Employee Absenteeism

Sai Bruhath

18 February 2019

****

**Table of Contents**

**1 Introduction**

1.1Background................................................................................................................................... 2 1.2ProblemStatement..........................................................................................................................2

**2 Exploring Data**

2.1 Data...............................................................................................................................................3 2.2 Summary of Each variable ...........................................................................................................4

2.3 Pre Processing……………………………………………………………………………………..5

2.3.1Missing Value Analysis..............................................................................................................5

2.3.2Missing Value treatment.............................................................................................................5

2.3.3 Outlier Analysis…….................................................................................................................6

2.3.4 Feature Selection……………………………………………………………………………….....................................9

2.3.5 Variable Importance……………………………………………………………………………………………………………….11

2.3.6 Feature scaling……………………………………………………………………………………………………………………….14

2.3.7 Distribution check…………………………………………………………………………………………………………………..15

**3 Sampling Techniques**......................................................................................................................18

**4 Modelling**

4.1 Random Forest ……………………………………………………………………………………19

4.2 Multiple Linear Regression……………………………………..................................................20

**5 Model Selection**……………………………………………………………………………………23

**6 Problems**...........................................................................................................................................24

**Appendix**

RandPythonCode.................................................................................................................................26

References……………………………………………………………………………………………..28

**1 Introduction**

**1.1 Background**

Absenteeism is one of the major threats to any industry. Absenteeism is the failure of employees to report for work when they are scheduled to work. In this implementation the focus is what changes has to be made by company to reduce employee absenteeism and on determining important factors that causes more absenteeism hours.

**1.2 Problem Statement**

The objective of this Case is to predict the reasons which causing high absenteeism hours, to overcome that what changes has to be made by the company. Provided dataset and its attributes are as follows :

* Individual identification (ID)
* Reason for absence (ICD).
* Month of absence
* Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
* Seasons (summer (1), autumn (2), winter (3), spring (4))
* Transportation expense
* Distance from Residence to Work (kilometers)
* Service time
* Age
* Work load Average/day
* Hit target
* Disciplinary failure (yes=1; no=0)
* Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
* Son (number of children)
* Social drinker (yes=1; no=0)
* Social smoker (yes=1; no=0)
* Pet (number of pet)
* Weight
* Height
* Body mass index

Target Variable : Absenteeism time in hours

Reason for absence has the following sub categories :

Absences attested by the International Code of Diseases (ICD) stratified into 21

categories (1 to 21) as follows:

1. Certain infectious and parasitic diseases
2. Neoplasms
3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4. Endocrine, nutritional and metabolic diseases
5. Mental and behavioural disorders
6. Diseases of the nervous system
7. Diseases of the eye and adnexa
8. Diseases of the ear and mastoid process
9. Diseases of the circulatory system
10. Diseases of the respiratory system
11. Diseases of the digestive system
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Diseases of the genitourinary system
15. Pregnancy, childbirth and the puerperium
16. Certain conditions originating in the perinatal period
17. Congenital malformations, deformations and chromosomal abnormalities
18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere
19. classified
20. Injury, poisoning and certain other consequences of external causes
21. External causes of morbidity and mortality
22. Factors influencing health status and contact with health services.

And 7 categories without (CID) patient follow-up ,

1. medical consultation
2. blood donation
3. laboratory examination
4. unjustified absence
5. physiotherapy
6. dental consultation

**2 Exploring Data**

**2.1 - Data size and structure**

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 740 obs. of 21 variables:

$ ID : num 11 36 3 7 11 3 10 20 14 1 ...

$ Reason for absence : num 26 0 23 7 23 23 22 23 19 22 ...

$ Month of absence : num 7 7 7 7 7 7 7 7 7 7 ...

$ Day of the week : num 3 3 4 5 5 6 6 6 2 2 ...

$ Seasons : num 1 1 1 1 1 1 1 1 1 1 ...

$ Transportation expense : num 289 118 179 279 289 17955 235 ...

$ Distance from Residence to Work: num 36 13 51 5 36 51 52 50 12 11 ...

$ Service time : num 13 18 18 14 13 18 3 11 14 14 ...

$ Age : num 33 50 38 39 33 38 28 36 34 37 ...

$ Work load Average/day : num 239554 239554 239554 239554 23955

$ Hit target : num 97 97 97 97 97 97 97 97 97 97 ...

$ Disciplinary failure : num 0 1 0 0 0 0 0 0 0 0 ...

$ Education : num 1 1 1 1 1 1 1 1 1 3 ...

$ Son : num 2 1 0 2 2 0 1 4 2 1 ...

$ Social drinker : num 1 1 1 1 1 1 1 1 1 0 ...

$ Social smoker : num 0 0 0 1 0 0 0 0 0 0 ...

$ Pet : num 1 0 0 0 1 0 4 0 0 1 ...

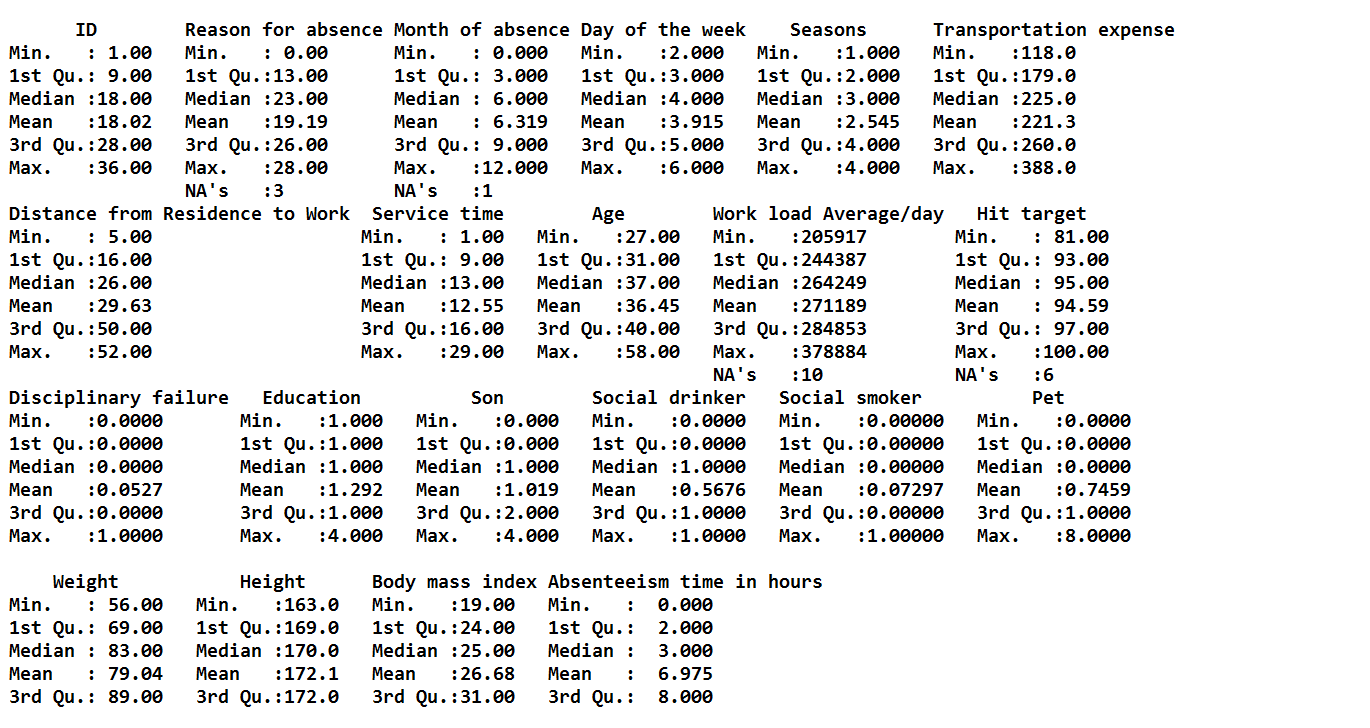
$ Weight : num 90 98 89 68 90 89 80 65 95 88 ...

$ Height : num 172 178 170 168 172 170 172 168 1

$ Body mass index : num 30 31 31 24 30 31 27 23 25 29 ... $ Absenteeism time in hours : num 4 0 2 4 2 NA 8 4 40 8 ...

**2.2 Summary of Each Variable :**

**[1] 740 observations 21 variables**



**2.3 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 this process analysing the count of missing values in each variable.

**2.3.1Missing Value Analysis :**

|  |  |
| --- | --- |
| **Missing values %** | **columns** |
| 4.189189189 | Body mass index |
| 2.972972973 | Absenteeism time in hours |
| 1.891891892 | Height |
| 1.351351351 | Work load Average/day |
| 1.351351351 | Education |
| 0.945945946 | Transportation expense |
| 0.810810811 | Hit target |
| 0.810810811 | Disciplinary failure |
| 0.810810811 | Son |
| 0.540540541 | Social smoker |
| 0.405405405 | Reason for absence |
| 0.405405405 | Distance from Residence to Work |
| 0.405405405 | Service time |
| 0.405405405 | Age |
| 0.405405405 | Social drinker |
| 0.27027027 | Pet |
| 0.135135135 | Month of absence |
| 0.135135135 | Weight |
| 0 | ID |
| 0 | Day of the week |
| 0 | Seasons |

**2.3.2 Missing values Treatment** : In the given data

Age,servicetime,Body mass index,Height,Education,Transportaion expense,Son,Social smoker,Social drinker,Distance from residence to work ,pet and weight is same for each unique Id.

Missing values for the above variables imputed manually by referring to the given values for the respective Ids and for the remaining variables missing values are imputed by mean method.

**2.3.3 Outlier Analysis :**

One of the other steps of pre-processingis the removal of outliers. Observations inconsistent with rest of global dataset are called as outliers.

Outlier can be caused due to several reasons :

* **Data Entry Errors:-** Human errors such as errors caused during data collection, recording, or entry can cause outliers in data. For example: Annual income of a customer is $100,000. Accidentally, the data entry operator puts an additional zero in the figure. Now the income becomes $1,000,000 which is 10 times higher. Evidently, this will be the outlier value when compared with rest of the population.
* **Measurement Error:**It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty. For example: There are 10 weighing machines. 9 of them are correct, 1 is faulty. Weight measured by people on the faulty machine will be higher / lower than the rest of people in the group. The weights measured on faulty machine can lead to outliers.
* **Experimental Error:** Another cause of outliers is experimental error. For example: In a 100m sprint of 7 runners, one runner missed out on concentrating on the ‘Go’ call which caused him to start late. Hence, this caused the runner’s run time to be more than other runners. His total run time can be an outlier.

Impact to the model due to presence of outliers :

Outliers can drastically change the results of the data analysis and statistical modeling. It increases the error variance and reduces the power of statistical tests

* If the outliers are non-randomly distributed, they can decrease normality
* They can bias or influence estimates that may be of substantive interest
* They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

Outliers can be detected by using various methods :

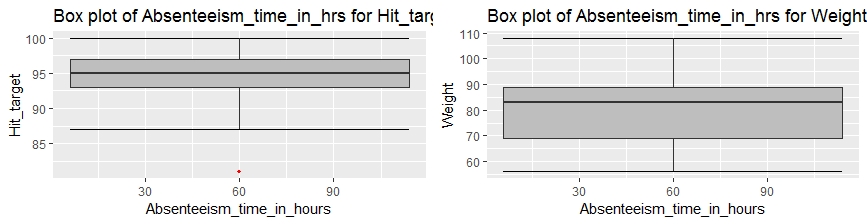
* Graphical too (Box Plot)
* Stastical Technique(Grabb’s test for outliers)
* R- package outlier
* Replace with NA

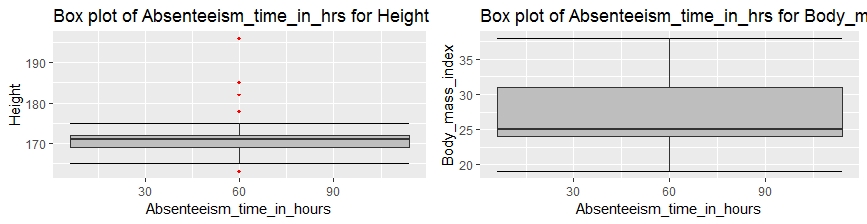
In this case we use a classic approach of removing outliers, Tukey’*s method*.

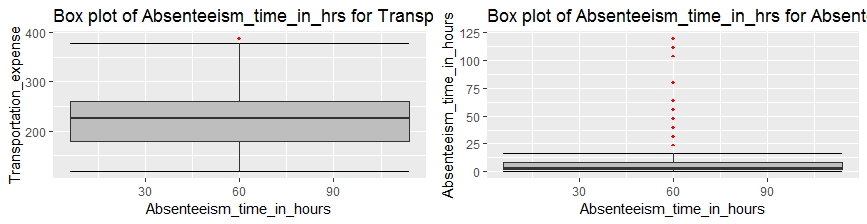
We visualize the outliers using boxplot*s* and remove the outliers.

In box plot values which are falling beyond the upper and lower fence is called as outliers.

Similar way outliers has been calculated for each variable.







Outlier length for each variable :

[1] "Transportation\_expense"

[1] 3

[1] "Distance\_from\_Residence\_to\_Work"

[1] 0

[1] "Service\_time"

[1] 5

[1] "Age"

[1] 8

[1] "Work\_load\_Average\_day"

[1] 29

[1] "Hit\_target"

[1] 19

[1] "Weight"

[1] 0

[1] "Height"

[1] 106

[1] "Body\_mass\_index"

[1] 0

[1] "Absenteeism\_time\_in\_hours"

[1] 25

**2.3.4 Feature Selection**

Feature selection is a subset of relevant features (Variables,Predictors) for use in model construction.

It is also called as variable selection or attribute selection.

Feature engineering is the science (and art) of extracting more information from existing data. You are not adding any new data here, but you are actually making the data you already have more useful

Feature selection can be done in two ways ::

* Correlation Analysis
* Chi-Square test
* Analysis of variance (Anova)

**Correlation Analysis** : Correlation can be derived using following formula:

**Correlation = Covariance(X,Y) / SQRT( Var(X)\* Var(Y))**

Various tools have function or functionality to identify correlation between variables. In Excel, function CORREL() is used to return the correlation between two variables and SAS uses procedure PROC CORR to identify the correlation.

It is applicable only for Numerical Variables.

Assumptions :

1. There should be no (or) less correlation exist between two independent variables.
2. High correlation should exist between independent and dependent variable.

Correlation value ranges from -1 to +1.

**Chi-Square test :** This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

It is applicable for categorical variables.

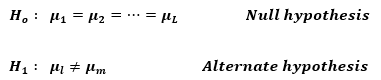
Assuumptions :

Null Hypothesis : Two variables are independent

Alternative hypothesis : Two variables are not independent.

**Analysis of variance** : ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

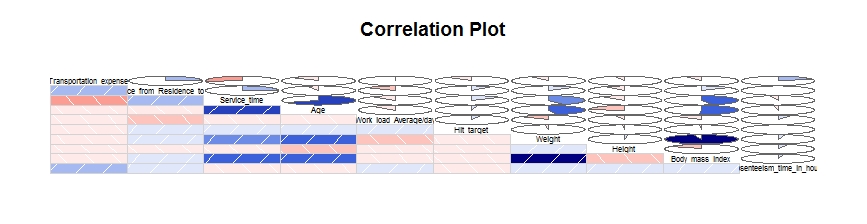
ANOVA also uses a Null hypothesis and an Alternate hypothesis. The Null hypothesis in ANOVA is valid when all the sample means are equal, or they don’t have any significant difference. Thus, they can be considered as a part of a larger set of the population. On the other hand, the alternate hypothesis is valid when at least one of the sample means is different from the rest of the sample means. In mathematical form, they can be represented as:



where https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/12/image0111.pngbelong to any two sample means out of all the samples considered for the test. In other words, the null hypothesis states that all the sample means are equal or the factor did not have any significant effect on the results. Whereas, the alternate hypothesis states that at least one of the sample means is different from another. But we still can’t tell which one specifically.

In this project performed correlation and ANOVA , Correlation is applied for numerical variables and Anova is for Categorical variables as Anova is used to compare one numerical and group of categorical variables.(Target variables is continuous) compared each categorical variable with Target variable.

**2.3.5 Variable Importance:**

Insights from Correlation Plot:

1. Weight and Body mass index are highly positive correlated.

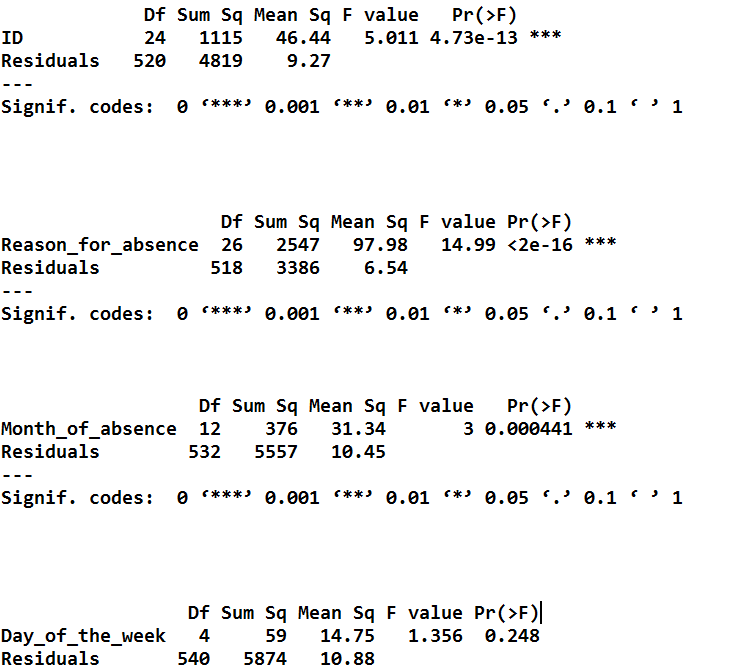
2. Age and Service time are highly positive correlated.

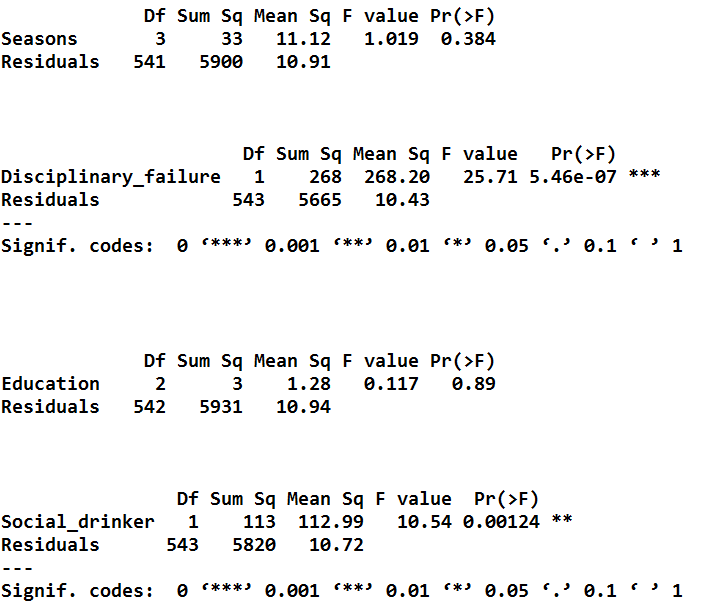
3. Age and Weight are positive correlated.

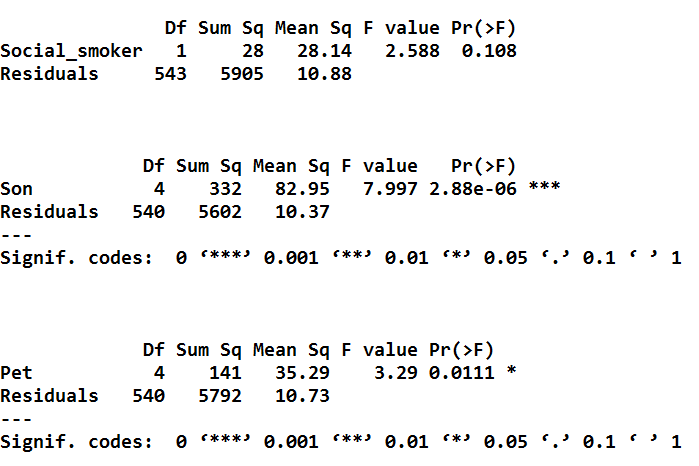
4. Body mass index and Age are positive correlated

5. Transportaion expense is negatively correlated with Service time, Service time and weight are correlated.

**ANOVA – Categorical Variables:**







Based on the Correlation and Anova test results removed variables which are not significant.

After clean up of insignificant variables from the datafeature scaling is applied for the continuous variables.

**2.3.6 Feature Scaling :**

When we want to change the scale of a variable or standardize the values of a variable for better understanding. While this transformation is a must if you have data in different scales, this transformation does not change the shape of the variable distribution.

Feature scaling is applicable only for Continuous variables.

Feature scaling has two methods to transform the data :

* Standardization
* Normalization

**Standardization /Z –score** :

Basically it converts each data point unique of standard deviation.

Z=

Z -represents the difference between raw score and population mean in the units of standard deviation.

If the data is normally distributed (or) uniformly distributed then standardization is used.

**Normalization :**

Normalization to bring all the variables into proportion with one another variable.

New value =(Value-Min value)/ (max value – minvalue).

Range is from 0 to 1.

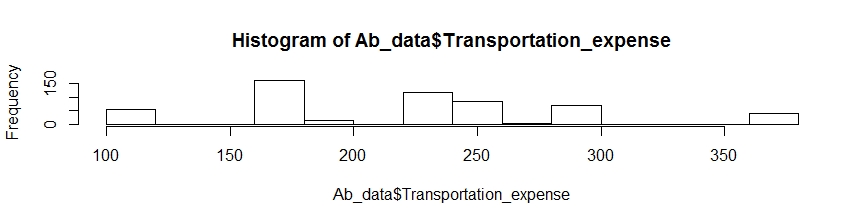
If the data is left skewed or right skewed other than unformally distrubuted then Nrmalization is used.

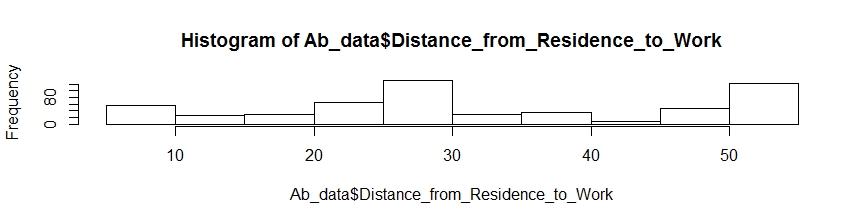
As our data is normally distributed standardization technique is used for most of the variables.

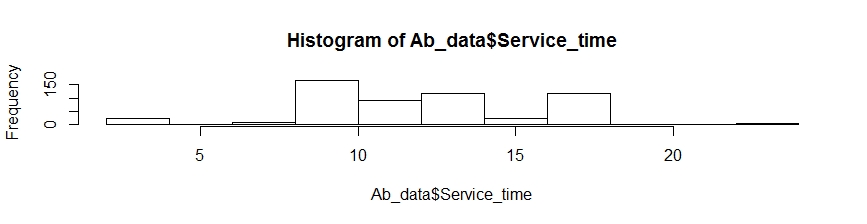
**2.3.7 Density Plot: Distribution of data points**

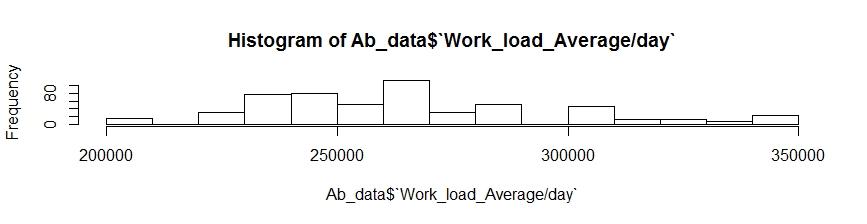
Checking the distribution of values for each variable. All the variable seems to be not normally distributed Hence Normaliation technique is used to change the scale of variables.

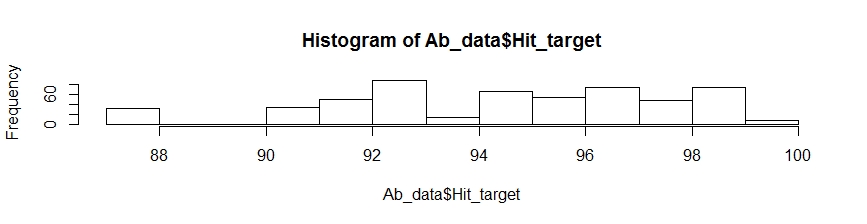
The below plots shows the variables pattern in the dataset.

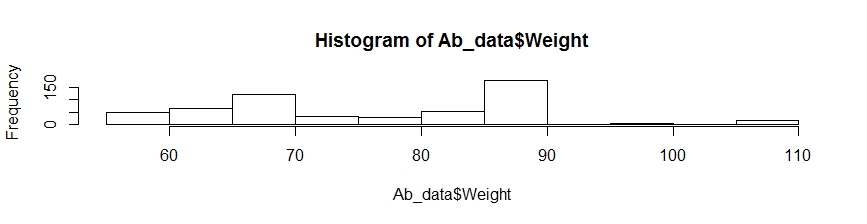








****

****

Now we have completed all the pre-processing techniques that has to be applied on the data.

Data available before pre-processing : 740 observations 21 variables

Data available after pre-processing : 545 observations 15 variables

**3.Sampling Techniques :**

Selection of Subset from the whole population is called sampling.

Sampling can be categorized into two categories

* Probability samples : Probability sampling is a sampling technique, in which the subjects of the population get an equal opportunity to be selected as a representative sample.

It is selected Randomly.

* Non – Probability samples: Nonprobability sampling is a method of sampling wherein, it is not known that which individual from the population will be selected as a sample.

It is selected Arbitarly.

Different types of Probability sampling are as follows :

1. Simple Random Sampling.
2. Systematic Radom Sampling.
3. Stratified Sampling.
4. Multi-stage cluster Sampling.

1. **Simple Random Sampling** : Random Sampling is purest form of probability sampling.

Selected by using chance or random numbers.

Each individual subject has an equal chance of being selected.

EX: Random Numbers, Drawing names from hat.

1. **Systematic Sampling** : It is also called kth name selection technique.

After required sample size has been calculated every kth record is selected from list of population numbers.

K=N/n ,where N= no of observations , n= Desired shape.

EX: If K=3

P1,P2,P3,P4,P5,P6,P7,P8,P9

P3,P6,P9 is selected.

This process is selected if the data desn’t contain an hidden order.

1. **Stratfied Sampling** : In stratified sampling we have to select categorical variables on equal propotions.

In this project simple random sampling is used.

After sampling distribution of the data are as follows :

Test : 109 Observations ,15Variables

Train : 436 Observations,15 Variables.

**4.Modeling :**

To make predictions, I have used Random Forest and Multiple linear regression,

* 1. **Random Forest :** Random forest is popular Regression method. random forest is better at fitting non-linear data. It can also work well even if there are correlated features, which can be a problem for interpreting Linear regression.

**Model summary :**

print(RF\_model)

Call:

randomForest(formula = Absenteeism\_time\_in\_hours ~ ., data = train, importance = TRUE, ntree = 500)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 4

Mean of squared residuals: 9.02763321

% Var explained: 88.79

rmse <- function(error)

{

sqrt(mean(error^2))

}

error <- test[,15] - RF\_predictions

> rmse(data.matrix(error))

[1] 0.1628441

> regr.eval(test[,15],RF\_predictions,stats = c('mae','rmse','mse'))

mae rmse mse

0.2869754 0.1600142 0.00253

**4.2 Linear Regression** :

Call:

lm(formula = Absenteeism\_time\_in\_hours ~ ., data = train)

Residuals:

Min 1Q Median 3Q Max

-0.36521 -0.08287 -0.00412 0.05789 0.77995

Coefficients: (15 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.2451295 0.1003767 2.442 0.015063 \*

ID3 0.0090824 0.0431854 0.210 0.833539

ID4 -0.2116155 0.1757671 -1.204 0.229367

ID5 0.0936561 0.0573403 1.633 0.103237

ID6 0.2163950 0.0699405 3.094 0.002123 \*\*

ID7 0.0005965 0.1199344 0.005 0.996034

ID10 0.0058049 0.0527652 0.110 0.912457

ID11 0.0803619 0.0480011 1.674 0.094931 .

ID13 0.0642022 0.0637132 1.008 0.314261

ID15 -0.0617789 0.0497269 -1.242 0.214879

ID16 0.2096078 0.1174824 1.784 0.075206 .

ID17 -0.0342318 0.0594652 -0.576 0.565189

ID19 -0.0841052 0.1162360 -0.724 0.469779

ID20 0.0780269 0.0474826 1.643 0.101164

ID22 -0.0217449 0.0468889 -0.464 0.643093

ID23 0.0205577 0.0795214 0.259 0.796149

ID24 0.0141717 0.0536884 0.264 0.791955

ID26 -0.0210946 0.0920300 -0.229 0.818827

ID27 0.0517914 0.0814279 0.636 0.525139

ID28 -0.0475602 0.0428701 -1.109 0.267967

ID29 0.0233604 0.1620661 0.144 0.885466

ID33 -0.0860157 0.0524602 -1.640 0.101919

ID34 -0.0029349 0.0458319 -0.064 0.948975

Reason\_for\_absence1 0.2373089 0.0987893 2.402 0.016784 \*

Reason\_for\_absence3 0.3962067 0.1740769 2.276 0.023407 \*

Reason\_for\_absence5 0.2904652 0.1751584 1.658 0.098093 .

Reason\_for\_absence6 0.1572239 0.1230150 1.278 0.202009

Reason\_for\_absence7 0.1402555 0.1017105 1.379 0.168727

Reason\_for\_absence8 0.1333666 0.1077954 1.237 0.216780

Reason\_for\_absence9 0.2743079 0.1737030 1.579 0.115139

Reason\_for\_absence10 0.1684634 0.0934017 1.804 0.072090 .

Reason\_for\_absence11 0.0740511 0.0932899 0.794 0.427829

Reason\_for\_absence12 0.4698012 0.1347252 3.487 0.000546 \*\*\*

Reason\_for\_absence13 0.1047790 0.0847895 1.236 0.217323

Reason\_for\_absence14 0.0781184 0.0973398 0.803 0.422753

Reason\_for\_absence15 0.2485974 0.1401901 1.773 0.076993 .

Reason\_for\_absence16 -0.1339623 0.1440305 -0.930 0.352919

Reason\_for\_absence18 0.1562354 0.0927738 1.684 0.093005 .

Reason\_for\_absence19 0.2380922 0.0880698 2.703 0.007175 \*\*

Reason\_for\_absence21 0.1189796 0.1085718 1.096 0.273843

Reason\_for\_absence22 0.2110229 0.0867156 2.434 0.015420 \*

Reason\_for\_absence23 -0.0237089 0.0816471 -0.290 0.771684

Reason\_for\_absence24 0.1910111 0.1750087 1.091 0.275781

Reason\_for\_absence25 0.0350605 0.0895248 0.392 0.695555

Reason\_for\_absence26 0.2068692 0.0872514 2.371 0.018247 \*

Reason\_for\_absence27 -0.0800260 0.0823298 -0.972 0.331668

Reason\_for\_absence28 -0.0509769 0.0817947 -0.623 0.533513

Month\_of\_absence1 -0.0300228 0.0444164 -0.676 0.499496

Month\_of\_absence2 0.0196419 0.0393278 0.499 0.617763

Month\_of\_absence3 0.0900169 0.0352870 2.551 0.011138 \*

Month\_of\_absence4 0.0644206 0.0395473 1.629 0.104163

Month\_of\_absence5 0.0522187 0.0435120 1.200 0.230858

Month\_of\_absence6 -0.0045758 0.0428036 -0.107 0.914924

Month\_of\_absence7 0.0189944 0.0375233 0.506 0.613011

Month\_of\_absence8 -0.0142095 0.0409919 -0.347 0.729055

Month\_of\_absence9 -0.0229723 0.0474620 -0.484 0.628658

Month\_of\_absence10 0.0106455 0.0429439 0.248 0.804352

Month\_of\_absence11 0.0049102 0.0364393 0.135 0.892882

Month\_of\_absence12 NA NA NA NA

Transportation\_expense NA NA NA NA

Distance\_from\_Residence\_to\_Work NA NA NA NA

Age NA NA NA NA

Work\_load\_Average\_day 0.0375235 0.0432867 0.867 0.386573

Hit\_target -0.0834537 0.0524353 -1.592 0.112327

Disciplinary\_failure1 -0.2231356 0.0894106 -2.496 0.013002 \*

Son1 NA NA NA NA

Son2 NA NA NA NA

Son3 NA NA NA NA

Son4 NA NA NA NA

Social\_drinker1 NA NA NA NA

Pet1 NA NA NA NA

Pet2 NA NA NA NA

Pet4 NA NA NA NA

Pet8 NA NA NA NA

Weight NA NA NA NA

Height NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1499 on 375 degrees of freedom

Multiple R-squared: 0.7806, Adjusted R-squared: 0.6555

F-statistic: 7.064 on 60 and 375 DF, p-value: < 2.2e-16

**5.Model Selection :**

By comparing both the models we could see the accuracy for Random Forest model is higher compared to Linar regression Model.

Conclusion :

Forecasting for the future data should be done with random forest model as this model gives higher accuracy.

.

**6 . Problems :**

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

Ans : On Analysing the data Nearly 53% of Absenteeism hours are from the below Id’s:

**ID : 11,3,14,28,34,36,20**

Further analysing reasons for the absence of these Id’s

We could see the majority of reasons are due to health issues.

Company should take Proper action on the above mentioned Id’s.

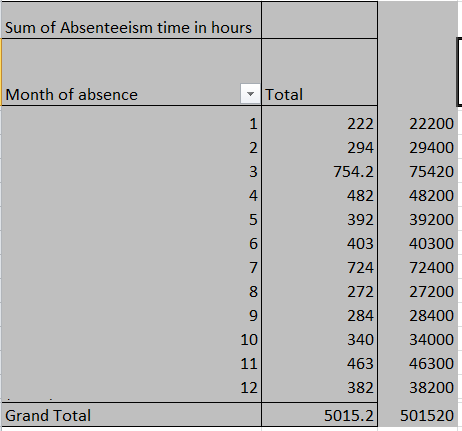
Conducting blood donation camp once in a month in the office can reduce the cause of this type of absenteeism and increase in working hours.

**2.** How much losses every month can we project in 2011 if same trend of

absenteeism continues?

Ans: As the same trend of absenteeism continues in 2011 so calculating the loss with the present provided data.

Assuming company will have a loss of 100 rs for each one our of absenteeism and calculating the loss as follows :



Explanation : In the above snippet first column represents Month .Second column represents Total no of Absenteeism hours in each month.and the Third column represents loss for each month.

Loss for each month = Total no of Absenteeism hours \*100

Ex : For January loss = 222\*100 =22200

High amount of loss occurred in March as total no of Absenteeism hrs is high =Rs75,420

Low amount of loss occurred in January =RS22,200

Total loss occurs if the same trend of absenteeism continues in 2011 =RS 5,01,520

**Appendix:**

R and Python code can be found in the submission folder.

**References :**

Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.

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