PROJECT REPORT

Employee Absenteeism

(BASED ON R AND PYTHON)

Presented by

SINGH SAURABH VINOD

(saurabh123.adbudds@gmail.com)

18th June 2019

CONTENTS

Page No.

LIST OF TABLES

LIST OF FIGURES

Page no

ABSTRACT

Absenteeism means pattern of absence from a duty without any good reason. Absenteeism is unplanned absences from ones duty and is indicator poor individual performance which in turn leads to poor performance of the company. In this project XYZ a courier company facing a severe issue of absenteeism has shared its dataset which contains the information about the employees of that company of previous year. On the basis of the previous year dataset and observed pattern and insights we have to predict the future values and help them to know the factors which help them to reduce the number of absenteeism as well as the monthly loss company is going to face in the coming year i.e. 2011.

We are using both R and Python to build the suitable model according to the company's problem statement. We will try different ML Algorithm and will choose the best model accordingly to help them to know the answers for the questions mentioned above.

CHAPTER – 1

INTRODUCTION

Absenteeism means unplanned absence from work without any good reason, company faces major losses due to this problem.

We will try to figure out what reason is playing the main cause and predict the measures the company must carry out to avoid further losses.

1. PROBLEM STATEMENT

Absenteeism leads to severe business loss. XYZ is a courier company and is facing a genuine issue of absenteeism. The aim of this project is to predict the factors to reduce the number of absenteeism and the work loss company is going to face next year if same trend of absenteeism continues. To make the predictions we used R and Python codes and algorithms.

2. DATA

According to the problem statement the given data comes under **"REGRESSION" (forecasting)** category as our target variable **"Absenteeism time in hours"** is continuous in nature. Depending on the problem statement we developed and selected algorithms for the prediction.

There are 21 predictors and 1 target variable in our dataset listed in table 1.1. These predictors are very useful in predicting the results.

Table 1.1: Variables name

No.	Variable name
1	Individual Identification(ID)
2	Reason for absence
3	Month of absence
4	Day of the week
5	Seasons
6	Transportation expense

7	Distance from residence to work
8	Service time
9	Age
10	Work load average/day
11	Hit target
12	Disciplinary failure
13	Son
14	Pet
15	Social smoker
16	Social drinker
17	Height
18	Weight
19	Body mass index
20	Education
21	Absenteeism time in hours

There are several categories in the variable “Reason for absence” according to International Code of Diseases (ICD) which is listed below in table 1.2.

Table 1.2: Levels of Reason for absence

No	Reason for absence
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 respiratory system.
XII	Diseases of the skin and subcutaneous tissue.
XIII	Diseases of the musculoskeletal system and connective tissue.
IV	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.
XI	Factors influencing health status and contact with health services.

There are 7 more categories without COD and different levels in few variables like day of the week, month of absence, seasons, education, social smoker, Disciplinary failure and social drinker which are mentioned in appendix B.

Table 1.3 represents the sample of the dataset on which we applied the preprocessing techniques and algorithms for further prediction of the target class.

Table 1.3: Sample of Employee Absenteeism Dataset (Column 1 to 7)

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from work to home
11	26	7	3	1	289	36
36	0	7	3	1	118	13
3	23	7	4	1	179	31
7	7	7	5	1	279	5
11	23	7	5	1	289	36
3	23	7	6	1	179	51

Table 1.4: Sample of Employee Absenteeism Dataset (Column 8 to14)

Service time	Age	Workload	Hit target	Disciplinary Failure	Education	Son
13	33	239554	97	0	1	2
18	50	239554	97	1	1	1
18	38	239554	97	1	1	0
14	39	239554	97	1	1	2
18	33	239554	97	1	1	2
3	38	239554	97	1	1	0

Table 1.5: Sample of Employee Absenteeism Dataset (Column 15 to 21)

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

1	0	1	90	172	30	2
1	0	4	89	170	31	NA

CHAPTER- 2

METHODOLOGY

We know that data is backbone of data science is Data. We collect data from different sources and converting data I proper format is very necessary. When any new project comes in we spend 80% time in understanding, cleaning and preparing the data as driving the data according to problem is very important. The whole data process is divided into six phases.

a) Business understanding: When any client comes in we should try to understand their problem statement first. It helps us to get proper data for better results.

b) Data understanding: In this we use many statistical techniques, Graphs and visualizations to understand the data so that we can understand the data well and can get relevant data from the client.

c) Data Preparation: This means exploring the raw data we receive from client and understanding what data speaks out. In data science 80% of our time goes in data understanding, cleaning and preparation and 20% in model development and model evaluation. If the quality of data is good the model will predict better and results in high accuracy.

d) Data modeling: There are many machine learning algorithms and we have to select the most appropriate algorithm according to our problem statement.

Evaluation: It helps us to evaluate our model. It tells us whether our model is able to accomplish the business objective or not.

e) Deployment: This is the final phase in which we deploy our model in client premises.

2.1 Data Exploration

In data exploration we try to understand the data. We should observe the data and understand them. Looking at data and understanding them with the help of different tools and graph and visualizations is called Exploratory Data Analysis. It is one of the very important steps as driving the data according to problem statement is very necessary and to drive the data we need to understand our data first.

In our project we check the data type dimensions, shape of the data, and count of unique values in each variable. In our dataset we have 740 observations and 21 variables. All the variables are in numeric format. Later in preprocessing technique we changed some of the variables such as reason for absence, seasons, and days of the week, months of absence, disciplinary failure, education, social smoker and social drinker into required format accordingly.

2.2 Pre Processing techniques

Data preprocessing is a data mining technique that involves transforming raw data into proper format. It prepares raw data for further processing. Data preprocessing includes cleaning, normalization, transformation, feature extraction and selection etc. The product of data preprocessing is final training set. Data we receive from client is messy data. If there is much irrelevant and redundant information present in our dataset it will make our model inconsistent which results in poor and low results.

2.2.1 Missing value analysis

Missing value is the values which are not present or missing from the dataset. Missing values appears in our dataset due to various reasons like human error, refuse to answer the questions in a survey or optional box questionnaire. The skipped or unanswered questions appear in form of missing values. Missing values can be treated either by dropping the variable or by imputing the missing values.

With the help of domain knowledge we need to understand why there is missing value in a dataset and then it is suitable to know whether to ignore or impute the missing values.

When to ignore the missing values: First we will create a data frame which tells us amount of missing values present in each variable. Drop the variables which consists more than 30% (according to industry standards) of missing values.

When to impute missing values and methods of imputation: We will impute those variables whose missing percentage is less than 30%. There are three methods to impute missing values: **a)** Fill with central statistics method i.e. mean and median for continuous variable and mode (majority minority rule) for categorical variable. **b)** Distance based or Data mining method which includes KNN imputation. **c)** The last method is prediction method which is based on ML algorithms.

In our dataset we have 135 missing values. All the variables have less than 10% of missing values so we have imputed them using mean and median accordingly. The graphs of missing values percentage are presented below.

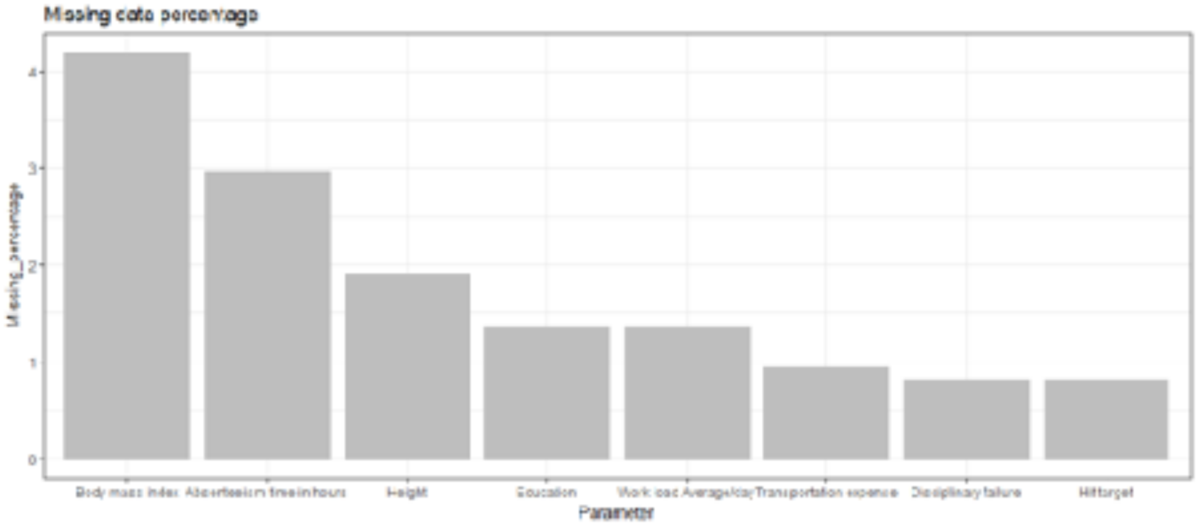


Fig 2.1: Missing percentage graph

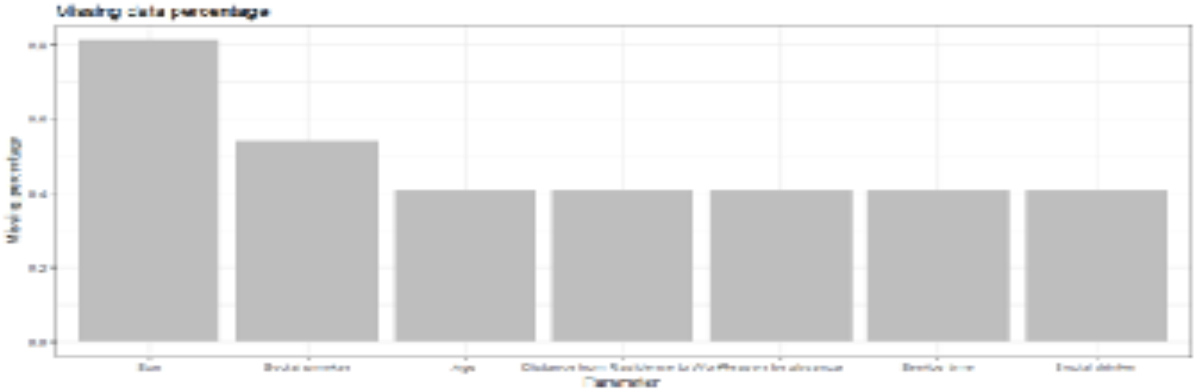


Fig 2.2: Missing value percentage

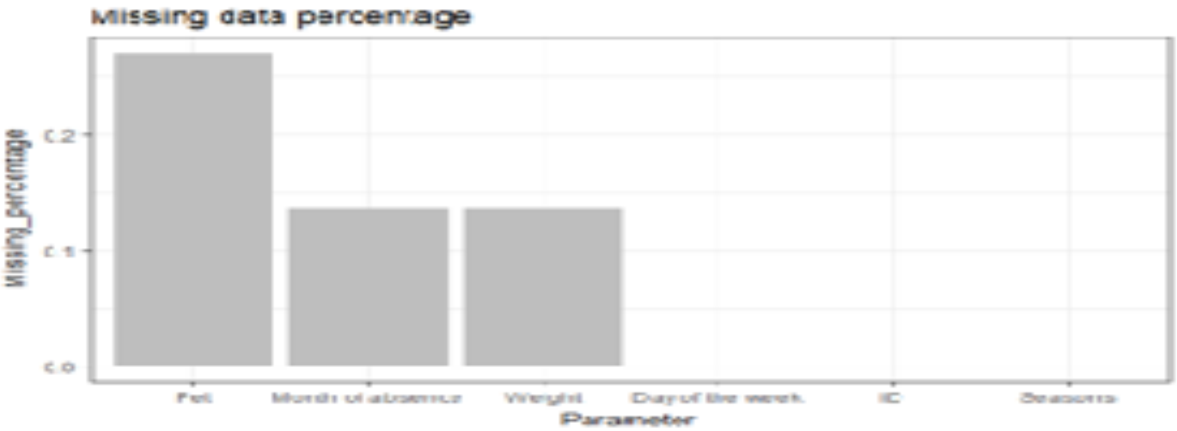


Fig 2.3: Missing value percentage

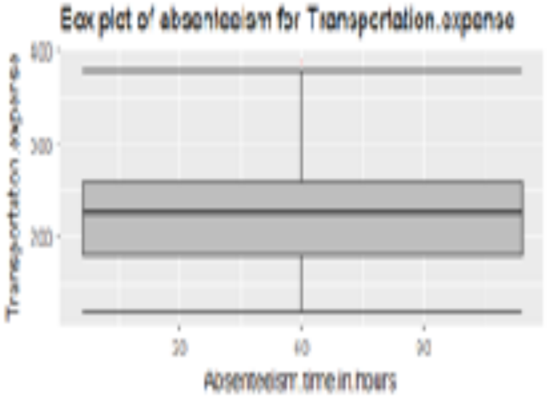
The graph shows all the missing values lies between 4.18% and 0.135%. The highest missing value is in Body mass index where is lowest is in month of absence and weight. Our target variable i.e. Absenteeism time in hours consists of 2.97% of missing values whereas three variables (Day of week, ID, Seasons) does not contain any missing values.

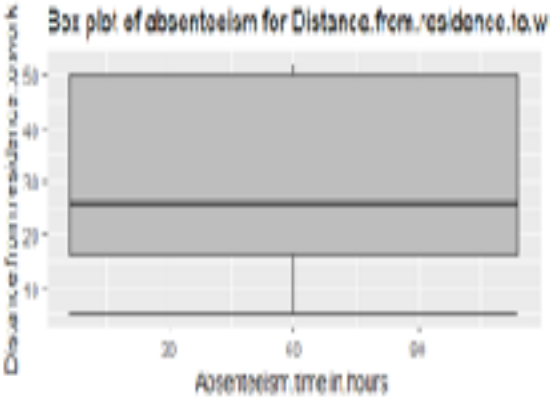
2.2.2 Outlier Analysis

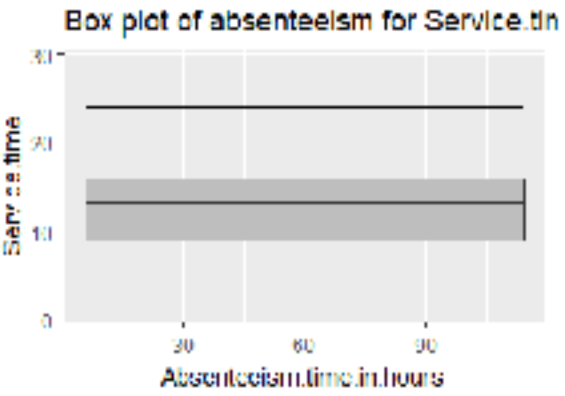
Outlier analysis is one of the preprocessing techniques used to check for abnormal values in the data set clean them and transform the data into a proper shape. There are many different techniques like Graphical Tools (Box Plot Method), Statistical technique (Grubbs test), R Package outlier for outlier analysis and replace with NA which will be treated as missing value analysis and will be imputed using suitable method mentioned in missing value analysis. Presence of outlier in our data leads to poor data quality and contamination, low quality measurement and manual errors. The best way to look at outlier is to understand business process i.e. how data is generated and how is the business flow.

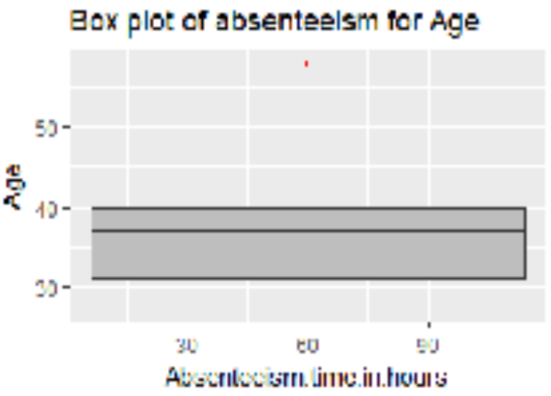
We know that outlier analysis is applicable only on numerical variable so we have converted the entire variable in their appropriate data types and separated out numerical variables for outlier analysis. We have plotted box plot for each numerical variable which is shown below.

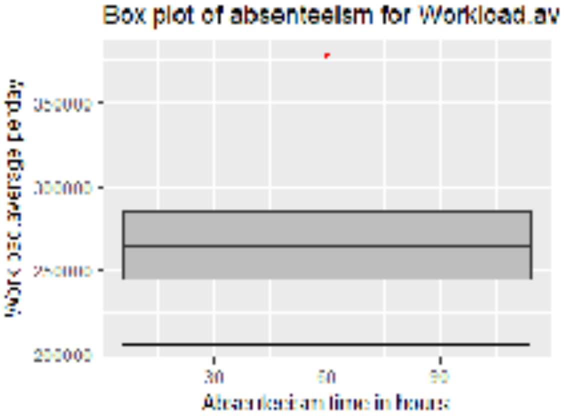
From the boxplot almost all the variables **except "Distance from residence to work", "Son", "Weight" and "Body mass index"** consists of outliers. We have converted the outliers (data beyond minimum and maximum values) as NA i.e. missing values and fill them with mean and median accordingly.

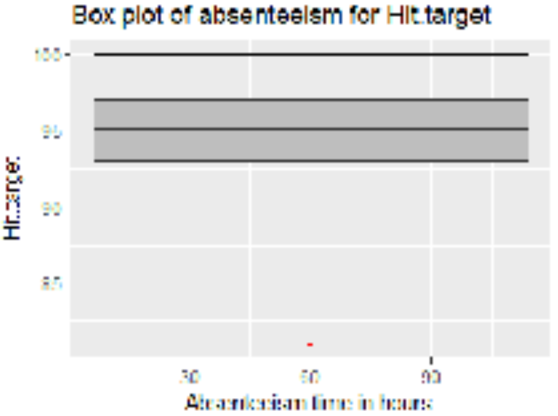


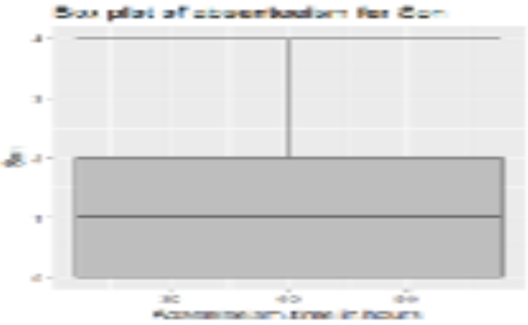


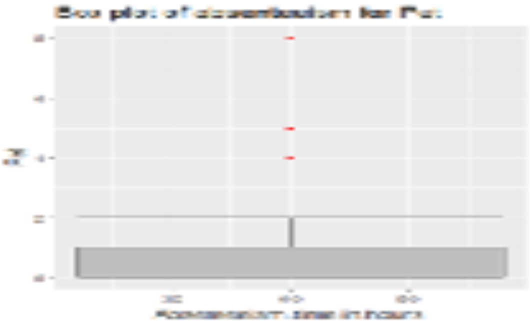


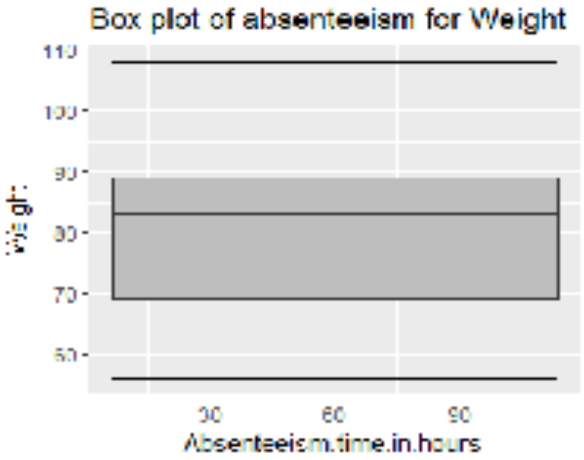


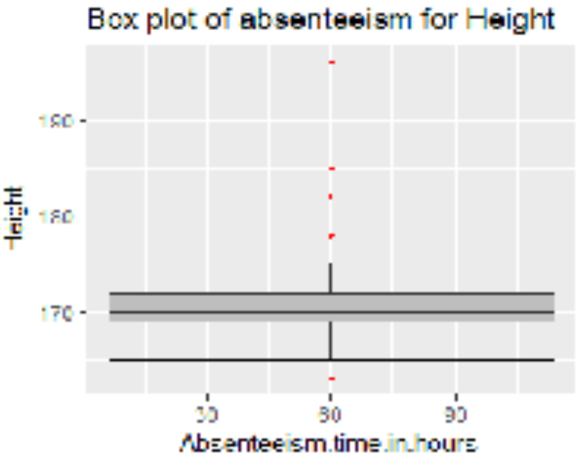

Box plot of absenteeism for Workload.av
Work.load.average.per.day
350000
300000
250000
200000
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Hit.target
Hit.target
100
95
90
85
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Son
Son
4
3
2
1
0
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Pet
Pet
8
6
4
2
0
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Weight
Weight
110
100
90
80
70
60
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Height
Height
190
180
170
30
60
90
Absenteeism.time.in.hours

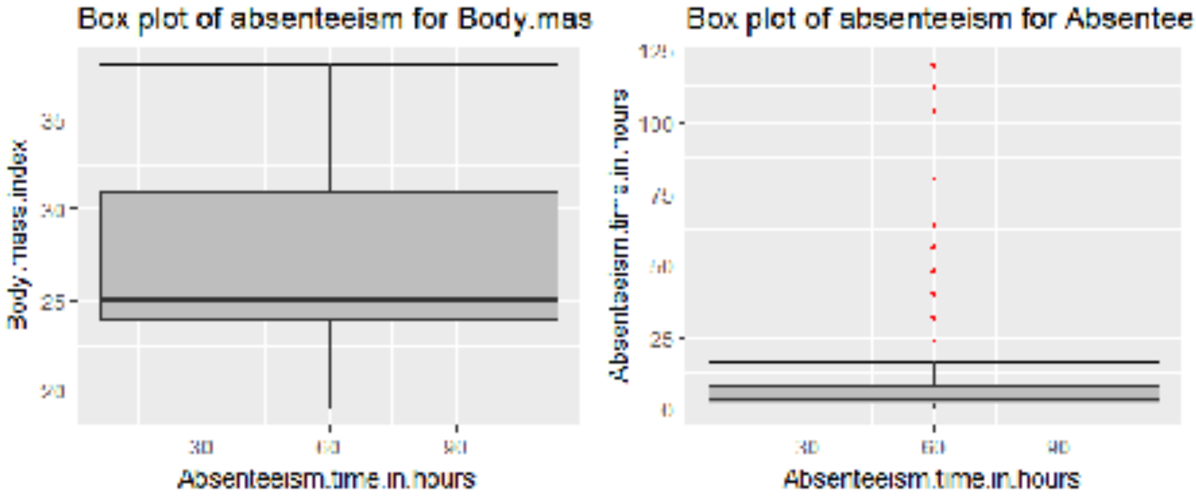


Fig 2.4: Box plot of predictors versus absenteeism of train data (with outliers)

2.2.3 Feature Selection

Selecting subset of relevant features for model construction is known as Feature Selection. When we get raw data we have multiple variables and with the help of variable selection we have to extract relevant data. We cannot use all the features because some features may be carrying the same information or irrelevant information which does not impact the business solution. To reduce complexity we adopt feature selection technique to extract meaningful features out of it. This in turn helps us to avoid the problem of multi colinearity. Correlation Analysis (for numerical variable), Chi Square Test (for both target and predictor as categorical) and other ML algorithm like Random Forest are some of the methods of feature selection.

Correlation analysis: It is one of the methods for feature selection technique applied only on numerical data. Correlation tells us the association between two continuous variables. It ranges for -1 to +1.

-1: Highly negatively correlated

0: No correlation

+1: Highly positively correlated

In correlation there is an assumption that there should be high dependency between predictor and target variable but there should be low dependency between two predictors.

ANOVA (Analysis of Variance): It is applied on one categorical and one continuous variable. It is a statistical technique used to compare means of two or more group. We get results in form of p values. If p value is less than 0.05 we will reject NULL HYPOTHESIS (result purely from chance) and accepts ALTERNATE HYPOTHESIS (influenced by some non-random cause).

In our project we have used Correlation plot and ANOVA to select important variables and selected the variables according to the values of correlation and ANOVA. The correlation plot and ANOVA values are given below.

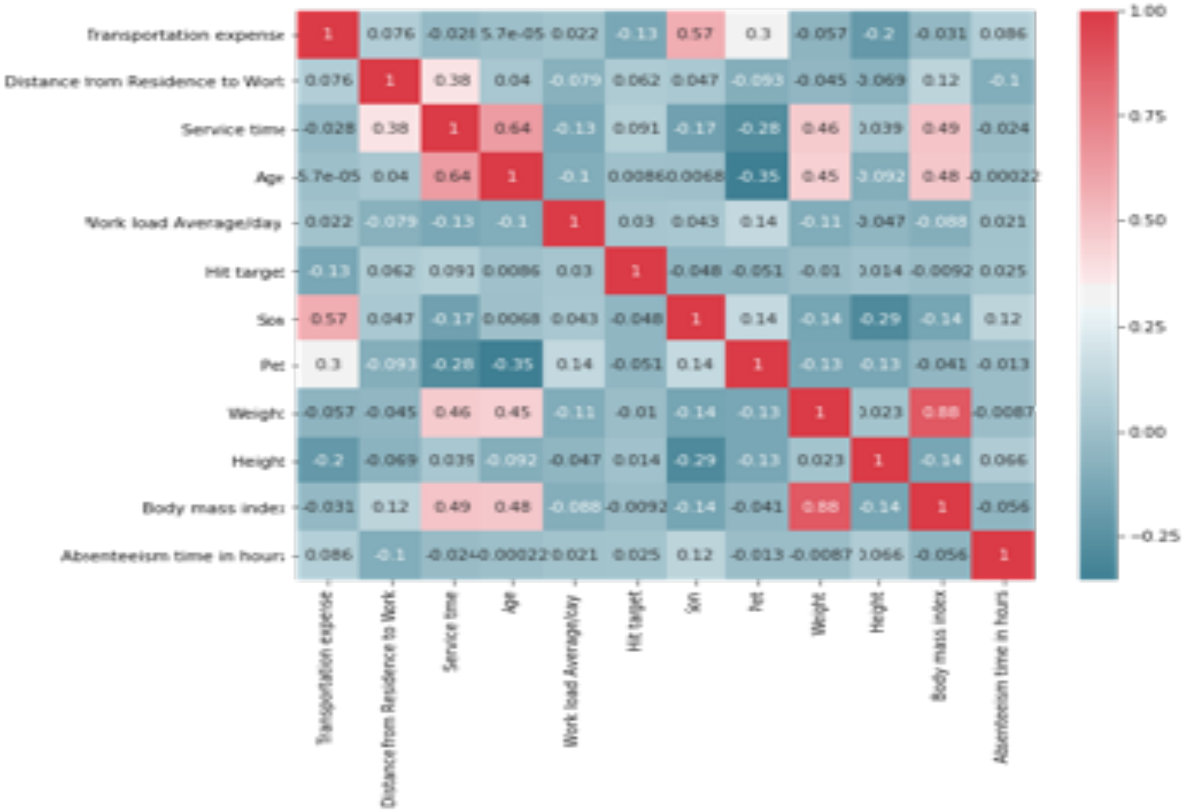


Fig 2.5: Correlation between the variables

Table 2.1: ANOVA for categorical variable

Variable name	p- value
ID	0.000149 ***
Reason for absence	<2e-16***
Month of absence	0.0236 *
Day of the week	0.0159 *
Seasons	0.384

Disciplinary Failure	0.109
Education	0.694
Social smoker	0.28
Social drinker	0.0873

After applying correlation analysis and ANOVA out of 20 predictors we are left with 9 predictor variables and the target variable i.e. Absenteeism time in hours. The variables we are considered after applying correlation analysis and ANOVA are:

1) Reason for absence
2) Transportation expenses
3) Distance from residence to work
4) Day of the week
5) Social drinker
6) Workload average/day
7) Service time
8) Height
9) Absenteeism time in hours

2.2.4 Feature Scaling

It comes into an action when we are dealing with parameters of different units and scales. It is also known as variable scaling. It is used to limit the range of range of variables so that they can be compared on common basis. Feature scaling is performed only on continuous data. There are two methods to scale the data **Normalisation and Standardisation.** Normalisation is the process of reducing unwanted variation either within or between variables. Normalisation brings all the variables into proportion with one another. It ranges between **0 and 1** and are sensitive to outliers. Normalisation works on all kind of continuous data whereas standardisation works well when data is uniformly distributed.

In our project we have used normalisation as our data is not uniformly distributed. The below plotted histograms shows our data are not uniformly distributed.

Before Normalisation after Normalisation

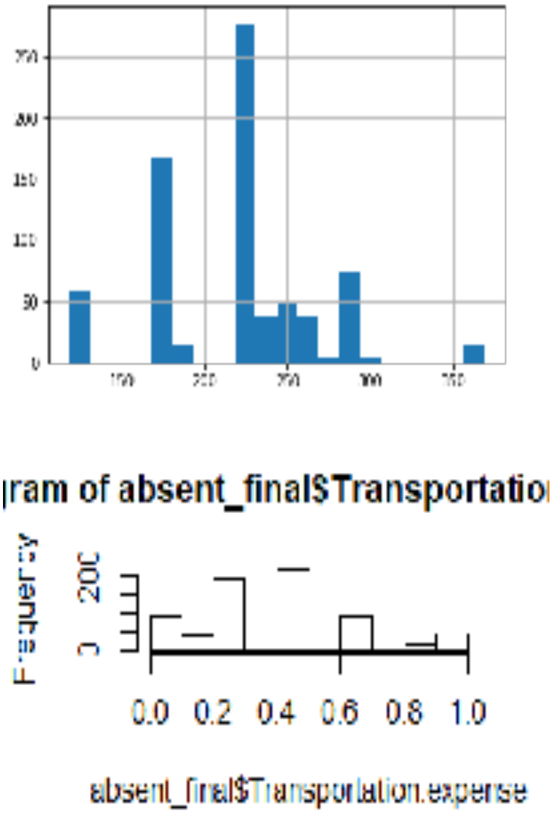


Fig 2.6: Normality distribution of transportation expense

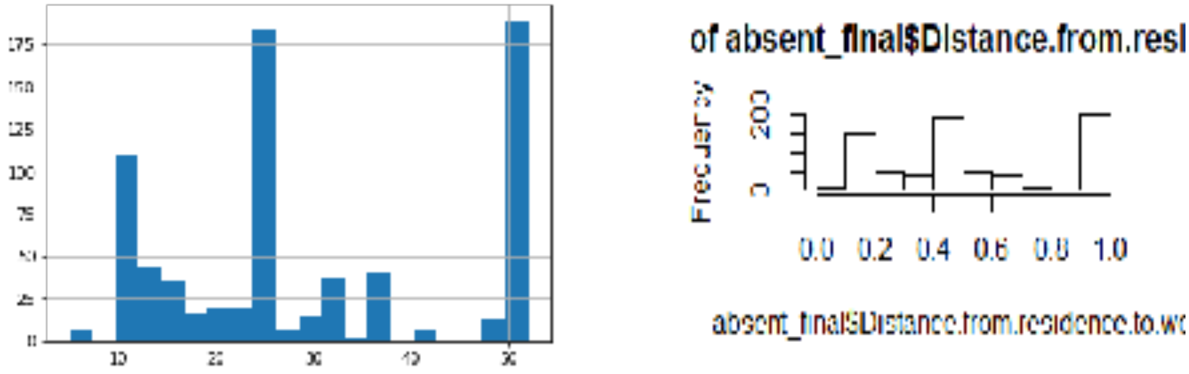


Fig 2.7: Normality dist. of distance from residence to work

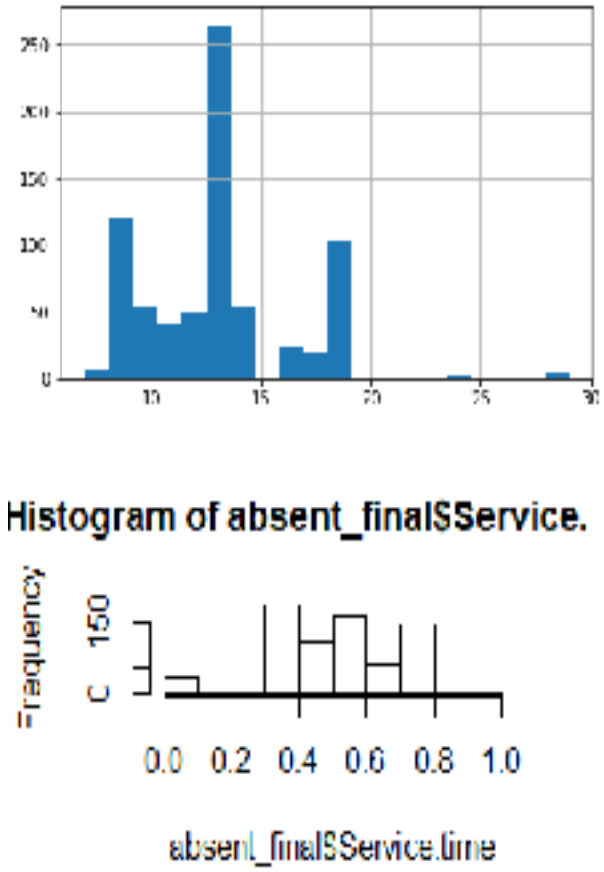


Fig 2.8: Normality dist. of service time

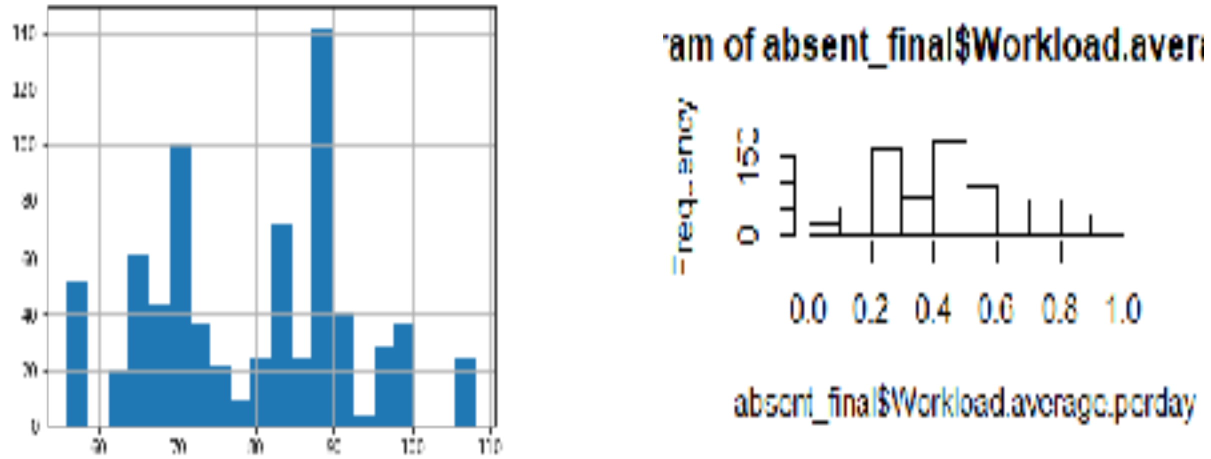


Fig 2.9: Normality dist. of weight

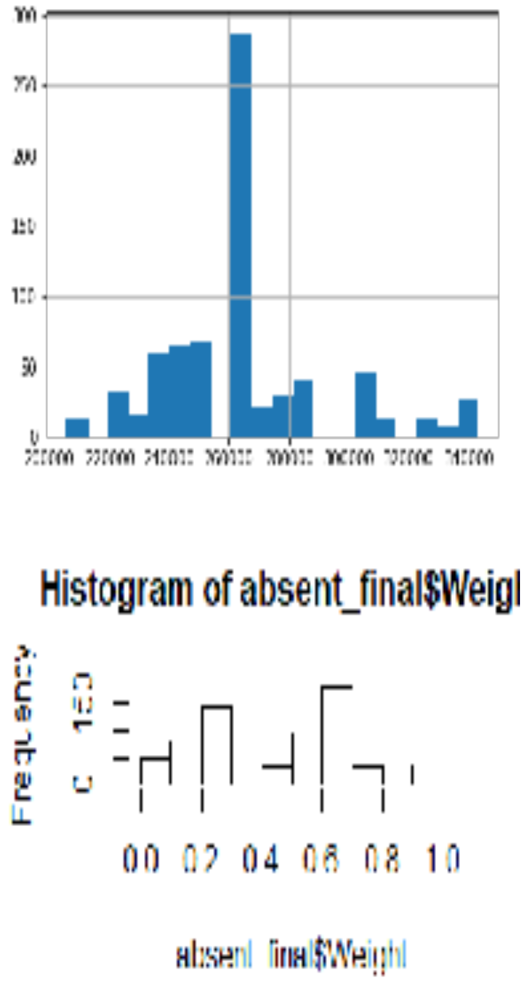


Fig 2.10: Normality dist. of work load avg/day

CHAPTER - 3

MODELING

Data modeling means selecting appropriate ML Algorithm according to our problem statement. There are more than 100 algorithm and we need to select an appropriate one.

3.1 MACHINE LEARNING ALGORITHM

Machine Learning means programming computers to optimize a performance criterion using example data or past experience. With help of historical data we try to extract patterns and save it in ML itself. Once a new test case comes in we use that historical pattern to apply on new data to predict its class level.

3.2 MODEL SELECTION

There are multiple parameter based on which we select an algorithm to be developed for particular dataset.

(a) It depends on problem statement as once we define problem statement then only we will be able to find out which type of problem statement it is.

(b) Every ML Algorithm has three components: **Representation, Evaluation and Optimisation.** Some algorithm deliver output in form of business rules some in form of numbers, probability and visualisations.

It depends on client requirement that in which form they want the output. If they want output in form numbers then we will go for regression algorithms.

In our early stages of analysis during pre-processing we have come to understand the customer behaviour pattern on test data. Here in our case the dependent variable “Absenteeism time in hours” is continuous in nature. The table 2.2 represents count of each unique value in the target variable.

Table3.1: Absent vs. count of eemployee

No of hours absent	No of employee absent
8	199
2	155
3	110
1	88

4	60
0	36
16	19
24	16
40	7
5	7
32	5
64	3
80	3
120	3
56	2
112	2
104	1
7	1
48	1

If the dependent variable is categorical the only predictive analysis that we can perform is **Classification** and if the dependent variable is Interval or Ratio i.e. continuous the normal method is to do a **Regression** analysis or classification after binning. The dependent variable we are dealing with is Regression, for which regression is preferred according to problem statement. We should always start our model building from the simplest to more complex. Here we have used different ML Algorithm on trained data and then applied it on test to predict the future values in both R and Python. Different algorithm gave different results with different accuracy and RMSE.

Before developing any model we need to divide model into train (development) and test (validation). To validate our model first we build our model on training data and then we apply the same model on the test data to predict its target variable. This is because for test data we already have target value. Then we compare its predicted value with the actual value of test data and then we will try to extract accuracy of that model.

The Machine Learning Model we applied in R and Python are as follows:

3.2.1 *Decision Tree Regression*

Decision Tree is a predictive model based on branching series of Boolean tests. It can be used for both classification and regression. It is one of the most powerful and popular algorithm and belongs to family of supervised learning algorithm. Decision tree is a rule and output of decision tree is in form of simple business rules which is extremely easy to understand by business users.

In our project we have used Decision Tree for both R and Python to predict the output.

Table 3.2: Decision tree accuracy table

Decision Tree	R	PYTHON
RMSE	11.620	10.516
ACCURACY	88.38	89.484

3.2.2 *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 method combines Breimans "bagging" idea and the random selection of features.

Table 3.3: Random Forest accuracy table

Random Forest	R	PYTHON
RMSE (n = 100)	11.105	11.519
ACCURACY (n = 100)	88.895	88.481
RMSE (n = 200)	11.093	11.314
ACCURACY (n = 200)	88.907	88.686
RMSE (n = 300)	11.114	11.359
ACCURACY (n = 300)	88.853	88.641
RMSE (n = 500)	11.172	11.425
ACCURACY (n = 500)	88.828	88.575

3.2.3 *Linear 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 relationship: It assumes that the data which is fed in linear regression model have linear relationship between dependent and independent variable.

Multivariate normality: Linear regression assumes that our target variable is normally distributed. It means that it is following the normality assumption.

No or little Multicollinearity: Two highly correlated variables in a dataset lead to multicollinearity effect. There is one test called VIF test (Variance Inflation Factor test). We need to run this test before feeding the data to the model to know either our data contains the correlated independent variables or not.

No Auto Correlation: It means there should be no correlation between the residuals. When we build a linear regression model we will get residuals (range of errors). Here we assume that there is no auto correlation it means error are independent.

Once our data satisfy these assumptions we go ahead and build linear regression model. Under this model on training data we build equation which carries an intercept and coefficient for all independent variable. Then we save that equation and then once new test data comes in then we allow passing the test case on linear regression equation to estimate the predicted value. Whatever the value predicted that will hold a target value to the new test data.

Table 3.4: Linear Regression accuracy table

Linear Regression	R	PYTHON
RMSE	11.052	9.608
ACCURACY	88.948	90.392

Table3.5: Accuracy of all the Models

Model Name	R		PYTHON	
	RMSE	ACCURACY	RMSE	ACCURACY
Decision Tree	11.620	88.38	10.516	89.48
Random Forest n = 100	11.105	88.89	11.519	88.48
n = 200	11.093	88.90	11.314	88.68
n = 300	11.114	88.85	11.359	88.64
n = 400	11.172	88.82	11.425	88.57
Linear Regression	11.052	88.94	9.608	90.39

*Blue color shows the best accuracy among all the given above models.

CHAPTER - 4

Model Evaluation

After building number of regression models there are criteria by which they can be evaluated and compared. Model evaluation tells us whether our model is able to accomplish the business object or not. There are different metrics for regression model like **MSE (Mean Square Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error) etc.** MSE and RMSE are used for **transition or time series data also called time series analysis** whereas MAPE and MAE are used for normal regression data.

If dataset is transitioned or time based then we go for RMSE. If we want to convert error number in particular percentage we should go for MAPE. Our project is time series multivariate so we have used RMSE as error metric. Accuracy can be calculated as:

Accuracy = 100 - RMSE

4.1 RMSE (Root Mean Square Error)

RMSE is a popular metric to measure the error rate of time series or transition regression model. It can be only compared between models whose errors are measured in the same units. It can be calculated by squaring the errors, finding their average and taking their square root. It can be mathematically represented as:

$$RMSE = \sqrt{\frac{\sum_{i-1}^{n}(p_i - a_i)^2}{n}}$$

a – actual target

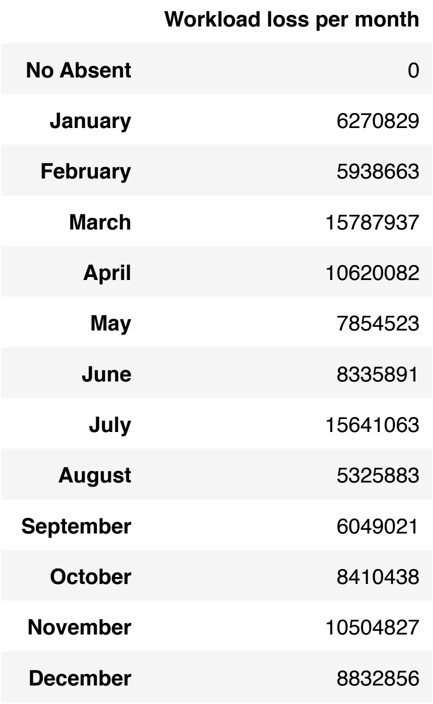
p – predicted target

CHAPTER - 5

WORK LOSS

The low work performance of the company leads to work loss. One of the major factors of work loss is employee absenteeism. In our project we have predicted the work loss faced by the company in year 2011. We have computed the monthly work loss the company is going to face in coming year with the help of the formula give below. Table 5.1 shows the monthly work loss the company is going to face in year 2011. Work loss can be calculated as:

Work loss = (Absenteeism time in hours * Work load average/day)/ Service time



	Workload loss per month
No Absent	0
January	6270829
February	5938663
March	15787937
April	10620082
May	7854523
June	8335891
July	15641063
August	5325883
September	6049021
October	8410438
November	10504827
December	8832856

Table 5.1: Work loss of all the months

CHAPTER - 6

CONCLUSION

In this project **“Employee Absenteeism”** we have applied different models to predict the final result. For each algorithm we got different accuracy and root mean square error which are acceptable in some case and not considerable in some cases. We applied the algorithm on both R and Python for the same dataset as mentioned in the project.

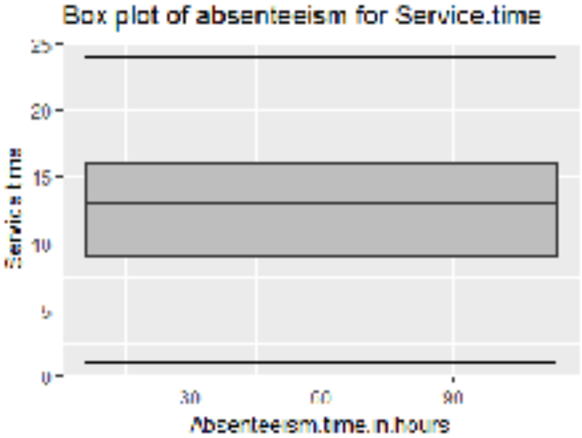
In R and Python we have selected Linear Regression for predictions as it gave the highest accuracy and rmse which **is 88.94, 90.39 (accuracy) and 11.052, 9.608 (rmse)** respectively. After applying linear regression in both Rand Python we concluded that most of the employees were absent due to health issues and the second reason for their absence was service time. Diseases of nervous system, respiratory system and skin disease are major cause of employee absenteeism. Through this study it is identified that the employees were mostly suffered from health issues. So the company might focus on these two given reasons to minimize absenteeism. Even it is possible to eliminate absenteeism completely by the way of providing valuable means to their internal resources i.e. employees by providing some medical facilities and by creating awareness regarding health. Though absenteeism is invisible but proves fatal for entire company. The work loss the company is going to face every month in year 2011 is given in table 5.1. So by taking preventive measure and reducing absenteeism will help to improve company’s performance.

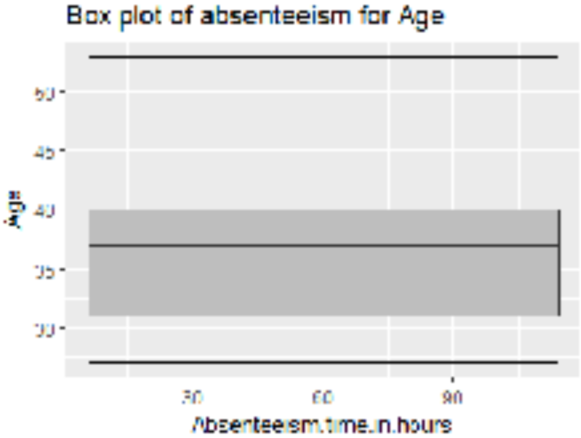
.

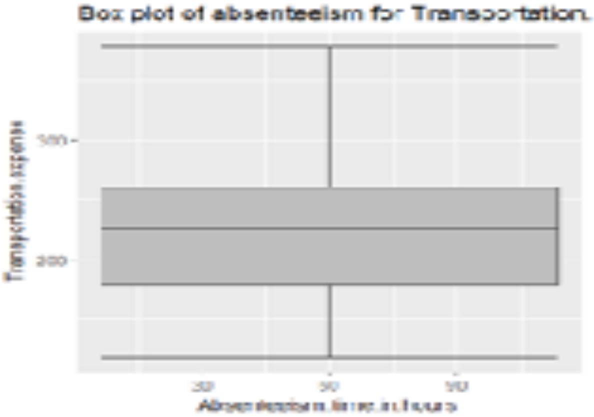
APPENDIX – A

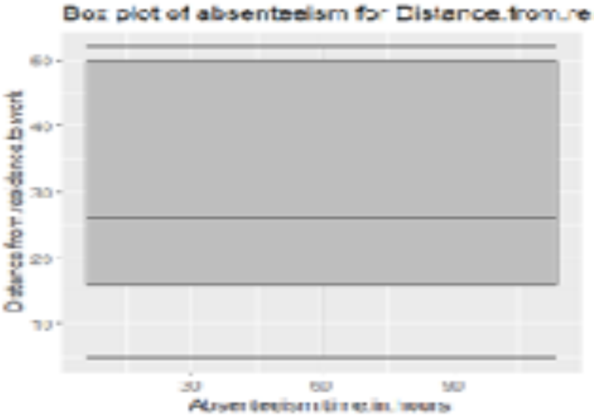
Extra Figures

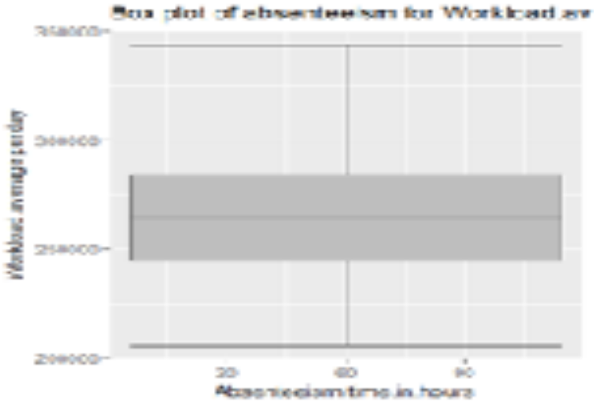
Plots after removing outliers

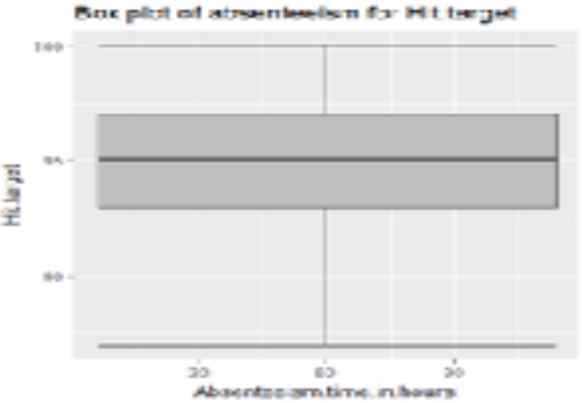

Box plot of absenteeism for Service.time
25
20
15
10
5
0
Service.time
30
60
90
Absenteeism.time.in.hours

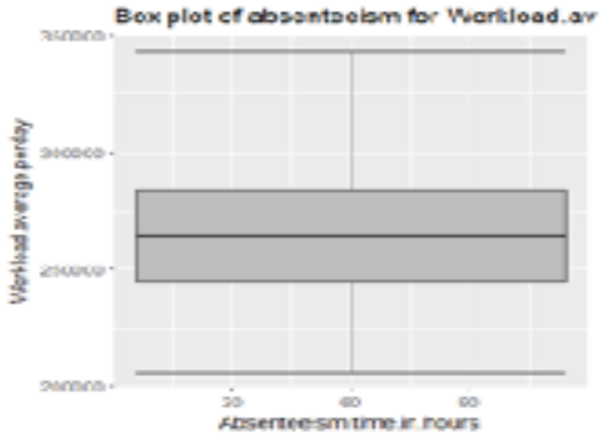

Box plot of absenteeism for Age
50
45
40
35
30
Age
30
60
90
Absenteeism.time.in.hours

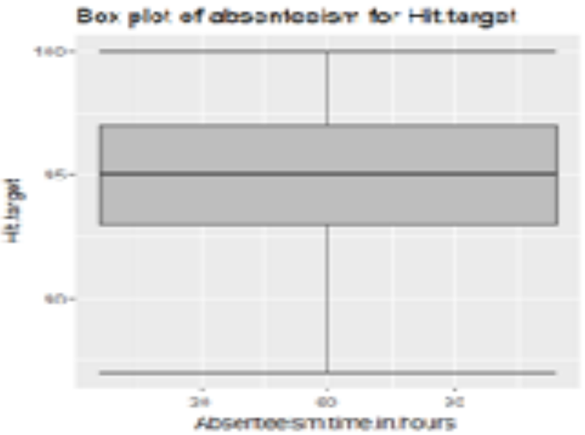

Box plot of absenteeism for Transportation.
300
200
Transportation.expense
30
60
90
Absenteeism.time.in.hours

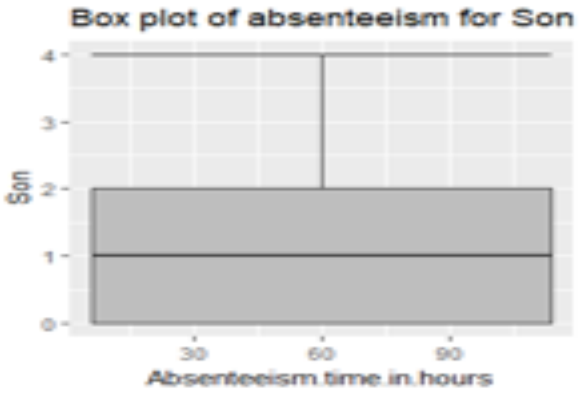

Box plot of absenteeism for Distance.from.re
50
40
30
20
10
Distance.from.residence.to.work
30
60
90
Absenteeism.time.in.hours

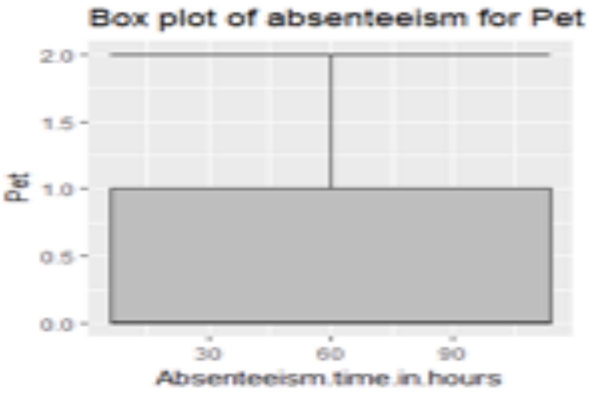

Box plot of absenteeism for Workload.av
350000
300000
250000
200000
Workload.average.per.day
30
60
90
Absenteeism.time.in.hours

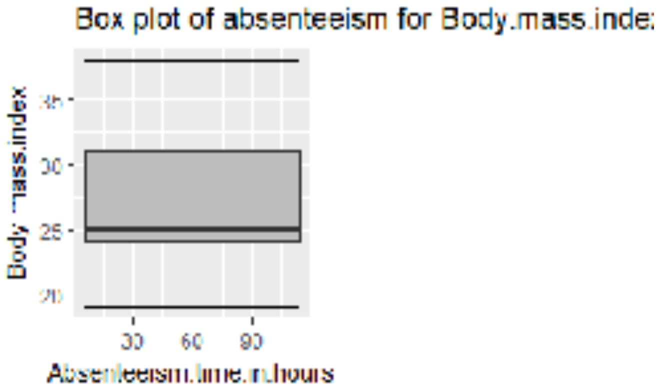

Box plot of absenteeism for Hit.target
100
95
90
Hit.target
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Workload.av
350000
300000
250000
200000
Workload.average.perday
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Hit.target
100
95
90
Hit.target
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Son
4
3
2
1
0
Son
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Pet
2.0
1.5
1.0
0.5
0.0
Pet
30
60
90
Absenteeism.time.in.hours


Box plot of absenteeism for Body.mass.inde:
35
30
25
20
Body.mass.index
30
60
90
Absenteeism.time.in.hours

APPENDIX - B

VARIANCE INFLATION FACTOR

VIF (Variance Inflation Factor) is used to detect and remove multicollinearity. It is one of the assumptions of linear regression. VIF is used only on independent variable. It is calculated by the formula,

$$VIF = 1/1-r^2$$

Where, r^2 = % variance in variables & $1-r^2$ also called tolerance of the model.

If r^2 is high it means the given variable is redundant. So we need not to bring the given variable in the model. It means the given variable is highly correlated. If r^2 is low it means the given variable is not redundant and we should include that variable in our model. It means the given variable is less correlated.

Higher the VIF more collinear is the variable which means we should not include that variable in our model. Lower the VIF less collinear is the variable which means it can be included in our model. The VIF values we obtained in our linear regression model are given in table below.

Variables Name	GVIF	Df	GVIF^(1/(2*Df))
Reason for absence	5.210478	27	1.031040
Month of absence	4.494072	12	1.064617
Day of the week	1.368016	4	1.039947
Transportation expenses	1.729534	1	1.315117
Distance from home	1.398463	1	1.182566
Service time	1.603039	1	1.266112
Workload avg per day	1.646295	1	1.283080
Weight	1.417460	1	1.190571

APPENDIX - B

Basic Output Terms

Residual standard error: It is also called standard deviation error. It measures the average amount that the coefficient estimates vary from actual average value of our response variable. It helps in calculation of p-value.

t- value: It measures how many standard deviation our coefficients are away from 0. Coefficients should be far away from zero because if coefficient of any variable is near to 0 it means that variable is not able to explain the target variable i.e. that variable is an irrelevant variable. With help of t-value we calculate p-value.

p-value: It helps us to decide whether to accept or reject the variable i.e. a variable is contributing much information or not.

F-statistics: It is a good indicator of whether there is a relationship between our predictor and the response variable. F-statistics should be greater than 1.

Degrees of Freedom: Number of observation (training data) – Total number of variable

R Square: It is numerical value which tells us the amount of variance of the dependent variable is explained by all independent variable. It tells us how much our model is robust and what the strength of model on training data is.

Adjusted R Square: It is derived from R-Square values. Adjusted R Square will penalize the effect of additional variables which are not carrying much information. It should be always less than R Square.

AIC (Alkaline Information Criteria): It adjusts the loc likelihood based on the number of observation and complexity of the model.

BIC (Baisen Information Criteria): It is similar to AIC but has high penalty for models.

Omnibus: Provides combined statistical test for the presence of skewness and kurtosis. Basically it is breakdown of skewness and kurtosis.

Skew and Kurtosis: These tests are basically for time series dataset.

Null Deviance: It tells us how well the response variable is predicted by the model with intercept only.

Residual Deviance: It tells us how well the response variable is predicted by using null deviance and all other independent variables.

REFERENCE

1) “*Edwisor Videos”*

2) “*Bharatendra Rai”* – *“YouTube Channel”.*

3) “Machine Learning with R” – by *“Brett Lantz”.*

4) “An Introduction to Data Cleaning with R” – by *“Edwin de Jonge and Mark van der Loo”.*