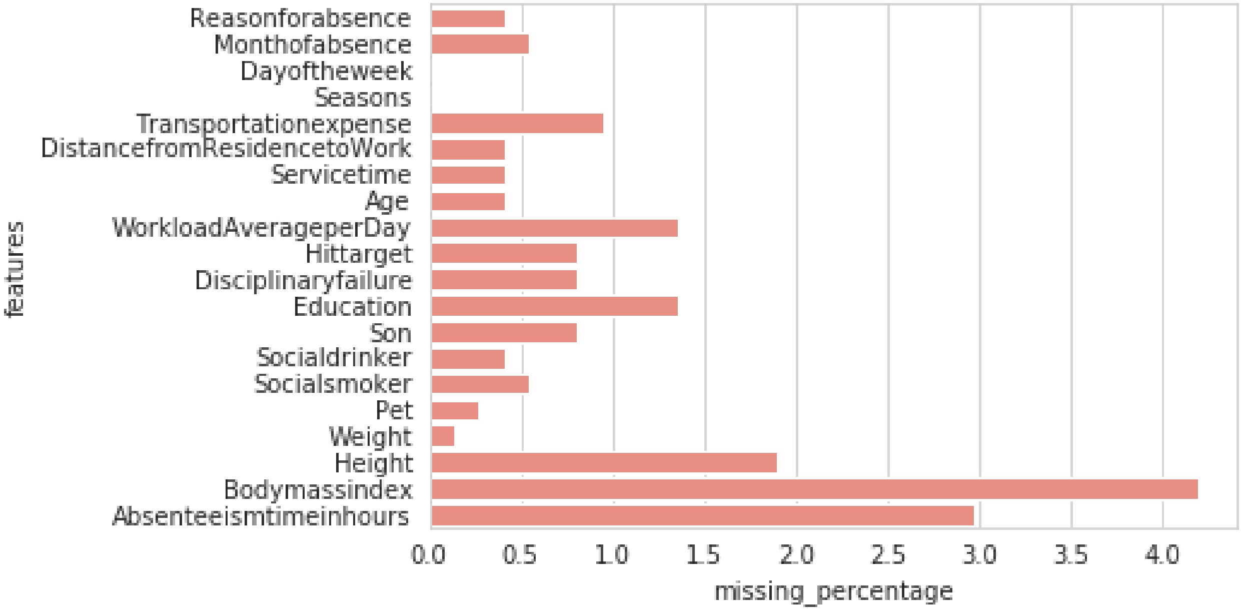
First we will do the imputation of the categorical missing values with the help of mode imputation. Mode imputation is the most common method for imputing categorical data.

Thereafter we will try different imputation method for our continuous data. These methods include mean, median and knn imputation.

In our test knn imputation was the best one, which gave more accurate values so we froze knn imputation for the whole dataset.



features	missing_percentage
Body.mass.index	4.1891892
Absenteeism.time.in.hours	2.9729730
Height	1.8918919
Education	1.3513514
Transportation.expense	0.9459459
Hit.target	0.8108108
Disciplinary.failure	0.8108108
Son	0.8108108
Month.of.absence	0.5405405
Social.smoker	0.5405405
Reason.for.absence	0.4054054
Distance.from.Residence.to.Work	0.4054054
Service.time	0.4054054
Age	0.4054054
Social.drinker	0.4054054
Pet	0.2702703
Weight	0.1351351
Day.of.the.week	0.0000000
Seasons	0.0000000
Work.load.Average.day.	0.0000000

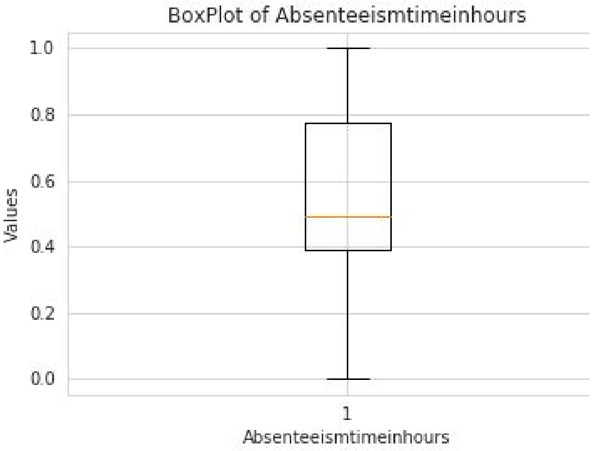


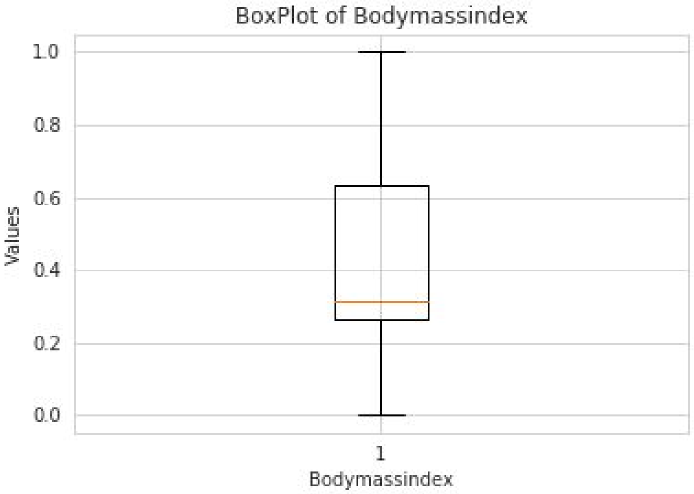
2.3 OUTLIER ANALYSIS

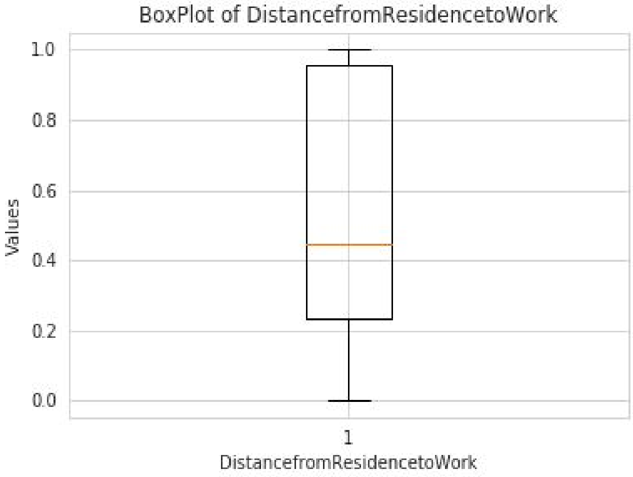
In outlier analysis we check the data which does not lie in the normal range of data. Either it is too much or too less than the data. Outliers are the observations which are inconsistent with the rest of the data.

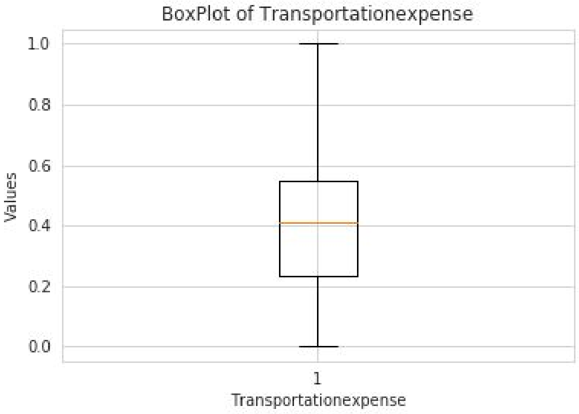
There are different techniques to detect outliers.

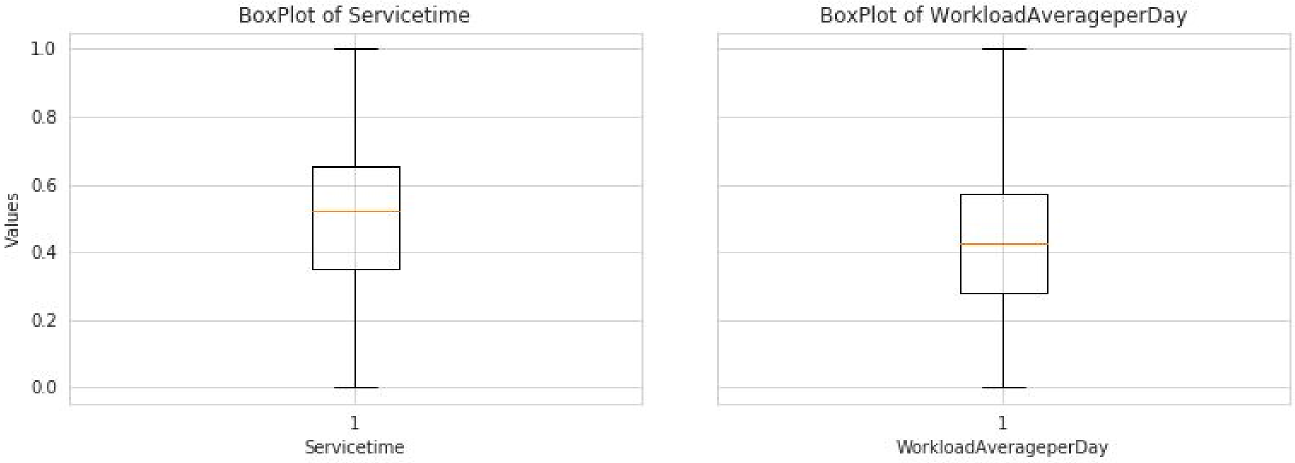
- **Box Plot Method**: Data above the upper fence and lower fence will be calculated as outliers.
- **Statistical Test - Grubbs test:** It has an assumption that the data is normally distributed, but in real cases it is very rare.
- **R -Package Outlier:** It considers mean algorithm to check for outliers. Data points which deviates from the mean are called outliers.
- **Replace with NA:** In this method we replace outliers with NA, and then do further imputation.











2.4 FEATURE SELECTION

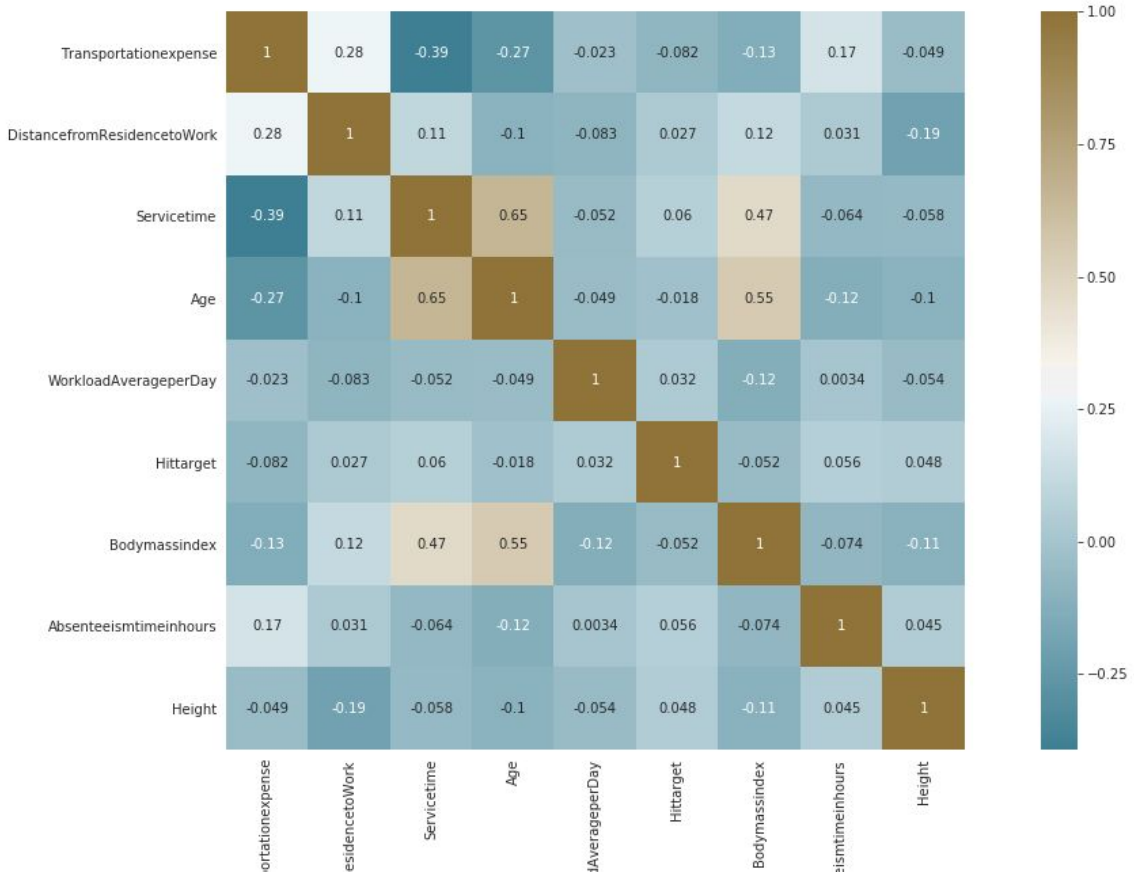
In feature selection we select a subset of relevant features or variables for use in model construction. Basically we check which features of the dataset do not contain the correlation with the dependent variable (Absenteeism time in hours).

As the dataset is hybrid and contain both categorical and continuous data, we will use separate test for checking the multicollinearity in the dataset.

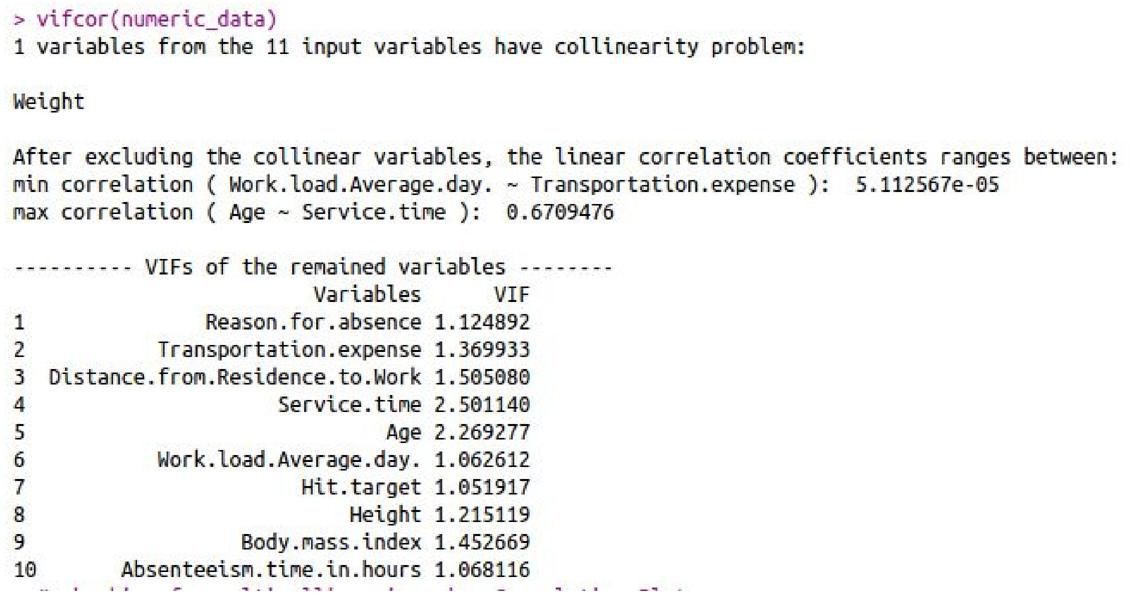
In categorical type of data we use and in continuous type of data we make correlation graph.

- **Correlation Analysis:**
  This method is used for continuous or numerical variables. The range of correlation values is -1 to +1.

  So at first I made the correlation graph and checked the correlation between different features with the target variable. The feature which has the least relation with the target variable is dropped out.



We further check for the **variance inflation factor.**



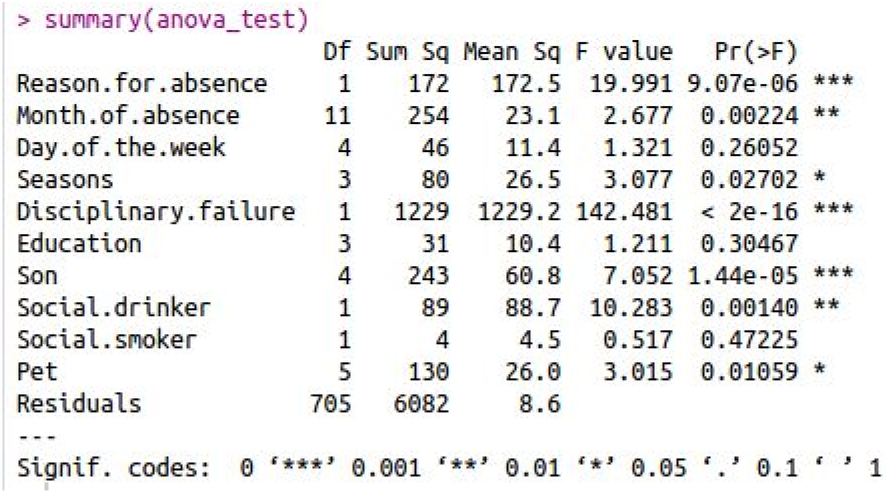
```
> vifcor(numeric_data)
1 variables from the 11 input variables have collinearity problem:

Weight

After excluding the collinear variables, the linear correlation coefficients ranges between:
min correlation ( Work.load.Average.day. ~ Transportation.expense ):  5.112567e-05
max correlation ( Age ~ Service.time ):  0.6709476

---------- VIFs of the remained variables --------
                        Variables      VIF
1               Reason.for.absence 1.124892
2           Transportation.expense 1.369933
3  Distance.from.Residence.to.Work 1.505080
4                     Service.time 2.501140
5                              Age 2.269277
6           Work.load.Average.day. 1.062612
7                       Hit.target 1.051917
8                           Height 1.215119
9                  Body.mass.index 1.452669
10       Absenteeism.time.in.hours 1.068116
```

- **ANOVA Test:** It is a statistical test which is used to test the mean of two or more group. Since some features has the probability value greater than 0.5, so we will drop those variables.



```
> summary(anova_test)
                      Df Sum Sq Mean Sq F value   Pr(>F)    
Reason.for.absence     1    172   172.5  19.991 9.07e-06 ***
Month.of.absence      11    254    23.1   2.677  0.00224 ** 
Day.of.the.week        4     46    11.4   1.321  0.26052    
Seasons                3     80    26.5   3.077  0.02702 *  
Disciplinary.failure   1   1229  1229.2 142.481  < 2e-16 ***
Education              3     31    10.4   1.211  0.30467    
Son                    4    243    60.8   7.052 1.44e-05 ***
Social.drinker         1     89    88.7  10.283  0.00140 ** 
Social.smoker          1      4     4.5   0.517  0.47225    
Pet                    5    130    26.0   3.015  0.01059 *  
Residuals            705   6082     8.6                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Updated Variables are following,

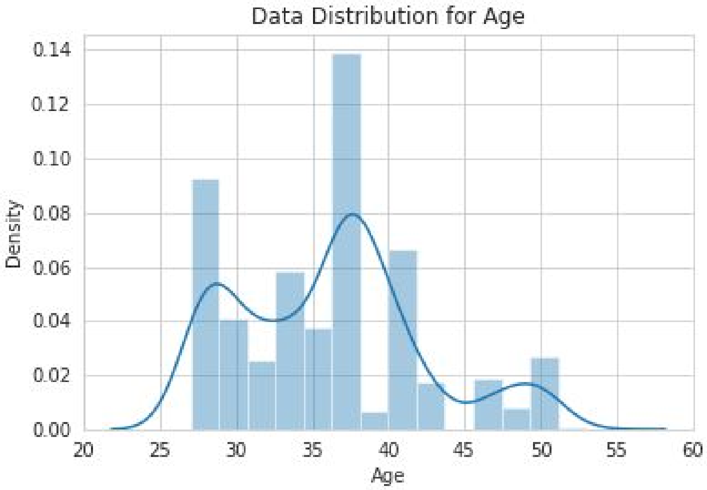
Categorical variables: Reasonforabsence, Dayoftheweek, Son, Disciplinaryfailure, Socialdriner

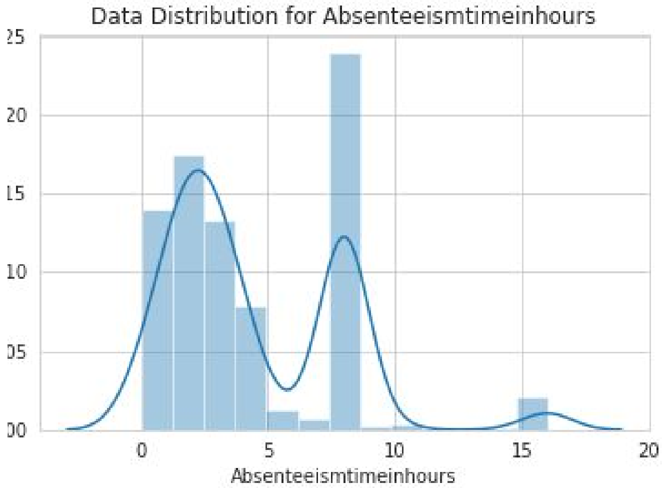
Numerical variables: Transportationexpense, DistancefromResidencetoWork,Servicetime, Age, WorkloadAverageperDay, Hittarget, Bodymassindex, Absenteeismtimeinhours,Height

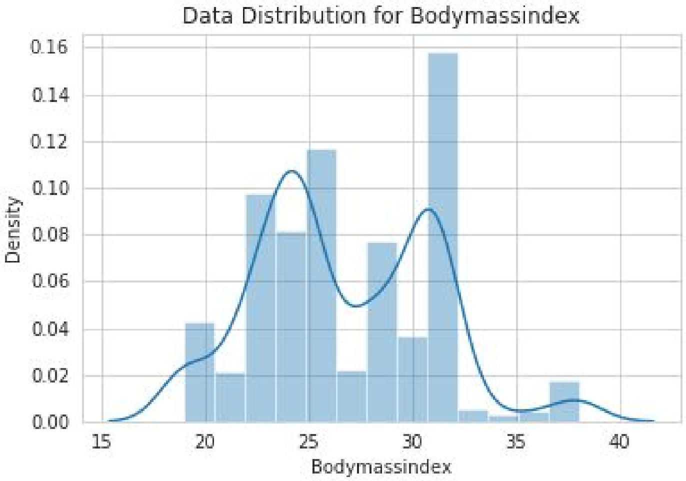
2.5 FEATURE SCALING:

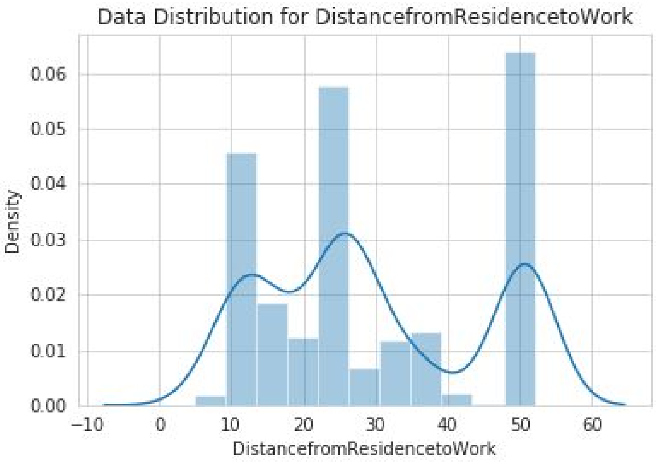
After the selection of important variable is done, then further skewness of the data is checked. Firstly the density plot of all the variables is plotted and checked whether the distribution is normal or not. There are two methods of feature scaling, one is standardisation and other one is normalization. When the density distribution of a dataset is normalized, we use standardization and if the density distribution is not normalized, then normalization is done.

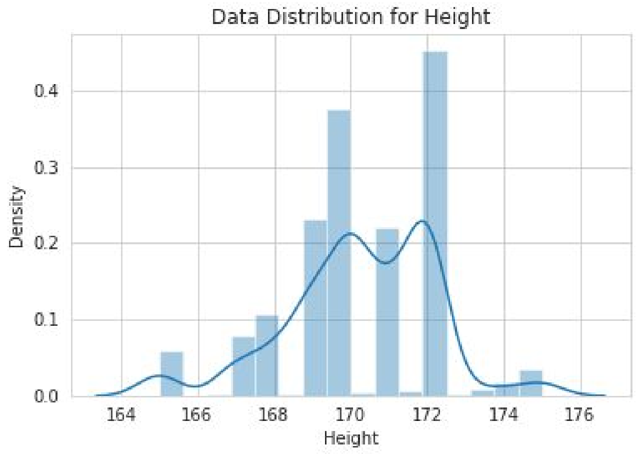
Density Plots:

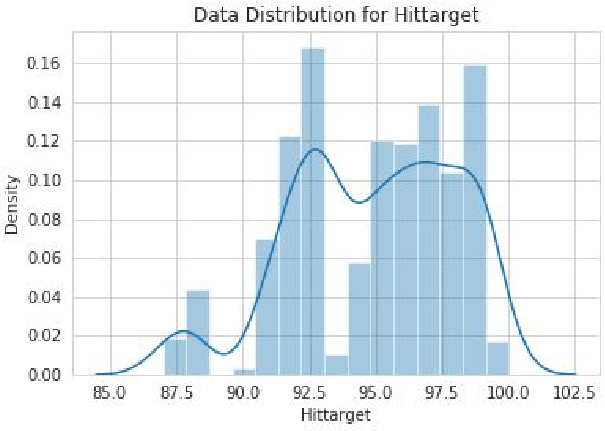


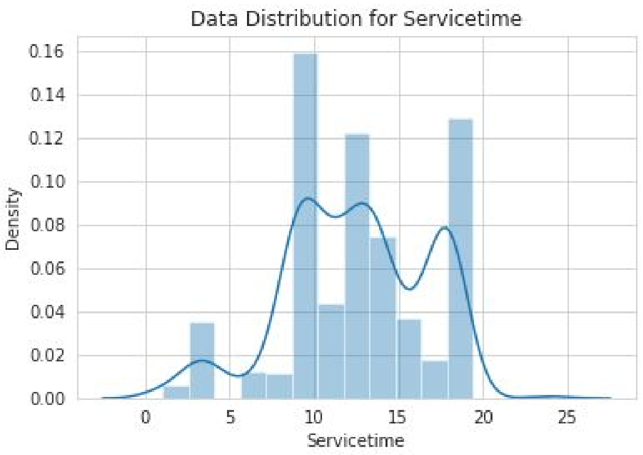


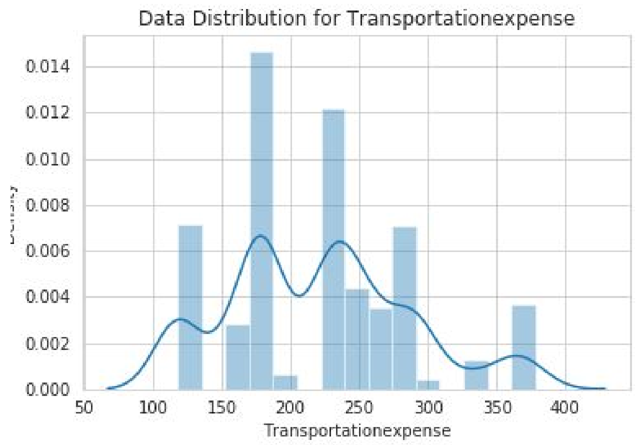


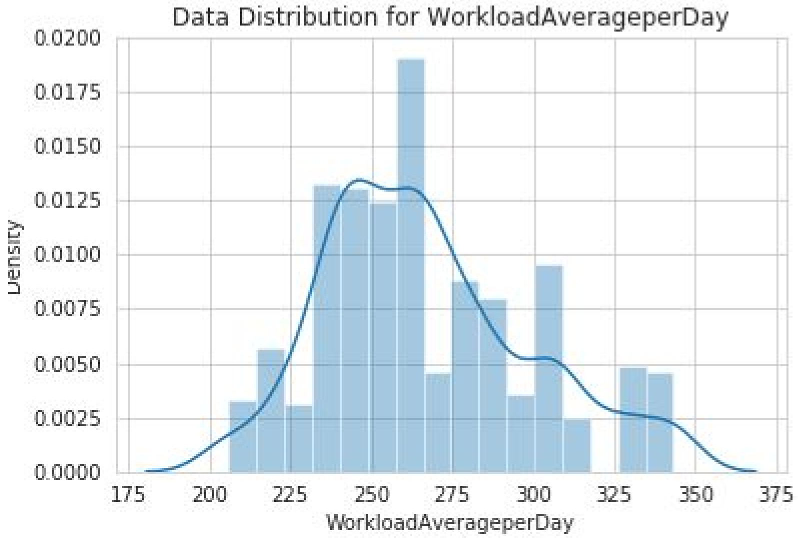












Since we can see that there is so much skewness in the data, we will have to use normalization. Data is not normally distributed.

CHAPTER 3 MODELLING

3.1 DUMMY VARIABLES

So the first question arrives in mind is that why we need dummy variables. Dummy variables are basically required for the modelling of categorical variables. As in categorical variables there are multiple levels, ranging from 1 to different levels. So machine learning algorithm will give weightage to the level which is large and give less weightage to the level which is small. So it can lead to the biasing of the machine learning algorithm.

So to avoid this situation we use dummy variable which make feature for every level of categorical and then feed it into machine learning algorithm.

3.2 SPLITTING DATA

As evaluating data on the train set can lead to overfitting of the model, we divide the dataset into two parts one is training dataset and another one is test dataset. So we started with splitting the data, putting 60% in the training set and 40% in test set. We have already pre processed the data so that there comes no problem in putting this model to any machine learning algorithm.

3.3 RANDOM FOREST

Random forests or random decision forests are an ensemble learning methods for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forest is very popular model in predicting the data.

Here is the result of the model,

- RMSE: 0.17755826
- MAE: 0.03152693
- R2 Score: 0.412840

3.2 DECISION TREE

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Here is the decision tree result of the method,

- RMSE : 0.2221078
- MAE: 0.049331886
- R2 Score : 0.00752315

3.3 LINEAR REGRESSION

Linear regression is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Here is the result of the linear regression method,

- RMSE: 0.17755826
- MAE : 0.0315269
- R2 Score: 0.41284

3.4 MODEL EVALUATION

Since in our dataset the data was continuous, so we consider basically two type of error matrices here.

1. RMSE error matrix
2. MAE error matrix
3. R squared matrix

1. RMSE error matrix:

MAPE is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation, also used as a loss function for regression problems in machine learning.It usually express accuracy as a percentage.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}{n}}$$

2. MAE error matrix:

In statistics, mean absolute error (MAE) is a measure of difference between two continuous variables. Assume *X* and *Y* are variables of paired observations that express the same phenomenon. Examples of *Y* versus *X* include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}.$$

3. R Squared Error:

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

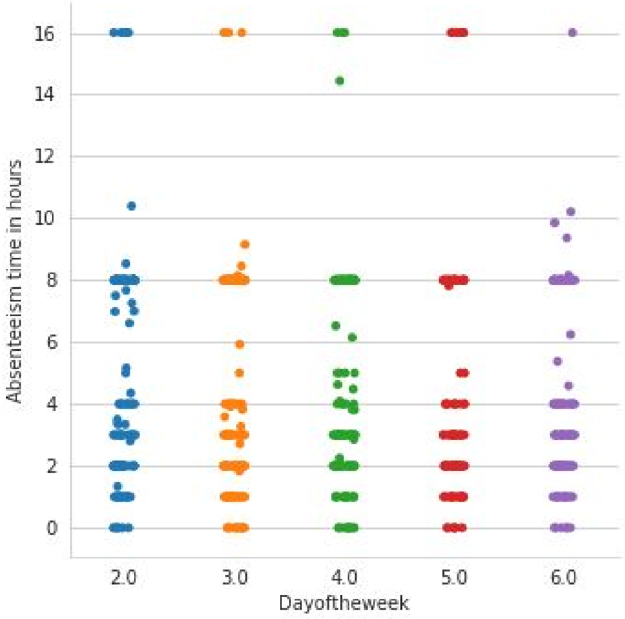
The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model.

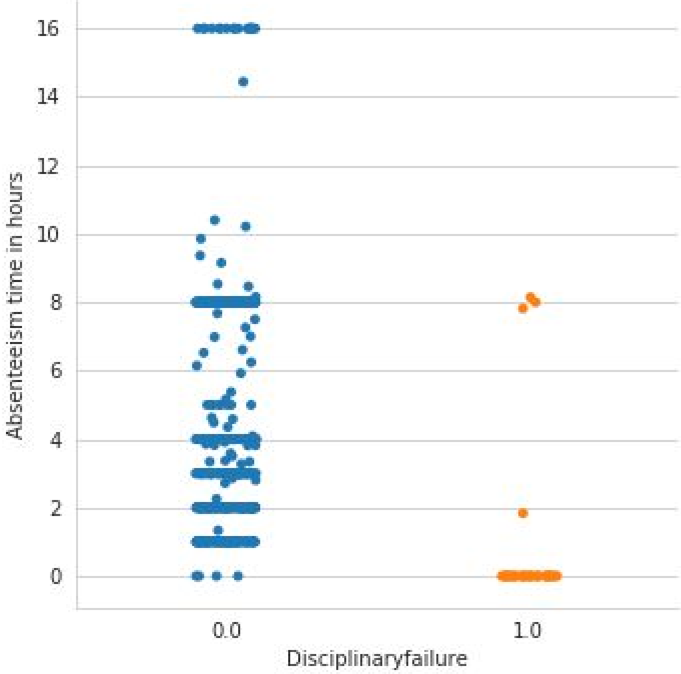
R-squared = Explained variation / Total variation

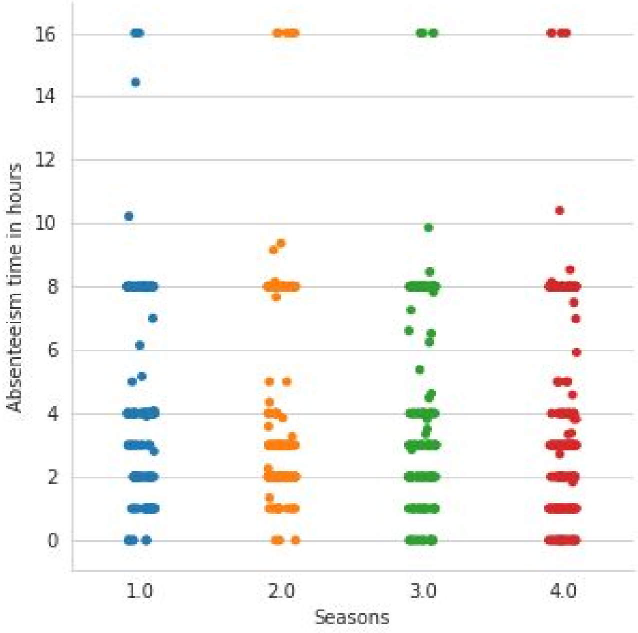
ANSWERS TO THE ASKED QUESTIONS:

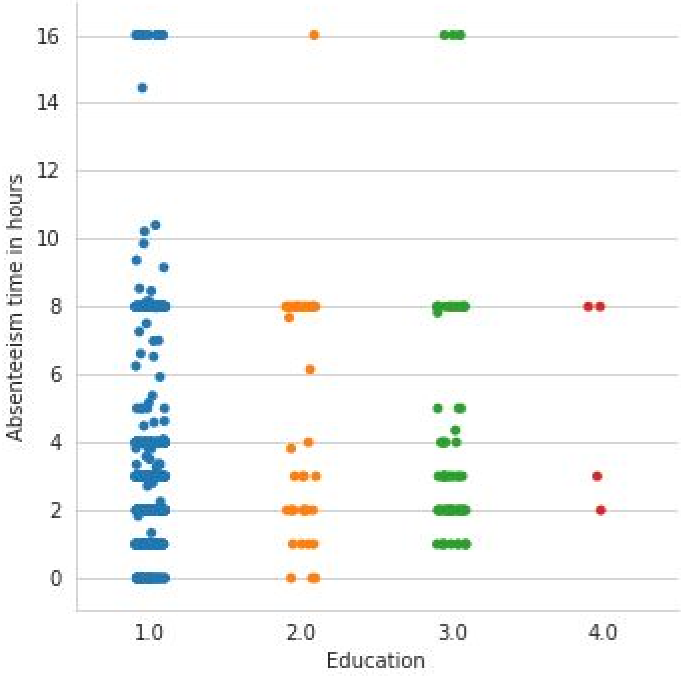
1. **What changes company should bring to reduce the number of absenteeism?**

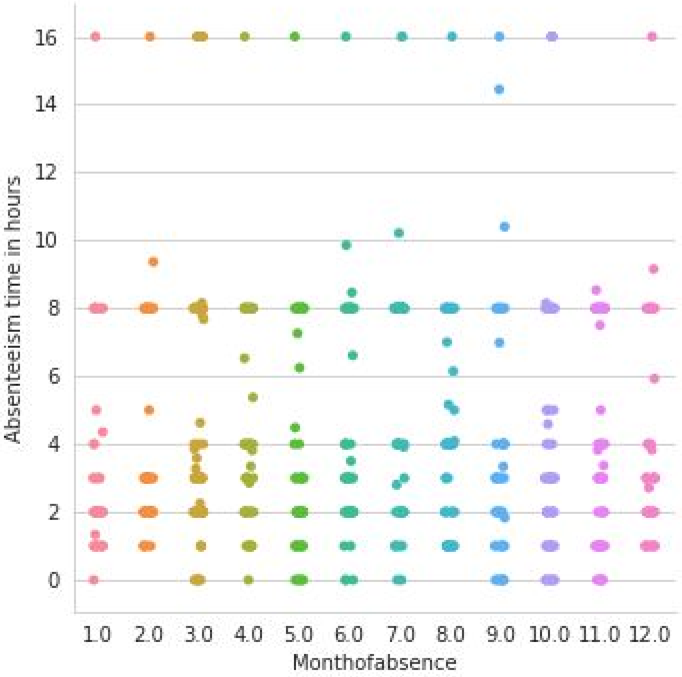
   Here is the distribution of data with different variables:

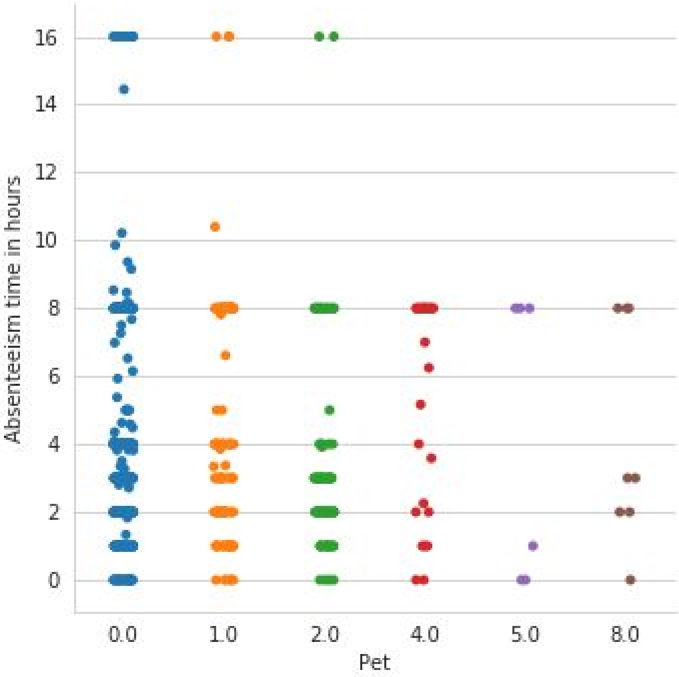


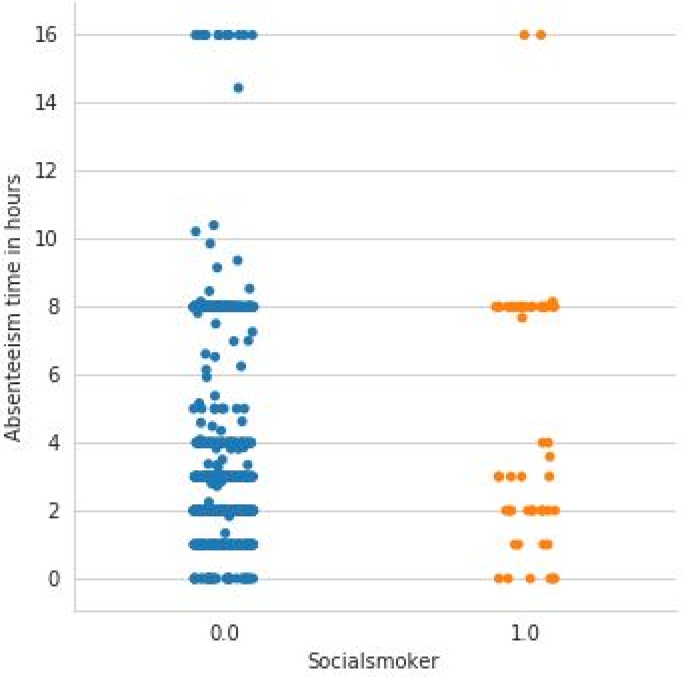


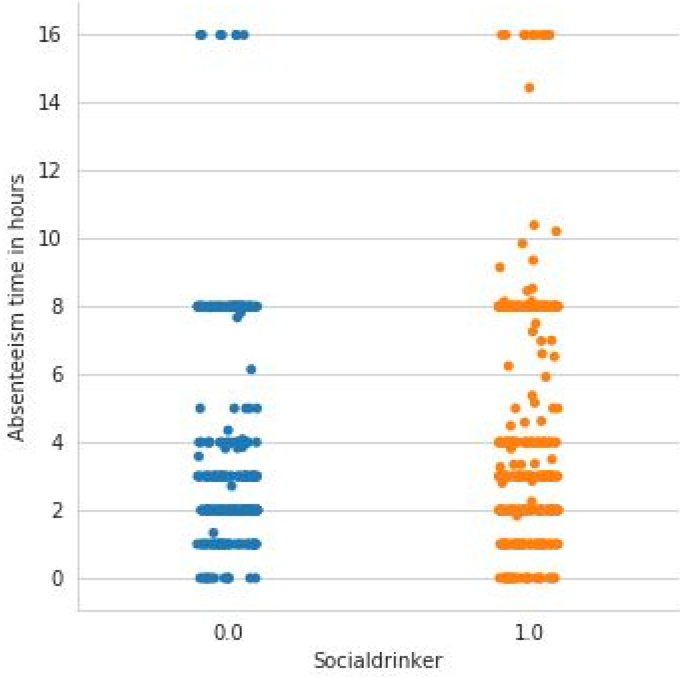


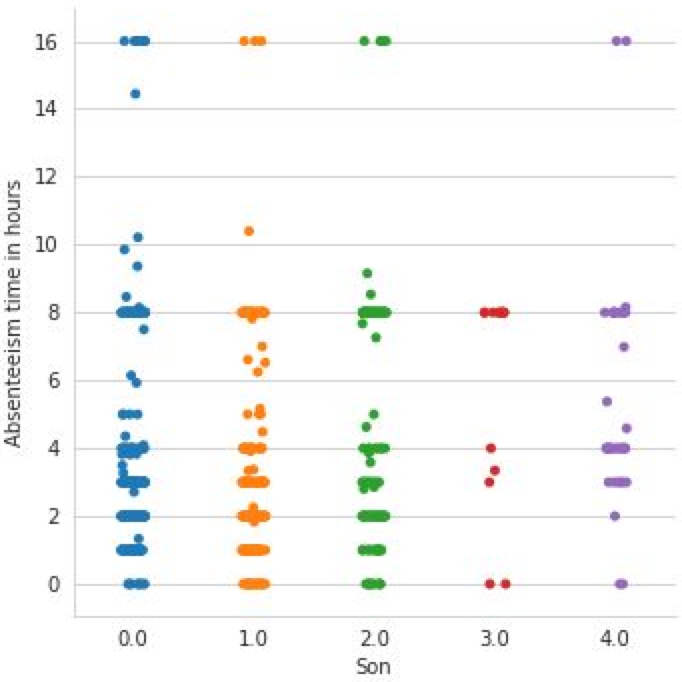


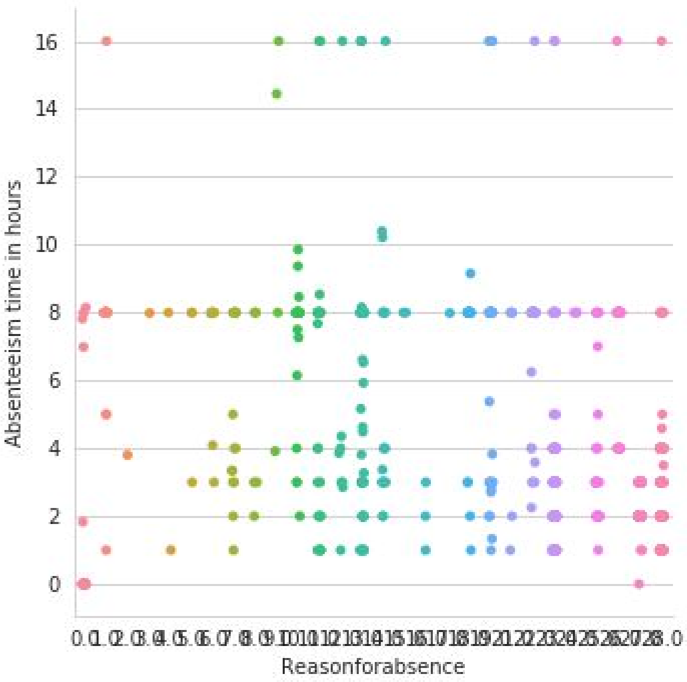

Absenteeism time in hours
0
2
4
6
8
10
12
14
16
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
Monthofabsence


Absenteeism time in hours
0
2
4
6
8
10
12
14
16
0.0
1.0
2.0
4.0
5.0
8.0
Pet


Absenteeism time in hours
0
2
4
6
8
10
12
14
16
0.0
1.0
Socialsmoker


Absenteeism time in hours
0
2
4
6
8
10
12
14
16
0.0
1.0
Socialdrinker





From analyzing these graphs we observed that,

1. High School employee has the maximum absentee time.
2. Employee with no disciplinary failure has maximum absentee time
3. In March and September, employees are more absent
4. Employee with no pet has maximum absentee time.
5. Social drinker has more absentee time.