Project Report

Employee absenteeism

*Madhur Dhawan*

*4 jan 2019*

**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 details of data attributes in the dataset are as follows –**

* Individual identification (ID)
* Reason for absence (ICD).
  + Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:
    - Certain infectious and parasitic diseases
    - Neoplasms
    - Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
    - Endocrine, nutritional and metabolic diseases
    - Mental and behavioural disorders
    - Diseases of the nervous system
    - Diseases of the eye and adnexa VIII Diseases of the ear and mastoid process
    - Diseases of the circulatory system
    - Diseases of the respiratory system
    - Diseases of the digestive system
    - Diseases of the skin and subcutaneous tissue
    - Diseases of the musculoskeletal system and connective tissue
    - Diseases of the genitourinary system
    - Pregnancy, childbirth and the puerperium
    - Certain conditions originating in the perinatal period
    - Congenital malformations, deformations and chromosomal abnormalities
    - Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
    - Injury, poisoning and certain other consequences of external causes
    - External causes of morbidity and mortality
    - 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 (7)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 (kilometres)

\* 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

\* Absenteeism time in hours

**The company has shared it dataset and requested to have an answer on the following areas:**

In this dataset “**Absenteeism time In hours**” is our target varable which is continues varable so it is clear that this is regression problem.

**Check Type of each varable**

**Str(data)**

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 289 118 179 279 289 179 NA 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 ...

$ 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 ...

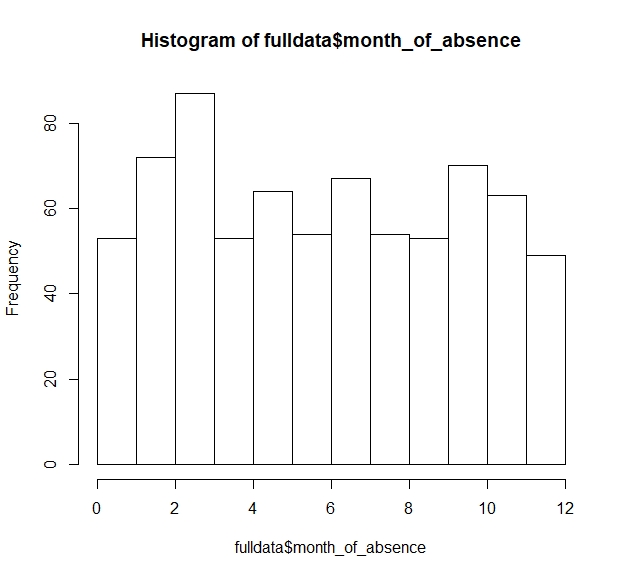
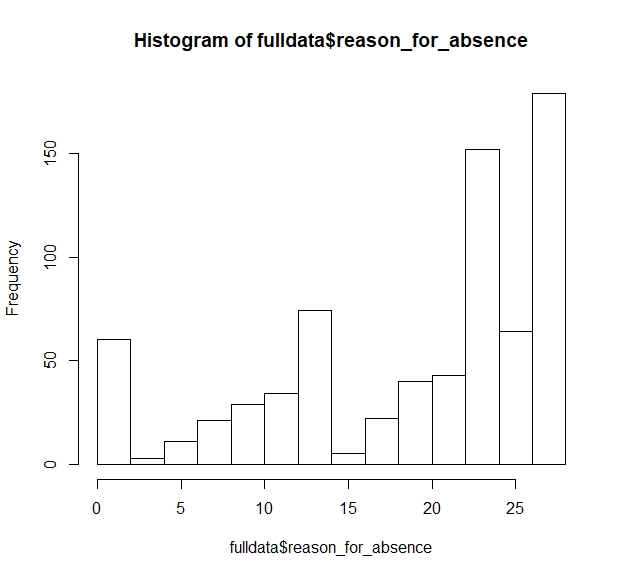
$ Pet : num 1 0 0 0 1 0 4 0 0 1 ...

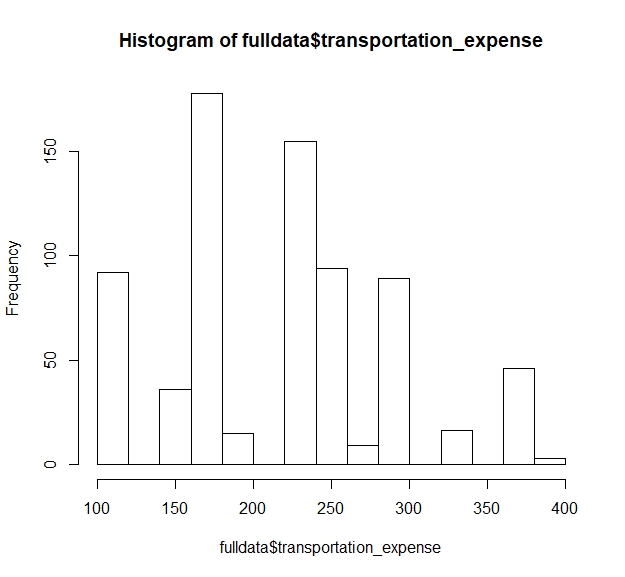
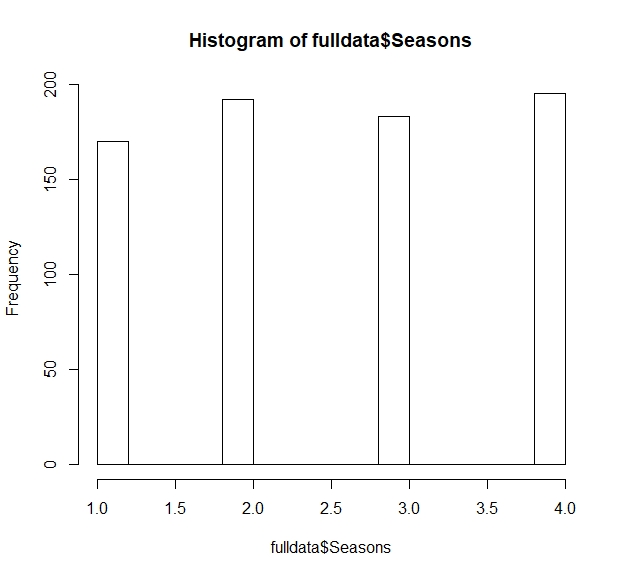
$ Weight : num 90 98 89 68 90 89 80 65 95 88 ...

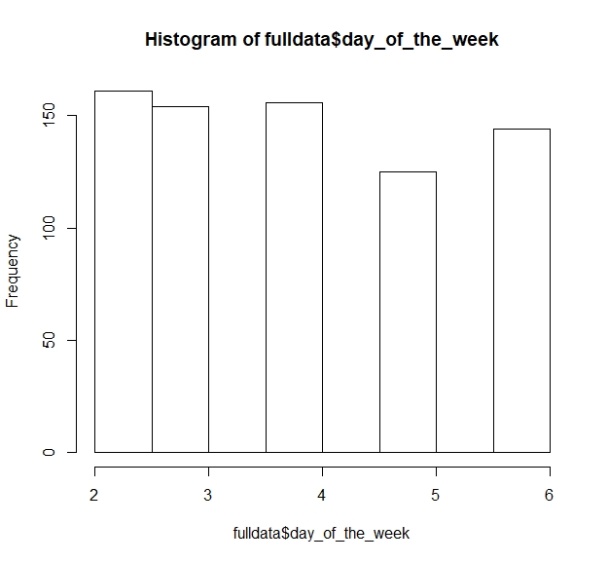
$ Height : num 172 178 170 168 172 170 172 172 ...

$ 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 ...

To understand distribution of numeric varables let us draw histogram of each varable:-





* It is clear from plot that varable **Reason of absence, month of absence,day of the week, seasons, education, disciplinary failure, social smoker, social drinker, pet, son,** are categorial type so convert it to factor type.
* Also clear that varable **id, transportation expenses, distance from residence to work, service time, age, working load average day, hit target** contains numeric data so there type will be numeric.

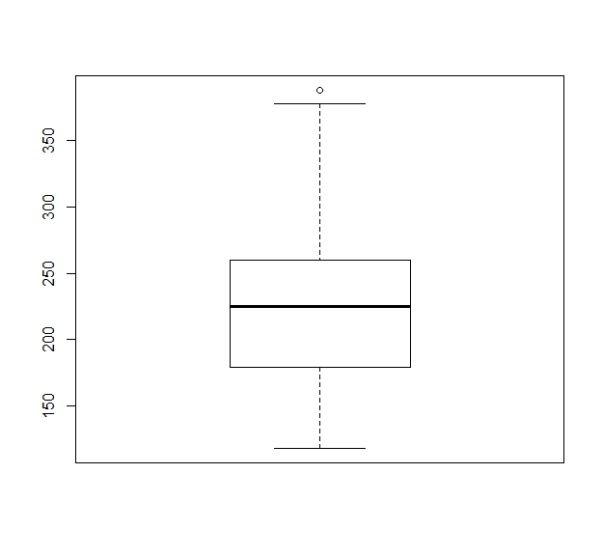
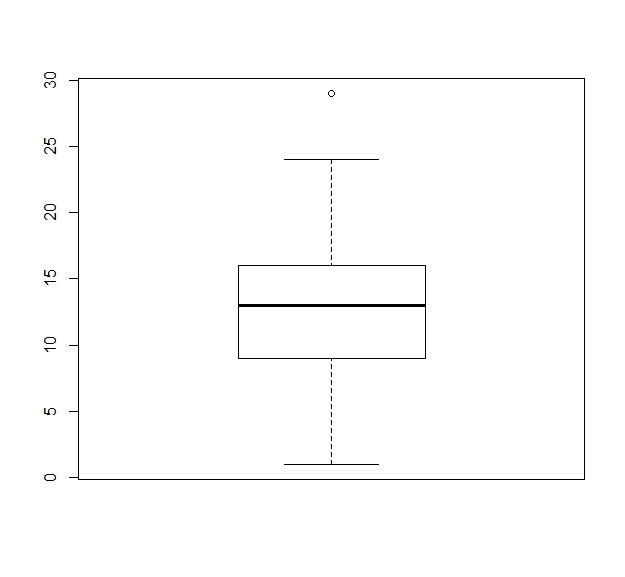
**Missing Value in dataset :-** First of all we have to check for missing values present in the dataset.

sum(is.na(day\_data)) 🡪 135

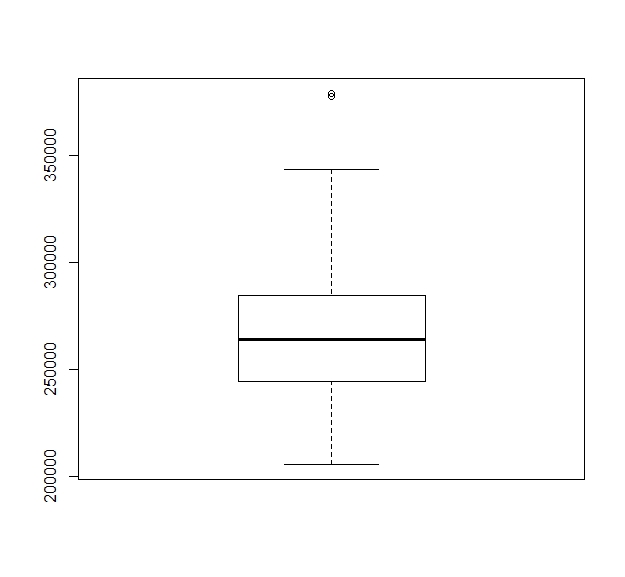
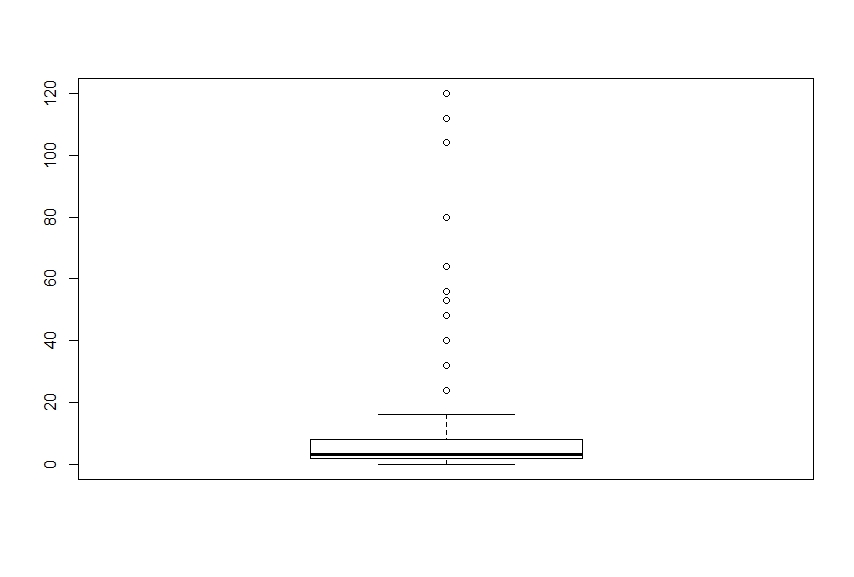
There are 135 values are missing in the whole dataset

So with the help of **knn** we will impute the missing values.

**Detecting Outlier in numeric varables :-** With the help of Boxplot we can detect the outlier present in dataset.



**Outlier in service time Outlier in work load average day**

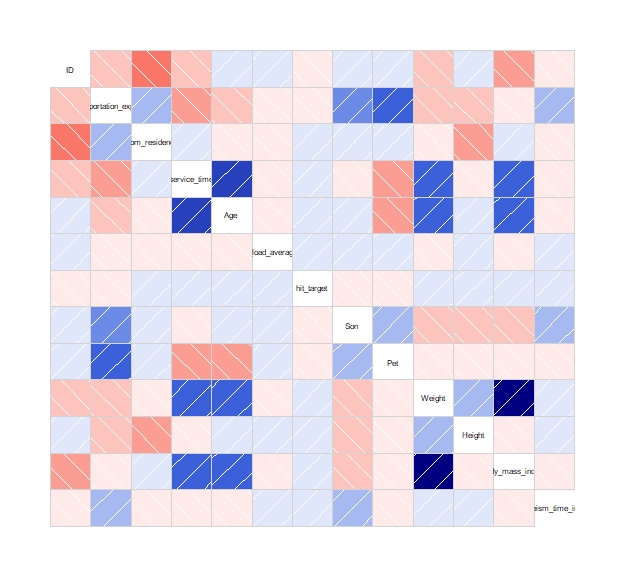
**Outlier in transportation expenses Outlier in Absentees varable**

* As it is clear from the boxplot that varable **service time**, **absentees**, **transportation expenses and work load** varable contains some outliers and to remove them replace it with the **maximum value and minimum value** of that particular varable.

**Features Selection among varables :-**  In this part we check which varable is contributing to tell about the target varable. This can be done in two parts :-

1. **For Continues varable :-** In this we will check only continues varables i.e we will select those varable who have highly correlate with target varable and less correlate with each other.

**Correlation plot is very helpful to check the correlation between to numeric varables**

****

It is clear from plot that varable **weight & body mass index**  are highly correlated.

so its better to remove one varable from the dataset.

Corr(weight,body mass index) 🡪 0.991

1. **For Catagorial Varables:-** In this we check the dependencies on categorial varables only. For this we will go with chi-square test.

In the chi-square test we compare the two varables in the contingency table to see if they are related to each other or not.

In this we consider two hypothesis ie:-

**Null Hypo**:- Two varables are independent

**Alternative Hypo**:- Two varables are not independent (Dependent)

If chi-squre value is greater than critical value then we have to reject the null hypothesis other wise acept it. We also consider p-value ie. If p-value is less than .05 then we reject the null hypothesis.

* After this step we drop **weight** varable because it is highly correlated with **body mass index** .

**Summary of Anova test:**

* Df Sum Sq Mean Sq F value Pr(>F)
* reason\_for\_absence 27 6111 226.33 19.058 < 2e-16 \*\*\*
* month\_of\_absence 12 153 12.73 1.072 0.38135
* day\_of\_the\_week 4 67 16.70 1.406 0.23013
* Seasons 3 98 32.57 2.743 0.04234 \*
* disciplinary\_failure 1 2 2.03 0.171 0.67914
* Education 3 26 8.83 0.743 0.52658
* social\_drinker 1 105 105.13 8.853 0.00303 \*\*
* social\_smoker 1 30 30.02 2.528 0.11231
* Residuals 687 8159 11.88
* ---
* Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Sampling Technique :-**

Divide the data into two parts **Train** and **Test.** Train part is used for train the model and we can check the accuracy of our model by applying the test part. We use **Simple Random Sampling** to divide the data into train and test.

So 80% of data is for training purpose and rest for testing purpose.

**Model Evaluation**

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models.

* **Decision Tree**

With using Algorithm **rpart**:-

n= 592

node), split, n, deviance, yval

\* denotes terminal node

1) root 592 11905.7200 5.131757

2) reason\_for\_absence=0,4,16,23,25,27,28 330 1723.4550 2.727273 \*

3) reason\_for\_absence=1,2,3,5,6,7,8,9,10,11,12,13,14,15,17,18,19,21,22,24,26 262 5871.2670 8.160305

6) reason\_for\_absence=1,3,5,6,7,8,10,11,13,14,15,17,18,21,22,24,26 222 4167.7120 7.630631

12) social\_drinker=0 93 1453.8280 6.376344

24) reason\_for\_absence=7,13,14,21,26 39 516.3077 4.692308 \*

25) reason\_for\_absence=1,3,5,6,8,10,11,17,18,22 54 747.0370 7.592593

50) month\_of\_absence=1,2,4,5,6,8,9,10,11 39 385.5897 6.564103 \*

51) month\_of\_absence=3,7 15 212.9333 10.266670 \*

13) social\_drinker=1 129 2462.0930 8.534884

26) reason\_for\_absence=1,3,5,6,8,11,14,15,22,24,26 75 749.5200 7.080000

52) service\_time>=17 13 109.6923 3.846154 \*

53) service\_time< 17 62 475.3710 7.758065 \*

27) reason\_for\_absence=7,10,13,18 54 1333.3330 10.555560

54) month\_of\_absence=1,2,3,7,8,10,12 30 714.9667 9.033333 \*

55) month\_of\_absence=4,5,6,9,11 24 461.9583 12.458330 \*

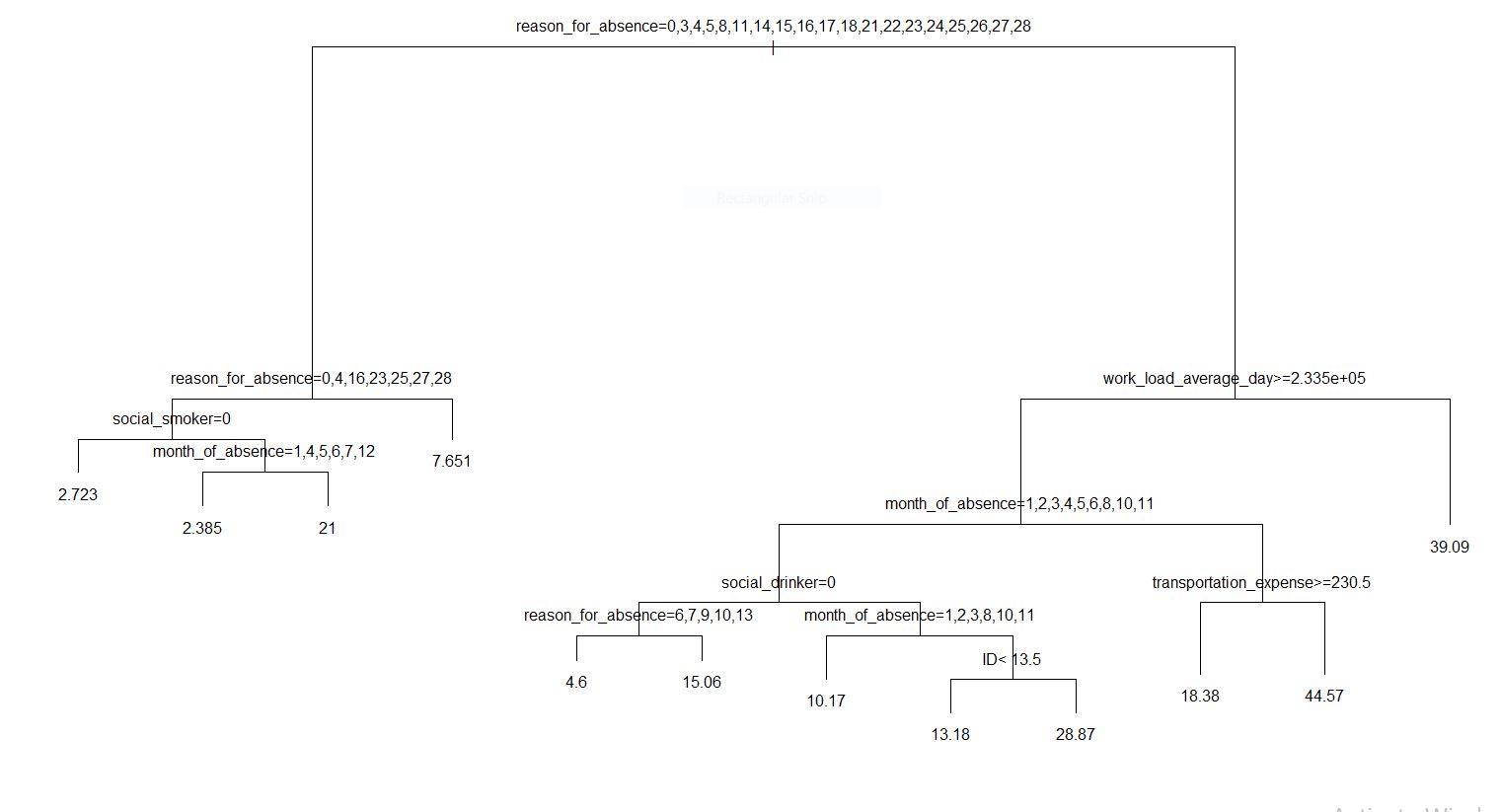
7) reason\_for\_absence=2,9,12,19 40 1295.6000 11.100000

14) hit\_target< 95.5 18 448.0000 8.333333 \*

15) hit\_target>=95.5 22 597.0909 13.363640

30) month\_of\_absence=2,3,4,5,11 12 379.0000 10.500000 \*

31) month\_of\_absence=1,6,7,9,12 10 1.6000 16.800000 \*



mape = 73.62 mae = 2.67

mse = 16.58 rmse = 4.07

* **Esambling Technique**

With using Algorithm **Randam forest**(ntrees=200)

**By extracting first two Rules from model :-**

[1] "reason\_for\_absence %in% c('0','16','27') & month\_of\_absence %in% c('0','1','2','4','5','6','7','8','9','10','11','12') & Age<=38.5 & Pet<=0.5 & Height<=174"

[2] "reason\_for\_absence %in% c('0','16','27') & month\_of\_absence %in% c('0','1','2','4','5','6','7','8','9','10','11','12') & Age<=38.5 & Pet<=0.5 & Height>174"

**Getting Rules Matric:-**

len freq err

[1,] "4" "0.003" "0"

[2,] "4" "0.01" "40.25"

**Condition**

[1,]"reason\_for\_absence %in% c('0','2','3','7','8','11','16','17','23','25','27','28') & distance\_from\_residence\_to\_work<=34 & service\_time<=5 & disciplinary\_failure %in% c('1')"

[2,]"reason\_for\_absence %in% c('0','2','3','7','8','11','16','17','23','25','27','28') & distance\_from\_residence\_to\_work<=34 & service\_time<=5 & disciplinary\_failure %in% c('0')"

pred

[1,] "1"

[2,] "8.5"

mape = 67.30 mae = 2.47

mse = 15.39 rmse = 3.92

**Regression Technique**

With using Algorithm **Linear regression**

**Vifcor()**

No variable from the 9 input variables has collinearity problem.

The linear correlation coefficients ranges between:

min correlation ( work\_load\_average\_day ~ transportation\_expense ): -0.002942905

max correlation ( body\_mass\_index ~ service\_time ): 0.5263395

---------- VIFs of the remained variables --------

Variables VIF

1 transportation\_expense 1.636249

2 distance\_from\_residence\_to\_work 1.344955

3 service\_time 1.828837

4 work\_load\_average\_day 1.027232

5 hit\_target 1.033659

6 Son 1.283265

7 Pet 1.336913

8 Height 1.241968

9 body\_mass\_index 1.492741

Summary of Linear Regression Model

Residuals:

Min 1Q Median 3Q Max

-8.0087 -1.3899 -0.2077 0.9329 14.2614

Coefficients:

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

(Intercept) -2.694e+01 1.271e+01 -2.120 0.034447 \*

ID -8.230e-02 2.187e-02 -3.763 0.000187 \*\*\*

reason\_for\_absence1 4.921e+00 3.447e+00 1.428 0.153976

reason\_for\_absence2 1.306e+01 4.685e+00 2.789 0.005482 \*\*

reason\_for\_absence3 5.925e+00 4.662e+00 1.271 0.204288

reason\_for\_absence4 2.648e+00 4.068e+00 0.651 0.515393

reason\_for\_absence5 3.130e+00 4.019e+00 0.779 0.436424

reason\_for\_absence6 4.225e+00 3.608e+00 1.171 0.242099

reason\_for\_absence7 4.921e+00 3.424e+00 1.437 0.151342

reason\_for\_absence8 3.582e+00 3.609e+00 0.993 0.321387

reason\_for\_absence9 1.371e+01 3.799e+00 3.610 0.000335 \*\*\*

reason\_for\_absence10 5.365e+00 3.375e+00 1.589 0.112577

reason\_for\_absence11 2.653e+00 3.373e+00 0.786 0.431942

reason\_for\_absence12 6.140e+00 3.520e+00 1.744 0.081705 .

reason\_for\_absence13 4.815e+00 3.330e+00 1.446 0.148757

reason\_for\_absence14 3.988e+00 3.424e+00 1.165 0.244625

reason\_for\_absence15 5.021e+00 3.998e+00 1.256 0.209718

reason\_for\_absence16 -2.155e-01 3.820e+00 -0.056 0.955037

reason\_for\_absence17 4.811e+00 4.680e+00 1.028 0.304441

reason\_for\_absence18 4.890e+00 3.389e+00 1.443 0.149628

reason\_for\_absence19 7.594e+00 3.333e+00 2.278 0.023108 \*

reason\_for\_absence21 3.366e+00 3.565e+00 0.944 0.345563

reason\_for\_absence22 4.448e+00 3.341e+00 1.331 0.183669

reason\_for\_absence23 4.561e-01 3.296e+00 0.138 0.889977

reason\_for\_absence24 4.501e+00 3.783e+00 1.190 0.234739

reason\_for\_absence25 8.772e-01 3.357e+00 0.261 0.793977

reason\_for\_absence26 3.555e+00 3.361e+00 1.058 0.290649

reason\_for\_absence27 3.318e-01 3.342e+00 0.099 0.920958

reason\_for\_absence28 -1.505e-01 3.299e+00 -0.046 0.963627

month\_of\_absence1 -6.056e-01 3.843e+00 -0.158 0.874852

month\_of\_absence2 -1.255e-01 3.825e+00 -0.033 0.973844

month\_of\_absence3 1.043e+00 3.826e+00 0.273 0.785269

month\_of\_absence4 5.238e-01 3.839e+00 0.136 0.891521

month\_of\_absence5 8.978e-02 3.830e+00 0.023 0.981306

month\_of\_absence6 9.383e-01 3.824e+00 0.245 0.806240

month\_of\_absence7 4.745e-01 3.839e+00 0.124 0.901675

month\_of\_absence8 5.862e-01 3.847e+00 0.152 0.878951

month\_of\_absence9 1.050e+00 3.768e+00 0.279 0.780571

month\_of\_absence10 7.467e-01 3.840e+00 0.194 0.845883

month\_of\_absence11 2.251e-01 3.826e+00 0.059 0.953118

month\_of\_absence12 6.825e-01 3.842e+00 0.178 0.859072

transportation\_expense 7.732e-03 2.981e-03 2.594 0.009749 \*\*

distance\_from\_residence\_to\_work -3.164e-02 1.316e-02 -2.405 0.016514 \*

service\_time 5.265e-02 5.829e-02 0.903 0.366770

Age -3.473e-02 3.827e-02 -0.907 0.364614

work\_load\_average\_day 1.012e-06 4.648e-06 0.218 0.827726

hit\_target 2.791e-02 5.144e-02 0.542 0.587719

disciplinary\_failure1 -1.686e+00 3.336e+00 -0.505 0.613543

Education2 -1.946e+00 7.466e-01 -2.606 0.009414 \*\*

Education3 -2.918e+00 6.869e-01 -4.249 2.54e-05 \*\*\*

Education4 -3.703e-01 1.777e+00 -0.208 0.835013

Son 3.896e-01 1.618e-01 2.407 0.016416 \*

social\_drinker1 -1.226e+00 5.583e-01 -2.196 0.028494 \*

social\_smoker1 -6.146e-01 6.584e-01 -0.933 0.350987

Pet -9.677e-01 2.580e-01 -3.751 0.000196 \*\*\*

Height 1.811e-01 6.752e-02 2.682 0.007549 \*\*

body\_mass\_index -7.739e-02 5.328e-02 -1.452 0.146958

---

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

Residual standard error: 3.204 on 535 degrees of freedom

Multiple R-squared: 0.5021, Adjusted R-squared: 0.45

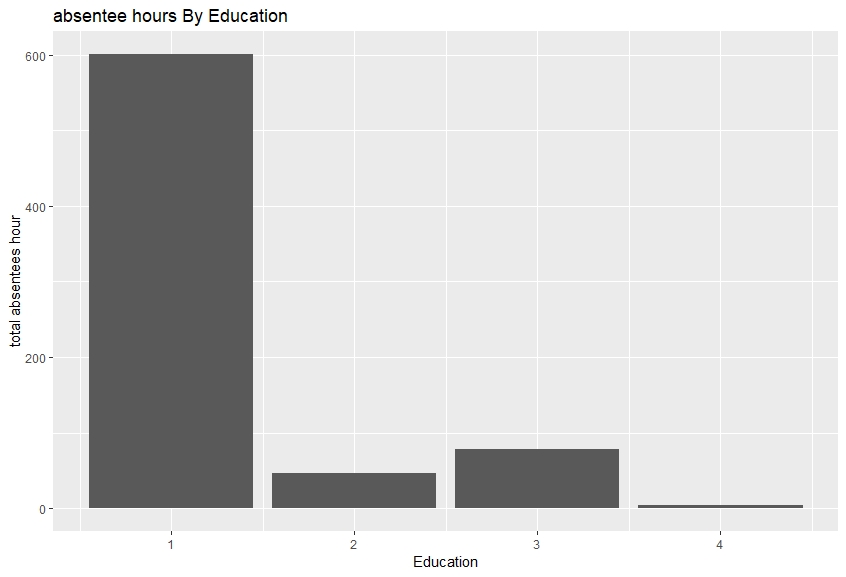
F-statistic: 9.635 on 56 and 535 DF, p-value: < 2.2e-16

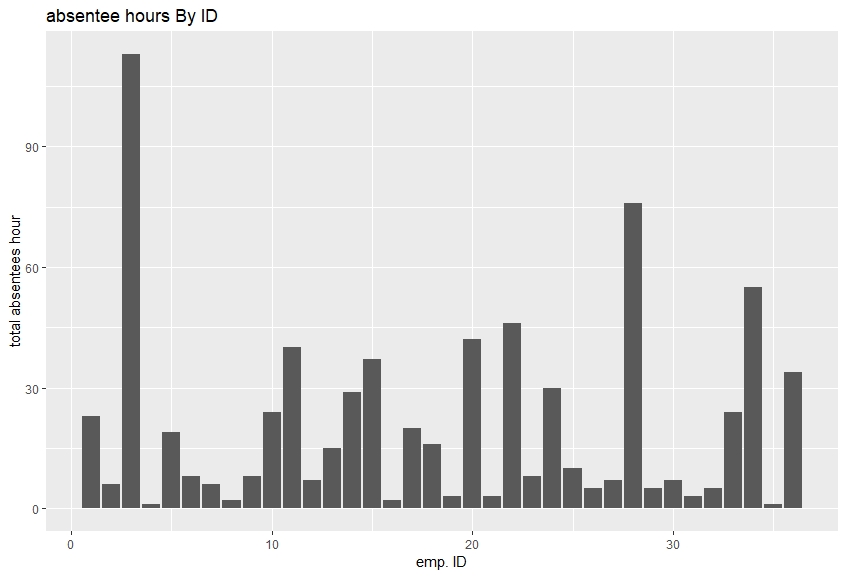
mape = 70.05 mae = 2.71

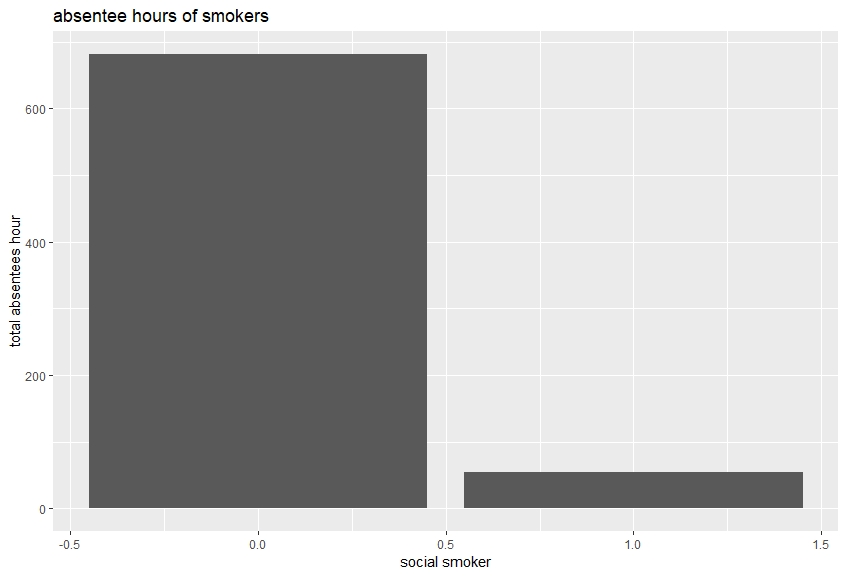
mse = 17.67 rmse = 4.20

**Answers of asked questions**

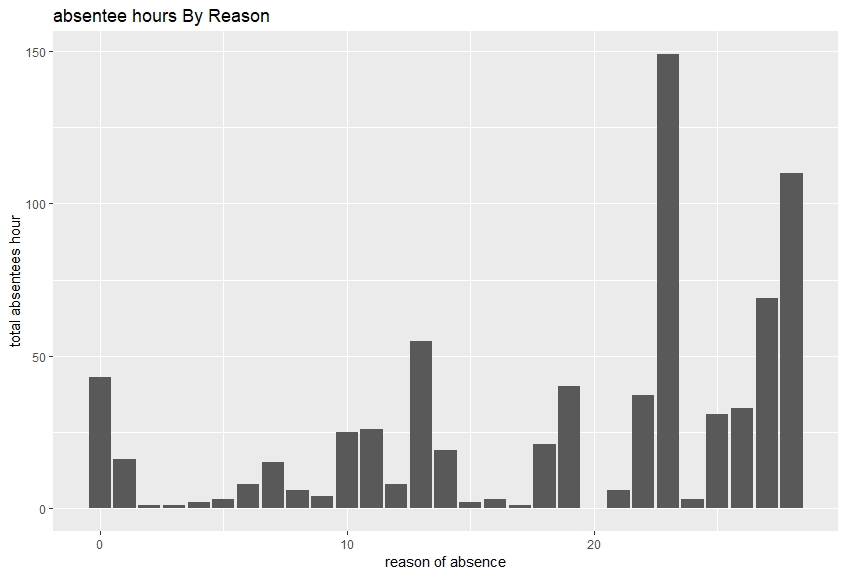
1. It is observed that employee with low education have maximum absentee time.



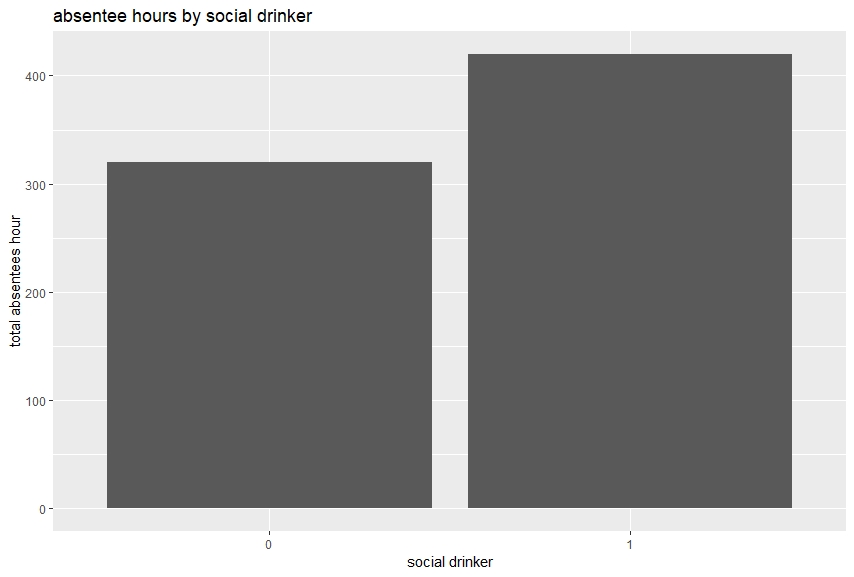
1. Some employee with **ID 3, 28, 34** are often absent from work as it is shown****
2. Employees who are social smoker have more absentee hour than who are not social smoker.

****

1. Most often Reason for absence are medical consultation and dental consultation.

****

1. Employees who are social drinker have more absentee hour than who are not social drinker.

****