## Edwisor Project Report

**Absenteeism @Work**

****

**Submitted by**

**-Anchal Sharma**

**Contents**

**1 Introduction ……………………………………………………12**

* 1.1 Problem Statement……………………………….….………. 12
* 1.2 Data………………………………………………..………… 15

**2 Methodology**

2.1  PreProcessing ………………………………………….………16

* 2.1.1 Histogram of Data ………………….……………………. 26
* 2.2 Missing Value Analysis……………………………….……….27
* 2.3 Feature Selection………………………………………………28
* 2.4 Feature Scaling………………………….……………….…….29

**3 Modelling ………………………………….…………………..30**

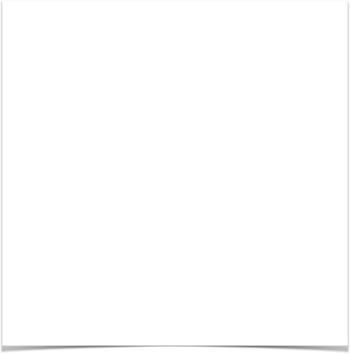
**Appendix………………………………………………………………….33**

**References**………………………………………………………………..34

Absenteeism at Workplace!

*How to reduce absenteeism at workplace?*

Absenteeism is an employee's intentional or habitual absence from work. While employers expect workers to miss a certain number of workdays each year, excessive absences can equate to decreased productivity and can have a major effect on company finances, morale and other factors. This article looks at the causes of absenteeism, the costs of lost productivity and what employers can do to reduce absenteeism rates in the workplace.



Absenteeism is an employee's intentional or habitual absence from work. While employers expect workers to miss a certain number of workdays each year, excessive absences can equate to decreased productivity and can have a major effect on company finances, morale and other factors.

**Causes of Absenteeism**

People miss work for a variety of reasons, many of which are legitimate and others less so. Some of the common causes of absenteeism include (but are not limited to):

* **Bullying and harassment** - Employees who are bullied or harassed by coworkers and/or bosses are more likely to call in sick to avoid the situation
* **Burnout, stress and low morale** - Heavy workloads, stressful meetings/presentations and feelings of being unappreciated can cause employees to avoid going into work. Personal stress (outside of work) can lead to absenteeism.
* **Childcare and eldercare** - Employees may be forced to miss work in order to stay home and take care of a child/elder when normal arrangements have fallen through (for example, a sick caregiver or a snow day at school) or if a child/elder is sick.
* **Depression** - According to the National Institute of Mental Health, the leading cause of absenteeism in the United States is depression. Depression can lead to substance abuse if people turn to drugs or alcohol to self-medicate their pain or anxiety.
* **Disengagement** - Employees who are not committed to their jobs, coworkers and/or the company are more likely to miss work simply because they have no motivation to go.
* **Illness** - Injuries, illness and medical appointments are the most commonly reported reasons for missing work (though not always the actual reason). Not surprisingly, each year during the cold and flu season, there is a dramatic spike in absenteeism rates for both full-time and part-time employees.
* **Injuries** - Accidents can occur on the job or outside of work, resulting in absences. In addition to acute injuries, chronic injuries such as back and neck problems are a common cause of absenteeism.
* **Job hunting** - Employees may call in sick to attend a job interview, visit with a headhunter or work on their resume/CVs
* **Partial shifts** - Arriving late, leaving early and taking longer breaks than allowed are considered forms of absenteeism and can affect productivity and workplace morale.

**Costs of Lost Productivity**

The Gallup-Healthways Well-Being Index surveyed 94,000 workers across 14 major occupations in the U.S. Of the 77% of workers who fit the survey's definition of having a chronic health condition (asthma, cancer, depression, diabetes, heart attack, high blood pressure, high cholesterol or obesity), the total annual costs related to lost productivity totaled $84 billion. According to the survey, the annual costs associated with absenteeism vary by industry, with the greatest loss occurring in professional occupations (excluding nurses, physicians and teachers); the 14 occupations and corresponding costs of lost productivity are shown in Figure 1.

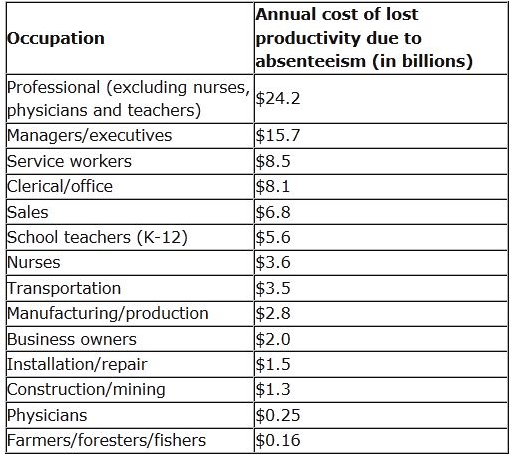


Figure 1: Annual cost of lost productivity by major U.S. occupations

According to Absenteeism: The Bottom-Line Killer, a publication of workforce solution company Circadian, unscheduled absenteeism costs roughly $3,600 per year for each hourly worker and $2,650 each year for salaried employees. The costs can be attributed to many factors including:

* Wages paid to absent employees
* High-cost replacement workers (overtime pay for other employees and/or temporary workers)
* Administrative costs of managing absenteeism

Other indirect costs and effects of absenteeism include:

* Poor quality of goods/services resulting from overtime fatigue or understaffing
* Reduced productivity
* Excess manager time (dealing with discipline and finding suitable employee replacements)
* Safety issues (inadequately trained employees filling in for others, rushing to catch up after arriving as a replacement, etc)
* Poor morale among employees who have to "fill in" or do extra work to cover absent coworkers

**What Employers Can Do**

Absenteeism is an especially difficult problem to tackle, because there are both legitimate and poor excuses for missing work - and it can be challenging for employers to effectively monitor, control and reduce absenteeism. Unless a company requires a written excuse from a doctor, for example, it can be difficult to determine if an employee is actually sick when missing work. At the same time, however, it is important for employers to consider the added costs associated with a sick employee who spreads an illness that gets the whole division - or a lot of customers - sick.

To address problems like this, some companies, cities and states have moved toward a mandatory paid sick leave policy, where each employee receives a specified number of days each year to use when sick.

Opponents of mandatory sick leave argue that it will ultimately cost businesses more money and lead to increased layoffs. In addition, opponents have concerns that employees will use all their sick days whether or not they need them. Advocates of such a move, however, argue that paid sick leave makes economic sense because it will help stop the spread of communicable diseases in the workplace and in schools - resulting in fewer instances of absenteeism in the long run - and that sick employees will be able to recover sooner.

The Centers for Disease Control, for example, states that paid sick leave could have an especially significant impact in the food service industry, where it estimate that sick food handlers are responsible for 53% of norovirus (a particularly nasty form of stomach virus) outbreaks. - One sick food handler could theoretically infect dozens or even hundreds of people, resulting in a large number of absences that could have been avoided if that employee had simply stayed home. Unfortunately, workers often either need the money or are worried about being terminated for calling in sick - even if it's unpaid leave - so they go to work even if they know they are contagious.

In an effort to reduce absenteeism, some companies offer incentives for going to work, such as earned time off or lotteries for workers who do not have any unexcused absences within a certain period. Other firms might try a more proactive approach, putting policies in place to focus on responses to employee health concerns, including:

* Physical health
* Psychological health
* Work-home balance
* Environmental health
* Economic health

The logic with this approach is that healthier, happier employees will be more able and motivated to go to work each day, resulting in increased productivity and higher morale for the individual workers as well as the entire team. Although these employee wellness strategies may be expensive to implement and maintain, they can have a net positive effect on a company's bottom line - and that's good for business.

**The Bottom Line**

Absenteeism costs companies billions of dollars each year in lost productivity, wages, poor quality of goods/services and excess management time. In addition, the employees who do show up to work are often burdened with extra duties and responsibilities to fill in for absent employees, which can lead to feelings of frustration and a decline in morale.

Occasional absences from work are inevitable - people get sick or injured, have to take care of others, or need time during business hours to handle personal business. It is the habitual absences that are most challenging to employers, and that can have the greatest negative effect on coworkers. Because missed work days have a profound financial effect on a company's bottom line, it is beneficial for most businesses to implement strategies to equitably monitor, reduce and respond to absenteeism.

**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?**   
   **The given predictors are as follows:**

**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

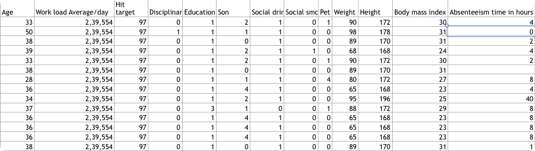
20. Body mass index  
21. Absenteeism time in hours (target)

**1.2 Data**

Given below is the sample of the Absenteeism at work excel file. There are 20 dependent variables and 1 independent variable for a total of 21 variables. The excel sheet comprises of 740 observations.

Table 1.1: Columns 1-8



Table 1.2 Columns 9-21

**Chapter 2**

**Methodology**

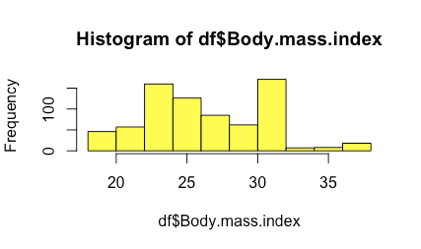
**2.1 Pre Processing**

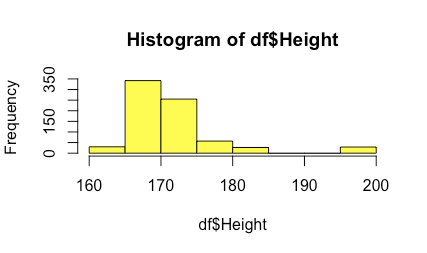
We require Pre-Processing as the data that we generally receive from the customer is noisy and not normalised. There could be a lot many missing values in the data, and accordingly we need to modify our dataset so that it could be fed into the required model , which in turn would increase the efficiency of the model in the long run, if it were to be deployed in the future.

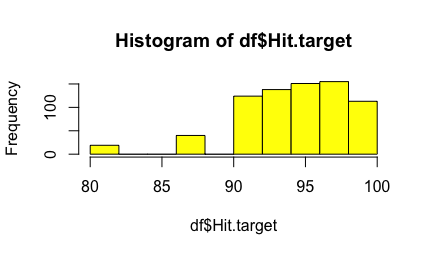
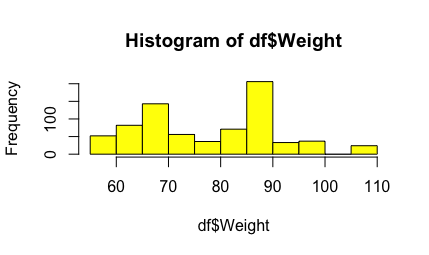
There are multiple steps that we need to follow while performing pre-processing such as:

* Data Cleaning
  + Filling in missing Value
  + Identifying outliers
  + Data transformation: Normalisation and standardisation
  + Data reduction

But before we dive in the above mentioned pre-processing techniques it is always better to visualise our data and get a rough idea of what we are dealing with. This process is also called Exploratory Data Analysis and is usually the first step in Data mining.

**Exploratory Data Analysis:**

****



# 

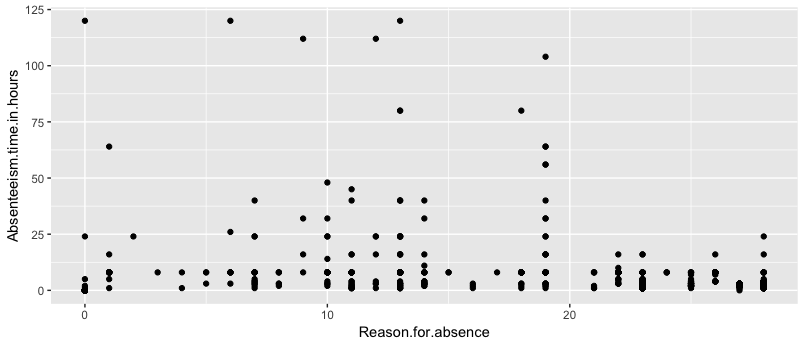
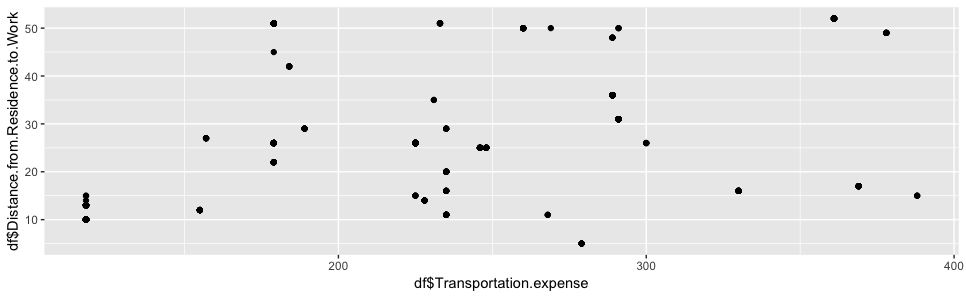
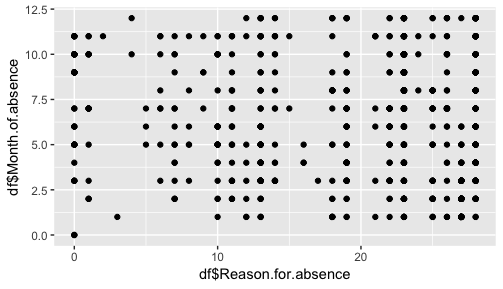
# 

# 

# 

# 

# 



The graphs above provide huge insights on the data and its distribution. For instance, we could see from the last scatterplot that the Reason for absence in the range 10-20 holds much of the distribution in terms of absenteeism hours, with reason 19 being the most scattered one. However, when we look at the histogram of Reason for absence, we could see that the reasons in the range of 20-25 dominates the dataset in terms of frequency when related with other reasons for absence. Hence, this gives us an insight that people are taking non ICD classifies leaves frequently.

This could be due to the possibility that non ICD types are unavailable at the office premise and doesn’t require doctor’s prescription. This opens a door of possibility that employees might be just faking the disease, so as to take leave.

Employers could subsequently try and introduce diagnosis of non ICD types in the office premise so that the employees affected by it could get help in the office itself and thereby being present at the workplace.

From the first scatterplot of reason for absence versus the month of absence, we can clearly see that there is no such correlation between the two and that one is not dependent on another in some way as the data is non uniformly distributed and doesn’t give much insight.

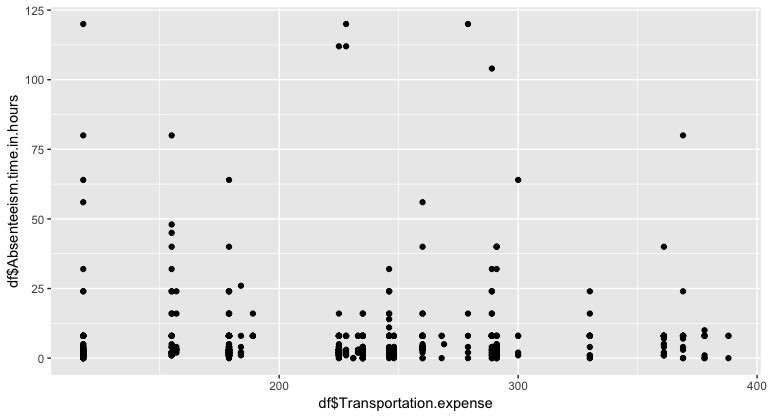
Heat map gives information on the relative collinearity amongst different variables and hence is very helpful in eliminating highly collinear variables. As we have only limited data (740 observations) we would want that the models learn as much as possible and therefore its better that we keep as much variables as possible so that the models’ efficiency/accuracy won’t be affected.

The two peaks in the weight histogram could be due to the average weight of man and woman. There’s a peak at 60-70 bin and a subsequent peak at 80-90 bin of the graph. The former could be an average of women’s weight and the latter could comprise the weight of men in our dataset.

There’s also a relation between absenteeism and transportation expense. As we can see that in the range 200-300 the frequency is maximum than that in other two ranges, with maximum number of points. The XYZ courier company could provide employees with transportation expense and/or company run buses so that the expense is not incurred from employees. This could result in an increased revenue from employees’ being present in the workplace and could offset the resulting expenditure of incurring the transportation expense.

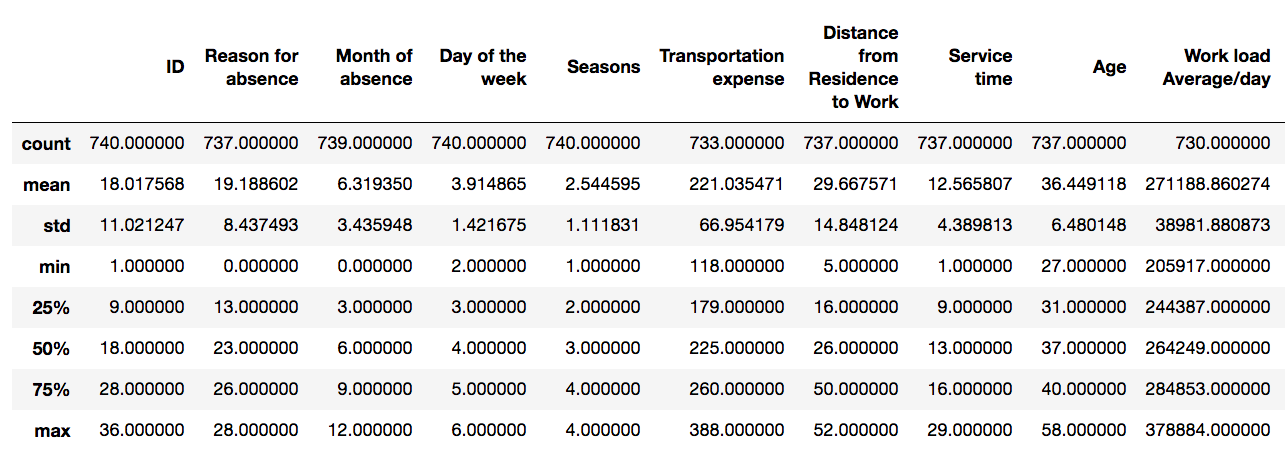
There seems to be a trend amongst people who have worked for a particular duration of time and their respective absenteeism hours. Those who have served for 8-18 years to the company, tend to be absent from the workplace more often than their colleagues who belong to other such bins.

Employees doesn’t seem to be affected much by the seasons , however there is a slight deviation from Spring to Summer. People tend to take leaves more often than they are to do so during summer. It could be due to the vacation practices or client shutdown.



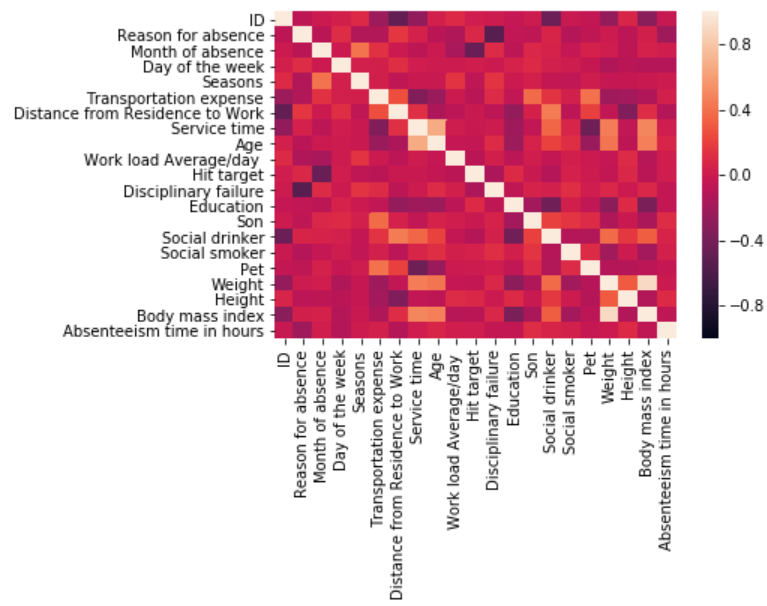
2.2 Missing Value Analysis

**missing\_val = data.frame(sapply(data,function(x){sum(is.na(x))}))**

This line of code in R will return the missing values(if any) in the dataset. In here we have used the sapply function which encapsulates data frame and a function body, which returns the sum of the ‘na’ , if any, in the dataset. The default method for is.na applied to an atomic vector returns a logical vector of the same length as its argument x, containing TRUE for those elements marked NA or, for numeric or complex vectors, [NaN](https://www.rdocumentation.org/link/NaN?package=base&version=3.5.1), and FALSE otherwise. After analysing the data, we have encountered that there are numerous missing values in our dataset and hence we have implemented the KNN Imputation method over the dataset to impute the missing values:

By using df.describe() in python, we can see the total count of all the variables.As we have 740 observations, we can clearly see that some of the variables have less than that. we could calculate the number of missing values in a particular variable by subtracting 740 with the count of that particular variable in the set. and the result would give us the missing value count.

2.3 Feature Selection

Once we receive the dataset from the client and after defining the problem category and forming the problem statement, we perform the above mentioned steps to get some sense of the data, but all that sense is intuitional and exploratory. We need to use Feature Selection in order to explore further trends and insights into out continuous and categorical variables so as to see the hidden dependencies between different such variables. As such in our dataset, we have performed two feature selection techniques: One, for the categorical variables and another one for the continuous variables.

As we could see from the correlation plot of our continuous variables, most of the data is independent of each other and the only ones which are showing dependencies are the ones who are kind of similar in a manner. There are a few variables which are highly negatively correlated, but as the number of observations is less, we decide to just feed our dataset to the model so as to train the model to adapt to the data.

For Categorical variable, we need to follow a different approach: **Chi-Square Test of Independence**

In Chi-square test of independence, we need to make two hypothesis: Null and alternated.

Out Null Hypothesis holds true if the p-value of the data is less than 0.05, and alternate hypothesis is true is p-value is greater than 0.05.

Here our null hypothesis says that the there is high dependency between Independent variable and dependent variable and low dependency between two independent variable. Vice-Versa is true for alternate hypothesis

**2.4 Feature Scaling**

As we can see from the histogram plots of all the predictor variables, the frequency as well as scale of different variables vary. For example if we take the height variable, the unit of which is centimetres, whereas if we were to take Distance from work, which has unit of kilometres. As we have to deal with different variables , almost all of which have separate units, its better that we normalise the data such that the range of all the variables will be the same (0-1). For such scale would lead to better interpretation of the data by out machine learning algorithm and would result in less anomalies. This process comes under Feature Scaling and is preferred most of the time if the dataset is not normalised. However, if you have a dataset that is already normalised then its better to go for Standardisation(**Z-Score**), a process which, instead of normalising the range from 0-1, focuses on the standard deviation of the particular value from mean of its comprising variable. Mostly the data that we receive from client is not normalised and needs lot many pre-processing steps , hence in those cases, standardisation is a no-no.

**Chapter 3**

**Modelling**

Once we are done with above mentioned steps, we need a model to feed out data and get the test result ,all with better accuracy. A good model is the one which can predict test data from the train dataset with higher accuracy.

We have divided our dataset into train and test with test data representing 20% of the dataset, while the train data represents 80% of our dataset.

We have implemented three different algorithms to our data and subsequently noted the RMSE and accuracy of the algorithm.

1. **Linear Regression**

A type of statistical model which stores coefficients and thereby predicts target variable on the basis of the coefficients.

**lm\_model = lm(Absenteeism.time.in.hours ~., data = train)**

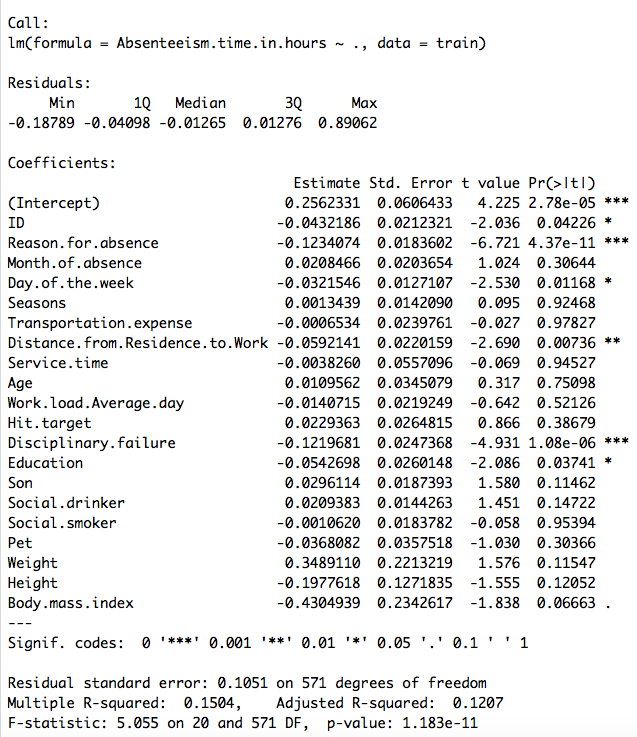
**#Error rate** 10.45% **#Accuracy**: 89.54%

The respective coefficients for our dataset is as mentioned below:

**2. Decision Tree**

Here we have employed the r-part function in R so as to impute the Decision tree on our continuous target variable. Method used is Anova. The syntax for the same is as mentioned below:

**fit = rpart(Absenteeism.time.in.hours ~ ., data = train, method = “anova")**

**#Error rate** 10.44% **#Accuracy:** 89.56

**3. K- Nearest Neighbour**

**KNN\_Predictions = knn(train[,1:20], test[,1:20], train$Absenteeism.time.in.hours, k=5)**

Using K- Nearest Neighbour method, we have tried imputing the values for our test data and found that accuracy of the model has increased and a subsequent decrease in the error rate can also be seen.

**#Error rate** 5.88% **#Accuracy:** 94.12%

Therefore we will freeze the KNN Model for our dataset and since there are not many assumptions that need to be taken into consideration while implementing KNN, it works best with our dataset, which is not normally distributed has collinear variables with limited observations.

**Appendix**

**R code for respective plots:**

**p <- ggplot(df, aes(x = Pet, fill = Pet)) + geom\_bar()**

**s <- ggplot(df, aes(x = Son, fill = Son)) + geom\_bar()**

**t <- ggplot(df, aes(x = Seasons, fill = Seasons )) + geom\_bar()**

**r <- ggplot(df, aes(x = Education, fill = Education)) + geom\_bar()**

**u <- ggplot(df, aes(x = Disciplinary.failure, fill = Disciplinary.failure)) + geom\_bar()**

**v <- ggplot(df, aes(x = Social.drinker, fill = Social.drinker)) + geom\_bar()**

**w <- ggplot(df, aes(x = Social.smoker, fill = Social.smoker)) + geom\_bar()**

**x <- ggplot(df, aes(x = Day.of.the.week, fill = Day.of.the.week)) + geom\_bar()**

**z <- ggplot(df, aes(x = df$Hit.target, y = df$Absenteeism.time.in.hours)) + geom\_point()**

**grid.arrange(z, nrow=1)**

**grid.arrange(p,s, nrow=1)**

**grid.arrange(t,r, nrow=1)**

**grid.arrange(u,v, nrow=1)**

**grid.arrange(w,x, nrow=1)**

**hist(df$Reason.for.absence, col='yellow')**

**df$Transportation.expense <- log10(df$Transportation.expense)**

**cor(x=df$Transportation.expense, y= df$Absenteeism.time.in.hours, method="pearson")**

**References**

* **[www.edwisor.com](http://www.edwisor.com)**
* **[www.datacamp.com](http://www.datacamp.com)**
* **[www.kaggle.com](http://www.kaggle.com)**
* **[www.Forbes.com](http://www.Forbes.com) : Article on absenteeism at work**