A Project On

Employee Absenteeism

By

ABHISHEK KANKATI

on 25/09/2019

Table of Contents

Chapter I

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 Variables

Number of Attributes: 21
Missing Values: Yes
Attribute 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)

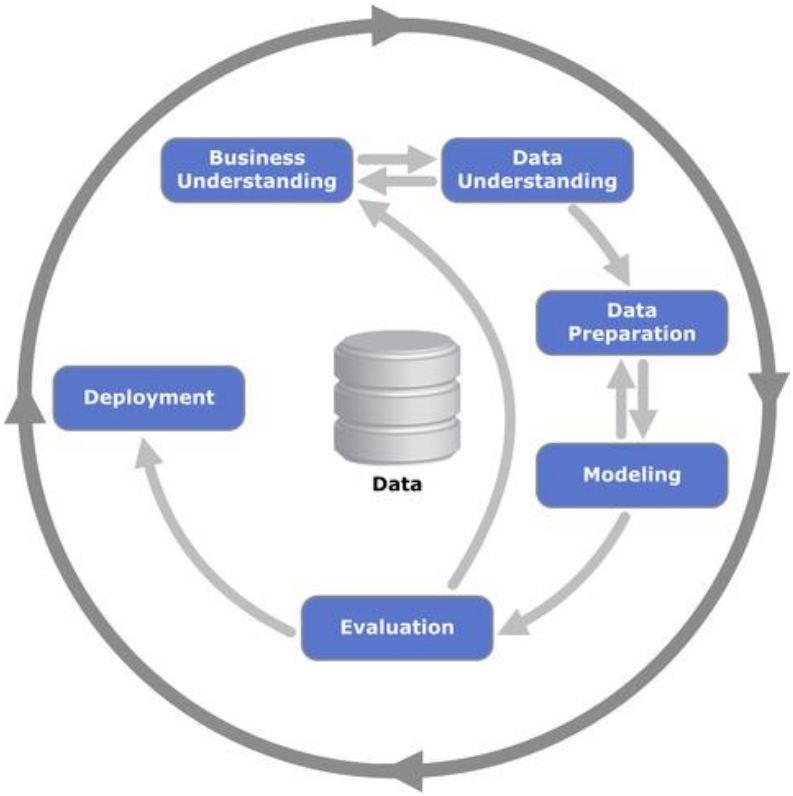
1.3 Sample Data

	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	...	Disciplinary failure	Education	Son	Social drinker	Social smoker
0	11	26.0	7.0	3	1	289.0	36.0	13.0	33.0	239554.0	...	0.0	1.0	2.0	1.0	0.0
1	36	0.0	7.0	3	1	118.0	13.0	18.0	50.0	239554.0	...	1.0	1.0	1.0	1.0	0.0
2	3	23.0	7.0	4	1	179.0	51.0	18.0	38.0	239554.0	...	0.0	1.0	0.0	1.0	0.0
3	7	7.0	7.0	5	1	279.0	5.0	14.0	39.0	239554.0	...	0.0	1.0	2.0	1.0	1.0
4	11	23.0	7.0	5	1	289.0	36.0	13.0	33.0	239554.0	...	0.0	1.0	2.0	1.0	0.0

Pet	Weight	Height	Body mass index	Absenteeism time in hours
1.0	90.0	172.0	30.0	4.0
0.0	98.0	178.0	31.0	0.0
0.0	89.0	170.0	31.0	2.0
0.0	68.0	168.0	24.0	4.0
1.0	90.0	172.0	30.0	2.0

1.3 CRISP DM Process (Cross Industry Standard Process Data Mining)

CRISP-DM is a cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. It can be explained by the below figure.



1. Business Understanding

This step mostly focuses on understanding the Business in all the different aspects. It follows the below different steps.

a. Identify the goal and frame the business problem.

b. Gather information on resource, constraints, assumptions, risks etc.

c. Prepare Analytical Goal

d. Flow Chart

As per the problem statement, XYZ is a courier company. We know that courier companies completely depend upon the manpower for fulfilling their daily works. It means that human capital plays an important role in collection, transportation and delivery. In the absence of manpower, they undergo a loss. The XYZ company is passing through a genuine issue of Absenteeism. So, let's work on their data, understand the

reasons for the absenteeism and provide them a permanent solution to decrease the absenteeism.

2. Data Understanding

Data Understanding phase of CRISP DM Framework focus on collecting the data, describing and exploring the data.

Exploring the data involves analyzing the data in hand for

· Dependent and Independent Variable Identification.

· Uni-variate Analysis – Exploring each independent variable

· Bi-variate Analysis – Exploring the different combination of two or more variable using Correlation, Chi-Square Test, T-Test, Z-Test etc. This step also involves subcategory analysis of each independent variable on the dependent variable.

· Aggregated data exploration

· Data Quality Check is also performed at the step.

In a ML Implementation, we may get numerous independent variables that may or may not contribute to the prediction of Dependent Variable. This step of data understanding at times even gives us a gist of attributes which may be important for predicting the dependent variable.

3. Data preparation

In this step, we prepare and clean the provided data. There are many steps that involves in this step as mentioned below.

a. The first and foremost step being the NA (null values) treatment. Normally the data at hand is not clean and at most of the times our data will have NA. We must identify such values and appropriately fill or impute them. There are many different techniques of NA treatment and there are packages in R and Python which automatically treat such variables based on some default logic. However, it is always good to do it manually, as this way we get to understand the data even further and can replace these NA's with our understanding of Business Requirement.

Imputing Missing Values

b. The next step would be to treat Null's. This step is equally important as NA treatment and as per my experience, I have below steps for the Null treatment.

i. If the variable is Continuous in Nature (Numeric Variable), we can use Mean/Median/Mode/KNN imputation technique for the missing value treatment.

ii. If the variable is categorical in nature, we can impute the Nulls with "Unavailable" as these Null values or Unavailable values may contribute significantly to the Model creation and we do not want to lose any important attribute.

Outlier Treatment

c. Outliers are the values in continuous variables that are inconsistent (may be very far from the mean) with data. These outliers will drastically impact our mean value. So, treating outlier will be our next step post imputing missing values. We have a few packages in R and python to remove the outliers or try to impute them with again mean/median/mode/KNN methods. We can display such variables using Boxplot of the attribute.

Feature Scaling

d. Scaling of the data – This step is used to scale up the values of the attribute so that they lie between 0 to 1. With scaling, the range of the variables get reduced and result in a better predicting variable. This could be done on Continuous Variables. If the data is normally distributed, we will use Standardization technique. Otherwise we will go with Normalization technique.

Feature Engineering

e. Feature Engineering: One of the most important steps and can be clubbed as the combination of Feature Transformation and Feature Extraction. In this step, we try to create or extract more attributes from the available attributes with some business sense. The more we explore, the better we can extract with different relations from the existing variables. The examples are below,

i. Create Value Transformation like Square or Cube or even Square-root or Cube root or Log of certain columns, as it has been seen that such derived columns contribute in algorithm then the deriving column.

ii. Variable Creation like creating dummy variables from the categorical variables, Data Split etc. also contribute to Feature Engineering.

Feature Engineering step is one such step which can be explored more and can contribute significantly to the outcome.

Dimensionality Reduction (Feature Selection)

f. Dimensionality reduction is a series of techniques in machine learning and statistics to reduce the number of random variables to consider. It involves feature selection and feature extraction. Dimensionality reduction makes analyzing data much easier and faster for machine learning algorithms without extraneous variables to process, making machine

learning algorithms faster and simpler in turn. Assume we have 1000+ variables in our dataset. Methods like Principal Component Analysis and Factor Analysis will help us to find the most important variables that explain our target variable to the best in these 1000 variables and we can use those variables to train our data. Thereby reducing the complexity of our data and increase the model's speed and the performance.

4. Modeling

Once the above steps are done, we have implemented the necessity of machine learning and now we can proceed with the implementation of different ML algorithms. The algorithm to be selected depends completely on the business requirement, available data and the desired outcome. In an ideal situation, we should try different algorithm or combination of algorithm (ensembles) to arrive at our final best algorithm. For our problem statement, I have used linear regression, decision tree, random forest and Gradient Boosting algorithms.

5. Evaluation of the Model

There are many model evaluation techniques like Accuracy, Sensitivity, Specificity, F-Score, R-Squared, Adj R-SQ, RMSE (Root Mean Square Error), MAPE etc. I have considered RMSE as my evaluation metric.

6. Deployment

Finally, once the model is created, tested and evaluated on the Test and Validation data, this is presented to the business (with PPT). The model the undergoes different real time evaluation and testing like A/B Testing and after all the approval process, the code is pushed to the PROD/Live data.

Chapter II

METHODOLOGY

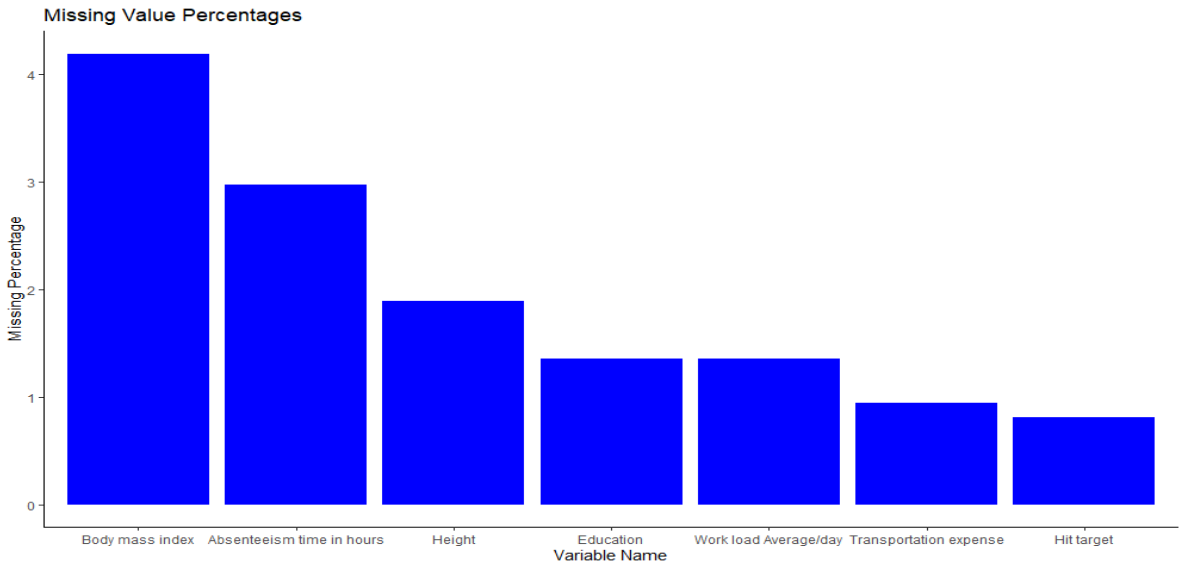
2.1 Pre – processing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

For achieving better results from the applied model in Machine Learning projects the format of the data must be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore, to execute random forest algorithm null values must be managed from the original raw data set. Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen. So, as part of pre – processing techniques, we should explore the data and analyze it, impute the missing values, treat the outliers, feature engineering, feature selection and reduce the dimensions of the dataset by selecting the important features of all the features.

2.2 Missing Value Analysis

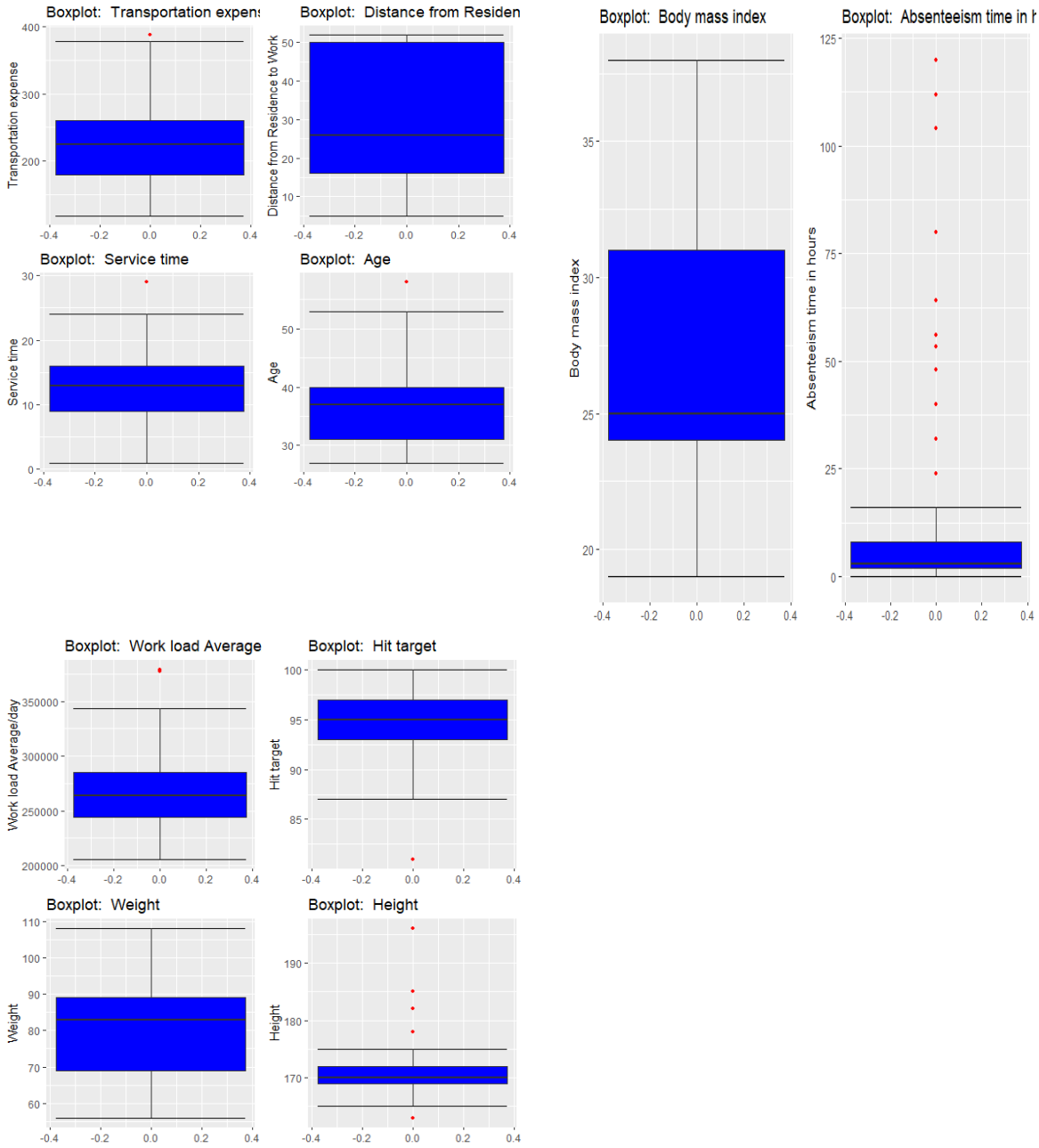
Missing values occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions or insights that can be drawn from the data. So, if a variable in a dataset has more than 30% of the missing values, then we can drop that variable. Otherwise these values can be filled by using central statistics or KNN imputation or prediction methods. The missing values for the given data are plotted and is mentioned below.



From the above plot, we can observe that variable, *Body Mass Index* has the highest percentage of the missing values with 4.18%. These missing values are imputed by using KNN method as it is giving the nearest value in our experiment.

2.3 Outlier Analysis

Outliers are the numeric values that are inconsistent with our data. Mean will have a drastic impact because of outliers. Outliers can explain the skewness in the data. So, these outliers are to be detected and should be removed. I have used *boxplot* method to detect the outliers. Then I have replaced them by NA and used KNN imputation technique to impute those missing values. Below are the boxplots plotted for each continuous variable.



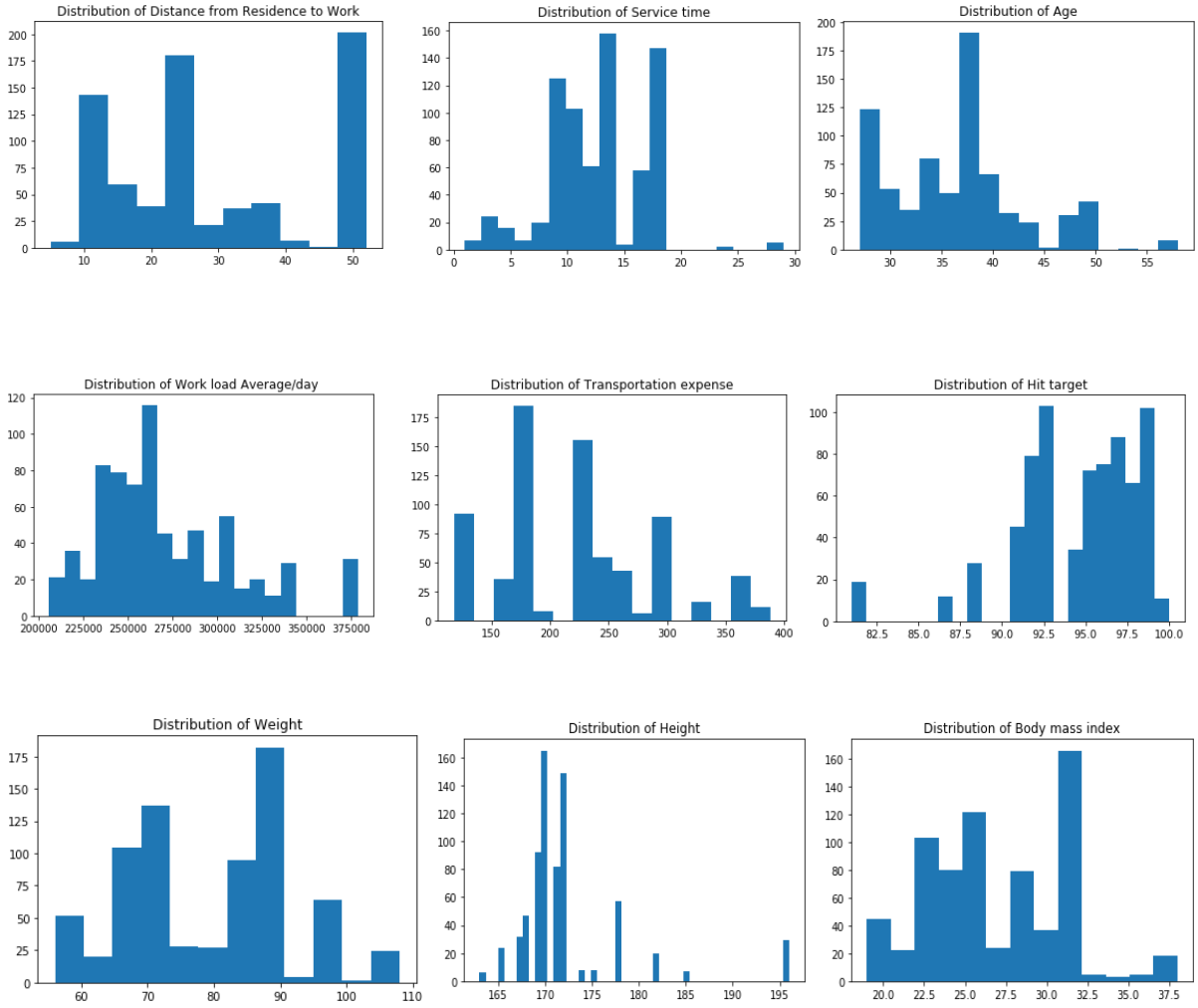
From the above plots it can be concluded that except for variables, Body Mass Index, Distance from residence to work and Weight, all other variables has outliers. These outliers are treated and imputed as mentioned above.

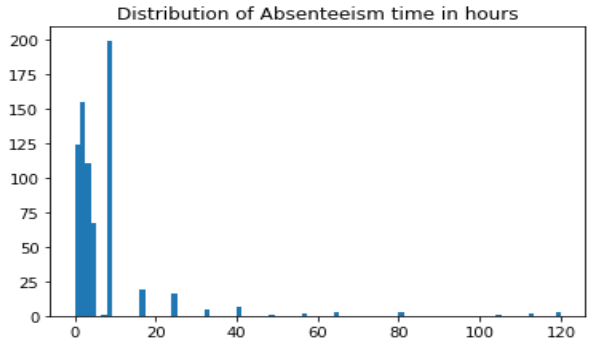
2.4 Distribution of the variables

By using various plots, we find the distribution of the variables. This is exploratory data analysis, from which we'll get a basic idea on how the variables are distributed, basic statistical parameters like mean of the variable, if the variable is uniformly or non-uniformly distribution etc.

2.4.1 Distribution of Continuous variables

To check the distribution for continuous variables, I have used histograms. The distribution plots are as below,

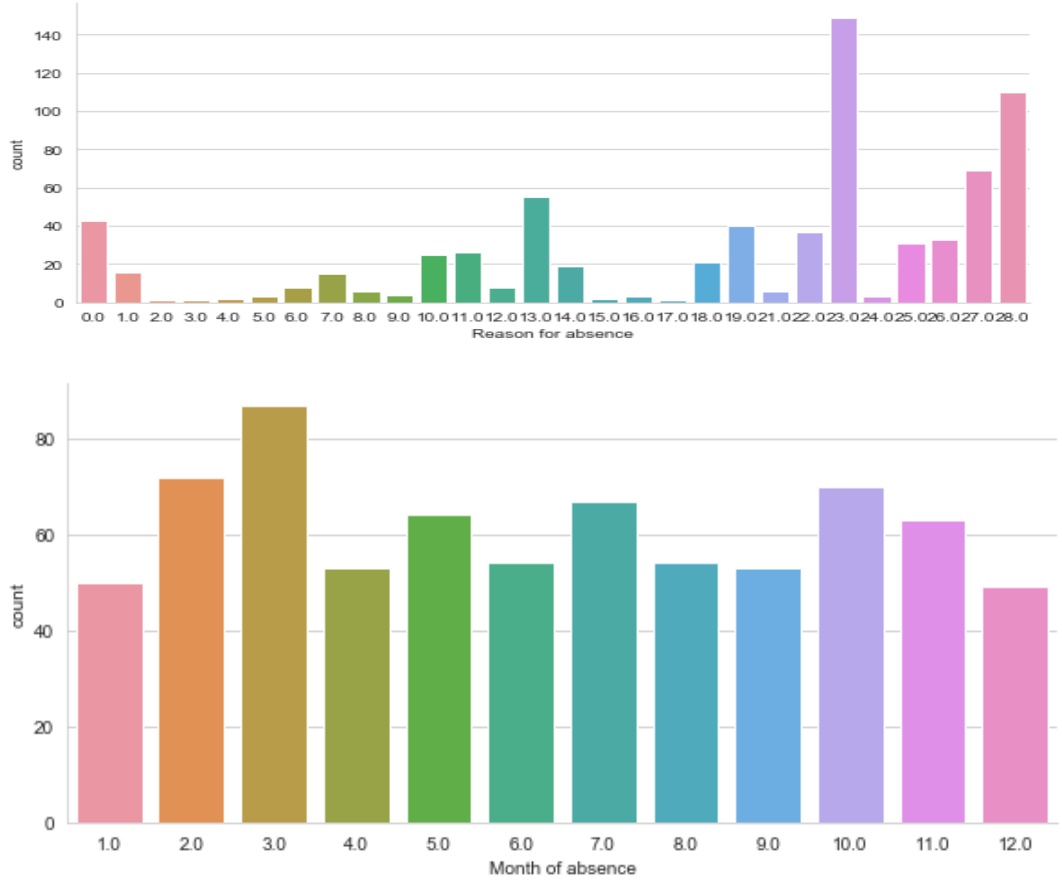


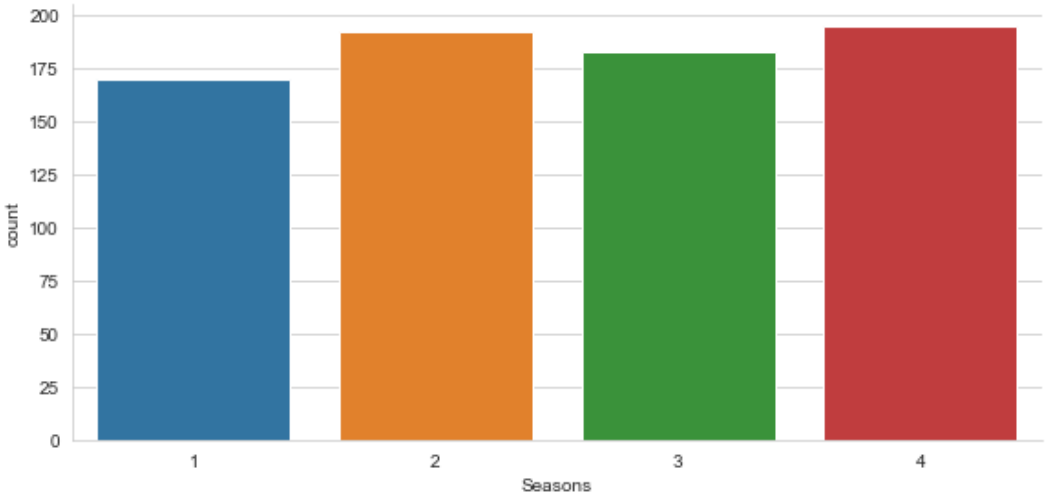


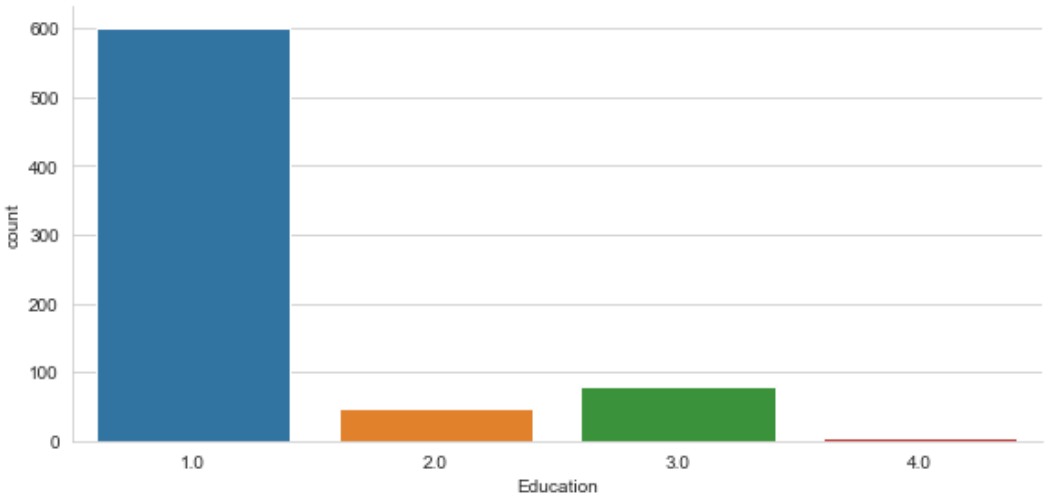
By observing the above plots, it is found that no variable has a normal distribution.

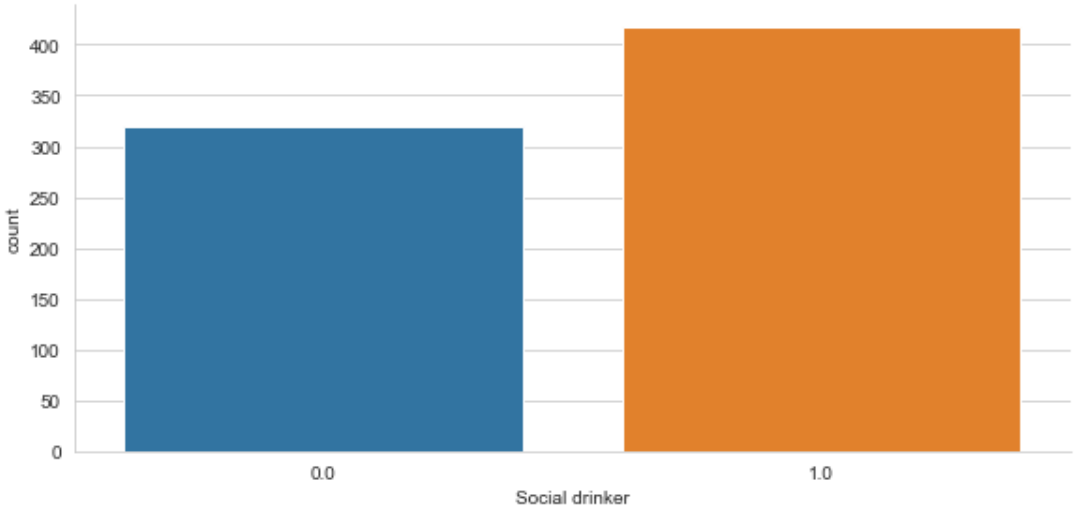
2.4.2 Distribution of Categorical Variables

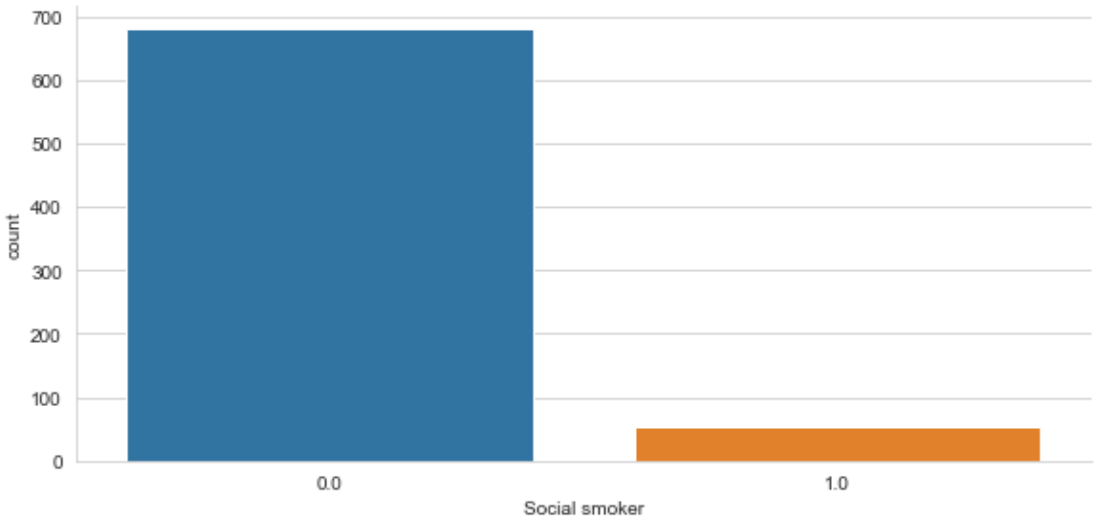
To check the distribution for categorical variables, I have used bar plots. The distribution plots are as below,

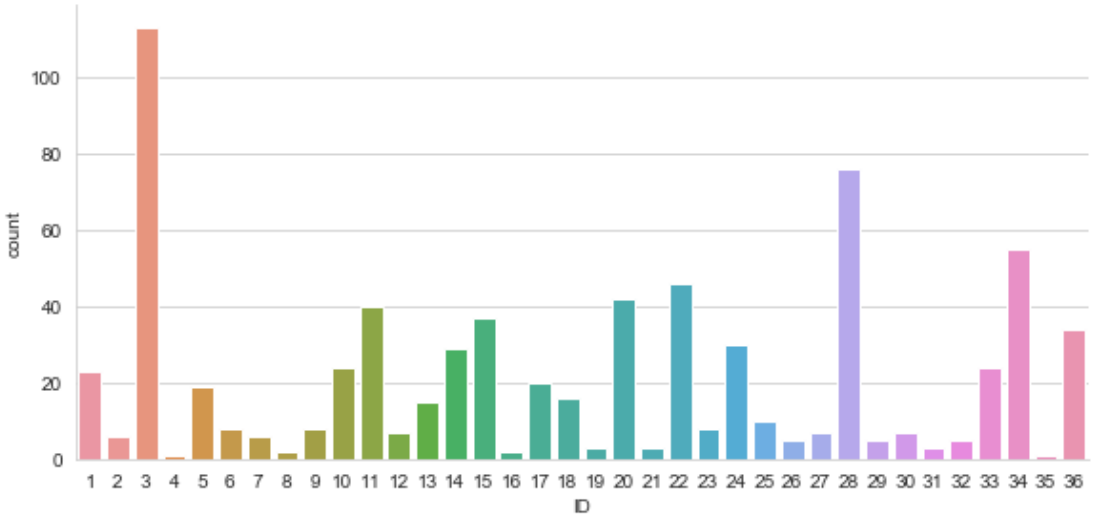


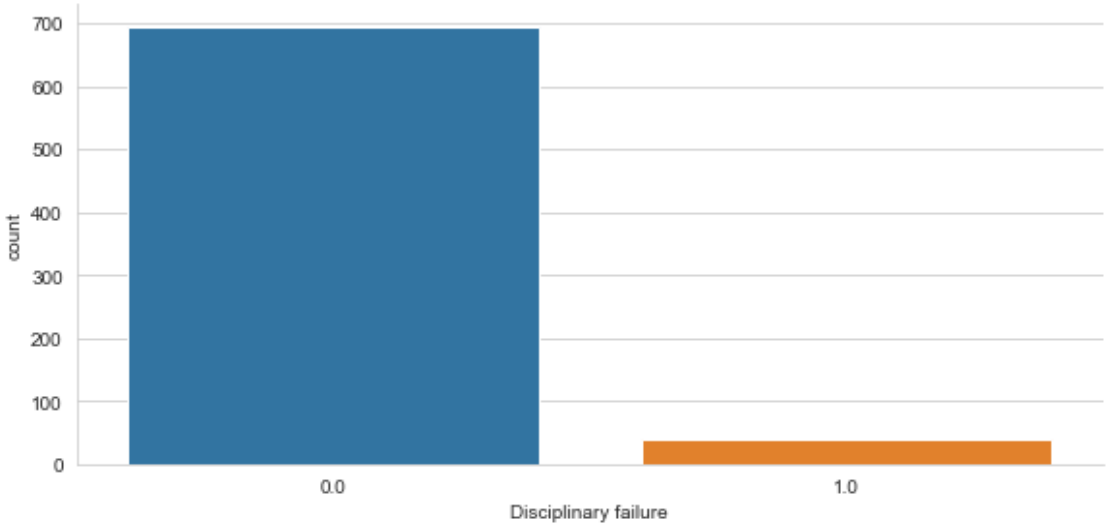

200
175
150
125
100
75
50
25
0
count
1
2
3
4
Seasons

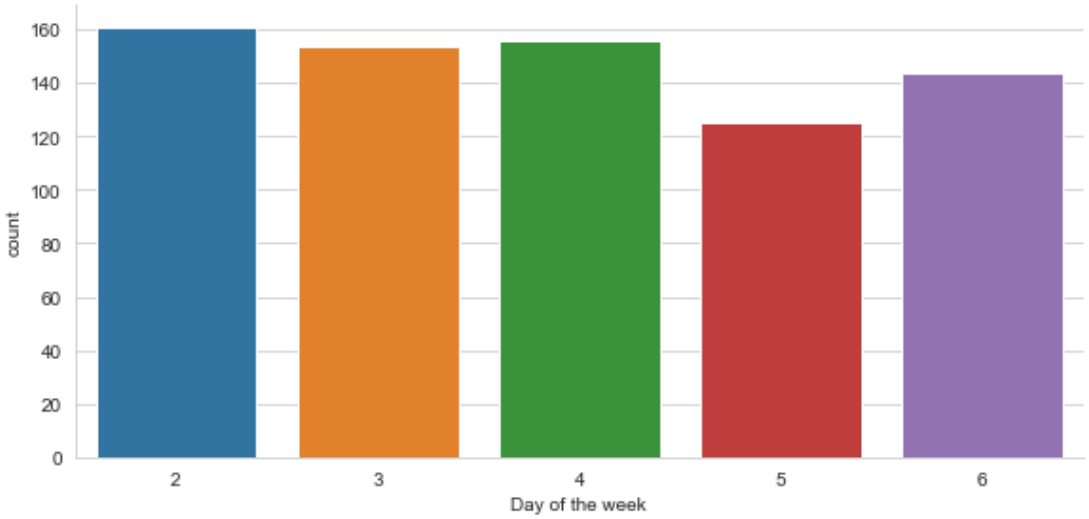

600
500
400
300
200
100
0
count
1.0
2.0
3.0
4.0
Education

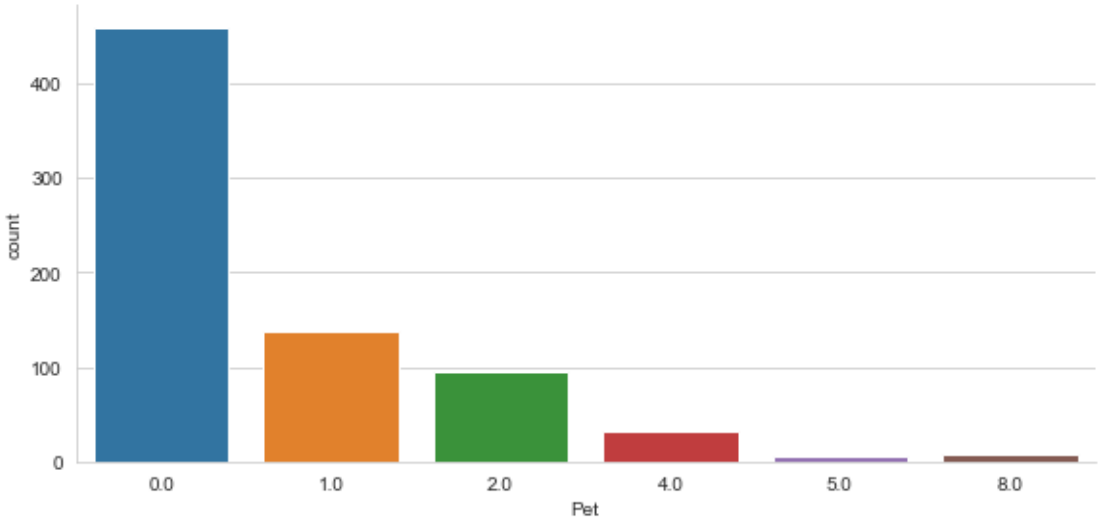

400
350
300
250
200
150
100
50
0
count
0.0
1.0
Social drinker

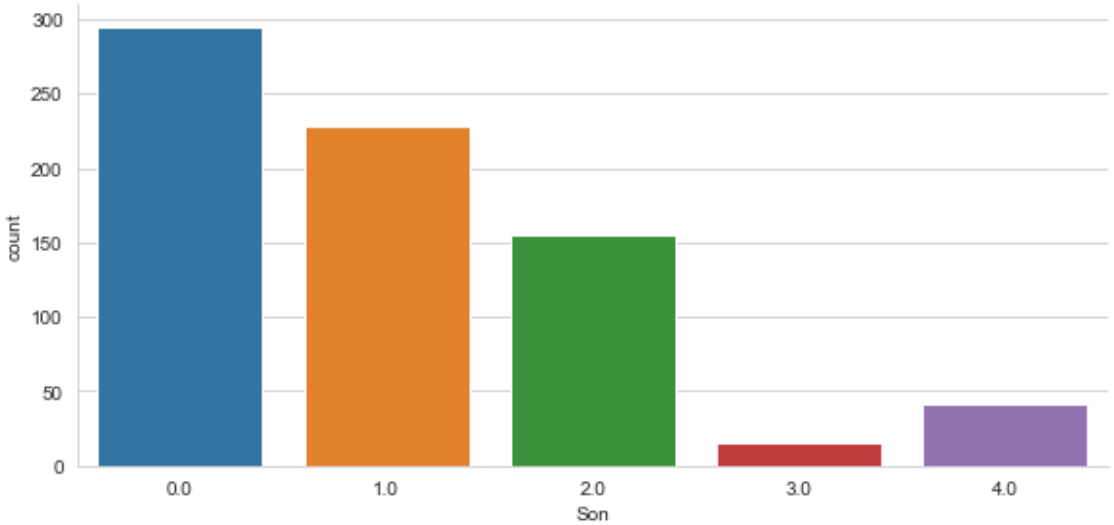

700
600
500
400
300
200
100
0
count
0.0
1.0
Social smoker


100
80
60
40
20
0
count
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
ID


700
600
500
400
300
200
100
0
count
0.0
1.0
Disciplinary failure


160
140
120
100
80
60
40
20
0
count
2
3
4
5
6
Day of the week

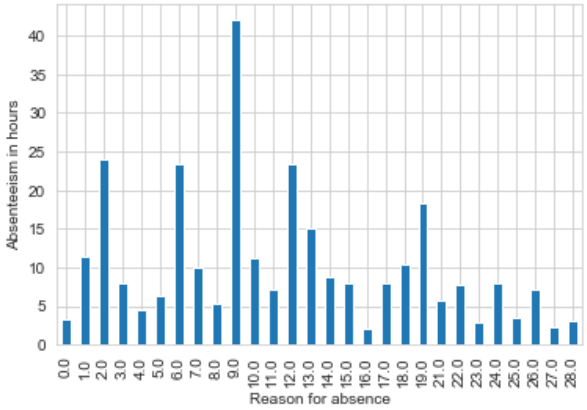

400
300
200
100
0
count
0.0
1.0
2.0
4.0
5.0
8.0
Pet

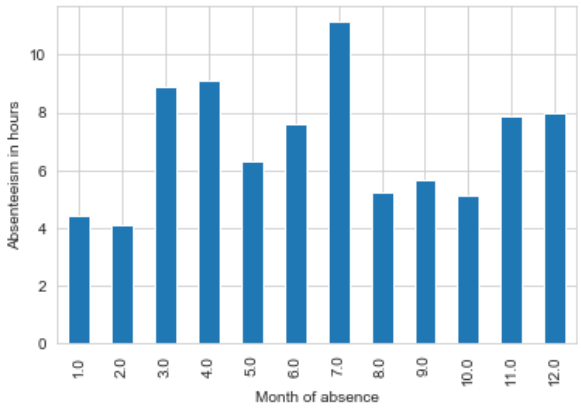

300
250
200
150
100
50
0
count
0.0
1.0
2.0
3.0
4.0
Son

Below are a few observations from the distribution of categorical variables.

1. Employees with ***zero*** sons has the highest absenteeism and then followed by ***one*** and ***two*** sons respectively. Employee with ***three*** sons has the least.
2. Employees having ***zero*** pets have more absenteeism than employees with one or two pets.
3. The absenteeism is found to be high on Mondays followed by Wednesday and least on Thursdays.
4. Employees with disciplinary failure and Social smoker category as ***NO*** have the highest absenteeism.
5. Employee with ***ID 3*** have the highest absenteeism, then followed by ID 3, ***ID 28*** have the highest.
6. Employees who are Social drinkers have the highest absenteeism.
7. Employees with Education as ***High School*** have the highest absenteeism and employees with ***Masters/Doctorate*** have the least absenteeism.
8. ***Spring*** season have the highest absenteeism and ***Summer*** has the least.
9. Highest absenteeism is found in March, following October and February. January and December have the least.
10. Reasons for absenteeism is highest for ***Medical consultation (23) and Dental consultation (28)*** respectively.

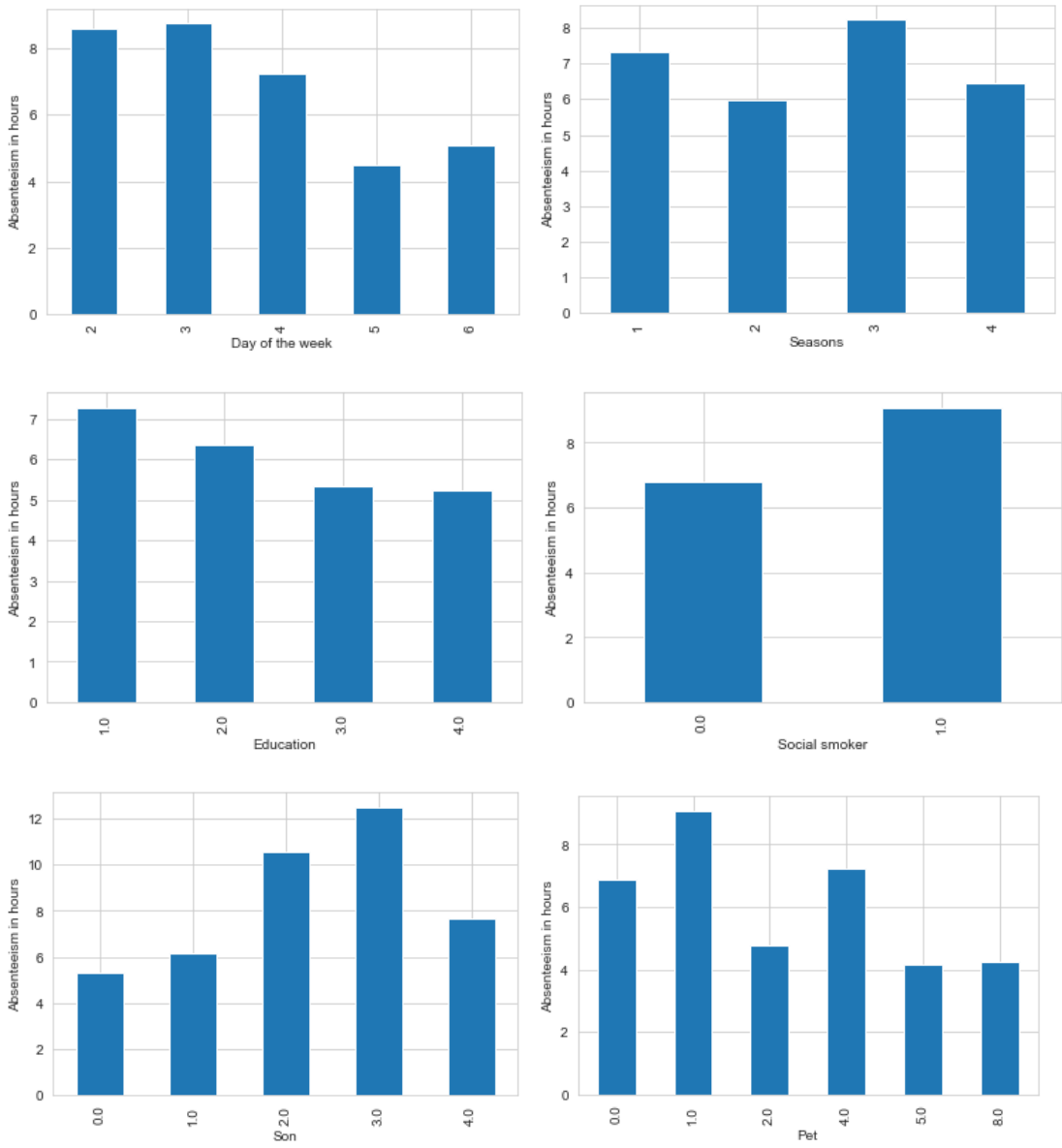
Let's also check the distribution of these categorical variables against our target variable. I have grouped the data with most of the categorical variables and find the *mean* of ***Absenteeism time in hours*** for each category. The plots are as below.





The observations are as below from the above plots.

1. Reason for absence - ***Diseases of the circulatory system*** is having the highest average absenteeism time.
2. Month of absence – ***July*** is having the highest average absenteeism and ***February*** is having the least average absenteeism.



3. On an average, the highest absenteeism is on ***Tuesdays*** and ***Mondays*** respectively and the least average is on ***Thursday***.
4. Average highest absenteeism is in ***Winter*** season and least average in ***Autumn***.
5. On an average, employees with ***three*** sons have the highest average of absenteeism.
6. Employees with ***one*** pet have the highest average absenteeism.

2.5 Feature Engineering

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. While programming in R, I have created two variables by using the existing variables. They are ***age_bin*** and ***distance_bin***.

By using the ***Age*** variable, I have labelled three categories of age into ***Young***, ***Middle_aged*** and ***Old_age***

By using the ***Distance from Residence to Work***, I have labelled two categories into ***Nearest*** and ***Farthest***.

2.5.1 Distribution of the new variables

The observations from the above plots are,

1. Most of our employees are middle-aged i.e., in between 35 years to 45 years.
2. Most of our employees are residing near to the company i.e., in between 0 to 30 kilometers.

2.6 Feature Selection

Feature selection reduces the complexity of our model. The lesser the variables, the higher the performance. It also reduces the over fitting of the model. Multicollinear variables are to be removed if correlation value, $r > 0.7$.

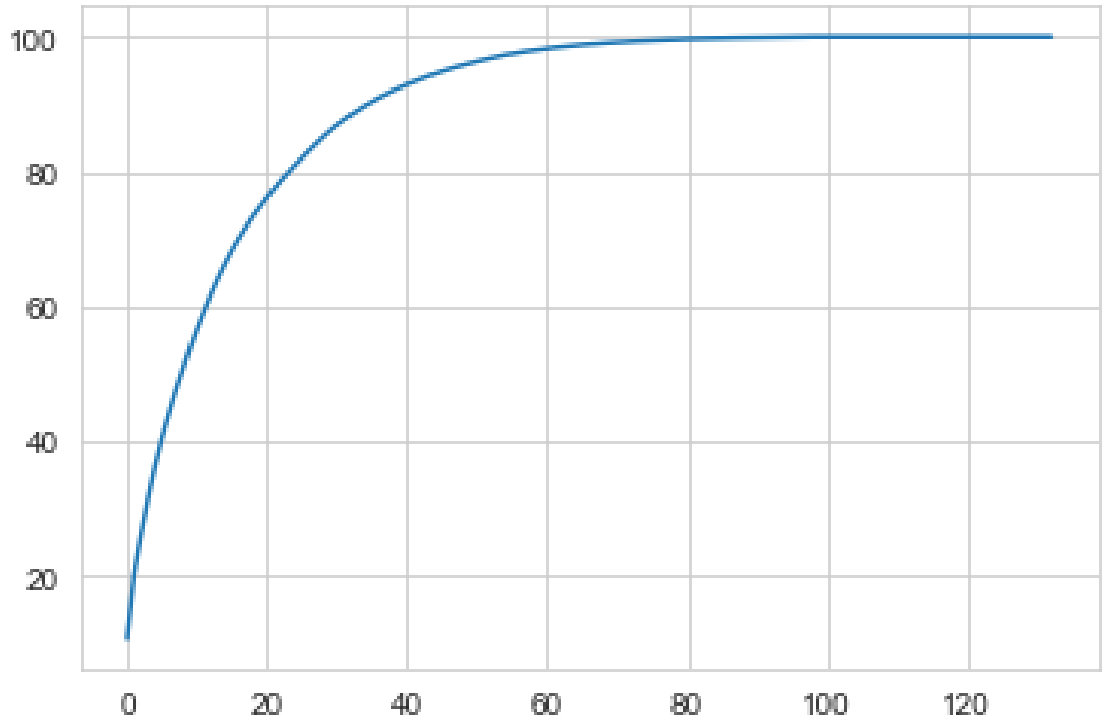
This can be found by using correlation plot for continuous variables and chi-square test for categorical variables. I have used one-way ANOVA for categorical variables while programming in python. The below is the correlation plot, which explains that ***Weight*** and ***Body Mass Index*** are highly correlated with each other with $r = 0.9$. Hence anyone variable can be removed.

Random forest will also help us to find the important variables in a dataset. We have used this method in R programming.

2.7 Principal Component Analysis

Principal Component Analysis (PCA) tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. It is used to explain the variance-covariance structure of a set of variables through linear combinations. It is often used as a dimensionality-reduction technique. We've used this technique while programming in python to find the most important variables (as components) which carries the maximum information to drive our target variable.

After creating dummy variables for our categorical variables, the data have 133 components (independent variables) and 740 observations. After applying PCA algorithm and by plotting a cumulative scree plot, we have found that approximately 50 variables explain almost 95% + of our data. Hence, we have selected those 50 variables and created a new model again.



2.8 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. It is also known as data normalization and is generally performed during the data pre-processing step. Since the raw data we receive from the client might have **n** number of variables with wide ranging values, our model will give high importance to the variable with high values. So all the independent continuous variables are to be scaled by using Standardization or Normalization techniques. If our data is normally distributed, we will use Standardization technique. As our data is not normally distributed, we have used Normalization technique to scale our variables. In some machine learning algorithms, objective functions may not work properly without normalization thereby leading to inaccurate results. Hence scaling of the variables is a mandatory.

Chapter III

Modelling

3.1 Modelling and Model Selection

After applying all the pre-processing techniques on our data, we should develop a regression model to predict our target variable. Our target variable is a continuous variable, hence the models that we've chose are Linear regression, Decision tree, Random forest and Gradient Boosting methods.

The whole data is divided into test and train data. We will train our model using train data and then apply and validate that model on our test data. Also, we have various metrics to validate our model like RMSE, MAPE, R-Squared, Adj R-Squared and etc.

3.2 Linear Regression

Linear regression is one of the statistical models that is used for prediction (regression only). It was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm.

Linear regression is a linear model, i.e., a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

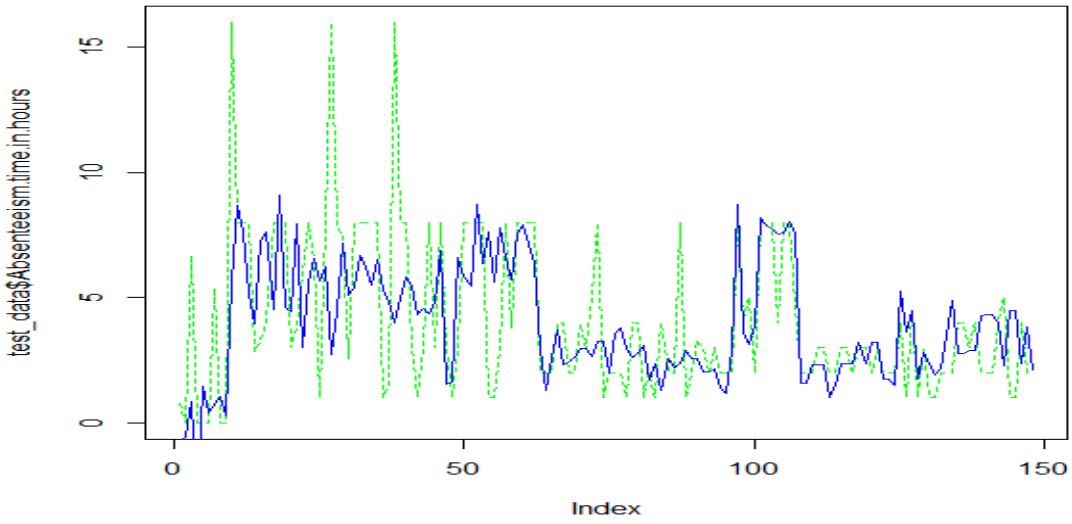
When there is a single input variable (x), the method is referred to as *simple linear regression*. When there are multiple input variables, it refers to the method as *multiple linear regression*.

Parameter	R	Python
RMSE	2.714622	1.7002E+12

Below is the line plot between actual and the predicted values.

Green line is the Actual values

Blue line is the Predicted values



3.3 Decision Tree

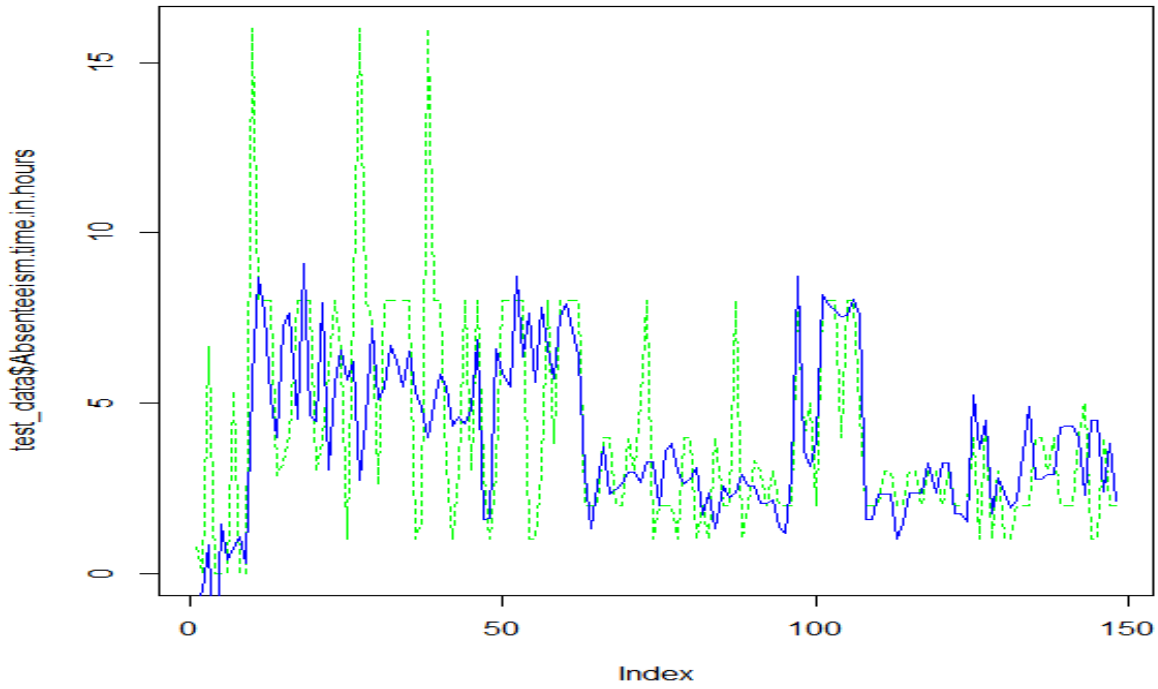
Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Each node in a decision tree represents a feature, each link represents a decision, or a rule and each leaf represent an outcome – *categorical value or continuous value*.

Parameter	R	Python
RMSE	2.679105	3.1792364568869935

Below is the line plot between actual and the predicted values.

Green line is the Actual values

Blue line is the Predicted values



3.4 Random Forest

Random Forest is one of the most popular and powerful supervised machine learning algorithms. It is a type of ensemble machine learning algorithm called Bootstrap Aggregation or bagging. This ensemble technique consists of many decision trees to improve the accuracy and reduce the weak learners to produce a strong learner from the model.

Let's assume we have 50000 observations in a dataset with 1000 features. We can ensure that one DT will give us the best results or rules or decisions with the minimal errors. In random forest, we do concatenate all the decision trees that we form on a dataset with different observations with different features to obtain the best results.

This method is the combination of *Bagging* idea and random selection of features. Bagging is feeding an error of one DT as an input to the next DT to improve the accuracy of the whole model. Likewise, we can improve the accuracy of the model.

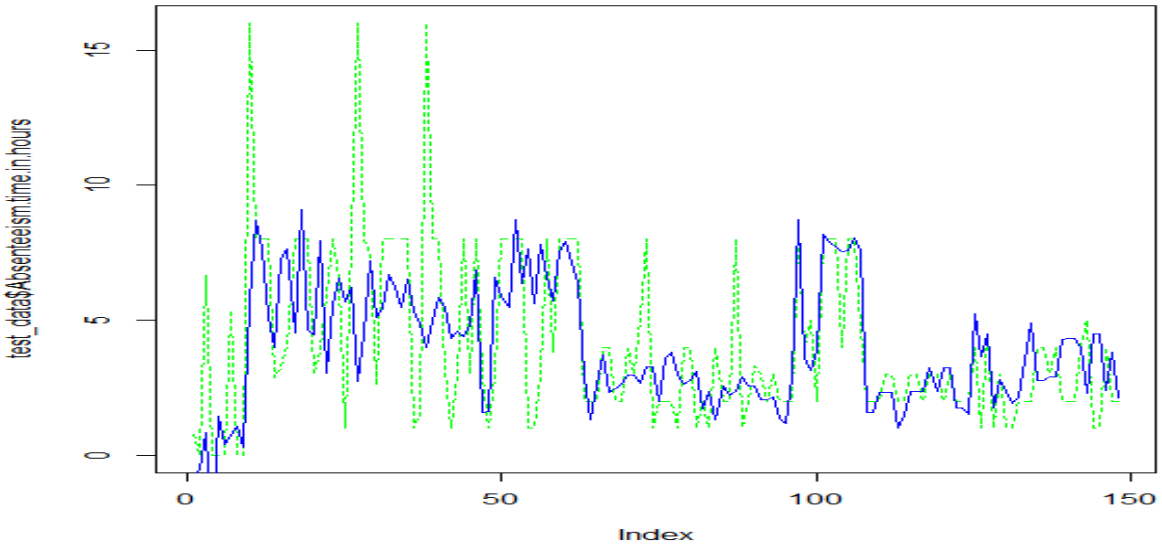
I have used 200 trees and 300 trees in R and Python respectively to predict our target variable.

Parameter	R	Python
RMSE	2.679105	2.6513491672836627

Below is the line plot between actual and the predicted values.

Green line is the Actual values

Blue line is the Predicted values



Chapter IV

Conclusion

4.1 Evaluation of the Model

Model evaluation metrics are used to assess goodness of fit between model and data, to compare different models, in the context of model selection and to predict how predictions (associated with a specific model and data set) are expected to be accurate. I have considered the RMSE (Root Mean Square Error). The lower RMSE shows the best fit.

Root Mean Square Error (RMSE) is the standard deviation of the residuals which are the prediction errors. *Residuals* are a measure of distance of the data points from the regression line. It is also a measure of how spread out these residuals are. RMSE is calculated by using the below formula,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Predicted – The value our model predicted

Actual – Actual value

4.2 Selection of the Model

From all the RMSE values we obtained from different models, ***Random forest*** gave us the least RMSE value. Hence, we can conclude that random forest is the best model for this dataset.

4.3 Solutions for the Problem Statement

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

As the absenteeism is high on Mondays, provide some additional incentives for the employees who work on Mondays or conduct a few recreational activities on Monday. There might be a chance to reduce the number of absenteeism.

Hire with people with Education qualification with at least Post Graduation and preferably hire employees with master's degree or Doctorate.

Individual employees with highest absenteeism are to be warned or reduce the incentives to them. Restrict some value for the absenteeism count to every employee, if that value is crossed, deduct their bonus.

Employees who are animal lovers can be considered for recruiting into the company.

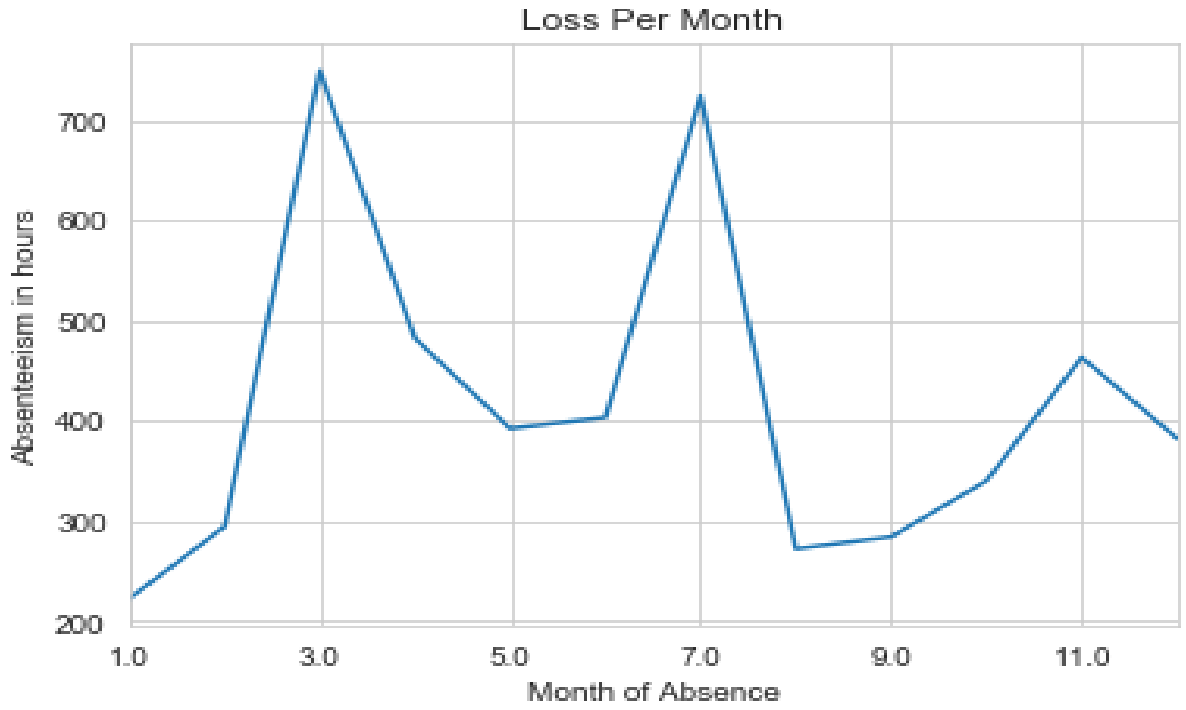
Reasons for absenteeism is highest for ***Diseases of the Circulatory system***, ***Medical consultation and Dental consultation***. Conduct some medical camps on quarterly basis and ensure if employees are having a good health. Conduct a few campaigns for the employees to take the preventive measures for their health issues.

Appoint a counselor once in a month and motivate employees with smoking and drinking habits to quit them.

March, October and July have the highest absenteeism. We can expect the same in the coming years. So, ensure to have some outsourcing labor to reduce the losses.

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

Let us consider the below graph.



As per the dataset given, we have the highest absenteeism in the month of March and July with 750 and 720 hours of absenteeism. Assume that if absenteeism hours are considered as losses and assume that our model's accuracy is 85%, then in 2011 which is our predicted values will have 600 – 630 losses in March and 590 – 610 hours in July.

/**********THE END**********/