PROJECT REPORT

ON

EMPLOYEE ABSENTEEISM

Mohit H. Kothari

25 Aug 2019

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 continuous in nature, 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 Variable)

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.

```
# Check count of unique values Present in data set
df.nunique()
```

```
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
dtype: int64
```

Fig 1: Count of unique values and column names of dataset

From here we have categories variable into 2 broad categories:

1. **category = ["Reason for absence","ID","Month of absence","Day of the week","Seasons","Disciplinary failure","Education","Social drinker","Social smoker","Son","Pet"] (These all variables will be classified as categorical variables)**

2. **numeric = ["Transportation expense","Distance from Residence to Work","Service time","Age","Work load Average/day ","Hit target","Weight","Height","Body mass index","Absenteeism time in hours"] (These all variables will be continuous variables)**

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. The basic step here will involve Missing Value analysis, Then followed by outlier analysis and then data visualizations followed with feature engineering.

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. I

In the given data many observations are given as 0, so we have treated them also as missing value and we have tried
Mean – Median Method for Numerical data and Mode method for categorical Data

Finally we imputed our value with Median-Mode Method

2.1.2 Outlier Analysis

For all continuous (numeric) variable we will do box plot analysis (outlier), in box plot it plots lower limit, upper limit and 25 quartile and 75 quartile depending on dataset values. Any value going out of the lower or upper fence is treated as outliers and then further treatment is done.

Here Many variables are having outliers but we wont remove them or replace with mean/median method because these outlier might represent extreme case of data upon which employee absenteeism is strongly related so we will keep as it is.

No Outlier is Removed or Replaced.

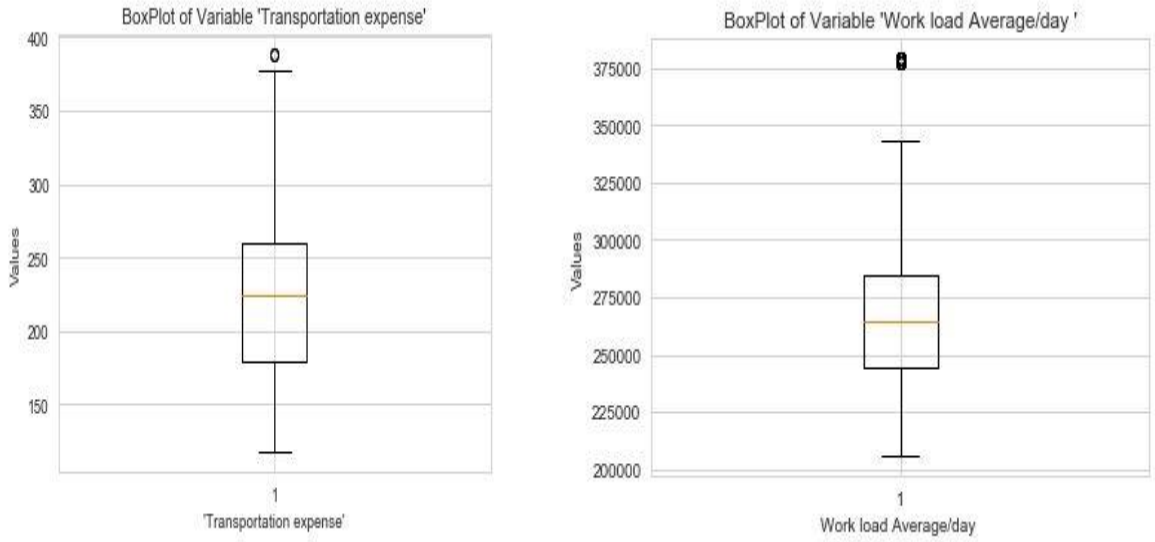


Fig2: Box plot of 'Transportation Expense' and 'Work load Average/day'

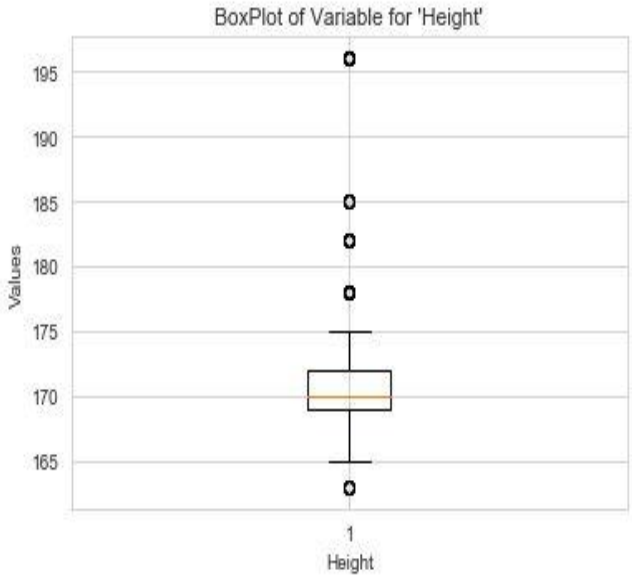


Fig3: Box Plot of Variable 'Height'

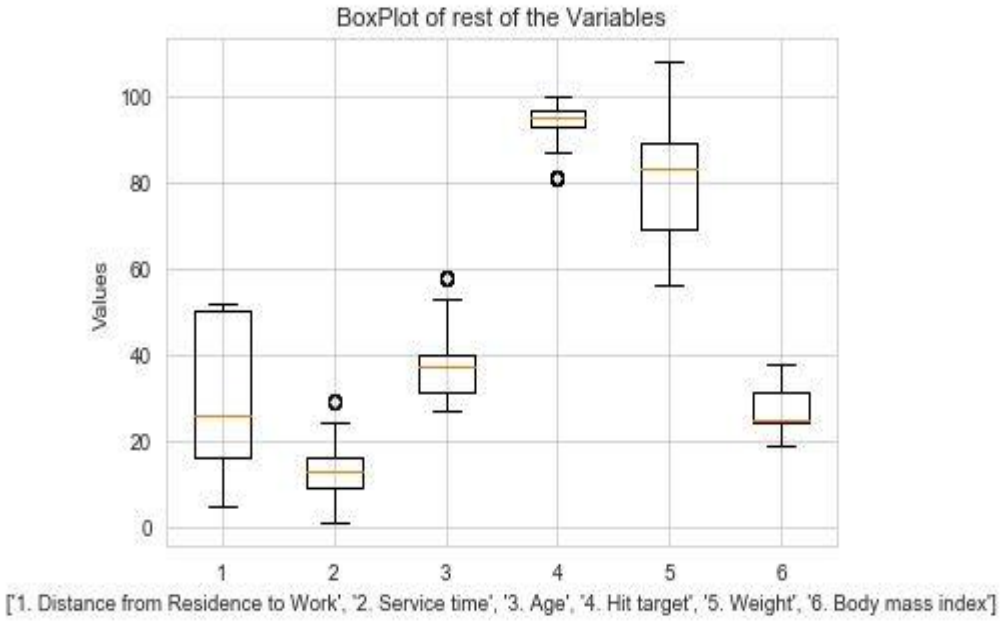
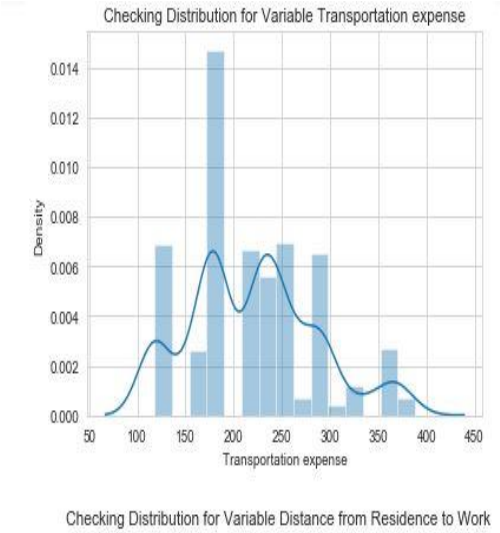


Fig4: Box Plot of variables 'Distance from Residence to work', 'Service time', 'Age', 'Hit target', 'Weight', 'Body Mass index'.

2.1.3 Visualizations – Frequency Distribution

Here to analyses how our continuous variable is distributed with outcome variable or what is frequency of continuous variables in our dataset we plotted histograms to explain their behavior.



Checking Distribution for Variable Distance from Residence to Work

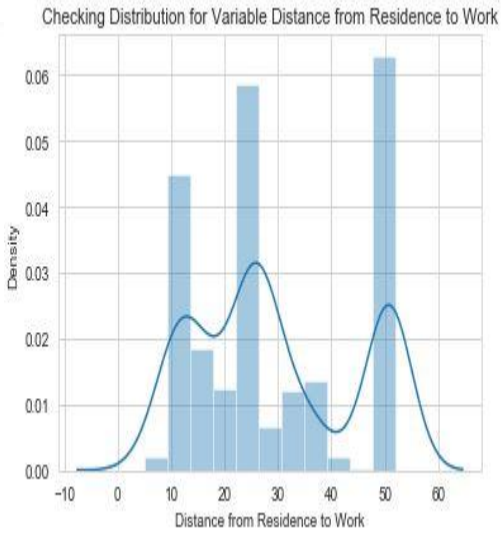


Fig5: Density Plot of 'Transportation Expense' and 'Distance from Residence to work'

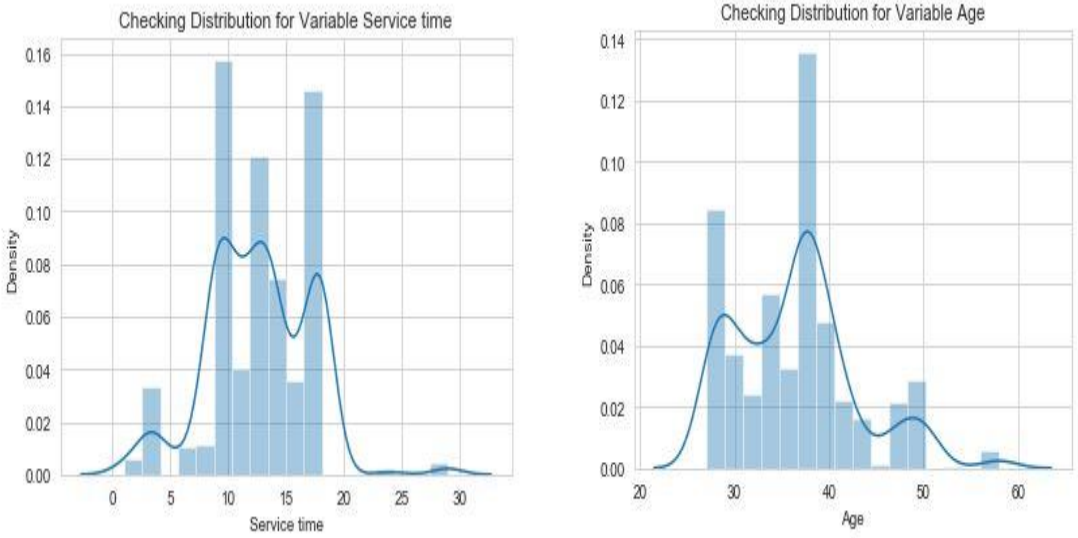


Fig6: Density Plot of 'Service Time' and 'Age'

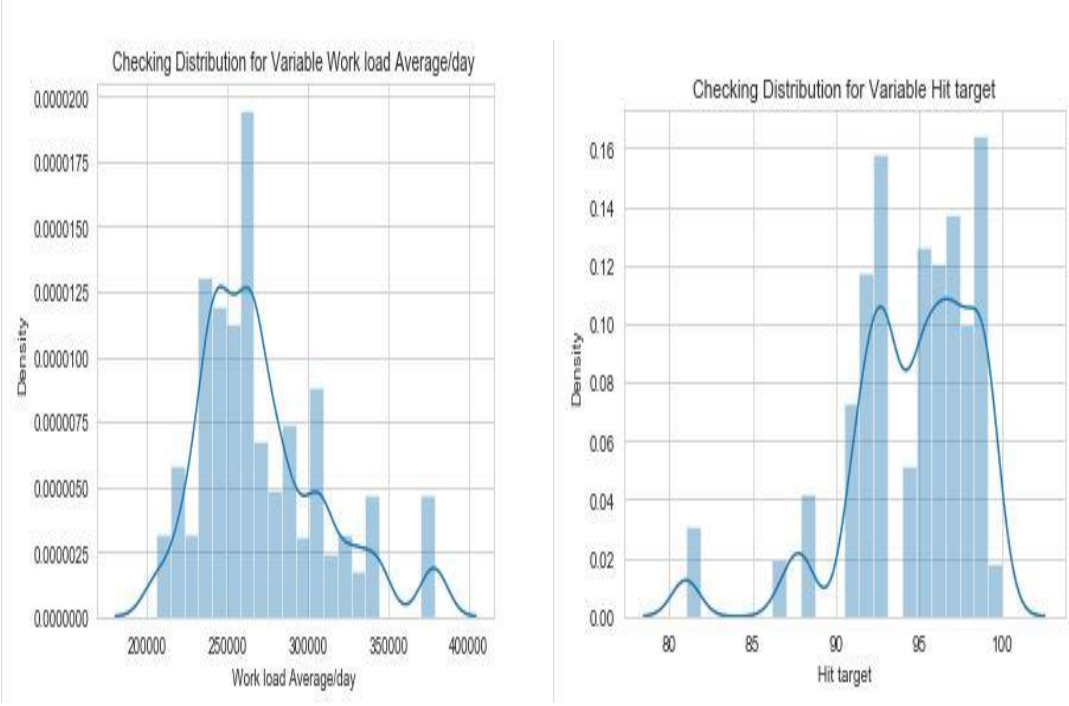


Fig7: Density Plot of 'Work load Average/day' and 'Hit target'

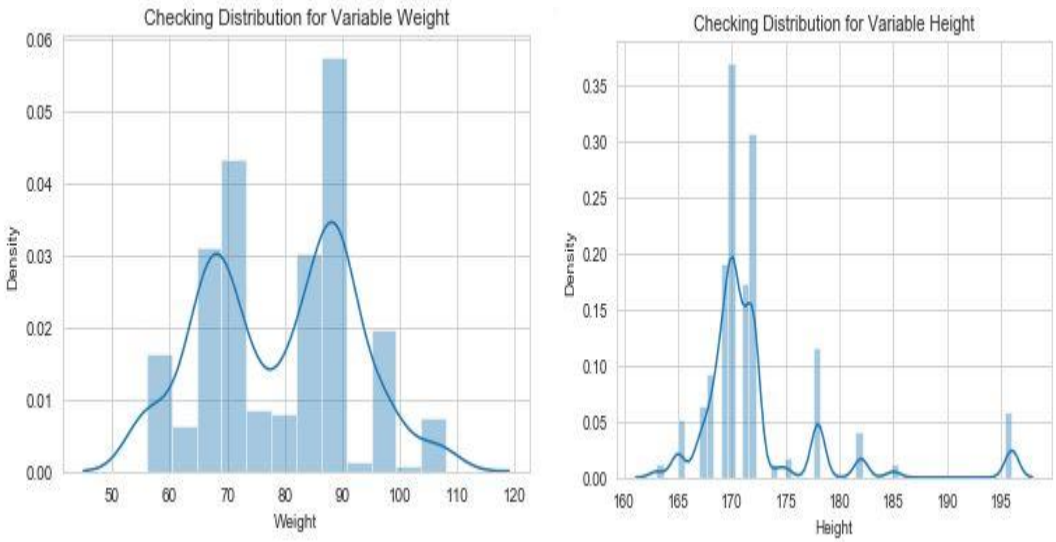


Fig8: Density Plot of 'Weight' and 'Height'

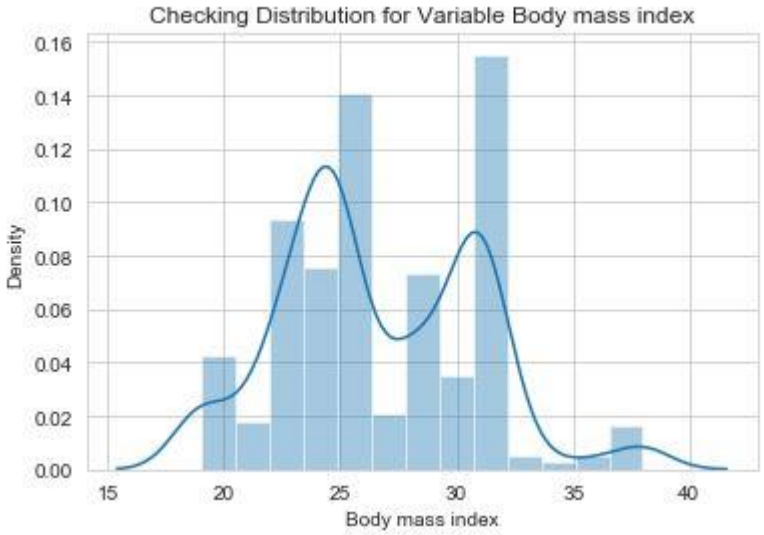


Fig7: Density Plot of 'Body Mass index'

2.1.4 Correlation Analysis

In Correlation analysis we have considered all continuous variable and plotted relation between them w.r.t output variable

So Finally weight was 0.9 correlated with BMI, so we removed weight.

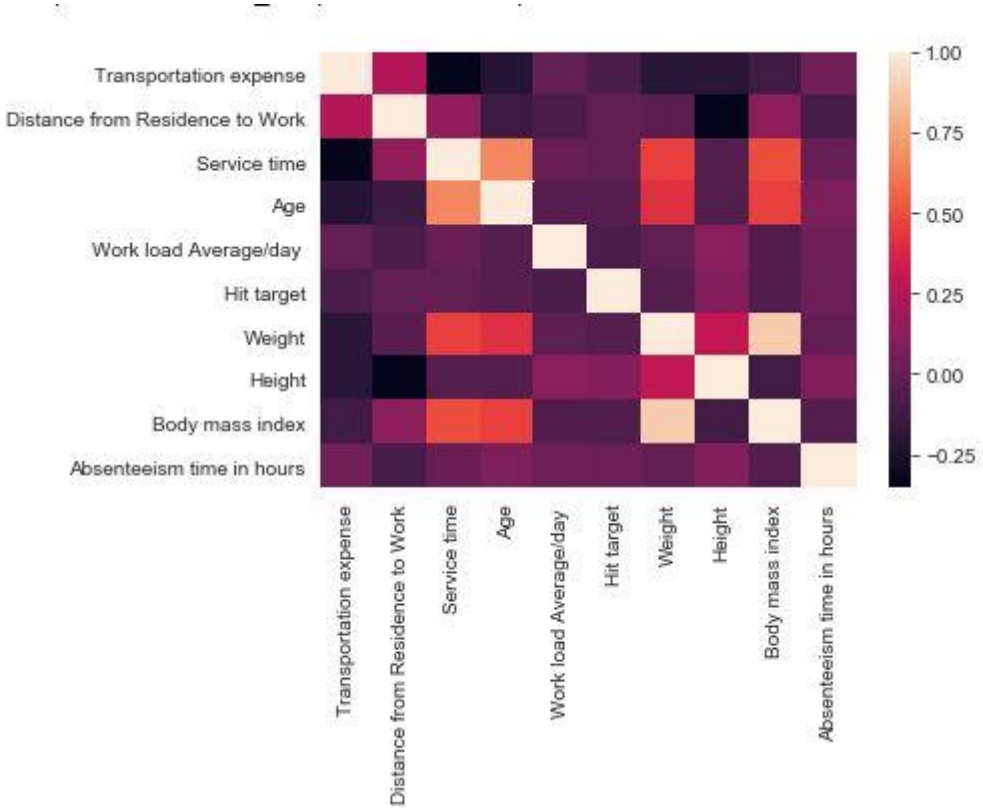


Fig8: Correlation Analysis of Continuous Variable

2.1.5 Normalization

We have cleaned the data copied it for future use and Since all continuous variable were not normalized, we normalized all continuous variables so when we apply linear regression model to it there is no problem of unequal weights in analysis.

So we normalized all Continuous variables and Converted all Categorical variable to dummy variables

2.2 Modeling

After a thorough preprocessing we will be using some regression models on our processed data to predict the target variable. Following are the models which we have built –

2.2.1 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	0.068	5.69
RMSE Test	0.103	11.7
R^2 Test	0.03	-0.04

2.2.2 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 Regression	R	PYTHON
RMSE Train	0.093	11.72
RMSE Test	0.1137	5.12e+15
R^2 Test	0.04	-2.063

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) 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

From the observation of all **RMSE Value** and **R-Squared** Value we have concluded that **Random Forest Regression Model** has minimum value of RMSE and it's **R-Squared** Value is also maximum (i.e. 0.04).

3.2 Answers of asked questions

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

1. It is observed that employee with Reason for **Absence 23 medical consultation and Reason of absence 28 Dental Consultation are most time absent.**
Sol: Company can provide monthly dental check up and Medical camps for their employees to reduce this reason.

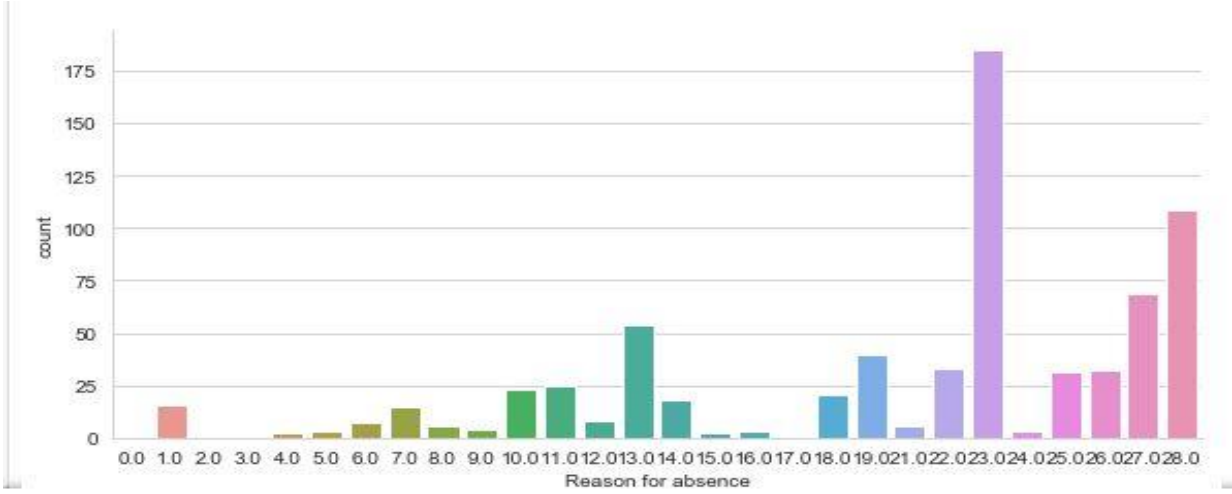


Fig9 : Graph of 'Reason of Absence' with respect to overall Count

2. Some employee with **ID 3, 28, 34** are often absent from work, company should take action against them.

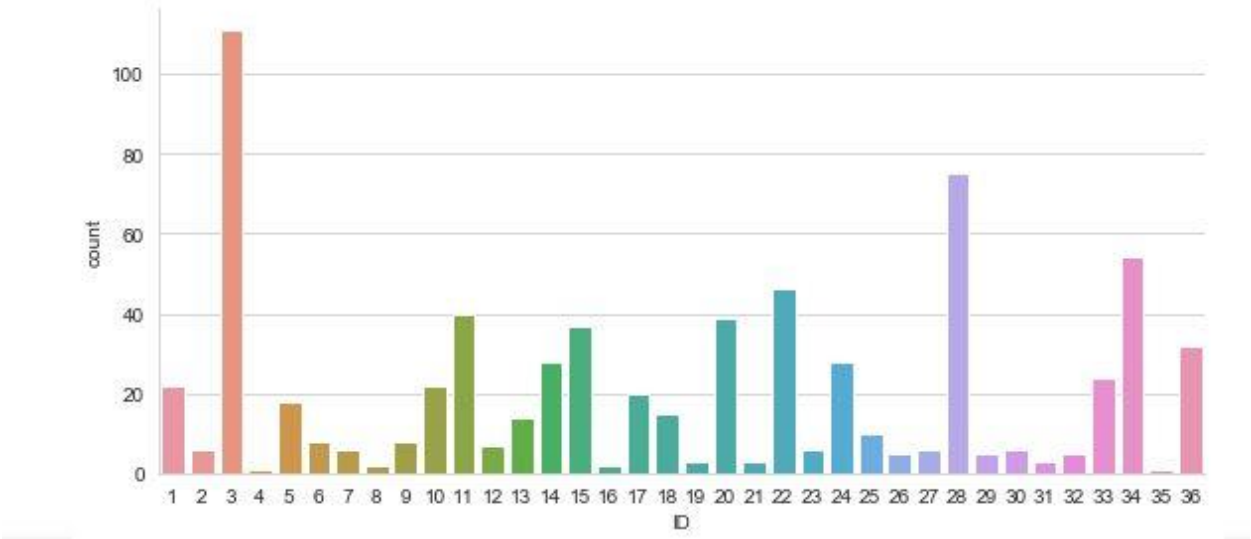


Fig10 : Graph of 'ID' with respect to overall Count

3. **Employee with high school education is more absent than other with graduate/post graduate.**

Sol : Company must hire Graduate employees at least, so that they have of responsibility and remain less absent.

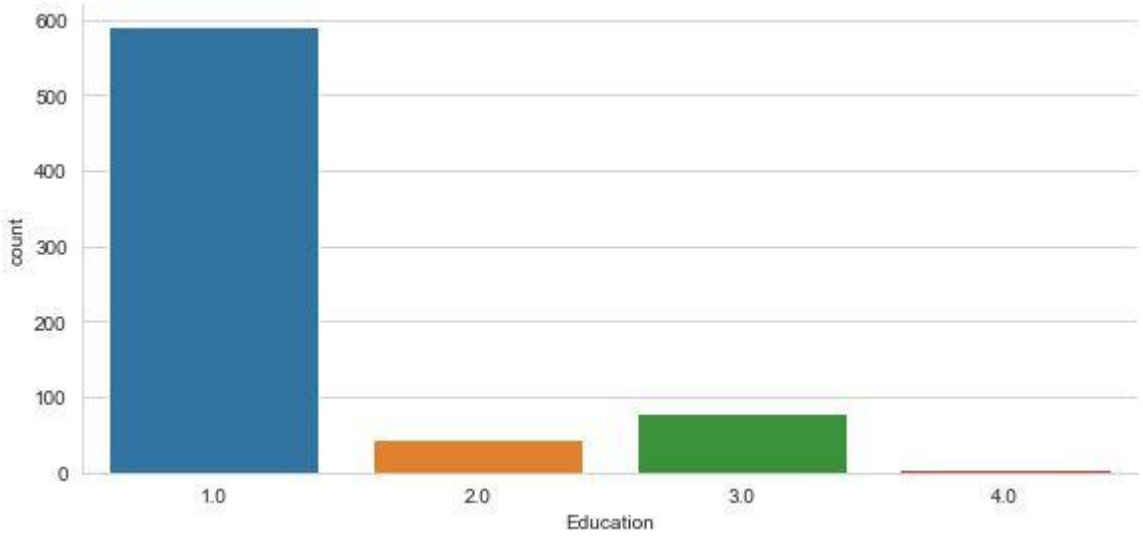


Fig11 : Graph of 'Education' with respect to overall Count

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

Here we calculated Loss work as: [((Work Load Average/day) / Service time) * (Absenteeism time in hours)].

Based on this we calculated the work loss for each month by using groupby command for month of absence variable.

Month of absence	Work Loss
0.0	0.000000e+00
1.0	6.312631e+06
2.0	8.268540e+06
3.0	1.574969e+07
4.0	1.099949e+07
5.0	9.326395e+06
6.0	1.436224e+07
7.0	1.901538e+07
8.0	8.791557e+06
9.0	6.482818e+06
10.0	8.895573e+06
11.0	1.225236e+07
12.0	1.223330e+07

Fig 12 : Work Loss for Each month