**Project Report on**

**Employee Absenteeism**

Anurag Bagadi

27th Sep 2018

**CONTENTS**

**Page No.**

1. **INTRODUCTION**…………………………………………………………………...4

**1.1.** Problem Statement …………………………………………………………...4

**1.2.** Data …………………………………………………………………………….4

**1.3** Exploratory Data Analysis …………………………………………………….7

**2. METHODOLOGY**…………………………………………………………………....8

**2.1.** Data Exploration……………………………………………………………8

**2.2. Preprocessing techniques**……………………………………………………...9

**2.2.1.** Missing value analysis…………………………………………………..9

**2.2.2.** Outlier Analysis……………………………………………………….…11

**2.2.3.** Feature selection………………………………………………………..15

**2.2.4.** Feature scaling…………………………………………………………..18

**2.2.5** Sampling Technique……………………………………………………..21

**3. MODELING**……………………………………………………………………………21

**3.1.** Decision Tree regression………………………………………………….22

**3.1.** Random Forest regression………………………………………………..23

**3.1.** Linear Regression………………………………………………………….24

**4. MODEL EVALUATION**……………………………………………………………...28

4.1 RMSE …………………………………………………………………….… 28

4.2 Model Selection ………………………………………………………….....29

4.3 Questions Answers………………………………………………………....29

**Appendix – A** ………………………………………………………………………30

**ABSTRACT**

Absenteeism means pattern of absence from a duty without any good reason. Absenteeism is unplanned absences from ones duty and is indicator poor individual performance which in turn leads to poor performance of the company. In this project XYZ a courier company facing a severe issue of absenteeism has shared its dataset which contains the information about the employees of that company of previous year. On the basis of the previous year dataset and observed pattern and insights we have to predict the future values and help them to know the factors which help them to reduce the number of absenteeism as well as the monthly loss company is going to face in the coming year i.e. 2011.

We are using both R and Python to build the suitable model according to the company’s problem statement. We will try different ML Algorithm and will choose the best model accordingly to help them to know the answers for the questions mentioned above.

1. **Introduction**:

Absenteeism is the unplanned absence from the duty without any good reason. It is one of the severe issues faced by many companies. We know that human capital plays an important role in any business. Being able to predict the employee absenteeism can prevent company from severe loss. It is very important to know the cause of absenteeism among the employee. One of the best ways to make the predictions is with the help of machine learning techniques.

* 1. **PROBLEM STATEMENT**

Absenteeism leads to severe business loss. XYZ is a courier company and is facing a genuine issue of absenteeism. The aim of this project is to predict the factors to reduce the number of absenteeism and the work loss company is going to face next year if same trend of absenteeism continues. To make the predictions we used R and Python codes and algorithms.

* 1. **DATA**

There are 21 variables in our data in which 20 are independent variables and 1 (Absenteeism time in hours) is dependent variable. Since our target variable is continuous in nature, this is a regression problem.

**Variables Information:**

**1.** Individual identification (ID)

**2.** Reason for absence (ICD) -

Absences attested by the **International Code of Diseases** (ICD) stratified into 21 categories (I to XXI) as follows:

**I**. Certain infectious and parasitic diseases

**II**. Neoplasms

**III.** Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

**IV**. Endocrine, nutritional and metabolic diseases

**V**. Mental and 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 (kilometres)

**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)

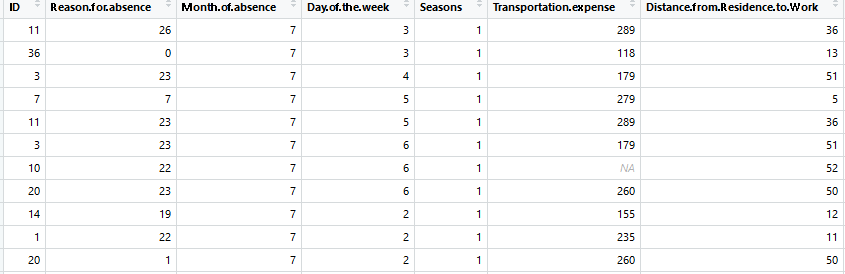
**17.** Pet (number of pet)

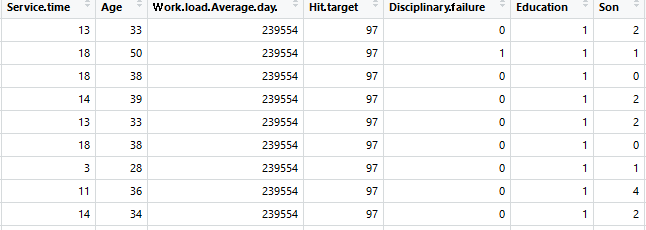
**18.** Weight

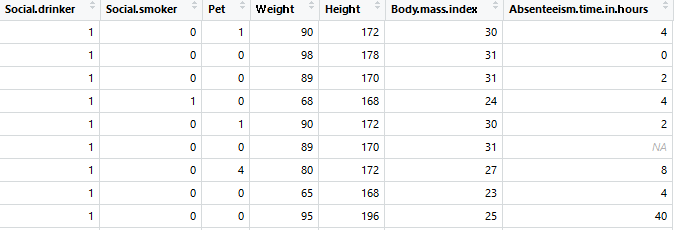
**19.** Height

**20.** Body mass index

**21**. Absenteeism time in hours (target)







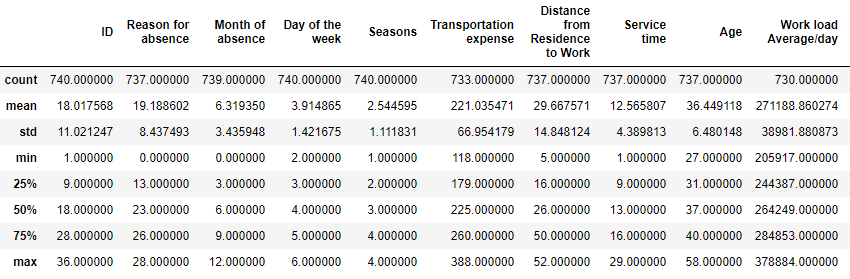
**1.2.1 Variable Detailed table**

**1.3 Exploratory Data Analysis (EDA):**

It is an approach to analysing data sets to summarize their main characteristics. In the given data set there are 21 variables and data types of all variables are either float64 or int64. There are 740 observations and 21 columns in our data set. Missing value is also present in our data.

**List of columns and their number of unique values** –

|  |  |
| --- | --- |
| **Variable** | **Number of unique value** |
| ID | 36 |
| Reason for absence | 28 |
| Month of absence | 13 |
| Day of the week | 5 |
| Seasons | 4 |
| Transportation expense | 24 |
| Distance from Residence to Work | 25 |
| Service time | 18 |
| Age | 22 |
| Work load Average/day | 38 |
| Hit target | 13 |
| Disciplinary failure | 2 |
| Education | 4 |
| Son | 5 |
| Social drinker | 2 |
| Social smoker | 2 |
| Pet | 6 |
| Weight | 26 |
| Height | 14 |
| Body mass index | 17 |
| Absenteeism time in hours | 19 |



1.3.1 **Summary of all columns**

From EDA we have concluded that there are 10 continuous variable and 11 categorical variable in nature.

1. **Methodology**
   1. **Data Exploration**

Before feeding the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project we spend almost 80% of time in it. The whole data process is divided into six phases.

* **Business understanding**: When any client comes in we should try to understand their problem statement first. It helps us to get proper data for better results.
* **Data understanding**: In this we use many statistical techniques, Graphs and visualizations to understand the data so that we can understand the data well and can get relevant data from the client.
* **Data Preparation**: This means exploring the raw data we receive from client and understanding what data speaks out. In data science 80% of our time goes in data understanding, cleaning and preparation and 20% in model development and model evaluation. If the quality of data is good the model will predict better and results in high accuracy.
* **Data modeling**: There are many machine learning algorithms and we have to select the most appropriate algorithm according to our problem statement.
* **Evaluation**: It helps us to evaluate our model. It tells us whether our model is able to accomplish the business objective or not.
* **Deployment**: This is the final phase in which we deploy our model in client premises

**2.2 Pre Processing Techniques**:

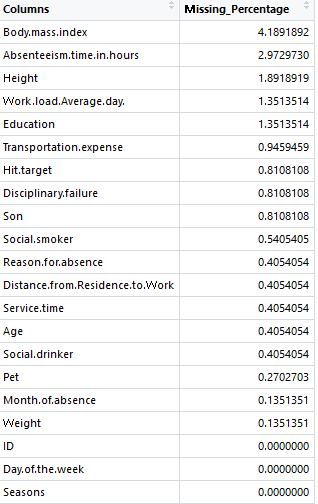
Any predictive modelling requires that we look at the data before we start modelling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process we will first try and look at all the probability distributions of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable.

**2.2.1** **Missing Value Analysis**

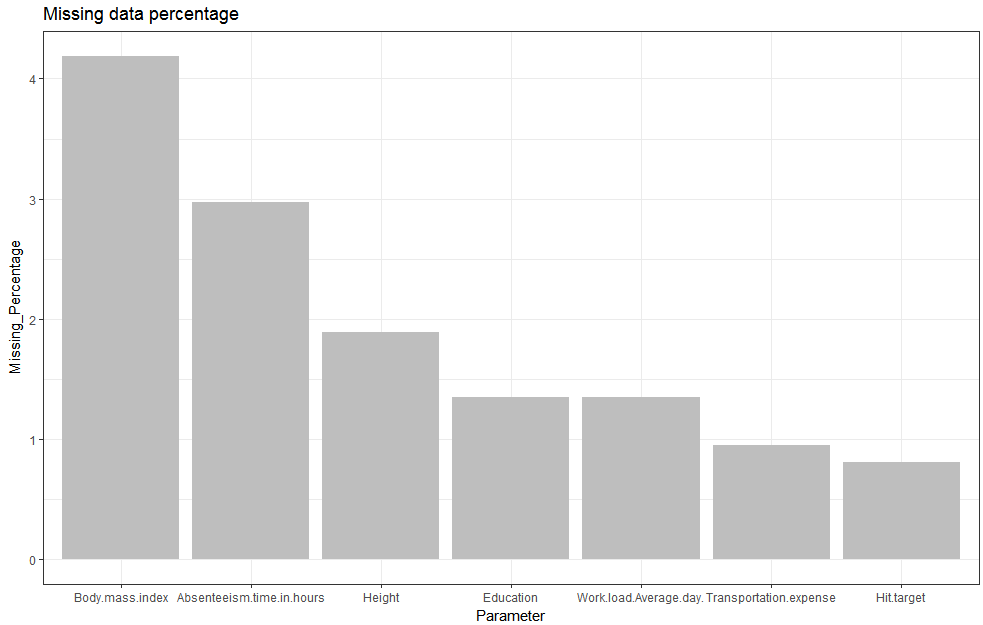
Missing value is the values which are not present or missing from the dataset. Missing values appears in our dataset due to various reasons like human error, refuse to answer the questions in a survey or optional box questionnaire. The skipped or unanswered questions appear in form of missing values. Missing values can be treated either by dropping the variable or by imputing the missing values.

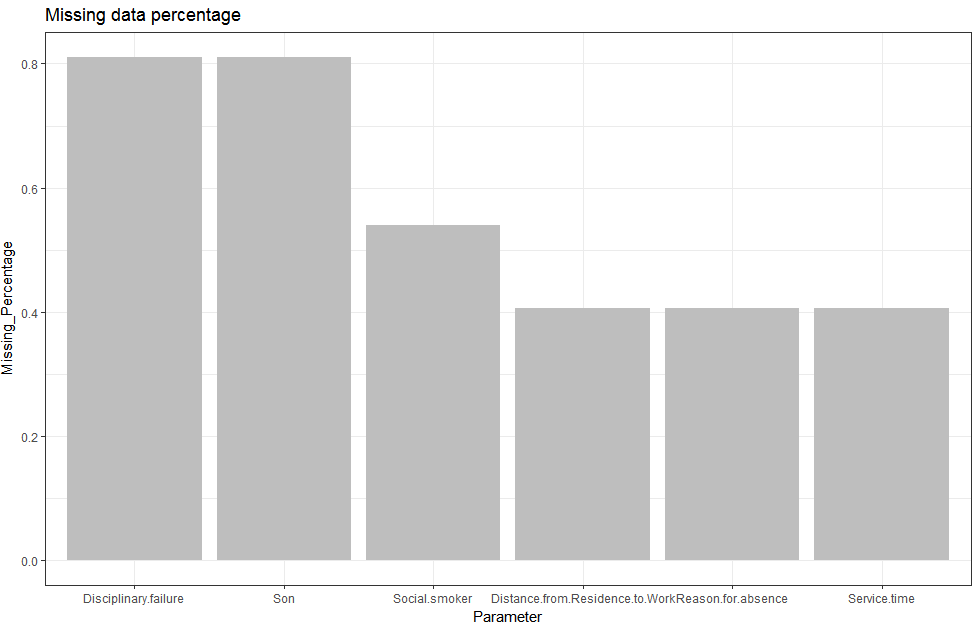
* **When to ignore the missing values:** First we will create a data frame which tells us amount of missing values present in each variable. Drop the variables which consists more than 30% (according to industry standards) of missing values.
* **When to impute missing values and methods of imputation:** We will impute those variables whose missing percentage is less than 30%. There are three methods to impute missing values:
  + Fill with central statistics method i.e. mean and median for continuous variable and mode (majority minority rule) for categorical variable.
  + Distance based or Data mining method which includes KNN imputation.
  + The last method is prediction method which is based on ML algorithms.

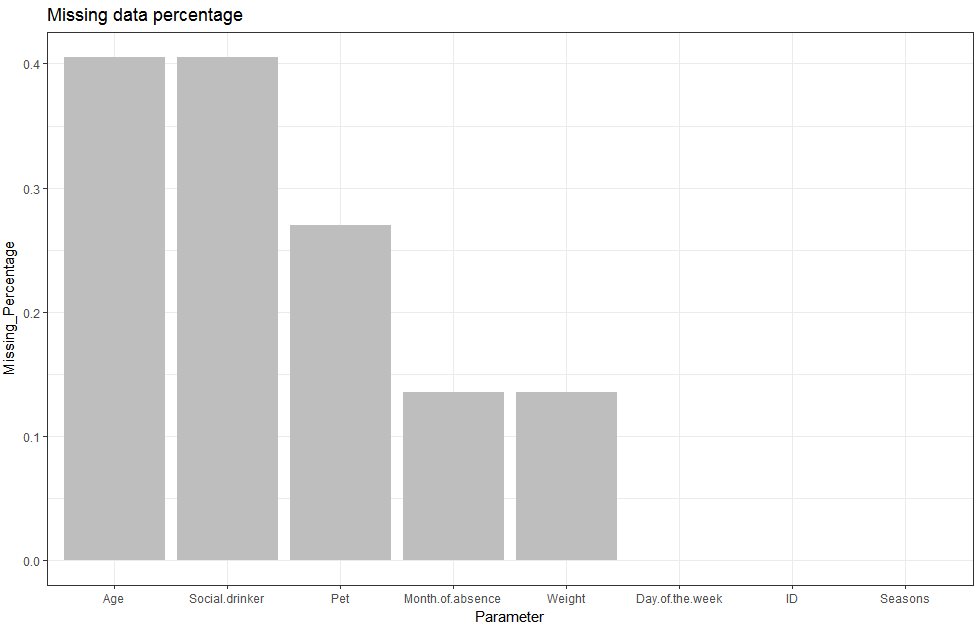
**In our dataset we have 135 missing values. All the variables have less than 10% of missing values so we have imputed them using KNN imputation**.The graphs of missing values percentage are presented below.



**2.2.1.1 Variable missing value percentage**







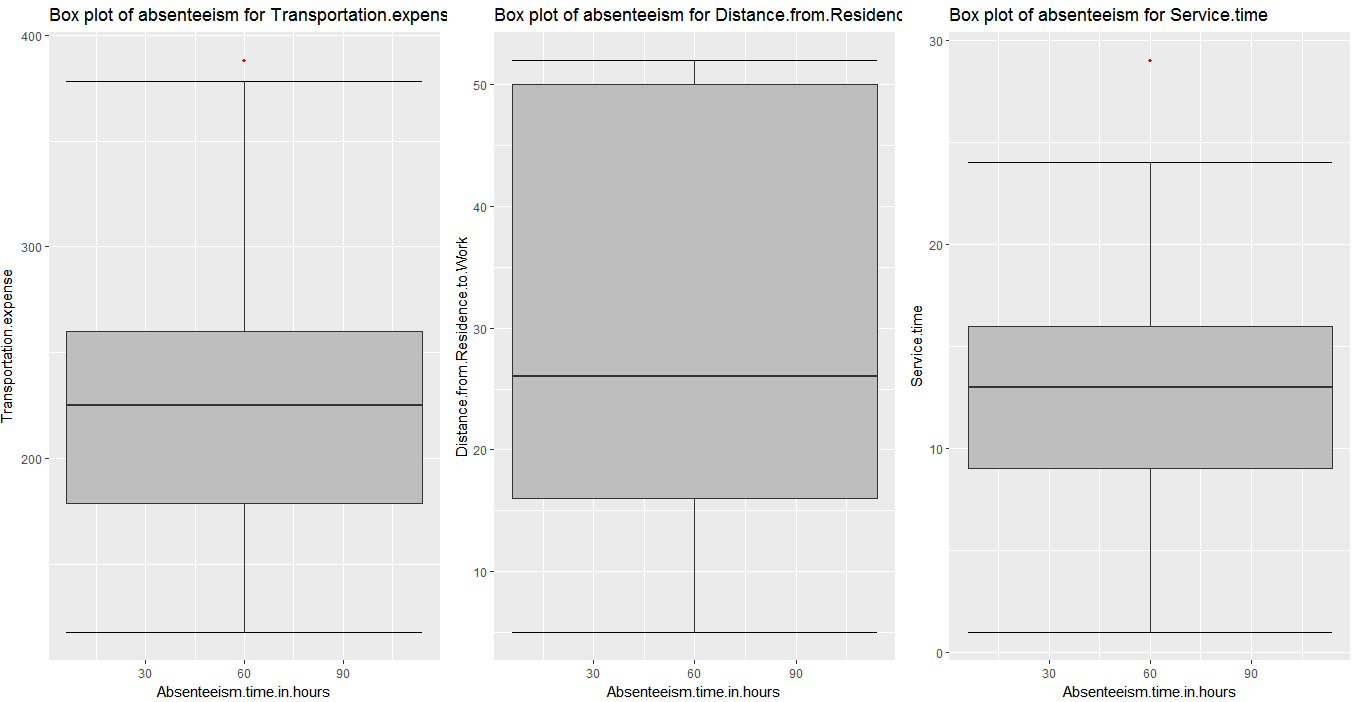
**Fig 2.2.1.1 Missing Data Percentage graph**

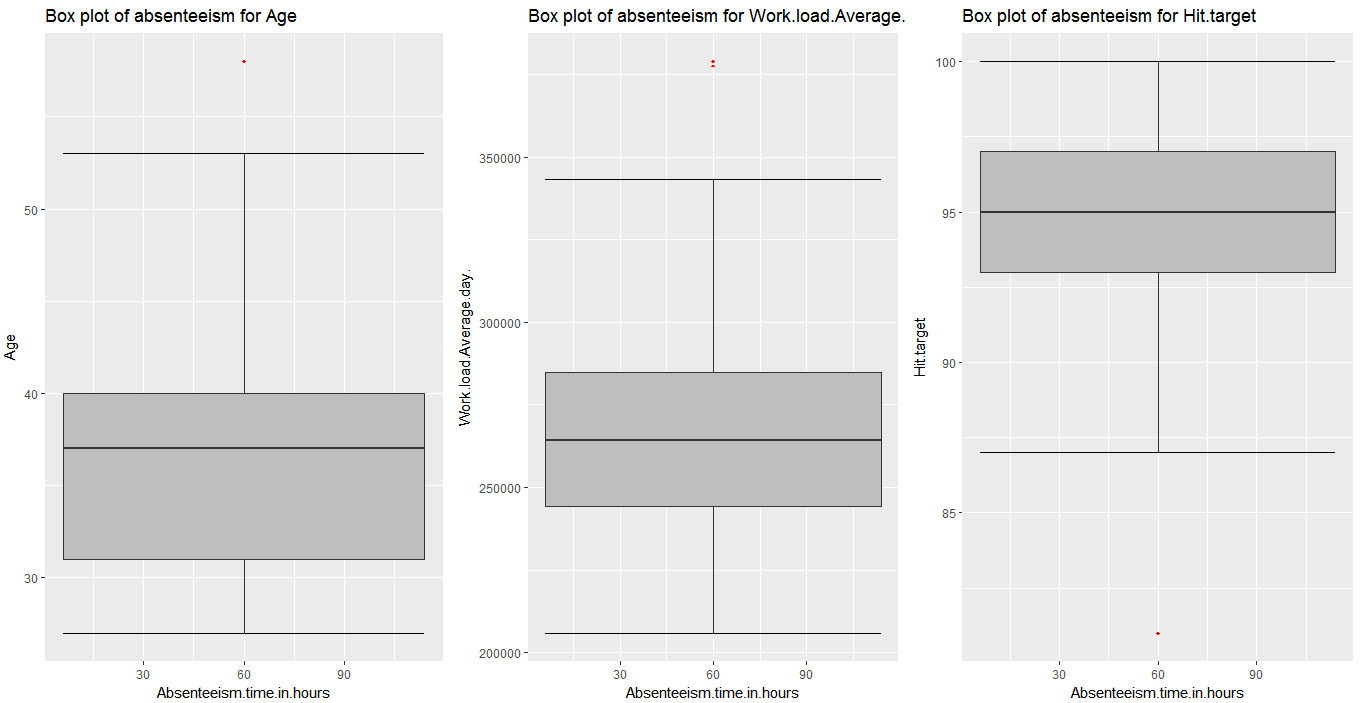
**2.2.2** **Outlier Analysis**

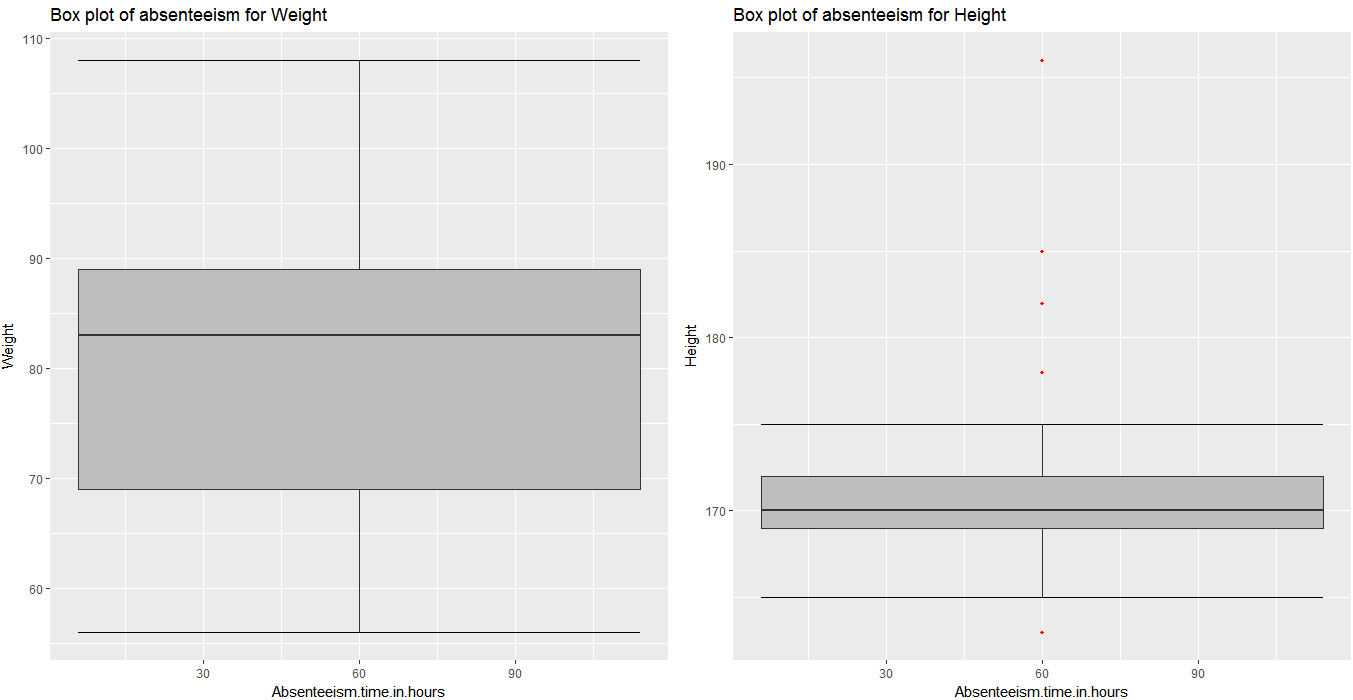
Outlier analysis is one of the pre-processing techniques used to check for abnormal values in the data set clean them and transform the data into a proper shape. Presence of outlier in our data leads to poor data quality and contamination, low quality measurement and manual errors. The best way to look at outlier is to understand business process i.e. how data is generated and how is the business flow. An outlier is an observation point that is distant from existing observations

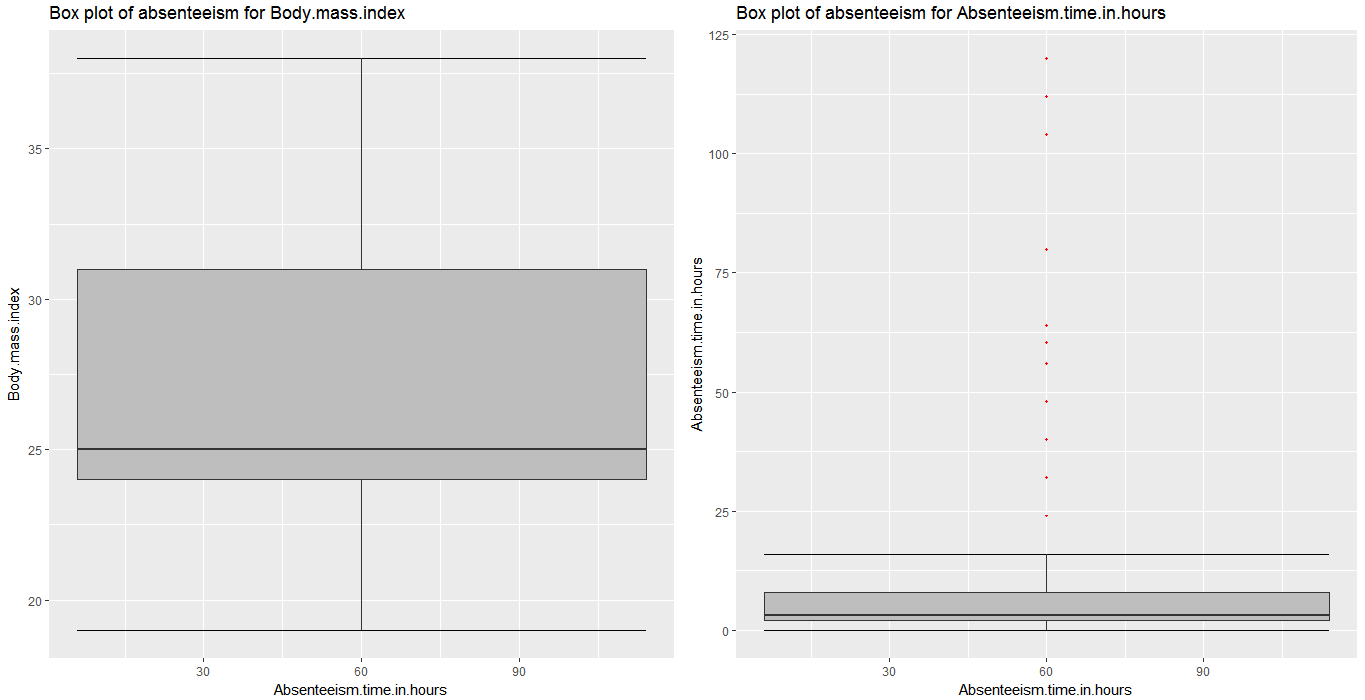
We know that outlier analysis is applicable only on numerical variable so we have converted the entire variable in their appropriate data types and separated out numerical variables for outlier analysis. One of the steps of pre-processing is to remove outliers. In this case we use a classic approach of removing outliers, Turkey’s method. We visualize the outliers using boxplots. We have plotted box plot for each numerical variable which is shown below.

From the boxplot almost all the variables **except “Distance from residence to work”, “Son”, “Weight” and “Body mass index”** consists of few outliers. So we no need to remove the outliers.

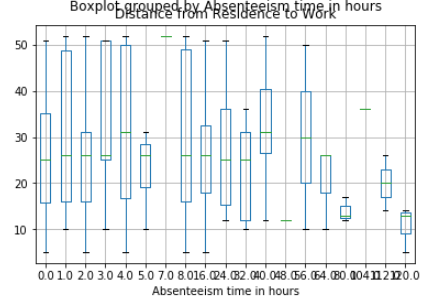
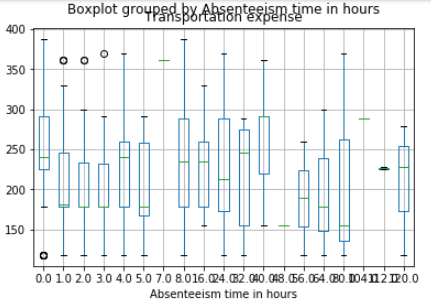


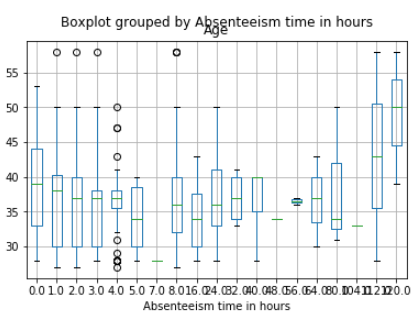
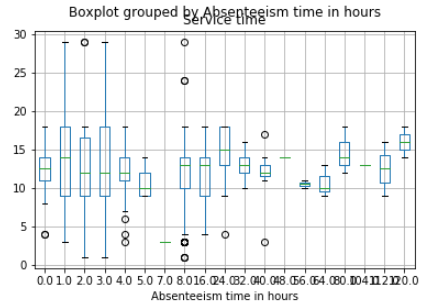


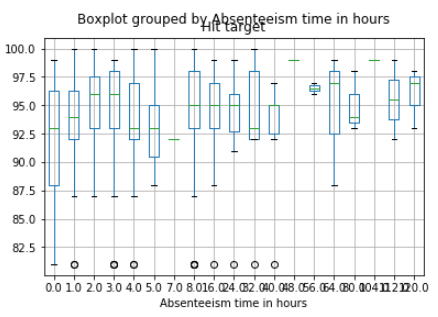
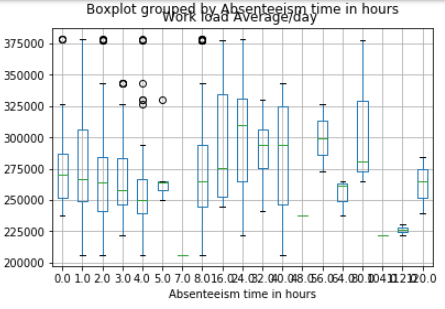


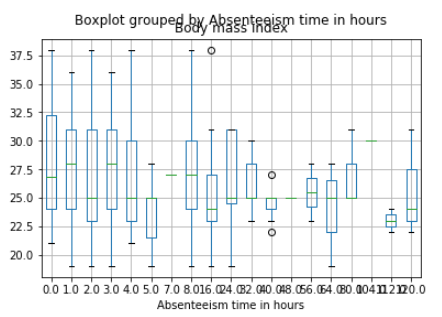
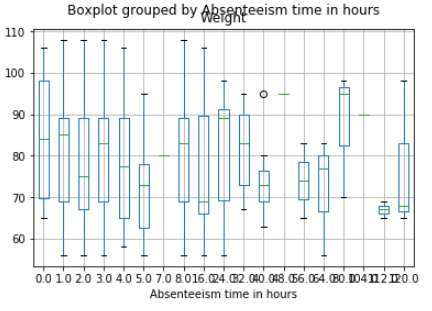
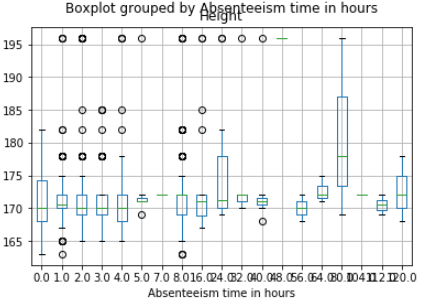


**Fig 2.2.2.1 Box plot of Numeric variable in R**









**Fig 2.2.2.2 Box plot of Numeric variable in Python**

**2.2.3** **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 problem of class prediction**. Selecting subset of relevant columns for the model construction** is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. We should use feature selection because of below reasons:

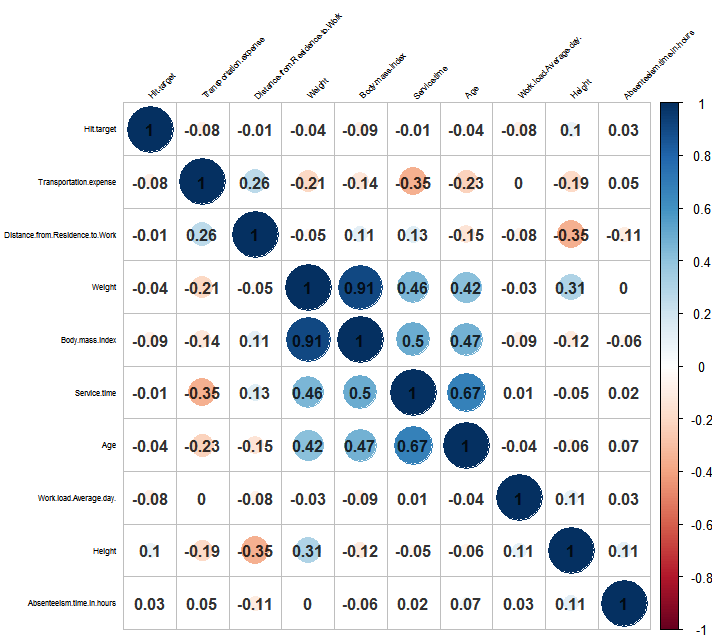
* It enables the machine learning algorithm to train faster.
* It reduces the complexity of a model and makes it easier to interpret.
* It improves the accuracy of a model if the right subset is chosen.
* It reduces over fitting.

In this project we have selected **Correlation Analysis** for numerical variable and **ANOVA** (Analysis of variance) for categorical variable.

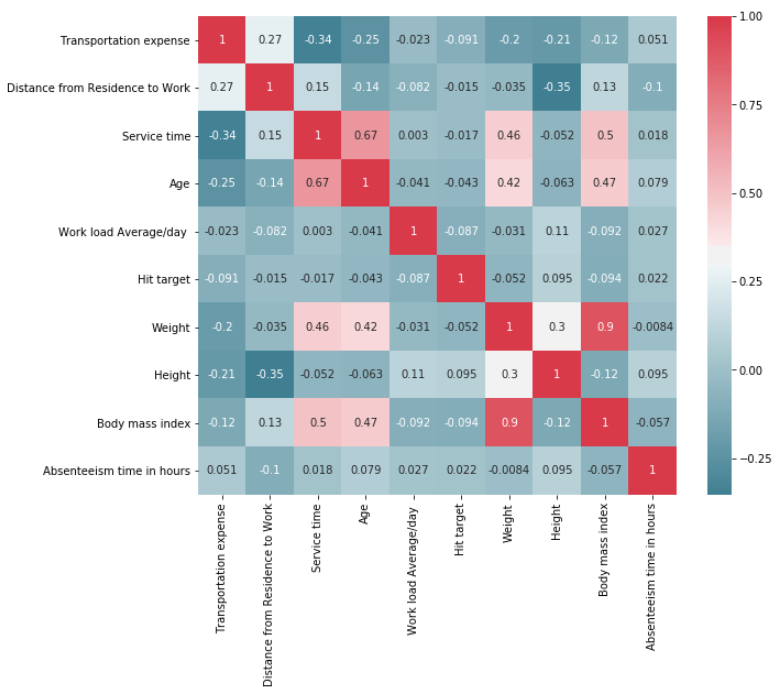
* **Correlation analysis:** It is one of the methods for feature selection technique applied only on numerical data. Correlation tells us the association between two continuous variables. It ranges for -1 to +1.
  + **-1: Highly negatively correlated**
  + **0: No correlation**
  + **+1: Highly positively correlated**

In correlation there is an assumption that there should be high dependency between predictor and target variable but there should be low dependency between two predictors. From the below diagram we can see that:

* Weight and body mass index are highly correlated with each other.
* Service time and Age are also close to threshold value of correlation.
* Body mass index is the calculated by dividing their weight in kilograms by the square of their height in metres. So we can also remove height from the dataset in model development.



**Fig 2.2.3.1 Correlation analysis in R**



**Fig 2.2.3.2 Correlation plot in Python**

**ANOVA (Analysis of Variance):** It is applied on one categorical and one continuous variable. It is a statistical technique used to compare means of two or more group. We get results in form of p values. If p value is less than 0.05 we will reject NULL HYPOTHESIS (result purely from chance) and accepts ALTERNATE HYPOTHESIS (influenced by some non-random cause). From the below anova table except ID, Reason for absence, Day of the week and Sons for remaining all the variable p-value is greater than 0.05 so we accept the alternate hypothesis. Below is the list of variable which will be removed from analysis:

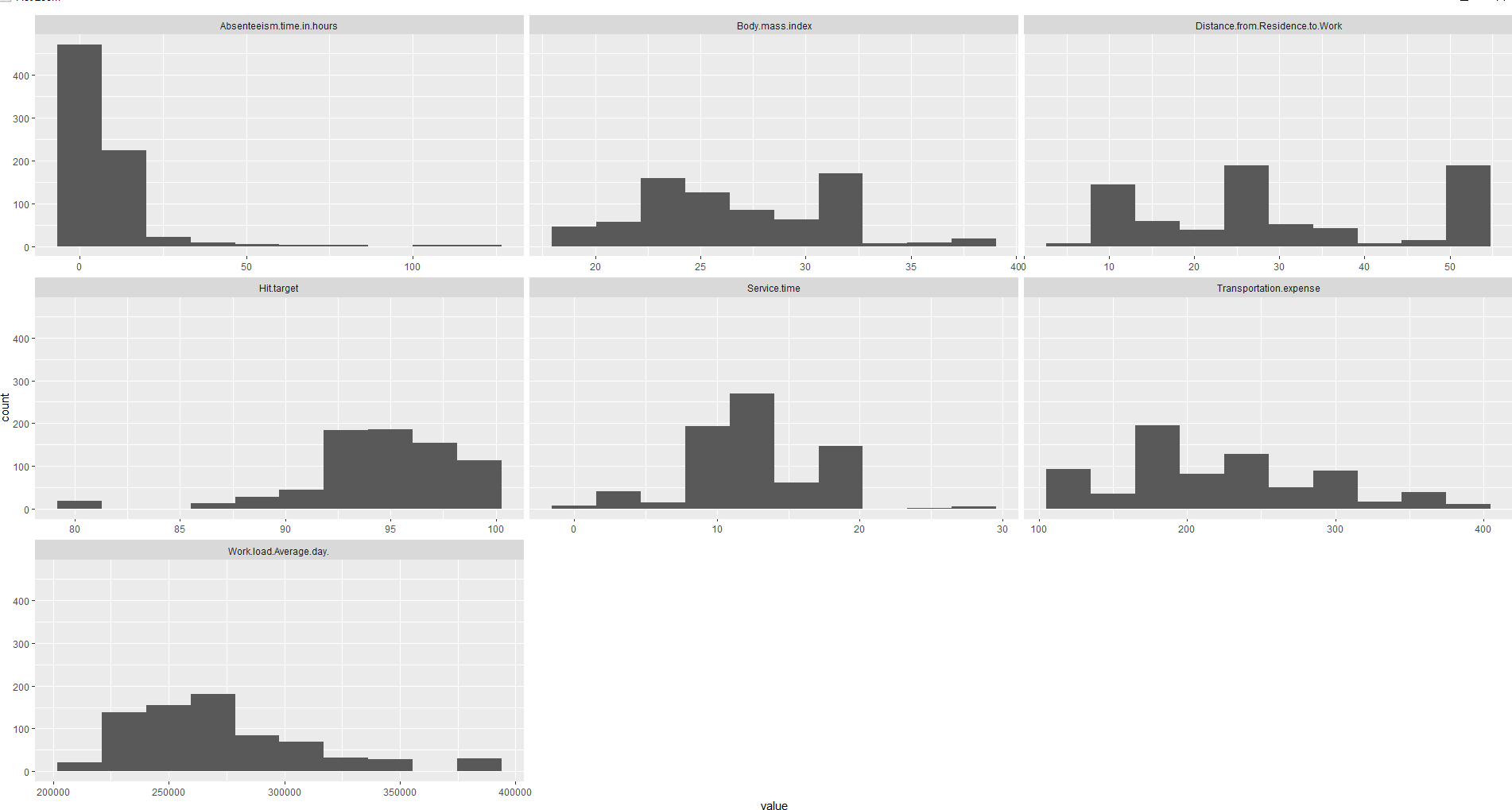
* Month of absence
* Seasons
* Disciplinary Failure
* Education
* Social smoker
* Social drinker
* Pet

|  |  |  |
| --- | --- | --- |
| **Variable name** | **p- value in R** | **p- value in Pythpn** |
| ID | 6.37e-05\*\*\* | 0.000211 |
| Reason for absence | <2e-16\*\*\* | 1.327218e-18 |
| Month of absence | 0.0831 | 0.136601 |
| Day of the week | 0.0134 \* | 0.017966 |
| Seasons | 0.406 | 0.393821 |
| Disciplinary Failure | 0.101 | 0.14492 |
| Education | 0.669 | 0.960249 |
| Social smoker | 0.28 | 0.24839 |
| Social drinker | 0.0912 | 0.086034 |
| Pet | 0.263 | 0.375844 |
| Sons | 0.000356\*\*\* | 0.002032 |

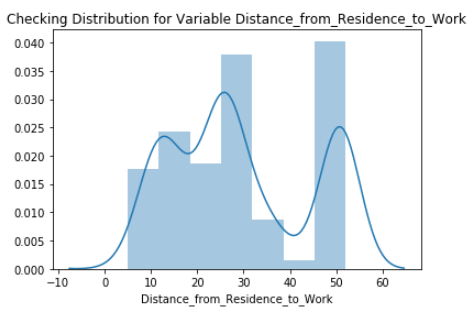
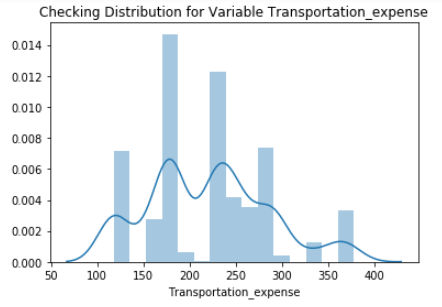
**Fig 2.2.3.3 Anova table**

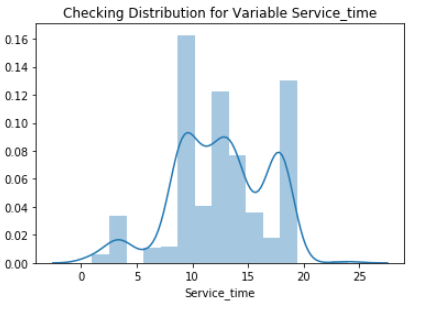
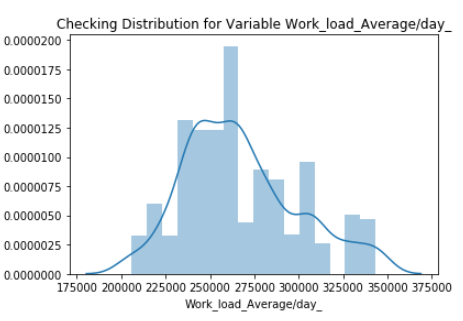
**2.2.4 Feature Scaling**

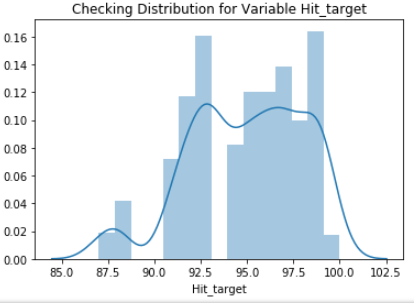
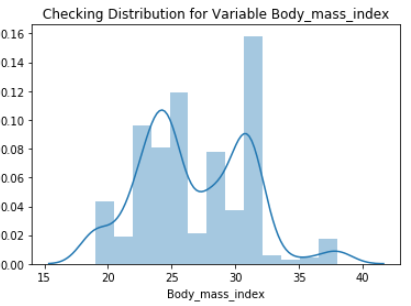
**Feature scaling** is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

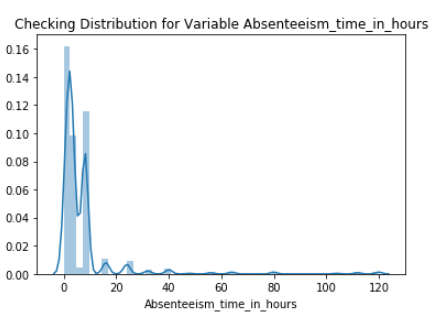


**Fig 2.2.4.1 Numeric variable Histogram plot in R**



**Fig 2.2.4.2 Numeric variable Histogram plot in Python**

**2.2.5 Sampling Technique:**

Sampling allows data analysts to work with a small, manageable amount of data in order to build and run analytical models more quickly, while still producing accurate findings. Sampling can be particularly useful with data sets that are too large to efficiently analyse in full.

There are many different methods for drawing samples from data, and the ideal one depends on the data set and situation. There are many types of sampling techniques for our problem we can use simple random sampling where we draws a sample from the population without any replacement.

1. **MODELING**

After a thorough pre-processing we will be using some regression models on our processed data to predict the target variable. Following are the models which we have built on the dataset after exploratory data analysis. The dependent variable can fall in either of the four categories:

1. Nominal

2. Ordinal

3. Interval

4. Ratio

If the dependent variable is nominal the only predictive analysis that we can perform is Classification, and if the dependent variable is Interval or Ratio the normal method is to do a Regression analysis, or classification after binning. But the dependent variable we are dealing with is Interval, so only regression predictive analysis can be done. We will start our modelling with simplest one and then switch to little complex models. Let’s us first divide the data in train and test using simple random sampling.

* 1. **Decision Tree Regression**

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each branch connects nodes with “and” and multiple branches are connected by “or”. It can be used for classification and regression. It is a supervised machine learning algorithm. Accept continuous and categorical variables as independent variables. Here Decision tree uses CART decision tree methodology.

Decision Tree is a recursive partitioning approach and CART split each of the input node into two child nodes, so CART decision tree is Binary Decision Tree.   At each level of decision tree, the algorithm identify a condition - which variable and level to be used for splitting input node (data sample) into two child nodes. The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any should be. The main elements of CART (and any decision tree algorithm) are:

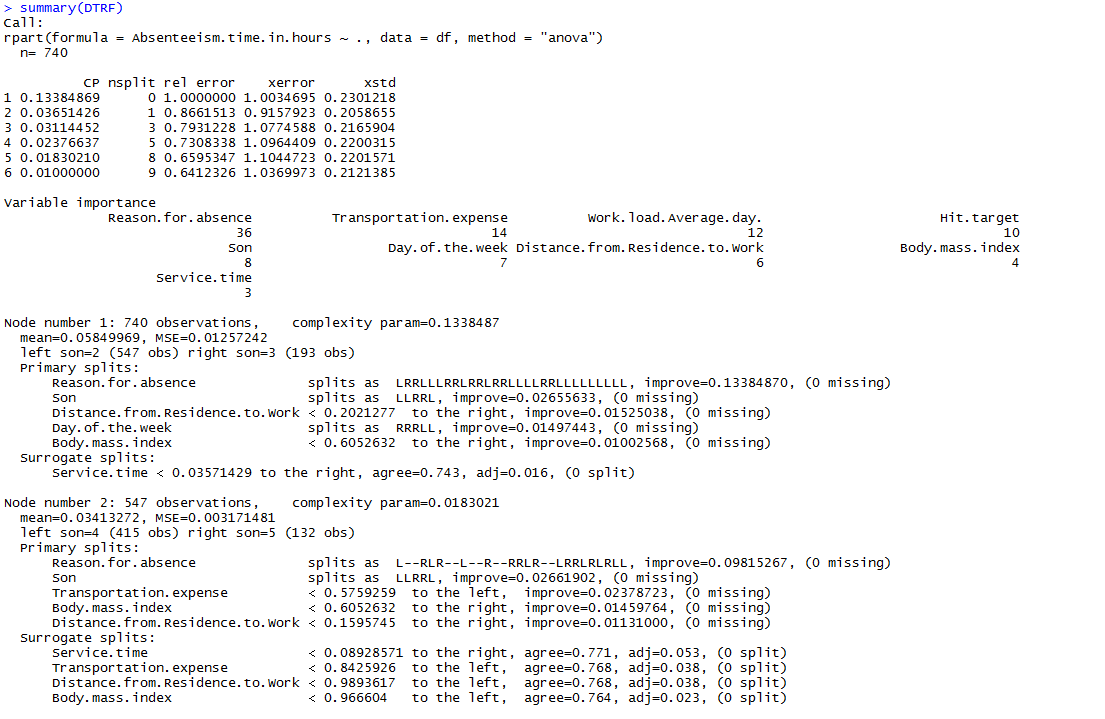
1. Rules for splitting data at a node based on the value of one variable;
2. Stopping rules for deciding when a branch is terminal and can be split no more; and
3. Finally, a prediction for the target variable in each terminal node.

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **R** | **PYTHON** |
| **RMSE** | 7.53 | 10.56 |
| **ACCURACY** | 92.47 | 89.44 |

**3.1.1 Decision Tree Regression Model Error Metrics**



**3.1.2 Decision Tree Regression Error Metrics**



**3.1.1 Decision Tree Regression Model Summary**

* 1. **Random Forest Regression**

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision. The method combines Breimans “bagging” idea and the random selection of features.

|  |  |  |
| --- | --- | --- |
| **Random Forest** | **R** | **PYTHON** |
| **RMSE (n = 100)** | 9.31 | 10.70 |
| **ACCURACY (n = 100)** | 91.69 | 89.30 |
| **RMSE (n = 200)** | 9.49 | 10.57 |
| **ACCURACY (n = 200)** | 91.51 | 89.43 |
| **RMSE (n = 300)** | 9.32 | 10.38 |
| **ACCURACY (n = 300)** | 91.68 | 89.62 |

**3.2.1 Random Forest Regression Model Error Metrics**



**3.2.2 Random Forest Tree Regression Error Metrics**

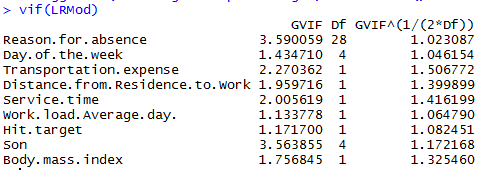
* 1. **Liner Regression**

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. To build any model we have some assumptions to put on data and model.

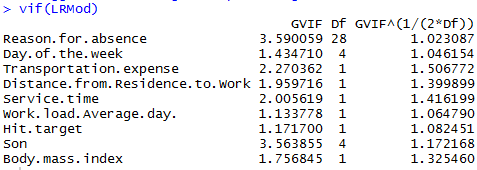
Here are the assumptions to the linear regression model.

* **Linear relationship**: It assumes that the data which is fed in linear regression model have linear relationship between dependent and independent variable.
* **Multivariate normality**: Linear regression assumes that our target variable is normally distributed. It means that it is following the normality assumption.
* **No or little Multicollinearity:** Two highly correlated variables in a dataset lead to multicollinearity effect. There is one test called VIF test (Variance Inflation Factor test). We need to run this test before feeding the data to the model to know either our data contains the correlated independent variables or not.
* **No Auto Correlation:** It means there should be no correlation between the residuals. When we build a linear regression model we will get residuals (range of errors). Here we assume that there is no auto correlation it means error are independent.

Once our data satisfy these assumptions we go ahead and build linear regression model. Under this model on training data we build equation which carries an intercept and coefficient for all independent variable. Then we save that equation and then once new test data comes in then we allow passing the test case on linear regression equation to estimate the predicted value. Whatever the value predicted that will hold a target value to the new test data.



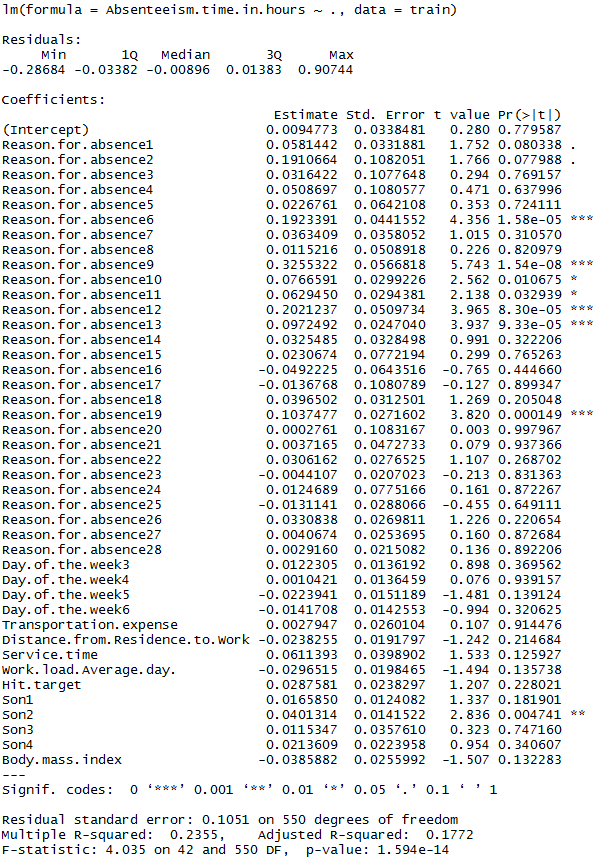
**3.3.1 VIF table to check multicollinearity in Python**



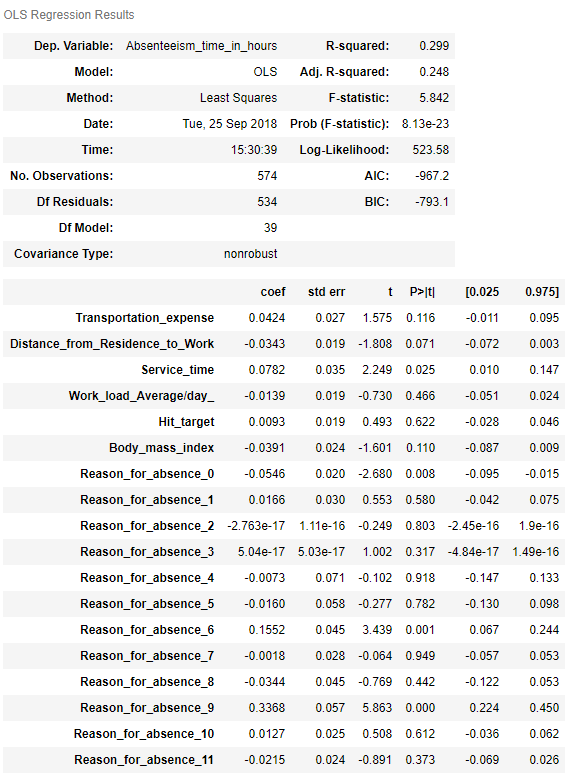
**3.3.2 VIF table to check multicollinearity in R**

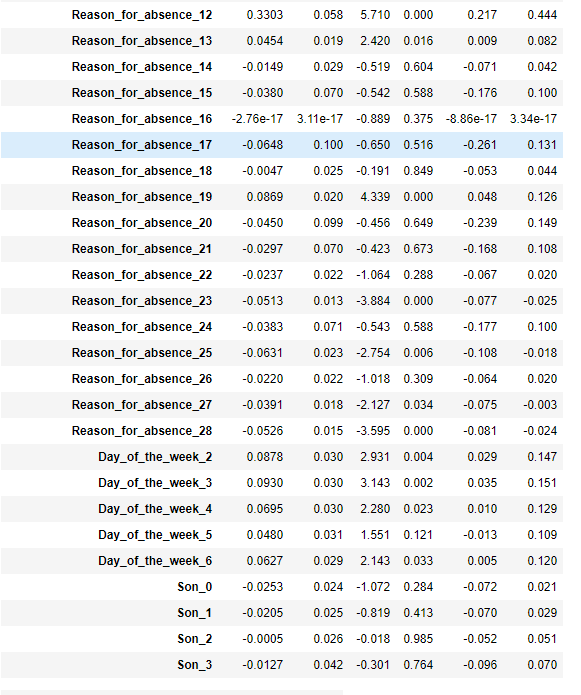


3.3.3 **Linear Regression Error Metrics**



**3.3.4 Linear Regression Model Summary in R**





**3.3.5 Linear Regression Model Summary in python**

|  |  |  |
| --- | --- | --- |
| **Linear Regression** | **R** | **PYTHON** |
| **RMSE** | 8.68 | 9.52 |
| **ACCURACY** | 91.32. | 91.48 |

**3.3.6 Linear Regression Model Error Metrics**

1. **Model Evaluation:**

After building number of regression models there are criteria by which they can be evaluated and compared. Model evaluation tells us whether our model is able to accomplish the business object or not. There are different metrics for regression model like **MSE (Mean Square Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error) etc.** MSE and RMSE are used for **transition or time series data also called time series analysis** whereas MAPE and MAE are used for normal regression data. Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Whereas **R**-**squared** is a relative measure of fit, **RMSE** is an absolute measure of fit.

As the square root of a variance, **RMSE** can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of **RMSE** and higher value of **R-Squared Value** indicate better fit.

If dataset is transitioned or time based then we go for RMSE. If we want to convert error number in particular percentage we should go for MAPE. Our project is time series multivariate so we have used RMSE as error metric. Accuracy can be calculated as:

**Accuracy = 100 - RMSE**

**4.1 RMSE (Root Mean Square Error)**

RMSE is a popular metric to measure the error rate of time series or transition regression model. It can be only compared between models whose errors are measured in the same units. It can be calculated by squaring the errors, finding their average and taking their square root. It can be mathematically represented as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| http://www.saedsayad.com/images/RMSE.png |  | |  | |
| http://www.saedsayad.com/images/actual_predicted.png  **4.2 Model Selection**  From the observation of all **RMSE Value** and **R-Squared** Value we have concluded that **Linear Regression Model** has minimum value of RMSE and its R-Squared value is also maximum. So Linear Regression Model is the best model among the all. | |  |  | |

**4.3 Questions Answers:**

**Q1. The Changes which company should bring to reduce the number of absenteeism:**

**Ans:** In this project **“Employee Absenteeism”** we have applied different models to predict the final result. In R and Python we have selected Linear Regression for predictions as it gave the highest accuracy and rmse.

* After applying the regression we conclude that Diseases of nervous system, circulatory system disease, respiratory system, musculoskeletal system and connective tissue disease and skin disease are major cause of employee absenteeism.
* Through this study it is identified that the employees were mostly suffered from health issues. So the company must focus on these above given reasons to minimize absenteeism.
* Even it is possible to eliminate absenteeism completely by the way of providing valuable means to their internal resources i.e. employees by providing some medical facilities and by creating awareness regarding health. Though absenteeism is invisible but proves fatal for entire company.
* The other reason for employee absenteeism is that few employee have more than 1 children’s so company can start in house day care facility for their employee to reduce the absenteeism rate.

Q2. **How much losses every month can we project in 2011 if same trend of absenteeism continues?**

**Ans:** The low work performance of the company leads to work loss. One of the major factors of work loss is employee absenteeism. In our project we have predicted the work loss faced by the company in year 2011. We have computed the monthly work loss the company is going to face in coming year with the help of the formula give below. Table 5.1 shows the monthly work loss the company is going to face in year 2011. Work loss can be calculated as:

**Work loss = (Absenteeism time in hours \* Work load average/day)/ Service time**

Refer to the below table for the work loss accumulated per month.

|  |  |  |
| --- | --- | --- |
| **Month** | **Work Loss/month in R** | **Work Loss/month in Python** |
| Jan | 6351681 | 6351550 |
| Feb | 8268540 | 8268542 |
| mar | 16330574 | 15707449 |
| Apr | 10999488 | 10999489 |
| May | 9668487 | 9326392 |
| Jun | 14757147 | 14362241 |
| Jul | 19249816 | 19015383 |
| Aug | 10517960 | 8791557 |
| Sep | 6630293 | 6482816 |
| Oct | 9129702 | 8931648 |
| Nov | 13691680 | 12244254 |
| Dec | 12352110 | 12299778 |

**Appendix A**

**Basic Analysis Terms Definition**

* **VIF (Variance Inflation Factor)** is used to detect and remove multicollinearity. It is one of the assumptions of linear regression. VIF is used only on independent variable. It is calculated by the formula,

VIF = 1/1-r2

Where, r2 = % variance in variables & 1-r2 also called tolerance of the model.

If r2 is high it means the given variable is redundant. So we need not to bring the given variable in the model. It means the given variable is highly correlated. If r2 is low it means the given variable is not redundant and we should include that variable in our model. It means the given variable is less correlated.

Higher the VIF more collinear is the variable which means we should not include that variable in our model. Lower the VIF less collinear is the variable which means it can be included in our model.

* **Residual standard error:** It is also called standard deviation error. It measures the average amount that the coefficient estimates vary from actual average value of our response variable. It helps in calculation of p-value.
* **t- value:** It measures how many standard deviation our coefficients are away from 0. Coefficients should be far away from zero because if coefficient of any variable is near to 0 it means that variable is not able to explain the target variable i.e. that variable is an irrelevant variable. With help of t-value we calculate p-value.
* **p-value:** It helps us to decide whether to accept or reject the variable i.e. a variable is contributing much information or not.
* **F-statistics:** It is a good indicator of whether there is a relationship between our predictor and the response variable. F-statistics should be greater than 1.
* **Degrees of Freedom:** Number of observation (training data) – Total number of variable
* **R Square:** It is numerical value which tells us the amount of variance of the dependent variable is explained by all independent variable. It tells us how much our model is robust and what the strength of model on training data is.
* **Adjusted R Square:** It is derived from R-Square values. Adjusted R Square will penalize the effect of additional variables which are not carrying much information. It should be always less than R Square.
* **AIC (Alkaline Information Criteria):** It adjusts the log likelihood based on the number of observation and complexity of the model.
* **BIC (Baisen Information Criteria):** It is similar to AIC but has high penalty for models.
* **Omnibus:** Provides combined statistical test for the presence of skewness and kurtosis. Basically it is breakdown of skewness and kurtosis.
* **Skew and Kurtosis:** These tests are basically for time series dataset.
* **Null Deviance:** It tells us how well the response variable is predicted by the model with intercept only.
* **Residual Deviance:** It tells us how well the response variable is predicted by using null deviance and all other independent variables.