**Project Name - Employee Absenteeism**

**-Abhishek**

**Index**

**1)Problem statement**

**2)EDA**

**3)Preprocessing**

**a)Missing Data Analysis**

**b)Outlier Analysis**

**c)Visual EDA**

**d)Feature Scaling**

**e)Feature Selection**

**4)PCA**

**5)Modeling**

**6)Conclusion**

**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?

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

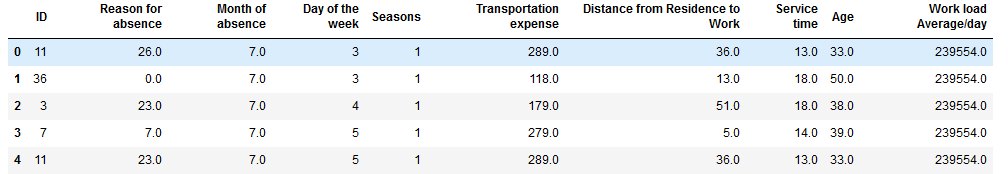
19. Height

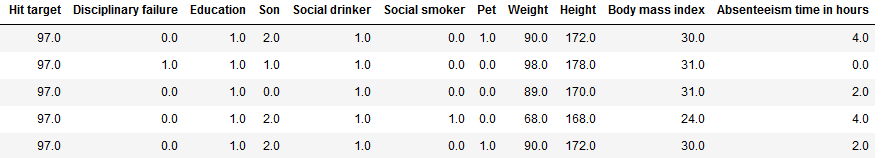
20. Body mass index

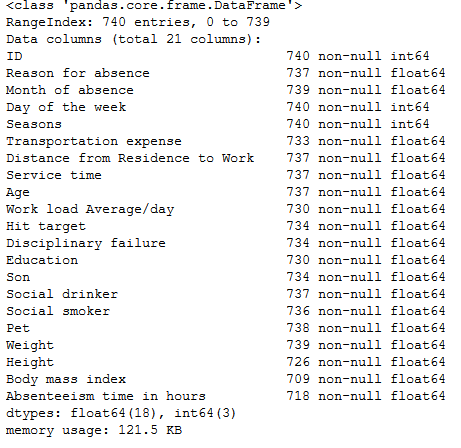
21. Absenteeism time in hours (target)

**EDA**

After importing the data, first lets look at the sample of data set that we have received and perform some basic Exploratory data analysis. EDA is a term for certain kinds of initial analysis and findings done with data sets, usually early on in an analytical process. Some experts describe it as “taking a peek” at the data to understand more about what it represents and how to apply it. Exploratory data analysis is often a precursor to other kinds of work with statistics and data.







From the Exploratory data Analysis, we can note that there are 21 columns and 740 rows in total, where ‘Absenteeism time in hours’ is our target variable. All other columns are our independent variables.

Further we can note that there are missing values in data, and the data types are of the type int64 or float64.

**Missing Value**

From EDA we have noted that there are some missing values in the data. Having missing data in our dataset may affect the conclusions that can be drawn from the dataset.

Missing values are a common occurrence, and we need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Below are common options to deal with missing values in dataset:

1)Drop the values

2)Fill the missing value with mean or median

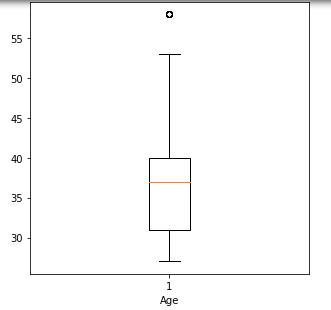
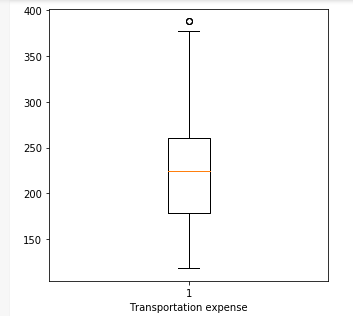
3)Fill the missing value with KNN imputation

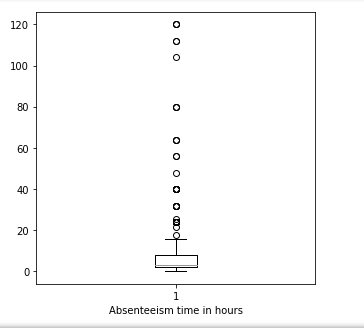
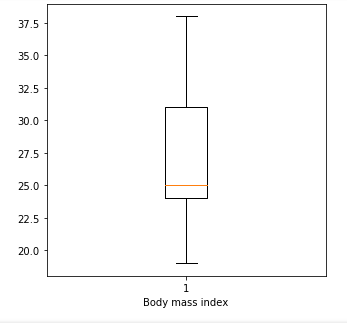
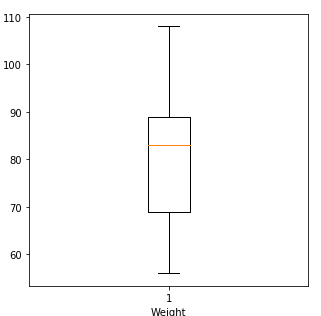
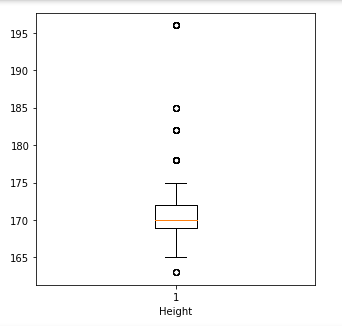
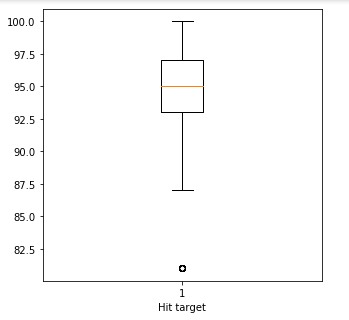
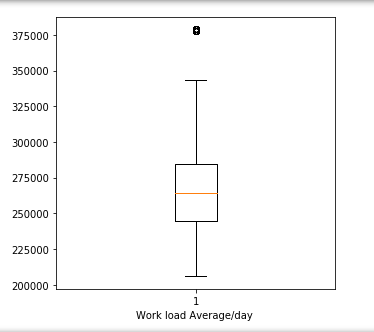
In this case KNN imputation seems to be a be better of the alternatives. This was concluded by randomly deleting a value in a column and trying out the methods mentioned above to note which method can give a result which is closest to the actual value.

**Outlier Analysis:**

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample.

We can make use of boxplots to detect outliers:





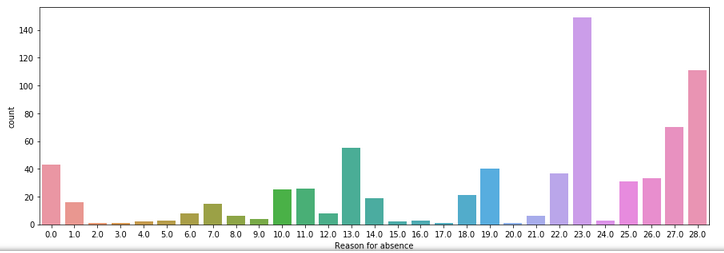
It can be noticed that there are outliers in all columns except Weight and Body mass index. We can also note that there are too many outliers in our target variables. Given that data set is daily data set with number of absent hours per day, the values seem redundant and we need to eliminate the outliers. In this case we have replaced the outliers as Nan values and we have imputed the nan values with KNN imputation.

**Visual EDA**

Perhaps one of the greatest disparities between those who live in the world of data science and those who don’t is the ability to translate the data so that everyone can understand it. One of the best ways for data scientists to present analysis of the data to those outside the industry is by generating visualizations. By visualizing data we can understand the data even better. Lets visualize and try to get a better sense of data:

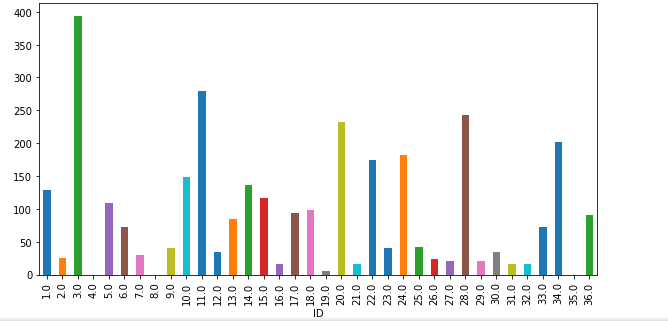
**Univariate analysis:**

Lets take a look at what was the most common reason used for absentee time in hours by looking at a count plot:

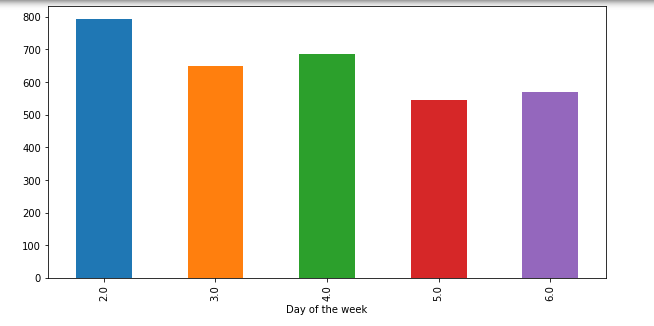


It looks like the most common reason was Medical consultation followed by dental consultation.

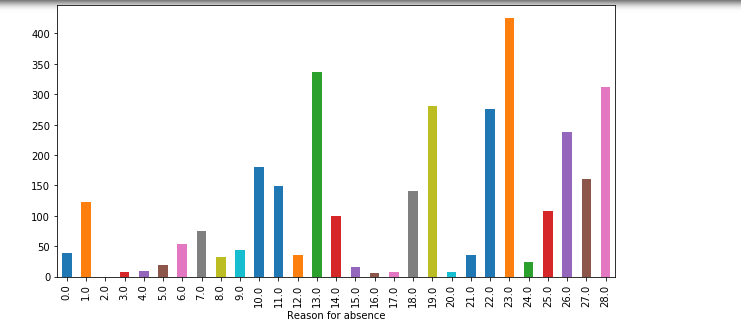
**Bivariate analysis, grouping by Absentee Time in Hours**



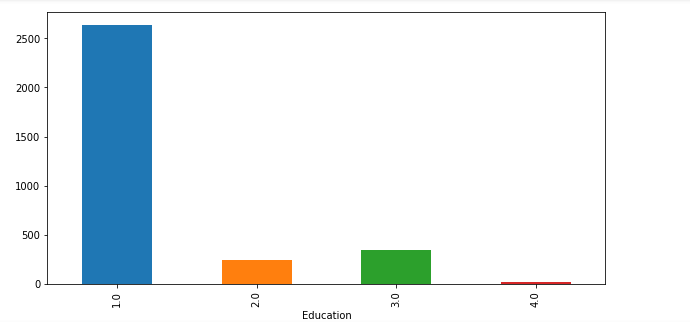
From the above visualization we can note that Employees with ID 3, 11, 20, 28 and 34 have absentee time in hours greater than 200.



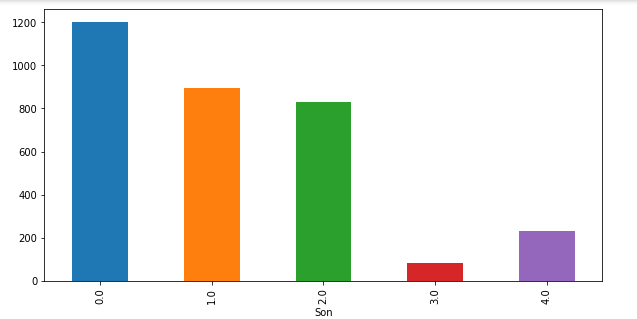
From the above graph we can conclude that absentee time in hours is the highest on Mondays.



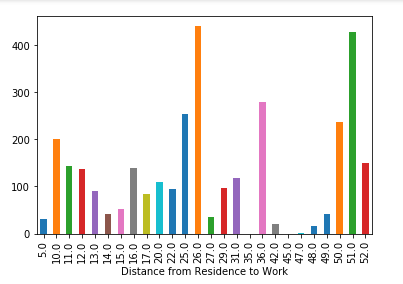
From the above graph we can see that Medical consultation was the commonly used reason to justify the absentee time in hours.



We can infer that lesser educated in the lot seems to having highest Absentee time in hours.



We can see that Employees with no children tend to be having highest Absentee time in hours.

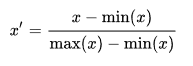


From the plot above we can infer that Employees travelling more than 50 kms are having high Absentee time in hours.

**Feature Scaling**

Feature scaling is a method used to standardize the range of independent variables or features of data. There are two types of Feature scaling:

Normalization: Also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula is given as:

x ′ = x − min ( x ) max ( x ) − min ( x ) {\displaystyle x'={\frac {x-{\text{min}}(x)}{{\text{max}}(x)-{\text{min}}(x)}}}

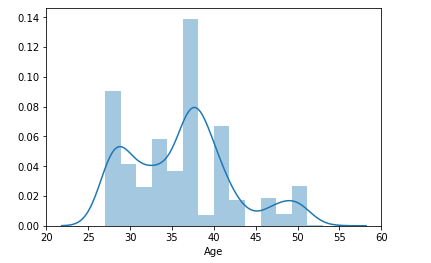
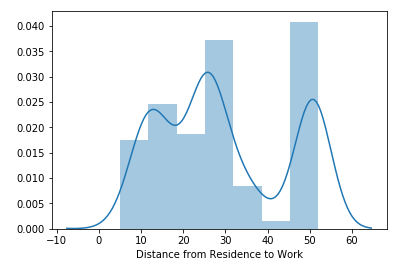
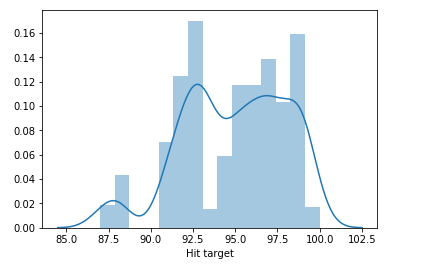
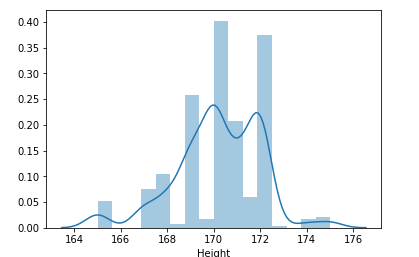
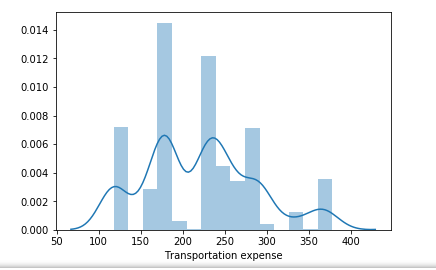
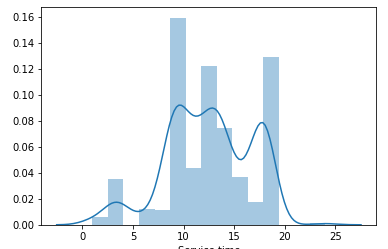
where x x {\displaystyle x} is an original value, x! x ′ {\displaystyle x'} is the normalized value.

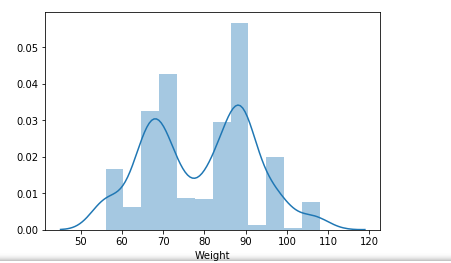
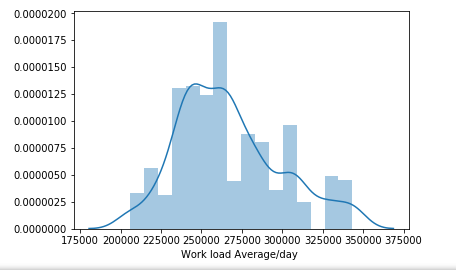
Standardization: We use this when data is normally distributed.

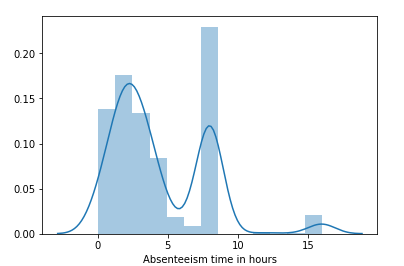
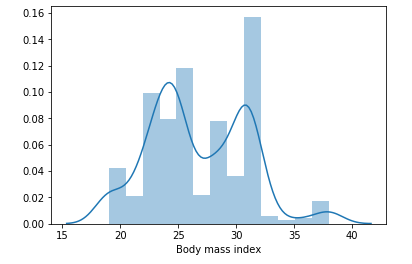


Where x {\displaystyle x} x is the original feature vector, x ¯ {\displaystyle {\bar {x}}} x-bar is the mean of that feature vector.

Now lets see how the data is distributed:







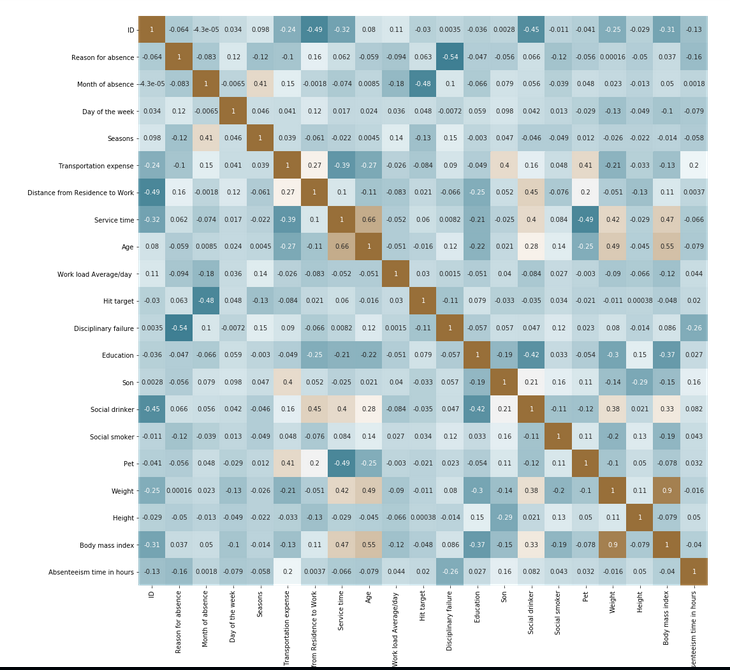
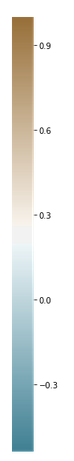
From the above plots, we can infer that data is not normally distributed. Hence the data was scaled with Normalization technique.

**Feature selection**

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

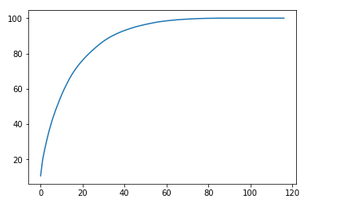
For this we will perform Correlation Analysis. Correlation tells you the association between two continuous variables. We can find if any two variables are highly correlated by viewing a heat map.

By looking at the heat map below it can be observed that Body Mass Index and Weight columns are highly correlated with each other. In this case as part of feature selection instead of keeping both the columns one column Weight has been dropped.

**PCA**

A more common way of speeding up a machine learning algorithm and extracting important variables is by using Principal Component Analysis (PCA). If your learning algorithm is too slow because the input dimension is too high, then using PCA to speed it up can be a reasonable choice. This is probably the most common application of PCA along with extracting important variables. Another common application of PCA is for data visualization.

****

We can observe that 50 variables are explains more than 95% of the data and hence only 50 variables are considered for modeling

**Model Selection**

Here three models have been taken for consideration, the output of which will be compared. Since this is a regression problem, the metric we consider is Root mean square error.

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. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

**1)Decision Tree**

A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node. Decision trees can handle both categorical and numerical data.

Results of our model are as below:

#Root Mean Squared Error = 0.04419942802838929(python)

#Root Mean Squared Error = 0.3629201(R)

#R^2 Score = 0.9536500781893115

# R^2 Score: 0.9838120

**2) Random Forest**

The **random forest** model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

*g*(*x*)=*f*0(*x*)+*f*1(*x*)+*f*2(*x*)+...

where the final model *g* is the sum of simple base models *fi*. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called [**model ensembling**](http://en.wikipedia.org/wiki/Ensemble_learning). In random forests, all the base models are constructed independently using a **different subsample** of the data.

Results of our model are as below:

#Root Mean Squared Error = 0.007627809966417199

#Root Mean Squared Error = 0.5633100

#R^2 Score = 0.9986195666691634

#R^2 Score = 0.9678819

**3)Linear Regression**

In statistics, linear regression is a linear approach to modelling the relationship between a dependent variable and one or more  independent variables. The case of one independent variable is called simple linear regression. For more than one independent variable, the process is called multiple linear regression. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula:

y = c + b\*x

where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Results of our model are as below:

#Root Mean Squared Error = 0.00947782685153262

#Root Mean Squared Error = 0.0012368689

#R^2 Score = 0.9978687556351598

#R^2 Score = 0.9999998113

**Conclusion**

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

1)As most common reasons of Absentee time in hours are Medical consultation followed by dental consultation, company should ask the employees to back their Medical reason with valid medical bills and prescriptions by doctor.

If this is a real concern company should provide employees with a medical plan of some sort.

2) Few employees have more absentee time in hours than other employees, the highest being the Employee with ID 3. Disciplinary action should be taken on such employees.

3) As lesser educated in the lot seems to having highest Absentee time in

hours, company should take actions to make the lesser educated

employees understand the importance of punctuality and the consequences of having more Absentee time in hours.

4) As Employees travelling more than 50 kms are having high Absentee time in hours, company should consider providing transportation facility.