Project Report

Employee Absenteeism

Vikrant Verma

30 October 2018

Contents

1. Introduction

2. Methodology

3. Conclusion

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

There are 21 variables in our data in which 20 are independent variables and 1 (Absenteeism time in hours) is dependent variable. Since our target variable is a continuous variable, this is a regression problem.

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 behavioral 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 (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)

1.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics. In the given data set there are 21 variables and data types of all variables are either float64 or int64. There are 740 observations and 21 columns in our data set. Missing value is also present in our data.

List of columns and their number of unique values -

ID	36
Reason for absence	28
Month of absence	13
Day of the week	5
Seasons	4
Transportation expense	24
Distance from Residence to Work	25
Service time	18
Age	22
Work load Average/day	38
Hit target	13
Disciplinary failure	2
Education	4
Son	5
Social drinker	2
Social smoker	2
Pet	6
Weight	26
Height	14
Body mass index	17
Absenteeism time in hours	19

From EDA we have concluded that there are 10 continuous variable and 11 categorical variable in nature.

Chapter 2

Methodology

Before feeding the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project we spend almost 80% of time in it.

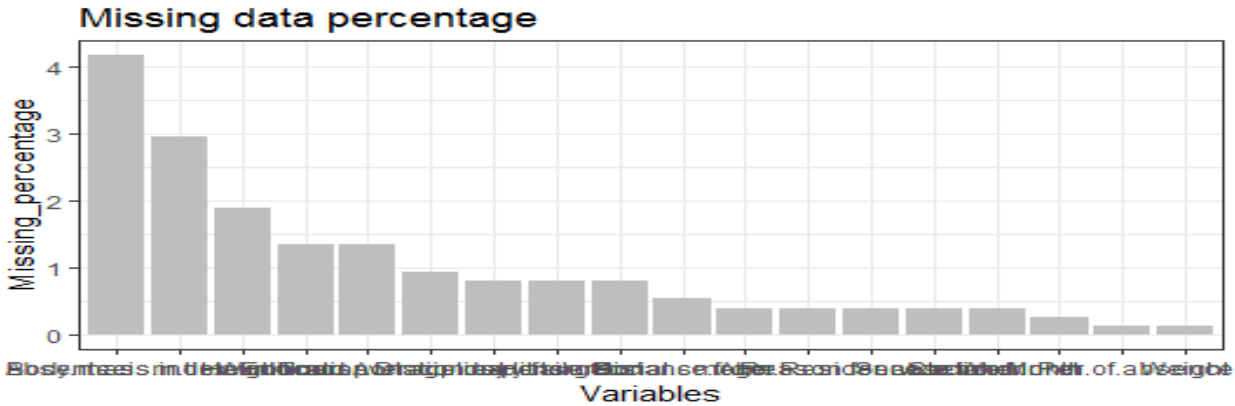
2.1 Pre Processing

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.

2.1.1 Missing Value Analysis

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a columns has more than 30% of data as missing value either we ignore the entire column or we ignore those observations. In the given data the maximum percentage of missing value is 4.189% for **body mass index** column. So we will compute missing value for all the columns.

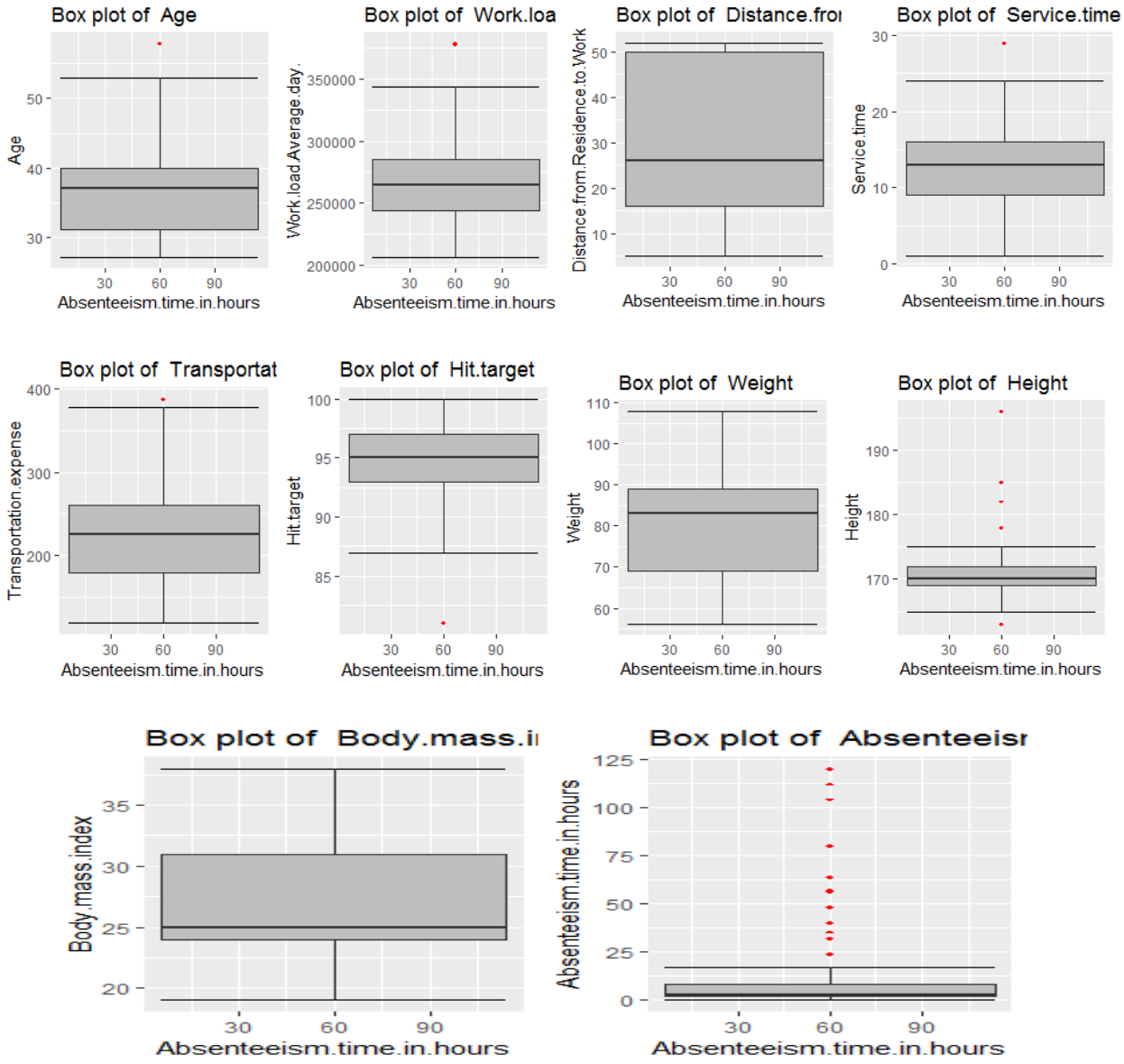
In this project we have used KNN imputation method to impute missing value.



2.1.2 Outlier Analysis

. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.

In figure we have plotted the boxplots of the 11 predictor variables with respect to **Absenteeism time in hour**. A lot of useful inferences can be made from these plots. First as you can see, we have a lot of outliers and extreme values in each of the data set.

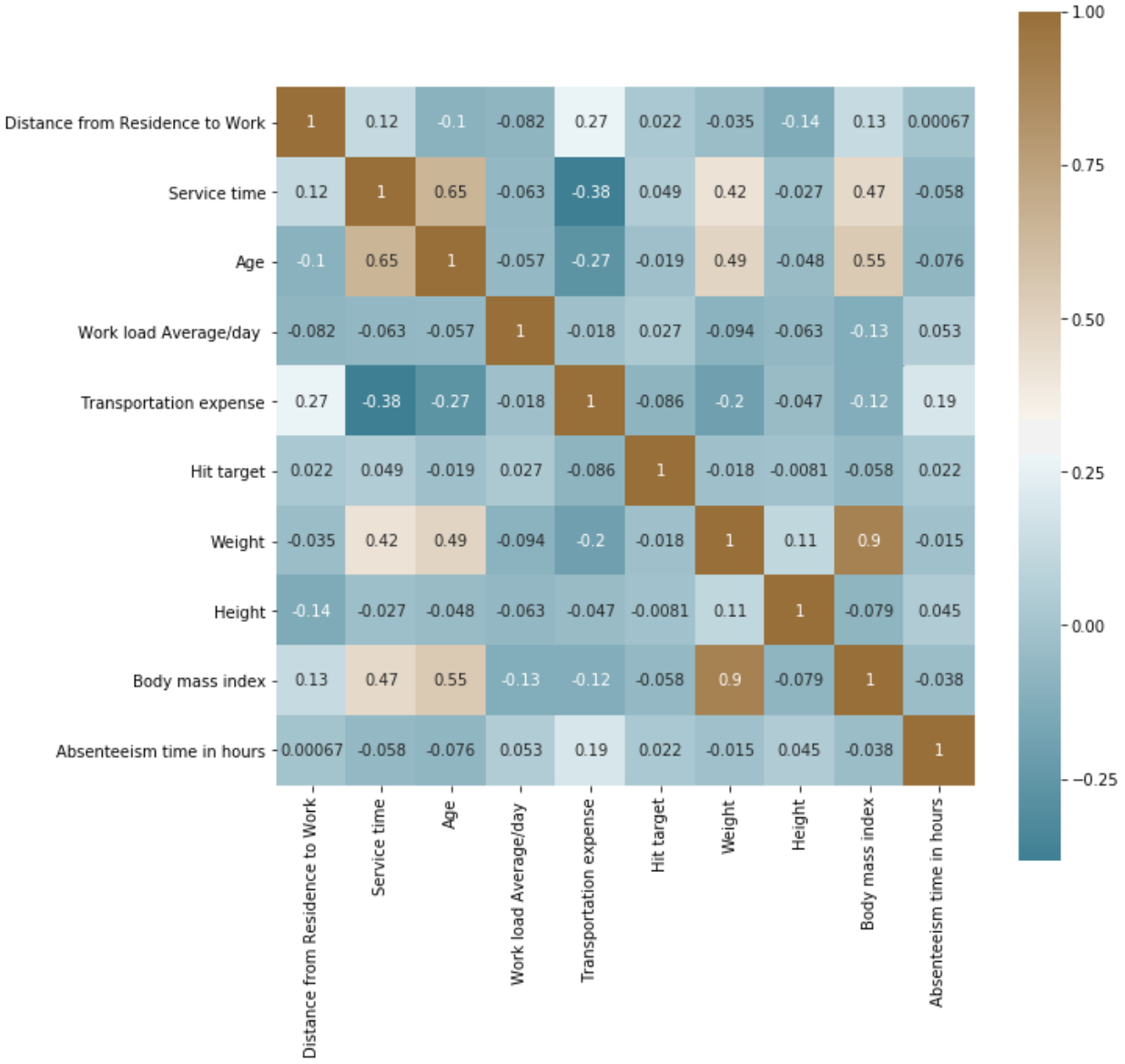


From the boxplot almost all the variables **except "Distance from residence to work", "Weight" and "Body mass index"** consists of outliers. We have converted the outliers (data beyond minimum and maximum values) as NA i.e. missing values and fill them by **KNN** imputation method.

2.1.3 Feature Selection

Before performing any type of modeling 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 important at all to the problem of class prediction. Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase

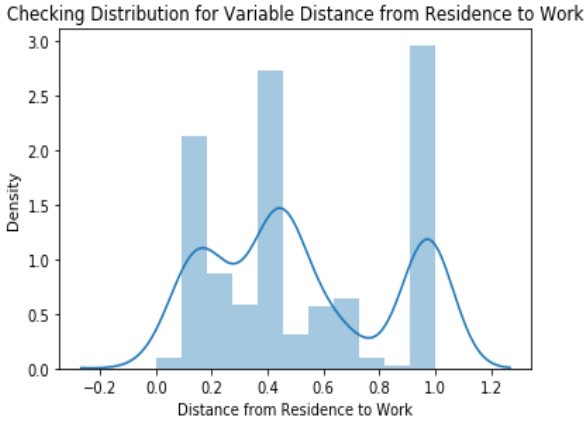
overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi collinearity. In this project we have selected **Correlation Analysis** for numerical variable

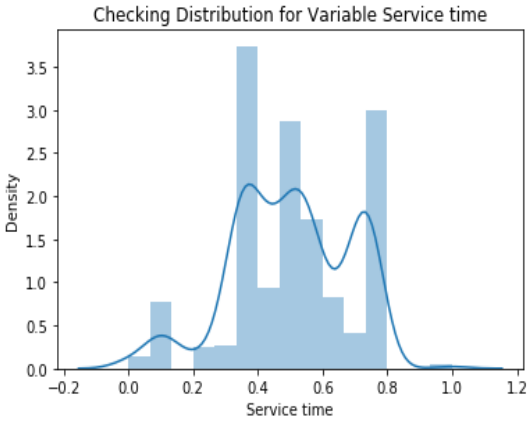


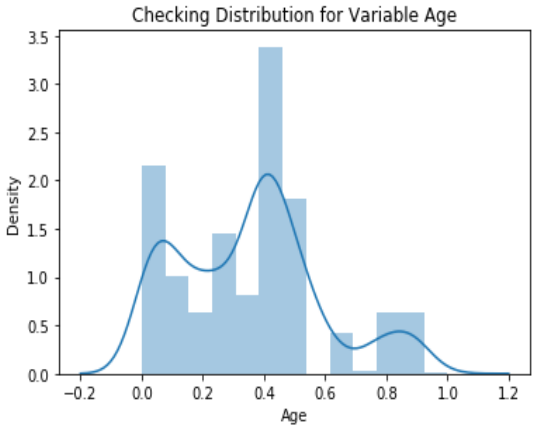
From correlation analysis we have found that **Weight** and **Body mass index** has high correlation (>0.7), so we have excluded the **Weight** column.

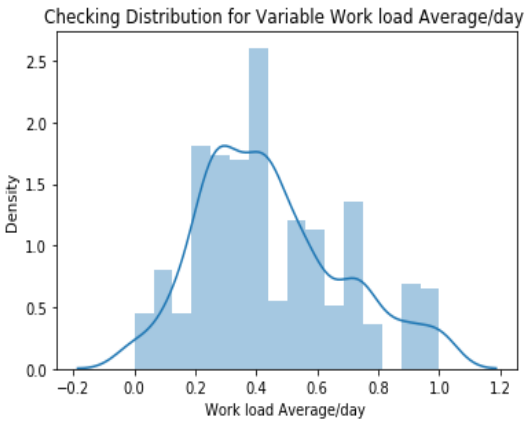
2.1.4 <u>Feature Scaling</u>

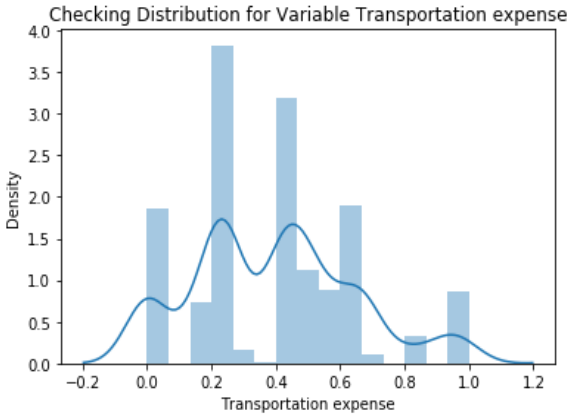
Checking distribution curve for all continuous variable

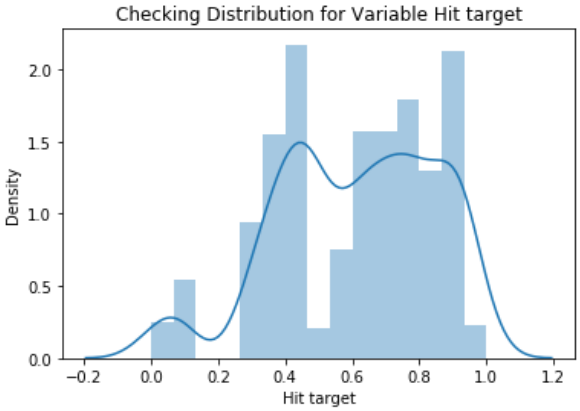

Checking Distribution for Variable Distance from Residence to Work
Density
Distance from Residence to Work

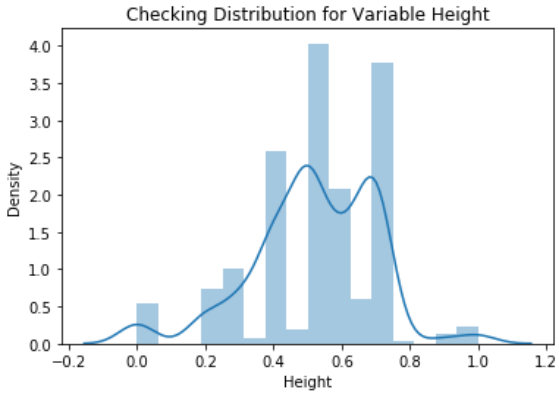

Checking Distribution for Variable Service time
Density
Service time

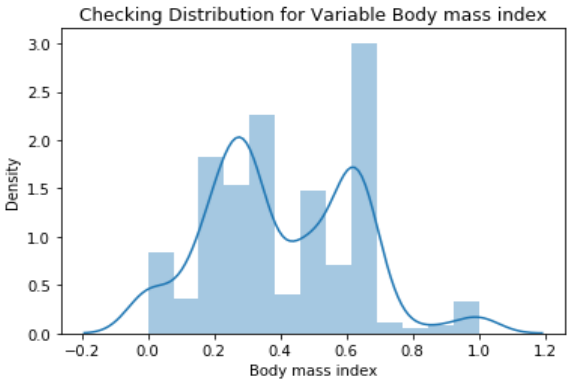

Checking Distribution for Variable Age
Density
Age


Checking Distribution for Variable Work load Average/day
Density
Work load Average/day


Checking Distribution for Variable Transportation expense
Density
Transportation expense


Checking Distribution for Variable Hit target
Density
Hit target





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 preprocessing 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.

2.1.5 Creating Dummy variables

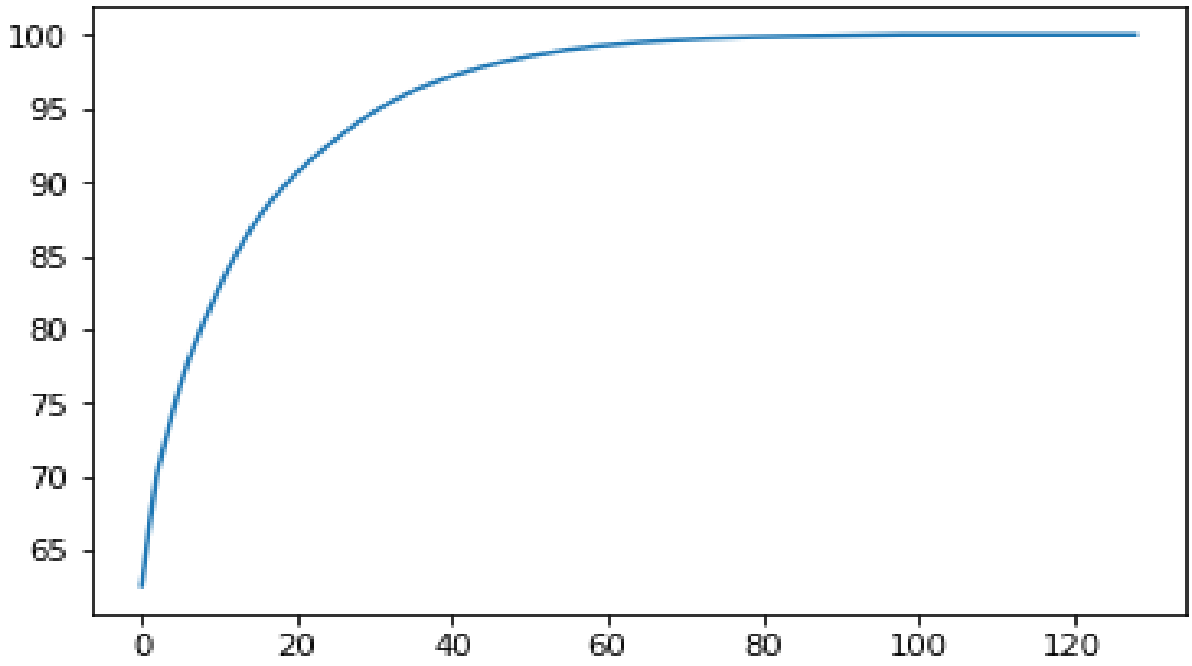
As in given data set target variable is a continuous variable so this is a regression problem so before developing a regression model we have to convert all multi-level categorical variable into binary dummy variable

2.1.6 Sampling

The dataset is been divided into 2 parts train data to train a machine learning model and test data to test the model accuracy
The data is divided into 8:2 ratio that means 80% of data is training data and rest 20% data is test data

2.1.7 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. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. 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 116 columns and 740 observations, this high number of columns leads to bad accuracy.



We have applied PCA algorithm on our data and from the above graph we have concluded that 45 variables out of 116 explains more than 95% of data. So we have selected only those 45 variables to feed our models.

2.2 Modeling

After a thorough preprocessing we will be using two regression model on our processed data to predict the target variable.

2.2.1 Liner Regression

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

Linear Regression	R	PYTHON
RMSE Train	0.002	1.9988146e-15
RMSE Test	0.003	3.4261115e-09

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 –

Random Forest	R	PYTHON
RMSE Train	0.249	0.044
RMSE Test	0.704	0.130

Chapter 3

Conclusion

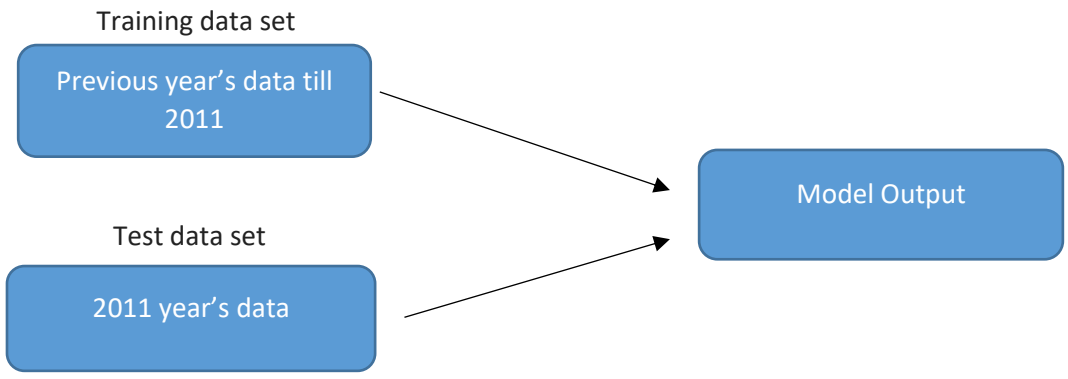
In this chapter we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

3.1 Model Evaluation

In the previous chapter we have seen the **Root Mean Square Error** (RMSE) 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. **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** indicate better fit.

3.2 Model Validation

OUT TIME VALIDATION is used for model validation
In out time validation process we have to compare performance major of both training data prediction and test data prediction with their actual values. For a model to pass this validation test the difference between performance major of training and test prediction with actual values should be minimum. Or in other words performance major of training and test prediction should be similar.



In our case we are assuming the train data is the previous year data and test data as the data for 2011

So on comparing both train and test performance majors

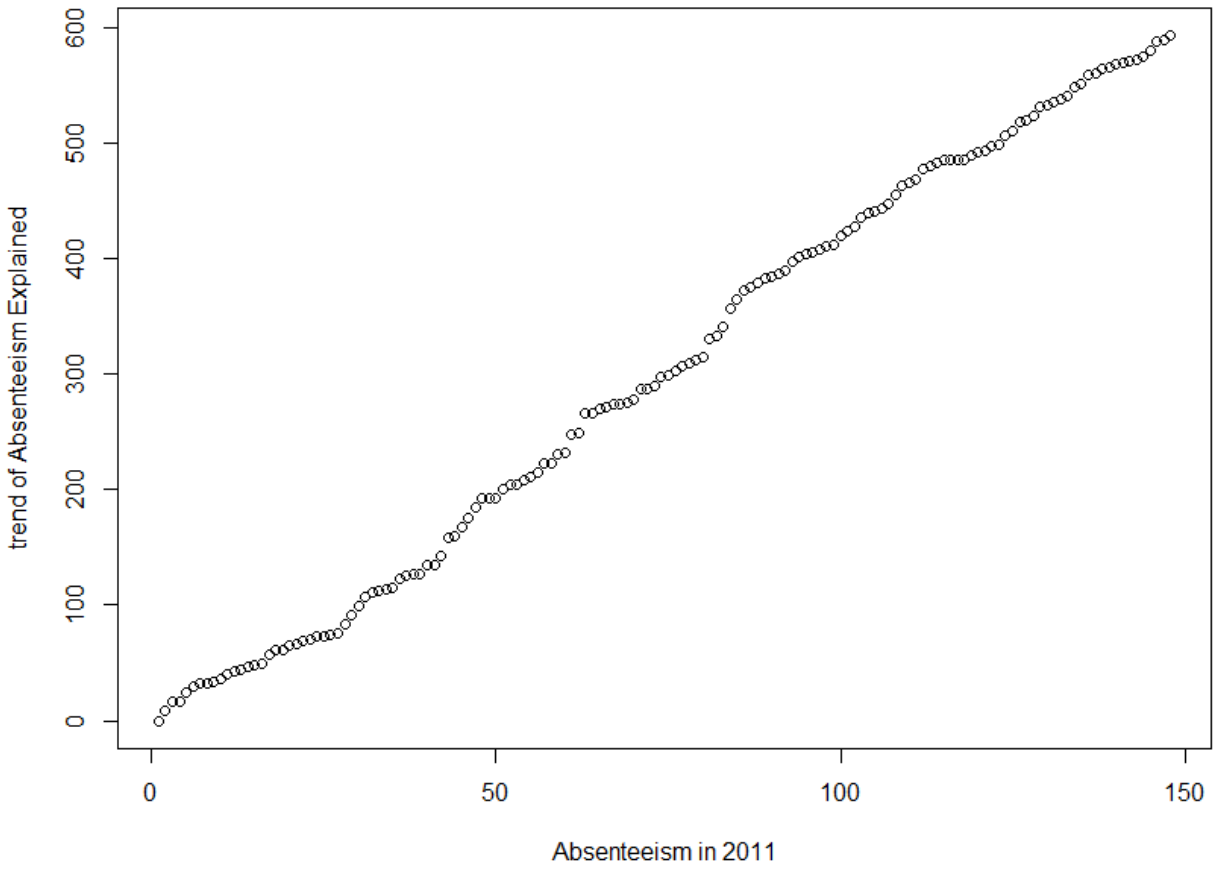
Linear Regression	R	PYTHON
RMSE Train	0.002	1.9988146e-15
RMSE Test	0.003	3.4261115e-09

Random Forest	R	PYTHON
RMSE Train	0.249	0.044
RMSE Test	0.704	0.130

From above observation it is clear that Linear regression model is best suited for this case of validation

3.3 VISUALIZATIONS

ANALYSING ABSENTEEISM TREND ON TEST DATA



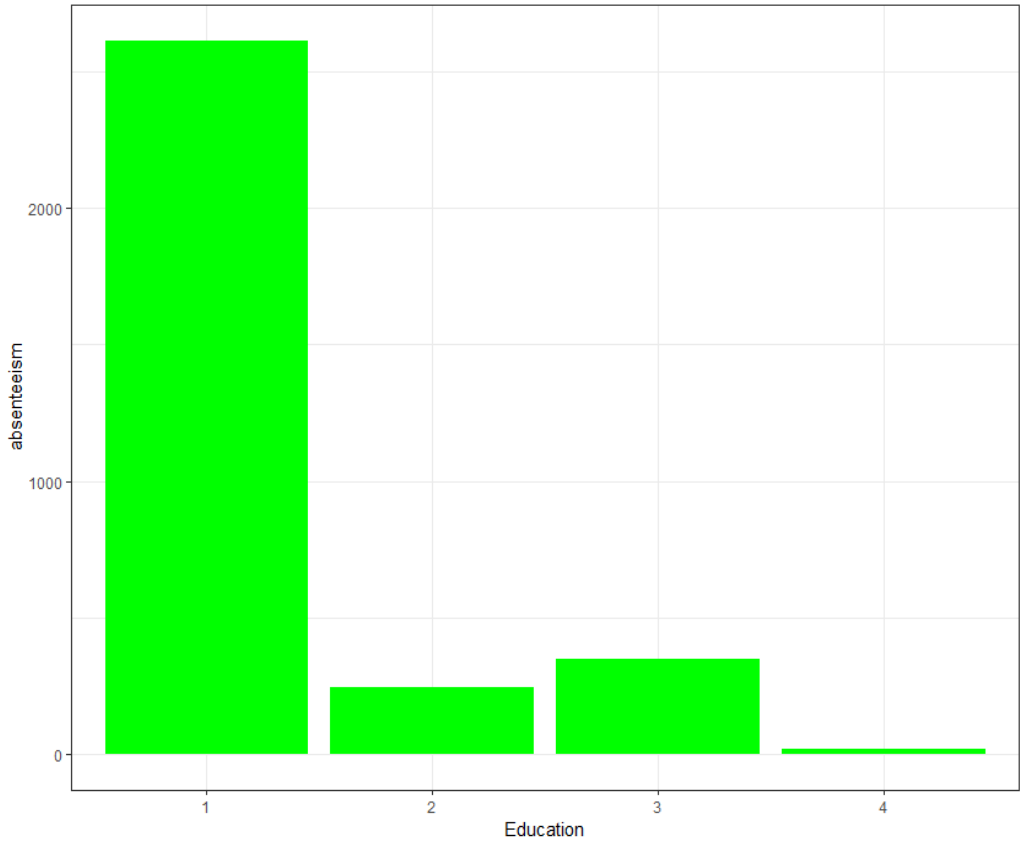
The above graph represents the trend of absenteeism in 2011 as the test data is assumed to be the data for 2011

As in the graph its clearly noticed that the trend of absenteeism is increasing at a constant rate and is forming almost a straight line. Now by a firm's management an employee's absenteeism is only taken care by deducting leaves from employee's kitty or by deducting salary for any absent day. But in actual The increase in absenteeism trend will directly lower the productivity rate which causes much Loss to company that cannot be compensated by deducting an employee's salary. Hence to cope up with the increasing rate of absenteeism a company should introduce some changes to reduce the increasing trend of absenteeism.

Below are some suggestions and visualizations that a company should take forward to reduce its rate of absenteeism based on the data provided

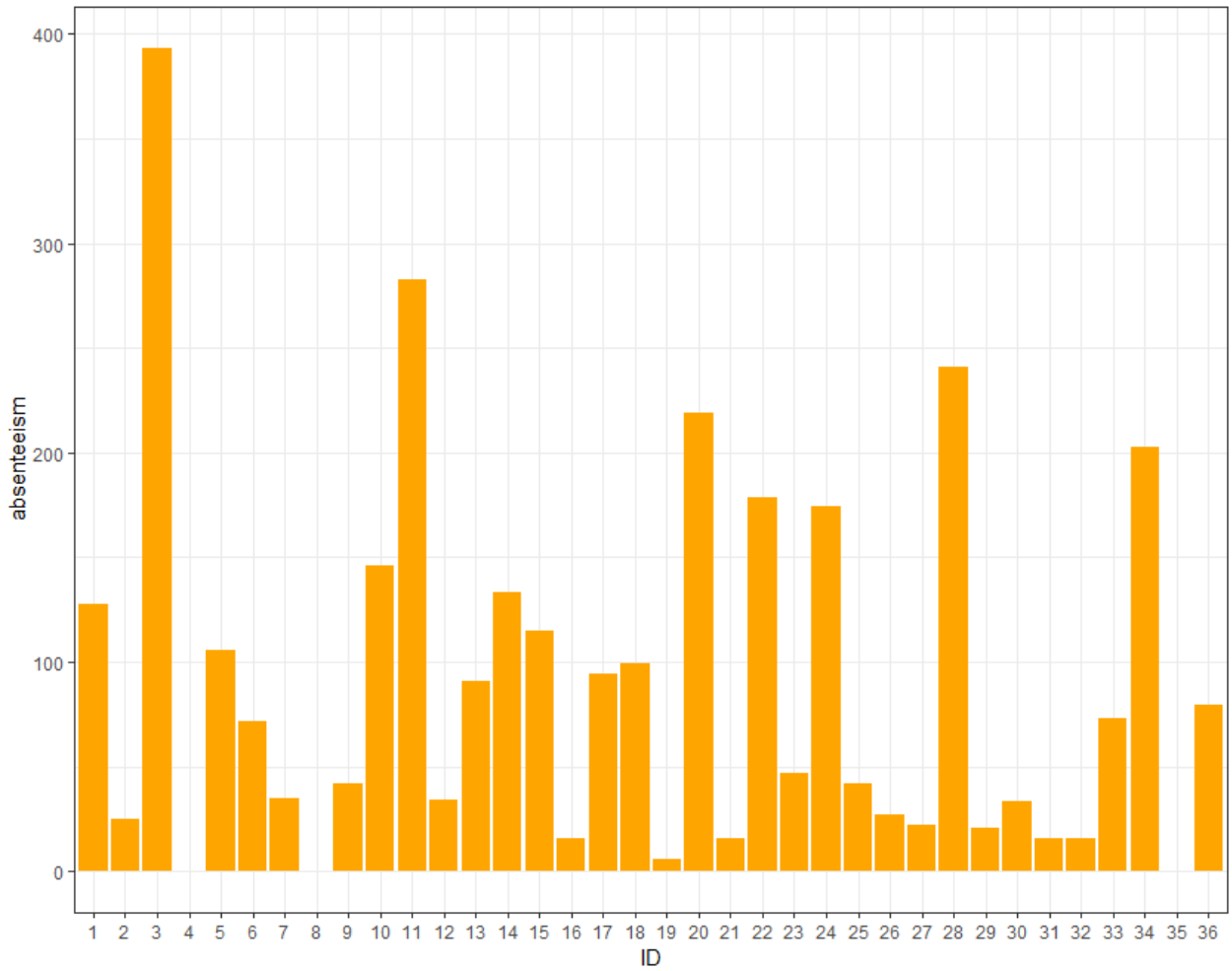
The Changes which company should bring to reduce the number of absenteeism –

1. It is observed that employee with low education have maximum absentee time.

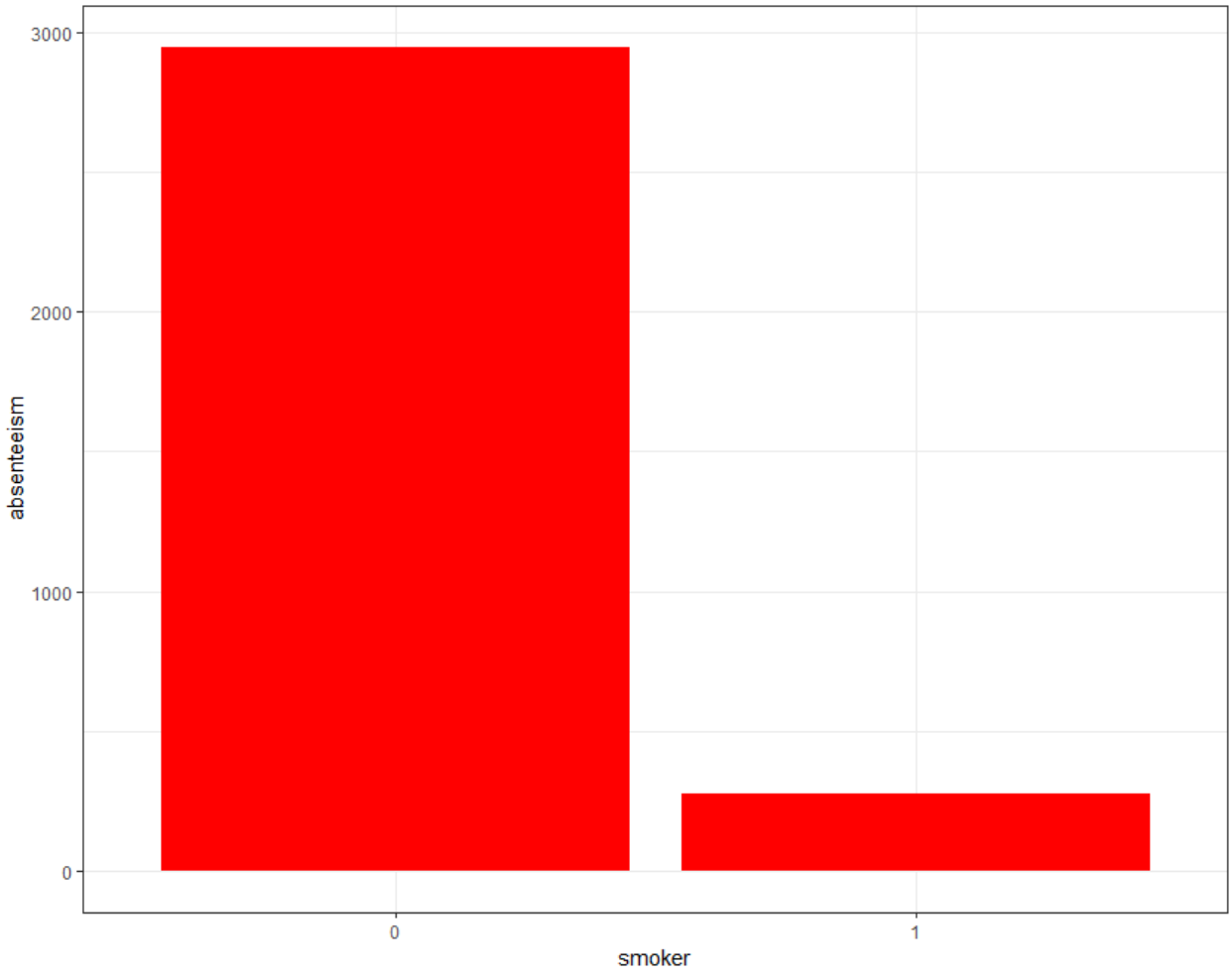


A firm should hire educated staff as staff with low education is having most absenteeism hours.

2. Some employee with **ID 3, 28, 34** are often absent from work, company should take action against them. Or even a warning to them might help

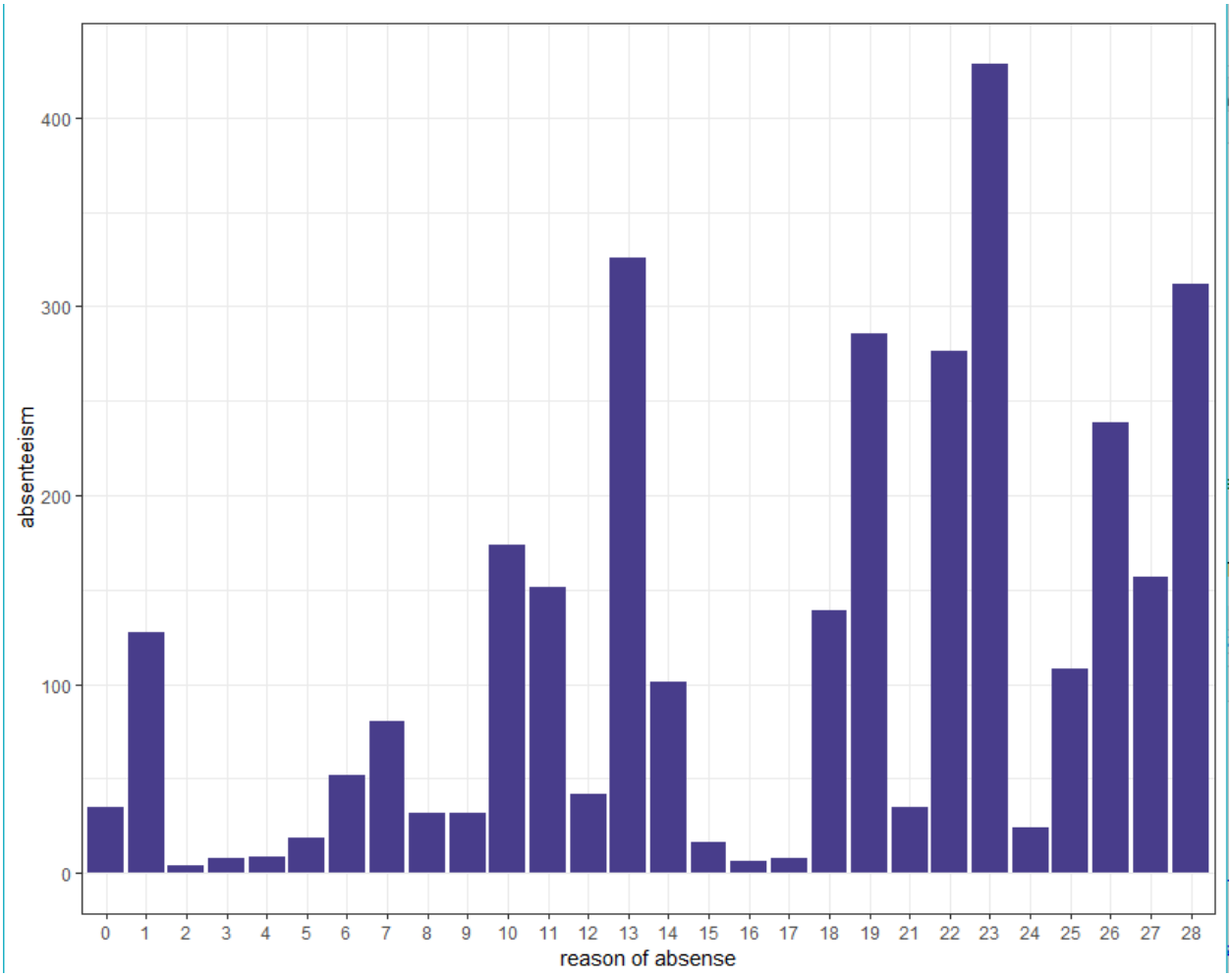


3. Employees who are social smoker have more absentee hour than who are not social smoker.



As a smoker is more prone to bad health condition so that causes a lot of absenteeism. So a firm should conduct health campaigns to educated employee about the harmful effects of smoking.

4. Most often Reason for absence are medical consultation and dental consultation, company should take care of it. The maximum people taking the absent hours are from category 23 followed by 28 and 27. These category are not attested by doctors. 23: Medical Consultation. 28: Dental consultation 27: Physiotherapy .



Other than the above statements

- A company should introduce 100% attendance incentive bonus.
- Should introduce new policy for salary deduction for un-informed or un-approved leaves
- A sandwich leave policy would be a good option to reduce absenteeism on day near to weekends