Employee Absenteeism

PROJECT REPORT

By

Esha Rudra

Date: 8th Sept 2019

1. Contents

2. Methodology

3. Conclusion

Appendix

References :

Edwisor learning path, Geeks for geeks website, Stacks overflow website, Youtube tutiorials.

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:

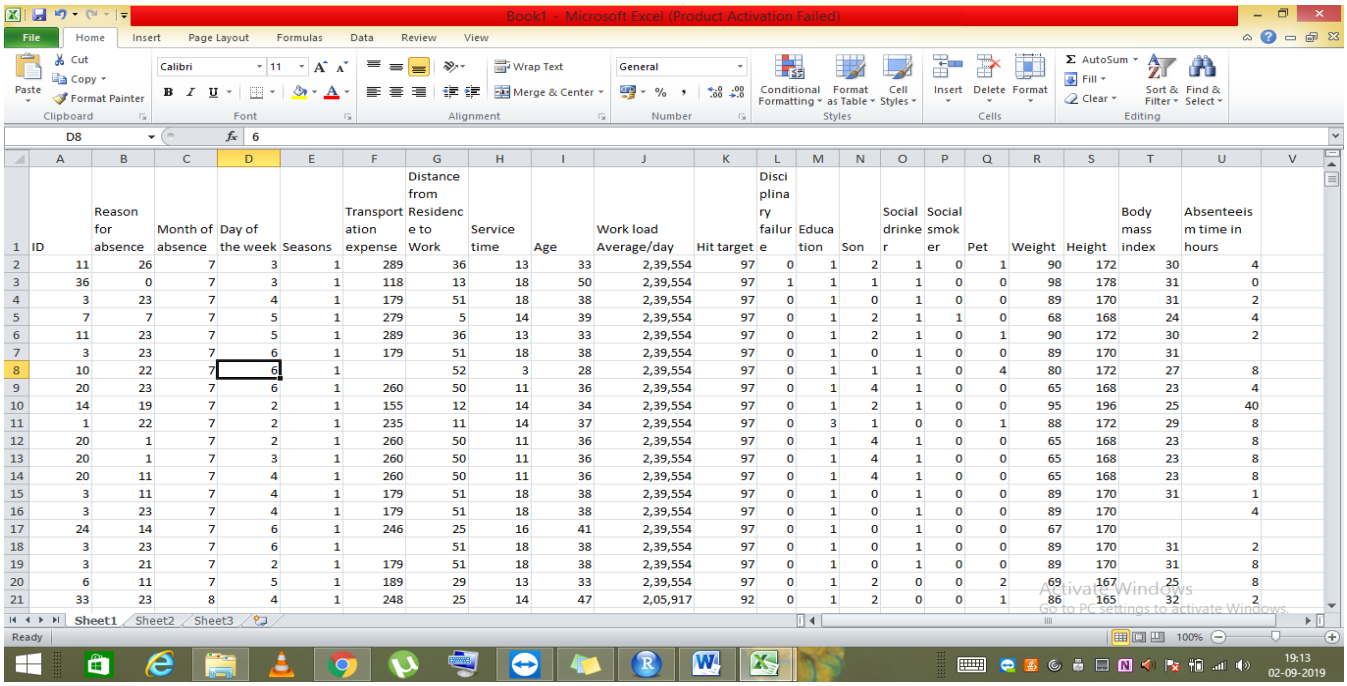
1. What changes company should bring to reduce the number of absenteeism?

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

The problem statement is to predict the loss to be incurred by the company monthly if same trend of absenteeism continues in the year 2011.

1.2 Data

We have data of employee absenteeism for 3 years (2007 to 2010) and we need to predict for 2011. We have to build machine learning models which can help the company in solving absenteeism issue & to optimize the work of the company and predict the losses if the issue continues. From the dataset, we can observe that there are 21 variables in our data which can be further identified as 20 independent variables and 1 dependent variable. Here our target variable i.e. **"Absenteeism time in hours"** is continuous in nature. Let's look at the dataset.



ID	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	Weight	Height	Body mass index	Absenteeism time in hours
11	26	7	3	1	289	36	13	33	2,39,554	97	0	1	2	1	0	1	90	172	30	4
36	0	7	3	1	118	13	18	50	2,39,554	97	1	1	1	1	0	0	98	178	31	0
3	23	7	4	1	179	51	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	2
7	7	7	5	1	279	5	14	39	2,39,554	97	0	1	2	1	1	0	68	168	24	4
11	23	7	5	1	289	36	13	33	2,39,554	97	0	1	2	1	0	1	90	172	30	2
3	23	7	6	1	179	51	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	
10	22	7	6	1		52	3	28	2,39,554	97	0	1	1	1	0	4	80	172	27	8
20	23	7	6	1	260	50	11	36	2,39,554	97	0	1	4	1	0	0	65	168	23	4
14	19	7	2	1	155	12	14	34	2,39,554	97	0	1	2	1	0	0	95	196	25	40
1	22	7	2	1	235	11	14	37	2,39,554	97	0	3	1	0	0	1	88	172	29	8
20	1	7	2	1	260	50	11	36	2,39,554	97	0	1	4	1	0	0	65	168	23	8
20	1	7	3	1	260	50	11	36	2,39,554	97	0	1	4	1	0	0	65	168	23	8
20	11	7	4	1	260	50	11	36	2,39,554	97	0	1	4	1	0	0	65	168	23	8
3	11	7	4	1	179	51	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	1
3	23	7	4	1	179	51	18	38	2,39,554	97	0	1	0	1	0	0	89	170		4
24	14	7	6	1	246	25	16	41	2,39,554	97	0	1	0	1	0	0	67	170		
3	23	7	6	1		51	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	2
3	21	7	2	1	179	51	18	38	2,39,554	97	0	1	0	1	0	0	89	170	31	8
6	11	7	5	1	189	29	13	33	2,39,554	97	0	1	2	0	0	2	69	167	25	8
33	23	8	4	1	248	25	14	47	2,05,917	92	0	1	2	0	0	1	86	165	32	2

Variables Information:

1. Individual identification (ID)

2. Reason for absence (ICD)

- Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I. Certain infectious and parasitic diseases

II. Neoplasms

III. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV. Endocrine, nutritional and metabolic diseases

V. Mental and behavioural disorders

VI. Diseases of the nervous system

VII. Diseases of the eye and adnexa

VIII. Diseases of the ear and mastoid process

IX. Diseases of the circulatory system

X. Diseases of the respiratory system

XI. Diseases of the digestive system

XII. Diseases of the skin and subcutaneous tissue

XIII. Diseases of the musculoskeletal system and connective tissue

XIV. Diseases of the genitourinary system

XV. Pregnancy, childbirth and the puerperium

XVI. Certain conditions originating in the perinatal period

XVII. Congenital malformations, deformations and chromosomal abnormalities

XVIII. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX. Injury, poisoning and certain other consequences of external causes

XX. External causes of morbidity and mortality

XXI. Factors influencing health status and contact with health services

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

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 (kilometres)

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

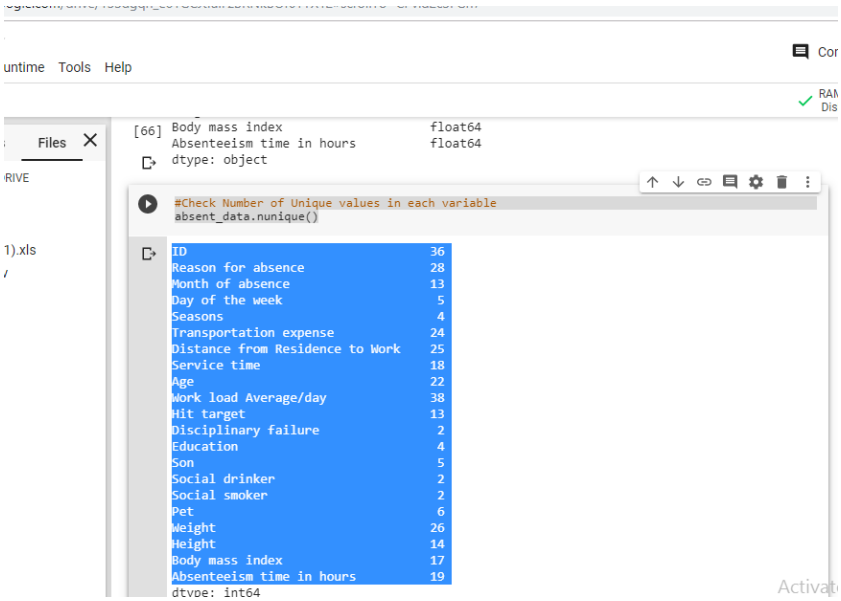
20. Body mass index

21. Absenteeism time in hours (target)

1.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a way to analyse data sets and get a better understanding about the data. It refers to performing initial investigations on data to find out trend, anomalies, draw inferences, test hypothesis.We have 21 variables and 740 observations in the data. It consists of both continuous and categorical variables. Let's do some more data analysis.

Variables and their number of unique values (so that we can categorise them into continuous and categorical variable)



After Exploratory Data Analysis, we concluded that there are 10 continuous variables and 11 categorical variables in the dataset which are as follows:

continuous_vars = ['Distance from Residence to Work', 'Service time', 'Age', 'Work load Average/day ', 'Transportation expense', 'Hit target', 'Weight', 'Height', 'Body mass index', 'Absenteeism time in hours']

categorical_vars = ['ID','Reason for absence','Month of absence','Day of the week', 'Seasons','Disciplinary failure', 'Education', 'Social drinker', 'Social smoker', 'Pet', 'Son']

2. Methodology

2.1 Pre Processing

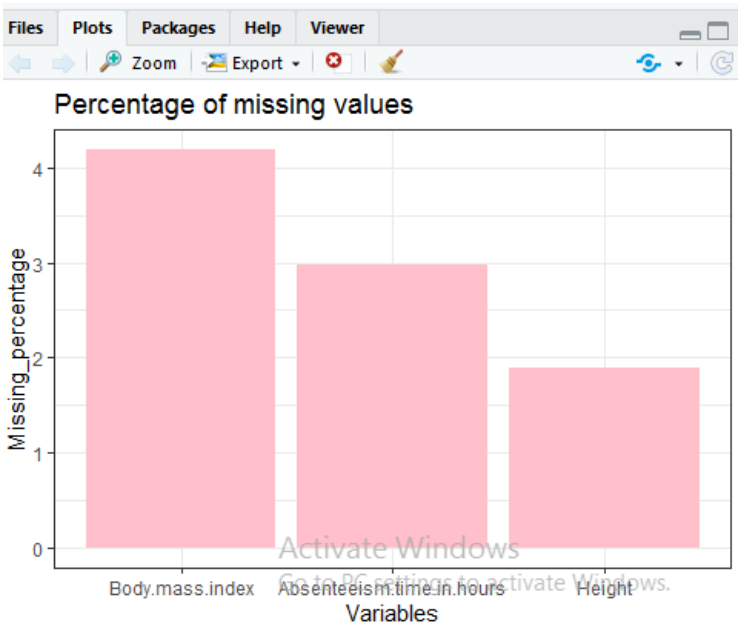
Data in real world is not clean and can't be used until we pre-process & clean the data i.e. transforming the data into a suitable format before feeding it to our machine learning model/ algorithm. With the help **data mining techniques** which involves transforming raw data into a machine learning model understandable format. In this process we clean the data, and work on data integration, transformation etc. It is the most crucial part of data science project and almost 80% of our time is spent on it. This is known as **Exploratory Data Analysis**. We first try and look at all the probability distributions of the variables. Most analysis like regression, require the data to be normally distributed. Here, we visualize and explore the data which will help us to build hypothesis, selecting proper machine learning algorithm etc.

2.2.1 Missing Value Analysis

In a dataset, if we find no value in an observation for a particular variable or variables we consider them as missing values. Most data sets have missing values for various reasons like human error, refuse to answer while surveying, optional box in questionnaire etc. On encountering missing values, we can either choose to impute them or ignore them. Depending on conditions we can either choose to drop the entire variable or entire observation or even consider imputing the missing

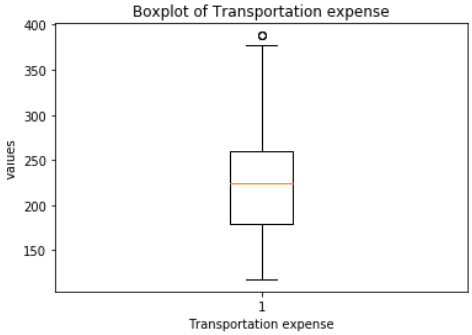
values. As per industry standards, we consider the variables to impute whose missing value is less than 30%. We plotted top 3 variables for missing values and the maximum percentage of missing value is 4.189% for “Body Mass Index” variable .So we will check for missing values in all variables and consider imputing them. **We will impute the values with KNN imputation method** as it works best on this dataset. We came to this conclusion after trying mean, median and KNN method.

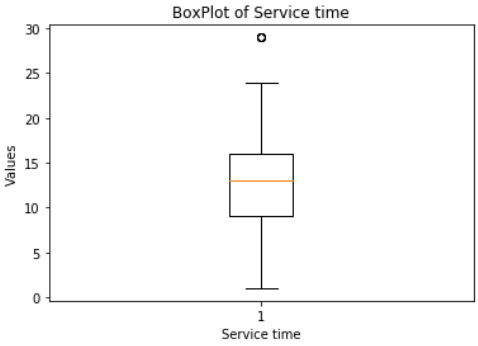
	Variables	Missing_percentage
0	Body mass index	4.189189
1	Absenteeism time in hours	2.972973
2	Height	1.891892
3	Work load Average/day	1.351351
4	Education	1.351351
5	Transportation expense	0.945946
6	Son	0.810811
7	Disciplinary failure	0.810811
8	Hit target	0.810811
9	Social smoker	0.540541
10	Age	0.405405
11	Reason for absence	0.405405
12	Service time	0.405405
13	Distance from Residence to Work	0.405405
14	Social drinker	0.405405
15	Pet	0.270270
16	Weight	0.135135
17	Month of absence	0.135135
18	Seasons	0.000000
19	Day of the week	0.000000
20	ID	0.000000

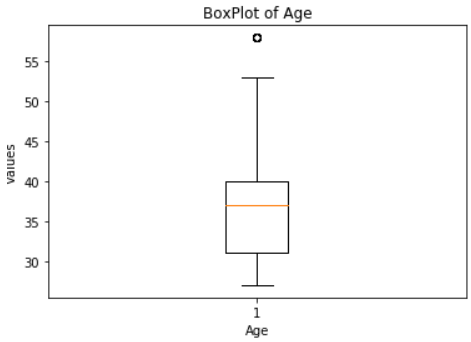


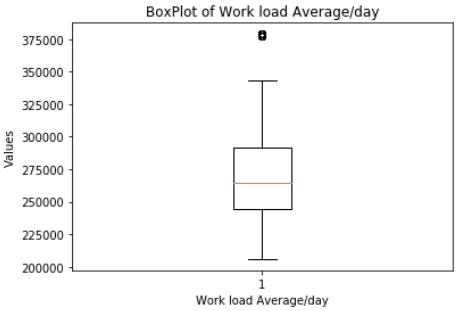
2.1.2 Outlier Analysis

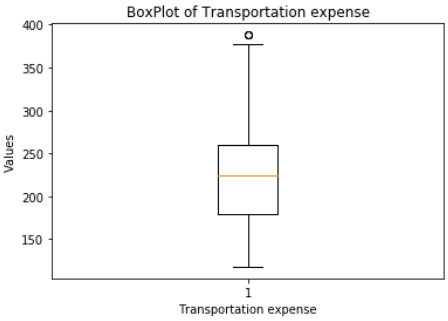
The outliers are the values of variables which fall beyond the normal range of the variable values and considered as exception. So it is better to remove them to make data normally distributed. But it is not the case always, sometimes outliers are telling something about the target variable. So we must check this before processing of the outliers. Now in our case, some of the variables are containing outliers. But we will not process the outliers of variable depending on ID to preserve the data integrity. But others can be processed using boxplot method.The below boxplot refers to outliers on the predictor variables, we can see various outliers associated with the features. Even though, the data has considerable amount of outliers, the approach is to retain every outlier and grab respective behaviour of all employees. As shown there are significant amount of outliers present in the target variable, which indicates a trend of Employee ' behaviour, there can be pattern , we need to treat those outliers. Below are the boxplots of the 11 independent continuous variables with respect to our target variable “Absenteeism time in hour”. These plots can help us with a lot of data understanding and give us knowledge to manipulate the data. Most of the variables except “**Distance from residence to work**”, “**Weight**” and “**Body mass index**” consist outliers. We treated the outliers by converting the outliers as NA i.e. missing values and filled them by KNN imputation method.

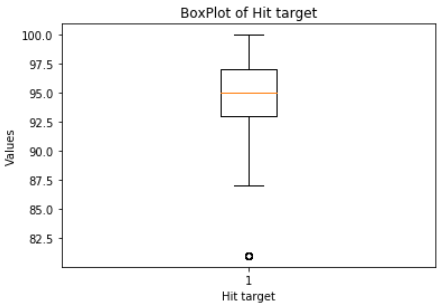

Boxplot of Transportation expense
values
Transportation expense

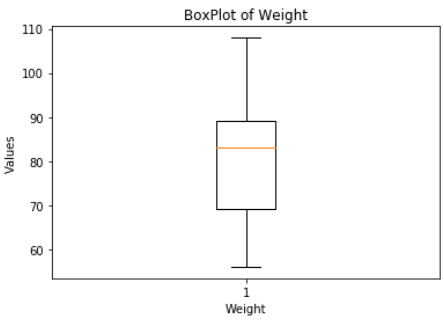

BoxPlot of Service time
Values
Service time

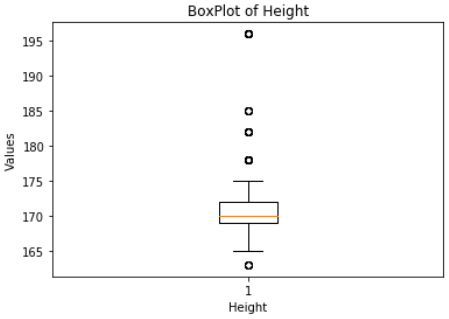

BoxPlot of Age
values
Age

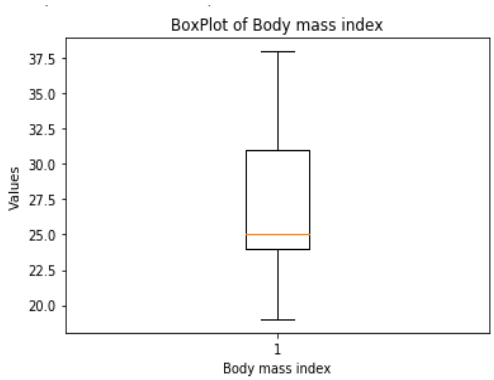

BoxPlot of Work load Average/day
Values
Work load Average/day

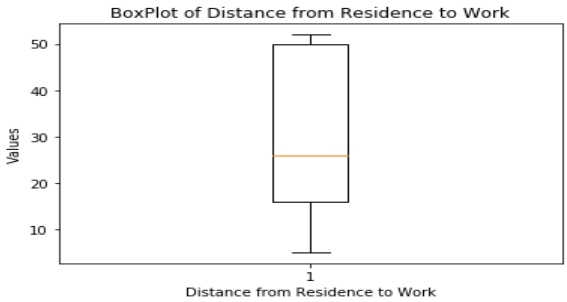

BoxPlot of Transportation expense
Values
Transportation expense


BoxPlot of Hit target
Values
Hit target


BoxPlot of Weight
Values
Weight


BoxPlot of Height
Values
Height



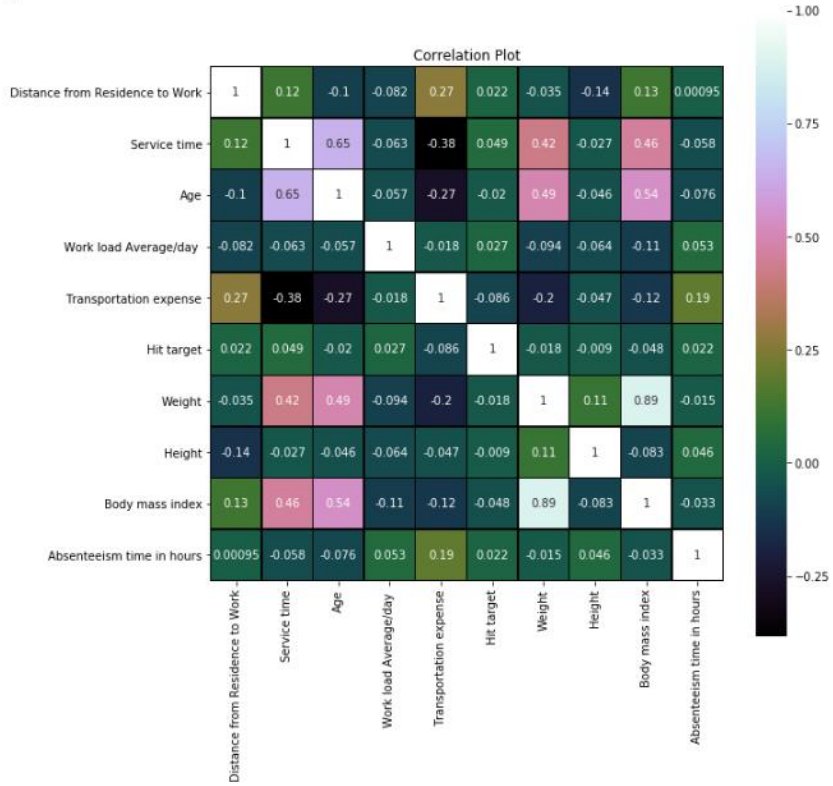


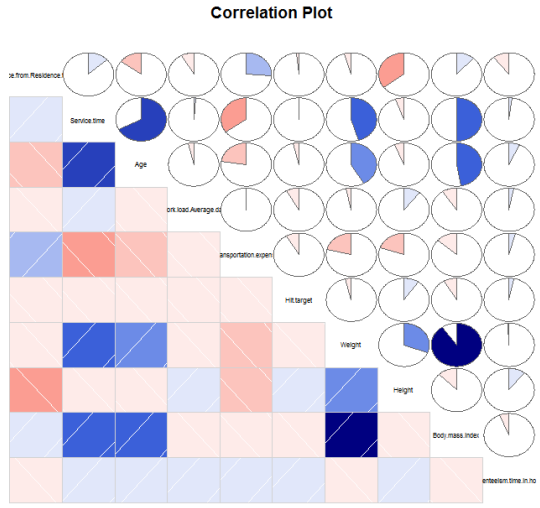
2.1.3 Feature Selection

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of our model. Machine learning works on a simple rule – if we put garbage in, we will only get garbage to come out. By garbage here, I mean noise in data. The data features that we use to train our machine learning models have a huge influence on the performance we can achieve. It is the process where we automatically or manually select those features which contribute most to our prediction variable or output in which we are interested in. Having irrelevant features in our data can decrease the accuracy of the models and make our model learn based on irrelevant features. There are several methods of doing that. We have used the correlation analysis to check collinearity between the variables and anova test to check dependence of target variable on the independent variables. There can be features that aren't relevant for the analysis, we can remove such variables using numerous ways. All the features may not be useful as they might be carrying the same information or irrelevant information and may cause multi- collinearity, and gradually overfit of model. ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. As our target variable is numerical we will use ANOVA for feature selection technique to see whether any categorical variable is related to target variable.

Correlation plot:

[ ]





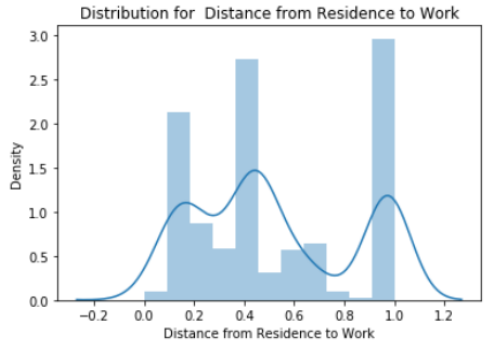
Anova test:

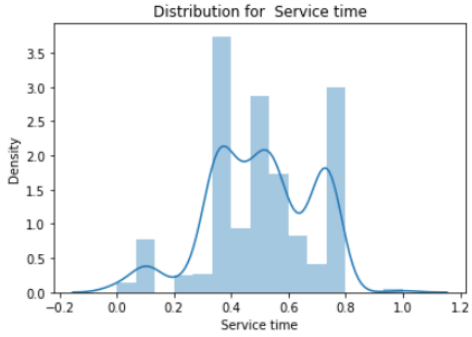
```
P value for ID = 8.762734377787565e-167
P value for Reason for absence = 1.0354352376410707e-276
P value for Month of absence = 3.459399600862653e-25
P value for Day of the week = 0.0008077239229031345
P value for Seasons = 3.0515782325937768e-40
P value for Disciplinary failure = 1.267243302334549e-185
P value for Education = 8.36466800112475e-105
P value for Social drinker = 1.3091608727705665e-150
P value for Social smoker = 9.469186658361471e-184
P value for Pet = 5.348130726125504e-127
P value for Son = 9.463924041308898e-116
```

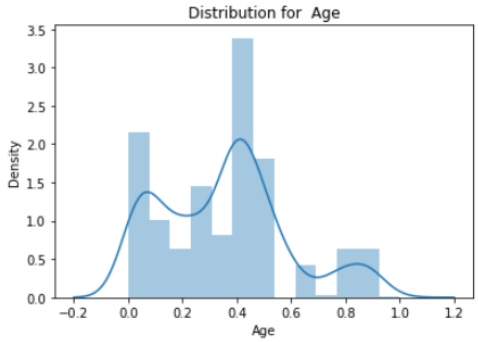
Weight and Body mass index have high correlation and hence we decided to drop the **weight**.

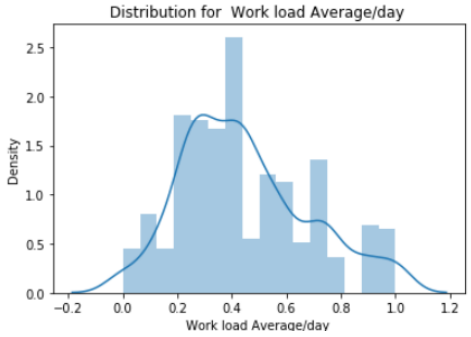
2.2.4 Feature Scaling

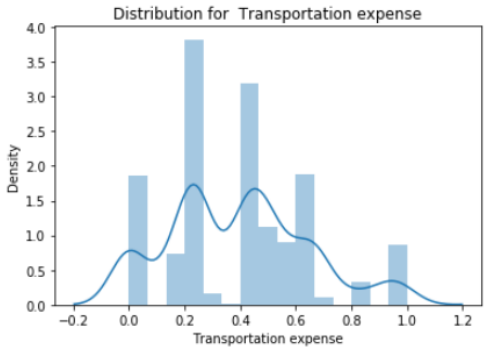
Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use Normalization as Feature Scaling Method.

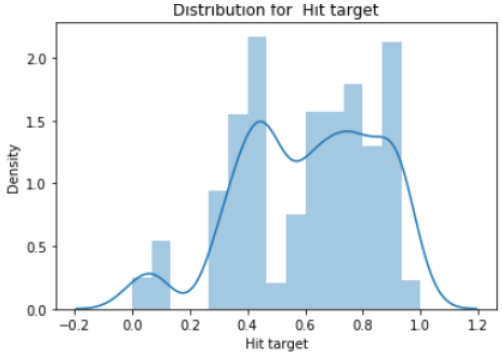


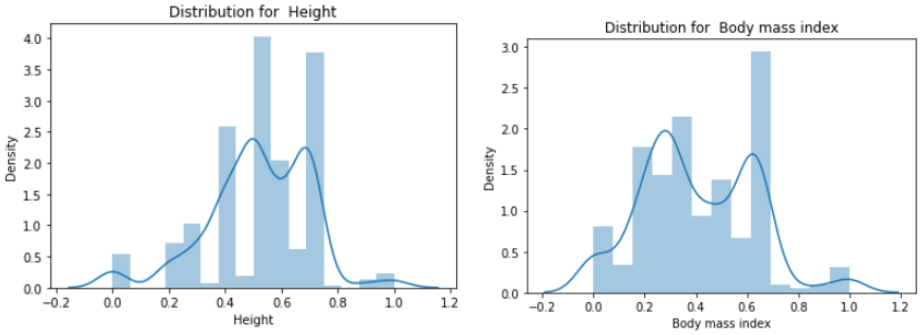






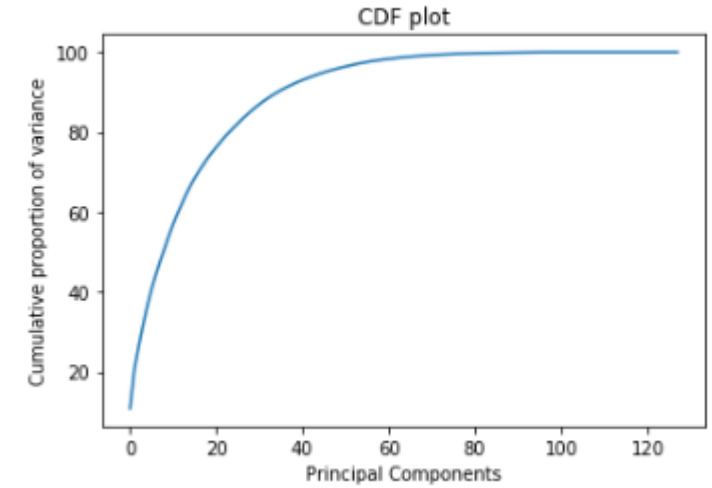






2.2.5 Principal Component Analysis

Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data. After creating dummy variable of categorical variables the shape of our data became 129 columns and 718 observations, this high number of columns leads to poor accuracy. So we are applying PCA technique to reduce the dimensions of our dataset.



We have applied PCA algorithm on our data and from the above graph we have concluded that more than 95% variance is explained by less than 50 variables. After applying PCA now we are left with 574 observations and 47 columns.

2.2 Modelling

We will apply some regression models on our processed data to predict the target variable. As we can see, we have applied all the possible pre-processing analysis to our dataset to make it suitable to input into machine learning models. We have also removed the missing values and outliers. Now since our data is a regression model, we have applied suitable models such as

Decision tree

Random forest

Linear Regression &

Gradient boosting.

2.2.1 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. Each branch connects nodes with “and” and multiple branches are connected by “or”. It can be used for classification and regression. It is a supervised machine learning algorithm. It accepts both continuous and categorical variables as independent variables. It’s easy to understand by the end business users. The RMSE value and R^2 value for our project in R and Python are –

2.2.2 Random Forest

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data. The RMSE value and R^2 value for our project in R and Python are –

2.2.3 Liner Regression

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous variables. To build any model we have some assumptions to put on data and model. Here are the assumptions to the linear regression model.

a)Linear relationship

b)No or little multi-collinearity

c)Multivariate normality

d)Homoscedasticity

2.2.4 Gradient boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Chapter 3

Conclusion

Here we will analyse our models and try to answer the questions asked.

3.1 Model Evaluation

We have calculated **Root Mean Square Error (RMSE)** and **R-Squared** Value of different models. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread

out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable**. Lower values of RMSE and higher value of R-Squared Value indicate better fit.**

3.2 Model Selection

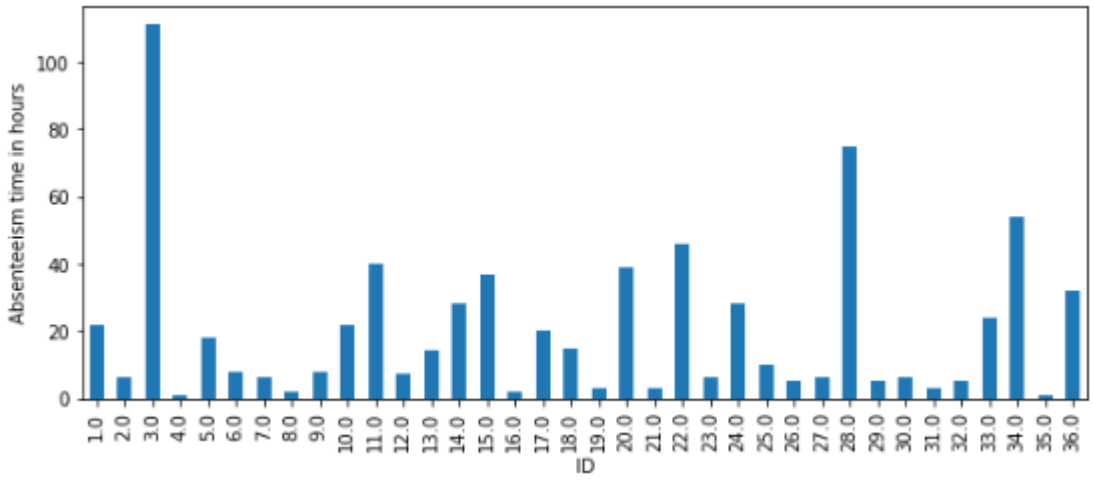
Here according to the problem statement, we are supposed to predict the loss incurred by the company if the same pattern of absenteeism continues. We are selecting **Random Forest** since it has the least RMSE and highest R square value.

	Model Name	RMSE Training	RMSE Test	R2 Score
0	Decision Tree	3.036527	3.368700	0.208495
1	Random Forest	1.052598	3.011765	0.367339
2	Linear Regression	2.416190	3.023604	0.362355
3	Gradient Boosting	1.279680	3.032757	0.358489

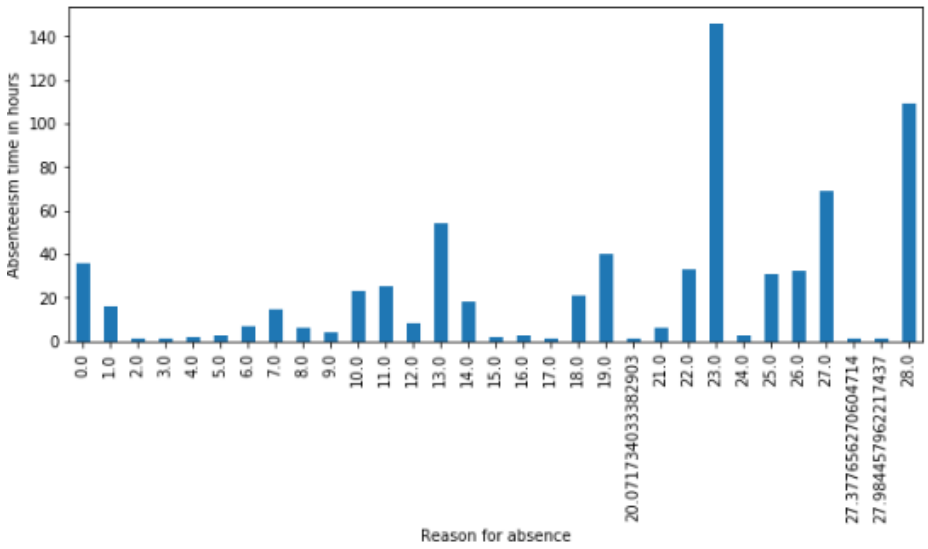
Answer to asked questions

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

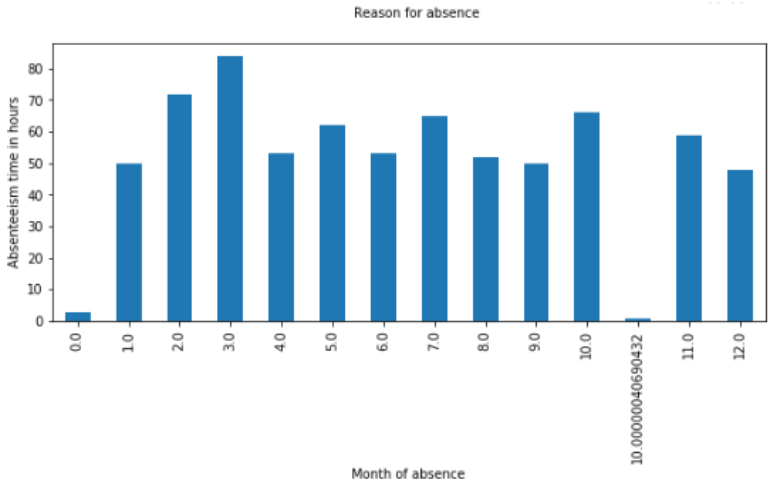
Answer: To answer this question we will do bivariate analysis to find the trend. Lets check the bar plots pasted below to get the inferences:



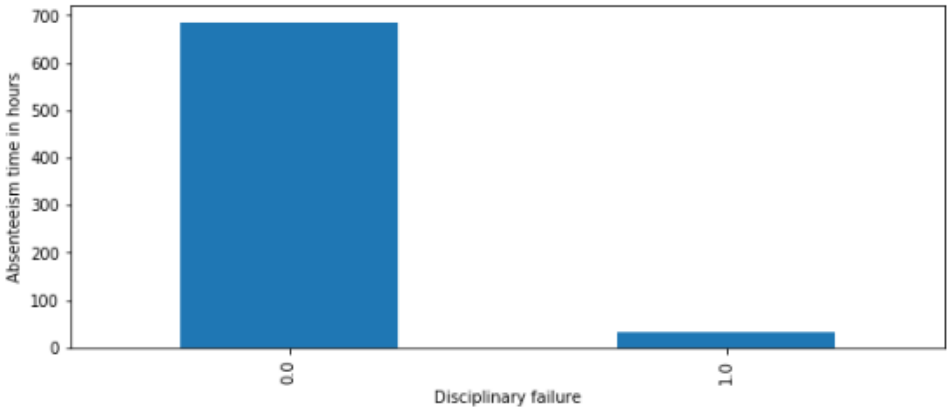
Form the above plot we can clearly see that employees with ID 3, 28, 34 have the most absenteeism hours. To tackle this company can do micro management by conducting meetings with those employees to know the root cause of their absenteeism.



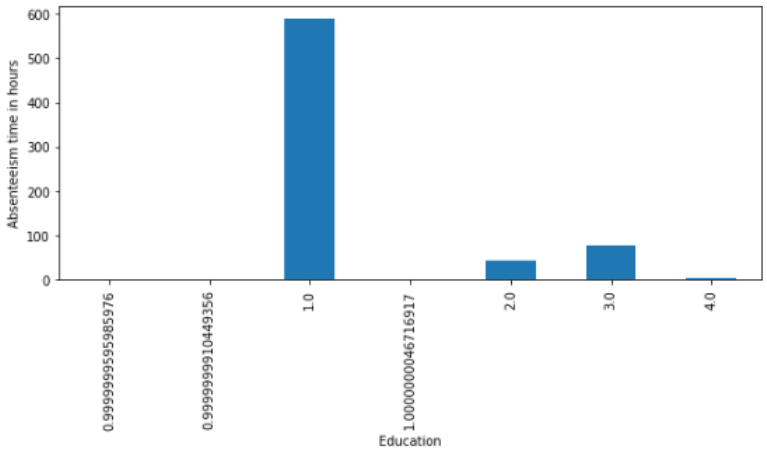
From the above graph it can be seen that reason for absence is mostly due to 23(Medical Consultation) , 27(Physiotherapy) & 28(Dental Consultation). So company can try to provide these three services in-house at discounted prices for employees by availing corporate discount etc. at regular intervals. Company should arrange for sick room and first aid kits.



No specific trend can be identified from the above graph except the fact that in March Employee Absenteeism is at its peak. Company can try to do create a buffer proactively for march workload so that leaves do not impact the work.

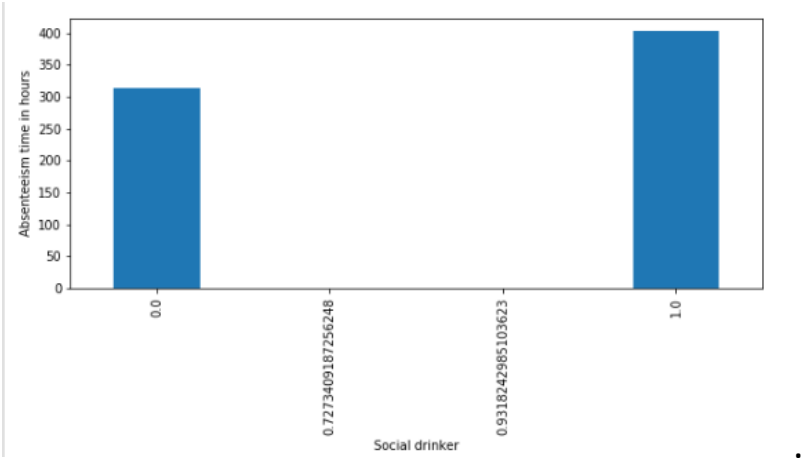


From the above graph we can conclude that disciplinary failure has no connection with employee absenteeism.

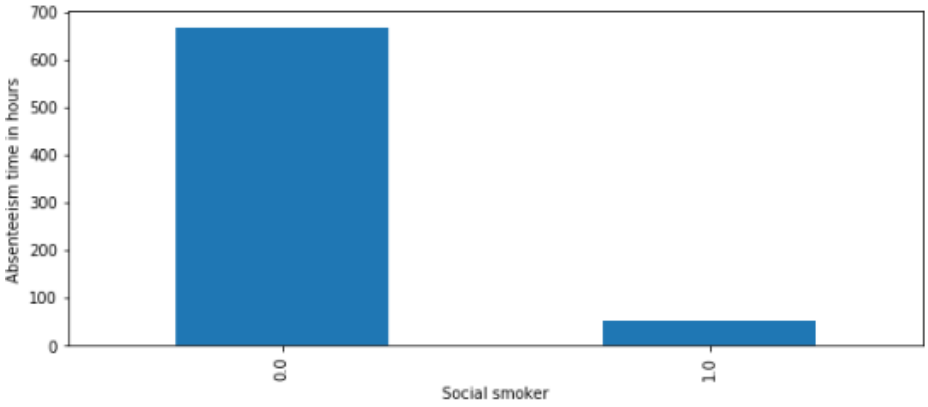


Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

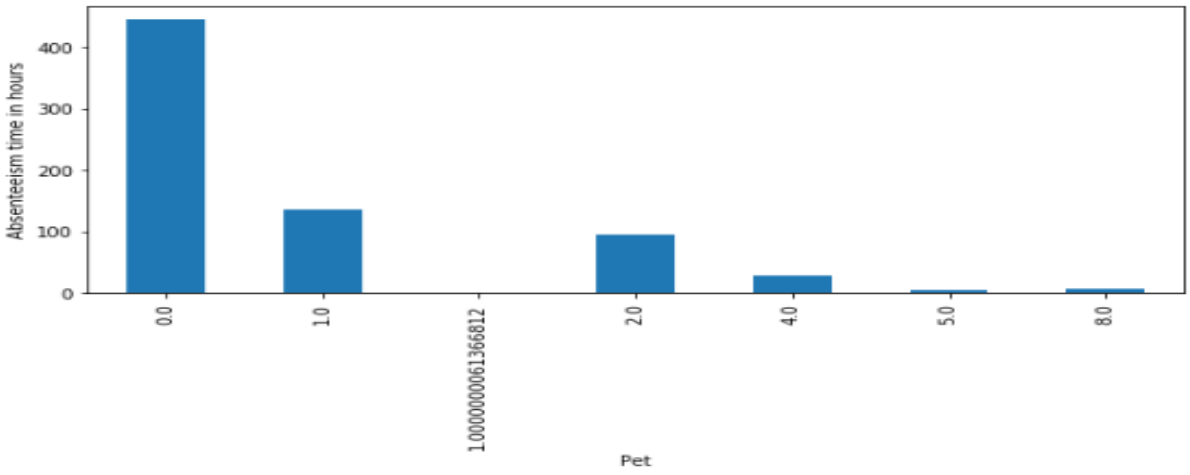
From the above graph we can conclude graduates (freshers) tend to take more leaves, since they are new to corporate world. Company can conduct trainings to teach them about professional behaviour, motivate them



Those who are social drinker tends to take more leaves, since they might suffer from hangover. Company should run campaigns to teach employees about the harmful effects of alcohol intake.



It can be seen from the above graph that those who smoke tend to take less leaves, they might be good at handling pressure by smoking it away. Company can look for alternative ways to reduce the stress of employees by conducting games and other forms of co-curricular activities.



From the above graph it can be clearly seen that employees with at least one pet tends to take less leaves than employees with no pets. Company can try to motivate employees to adopt pets.

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

Answer: Please see attachment **Employee_Absenteeism_2011.csv** and **Monthly_loss.csv**