Employee Absenteeism

Project Report

By

Ishan Sharma

Contents

.

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 loss every month can we project in 2011 if same trend of absenteeism continues?

Employee Absenteeism is the absence of an employee from work. It is an issue faced by every other employer. Absence of employees results in the fall of productivity for any company. Employee absenteeism leads to pending work, delayed deadlines which can really hamper the image of an organisation.

1.2 Data

Our task is to build a regression model which will predict the absenteeism in hours per employee based on the employee attributes and information in their work place and general information available to the company about them. Given below is the sample of data we have in hand:

Table 1.1: Employee Absenteeism Sample Data (Columns: 1-6)

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportati on expense
11	26	7	3	1	289
36	0	7	3	1	118
3	23	7	4	1	179
7	7	7	5	1	279
11	23	7	5	1	289

Table 1.2: Employee Absenteeism Sample Data (Columns: 7-12)

Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure
36	13	33	239554	97	0
13	18	50	239554	97	1
51	18	38	239554	97	0
5	14	39	239554	97	0
36	13	33	239554	97	0.

Table 1.3: Employee Absenteeism Sample Data (Columns: 13-18)

Education	Son	Social drinker	Social smoker	Pet	Weight
1	2	1	0	1	90
1	1	1	0	0	98
1	0	1	0	0	89
1	2	1	1	0	68
1	2	1	0	1	90

Table 1.4: Employee Absenteeism Sample Data (Columns: 19-21)

Height	Body mass index	Absenteeism time in hours
172	30	4
178	31	0
170	31	2
168	24	4
172	30	2

The details of data attributes in the dataset are as follows -
Dataset Characteristics: Timeseries Multivariant
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 behavioural disorders
VI Diseases of the nervous system
VII Diseases of the eye and adnexa
VIII Diseases of the ear and mastoid process
IX Diseases of the circulatory system
X Diseases of the respiratory system
XI Diseases of the digestive system
XII Diseases of the skin and subcutaneous tissue
XIII Diseases of the musculoskeletal system and connective tissue
XIV Diseases of the genitourinary system
XV Pregnancy, childbirth and the puerperium
XVI Certain conditions originating in the perinatal period
XVII Congenital malformations, deformations and chromosomal abnormalities
XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX Injury, poisoning and certain other consequences of external causes
XX External causes of morbidity and mortality
XXI Factors influencing health status and contact with health services.
And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).
3. Month of absence
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
5. Seasons (summer (1), autumn (2), winter (3), spring (4))
6. Transportation expense
7. Distance from Residence to Work (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)

From the table, we have the following 20 variables, using which we have to accurately estimate the hourly absent time:

Table 1.5: Predictor Variables

S.No.	Predictor
1	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	Education
14	Son
15	Social drinker
16	Social smoker
17	Pet
18	Weight
19	Height
20	Body mass index

Chapter 2

Methodology

2.1 Pre Processing

Any predictive modelling requires that we look at the data before we start modelling. 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**.
To start the process, we would first evaluate whether the given dataset has any kind of missing information.

2.1.1 Missing Value Analysis

If the records (observations) in a dataset contain any missing information then it is important to treat those missing values in order to build an optimised prediction model. The missing values, if present, can be imputed using several methods such as KNN imputation, mean imputation or mode imputation.
As a rule of thumb, if any variable in the dataset has more than 30% values missing, then that variable is removed altogether from the model as imputing such a variable would cause highly engineered results.

Missing values in a variable can be checked using **is.null()** function in Python and using **is.na()** in R.
The given dataset contains at the max 4% missing value for a variable. Since, this is way less than the 30% mark, the missing values are imputed using the KNN imputation method which works on the Euclidean distance.

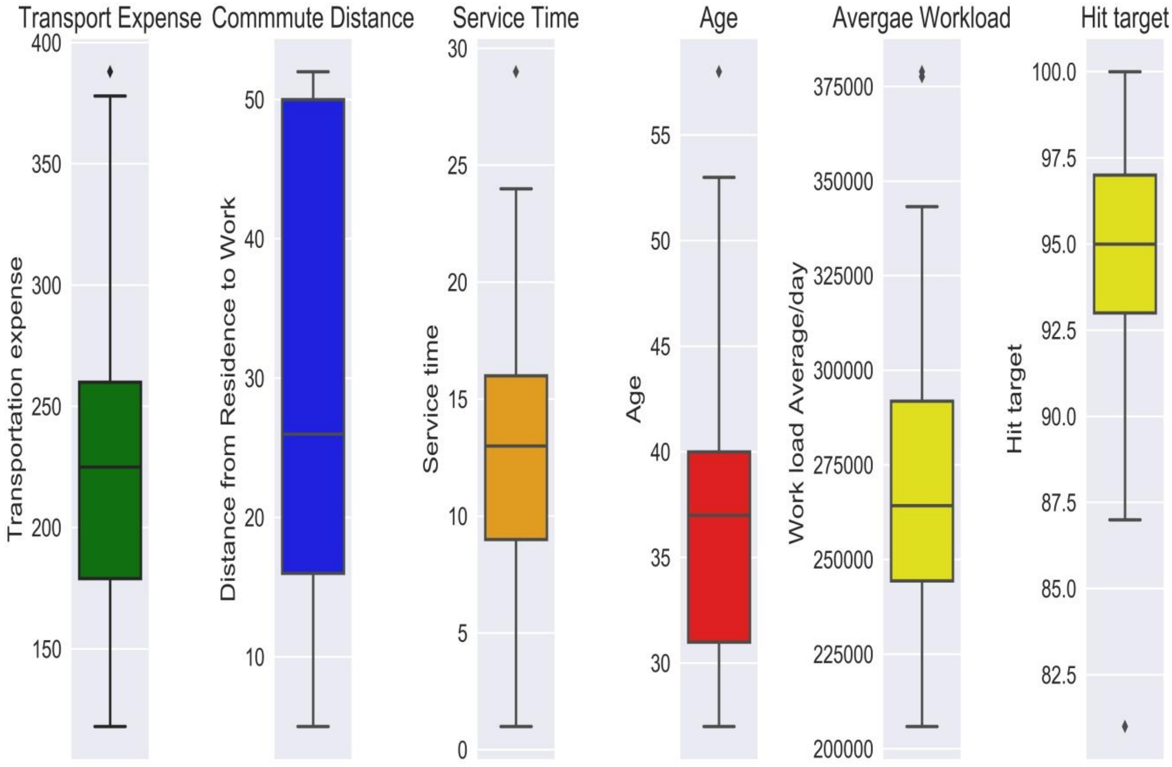
2.1.2 Datatype Analysis

For proper EDA, it is important that the variables present in the dataset have correct datatype associated with them.
Variables are of 2 types:
1) Numerical
2) Categorical
Categorical variables are further divided into 2 types: Ordinal & Nominal
Numerical variables are those which take numerical values and it makes sense to do arithmetic operations upon them.
Categorical variables are those which take distinct categories as their value and it doesn’t make sense to do arithmetic operations on these.
In the given dataset,
Variables: **ID**, **Reason for absence**, **Month of absence**, **Day of the week**, **Seasons**, **Disciplinary failure**, **Education**, **Social drinker**, **Social smoker** are categorical variables having integer encoded values.

Variables: **Transportation expense**, **Distance from Residence to Work**, **Service Time**, **Age**, **Work load Average/day**, **Hit target**, **Weigh**t, **Height, Body mass index** are numerical variables.

2.1.3 Outlier Analysis

Outliers are the data points which are significantly lesser or greater as compared to the other data points. Statistical parameters such as mean, standard deviation are highly sensitive to the outliers because they tend to shift the values towards themselves.
So, in our model, we use the approach of Flooring and Capping the outliers in which the outliers more above the upper fence are made equal to the upper fence and those below the lower fence are made equal to the lower fence. Also, KNN imputation for outlier analysis is used in the R code.
The Tukey's method where the outliers are defined as the data points which are ±1.5 *IQR*.
We visualize the outliers using *boxplots*.
In figure 2.1 we have plotted the boxplots of the 4 predictor variables. As shown in the figure 2.1, there are outliers present in the variables.



. Figure 2.1: Boxplots of Numerical Variables

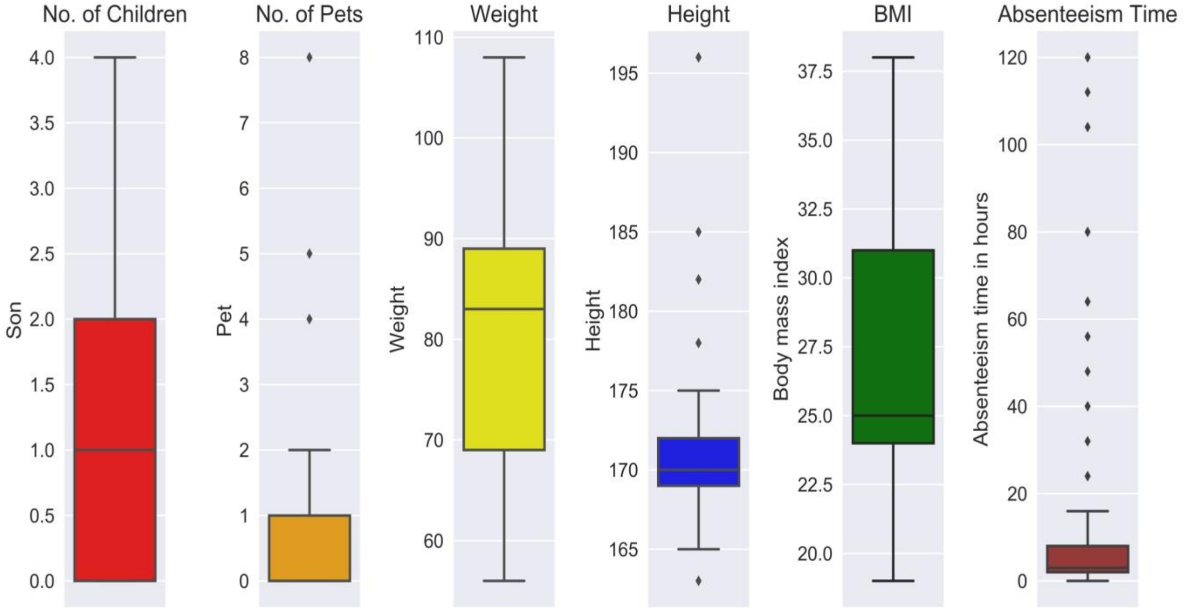


Figure 2.2: Boxplots of Numerical Variables

2.1.4 Feature Selection

Before performing any type of modelling 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 task of prediction of Absent time.

Numerical Feature Selection:

To asses, whether the predictor variables suffer from the problem of multicollinearity, we use correlation heatmap as show in Figure 2.2.

Note that correlation analysis can be done only for the numerical variables.

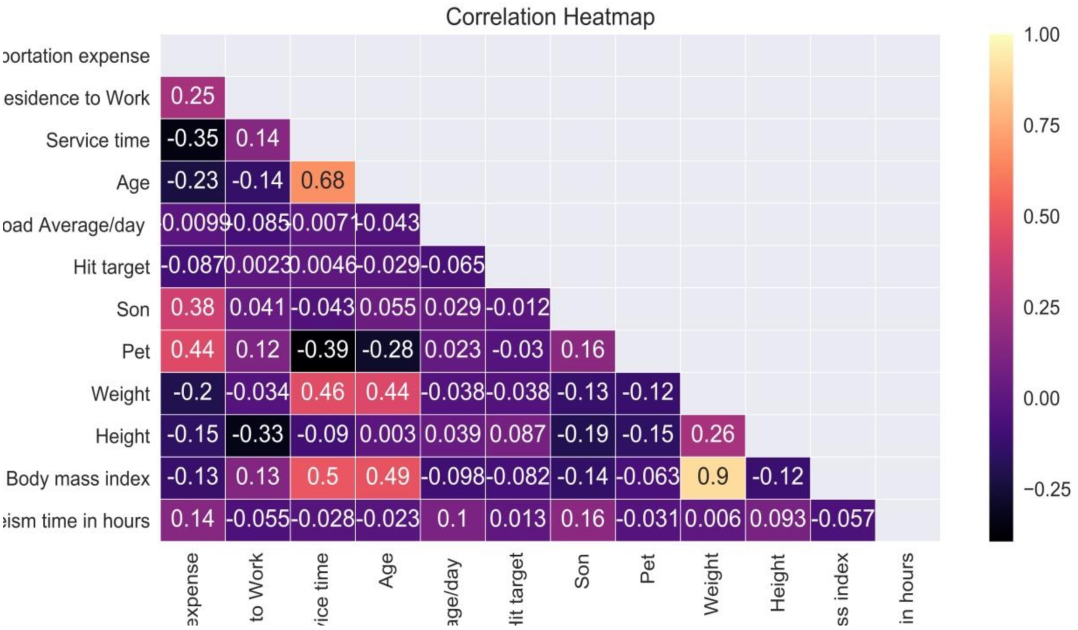


Figure 2.3: Correlation Heatmap

Since, the correlation coefficient of **Weight** and **Body mass index** variables is 0.9, which is unusually very high, one of these predictor variables should be dropped from the model. So, we drop **Weight** predictor variable.

Principal Component Analysis:

We use a more advanced technique for dimensionality reduction called PCA .While doing this; we first plot a cumulative distribution function plot to observe how much percentage of variance is explained by how many variables (Principle Components). The CDF plot for the same is plotted below:

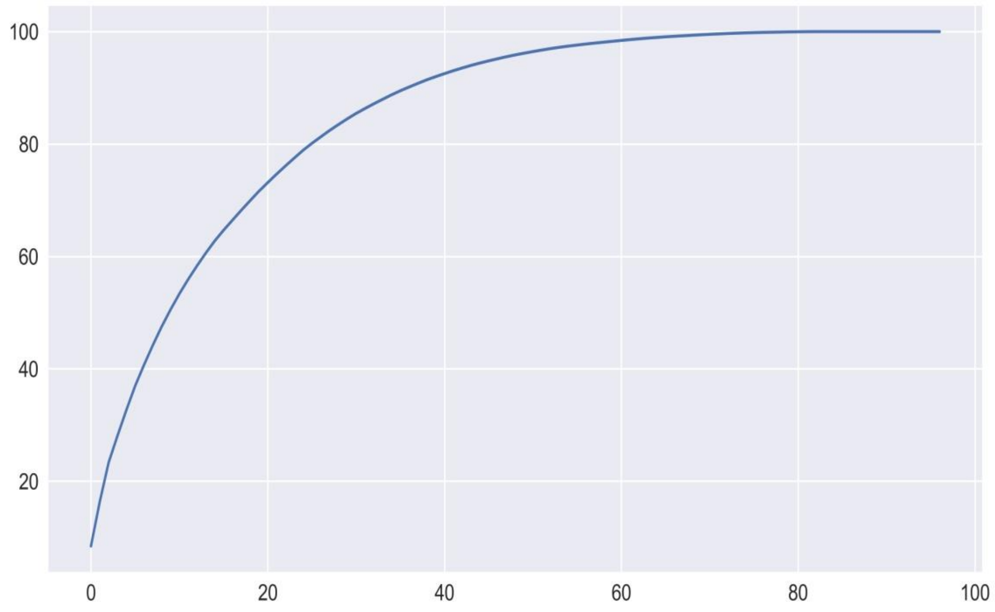


Figure 2.4: Explained Variance v/s N_Components

It is very clear from the above CDF plot for ‘Variance Explained’ Vs ‘Principle components’ that almost 95%+ variance is explained by just 45 variables (Principle components). We, can imagine how powerful PCA is, It just shrank down our feature space to just 45 from a total of 97 features. So, we will keep only 45 principle components in the data and will perform modeling on it.

2.2 Model Development

2.2.1 Model Selection

The dependent variable can fall in any of the four categories:
1. Numerical
2. Categorical

If the dependent variable, (in our case **Absenteeism time in hours)**, is Numerical, the only predictive analysis that we can perform is **Regression** and if the dependent variable is Categorical, the predictive analysis performed is **Classification.**

The dependent variable in our project is numerical, so regression models would be used.

Model development should always start from simpler models and then proceed to more complex ones.

In this project, we would be analysing three Regression models:

1) Linear Regression
2) Decision Tree Regression
3) Random Forest Regression

Based on the performance in accurately predicting the results, the best model out of the three will be selected.

2.2.2 Linear Regression

Linear regression is a simple approach to supervised learning. It assumes that the dependence of dependent variable, Y on predictor variables X1, X2, . . . Xp is linear.

Multiple Linear Regression algorithm using Ordinary Least Squares, the simplest of all.Ordinary least squares (OLS) minimises the squared distances between the observed and the predicted dependent variable. So, we get the following results after implementing the model :

```
# Creating Linear Model object:
lm = LinearRegression()

# Fitting the training data:
lm.fit(X_train,y_train)

Linear Regression:
The Mean Absoulte Error[MAE] is: 2.459648947751508
The Mean Squared Error[MSE] is: 13.525616891549129
```

2.2.3 Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

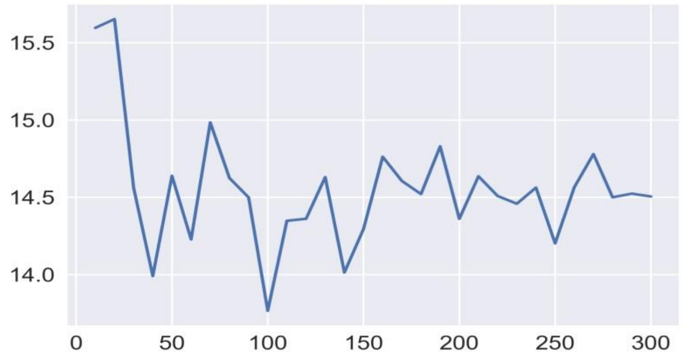
```
# Creating Decision Tree Regression Model object:
dt = DecisionTreeRegressor()

# Fitting the training data:
dt.fit(X_train,y_train)
```

```
Decision Tree:
The Mean Absoulte Error[MAE] is: 3.533333333333333
The Mean Squared Error[MSE] is:   31.022222222222222
```

2.2.4 Random Forest Regression

Random forest regression requires passing the value of **ntree** parameter which is the no. of tress employed in estimating the dependent variable.
Using the elbow method, which tests the error rate for various values of **ntree**, the optimum value of **ntree** can be used. (See full code in Appendix)



```
# Getting the Optimum value for ntree parameter:
for i in range(0,len(k)):
    if(error[i]== np.array(error).min()):
        opt = k[i]

rf = RandomForestRegressor(n_estimators= int(opt))


# Fitting the training data:
rf.fit(X_train,y_train)
```

```
Random Forest Regression:
The Mean Absoulte Error[MAE] is: 2.6891977383546006
The Mean Squared Error[MSE] is:   14.037212569677594
```

Chapter 3

Conclusion

3.1 Model Evaluation

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Predictive Performance
2. Interpretability
3. Computational Efficiency

In our case of Employee Absenteeism, the latter two, *Interpretability* and *Computation Efficiency*, do not hold much significance. Therefore we will use *Predictive performance* as the criteria to compare and evaluate models.

Predictive performance can be measured by comparing Predictions of the models with real values of the target variables, and calculating some average error measure.

3.1.1 Mean Absolute Error [MAE]

MAE is one of the error measures used to calculate the predictive performance of the model. We have applied this measure to our models that we have generated in the previous section. MAE calculates the total sum of individuals error in prediction of each observation and then averages it for the total no. of observations.

3.1.2 Mean Squared Error [MSE]

MSE takes the sum of the squares of the individual errors and then averages it for the total no. of observations.

Because of the square, large errors have relatively greater influence on MSE than do the smaller error. Therefore, MAE is more robust to outliers since it does not make use of square. On the other hand, MSE is more useful if we are concerned about large errors whose consequences are much bigger than equivalent smaller ones. MSE also corresponds to maximizing the likelihood of Gaussian random variables..

3.2 Model Selection

Error Metrics of Linear Regression are the best out of all the tested models, so Linear Regression model is the selected model for the Employee Absenteeism Analysis.

3.3 Questions/Answers

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

By the plotted visualizations and the EDA, following observations come out:
1) The rate of Absenteeism is maximum in Season 3: Winter followed by Season 1 : Summer , Season 4 : Spring, Season 2 : Autumn.

2) Also, We can say that the 'Absenteeism rate' is maximum in Month 7 : July followed by Month 4 : April , Month 3 : March, Month 12 : December, Month 11 : November , Month 6 : June , Month 5 : May etc.

3) Looking at the Bar plot of 'Absenteeism rate' Vs 'Day of the week', it can clearly be observed that the 'Absenteeism rate' is maximum on the third day of the week i.e Day 3 : Tuesday followed by Day 2 : Monday, Day 4 : Wednesday. Also, the 'absenteeism rate' is lowest on Day 6: Friday followed by Day 5: Thursday.

4) From the Bar plot of 'Absenteeism rate' Vs 'Reason of absence' we can observe that '9 : Diseases of the circulatory system' is the most frequent reason for the absence of the employees. The second most frequent reason given by the employees for their absence is '2 : Neoplasms' followed by '6 : Diseases of the nervous system', '12: Diseases of the skin and subcutaneous tissue', '19 : Injury, poisoning and certain other consequences of external causes' etc.

Steps for reducing the absenteeism time:

1) Since the diseases contribute significantly to the Absent time of employees in the company, health schemes can be provided to the employees in order to boost the productivity.
2) We can also we can come up with other ideas like: An incentive or conversion scheme for unused sick days.
3) Also, strict action could be taken towards the employee with high absence rate in the workplace without any valid reason for the absence and Employees with no absence or a minimum absence can be rewarded with perks.
4) Regular employee centric programmes and campaigns can be carried out. This would change the outlook of the employees towards the company.

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

Using basic unitary method, loss per month can be formulated as:
Loss = (Average work load/day * Absenteesim Time In Hours) / Service Time

Grouping the data by Month of absence variable and using the above formula, we can get monthly trend.
Since, the Absenteeism time in hours was highest for the months of July among all months and the workload per day and the service time were more or less similar, going by this trend, we can expect that the loss would be highest in the month of July.

Appendix A - Extra Figures

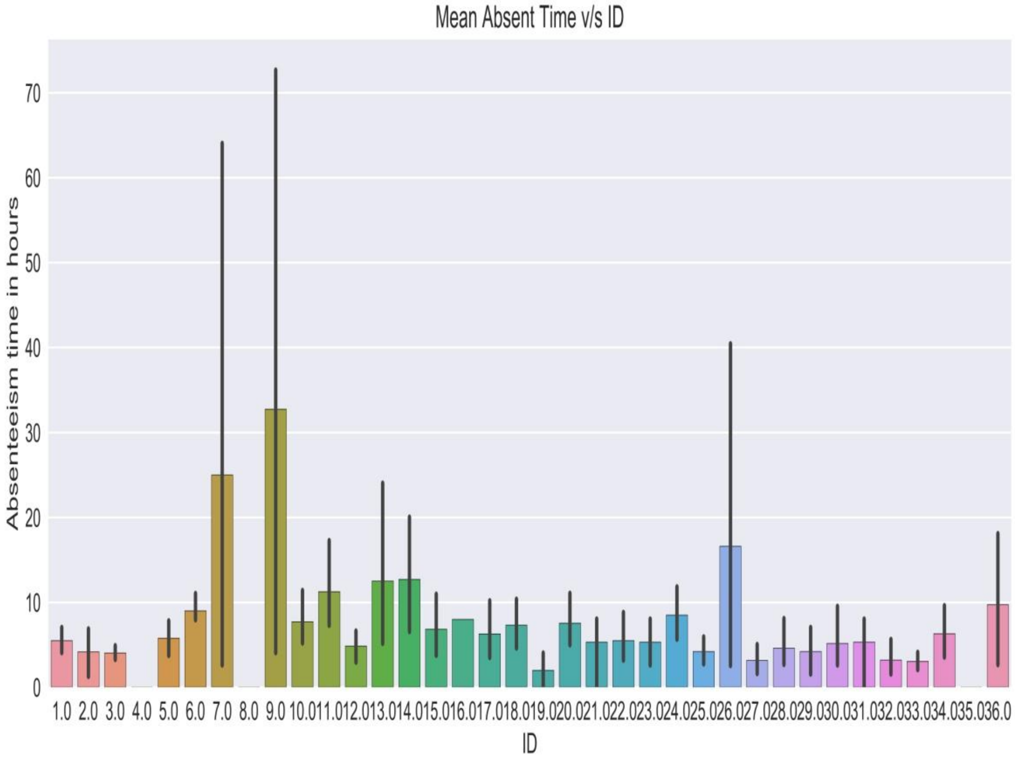


Figure 3.1: Bar plot of ID

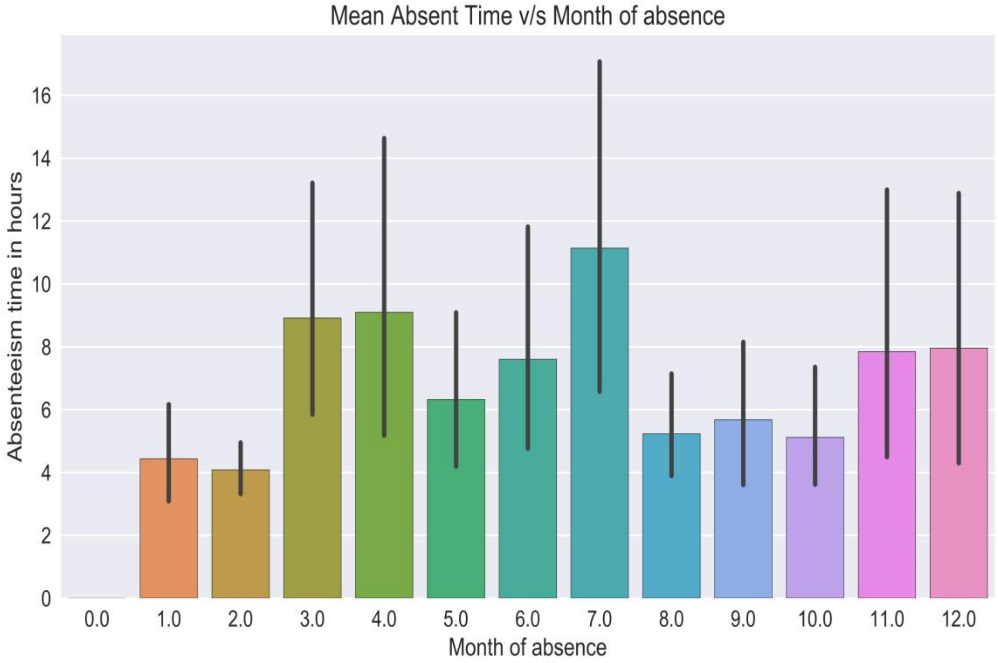


Figure 3.2: Barplot of Absent Month

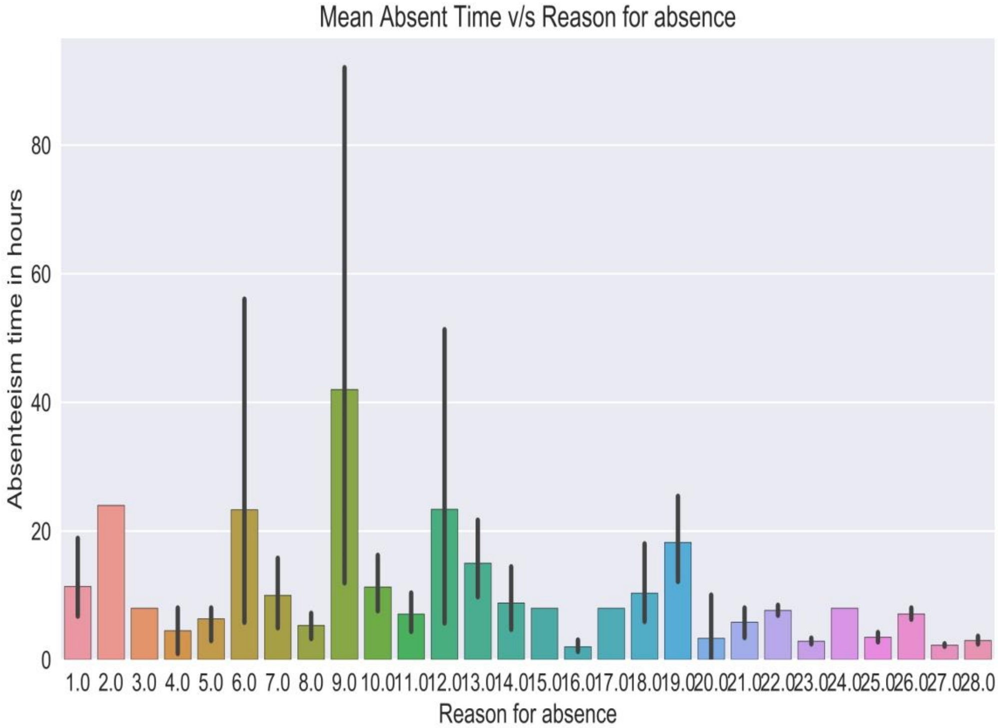


Figure 3.3: Barplot of Reason of Absence

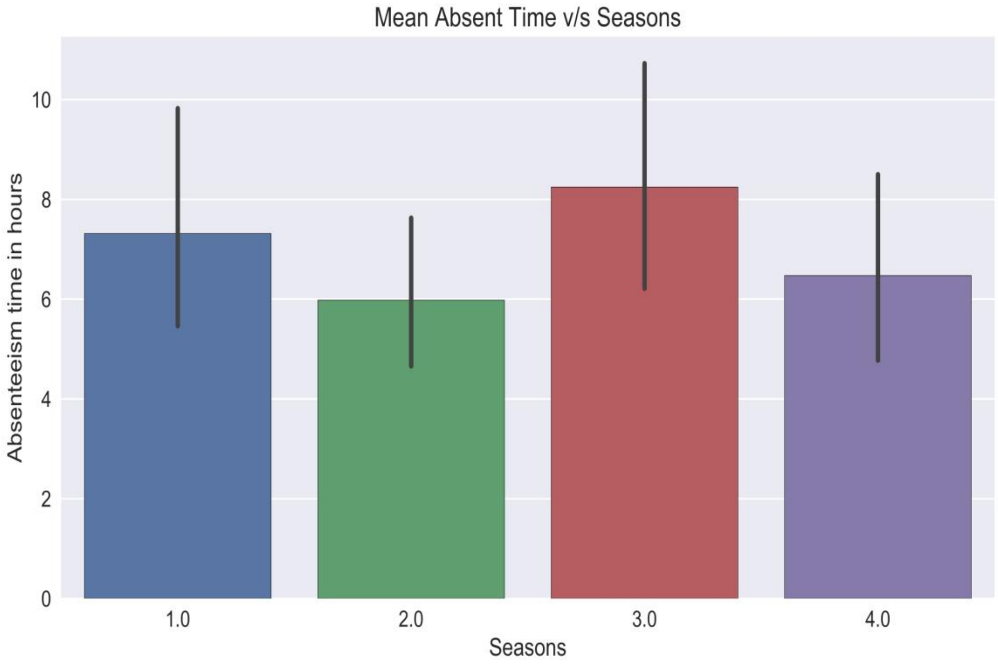


Figure 3.4: Barplot of Seasons

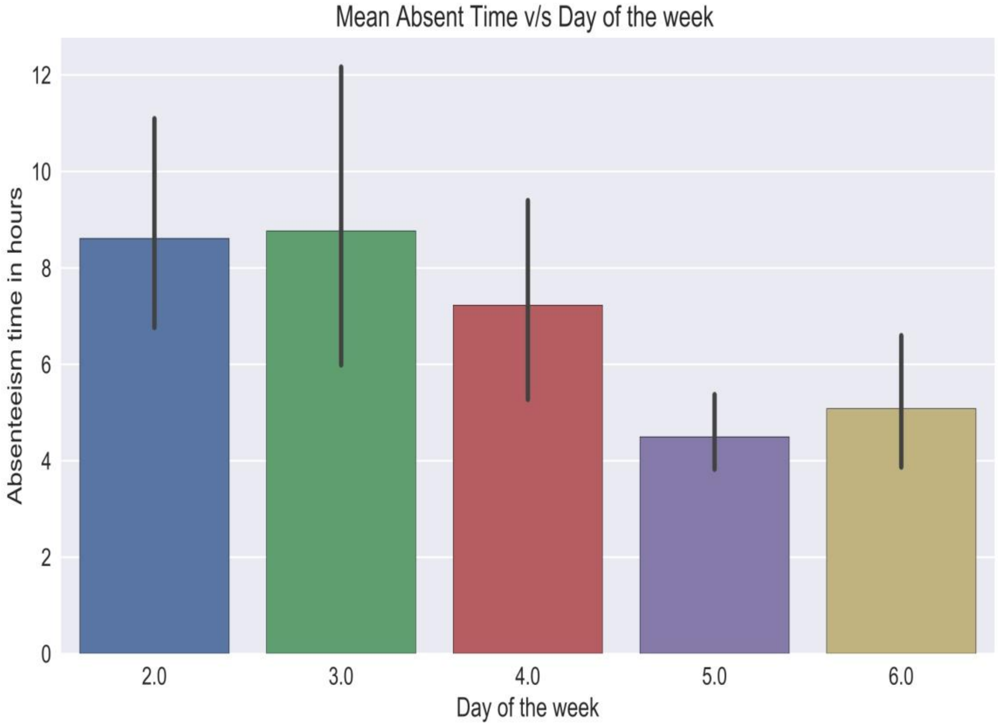


Figure 3.5: Barplot of Day of week

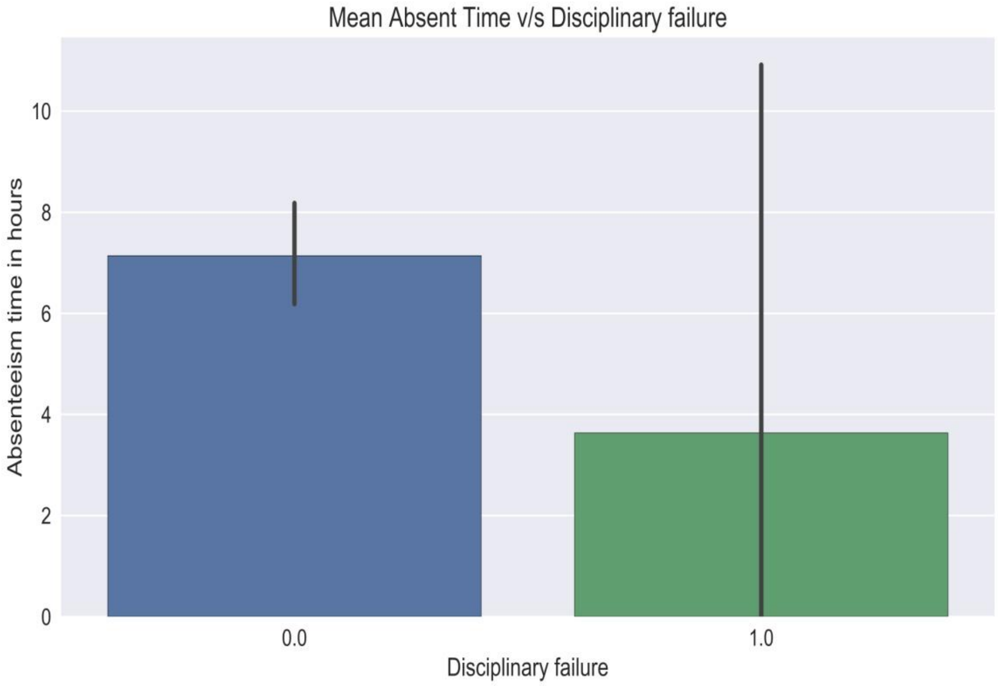


Figure 3.6: Barplot of Disciplinary Failure

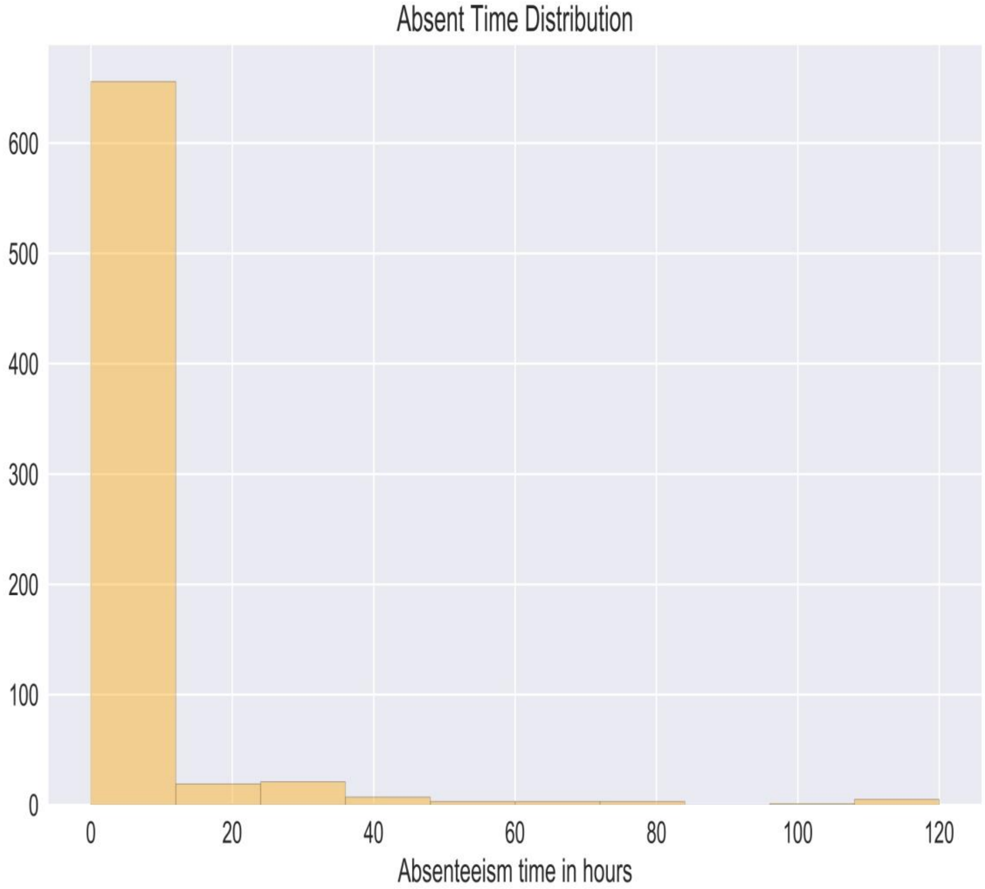


Figure 3.7: Target Variable Distribution

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 6. Springer.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer Science & Business Media.