Predicting Absenteeism Time

Goldie Sahni

3 September 2018

Contents

1 Problem Definition & Data Description

1.1 Problem Statement..2

1.2 Data Description..2

Modelling in Python

2 Exploratory Data Analysis

2.1 Missing Values Analysis..5

2.2 Distributions...9

2.3 Outlier Analysis..13

2.4 Correlation Analysis...16

2.5 Relationships of categorical independent variables with dependent variable...18

2.6 Relationships of continuous independent variables with dependent variable...24

2.7 Conclusions & Remedies..24

3 Modelling

3.1 Time Series - Arima Model..25

Modelling in R

4 Exploratory Data Analysis

4.1 Missing Values Analysis..32

4.2 Distributions...36

4.3 Outlier Analysis..41

4.4 Correlation Analysis...45

4.5 Relationships of categorical independent variables with dependent variable...47

4.6 Conclusions & Remedies..48

5 Modelling

5.1 Time Series - Arima Model..49

Chapter 1

Problem Definition & Data Description

1.1 Problem Statement

Absenteeism leads to loss of productive work hours directly affecting the business of a company. Absenteeism can be due to medical reasons, personal reasons or unforeseen circumstances.

The absenteeism data in this problem is from a courier company.

This problem has two parts:

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?

1.2 Data Description

The data given has 21 variables. A sample is shown below:

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense
11	26	7	3	1	289
36	0	7	3	1	118
3	23	7	4	1	179
7	7	7	5	1	279
11	23	7	5	1	289

Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure
36	13	33	239,554	97	0
13	18	50	239,554	97	1
51	18	38	239,554	97	0
5	14	39	239,554	97	0
36	13	33	239,554	97	0

Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
1	2	1	0	1	90	172	30	4
1	1	1	0	0	98	178	31	0
1	0	1	0	0	89	170	31	2
1	2	1	1	0	68	168	24	4
1	2	1	0	1	90	172	30	2

Variable description is as under:

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)

Categorical independent variables are:

1. ID
2. Reason for absence
3. Month of absence
4. Day of the week
5. Seasons
6. Disciplinary failure (yes=1; no=0)
7. Education
8. Social drinker (yes=1; no=0)
9. Social smoker (yes=1; no=0)
10. Son
11. Pet

Continuous independent variables are:

1. Transportation expense
2. Distance from Residence to Work (kilometers)
3. Service time
4. Age
5. Work load Average/day
6. Hit target
7. Weight
8. Height
9. Body mass index

Absenteeism time in hours is the target or dependent variable.

Commas have been removed from 'Work load Average/day ' variable values.

Observations having 'Absenteeism time in hours' equal to zero have been removed.

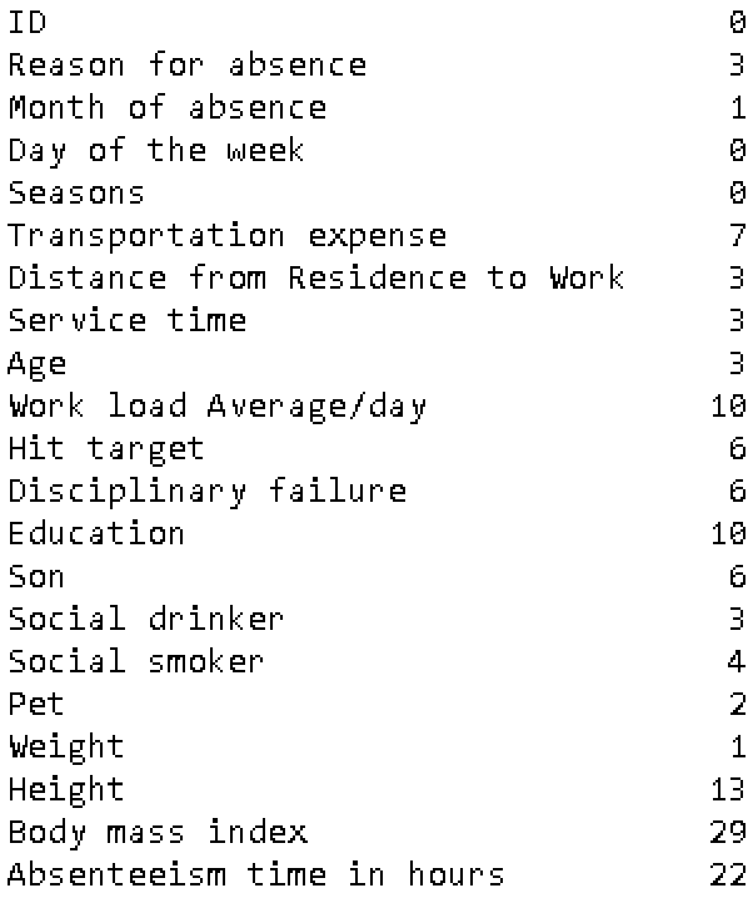
Modelling in Python

Chapter 2

Exploratory Data Analysis

2.1 Missing Values Analysis

Missing values of all variables are as under:

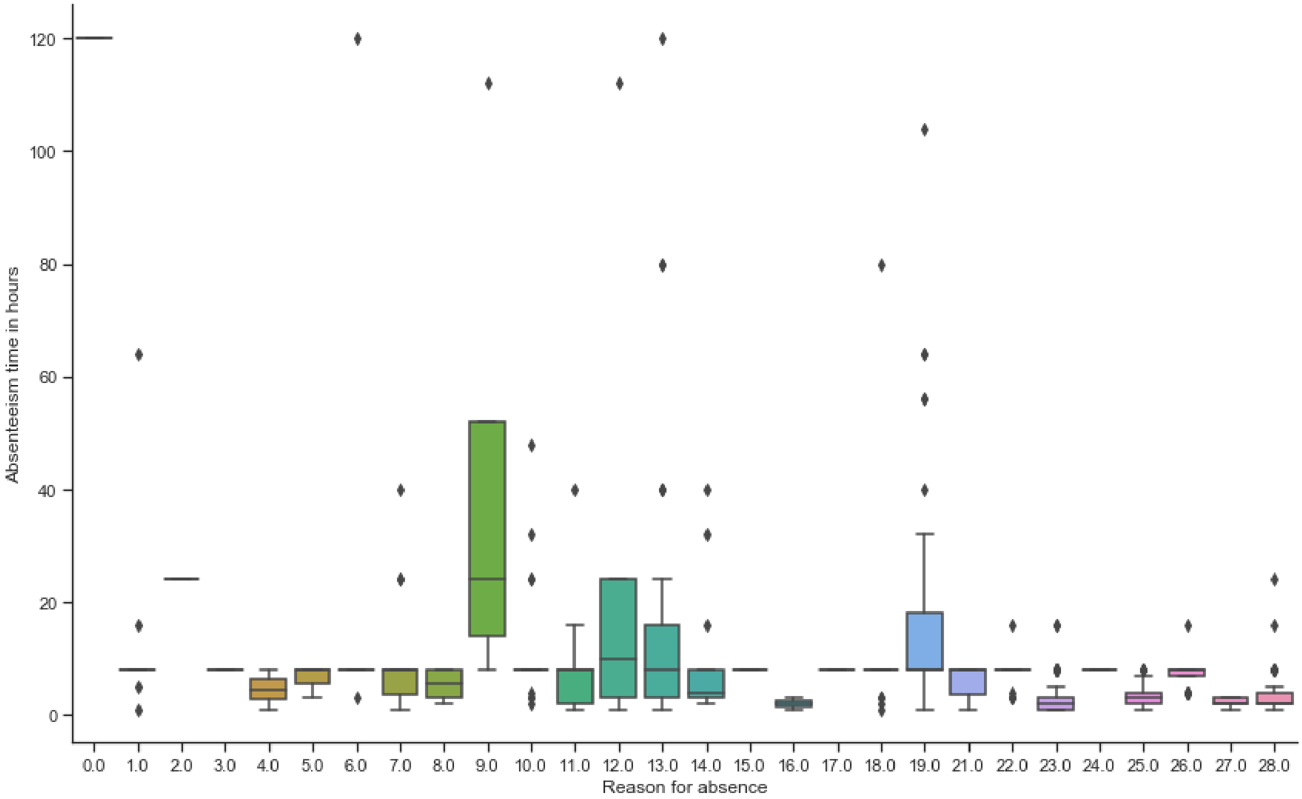


```
ID                                  0
Reason for absence                  3
Month of absence                    1
Day of the week                     0
Seasons                             0
Transportation expense              7
Distance from Residence to Work     3
Service time                        3
Age                                 3
Work load Average/day              10
Hit target                          6
Disciplinary failure                6
Education                          10
Son                                 6
Social drinker                      3
Social smoker                       4
Pet                                 2
Weight                              1
Height                             13
Body mass index                    29
Absenteeism time in hours          22
```

1. Reason for absence

   Pattern between ‘Reason for absence’ and ‘Absenteeism time in hours’ may be used to impute missing values in ‘Reason for absence’.

   Box plot for ‘Reason for absence’ and ‘Absenteeism time in hours’ :



Category 27 of ‘Reason for absence’ is taking < 10 hrs of ‘Absenteeism time in hours’. So, null values of ‘Reason for absence’ are put equal to 27 since ‘Absenteeism time in hours’ for the observations having null values is < 10 hrs.

Zero category of 'Reason for absence' column has been put equal to category 26(i.e. unjustified absence).

2. Month of absence

   Putting 'Month of absence' null value equal to 10.

3. Transportation expense

   'Transportation expense' depends on 'Distance from Residence to Work' so we will use 'Distance from Residence to Work' values to impute missing values in 'Transportation expense'.

If 'Distance from Residence to Work' value is 51, then ‘Transportation expense’ is equal to 179.
If 'Distance from Residence to Work' value is 50, then ‘Transportation expense’ is equal to 260.
If 'Distance from Residence to Work' value is 52, then ‘Transportation expense’ is equal to 361.
If 'Distance from Residence to Work' value is 11, then ‘Transportation expense’ is equal to 235.
If 'Distance from Residence to Work' value is 31, then ‘Transportation expense’ is equal to 291.

4. Distance from Residence to Work

   'ID' column has been used to impute missing value for 'Distance from Residence to Work'.

5. Service time

   'ID' column has been used to impute missing value for ‘Service time’.

6. Age

   'ID' column has been used to impute missing value for ‘Age’.

7. Work Load Average/day

   'Work load Average/day' values are dependent upon 'Month of absence' and 'Hit target' values.

8. Hit target

   'Hit target' values are dependent upon 'Month of absence' and 'Work load Average/day' values.

9. Disciplinary failure

   'Disciplinary failure' missing values have been put to 0.

10. Education

    'ID' column has been used to impute missing value for ‘Education’.

11. Son

    'ID' column has been used to impute missing value for ‘Son’.

12. Social drinker

    'ID' column has been used to impute missing value for ‘Social drinker’.

13. Social smoker

    'ID' column has been used to impute missing value for ‘Social smoker’.

14. Pet

    'ID' column has been used to impute missing value for ‘Pet’.

15. Weight

    'ID' column has been used to impute missing value for ‘Weight’.

16. Height

    'ID' column has been used to impute missing value for ‘Height’.

17. Body mass index

    'ID' column has been used to impute missing value for ‘Body mass index’.

18. Absenteeism time in hours

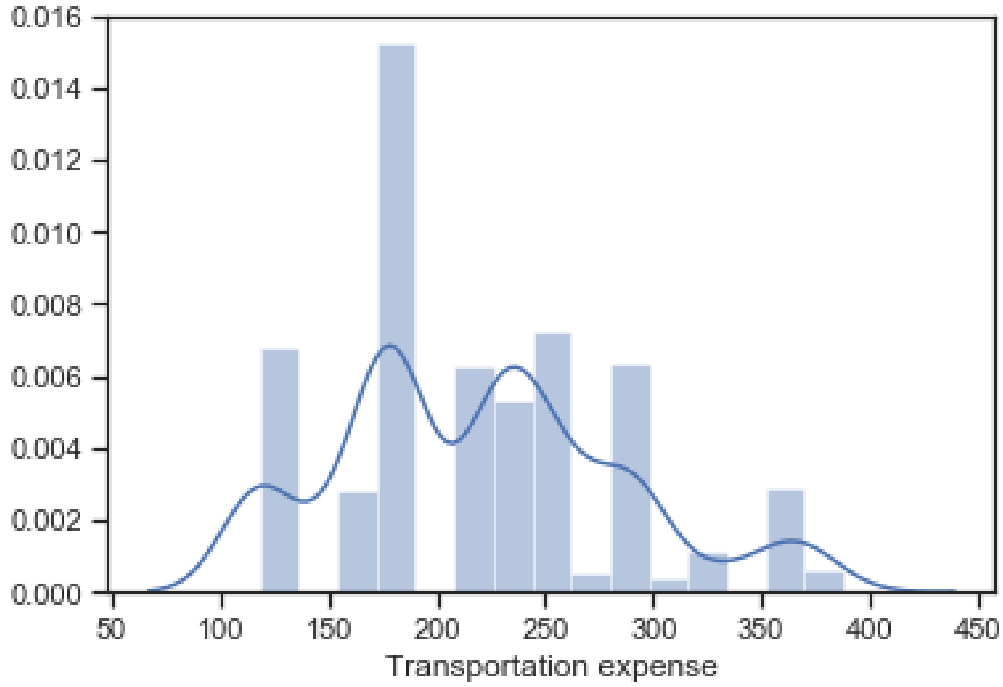
    'Reason for absence' column has been used to impute missing value for ‘Absenteeism time in hours’.

    All variables missing values have been imputed.

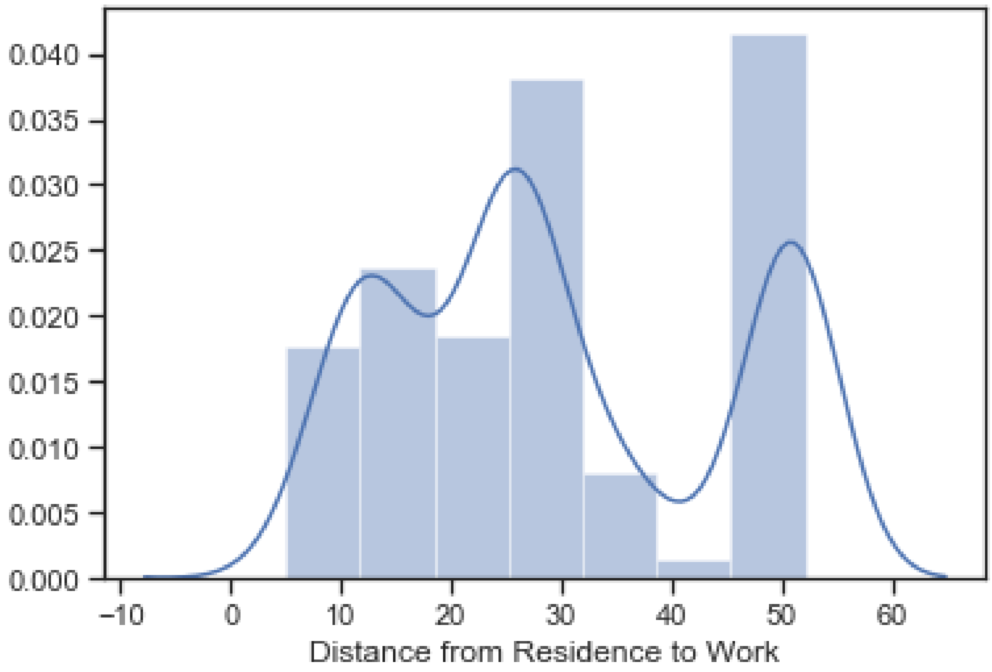
'Work load Average/day' variable has been converted from object to int type since it is a continuous variable.

2.2 Distributions

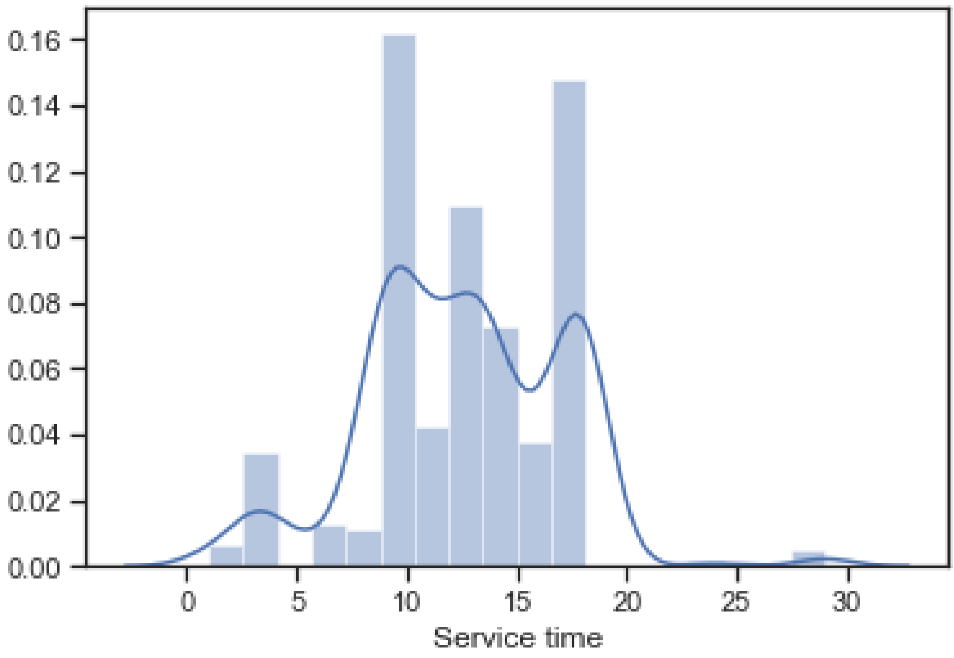
1. Transportation expense



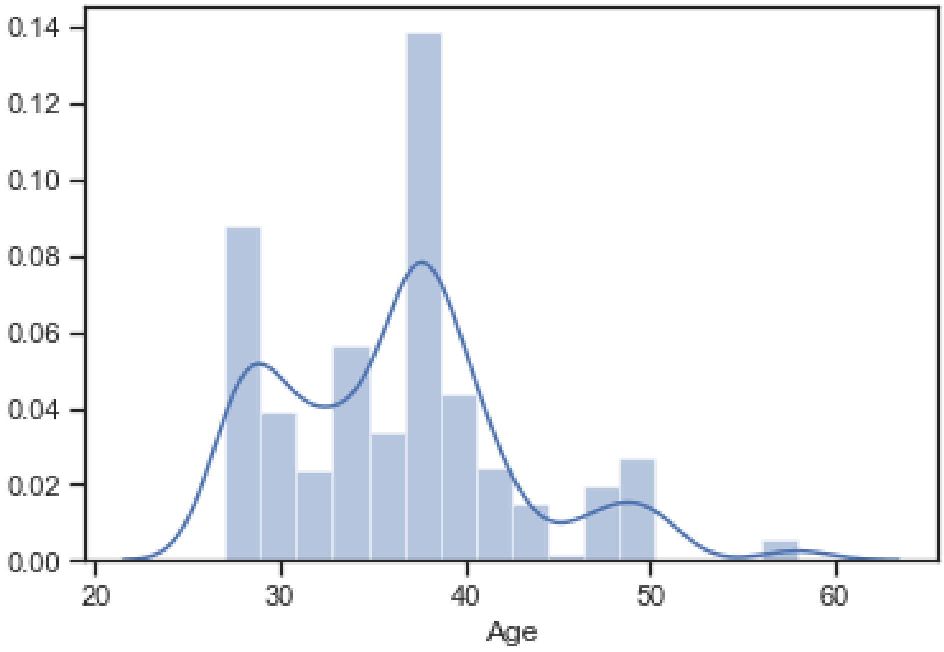
2. Distance from Residence to Work



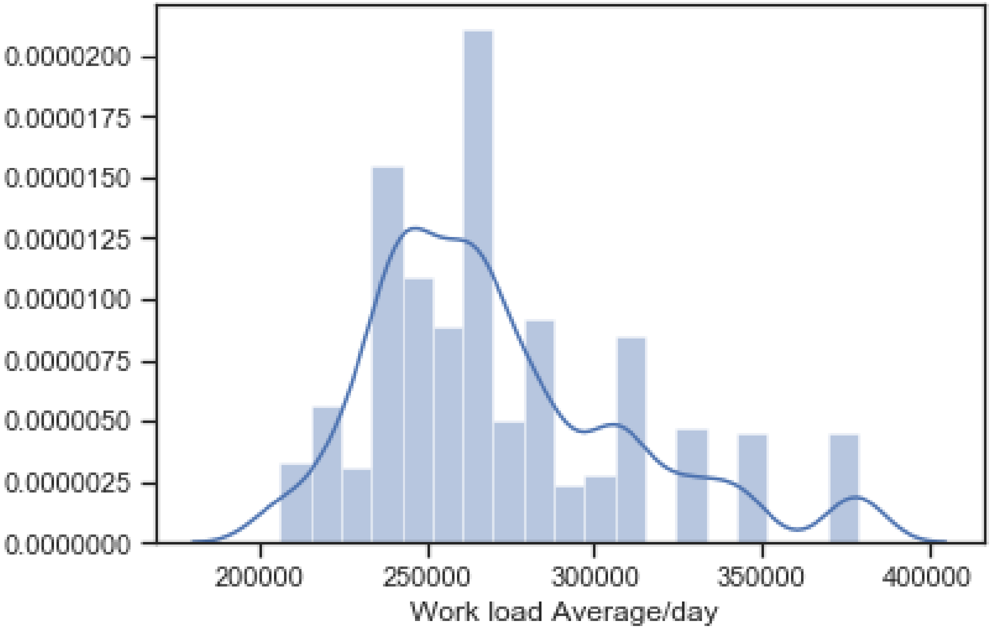
3. Service time



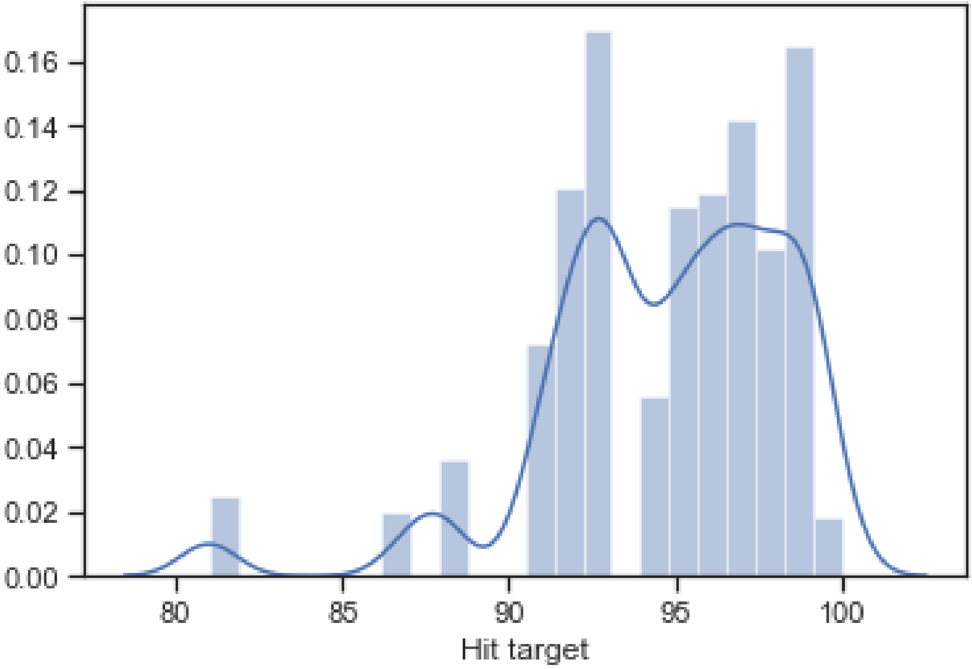
4. Age



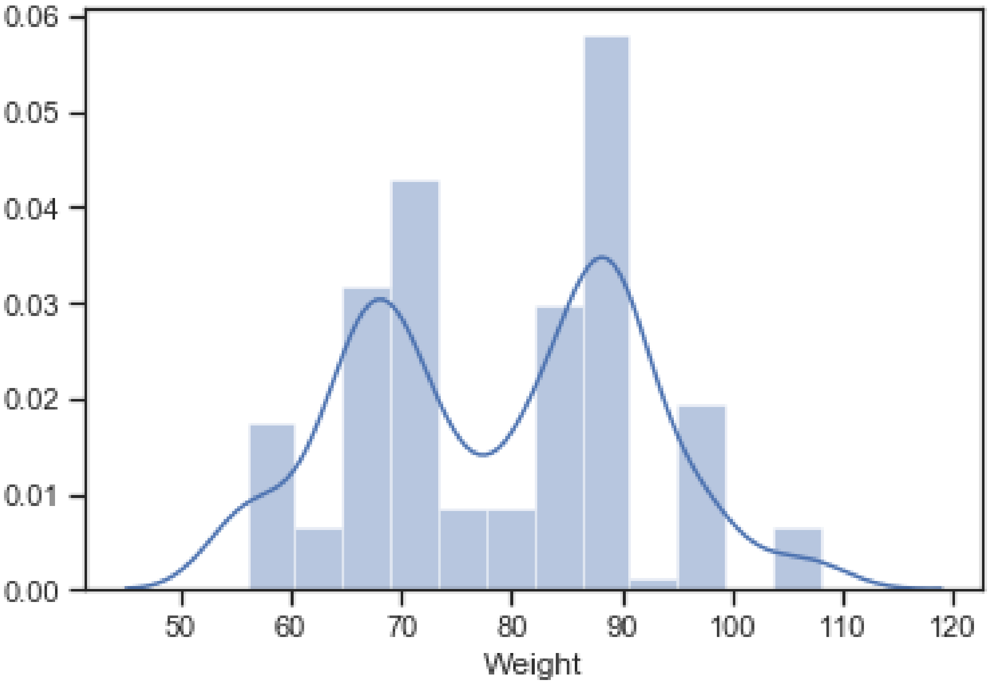
5. Work load Average/day



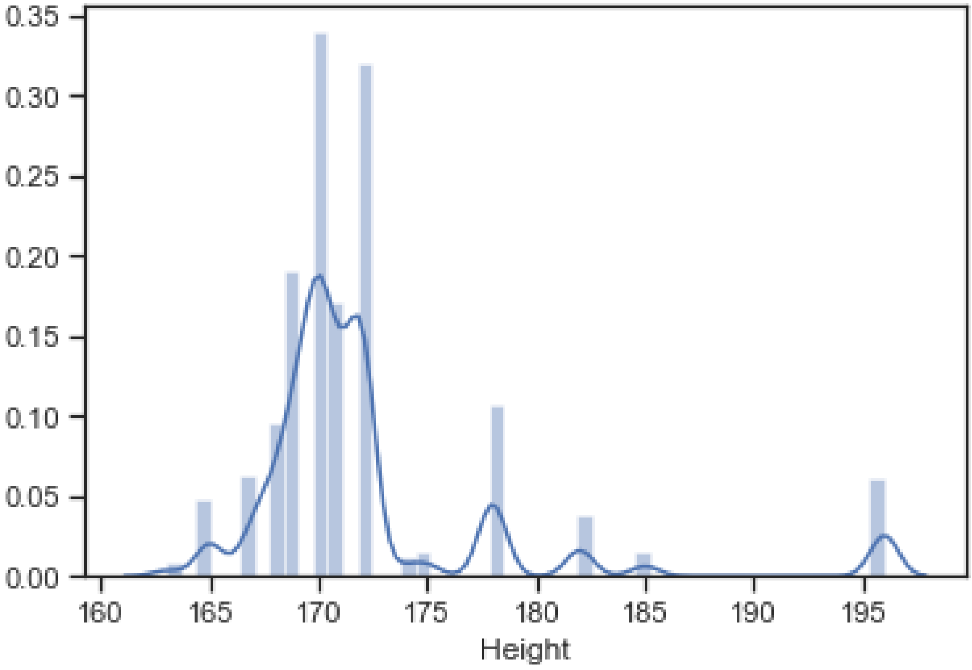
6. Hit target



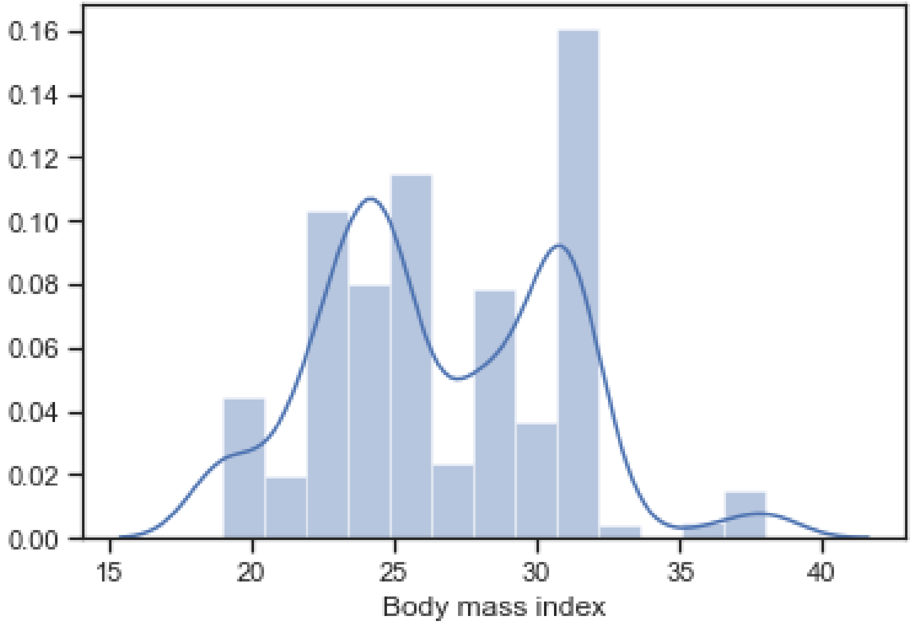
7. Weight



8. Height

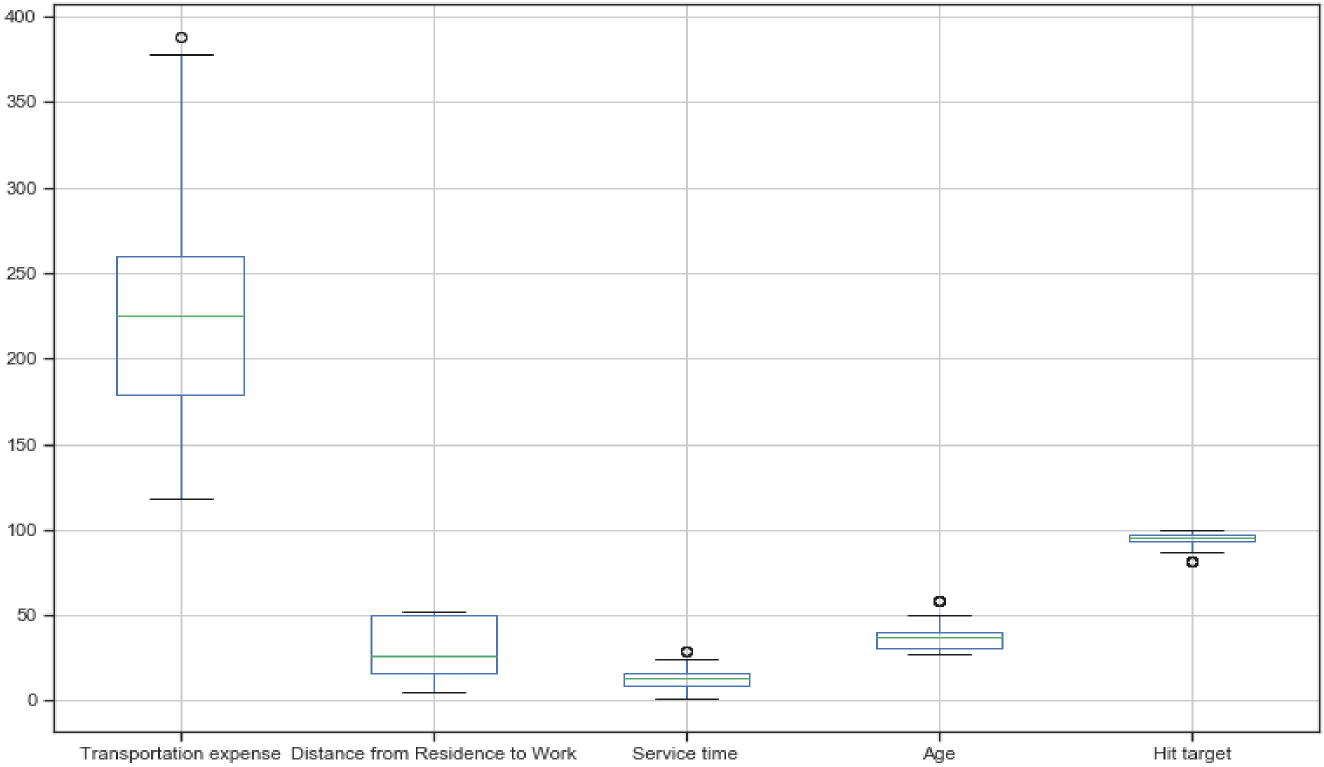


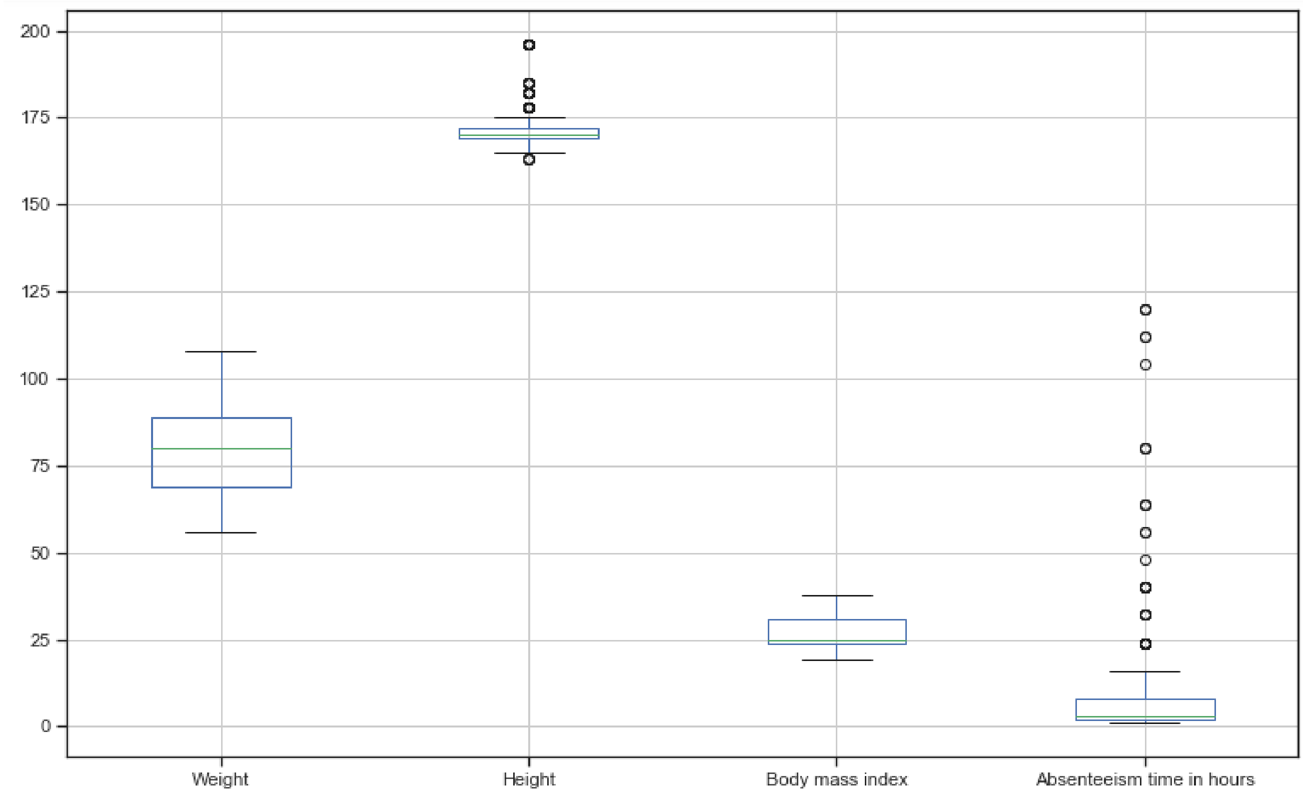
9. Body mass index

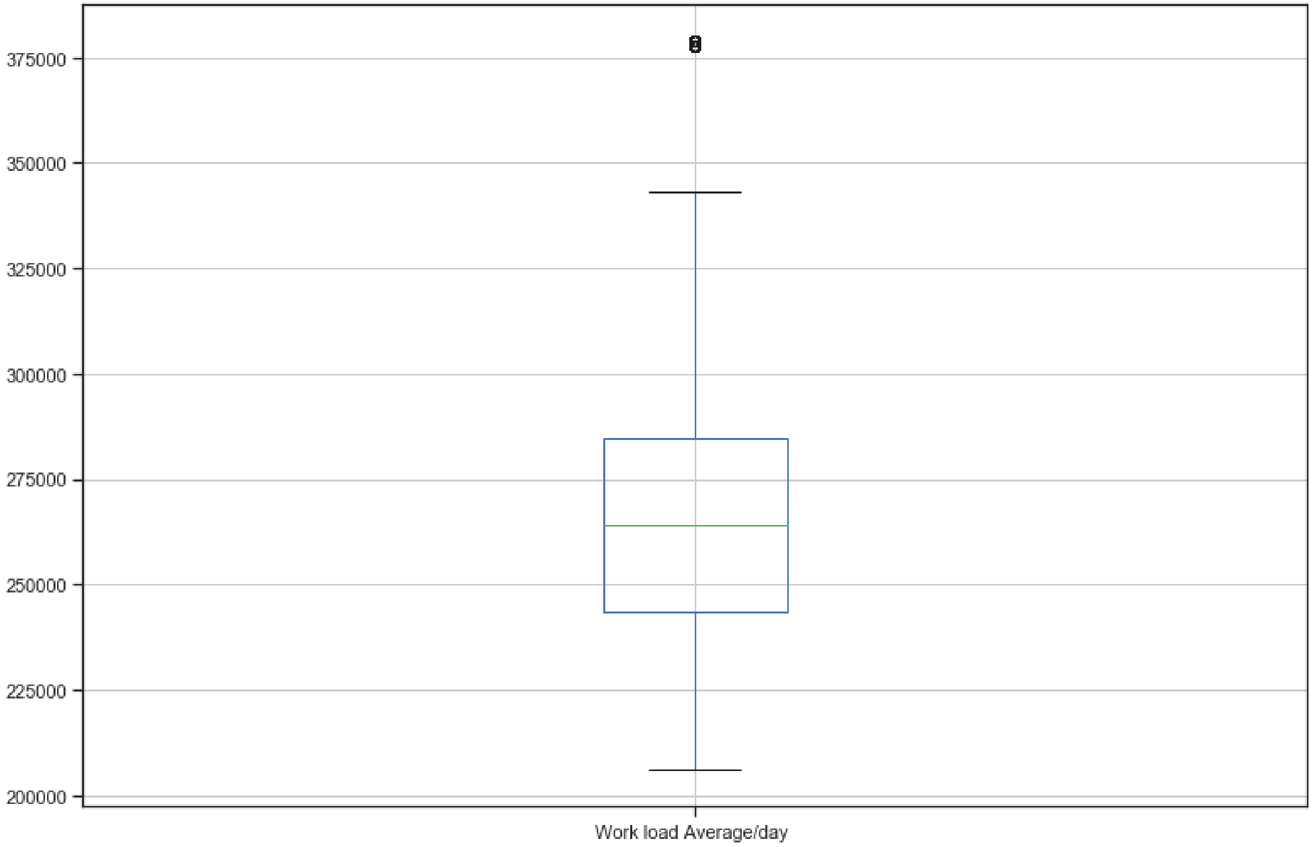


All continuous variables have skewed distributions.

2.3 Outlier Analysis

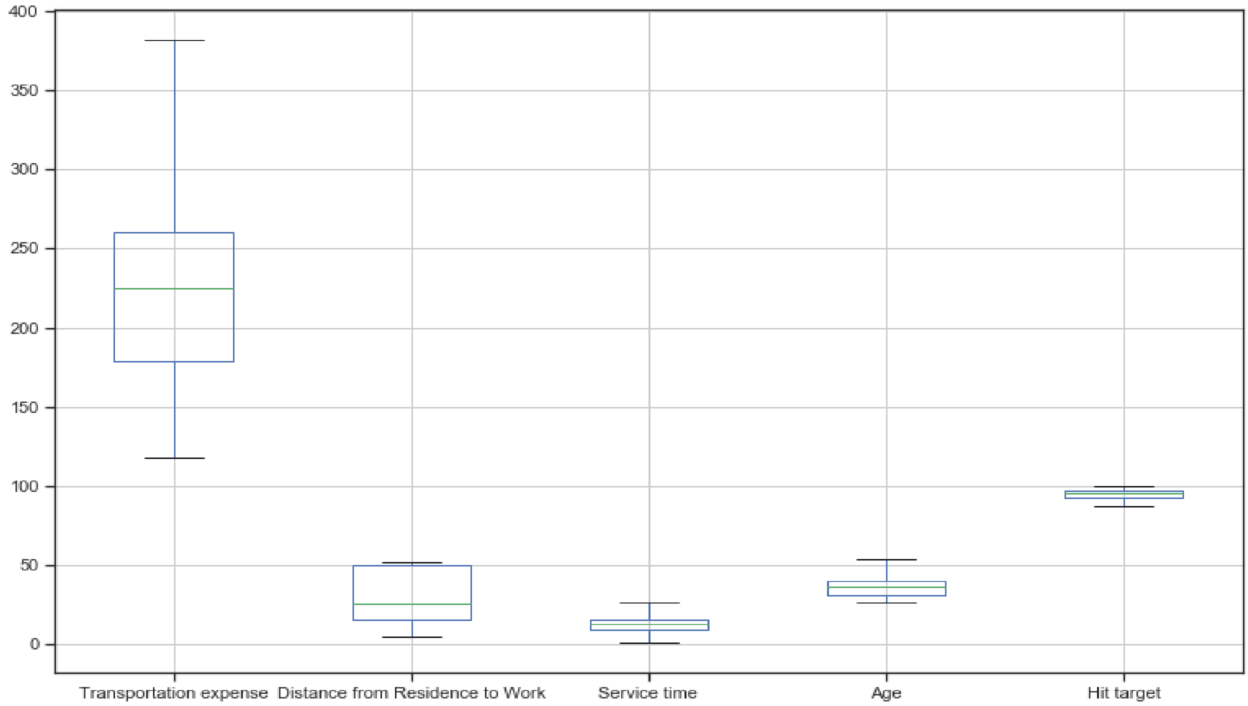


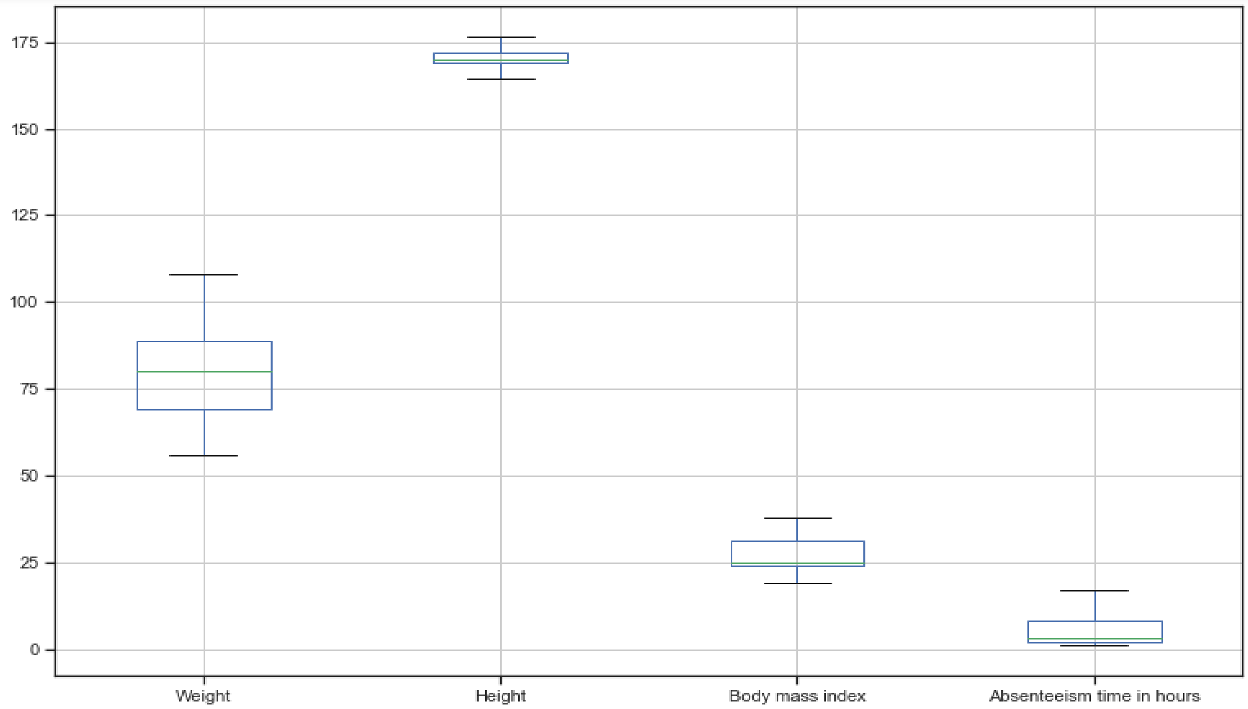

200
175
150
125
100
75
50
25
0
Weight
Height
Body mass index
Absenteeism time in hours

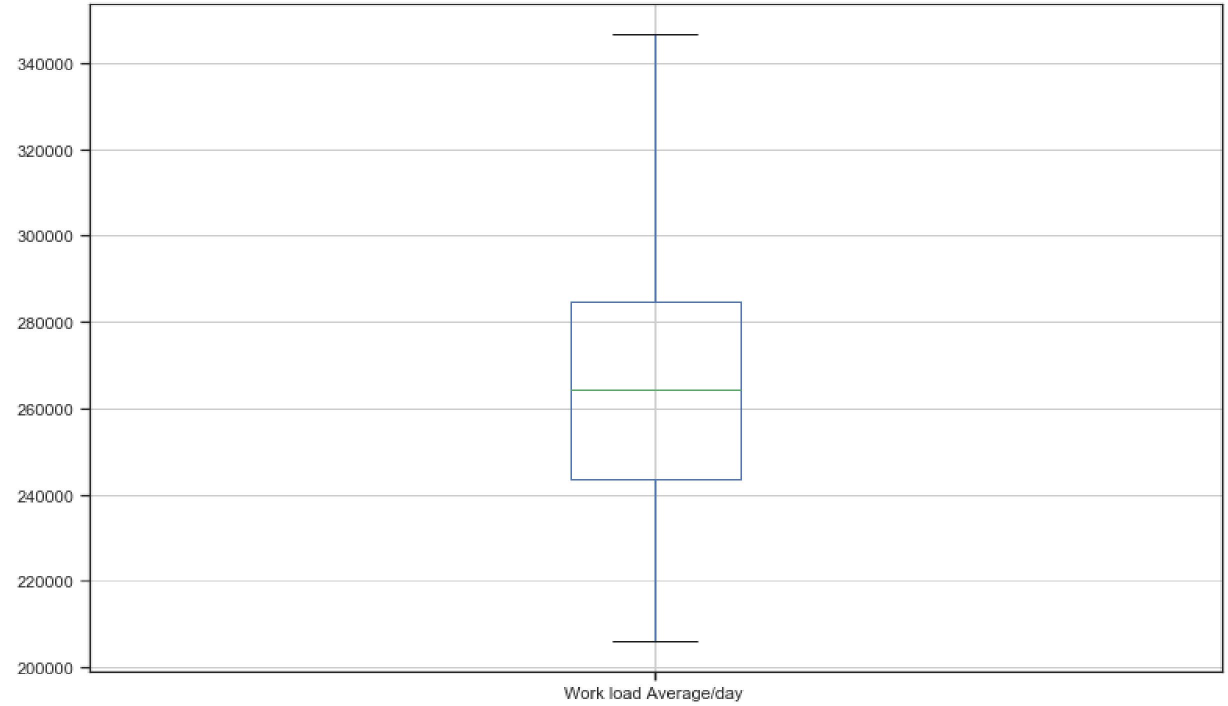

375000
350000
325000
300000
275000
250000
225000
200000
Work load Average/day

'Transportation expense', 'Service time', 'Age', 'Hit target', 'Height', 'Absenteeism time in hours',‘Work load Average/day’ have outliers.

Outliers have been capped as shown below:



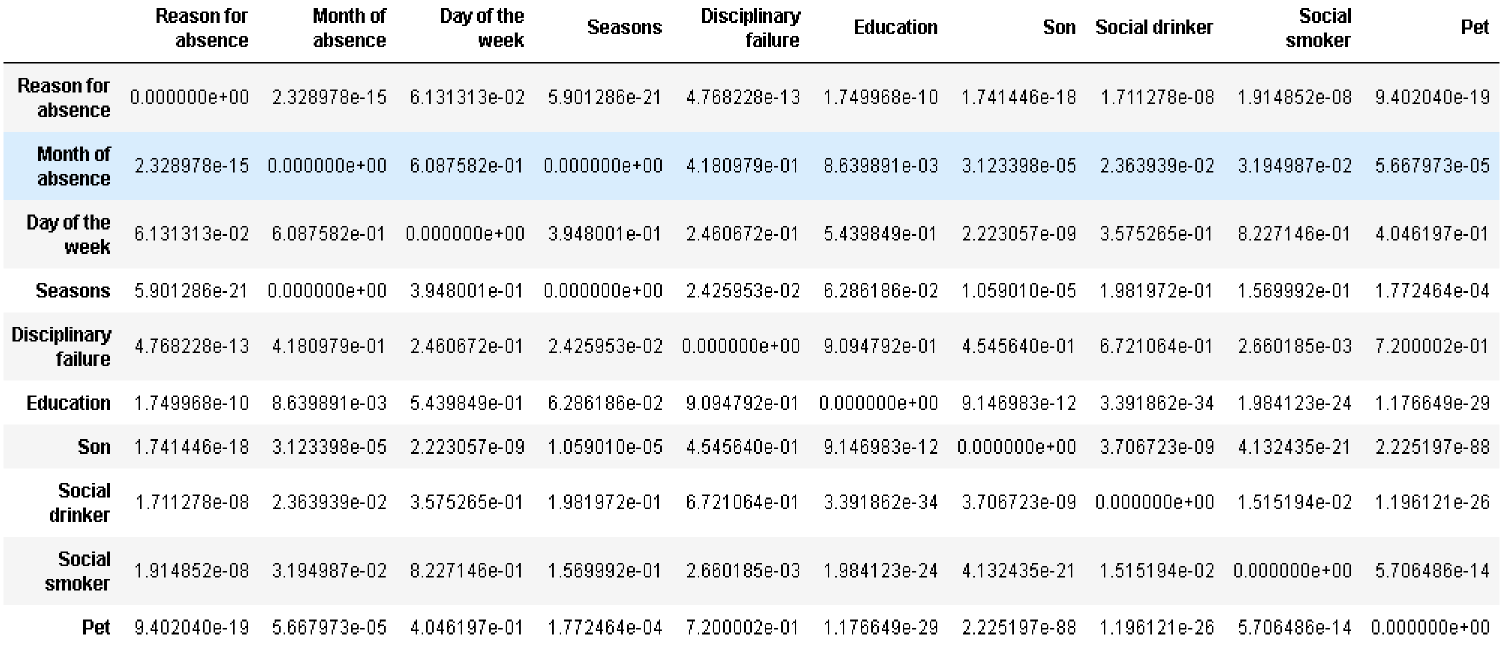




2.4 Correlation Analysis

Variables - 'Reason for absence', 'Month of absence', 'Day of the week', 'Seasons', 'Disciplinary failure', 'Education', 'Son', 'Social drinker', 'Social smoker', 'Pet' - have been converted to category type.

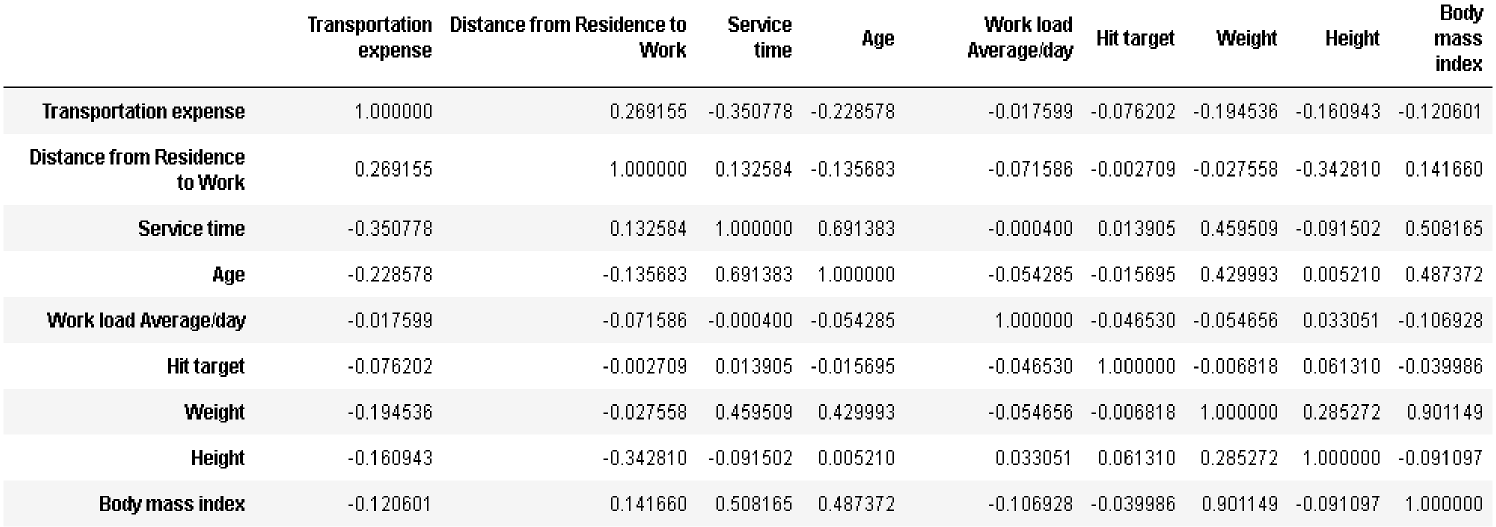
Chi-square test was done for correlation between categorical variables:



	Reason for absence	Month of absence	Day of the week	Seasons	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet
Reason for absence	0.000000e+00	2.328978e-15	6.131313e-02	5.901286e-21	4.768228e-13	1.749968e-10	1.741446e-18	1.711278e-08	1.914852e-08	9.402040e-19
Month of absence	2.328978e-15	0.000000e+00	6.087582e-01	0.000000e+00	4.180979e-01	8.639891e-03	3.123398e-05	2.363939e-02	3.194987e-02	5.667973e-05
Day of the week	6.131313e-02	6.087582e-01	0.000000e+00	3.948001e-01	2.460672e-01	5.439849e-01	2.223057e-09	3.575265e-01	8.227146e-01	4.046197e-01
Seasons	5.901286e-21	0.000000e+00	3.948001e-01	0.000000e+00	2.425953e-02	6.286186e-02	1.059010e-05	1.981972e-01	1.569992e-01	1.772464e-04
Disciplinary failure	4.768228e-13	4.180979e-01	2.460672e-01	2.425953e-02	0.000000e+00	9.094792e-01	4.545640e-01	6.721064e-01	2.660185e-03	7.200002e-01
Education	1.749968e-10	8.639891e-03	5.439849e-01	6.286186e-02	9.094792e-01	0.000000e+00	9.146983e-12	3.391862e-34	1.984123e-24	1.176649e-29
Son	1.741446e-18	3.123398e-05	2.223057e-09	1.059010e-05	4.545640e-01	9.146983e-12	0.000000e+00	3.706723e-09	4.132435e-21	2.225197e-88
Social drinker	1.711278e-08	2.363939e-02	3.575265e-01	1.981972e-01	6.721064e-01	3.391862e-34	3.706723e-09	0.000000e+00	1.515194e-02	1.196121e-26
Social smoker	1.914852e-08	3.194987e-02	8.227146e-01	1.569992e-01	2.660185e-03	1.984123e-24	4.132435e-21	1.515194e-02	0.000000e+00	5.706486e-14
Pet	9.402040e-19	5.667973e-05	4.046197e-01	1.772464e-04	7.200002e-01	1.176649e-29	2.225197e-88	1.196121e-26	5.706486e-14	0.000000e+00

Dropping Seasons since p-value of 'Seasons' versus 'Month of absence' is 0.00 (<0.05) rejecting null hypothesis that the two variables are independent.

Correlation test was done for continuous independent variables:

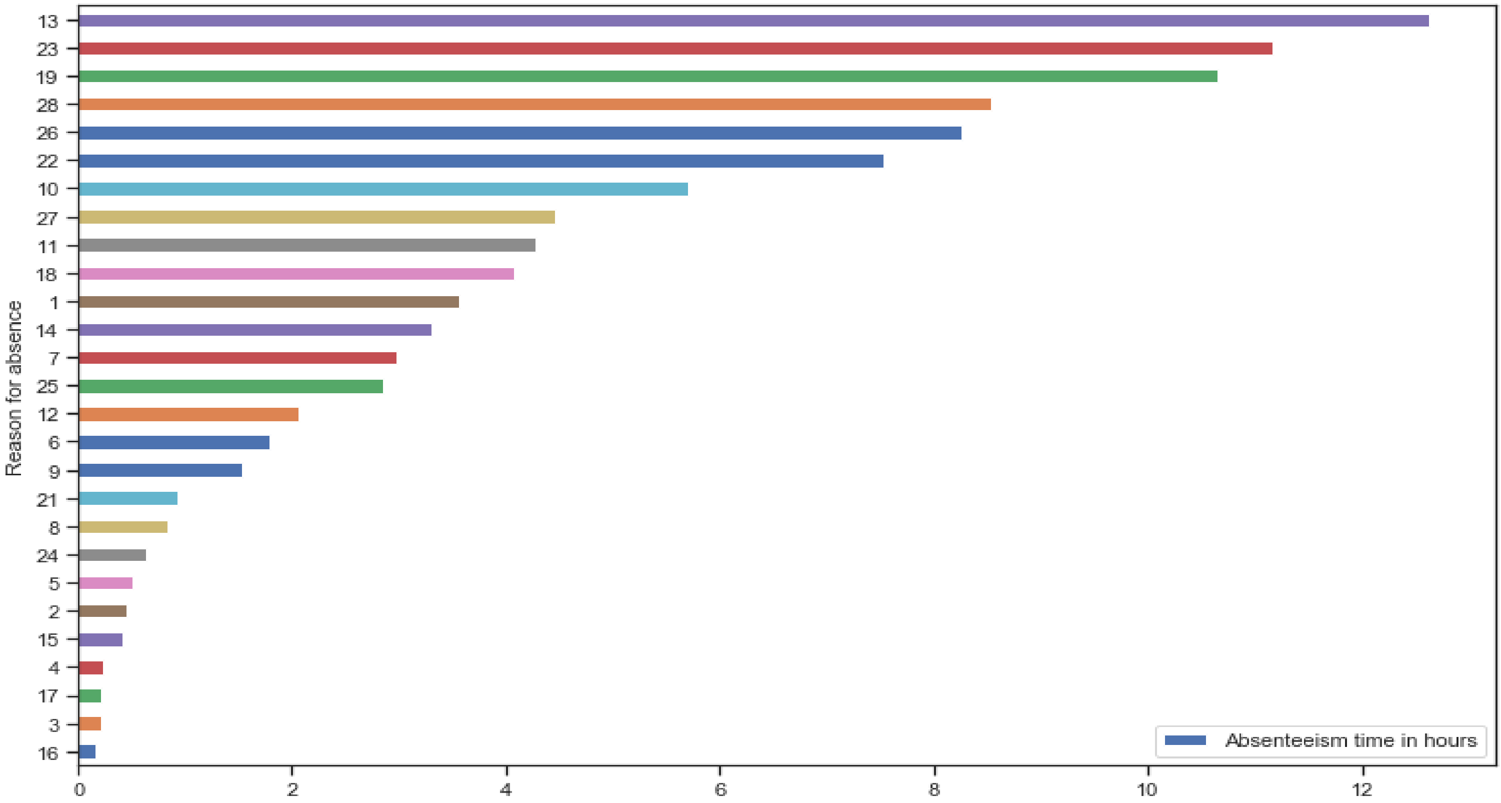


	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Weight	Height	Body mass index
Transportation expense	1.000000	0.269155	-0.350778	-0.228578	-0.017599	-0.076202	-0.194536	-0.160943	-0.120601
Distance from Residence to Work	0.269155	1.000000	0.132584	-0.135683	-0.071586	-0.002709	-0.027558	-0.342810	0.141660
Service time	-0.350778	0.132584	1.000000	0.691383	-0.000400	0.013905	0.459509	-0.091502	0.508165
Age	-0.228578	-0.135683	0.691383	1.000000	-0.054285	-0.015695	0.429993	0.005210	0.487372
Work load Average/day	-0.017599	-0.071586	-0.000400	-0.054285	1.000000	-0.046530	-0.054656	0.033051	-0.106928
Hit target	-0.076202	-0.002709	0.013905	-0.015695	-0.046530	1.000000	-0.006818	0.061310	-0.039986
Weight	-0.194536	-0.027558	0.459509	0.429993	-0.054656	-0.006818	1.000000	0.285272	0.901149
Height	-0.160943	-0.342810	-0.091502	0.005210	0.033051	0.061310	0.285272	1.000000	-0.091097
Body mass index	-0.120601	0.141660	0.508165	0.487372	-0.106928	-0.039986	0.901149	-0.091097	1.000000

No two variables have correlation coeff. > 0.95 so we will not drop any continuous independent variables.

2.5 Relationships of categorical independent variables with dependent variable

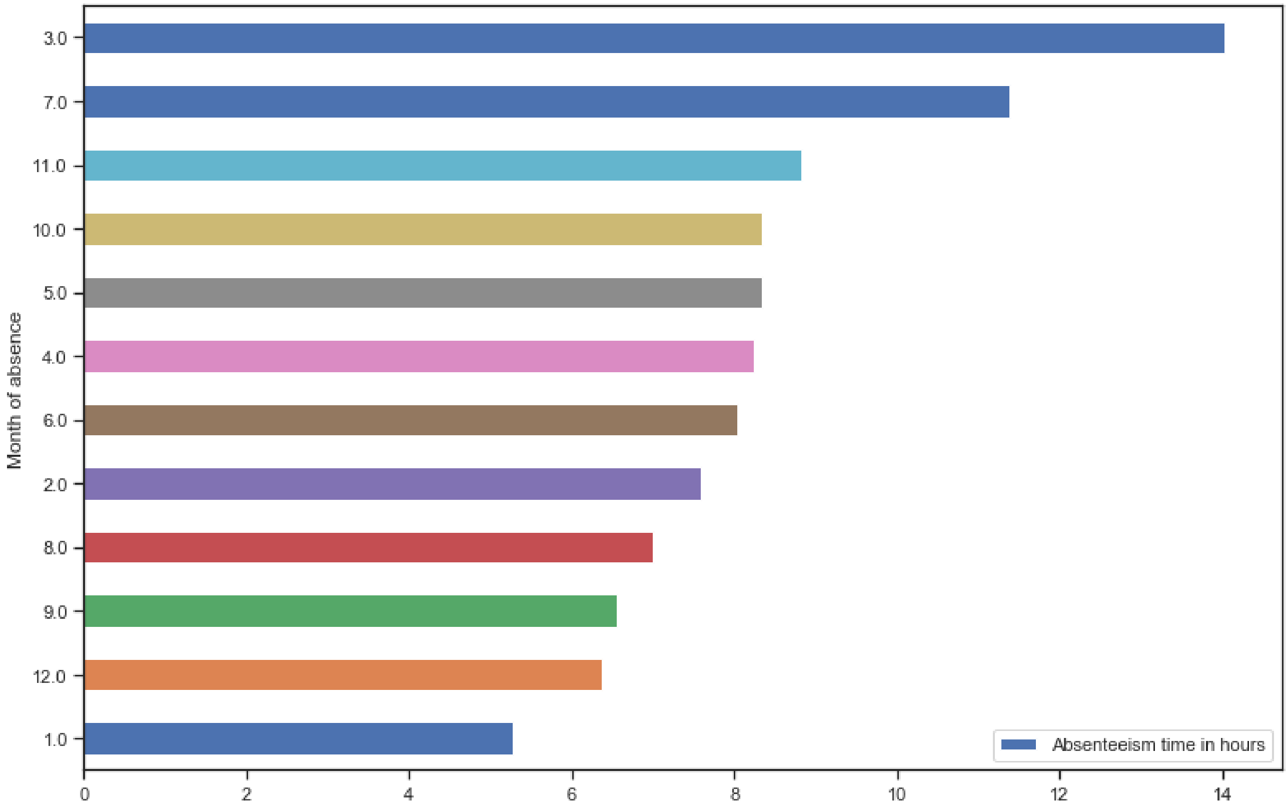
1. 'Reason for absence' Vs. 'Absenteeism time in hours'



Top 3 categories in order of Absenteeism time are:

A. Category 13 : Diseases of the musculoskeletal system and connective tissue - 12.62 % of total time
B. Category 23 : medical consultation - 11.17 % of total time
C. Category 19 : Injury, poisoning and certain other consequences of external causes - 10.64 % of total time
D. Category 28 : dental consultation - 8.53 % 0f total time
E. Category 26 : unjustified absence - 8.27 % of total time

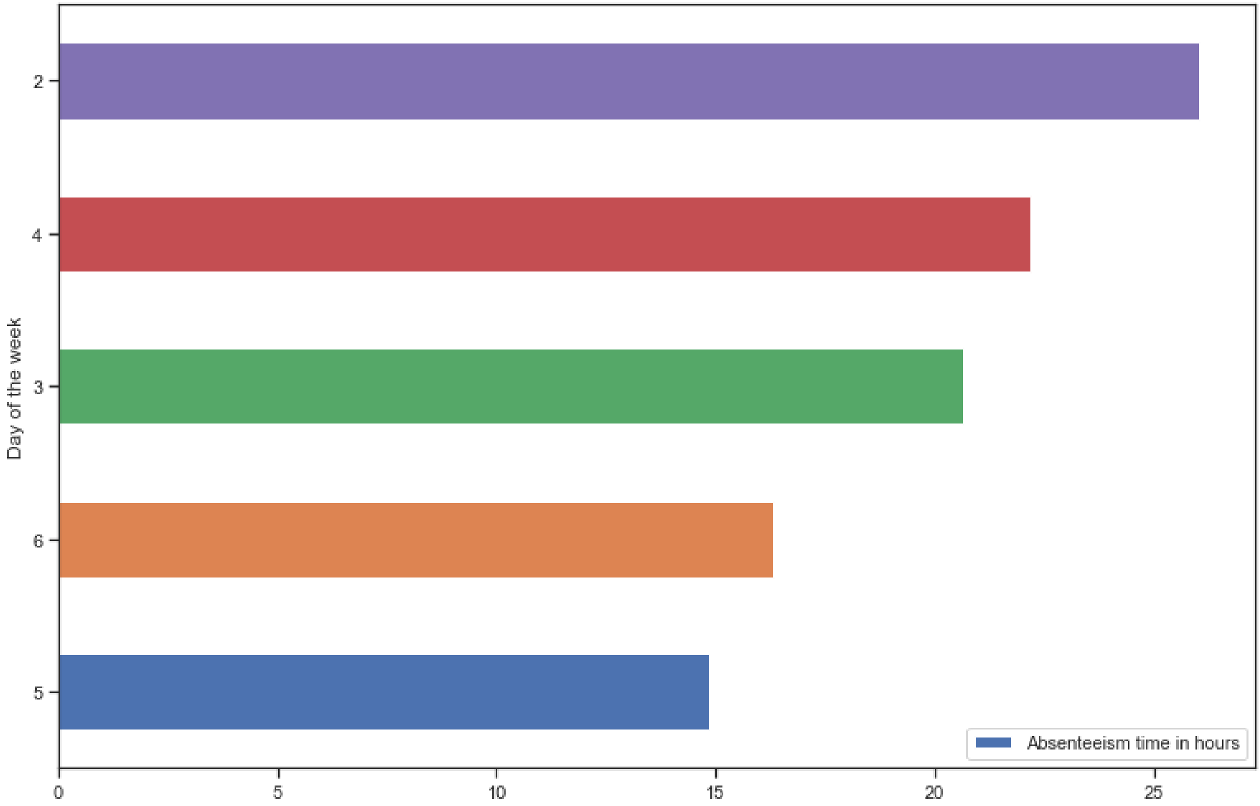
2. 'Month of absence' Vs. 'Absenteeism time in hours'



Top 3 months in order of Absenteeism time are:

A. Month 3 : March - 14.02 % of total time
B. Month 7 : July - 11.38 % of total time
C. Month 11 : November - 8.82 % of total time

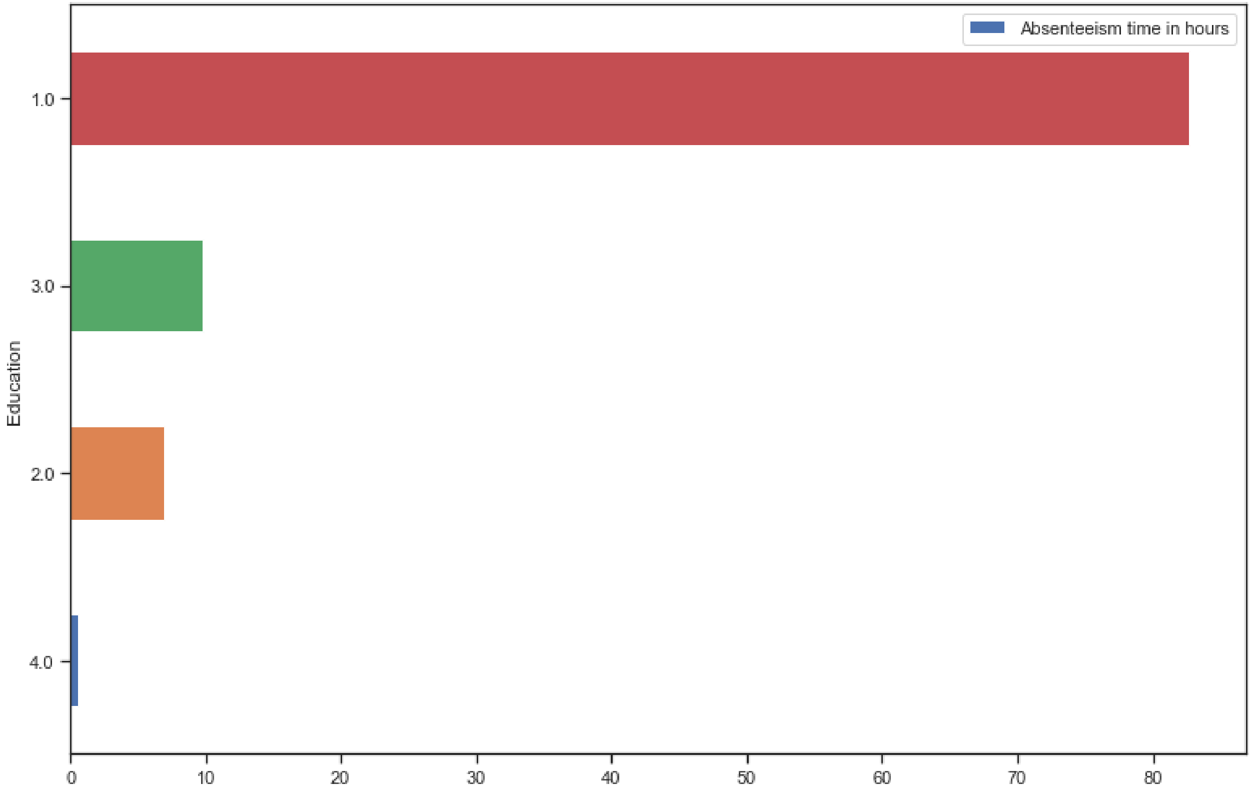
3. 'Day of the week' Vs. 'Absenteeism time in hours'



Top 3 days in order of Absenteeism time are:

A. Day 2 : Monday - 26.02 % of total time
B. Day 4 : Wednesday - 22.19 % of total time
C. Day 3 : Tuesday - 20.63 % of total time

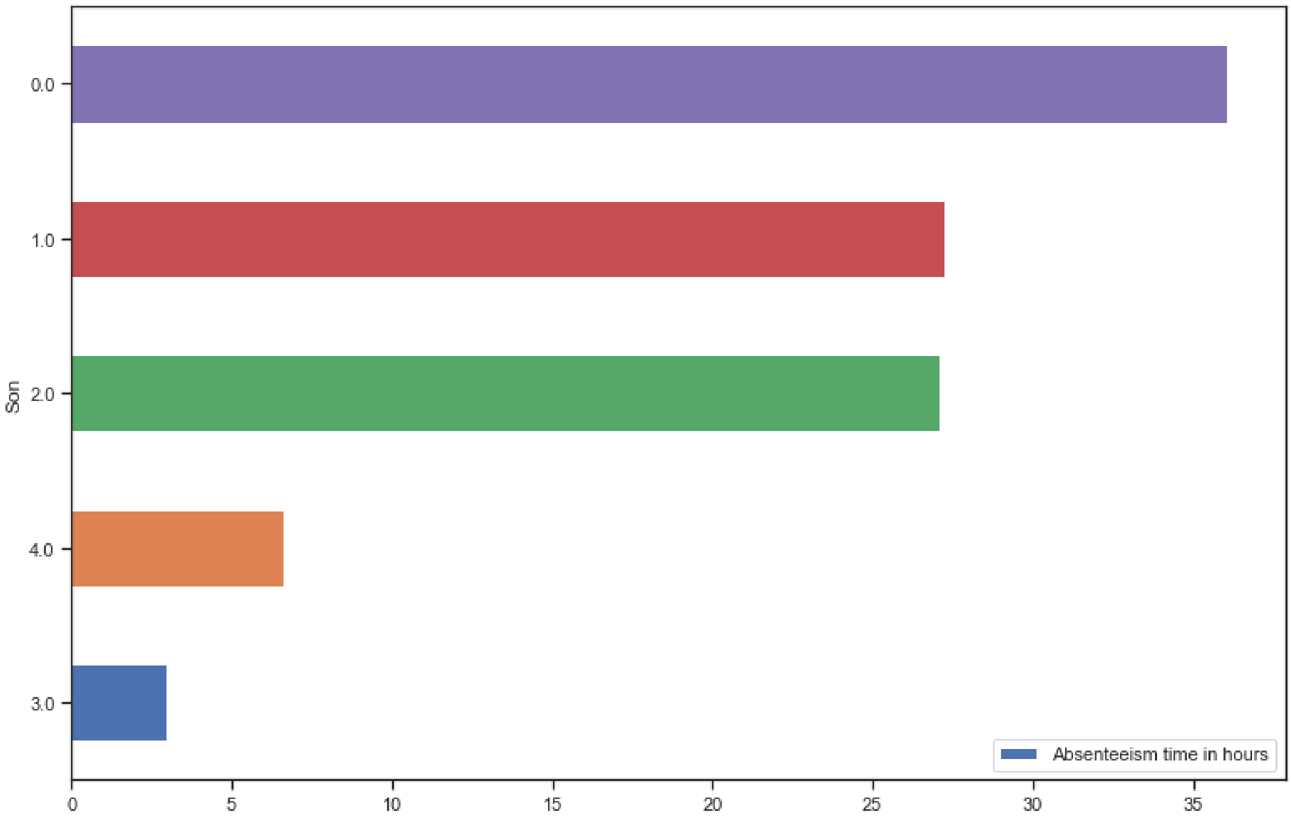
4. 'Education' Vs. 'Absenteeism time in hours'



82.69 % of absenteeism time is contributed by people having high school education.

This may be due to majority of people having high school education. No conclusion may be drawn from this graph.

5. 'Son' Vs. 'Absenteeism time in hours'

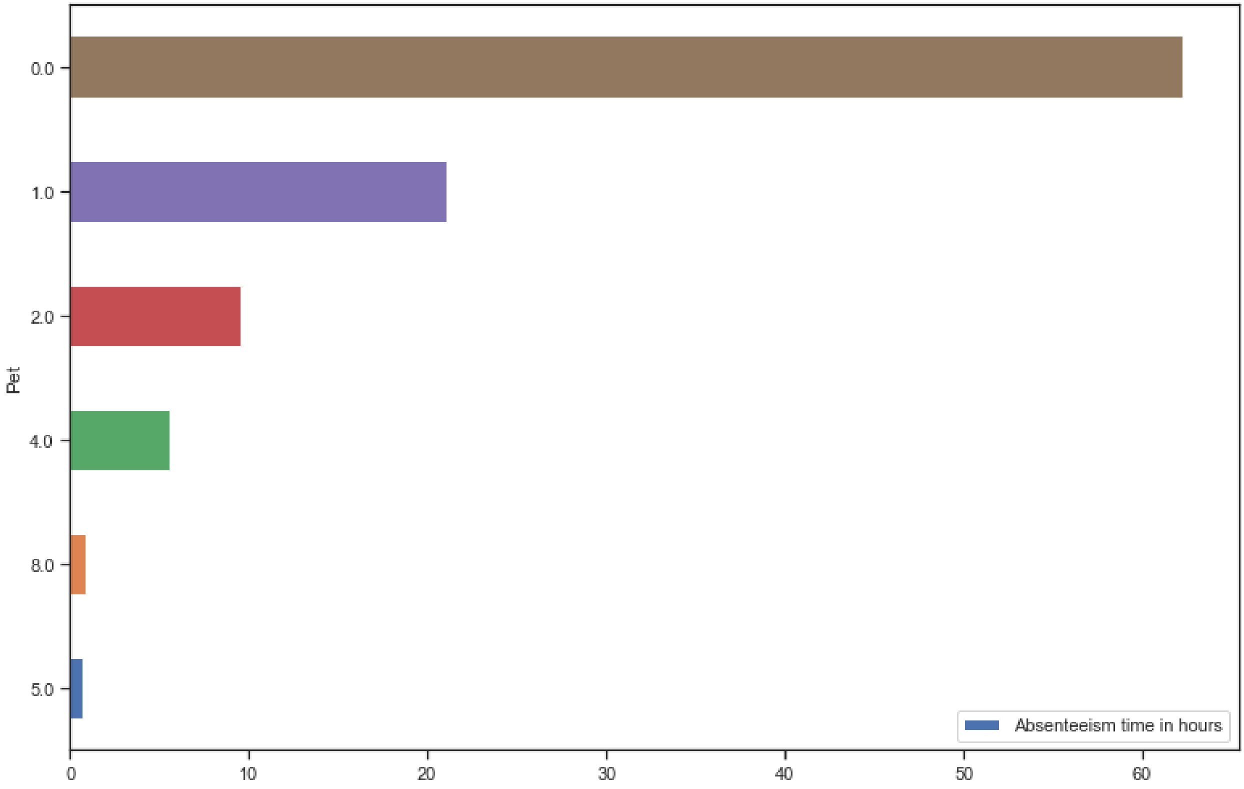


Top 3 categories in order of Absenteeism time are:

A. Category 0 : No son - 36.06 % of total time
B. Category 1 : One son - 27.23 % of total time
C. Category 2 : Two sons - 27.08 % of total time

People with no son are taking most of absenteeism time.

6. 'Pet' Vs. 'Absenteeism time in hours'



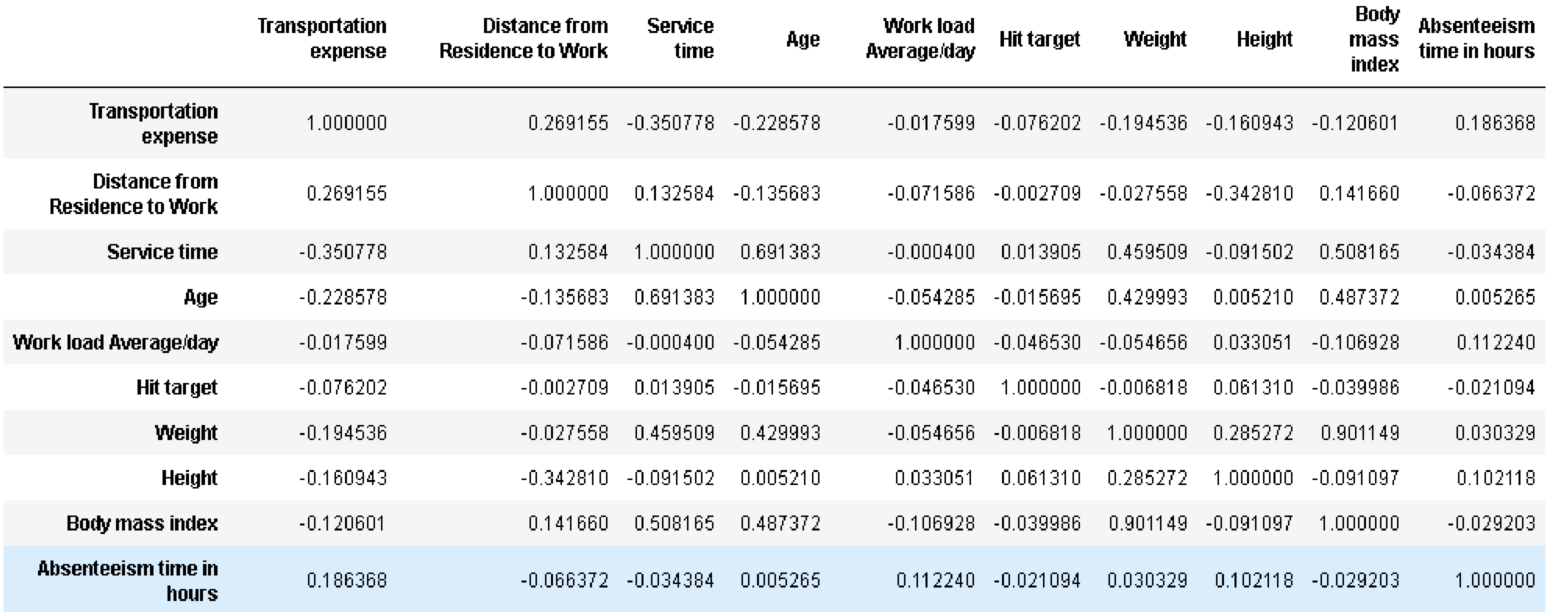
Top 3 categories in order of Absenteeism time are:

A. Category 0 : No pet - 62.24 % of total time
B. Category 1 : One pet - 21.08 % of total time
C. Category 2 : Two pets - 9.53 % of total time

People with no pet are taking most of absenteeism time.

2.6 Relationships of continuous independent variables with dependent variable

Correlation of independent variables with dependent variable



	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Weight	Height	Body mass index	Absenteeism time in hours
Transportation expense	1.000000	0.269155	-0.350778	-0.228578	-0.017599	-0.076202	-0.194536	-0.160943	-0.120601	0.186368
Distance from Residence to Work	0.269155	1.000000	0.132584	-0.135683	-0.071586	-0.002709	-0.027558	-0.342810	0.141660	-0.066372
Service time	-0.350778	0.132584	1.000000	0.691383	-0.000400	0.013905	0.459509	-0.091502	0.508165	-0.034384
Age	-0.228578	-0.135683	0.691383	1.000000	-0.054285	-0.015695	0.429993	0.005210	0.487372	0.005265
Work load Average/day	-0.017599	-0.071586	-0.000400	-0.054285	1.000000	-0.046530	-0.054656	0.033051	-0.106928	0.112240
Hit target	-0.076202	-0.002709	0.013905	-0.015695	-0.046530	1.000000	-0.006818	0.061310	-0.039986	-0.021094
Weight	-0.194536	-0.027558	0.459509	0.429993	-0.054656	-0.006818	1.000000	0.285272	0.901149	0.030329
Height	-0.160943	-0.342810	-0.091502	0.005210	0.033051	0.061310	0.285272	1.000000	-0.091097	0.102118
Body mass index	-0.120601	0.141660	0.508165	0.487372	-0.106928	-0.039986	0.901149	-0.091097	1.000000	-0.029203
Absenteeism time in hours	0.186368	-0.066372	-0.034384	0.005265	0.112240	-0.021094	0.030329	0.102118	-0.029203	1.000000

Correlation of every continuous independent variable with dependent variable ('Absenteeism time in hours') is < 0.2 which means that **no independent variable has strong relationship with dependent variable.**

2.7 Conclusions & Remedies

Conclusions & possible remedies are:

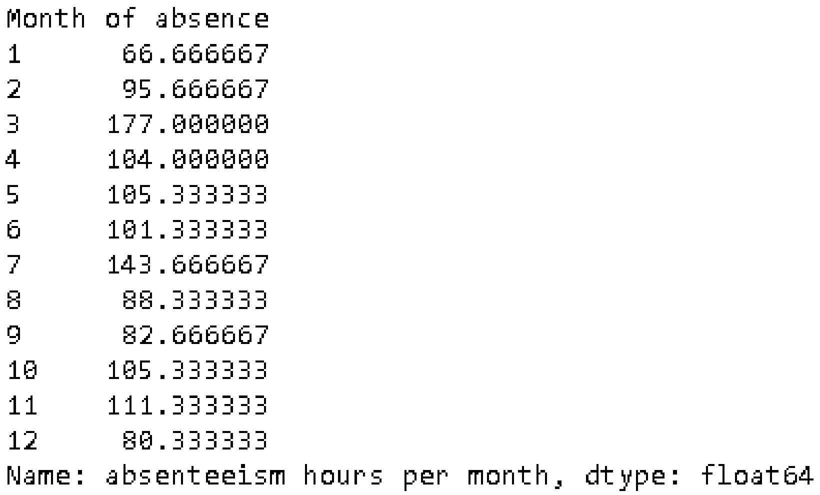
a. Musculoskeletal system disease is the major reason of absenteeism. Bad working posture & high workload are possible reasons for the high incidence of musculoskeletal disease. Company should conduct a study on the working postures of people and go for more ergonomic workplace design. Company should try to optimize workload keeping in mind occupational health of working people.
b. Medical consultation may be brought down by optimizing workloads.
c. Injury incidence may be reduced by creating proper ergonomic working setup.
d. Dental consultation time may be reduced by informing employees of the dental health guidelines so that they can take better care of their teeth.
e. Unjustified absence is too high. Company should try to reduce high workloads so that employees don't feel work stress to take unjustified absence leave.

Chapter 3

Modelling

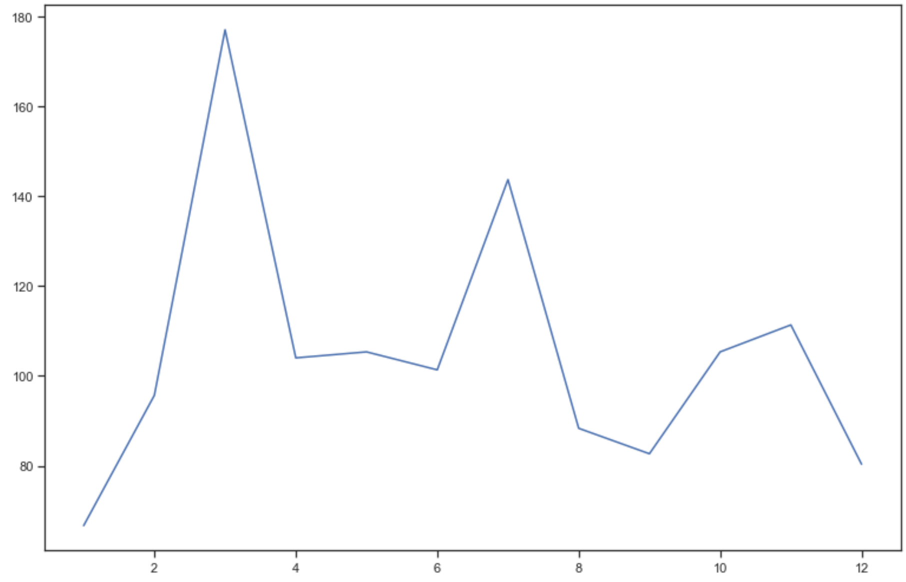
Time Series forecasting will be used to forecast the absenteeism hours for every month in 2011.Absenteeism hours has been aggregated by month to get total absenteeism hours for every month. Since this data is of 3 years from July,2007 to June,2010, Aggregated absenteeism hours has been divided by 3 to get average absent hours for every month.

Time Series is as under:



```
Month of absence
1       66.666667
2       95.666667
3      177.000000
4      104.000000
5      105.333333
6      101.333333
7      143.666667
8       88.333333
9       82.666667
10     105.333333
11     111.333333
12      80.333333
Name: absenteeism hours per month, dtype: float64
```

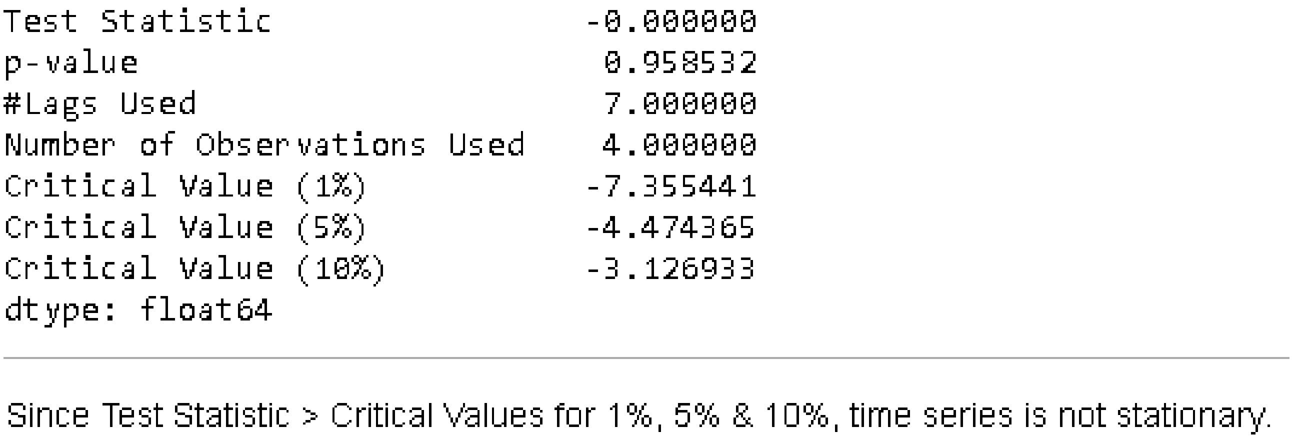
Plot of time series is as under:



3.1 Time Series - Arima Model

First, **Dickey Fuller test** has been done to check if time series is stationary or not.

Results were:

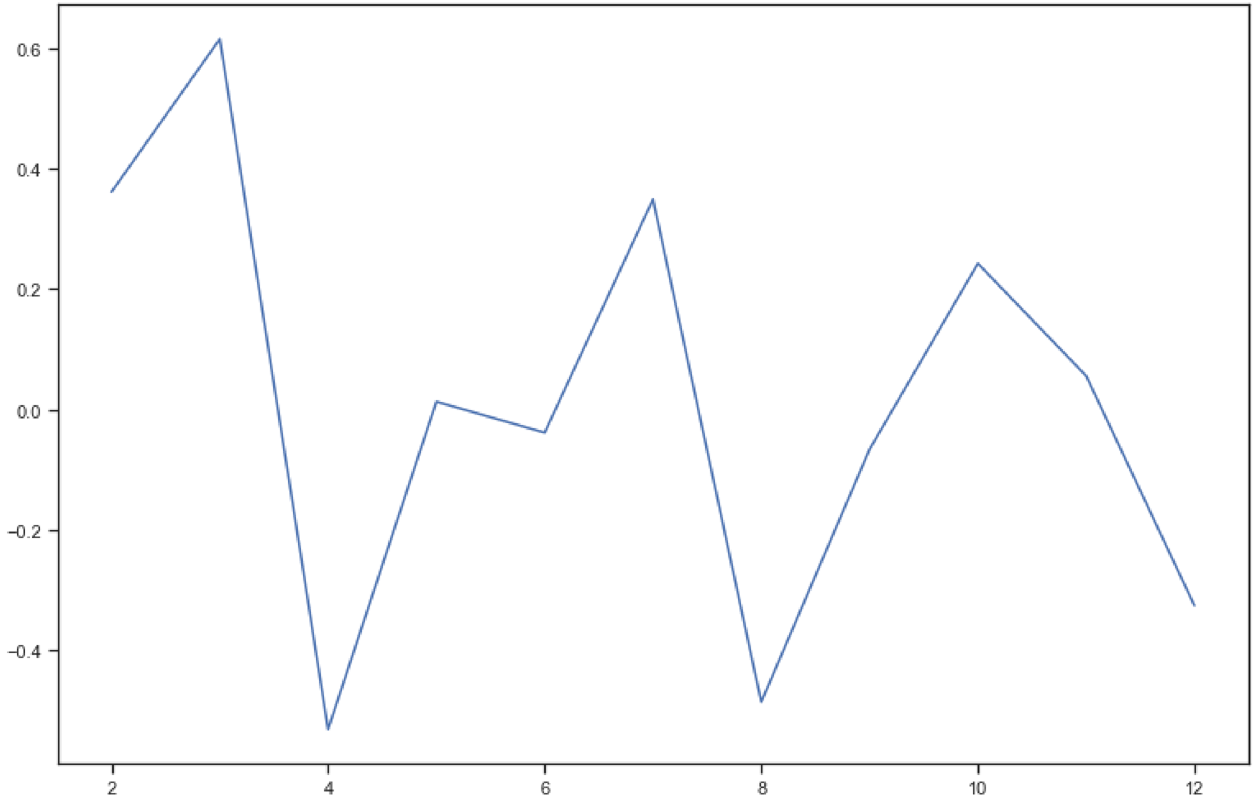


```
Test Statistic                  -0.000000
p-value                          0.958532
#Lags Used                       7.000000
Number of Observations Used      4.000000
Critical Value (1%)             -7.355441
Critical Value (5%)             -4.474365
Critical Value (10%)            -3.126933
dtype: float64
```

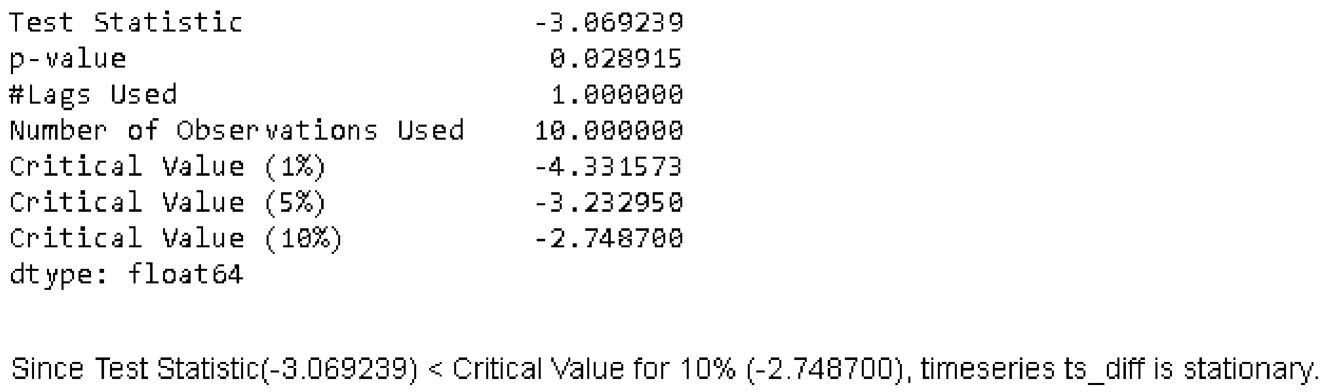
Since Test Statistic > Critical Values for 1%, 5% & 10%, time series is not stationary.

We will be first taking log of time series and then subtracting a shifted (single lag) log time series from log series.

Now, time series plot is as under:



Dickey Fuller test was done again after taking log & differencing on log.

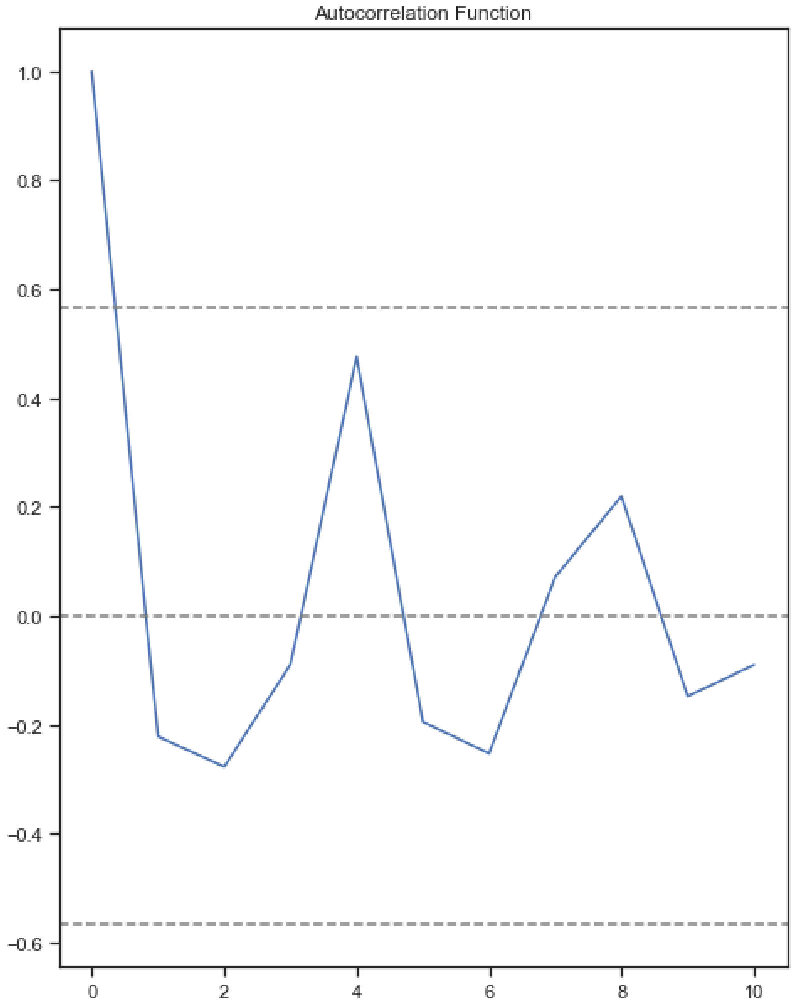


```
Test Statistic                   -3.069239
p-value                           0.028915
#Lags Used                        1.000000
Number of Observations Used      10.000000
Critical Value (1%)              -4.331573
Critical Value (5%)              -3.232950
Critical Value (10%)             -2.748700
dtype: float64
```

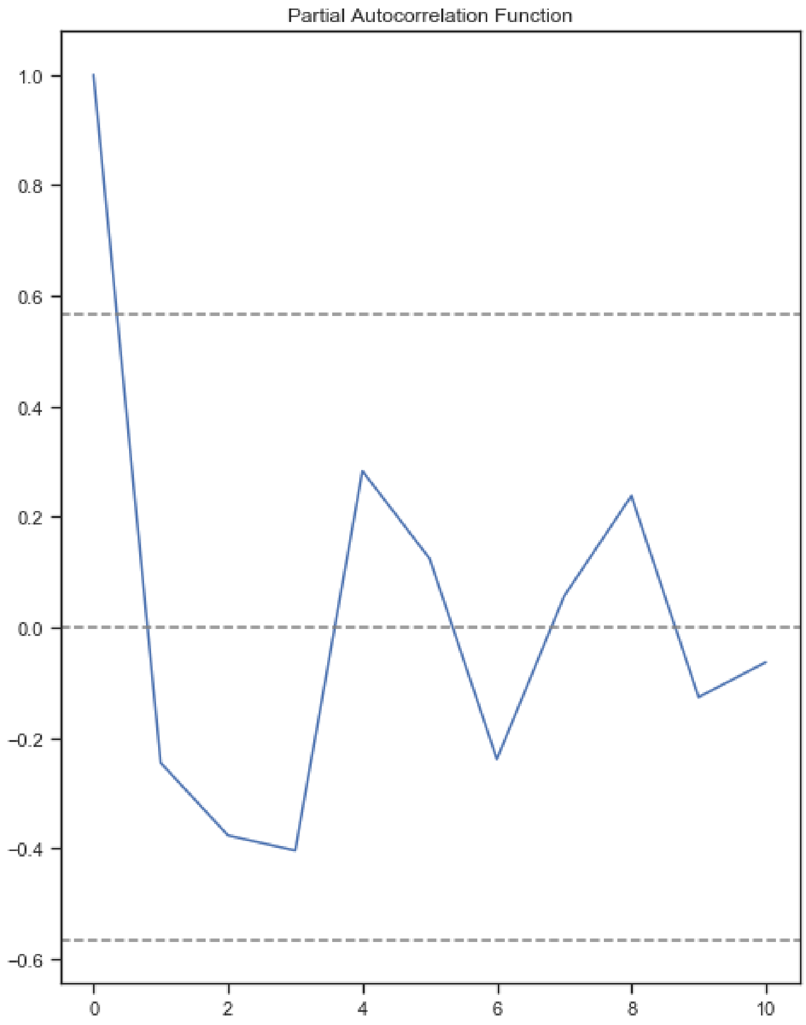
Since Test Statistic(-3.069239) < Critical Value for 10% (-2.748700), timeseries ts_diff is stationary.

Auto Correlation Function (ACF) and **Partial Auto Correlation Function (PACF)** were plotted to find the values of p (AR order) and q (MA order).

ACF plot



PACF plot



It was found from the above charts that p=0 & q=0.

ARIMA stands for **Auto-Regressive Integrated Moving Averages**. The predictors depend on the parameters (p, d, q) of the ARIMA model:

1. **Number of AR (Auto-Regressive) terms (p):** AR terms are just lags of dependent variable.
2. **Number of MA (Moving Average) terms (q):** MA terms are lagged forecast errors in prediction equation.
3. **Number of Differences (d):** These are the number of differences.

We will have to check for several combinations of p & q to decide which model is best for forecasting.

ARIMA model was applied for several combinations of p, d, q and Residual Sum of Squares(RSS) was calculated to check which combination gives lowest RSS.

RSS for various order combinations:

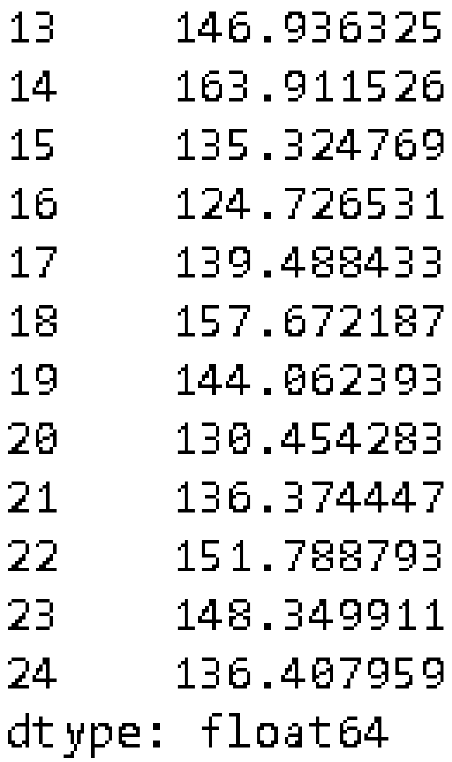
a. RSS = 6.179 for order = (1,2,1)
b. RSS = 1.051 for order = (1,0,1)
c. RSS = 2.277 for order = (1,1,1)
d. RSS = 1.018 for order = (2,0,1)
e. RSS = 2.082 for order = (2,1,1)
f. RSS = 6.802 for order = (2,2,1)
g. RSS = 0.944 for order = (3,0,0)

Order = (3,0,0) gives lowest RSS of 0.944. So, we will use order = (3,0,0)

Forecast for next 12 months was done with ARIMA (order=(3,0,0)) model using predict function.

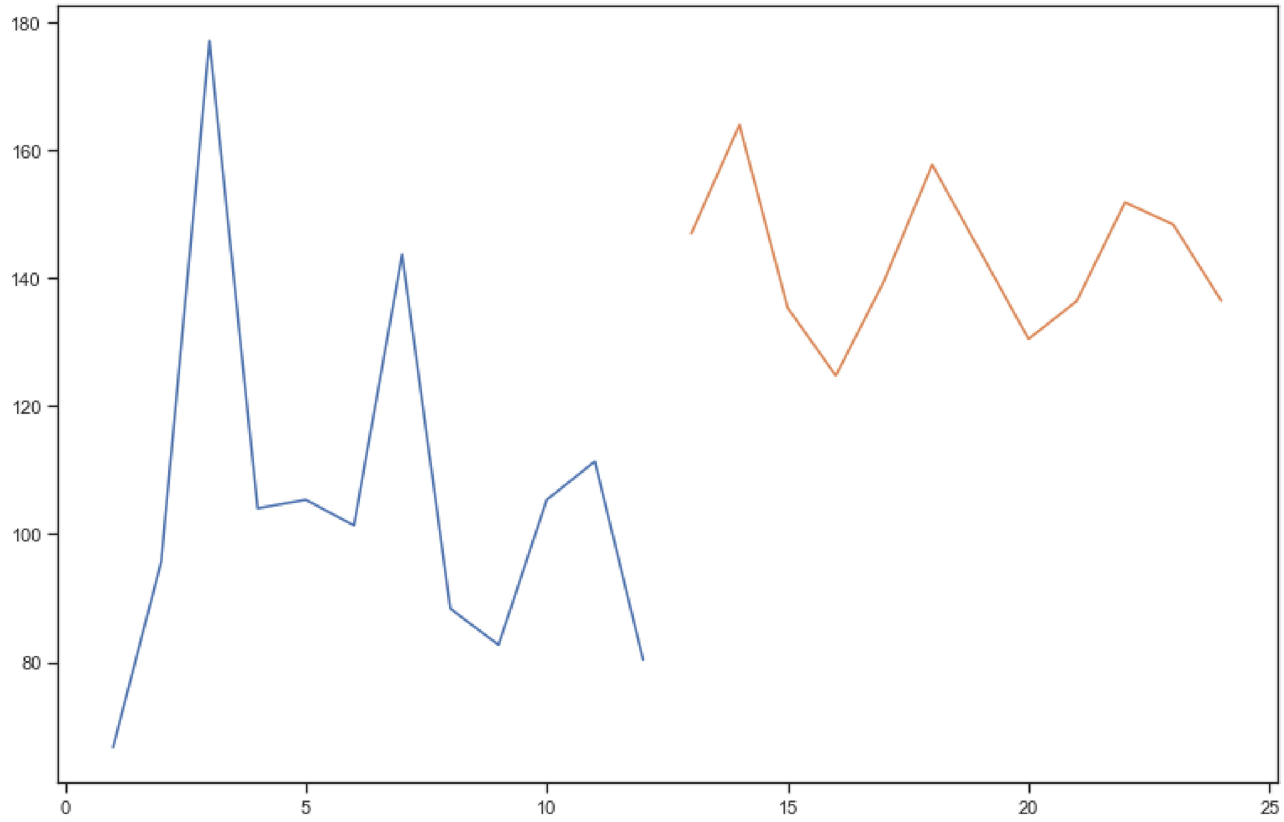
Predicted values were scaled back to original scale.

Prediction values (after scaling) are as under:



```
13    146.936325
14    163.911526
15    135.324769
16    124.726531
17    139.488433
18    157.672187
19    144.062393
20    130.454283
21    136.374447
22    151.788793
23    148.349911
24    136.407959
dtype: float64
```

Plot of original time series and forecast values:



Modelling in R

Chapter 4

Exploratory Data Analysis

Data has been read:

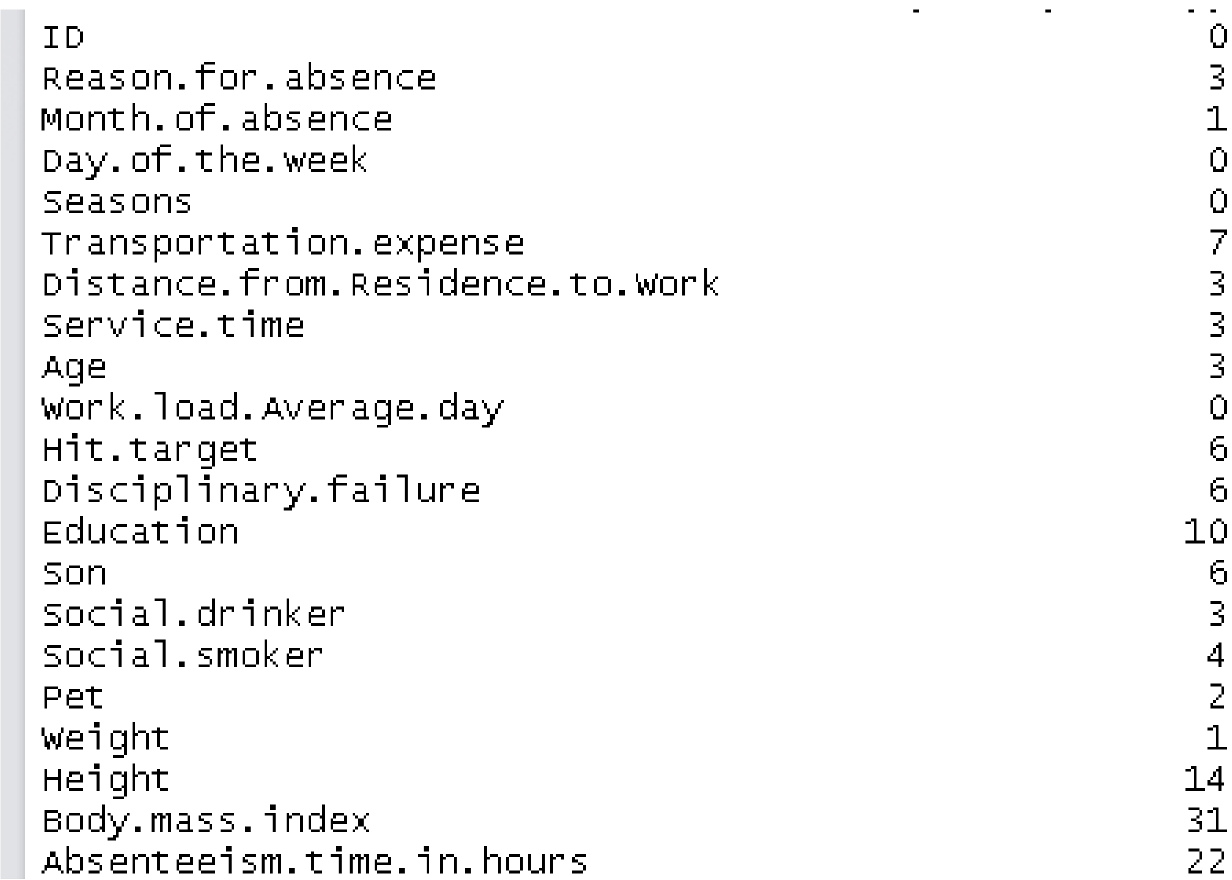
```
# 'data.frame':      740 obs. of  21 variables:
# $ ID                               : int  11 36 3 7 11 3 10 20 14 1 ...
# $ Reason.for.absence               : int  26 0 23 7 23 23 22 23 19 22 ...
# $ Month.of.absence                 : int  7 7 7 7 7 7 7 7 7 7 ...
# $ Day.of.the.week                  : int  3 3 4 5 5 6 6 6 2 2 ...
# $ Seasons                          : int  1 1 1 1 1 1 1 1 1 1 ...
# $ Transportation.expense           : int  289 118 179 279 289 179 NA ...
# $ Distance.from.Residence.to.Work :  int  36 13 51 5 36 51 52 50 12 11 ...
# $ Service.time                     : int  13 18 18 14 13 18 3 11 14 14 ...
# $ Age                              : int  33 50 38 39 33 38 28 36 34 37 ...
# $ Work.load.Average.day            : Factor w/ 39 levels "","205,917","222,196"
# $ Hit.target                       : int  97 97 97 97 97 97 97 ...
# $ Disciplinary.failure             : int  0 1 0 0 0 0 0 0 0 0 ...
# $ Education                        : int  1 1 1 1 1 1 1 1 1 3 ...
# $ Son                              : int  2 1 0 2 2 0 1 4 2 1 ...
# $ Social.drinker                   : int  1 1 1 1 1 1 1 1 1 0 ...
# $ Social.smoker                    : int  0 0 0 1 0 0 0 0 0 0 ...
# $ Pet                              : int  1 0 0 0 1 0 4 0 0 1 ...
# $ Weight                           : int  90 98 89 68 90 89 80 65 95 88 ...
# $ Height                           : int  172 178 170 168 172 170 172 168 ...
# $ Body.mass.index                  : int  30 31 31 24 30 31 27 23 ...
```

$ Absenteeism.time.in.hours   : int 4 0 2 4 2 NA 8 4 40 8 ...

Work.load.Average.day variable values had commas which have been removed.

4.1 Missing Values Analysis

Missing Values are as shown:

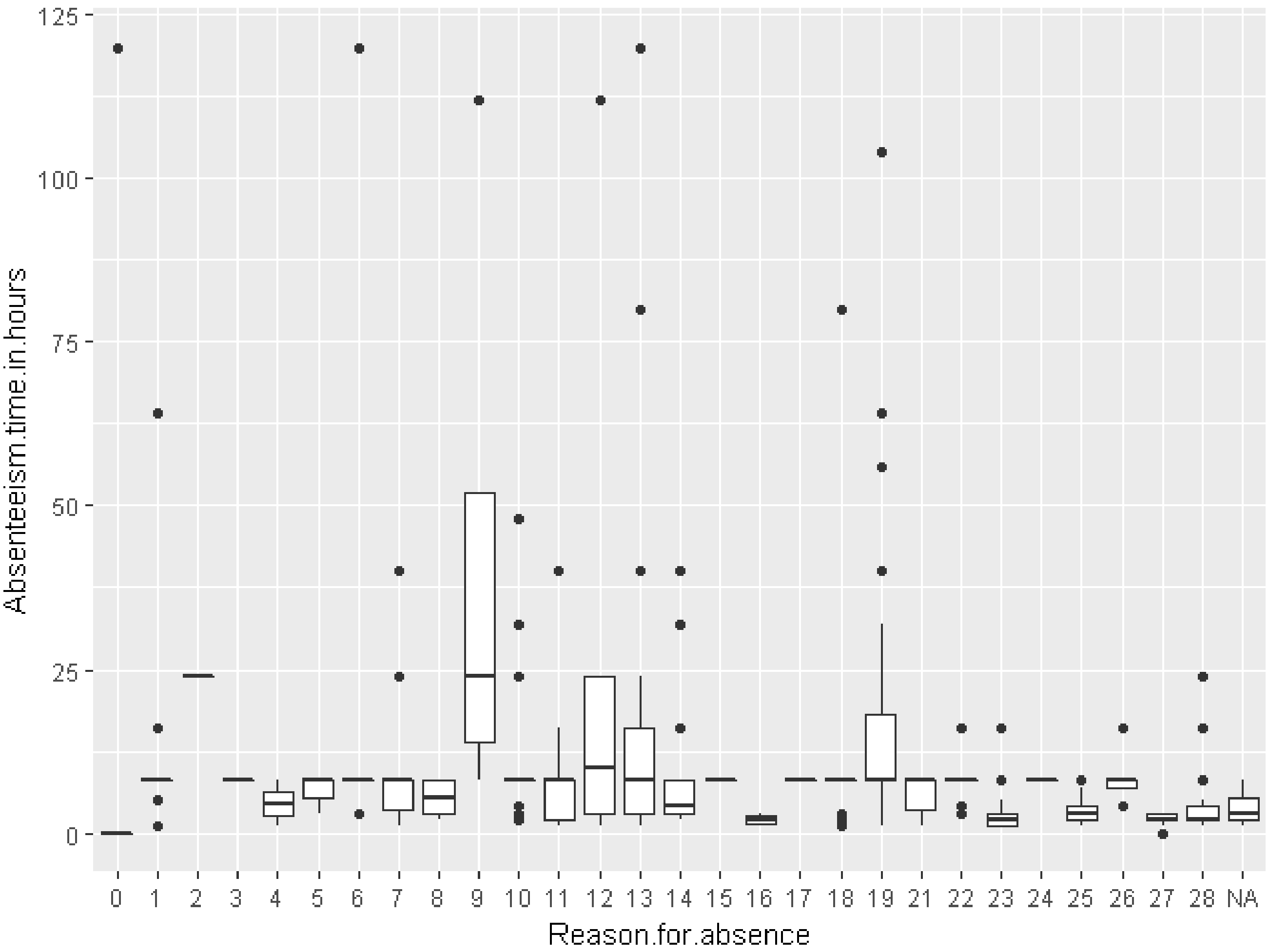


```
ID                                                0
Reason.for.absence                                3
Month.of.absence                                  1
Day.of.the.week                                   0
Seasons                                           0
Transportation.expense                            7
Distance.from.Residence.to.Work                   3
Service.time                                      3
Age                                               3
Work.load.Average.day                             0
Hit.target                                        6
Disciplinary.failure                              6
Education                                        10
Son                                               6
Social.drinker                                    3
Social.smoker                                     4
Pet                                               2
Weight                                            1
Height                                           14
Body.mass.index                                  31
Absenteeism.time.in.hours                        22
```

1. Reason for absence

   Pattern between 'Reason for absence' and 'Absenteeism time in hours' may be used to impute missing values in 'Reason for absence'.

   Box plot for 'Reason for absence' and 'Absenteeism time in hours' :



Category 27 of ‘Reason for absence’ is taking < 10 hrs of ‘Absenteeism time in hours’. So, null values of ‘Reason for absence’ are put equal to 27 since ‘Absenteeism time in hours’ for the observations having null values is < 10 hrs.

Zero category of 'Reason for absence' column has been put equal to category 26(i.e. unjustified absence).

2. Month of absence

   Putting 'Month of absence' null value equal to 10.

3. Transportation expense

   'ID' column has been used to impute missing value for 'Transportation expense'.

4. Distance from Residence to Work

   'ID' column has been used to impute missing value for 'Distance from Residence to Work'.

5. Service time

   'ID' column has been used to impute missing value for ‘Service time’.

6. Age

   'ID' column has been used to impute missing value for ‘Age’.

7. Work Load Average/day

   'Work load Average/day' values are dependent upon 'Month of absence' and 'Hit target' values.

8. Hit target

   'Hit target' values are dependent upon 'Month of absence' and 'Work load Average/day' values.

9. Disciplinary failure

   'Disciplinary failure' missing values have been put to 0.

10. Education

    'ID' column has been used to impute missing value for ‘Education’.

11. Son

    'ID' column has been used to impute missing value for ‘Son’.

12. Social drinker

    'ID' column has been used to impute missing value for ‘Social drinker’.

13. Social smoker

    'ID' column has been used to impute missing value for ‘Social smoker’.

14. Pet

'ID' column has been used to impute missing value for ‘Pet’.

15. Weight

'ID' column has been used to impute missing value for ‘Weight’.

16. Height

'ID' column has been used to impute missing value for ‘Height’.

17. Body mass index

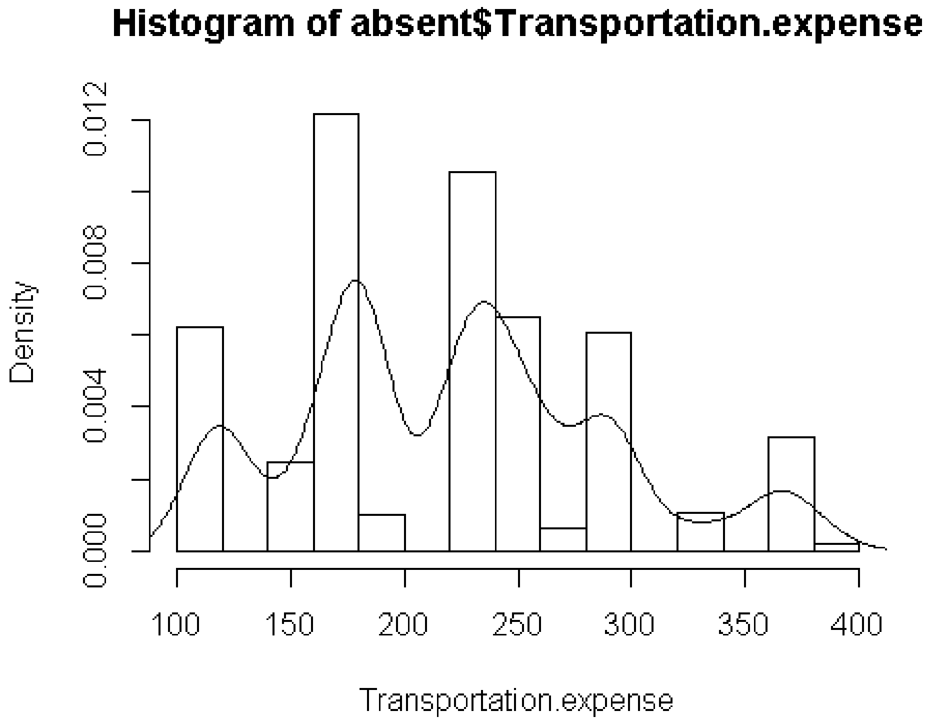
'ID' column has been used to impute missing value for ‘Body mass index’.

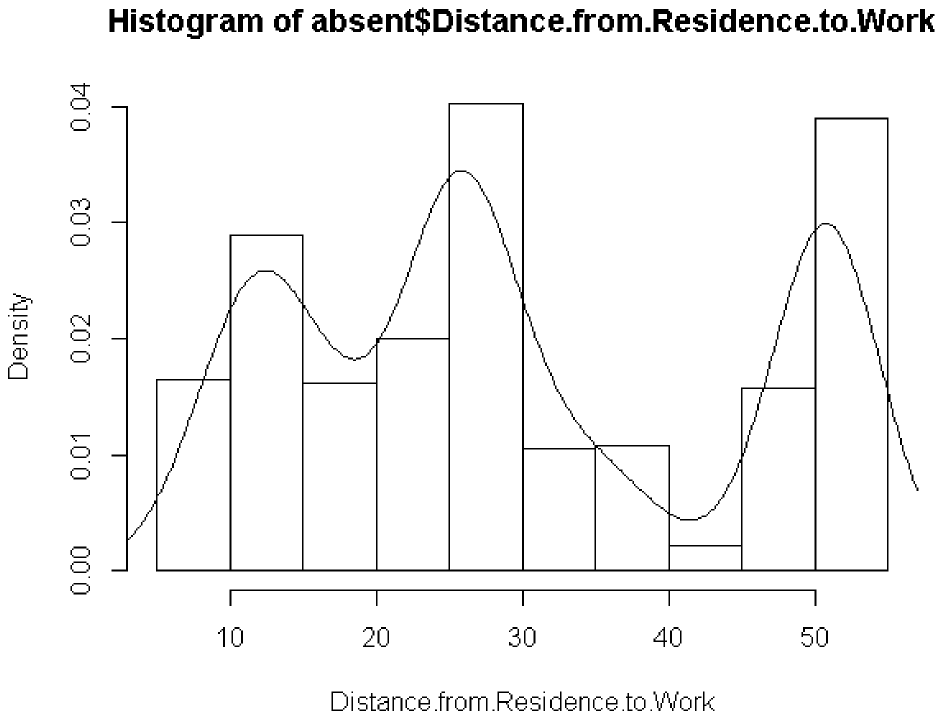
18. Absenteeism time in hours

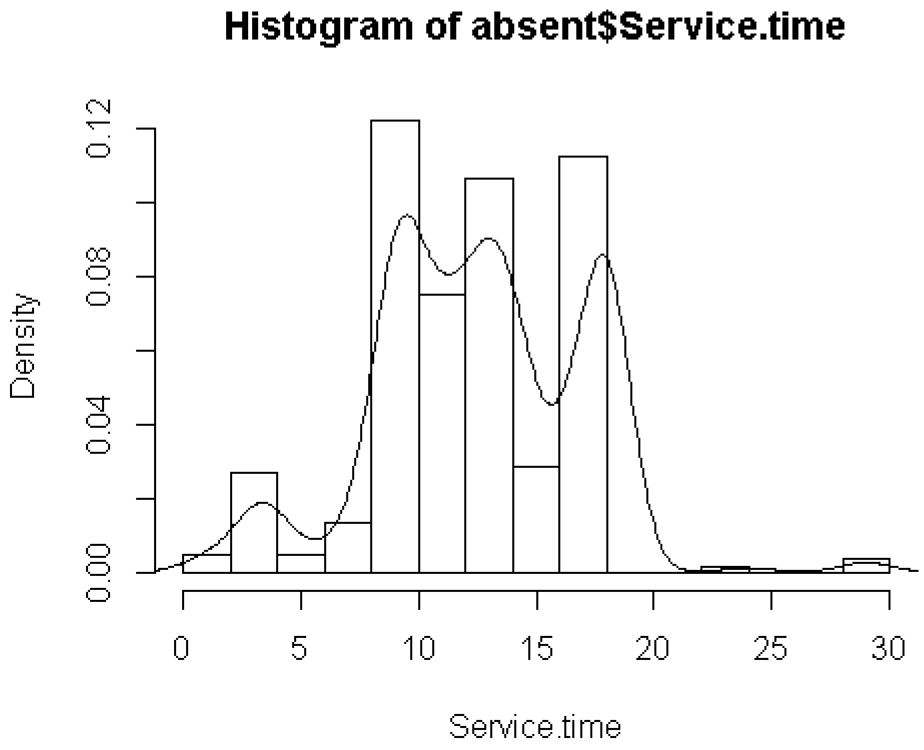
'Reason for absence' column has been used to impute missing value for ‘Absenteeism time in hours’.

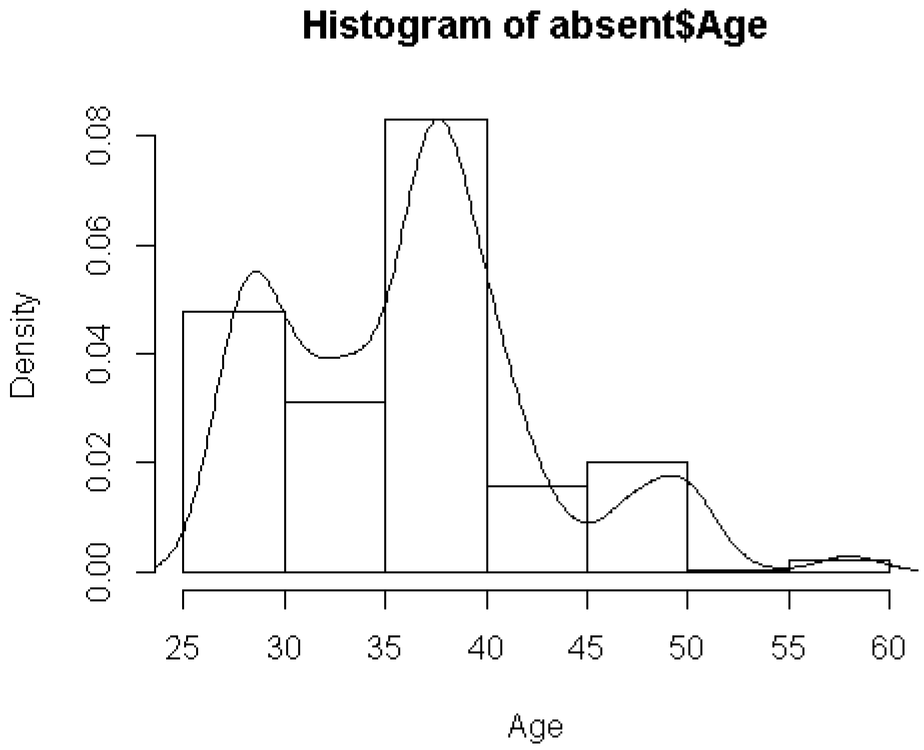
All variables missing values have been imputed.

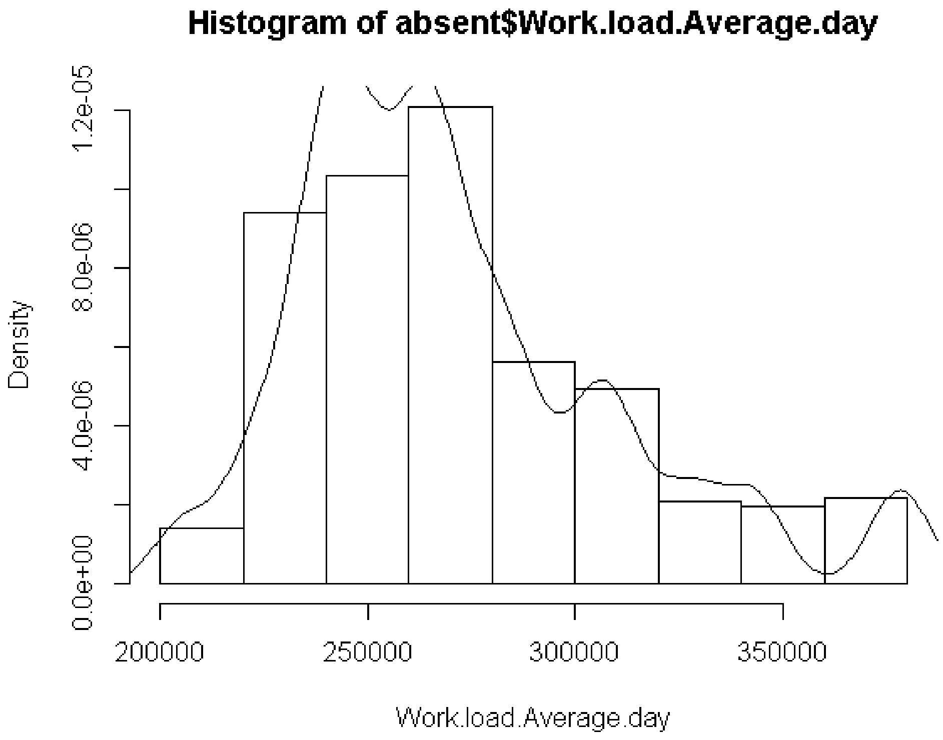
4.2 Distributions

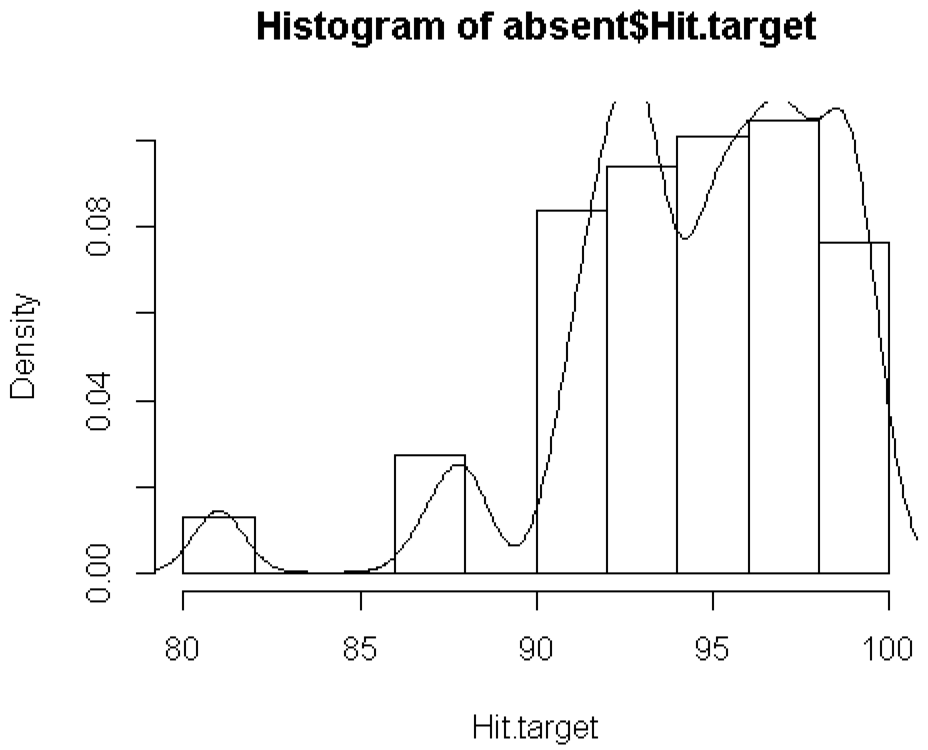


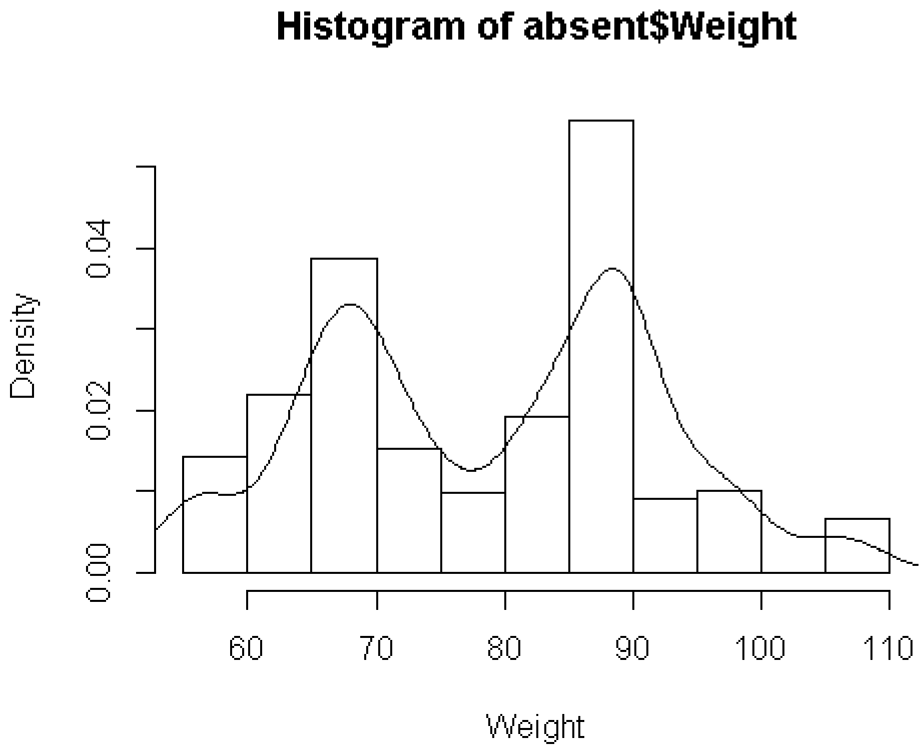


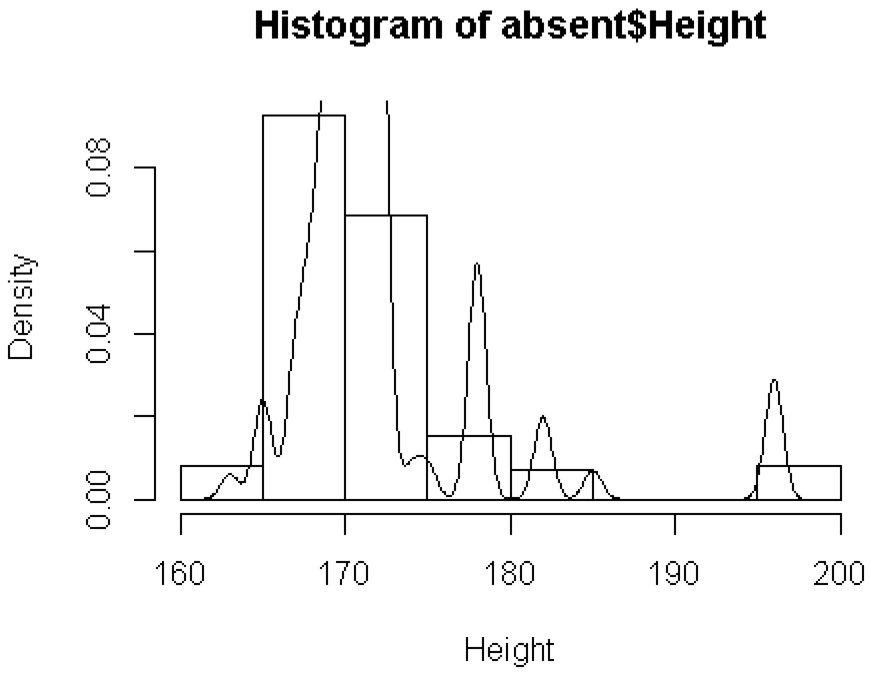

Histogram of absent$Service.time
Density
0.00
0.04
0.08
0.12
0
5
10
15
20
25
30
Service.time

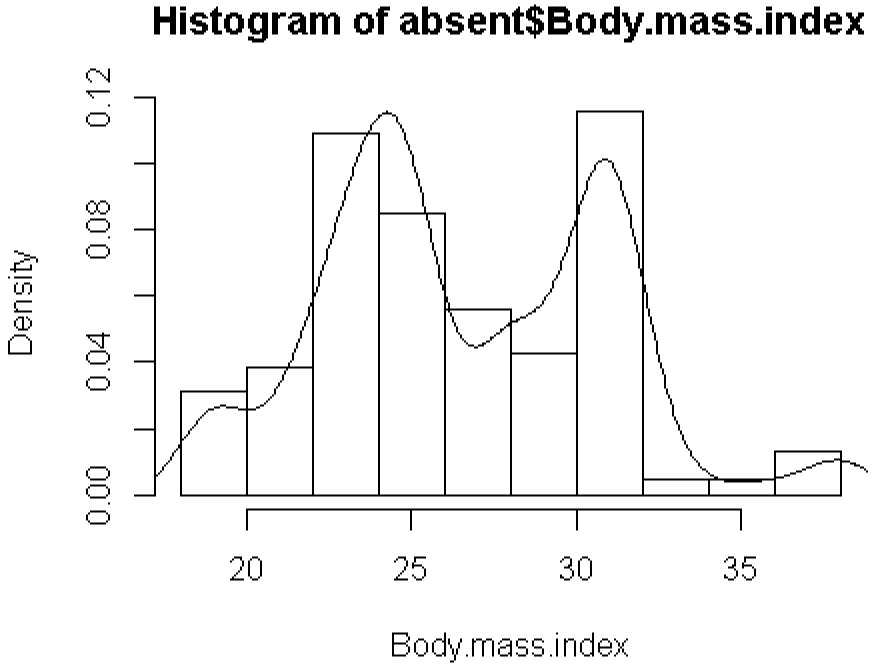

Histogram of absent$Age
Density
0.00
0.02
0.04
0.06
0.08
25
30
35
40
45
50
55
60
Age


Histogram of absent$Work.load.Average.day
Density
0.0e+00
4.0e-06
8.0e-06
1.2e-05
200000
250000
300000
350000
Work.load.Average.day


Histogram of absent$Hit.target
Density
0.00
0.04
0.08
80
85
90
95
100
Hit.target


Histogram of absent$Weight
Density
0.00
0.02
0.04
60
70
80
90
100
110
Weight

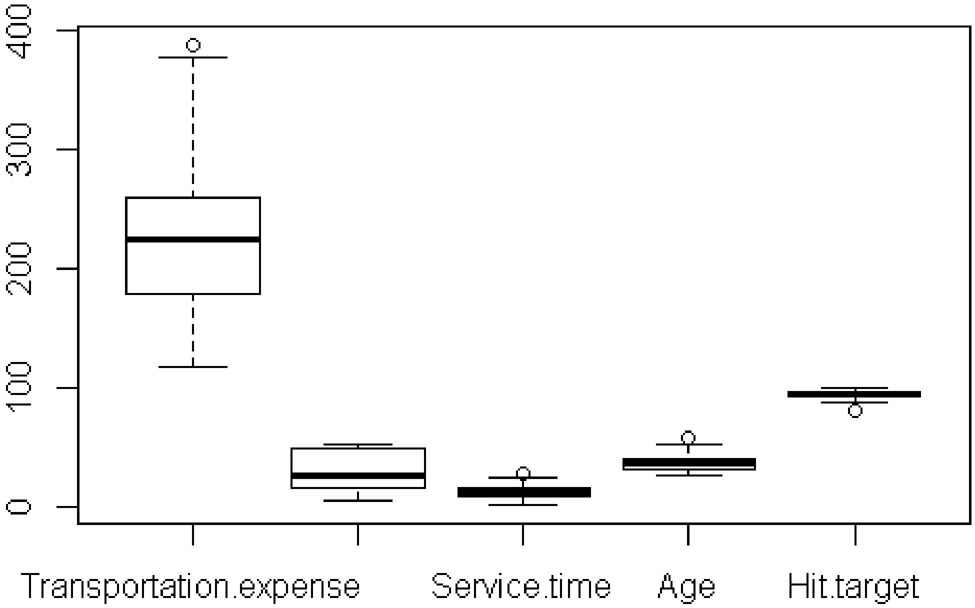

Histogram of absent$Height
Density
0.00
0.04
0.08
160
170
180
190
200
Height



All continuous variables have skewed distribution.

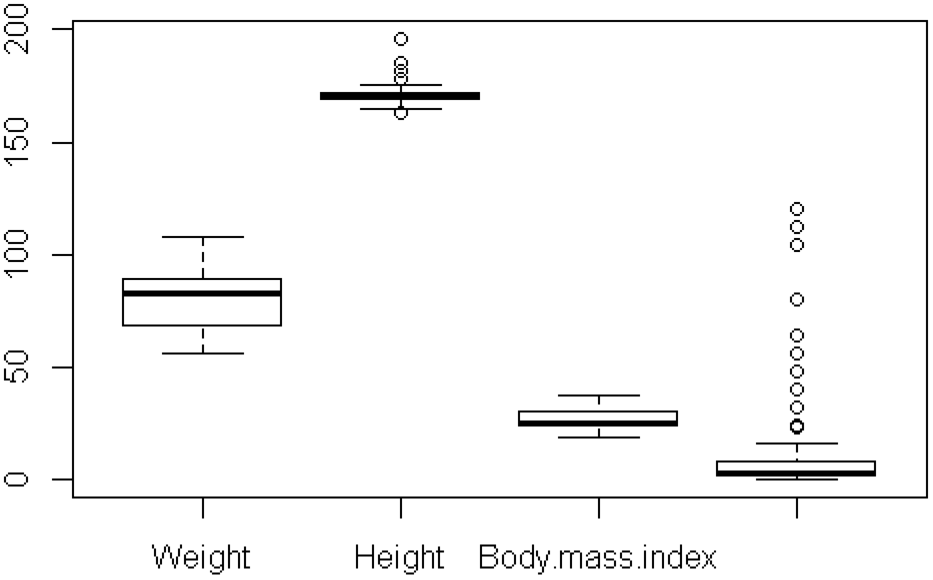
4.3 Outliers Analysis

Boxplots for variables – Transportation.expense, Distance.from.Residence.to.Work, Service.time, Age, Hit.target



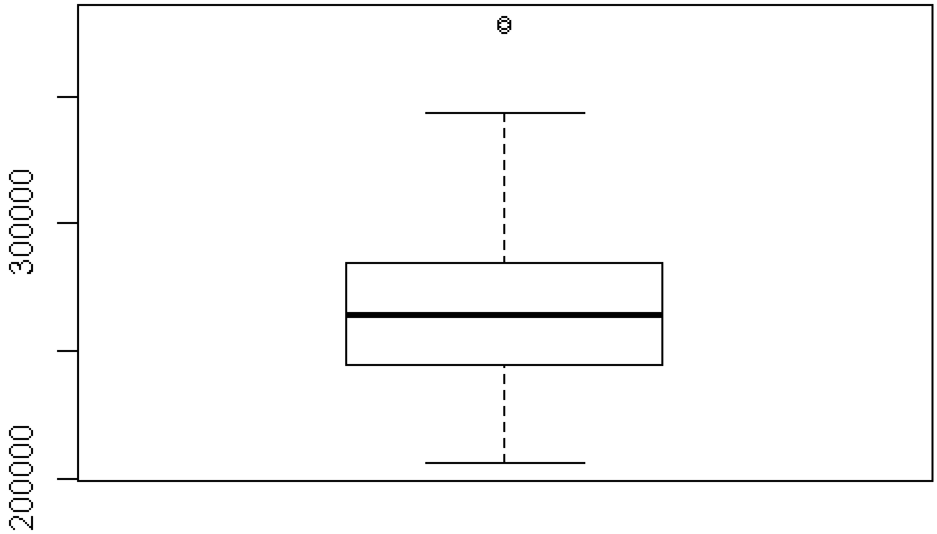
Transportation.expense, Service.time, Age, Hit.target have outliers.

Boxplot for variables - Weight, Height, Body.mass.index, Absenteeism.time.in.hours



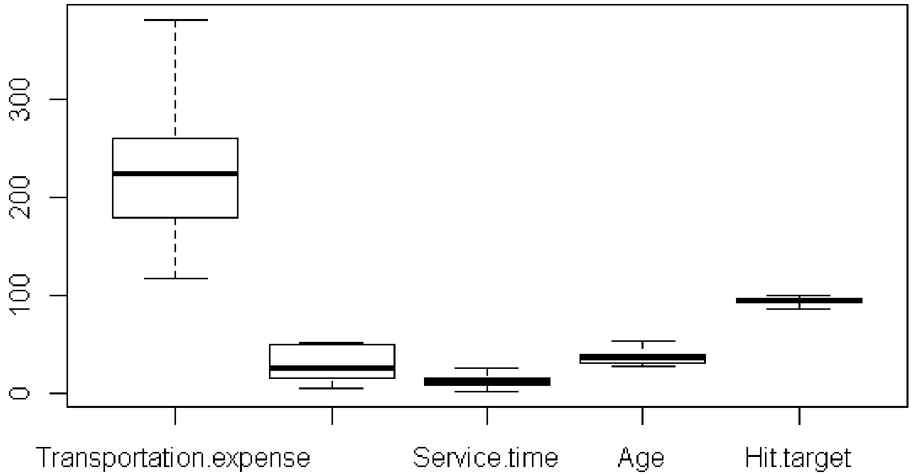
Height and Absenteeism.time.in.hours have outliers.

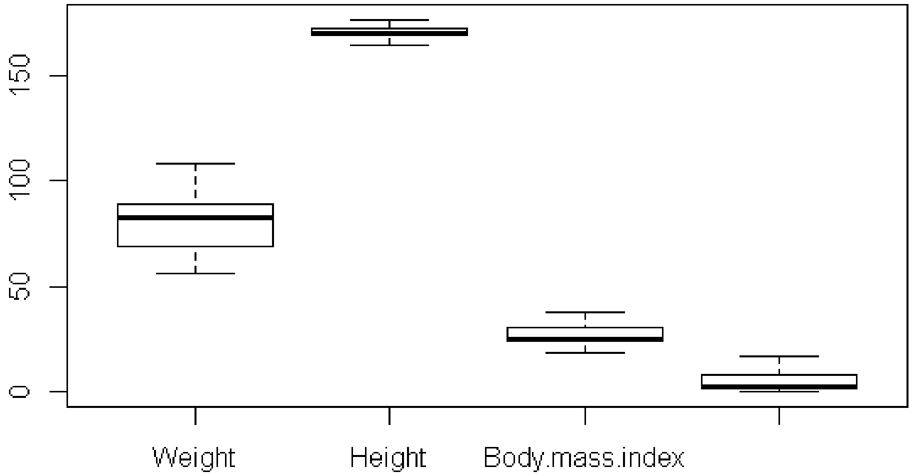
Boxplot for Work.load.Average.day

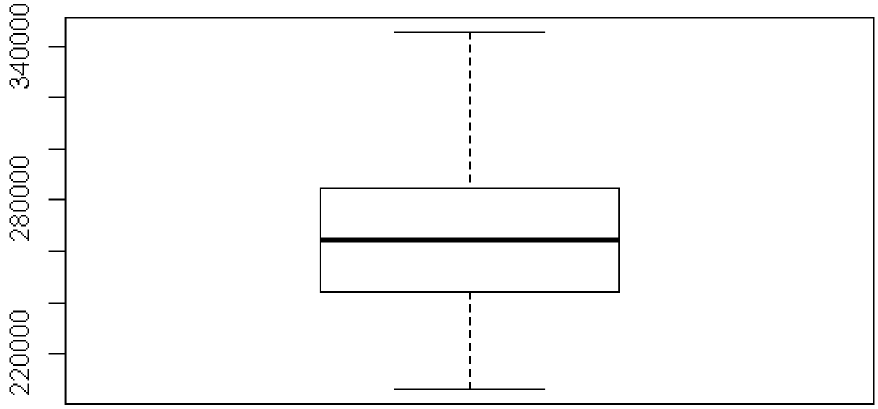


Work.load.Average.day has outliers.

All the outliers have been capped.

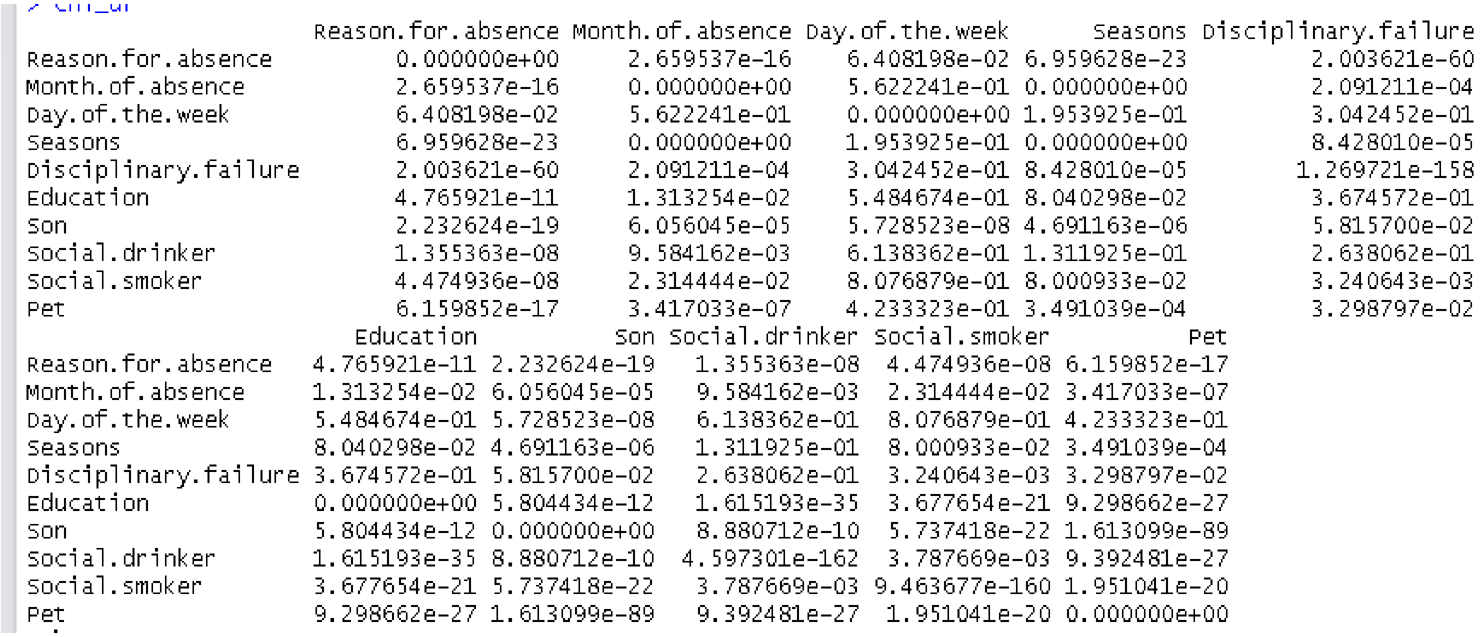

0
100
200
300
Transportation.expense
Service.time
Age
Hit.target


0
50
100
150
Weight
Height
Body.mass.index



4.4 Correlation Analysis

Variables - Reason.for.absence, Month.of.absence, Day.of.the.week, Seasons, Disciplinary.failure, Education, Son, Social.drinker, Social.smoker, Pet have been converted to factor. **Chi-square test has been done for finding correlation between factors. p-values are shown:**

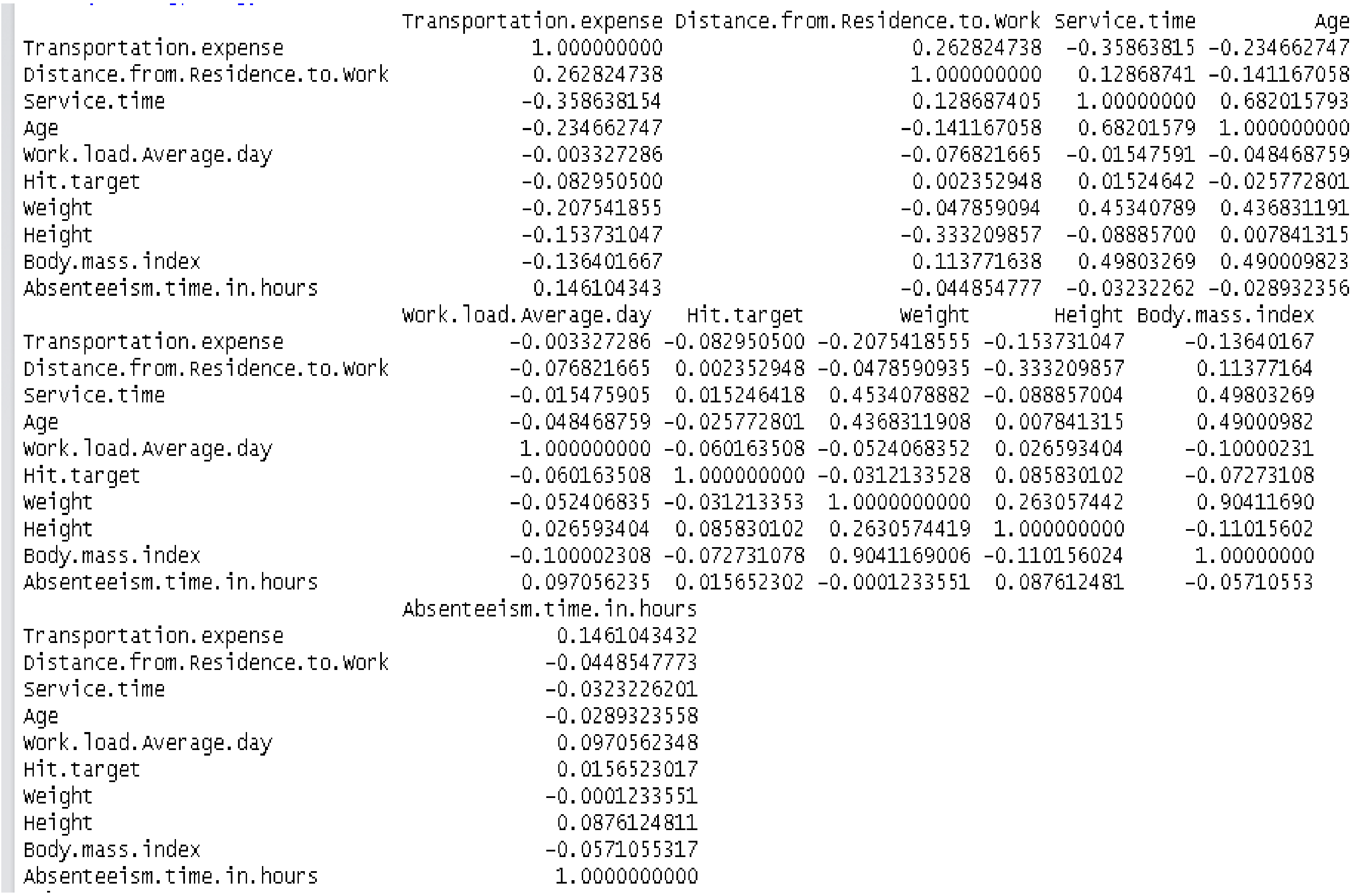


```
> chi_df
                     Reason.for.absence Month.of.absence Day.of.the.week      Seasons Disciplinary.failure
Reason.for.absence         0.000000e+00     2.659537e-16    6.408198e-02 6.959628e-23         2.003621e-60
Month.of.absence           2.659537e-16     0.000000e+00    5.622241e-01 0.000000e+00         2.091211e-04
Day.of.the.week            6.408198e-02     5.622241e-01    0.000000e+00 1.953925e-01         3.042452e-01
Seasons                    6.959628e-23     0.000000e+00    1.953925e-01 0.000000e+00         8.428010e-05
Disciplinary.failure       2.003621e-60     2.091211e-04    3.042452e-01 8.428010e-05        1.269721e-158
Education                  4.765921e-11     1.313254e-02    5.484674e-01 8.040298e-02         3.674572e-01
Son                        2.232624e-19     6.056045e-05    5.728523e-08 4.691163e-06         5.815700e-02
Social.drinker             1.355363e-08     9.584162e-03    6.138362e-01 1.311925e-01         2.638062e-01
Social.smoker              4.474936e-08     2.314444e-02    8.076879e-01 8.000933e-02         3.240643e-03
Pet                        6.159852e-17     3.417033e-07    4.233323e-01 3.491039e-04         3.298797e-02
                        Education          Son Social.drinker Social.smoker          Pet
Reason.for.absence   4.765921e-11 2.232624e-19   1.355363e-08  4.474936e-08 6.159852e-17
Month.of.absence     1.313254e-02 6.056045e-05   9.584162e-03  2.314444e-02 3.417033e-07
Day.of.the.week      5.484674e-01 5.728523e-08   6.138362e-01  8.076879e-01 4.233323e-01
Seasons              8.040298e-02 4.691163e-06   1.311925e-01  8.000933e-02 3.491039e-04
Disciplinary.failure 3.674572e-01 5.815700e-02   2.638062e-01  3.240643e-03 3.298797e-02
Education            0.000000e+00 5.804434e-12   1.615193e-35  3.677654e-21 9.298662e-27
Son                  5.804434e-12 0.000000e+00   8.880712e-10  5.737418e-22 1.613099e-89
Social.drinker       1.615193e-35 8.880712e-10  4.597301e-162  3.787669e-03 9.392481e-27
Social.smoker        3.677654e-21 5.737418e-22   3.787669e-03 9.463677e-160 1.951041e-20
Pet                  9.298662e-27 1.613099e-89   9.392481e-27  1.951041e-20 0.000000e+00
```

p-values are <0.05 for chi-square test of all categorical variables with Reason.for.absence except Day.of.the.week.This means that categorical variables having p-values<0.05 have dependence on Reason.for.absence. So, all categorical variables except Reason.for.absence and Day.of.the.week will be dropped.

Month.of.absence, Seasons, Disciplinary.failure, Education,Son, Social.drinker,
Social.smoker and Pet have been dropped.

Correlation between continuous independent & dependent variables



```
                                Transportation.expense Distance.from.Residence.to.Work Service.time          Age
Transportation.expense                     1.000000000                     0.262824738  -0.35863815 -0.234662747
Distance.from.Residence.to.Work            0.262824738                     1.000000000   0.12868741 -0.141167058
Service.time                              -0.358638154                     0.128687405   1.00000000  0.682015793
Age                                       -0.234662747                    -0.141167058   0.68201579  1.000000000
Work.load.Average.day                     -0.003327286                    -0.076821665  -0.01547591 -0.048468759
Hit.target                                -0.082950500                     0.002352948   0.01524642 -0.025772801
Weight                                    -0.207541855                    -0.047859094   0.45340789  0.436831191
Height                                    -0.153731047                    -0.333209857  -0.08885700  0.007841315
Body.mass.index                           -0.136401667                     0.113771638   0.49803269  0.490009823
Absenteeism.time.in.hours                  0.146104343                    -0.044854777  -0.03232262 -0.028932356
                                Work.load.Average.day   Hit.target        Weight       Height Body.mass.index
Transportation.expense                   -0.003327286 -0.082950500 -0.2075418555 -0.153731047     -0.13640167
Distance.from.Residence.to.Work          -0.076821665  0.002352948 -0.0478590935 -0.333209857      0.11377164
Service.time                             -0.015475905  0.015246418  0.4534078882 -0.088857004      0.49803269
Age                                      -0.048468759 -0.025772801  0.4368311908  0.007841315      0.49000982
Work.load.Average.day                     1.000000000 -0.060163508 -0.0524068352  0.026593404     -0.10000231
Hit.target                               -0.060163508  1.000000000 -0.0312133528  0.085830102     -0.07273108
Weight                                   -0.052406835 -0.031213353  1.0000000000  0.263057442      0.90411690
Height                                    0.026593404  0.085830102  0.2630574419  1.000000000     -0.11015602
Body.mass.index                          -0.100002308 -0.072731078  0.9041169006 -0.110156024      1.00000000
Absenteeism.time.in.hours                 0.097056235  0.015652302 -0.0001233551  0.087612481     -0.05710553
                                Absenteeism.time.in.hours
Transportation.expense                       0.1461043432
Distance.from.Residence.to.Work             -0.0448547773
Service.time                                -0.0323226201
Age                                         -0.0289323558
Work.load.Average.day                        0.0970562348
Hit.target                                   0.0156523017
Weight                                      -0.0001233551
Height                                       0.0876124811
Body.mass.index                             -0.0571055317
Absenteeism.time.in.hours                    1.0000000000
```

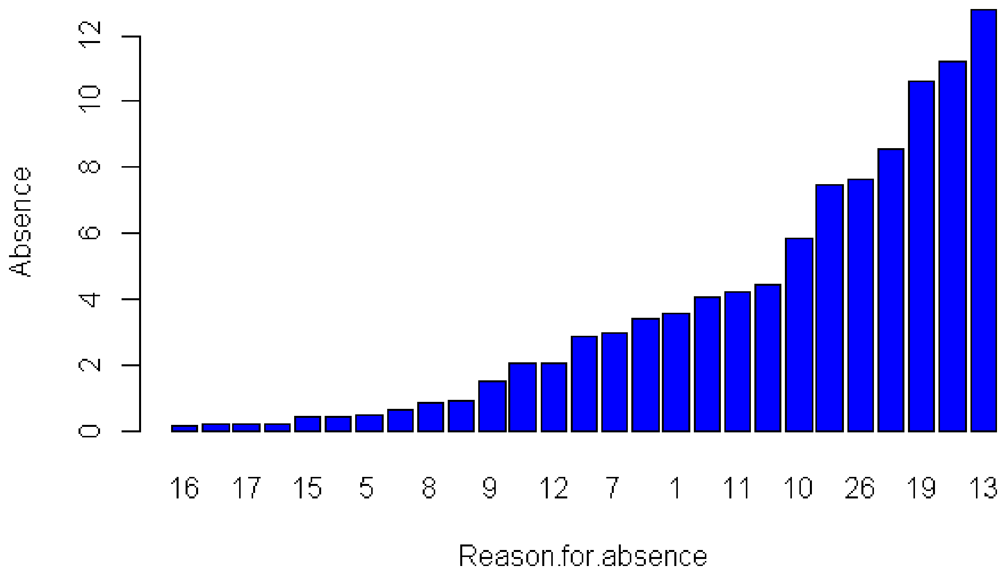
Correlation amongst continuous independent variables < 0.95.

Correlation between every independent variable & dependent variable < 0.2.

This means that there is no relationship between any independent continuous variable and dependent variable.

4.5 Relationships of categorical independent variables with dependent variable

Relationship of Reason.for.absence with Absence (absenteeism time as percent of total time)



Top 5 categories in terms of Absence time are:

1. Category-13 : Diseases of the musculoskeletal system and connective tissue - 12.79 % of total time

2. Category-23 : medical consultation - 11.22 % of total time

3. Category-19 : Injury, poisoning and certain other consequences of external causes - 10.63 % of total time

4. Category 28 : dental consultation - 8.54 % 0f total time

5. Category 26 : unjustified absence - 7.66 % of total time

4.6 Conclusions & Remedies

Conclusions & possible remedies are:

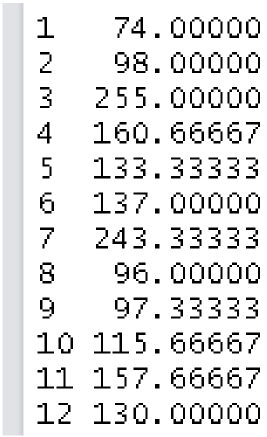
a. Musculoskeletal system disease comes out to be the major reason of absenteeism. Repetitive movement strain due to high workload is one possible reason for the high incidence of musculoskeletal disease. Company should conduct a study on the repetitive movements involved in working setup and try to go for minimizing strain on employees due to work posture. Company should try to optimize workload keeping in mind occupational health of working people.
b. Medical consultation can be brought down by stopping practice of setting unnecessary high targets resulting in high workloads. Target & workloads should be optimized keeping in mind health of employees.
c. Injuries may be reduced by creating proper ergonomic working setup and optimizing workloads.
d. Dental consultation time may be reduced by informing employees of the dental health guidelines so that they can take better care of their teeth.
e. Unjustified absence is too high. Company should try to reduce high workloads so that employees don't feel work stress to take unjustified absence leave. Company may go for giving monetary incentive to employees not taking any unjustified absence.

Chapter 5

Modelling

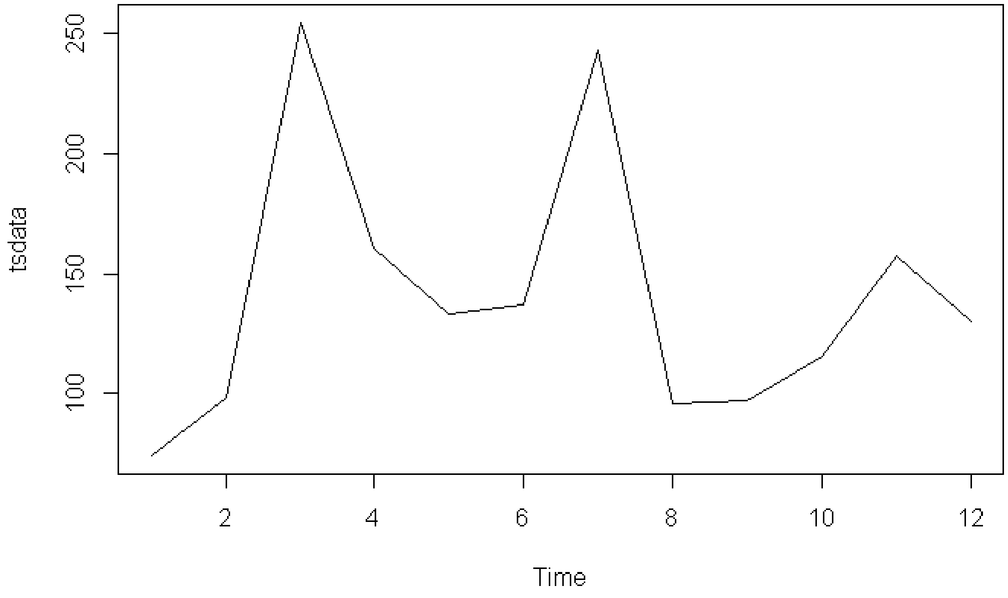
Time Series forecasting will be used to forecast the absenteeism hours for every month in 2011.Absenteeism hours has been aggregated by month to get total absenteeism hours for every month. Since this data is of 3 years from July,2007 to June,2010, Aggregated absenteeism hours has been divided by 3 to get average absent hours for every month.

Original time series is:



```
1    74.00000
2    98.00000
3   255.00000
4   160.66667
5   133.33333
6   137.00000
7   243.33333
8    96.00000
9    97.33333
10  115.66667
11  157.66667
12  130.00000
```

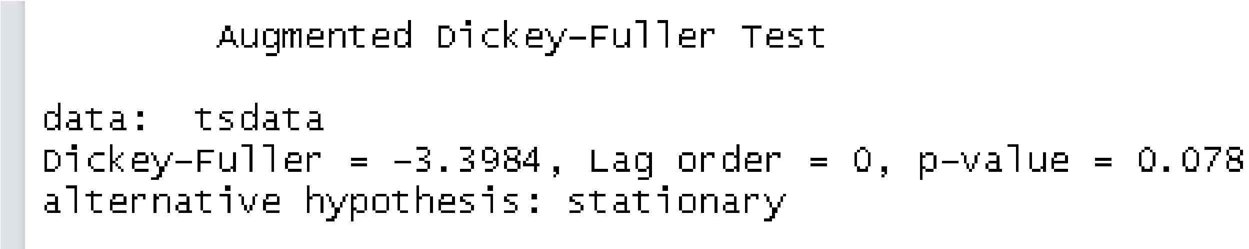
Time Series is as under:



5.1 Time Series - Arima Model

Dickey Fuller test has been done to check if time series is stationary or not.

Results were:



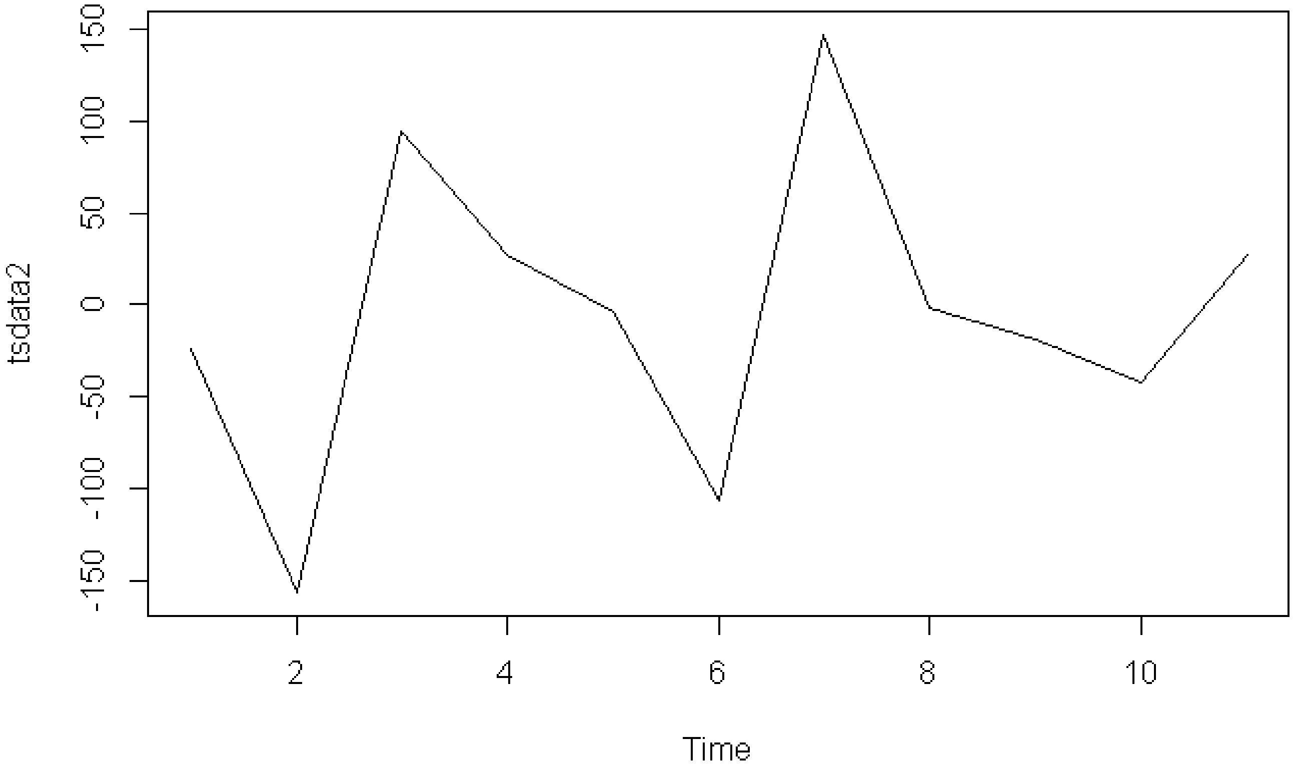
```
        Augmented Dickey-Fuller Test

data:  tsdata
Dickey-Fuller = -3.3984, Lag order = 0, p-value = 0.078
alternative hypothesis: stationary
```

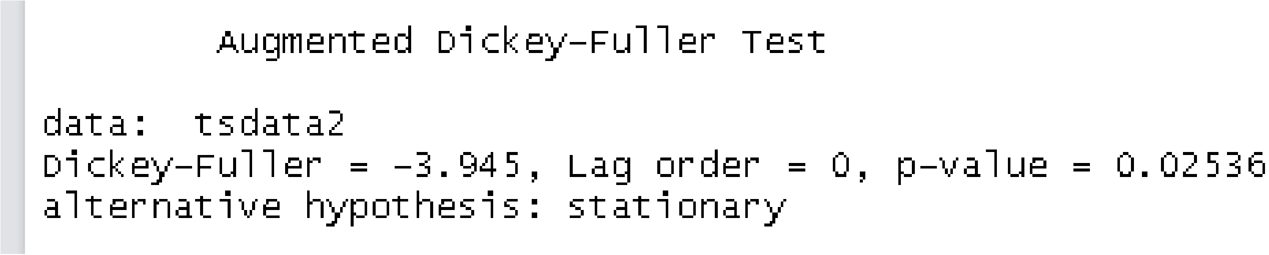
Time series is not stationary as determined by p-value of 0.078(>0.05).

We will be subtracting shifted (single lag) time series from original time series.

Now, modified time series plot is as under:



Augmented Dickey-Fuller Test was done again to check stationarity of modified time series tsdata2.



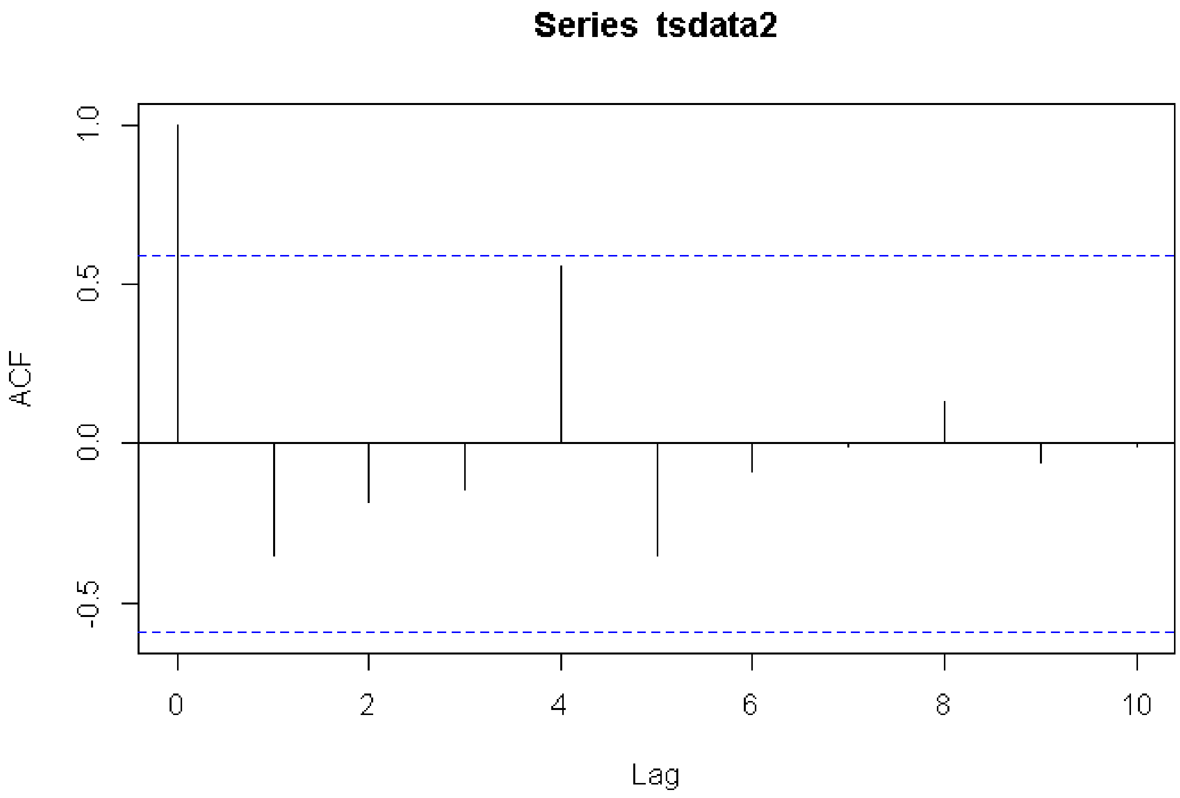
```
        Augmented Dickey-Fuller Test

data:  tsdata2
Dickey-Fuller = -3.945, Lag order = 0, p-value = 0.02536
alternative hypothesis: stationary
```

p-value < 0.05 which means that modified time series tsdata2 is stationary.

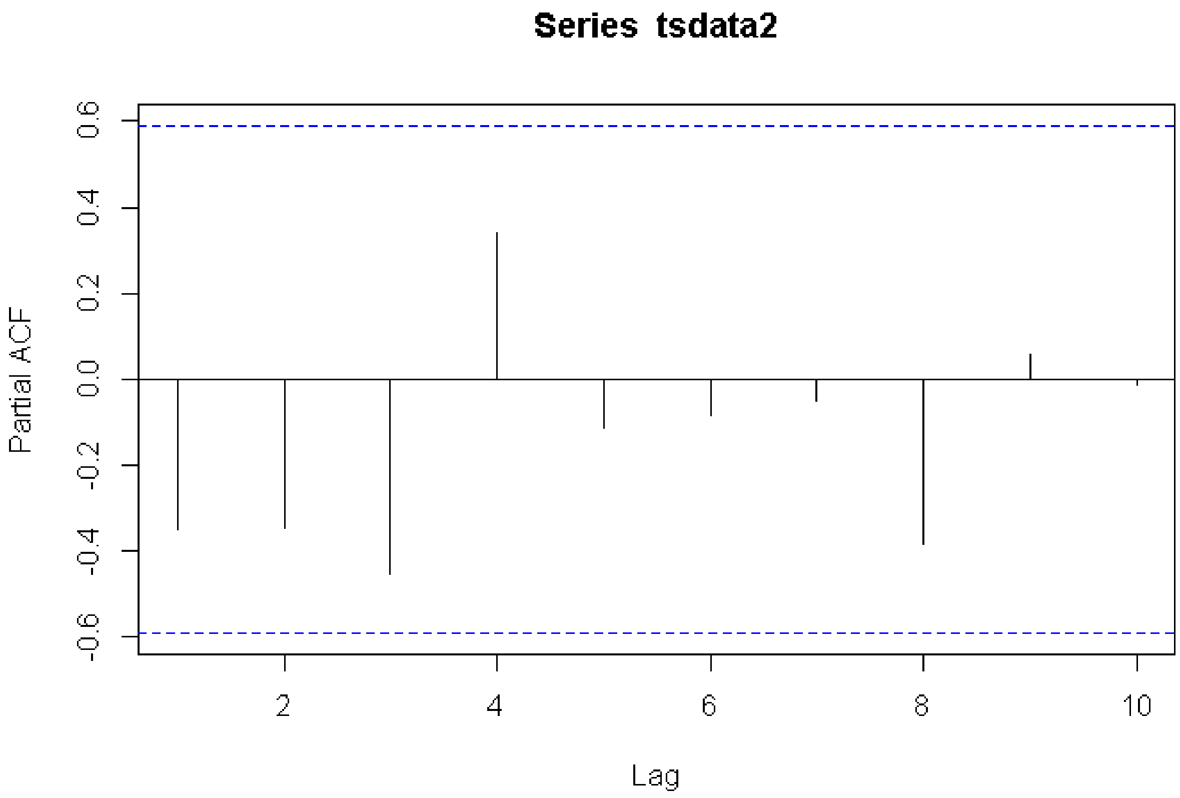
Auto Correlation Function (ACF) and **Partial Auto Correlation Function (PACF)** were plotted to find the values of p (AR order) and q (MA order).

ACF plot



Value of p (order of AR) should be 0 as ACF plot is getting cut off after first line.

PACF plot



Value of q (MA order) should be 0.

We will have to check for several combinations of p & q to decide which model is best for forecasting.

ARIMA model was applied for several combinations of p, d, q and Residual Sum of Squares (RSS) was calculated to check which combination gives lowest RSS.

RSS for various order combinations:

#RSS for order=(1,0,0):61679.74

#RSS for order=(2,0,0):47290.44

#RSS for order=(3,0,0):20928.83

#RSS for order=(4,0,0):20649.88

#RSS for order=(4,0,1):20653.29

#RSS for order=(4,0,2):16526.79

#RSS for order=(4,0,3):10442.2

#RSS for order=(4,0,4):8476.191

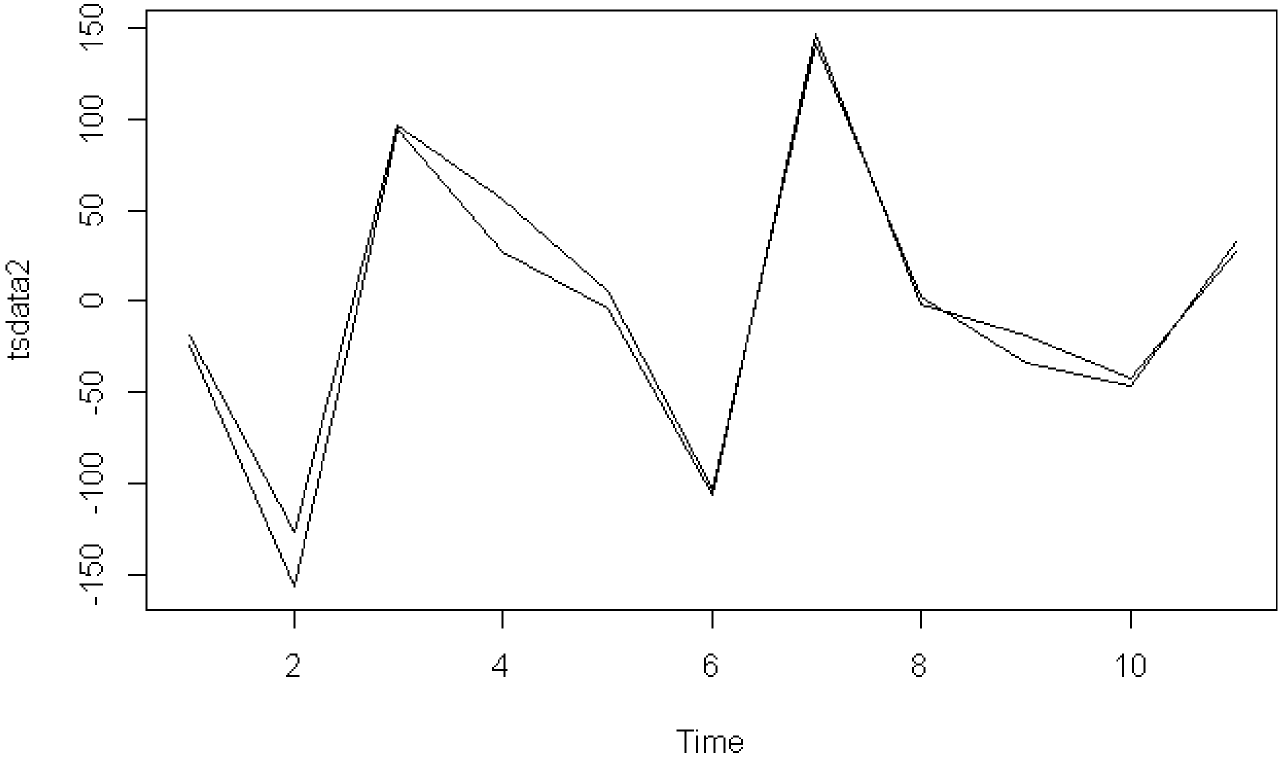
#RSS for order=(4,0,5):8104.803

#RSS for order=(4,0,7):6743.328

#RSS for order=(4,0,9):2222.32

Arima with order=(4,0,9) gives us lowest RSS of 2222.32 so order=(4,0,9).

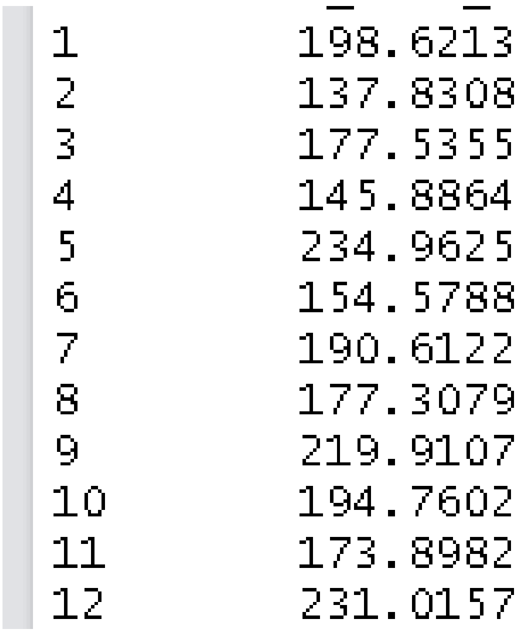
Plot of timeseries and fitted values:



Forecast for next 12 months was done with ARIMA (order=(4,0,9)) model using predict function.

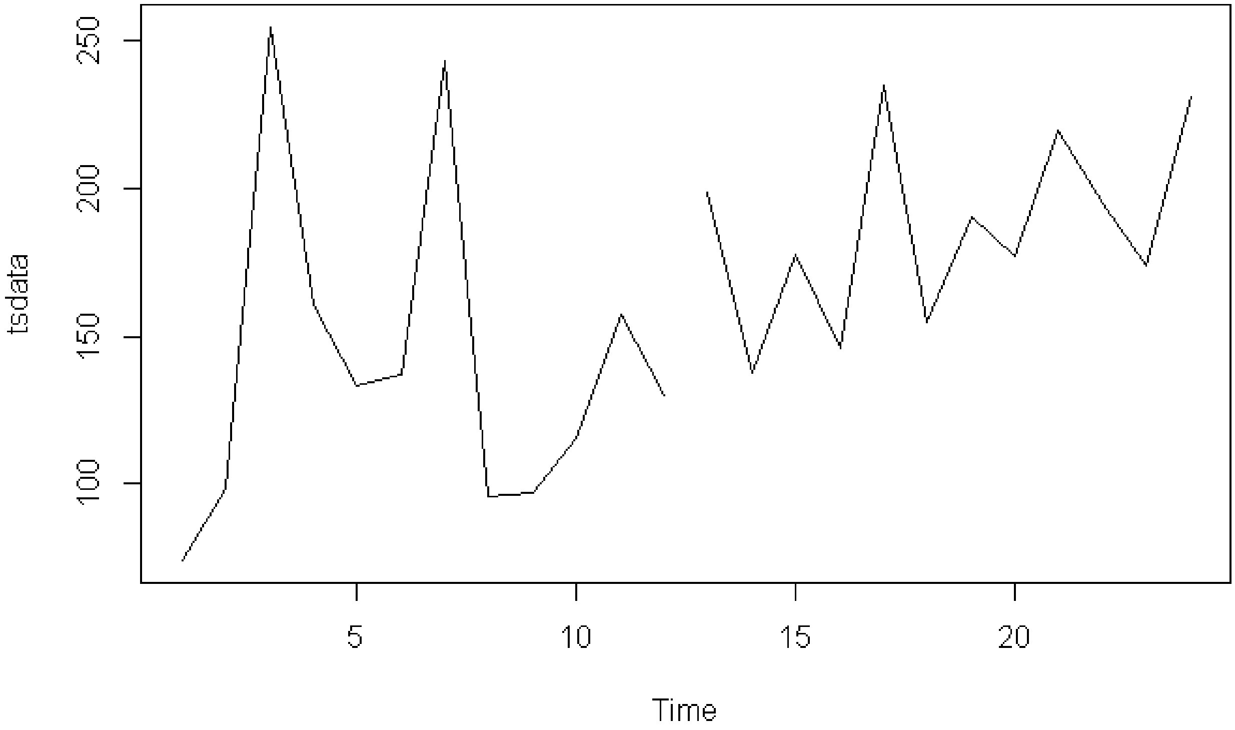
Predicted values were scaled back to original scale.

Predicted time series for 2011 months is:



```
1          198.6213
2          137.8308
3          177.5355
4          145.8864
5          234.9625
6          154.5788
7          190.6122
8          177.3079
9          219.9107
10         194.7602
11         173.8982
12         231.0157
```

Plot of original timeseries & forecast values:


tsdata
Time