ABSENTEESIM

**CONTENT:**

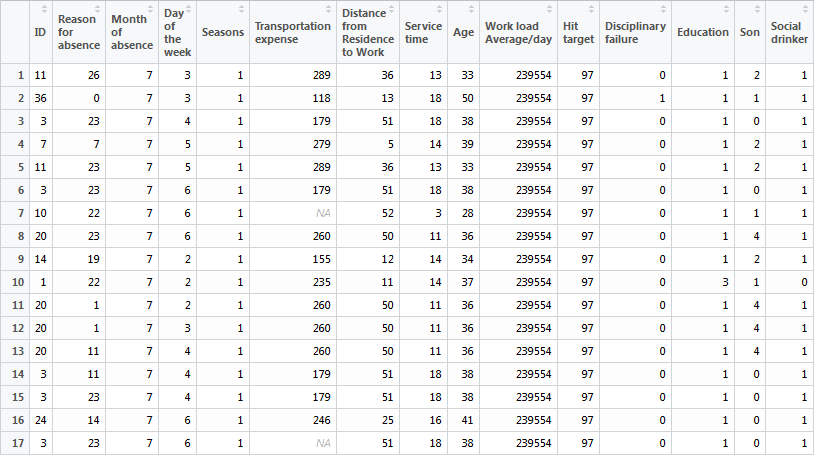
1. **Summary and Problem Statement**
2. **Data Overview**
3. **Feature Engineering** 
   1. **Data Conversion**
   2. **Outlier Analysis**
   3. **Data Imputation**
   4. **Correlation**
   5. **Scaling**
   6. **Sampling**
4. **Modelling** 
   1. **Linear model with R**
   2. **Linear model with python**
   3. **Cross Validation(Overfit or Underfit)**
   4. **Random Forest with R**
   5. **Random Forest with Python**
5. **Conclusion**
   1. **Model Evaluation**
   2. **Results Graphs**
   3. **Model Selection**

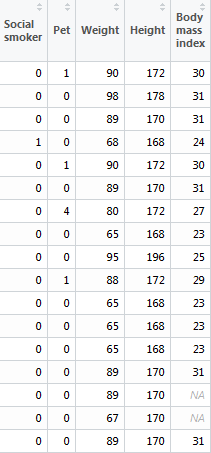
**1. Summary and Problem Statement:**

XYZ is a courier company. The human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The aim of the project is to suggest the company that what change company should bring to reduce the number of absenteeism and how many losses every month can company project in 2011 if same trend of absenteeism continues. We would like to predict the hours of absent based on the predictor variables.

**2. Data Overview:**

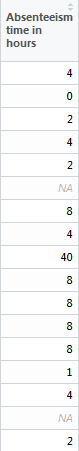
Our objective is to build the regression model which will be depending upon the various properties of the employ. Given below is sample of the data set that we are using to predict the hours of absenteeism in the company.

**Predictor Variable**

****

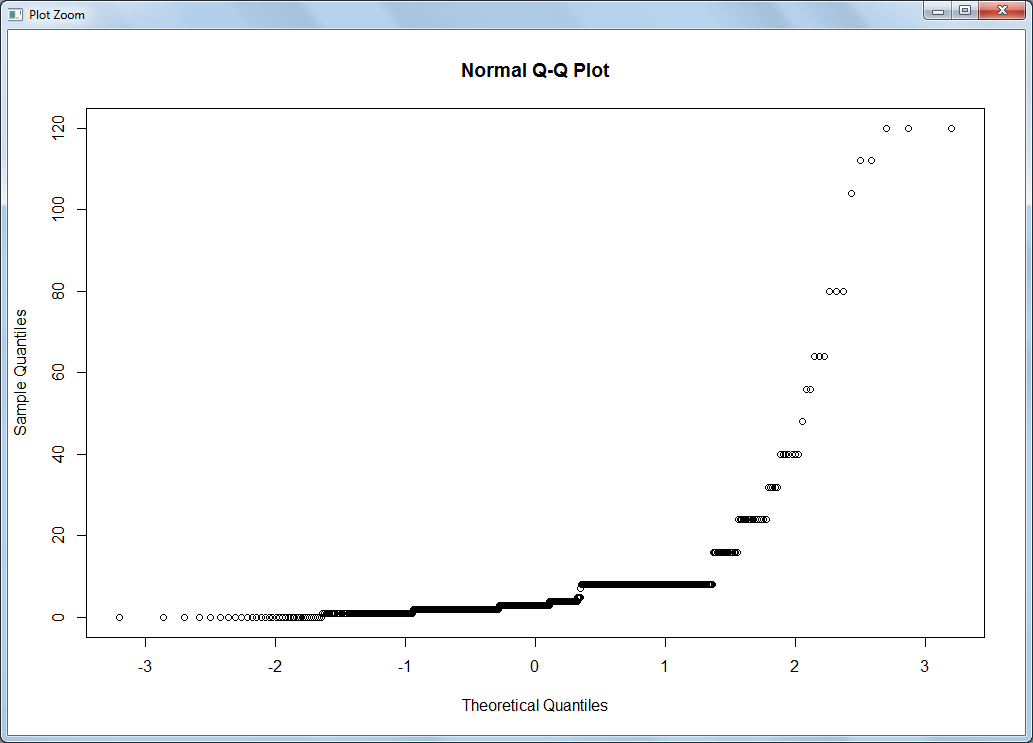
The above tables are the predictor variables which are used to predict the hours of absenteeism in the company. Below is the sample of dependent variable.

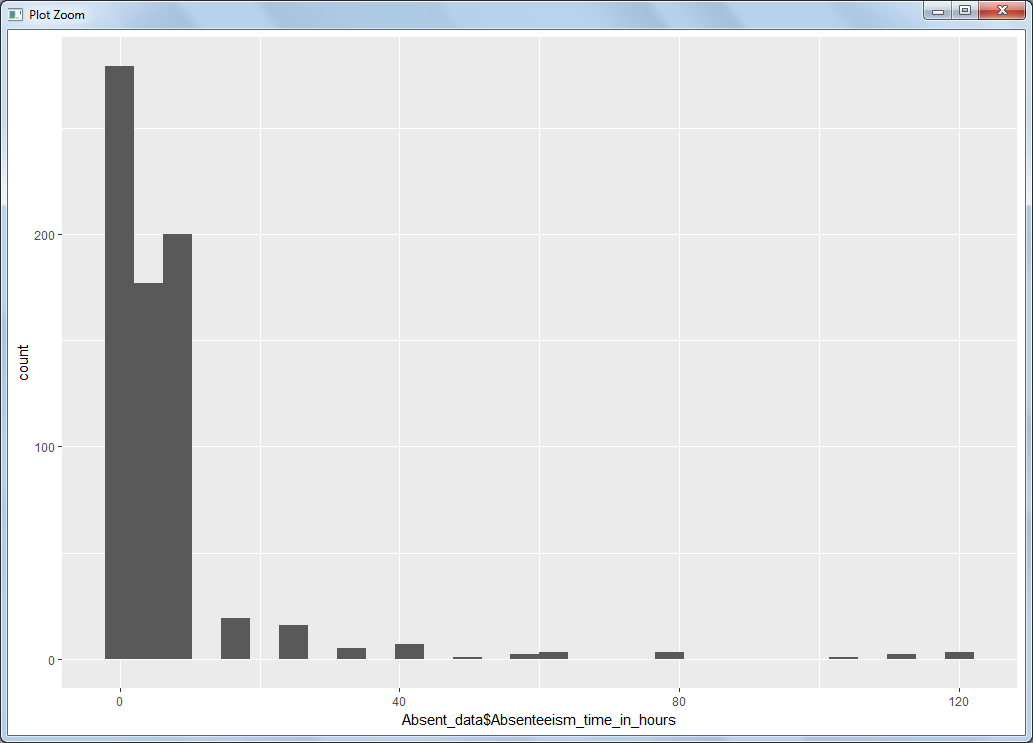
**DEPENDENT VARIABLE**

****

**DISTRIBUTION OF DEPENDENT VARIABLE**

Quantile wise distribution of dependent variable

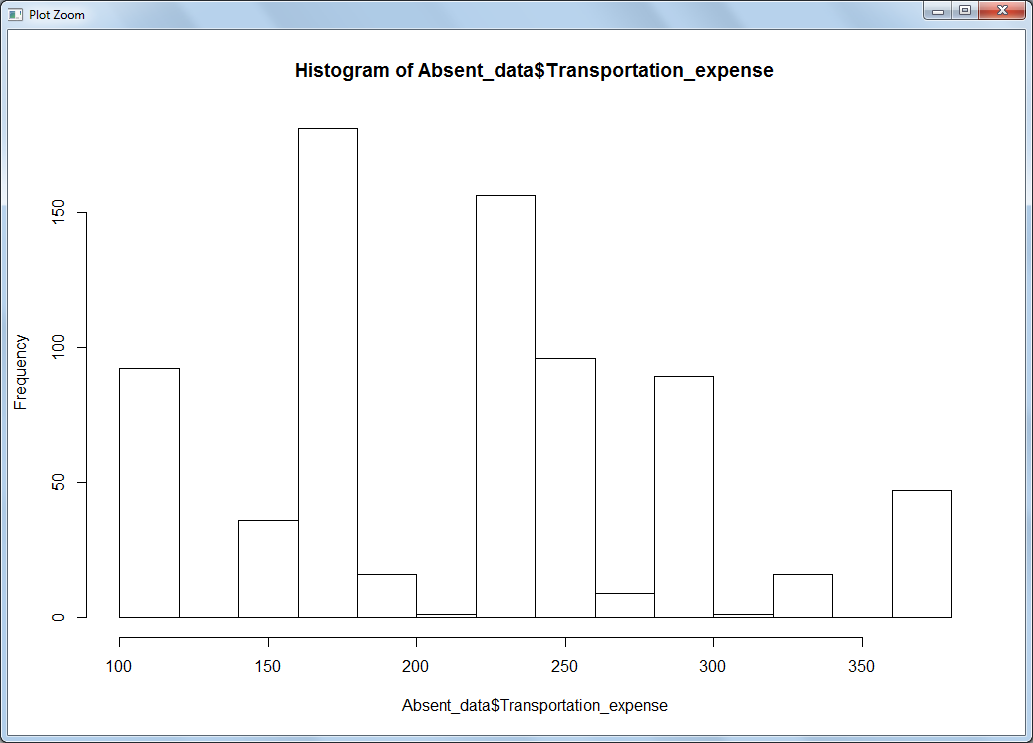




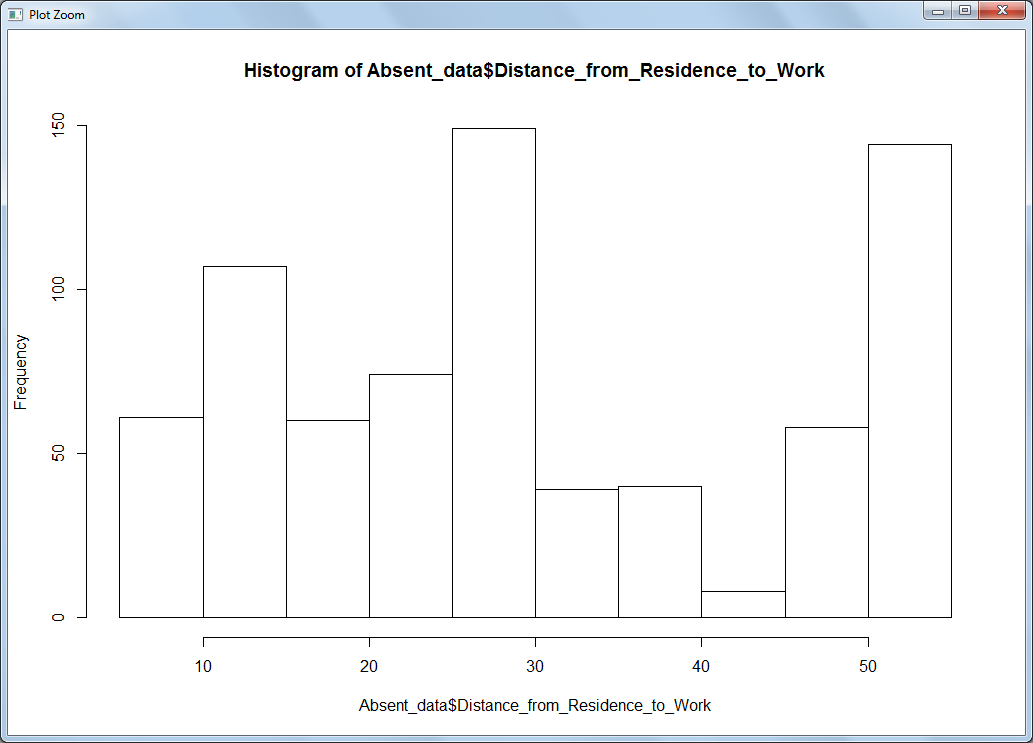
As we have seen the data is highly positively skewed and abnormally distributed.

Now we can check some other variables also and see their distribution in the data set

**Transportation expense**



**Distance from Residence to Work**



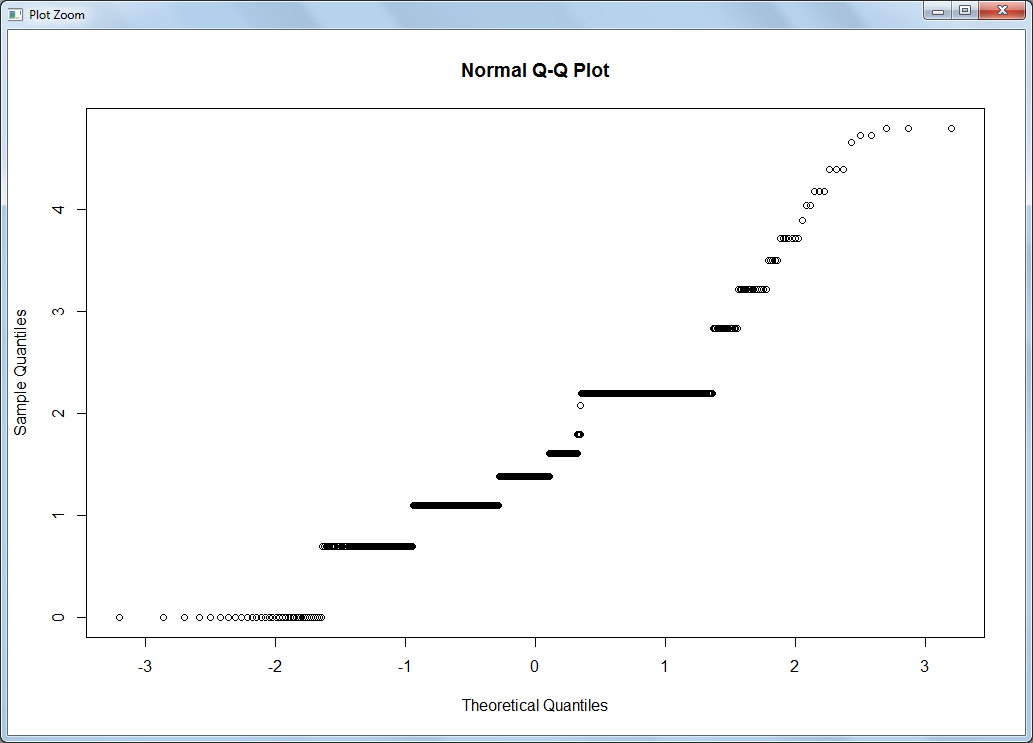
**3. Feature Engineering**

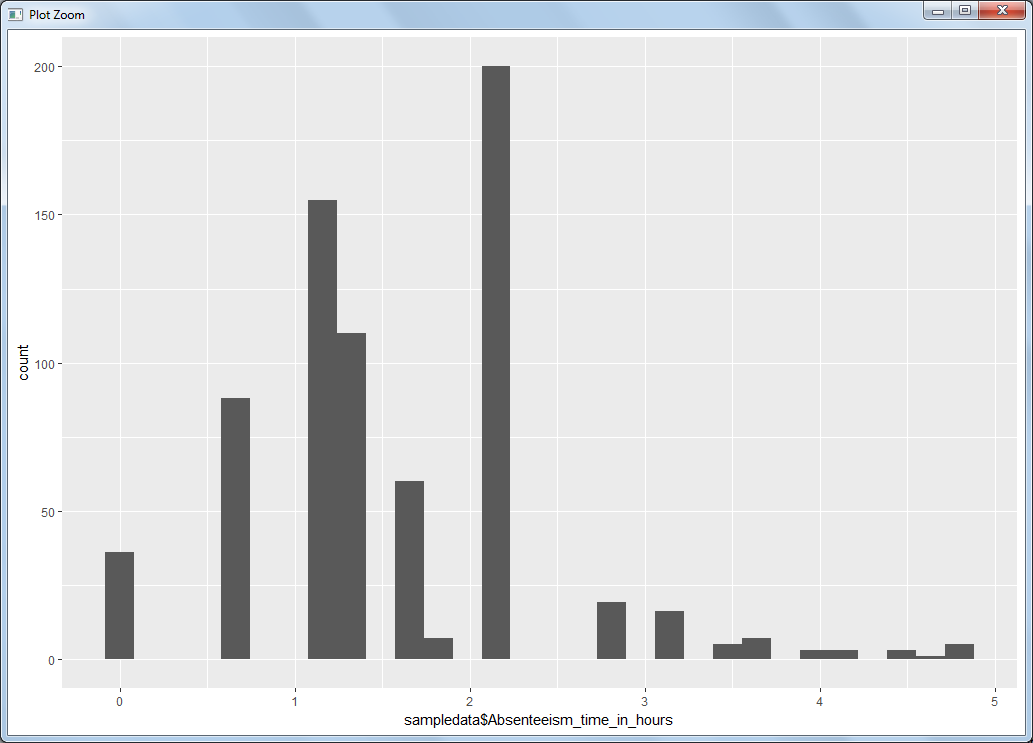
**3.1 Data Conversion**

a) Now we can see from above two graphs that our data is highly positively skewed.

We can transform our target variable by taking the log of target variable.

The residuals have a skewed distribution. The purpose of a transformation is to obtain residuals that are approximately symmetrically distributed the about zero.





Now we have seen that after taking the log of the dependent variable it became it slightly normally distributed.

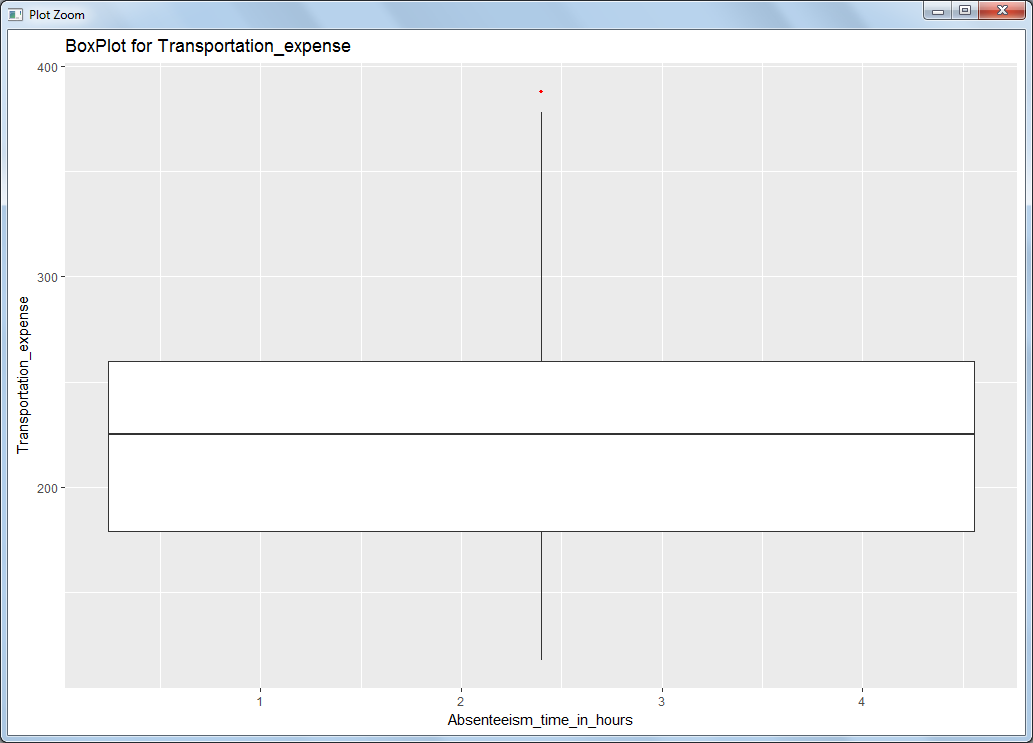
b) Convert the variable into suitable variable category for example Age is given as numerical variable but it has to be categorical.

c) Create the dummies for all the categorical values

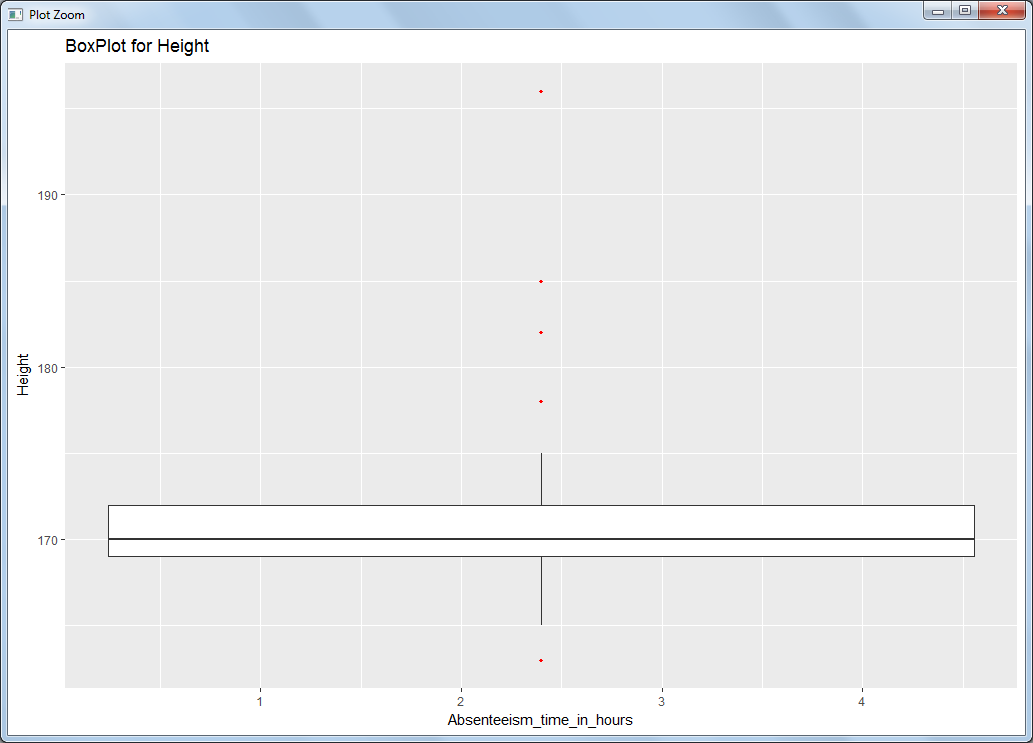
**3.2 Outlier Analysis**

Outlier is one of the steps of the data pre processing where we have identify the outlier and remove it. Box plot is one of the best ways to find out the outliers in the variable. Below are the outliers for the some of the continuous variable.

**Transportation Expense**



**Height**



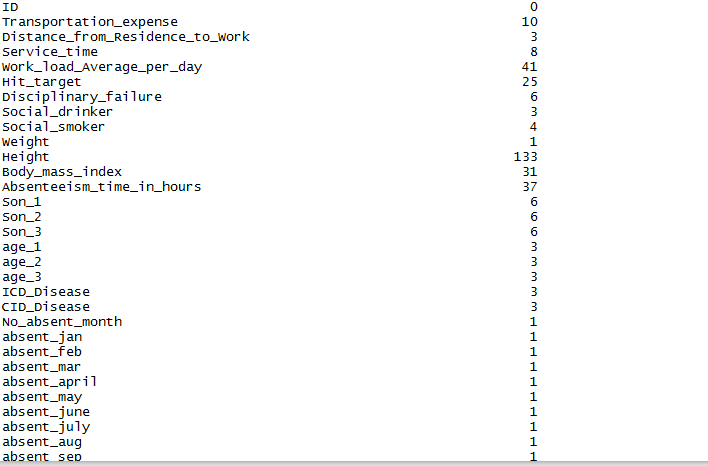
We have seen from the probability distribution of the variable that the data is skewed and now with the help of box plot we confirmed that there is outlier present in the some of the continuous variables. The data which is fall below or above the lower and upper fence is considers as outlier and should be removed as part of data pre processing.

As we don’t have lot of data and the percentage of the outlier data is significantly high we cannot simply delete the data.

We can replace the outlier which NA and later impute as missing value.

**3.3 Data Imputation**

Now we can check the total number of missing value for each of the variables.



From above table we can see that the there are missing value in many of the variables.

Impute the missing value using KNN imputation method.

**3.4 Correlation**

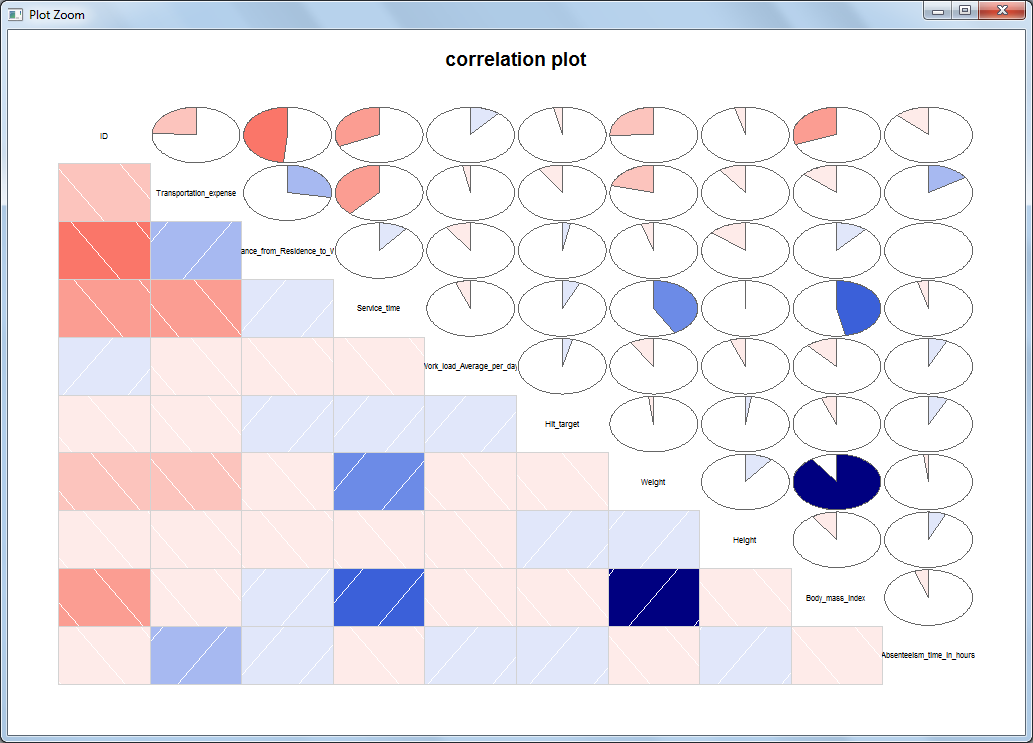
Now we have check for the correlations.

We have to make sure that independent variables should not be correlated, which means two different independent variable should not carry the same information, otherwise it cannot the explain the variance in the dependent variable.

We can drop any one of the such variable from the data set.

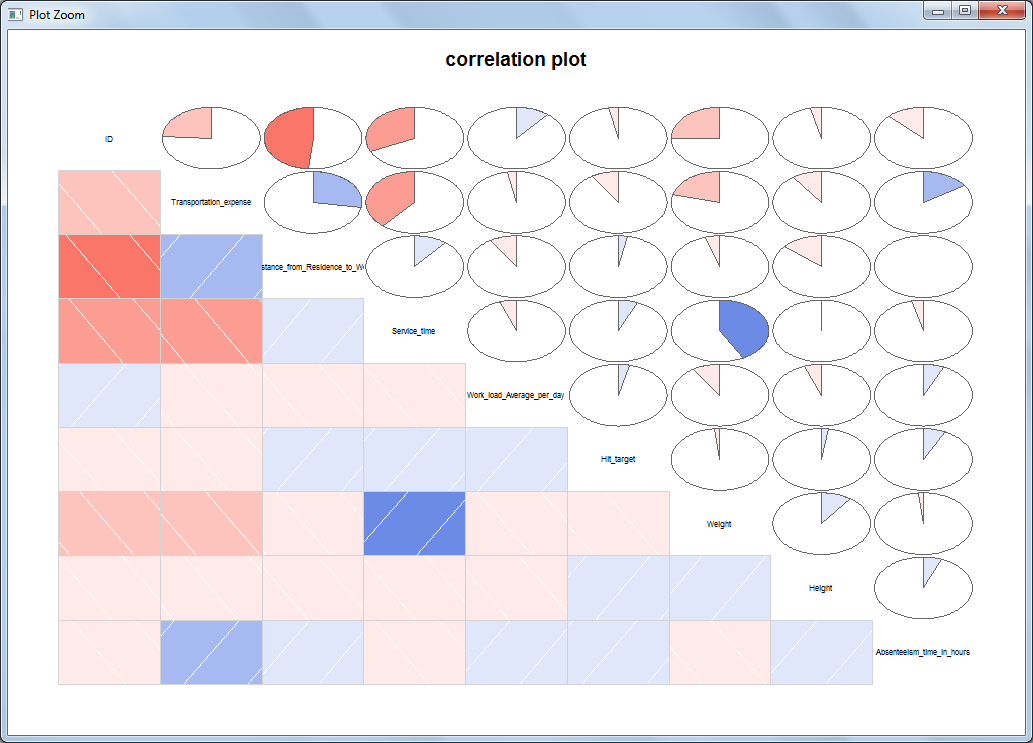
We will use the Scatter plot to visualize the correlation between the independent variables.

**CORRELATION PLOT**



We can see from above plot that body mass index and weight is highly positively correlated.

We can drop any one of the variable.



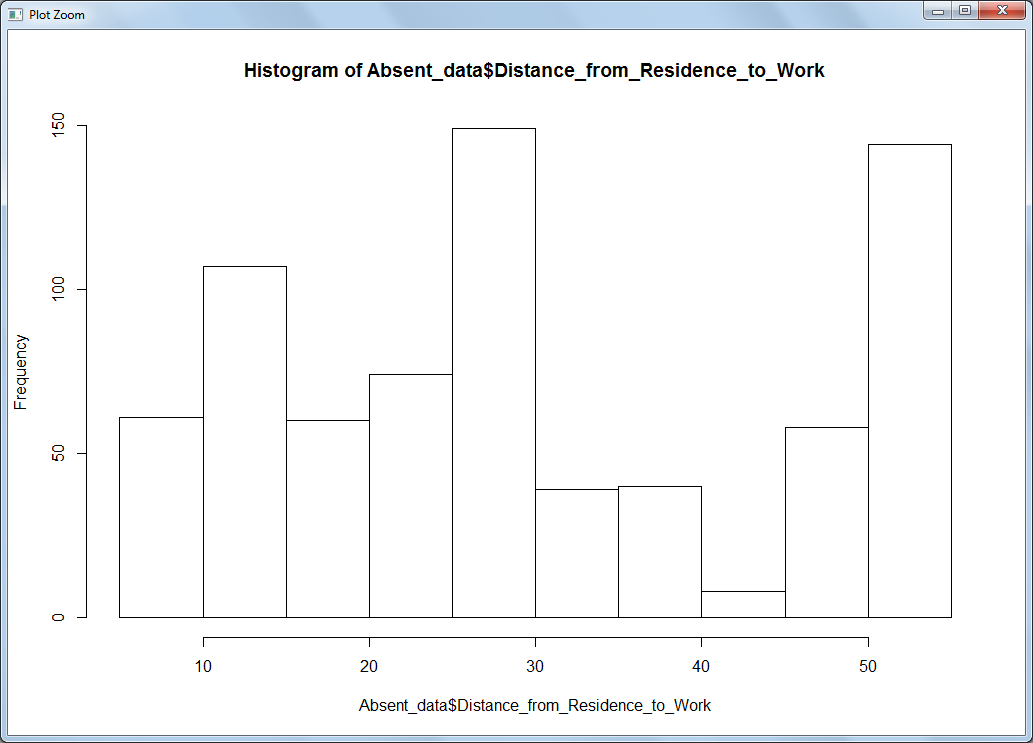
Now after dropping the body mass index we can see that there are no other correlated

Variables

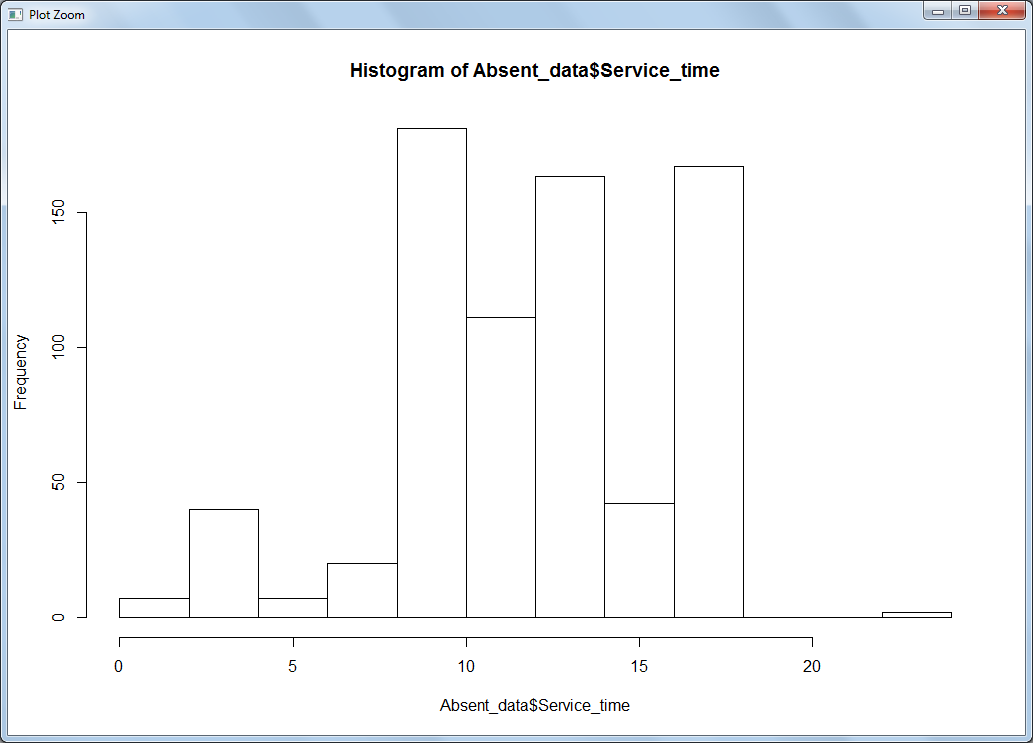
**3.4 Scaling**

Under the scaling we are going to use the normalization technique. As we have observed that the all the variable are of different scale. Please see below few of the independent variables.

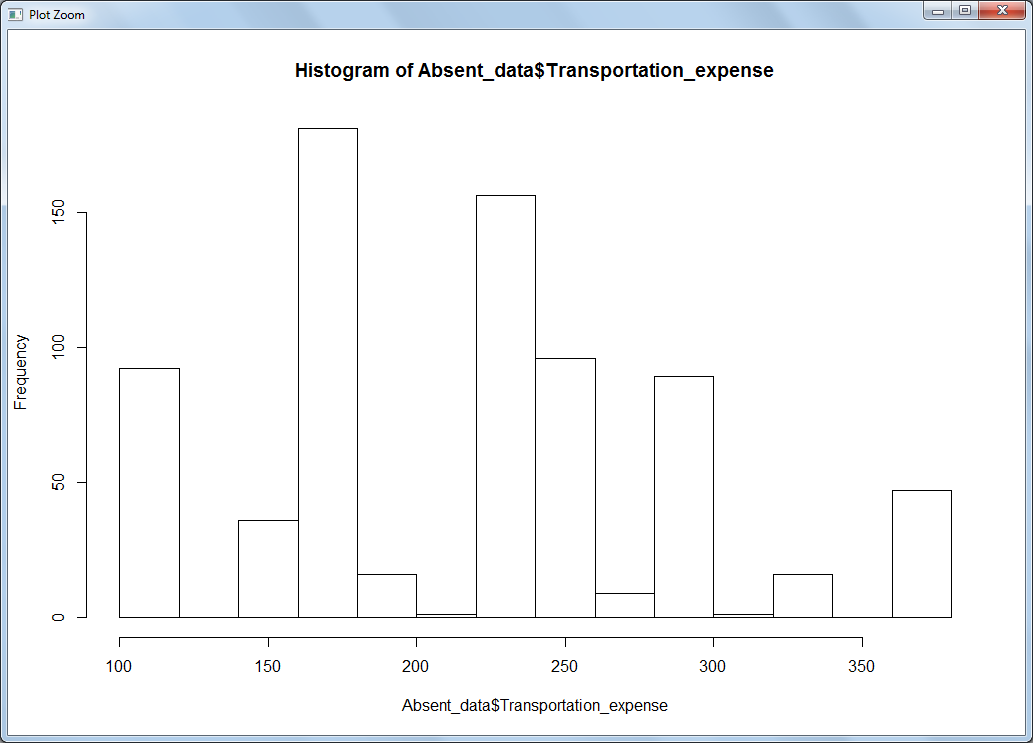
**Distance from residence to work**



**Service Time**



**Transportation Expense**



As seen from the above graphs, independent variables are of different scale. All the variables should be equal scale for the better predictions. If there are variables with high scale the output will be biased toward it. In normalization we convert all the observation into range of 0 to 1.

**Formula Used:**

Absent\_data[,i] = (Absent\_data[,i] - min(Absent\_data[,i]))/

(max(Absent\_data[,i] - min(Absent\_data[,i])))

**3.6 Sampling**

Sampling is the technique of dividing the data in training and testing data set. Let say we divide out data in ratio of 70:30 for training and testing data set

Training Data Set: Training data set is that data upon which we create out model.

Testing Data Set: Testing data set is that data set upon which we validate our model by means of accuracy matrix.

Here, we have used simple random sampling technique.

**4 Modelling:**

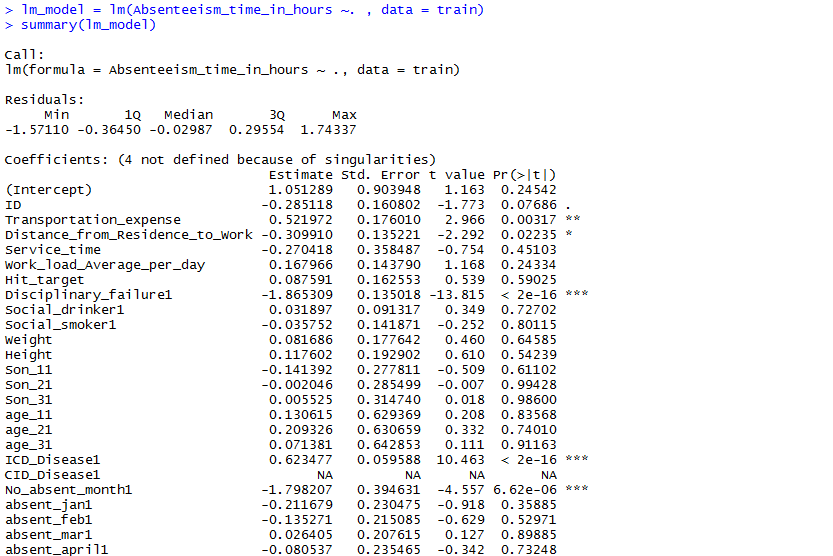
Now we are done with data pre processing. The dependent variable in our data set is a continuous variable therefore we have to build the regression model for that.

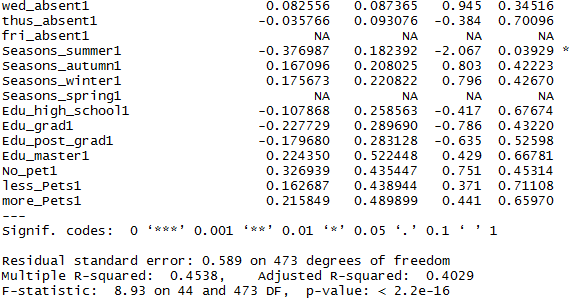
**1. Linear Model with R**

We will create the model by keeping all the independent variables into the model.

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

summary(lm\_model)





We can see that there are lots of variables which are not significant and will lead to the errors in the predictions. Star marked variables are most significant variables.

We will create another model with all the significant variables.

lm\_model = lm(Absenteeism\_time\_in\_hours ~ ID +Transportation\_expense + Disciplinary\_failure+

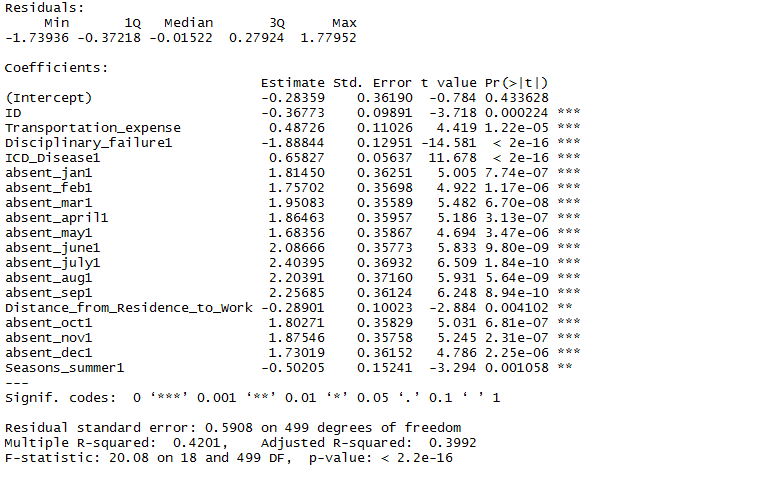
ICD\_Disease + absent\_jan + absent\_feb + absent\_mar+

absent\_april + absent\_may + absent\_june +

absent\_july + absent\_aug + absent\_sep +Distance\_from\_Residence\_to\_Work+

absent\_oct + absent\_nov + absent\_dec + Seasons\_summer + No\_absent\_month

, data = train)



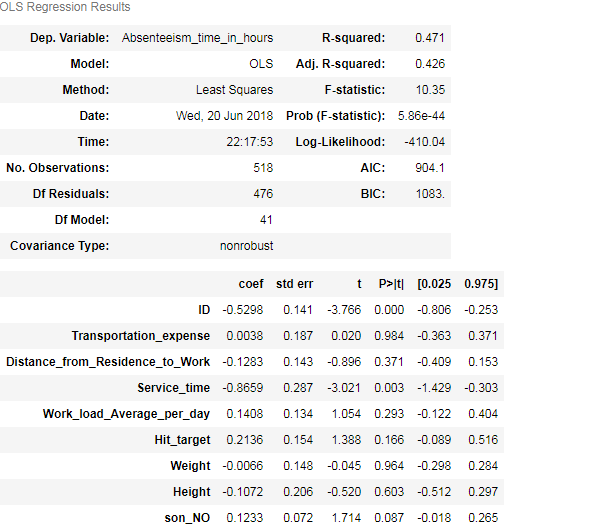
Looking at the significance values of some of the predictor we can see that there is some scope of improvement in this model.

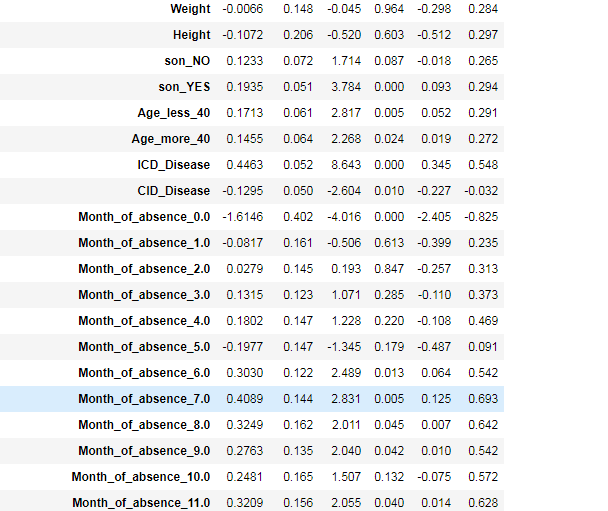
**2. Linear Model with Python**

We will create the model by keeping all the independent variables into the model.

lm\_model = sm.OLS(train.iloc[:,0],train.iloc[:,1:53]).fit()

lm\_model.summary()





Now we will select the significant variable and build the model.

**4.3 Cross Validation(Overfit or Underfit)**

NOTE: I have followed the same procedure for R and Python thats why not writing cross validation for separately but performed seprately

After building the model we have to check for the overfiiting and underfitting of the model

predict result with training dataset independent variable and then calculate error between training dataset actual value and predicted value.

do same thing with test dataset.

now if training dataset has very less error and test dataset has very high error , it means your model has **overfit** the data

if training dataset and test dataset error both are high than it is case of **bias/underfit**.

both error should be less and approx same(obv training error would be less than test error)

In our case we have found that both the training and testing data has high error rate. Check the below code:

#Make predictions with train data

prediction\_lm\_cv = predict(lm\_model\_cv,train[,-1])

#Calculate MAPE for train data set

MAPE(train[,1],prediction\_lm\_cv)

0.3558497

#ERROR on Training data : 35.5%

#Make predictions with test data

prediction\_lm\_cv\_test = predict(lm\_model\_cv,test[,-1])

#Calculate MAPE for test data set

MAPE(test[,1],prediction\_lm\_cv\_test)

0.3682947

#ERROR on Testing data : 36.8%

So, we decided to split the data more into ratio.

**4.4 Random Forest with R**

Random forest is tree based algorithm. For building the model with random forest we have taken the split of the data after cross validation.

First we have tune the mtry and found out the that mtry 16 is best fitted for the model and then create the model using RandomForest function

rf\_improve = randomForest(Absenteeism\_time\_in\_hours ~., data = train\_cv,

ntree = 300,

mtry = 16,

importance = TRUE,

proximity = TRUE)

**4.5 Random Forest with Python**

For building the model with random forest in python also we have taken the split of the data after cross validation and build the model on top of that only.

rf\_model = RandomForestRegressor(n\_estimators = 200,oob\_score = True ,random\_state = 0,max\_features = 'auto')

rf\_model.fit(train\_improve\_cv.iloc[:,1:53],train\_improve\_cv.iloc[:,0])

**5. Conclusion**

**5.1 Model Evaluation**

MAPE and MSE are one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our models and generate the error and accuracy.

**Error metric before cross validation**

**Linear Regression R:**

**MAPE :**

prediction\_lm = predict(lm\_model,test[,-1])

MAPE(test[,1],prediction\_lm)

0.3682947

Error : 36%

Accuracy : 64%

**Random Forest R:**

(Before and after will be same because already taken the train and test data after cross validation)

rf\_improve\_predict = predict(rf\_improve,test\_cv[,-1])

MAPE(test\_cv[,1],rf\_improve\_predict)

0.2723119

Error : 27%

Accuracy : 73%

**Linear Regression Python:**

lm\_prediction\_improve = lm\_model\_improve.predict(test\_improve)

**MAPE**

MAPE(test.iloc[:,0],lm\_prediction\_improve)

38.40149620576362%

Error : 38%

Accuracy : 62%

**MSE**

0.4079222686560284

Error : 40%

Accuracy : 60%

**Random Forest Python:**

(Before and after will be same because already taken the train and test data after cross validation)

Prediction = rf\_model.predict(test\_improve\_cv.iloc[:,1:53])

**MAPE:**

MAPE(test\_improve\_cv.iloc[:,0],Prediction)

8.643080083059074

ERROR : 8.64%

Accuracy: 91.36%

**MSE:**

mean\_squared\_error(test\_improve\_cv.iloc[:,0],Prediction)

0.006819425951461669

ERROR : 0.06%

Accuracy: 99.94%

**Error metric after cross validation**

**Linear Regression In R**

prediction\_lm\_new = predict(lm\_model\_cv\_train,test\_cv)

MAPE

MAPE(test\_cv[,1],prediction\_lm\_new)

0.2832943

MAPE ERROR : 28%

ACCURACY : 72%

**Linear Regression In Python**

**MAPE:**

lm\_prediction\_train\_cv = lm\_model\_train\_cv.predict(test\_improve\_cv)

MAPE(test\_improve\_cv.iloc[:,0],lm\_prediction\_train\_cv)

1.9274833148301912e-13

ERROR : 1.9274833148301912e-13

ACCURACY : 99.999999%

**MSE:**

mean\_squared\_error(test\_improve\_cv.iloc[:,0], lm\_prediction\_train\_cv)

6.052269135761492e-31

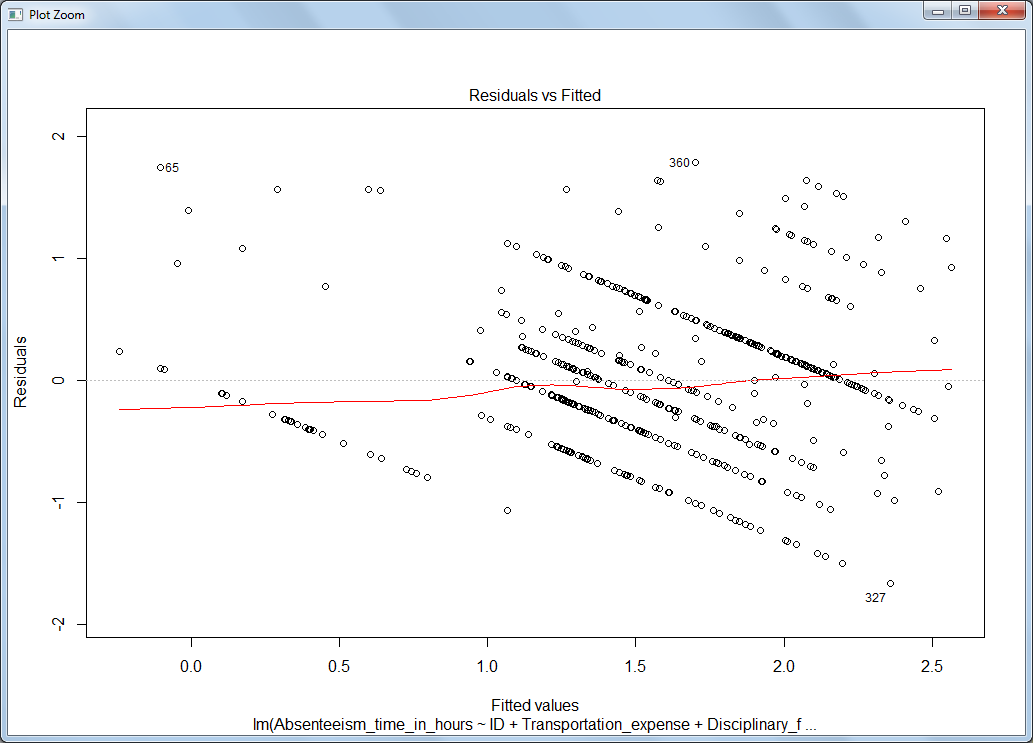
ERROR : 6.052269135761492e-31

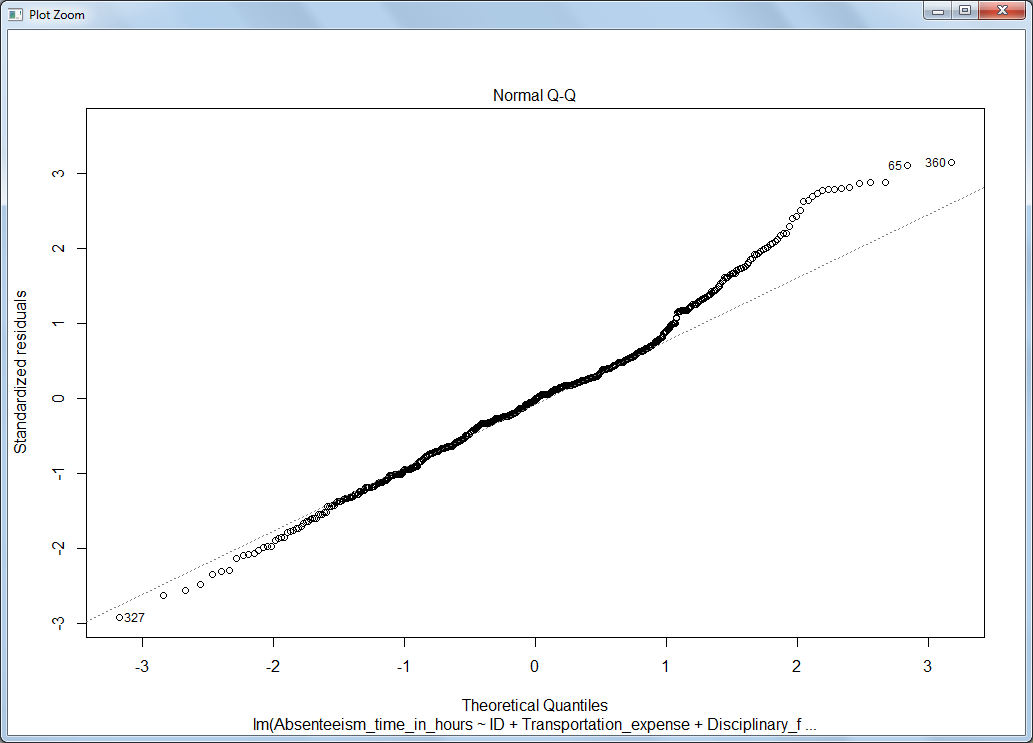
ACCURACY : 99.999999%

**5.2 Results Graphs**

GRAPH FOR LINEAR REGRESSION

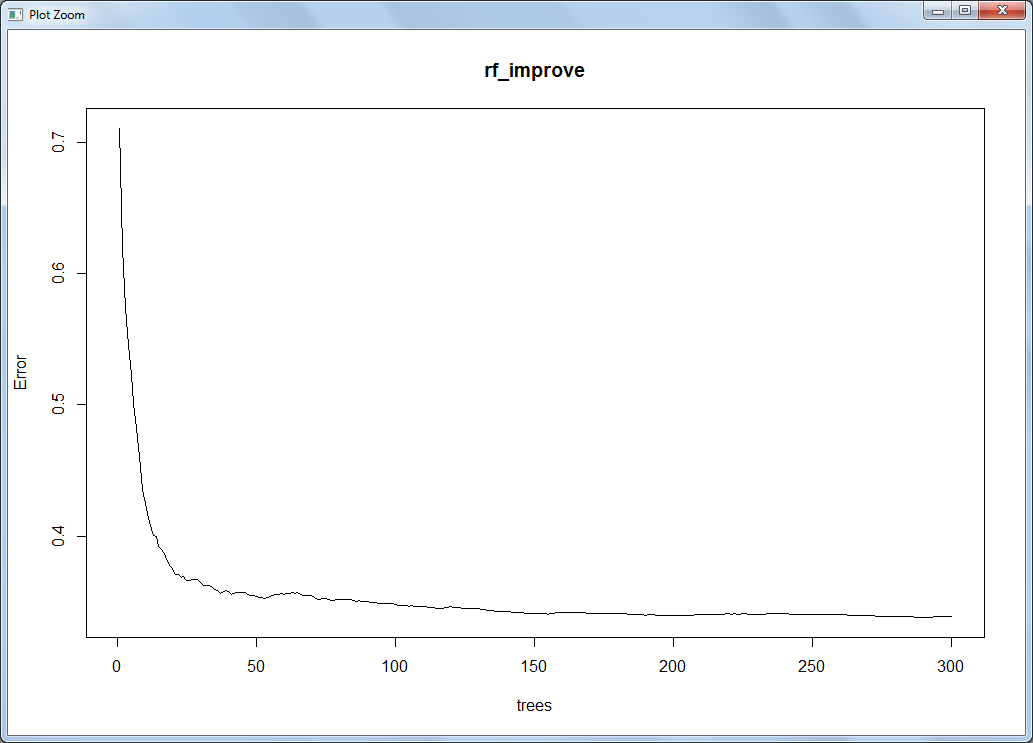
From below two graphs you can see that there is very less hetroskedasticity and error are normally distributed.





GRAPH FOR RANDOM FOREST

From the below graph we can see that as the number of trees increase the error became less.



**5.3 Model Selection**

We can see that the In R Random forest is working fine comparatively linear regression and in python linear regression works fine comparatively random forest.

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

As we have seen from the linear regression and random forest there are certain factor which company has keep in mind before hiring the new staff from going onwards.

a) Employee should be of nearby area so that expense cost of the travelling will not a issue for him.

b) Summer time is holiday time for employee children,they might want to spend time with their children. So company should not over burden them on summer’s time.

c) Company should be more careful about the ICD disease. Every employ should go for complete medical check up before hiring.

2) How much losses every month can we project in 2011 if same trend of absenteeism continues?

I have create a new data with predicted absent hours with that we can get to know about the absent hours going to be in 2011.