Employee Absenteeism

Piyush Prakash

28th July, 2020

Contents

Chapter 1

Assigned Problem & Accompanying Data

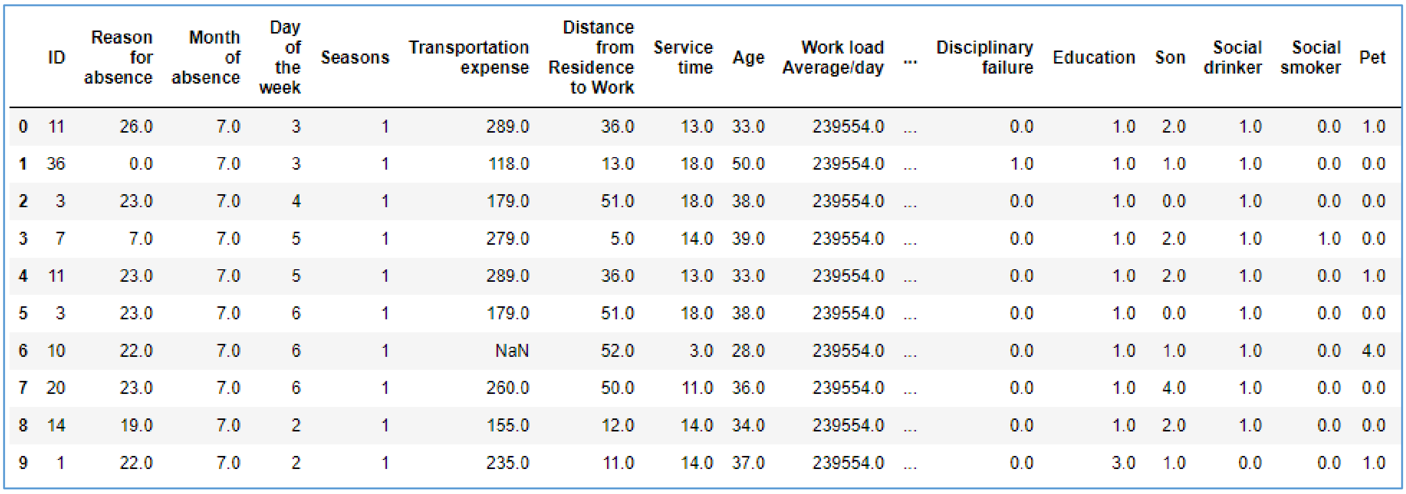
1.1 Problem Statement

XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery; the company is passing through a genuine issue of Absenteeism. The company has shared its dataset and requested to have an answer on the following areas:

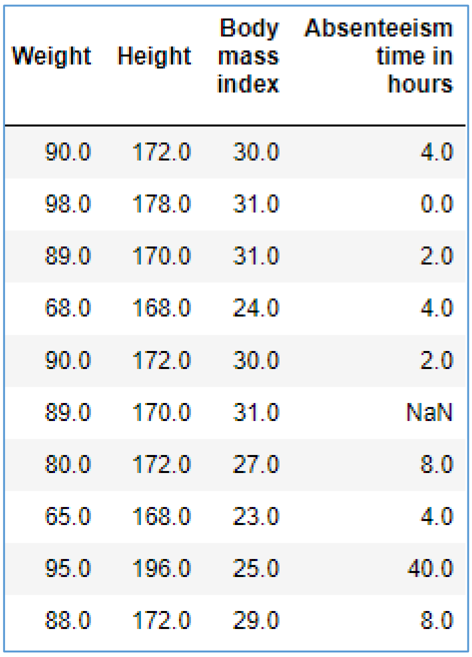
1. What changes the company should bring in to reduce the number of absenteeism hours?
2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

1.2 Data Description

The provided dataset has 21 variables, as shown in the following sample -



	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet
0	11	26.0	7.0	3	1	289.0	36.0	13.0	33.0	239554.0	...	0.0	1.0	2.0	1.0	0.0	1.0
1	36	0.0	7.0	3	1	118.0	13.0	18.0	50.0	239554.0	...	1.0	1.0	1.0	1.0	0.0	0.0
2	3	23.0	7.0	4	1	179.0	51.0	18.0	38.0	239554.0	...	0.0	1.0	0.0	1.0	0.0	0.0
3	7	7.0	7.0	5	1	279.0	5.0	14.0	39.0	239554.0	...	0.0	1.0	2.0	1.0	1.0	0.0
4	11	23.0	7.0	5	1	289.0	36.0	13.0	33.0	239554.0	...	0.0	1.0	2.0	1.0	0.0	1.0
5	3	23.0	7.0	6	1	179.0	51.0	18.0	38.0	239554.0	...	0.0	1.0	0.0	1.0	0.0	0.0
6	10	22.0	7.0	6	1	NaN	52.0	3.0	28.0	239554.0	...	0.0	1.0	1.0	1.0	0.0	4.0
7	20	23.0	7.0	6	1	260.0	50.0	11.0	36.0	239554.0	...	0.0	1.0	4.0	1.0	0.0	0.0
8	14	19.0	7.0	2	1	155.0	12.0	14.0	34.0	239554.0	...	0.0	1.0	2.0	1.0	0.0	0.0
9	1	22.0	7.0	2	1	235.0	11.0	14.0	37.0	239554.0	...	0.0	3.0	1.0	0.0	0.0	1.0



Weight	Height	Body mass index	Absenteeism time in hours
90.0	172.0	30.0	4.0
98.0	178.0	31.0	0.0
89.0	170.0	31.0	2.0
68.0	168.0	24.0	4.0
90.0	172.0	30.0	2.0
89.0	170.0	31.0	NaN
80.0	172.0	27.0	8.0
65.0	168.0	23.0	4.0
95.0	196.0	25.0	40.0
88.0	172.0	29.0	8.0

The variables are described as following:

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)

1.3 Classification of Variables:

Categorical Variables –

1. Individual identification (ID)
2. Reason for absence (ICD)
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. Disciplinary failure (yes=1; no=0)
7. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
8. Son (number of children)
9. Social drinker (yes=1; no=0)
10. Social smoker (yes=1; no=0)
11. Pet (number of pet)

Continuous Variables –

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

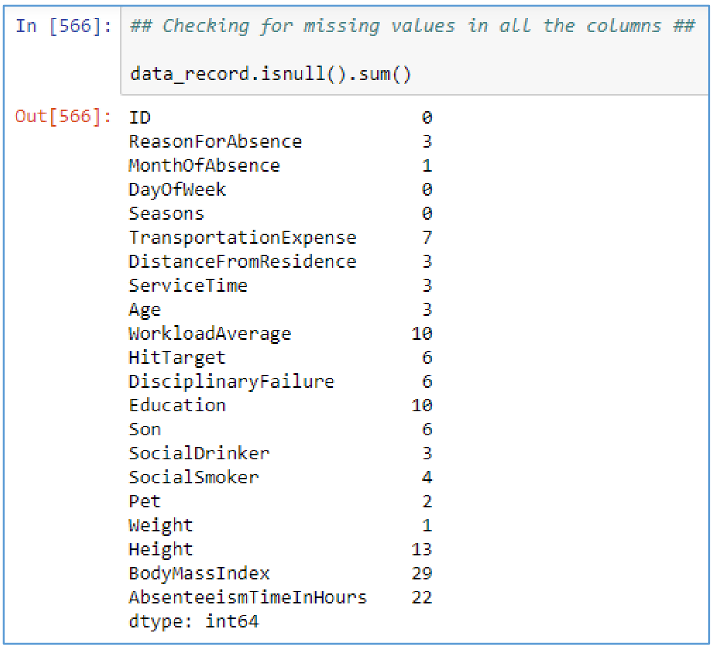
Here, Absenteeism Time In Hours is the target variable.

Chapter 2

Modelling In Python

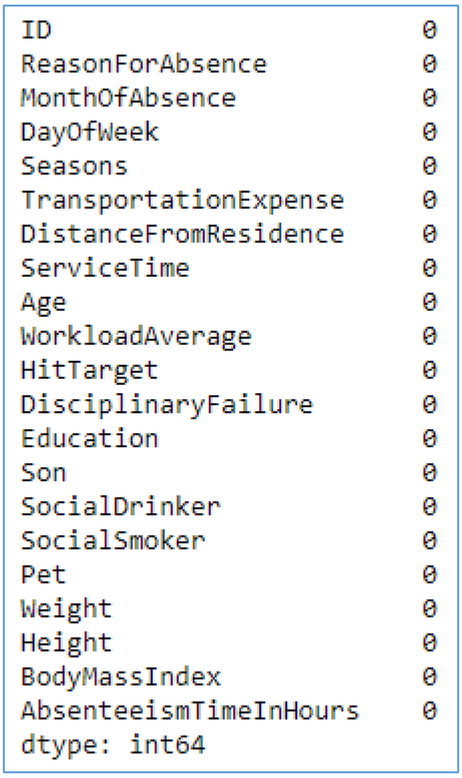
2.1 Missing Value Analysis

After loading the dataset and assessing it, the first order of business is to manage all the missing values in the dataset.



```
In [566]: ## Checking for missing values in all the columns ##

          data_record.isnull().sum()

Out[566]: ID                         0
          ReasonForAbsence           3
          MonthOfAbsence             1
          DayOfWeek                  0
          Seasons                    0
          TransportationExpense      7
          DistanceFromResidence      3
          ServiceTime                3
          Age                        3
          WorkloadAverage           10
          HitTarget                  6
          DisciplinaryFailure        6
          Education                 10
          Son                        6
          SocialDrinker              3
          SocialSmoker               4
          Pet                        2
          Weight                     1
          Height                    13
          BodyMassIndex             29
          AbsenteeismTimeInHours    22
          dtype: int64
```

The major imputation methods (mean, median, KNN) are tested for accuracy, and the method with the highest accuracy is chosen to impute the missing values. In this scenario, KNN (k-nearest neighbour) method had the most accuracy, and is used to impute the missing values.

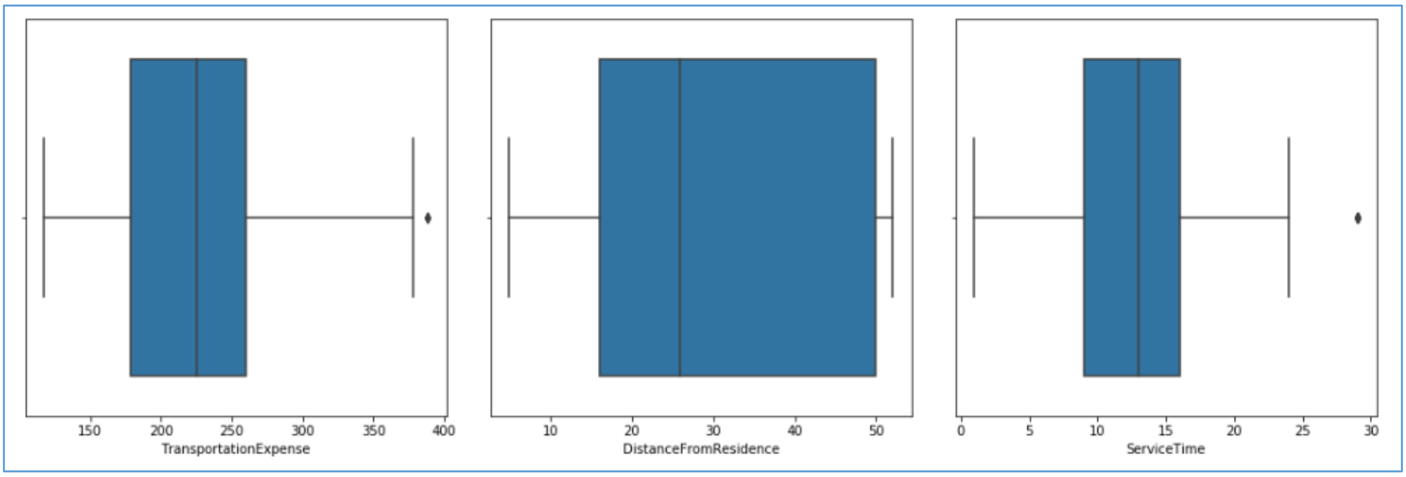


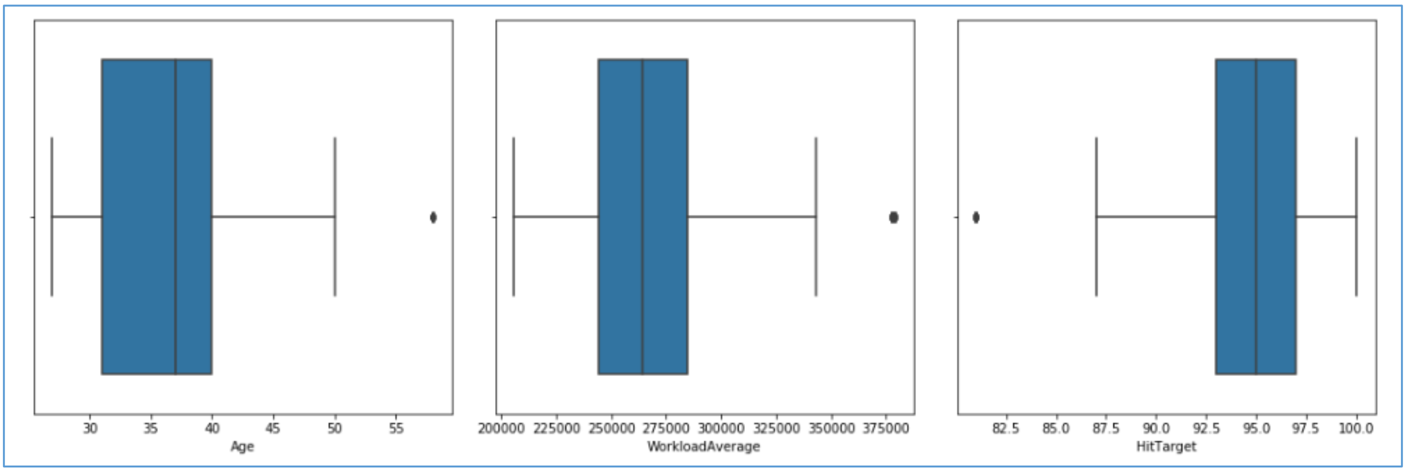
```
ID                        0
ReasonForAbsence          0
MonthOfAbsence            0
DayOfWeek                 0
Seasons                   0
TransportationExpense     0
DistanceFromResidence     0
ServiceTime               0
Age                       0
WorkloadAverage           0
HitTarget                 0
DisciplinaryFailure       0
Education                 0
Son                       0
SocialDrinker             0
SocialSmoker              0
Pet                       0
Weight                    0
Height                    0
BodyMassIndex             0
AbsenteeismTimeInHours    0
dtype: int64
```

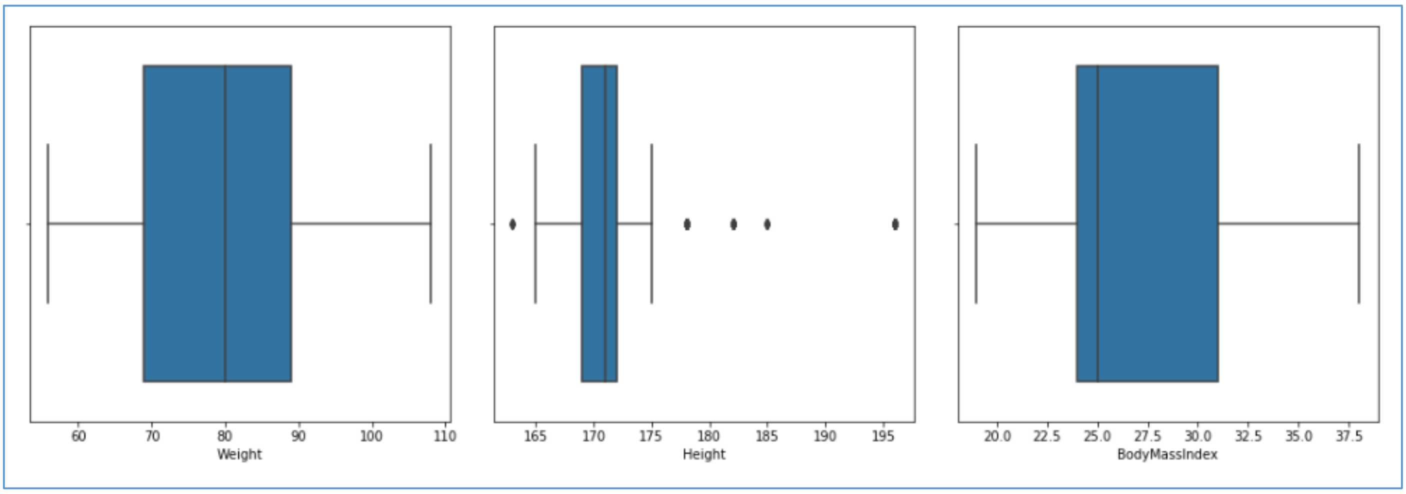
2.2 Outlier Analysis

The next task to handle is to check for outliers in the dataset, and replace those values to get rid of the skew in distribution. The detection of outliers is done by the boxplot graphical method.

Boxplots of all continuous variables (except target variable) are plotted to detect outliers.





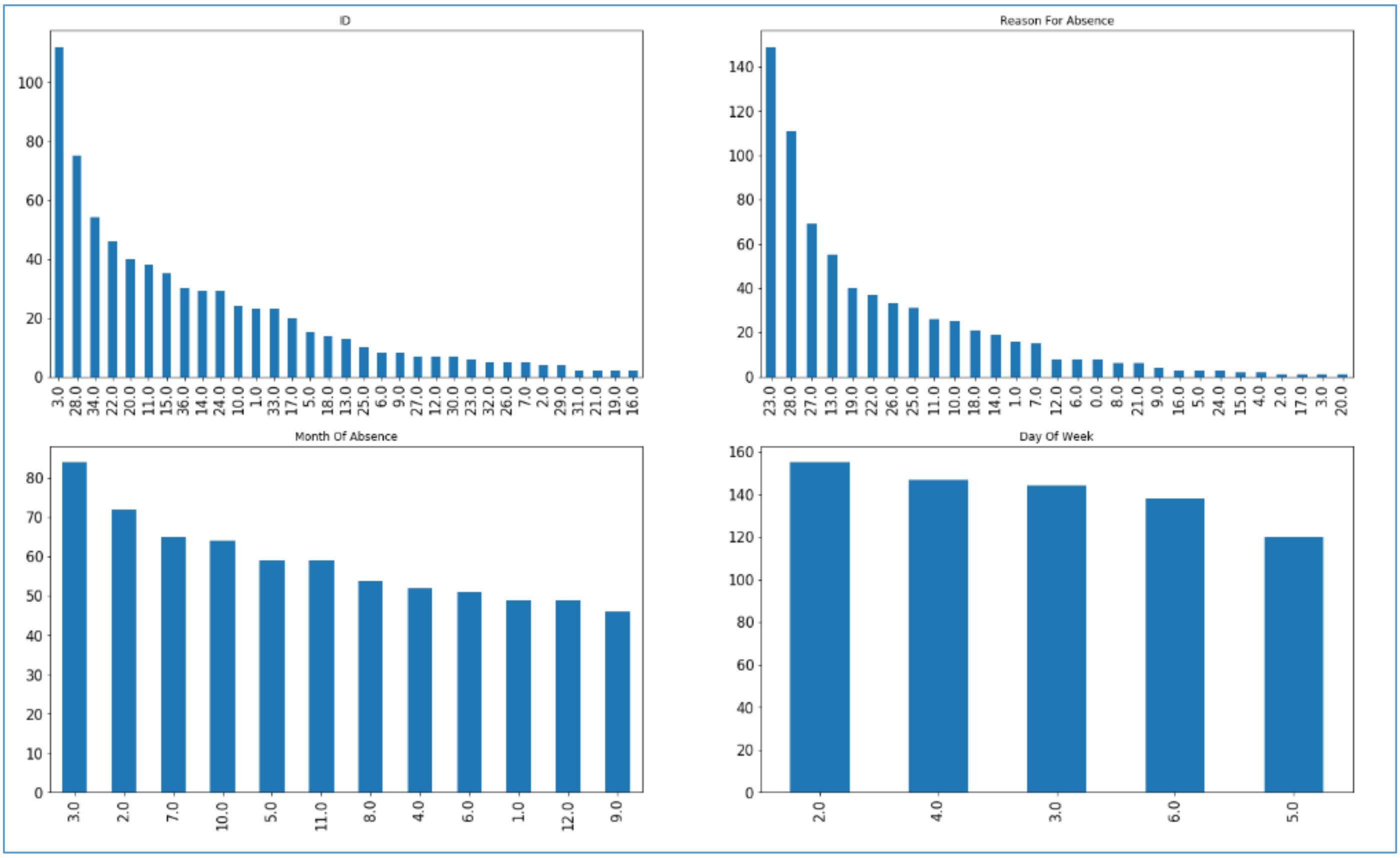


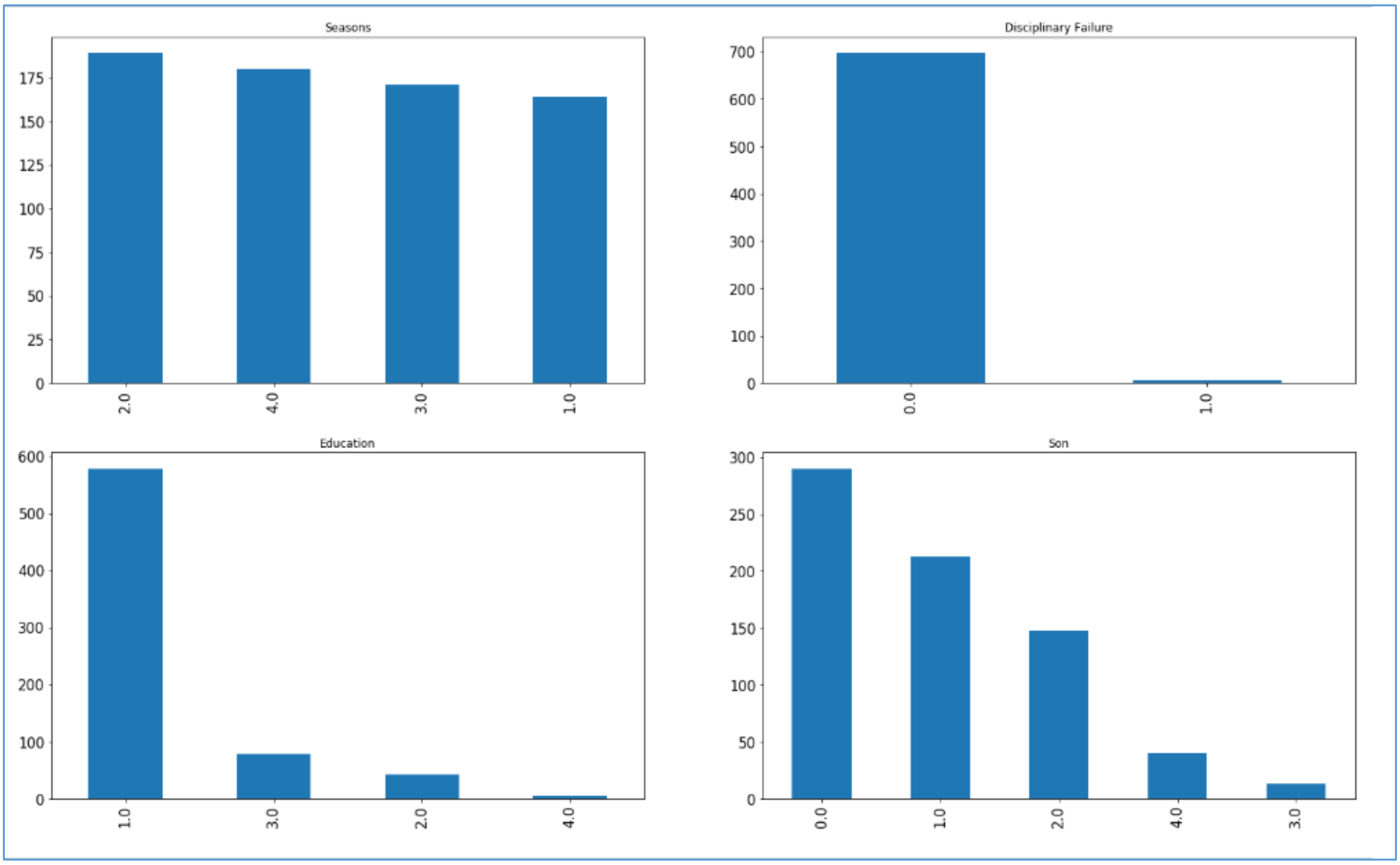
The outlying values are changed to NA to classify them as missing values, and then the missing values are imputed using KNN imputation.

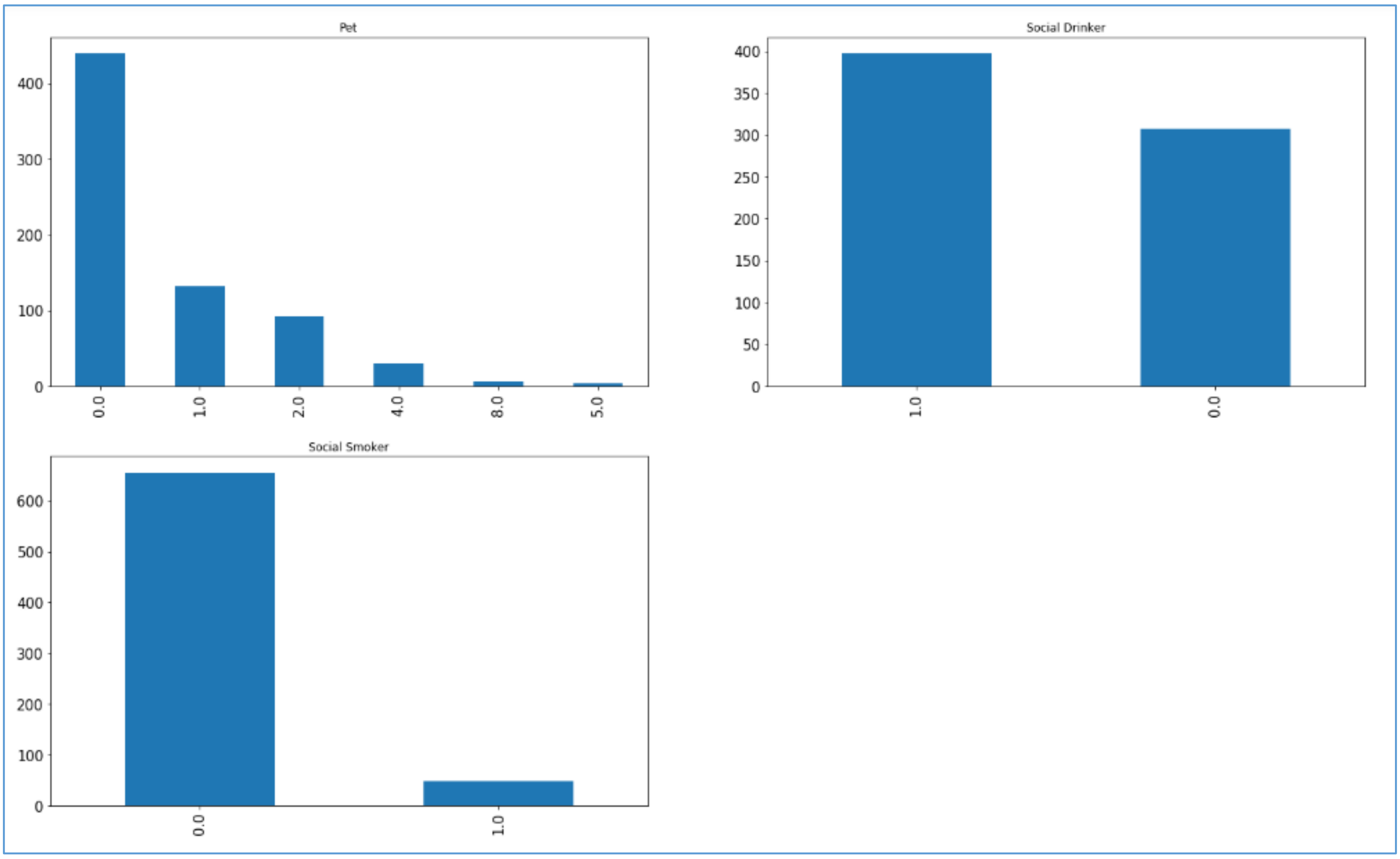
2.3 Data Analysis

The next step is to try and draw conclusions from the dataset. This can be performed by creating various plots of the variables.

Bar graphs of the categorical variables –







From these graphs some ***primitive conclusions*** can be drawn, such as:

# Employee with ID = 3 has the most entries for absenteeism.

# The top 5 reasons for absenteeism (in number of entries) are 23 (*Medical Consultation*), 28 (*Dental Consultation*), 27 (*Physiotherapy*), 13 (*Musculoskeletal Diseases*), and 19 (*Injury, poisoning or other external causes*).

# Most entries for absenteeism is in month number 3 (March).

# Most entries for absenteeism are for the second day of the week (Monday).

# Disciplinary Failures are rare.

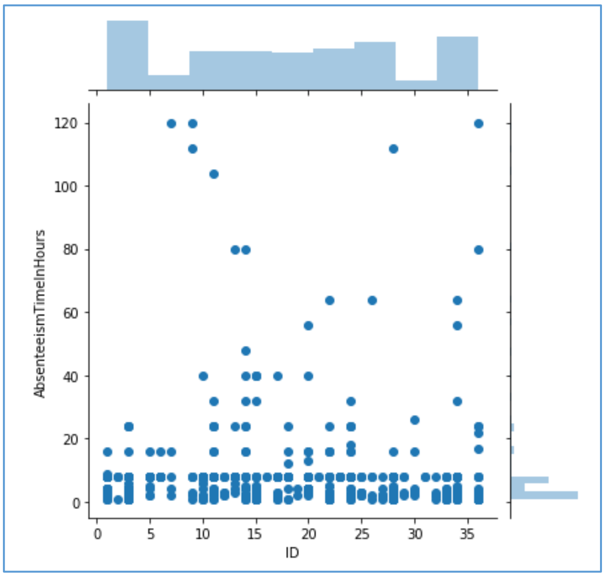
# A sizeable number of employees don't have kids and/or pets.

# There is a moderate number of social drinkers and a miniscule number of social smokers.

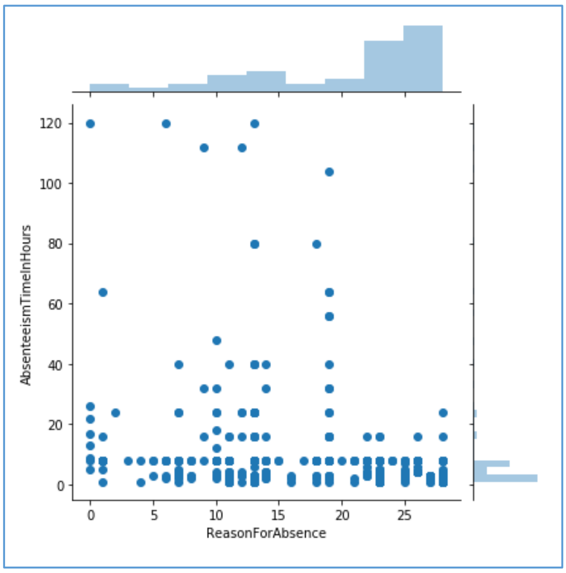
In order to understand the relation and draw conclusions in a better way, plots must be created which show the variables accounting for the highest counts in the target variable (Absenteeism time in hours).

Following are the joint scatter-histogram plots that show the correlation between the independent variables and the target variable:

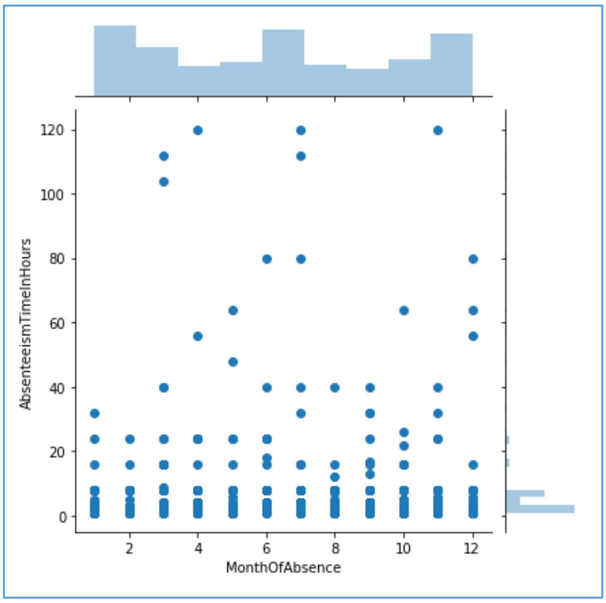
Categorical Variables



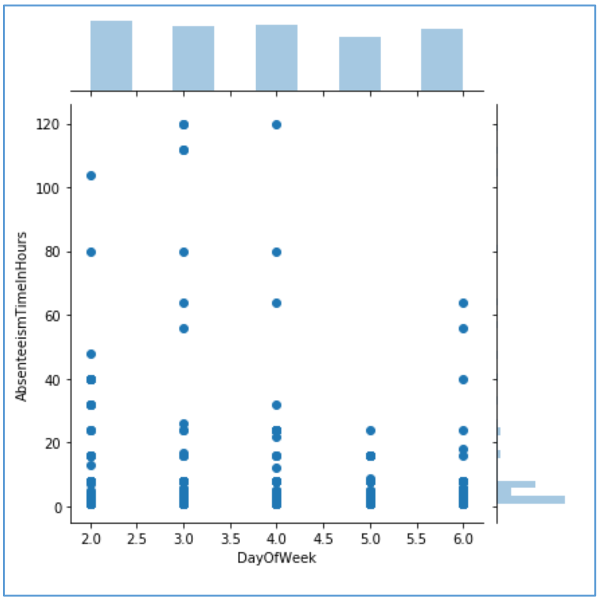
Absenteeism hours by ID



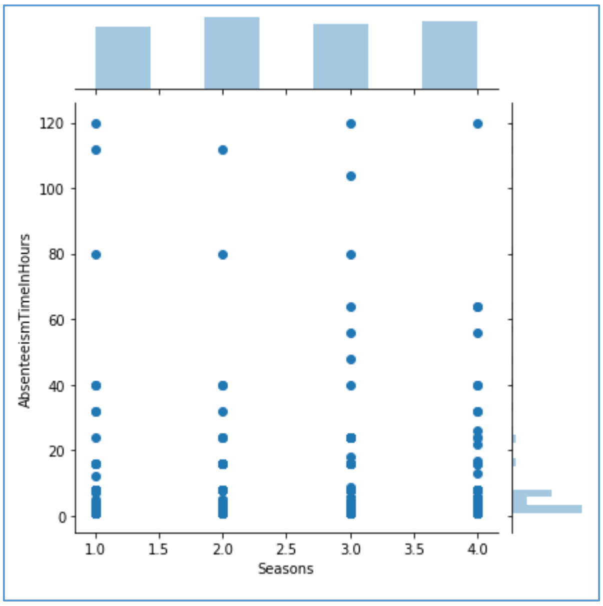
Absenteeism hours by Reason



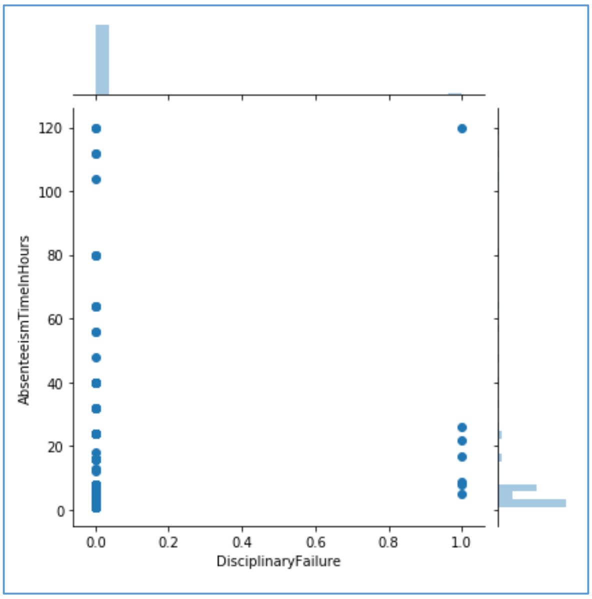
Absenteeism hours by Month



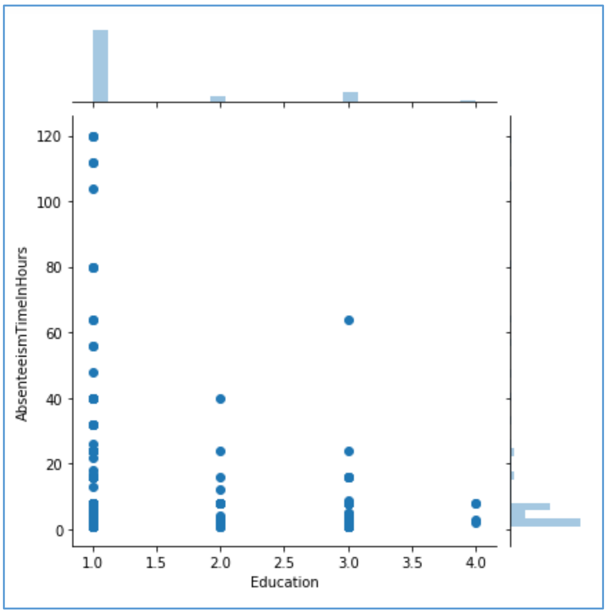
Absenteeism hours by Day Of Week



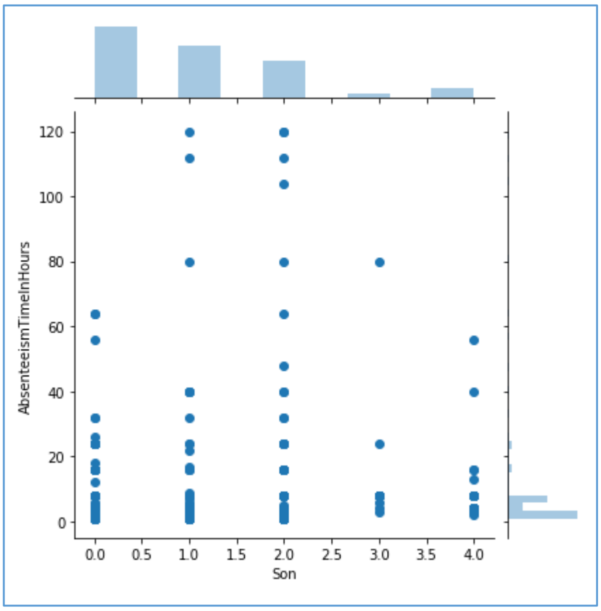
Absenteeism hours by Season



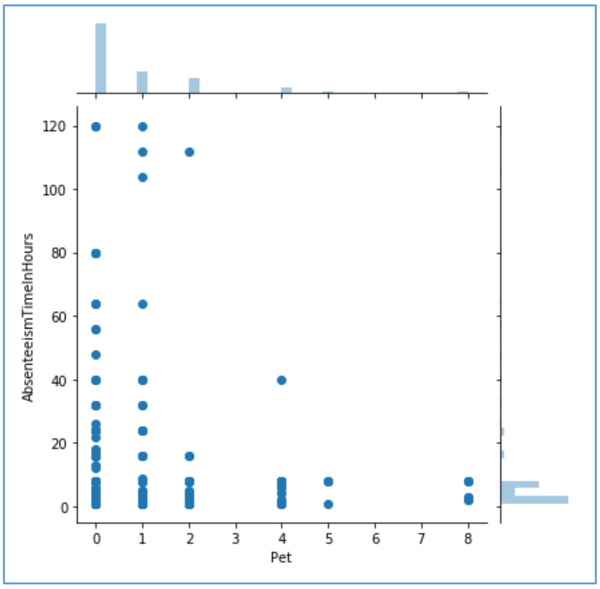
Absenteeism hours by Disciplinary Failures



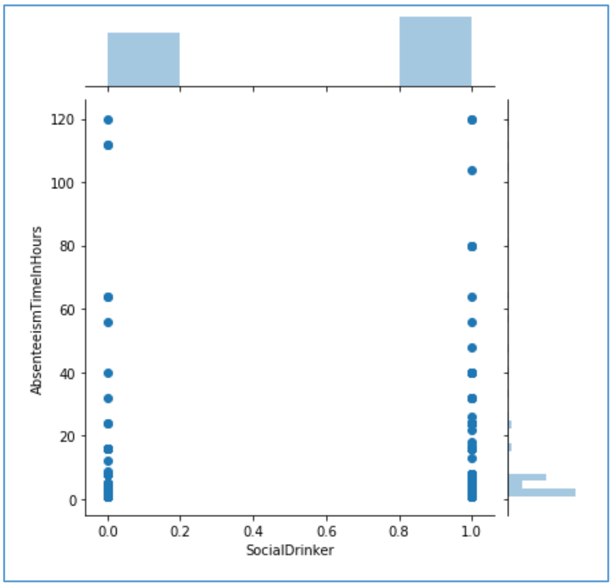
Absenteeism hours by Education Level



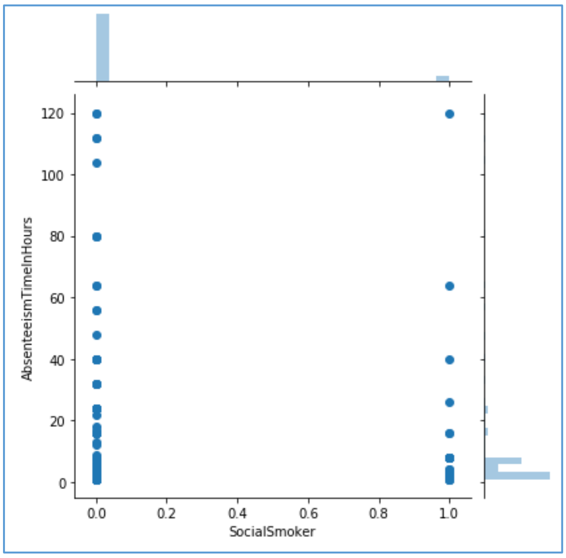
Absenteeism hours by No. Of Children



Absenteeism hours by No. Of Pets

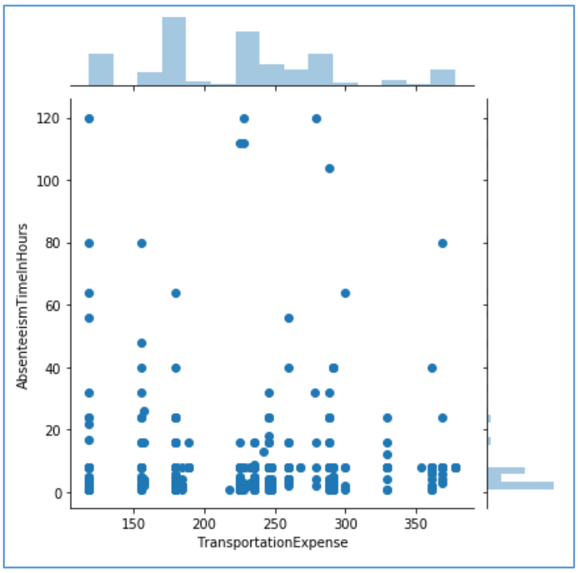


Absenteeism hours by No. Of Social Drinkers

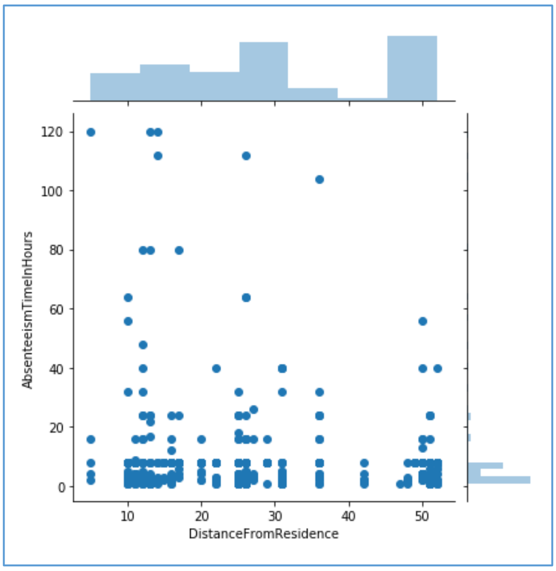


Absenteeism hours by No. Of Social Smokers

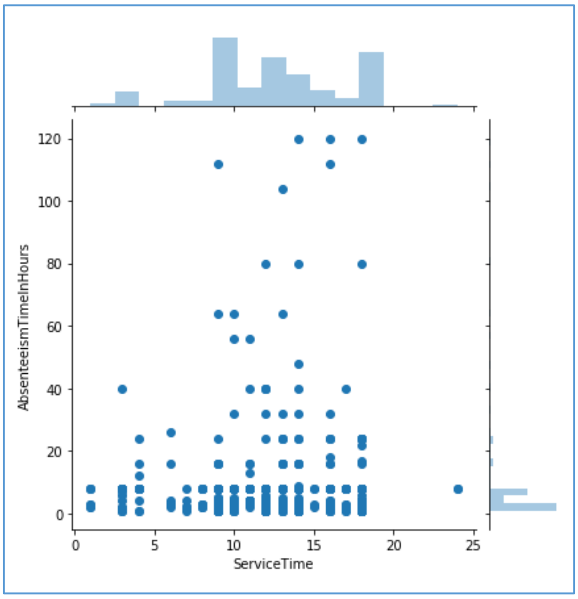
Continuous Variables



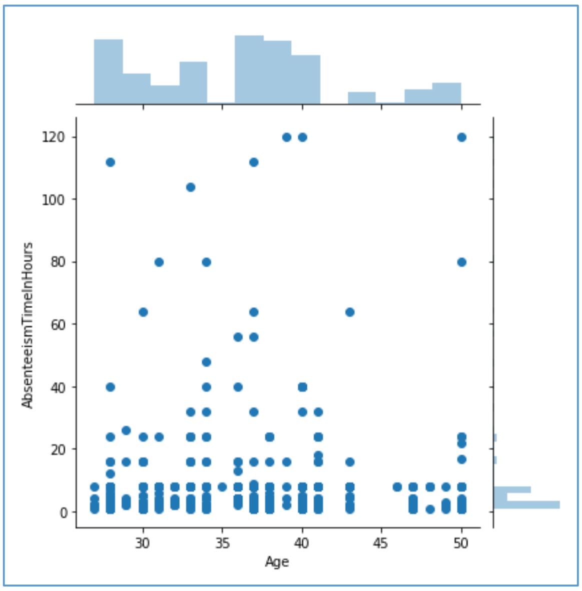
Absenteeism hours by Transportation Expense



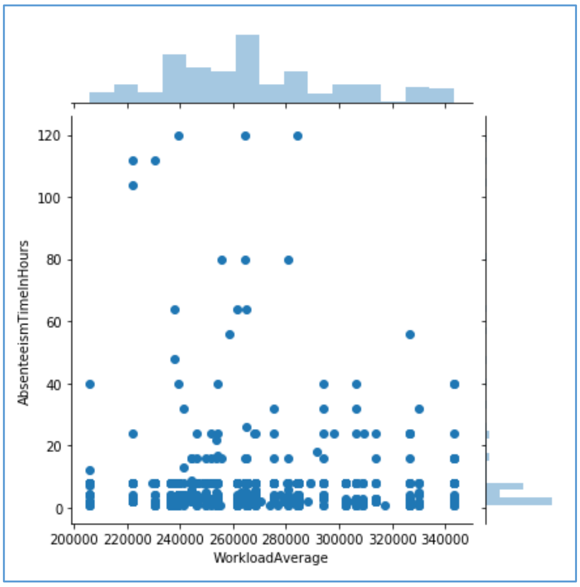
Absenteeism hours by Distance From Residence



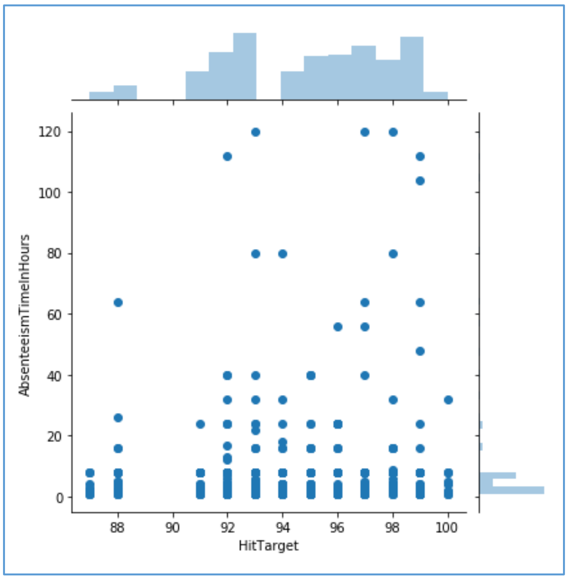
Absenteeism hours by Service Time (yrs)



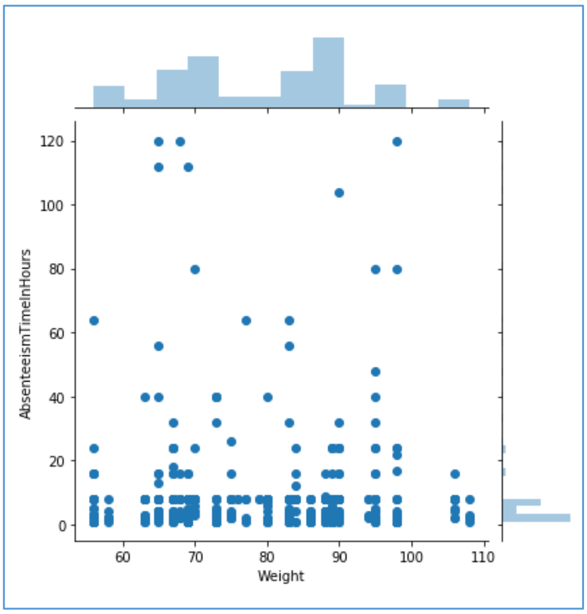
Absenteeism hours by Age (yrs)



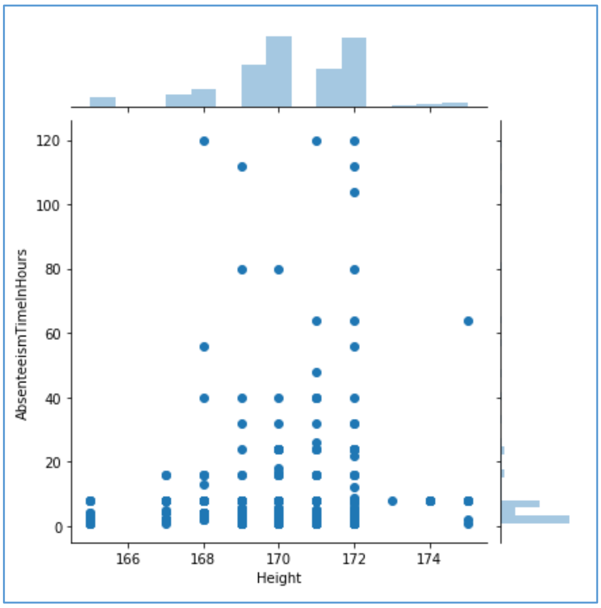
Absenteeism hours by Workload Average / Day



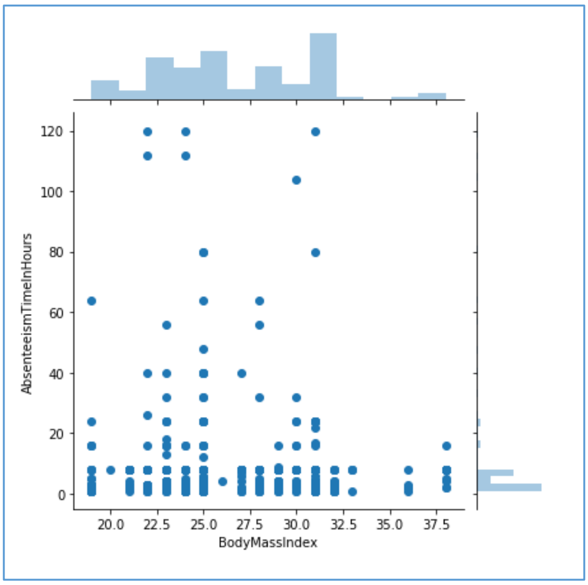
Absenteeism hours by Target Hit (%)



Absenteeism hours by Weight



Absenteeism hours by Height



Absenteeism hours by Body Mass Index

From these graphs, some *proper informed conclusions* can be drawn, such as:

The company lost most hours to absenteeism due to employees having *Physiotherapy* and *Medical Consultations*.

The company lost most hours to absenteeism in the months of *February*, *July* and *December*.

The company lost significant hours to absenteeism from employees who have residences *farthest from work*.

The company lost significant hours to absenteeism from employees who are nearing *10* or *20* years of service.

The company lost most hours to absenteeism from employees in the *36-42* age group, and lost significant hours to absenteeism from employees in the *28-29* age group.

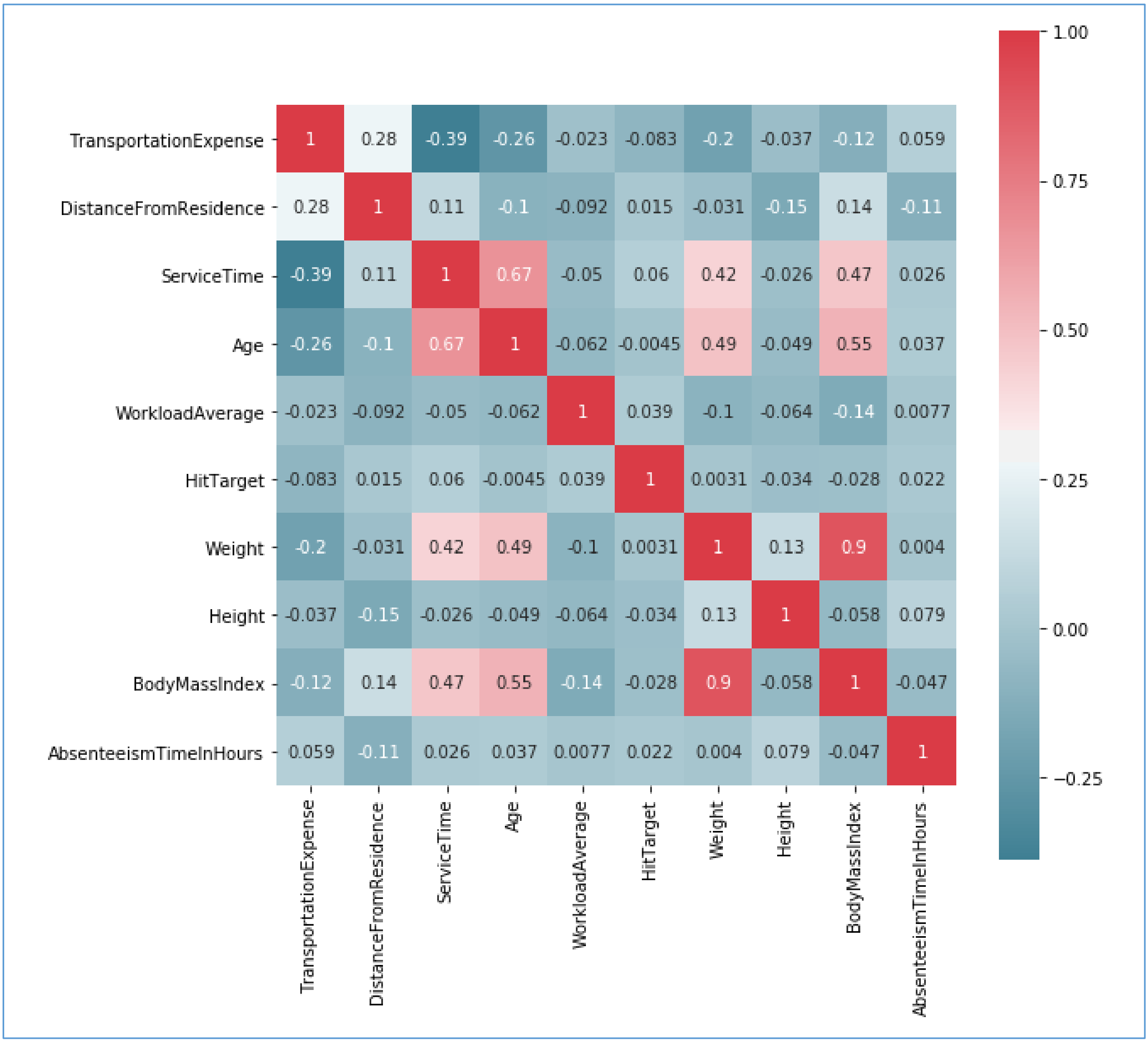
The company lost most hours to absenteeism from employees with average daily workload of *270000* units, and lost significant hours to absenteeism from employees with average daily workload of *240000* units.

The company lost most hours to absenteeism from employees who hit a target of *93%*, and lost significant hours to absenteeism from employees who hit the target range of *96-99%*.

2.4 Feature Selection

The variables are then tested for correlation so as to reduce their number for the ease of handling the dataset. Variables which are deemed to be heavily correlated to other variable/s are then removed form the dataset.

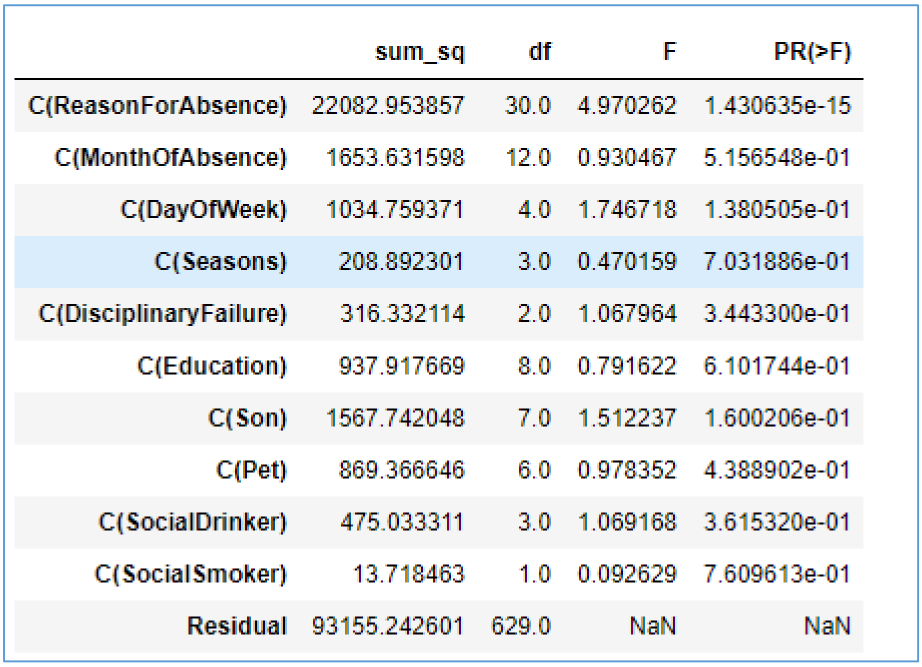
The correlation among continuous variables is checked by created a correlation matrix and depicting it using a heatmap:



	TransportationExpense	DistanceFromResidence	ServiceTime	Age	WorkloadAverage	HitTarget	Weight	Height	BodyMassIndex	AbsenteeismTimeInHours
TransportationExpense	1	0.28	-0.39	-0.26	-0.023	-0.083	-0.2	-0.037	-0.12	0.059
DistanceFromResidence	0.28	1	0.11	-0.1	-0.092	0.015	-0.031	-0.15	0.14	-0.11
ServiceTime	-0.39	0.11	1	0.67	-0.05	0.06	0.42	-0.026	0.47	0.026
Age	-0.26	-0.1	0.67	1	-0.062	-0.0045	0.49	-0.049	0.55	0.037
WorkloadAverage	-0.023	-0.092	-0.05	-0.062	1	0.039	-0.1	-0.064	-0.14	0.0077
HitTarget	-0.083	0.015	0.06	-0.0045	0.039	1	0.0031	-0.034	-0.028	0.022
Weight	-0.2	-0.031	0.42	0.49	-0.1	0.0031	1	0.13	0.9	0.004
Height	-0.037	-0.15	-0.026	-0.049	-0.064	-0.034	0.13	1	-0.058	0.079
BodyMassIndex	-0.12	0.14	0.47	0.55	-0.14	-0.028	0.9	-0.058	1	-0.047
AbsenteeismTimeInHours	0.059	-0.11	0.026	0.037	0.0077	0.022	0.004	0.079	-0.047	1

Here, it is apparent from the heatmap that the column BodyMassIndex is very highly correlated to the Weight column. Therefore, the BodyMassIndex column is dropped since it is the derived variable (BodyMassIndex = Weight/Height^2).

To check correlation for categorical variables, the ANOVA test (Analysis Of Variance) is performed:



	sum_sq	df	F	PR(>F)
C(ReasonForAbsence)	22082.953857	30.0	4.970262	1.430635e-15
C(MonthOfAbsence)	1653.631598	12.0	0.930467	5.156548e-01
C(DayOfWeek)	1034.759371	4.0	1.746718	1.380505e-01
C(Seasons)	208.892301	3.0	0.470159	7.031886e-01
C(DisciplinaryFailure)	316.332114	2.0	1.067964	3.443300e-01
C(Education)	937.917669	8.0	0.791622	6.101744e-01
C(Son)	1567.742048	7.0	1.512237	1.600206e-01
C(Pet)	869.366646	6.0	0.978352	4.388902e-01
C(SocialDrinker)	475.033311	3.0	1.069168	3.615320e-01
C(SocialSmoker)	13.718463	1.0	0.092629	7.609613e-01
Residual	93155.242601	629.0	NaN	NaN

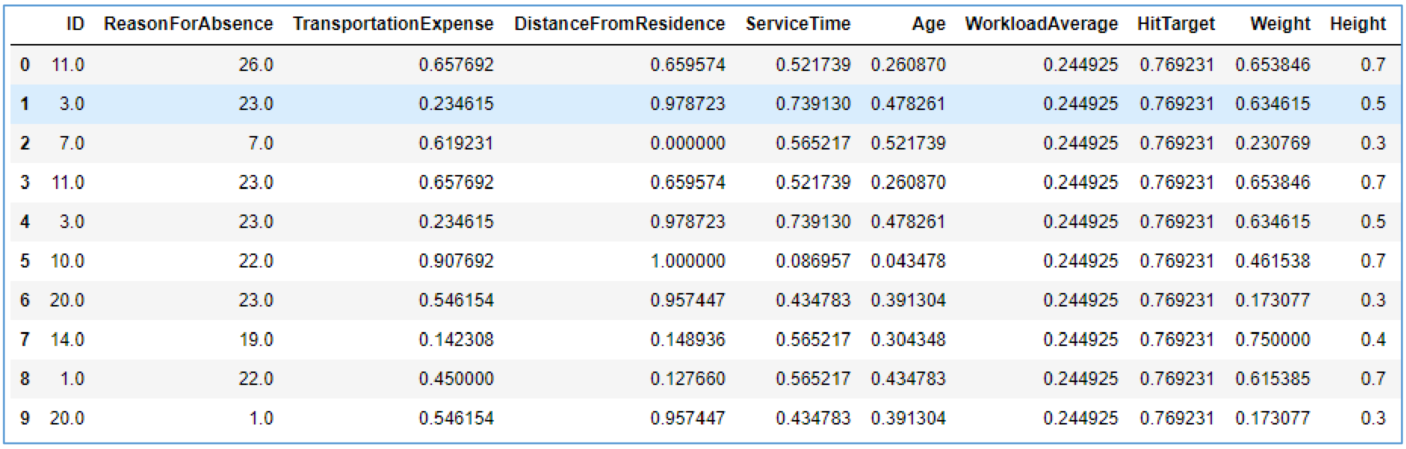
It is instantly observable that all columns except ReasonForAbsence have p-values more than 0.05. Therefore, all categorical variables except ReasonForAbsence are dropped from the dataset.

2.5 Feature Scaling

To get rid of the unwanted variation within or between variables, the observations need to be scaled to conform to uniform distribution. Of the two available options, Normalization is used for the continuous variables in the dataset (since there is only one categorical variable left after the Feature Selection process).

$$\text{Normalized value} = \frac{\text{Individual Observation} - \text{Min(Observations)}}{\text{Max(Observations)} - \text{Min(Observations)}}$$

After Feature Scaling is performed, the dataset looks like this:



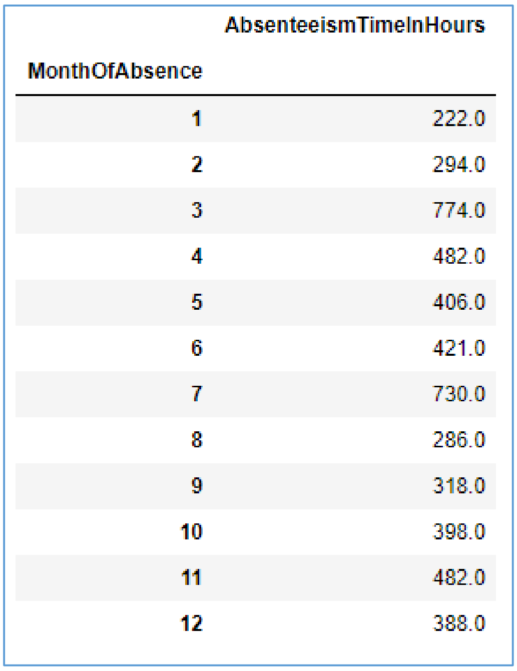
	ID	ReasonForAbsence	TransportationExpense	DistanceFromResidence	ServiceTime	Age	WorkloadAverage	HitTarget	Weight	Height
0	11.0	26.0	0.657692	0.659574	0.521739	0.260870	0.244925	0.769231	0.653846	0.7
1	3.0	23.0	0.234615	0.978723	0.739130	0.478261	0.244925	0.769231	0.634615	0.5
2	7.0	7.0	0.619231	0.000000	0.565217	0.521739	0.244925	0.769231	0.230769	0.3
3	11.0	23.0	0.657692	0.659574	0.521739	0.260870	0.244925	0.769231	0.653846	0.7
4	3.0	23.0	0.234615	0.978723	0.739130	0.478261	0.244925	0.769231	0.634615	0.5
5	10.0	22.0	0.907692	1.000000	0.086957	0.043478	0.244925	0.769231	0.461538	0.7
6	20.0	23.0	0.546154	0.957447	0.434783	0.391304	0.244925	0.769231	0.173077	0.3
7	14.0	19.0	0.142308	0.148936	0.565217	0.304348	0.244925	0.769231	0.750000	0.4
8	1.0	22.0	0.450000	0.127660	0.565217	0.434783	0.244925	0.769231	0.615385	0.7
9	20.0	1.0	0.546154	0.957447	0.434783	0.391304	0.244925	0.769231	0.173077	0.3

All values of the continuous variables are now situated between 0 and 1.

2.6 Time Series Forecasting

The problem statement specifies that the company has requested for forecasts regarding the absenteeism hours for each month of the year 2011. Time series modelling will be used to forecast absenteeism hours for each month.

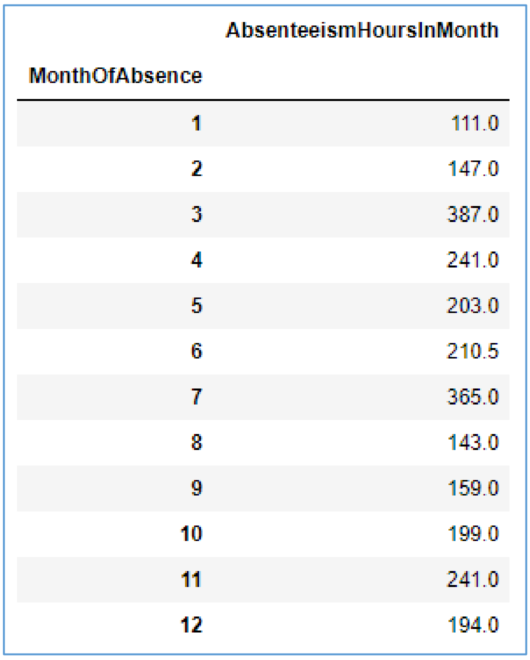
Absenteeism hours aggregated by month gives the following dataset:



	AbsenteeismTimeInHours
MonthOfAbsence	
1	222.0
2	294.0
3	774.0
4	482.0
5	406.0
6	421.0
7	730.0
8	286.0
9	318.0
10	398.0
11	482.0
12	388.0

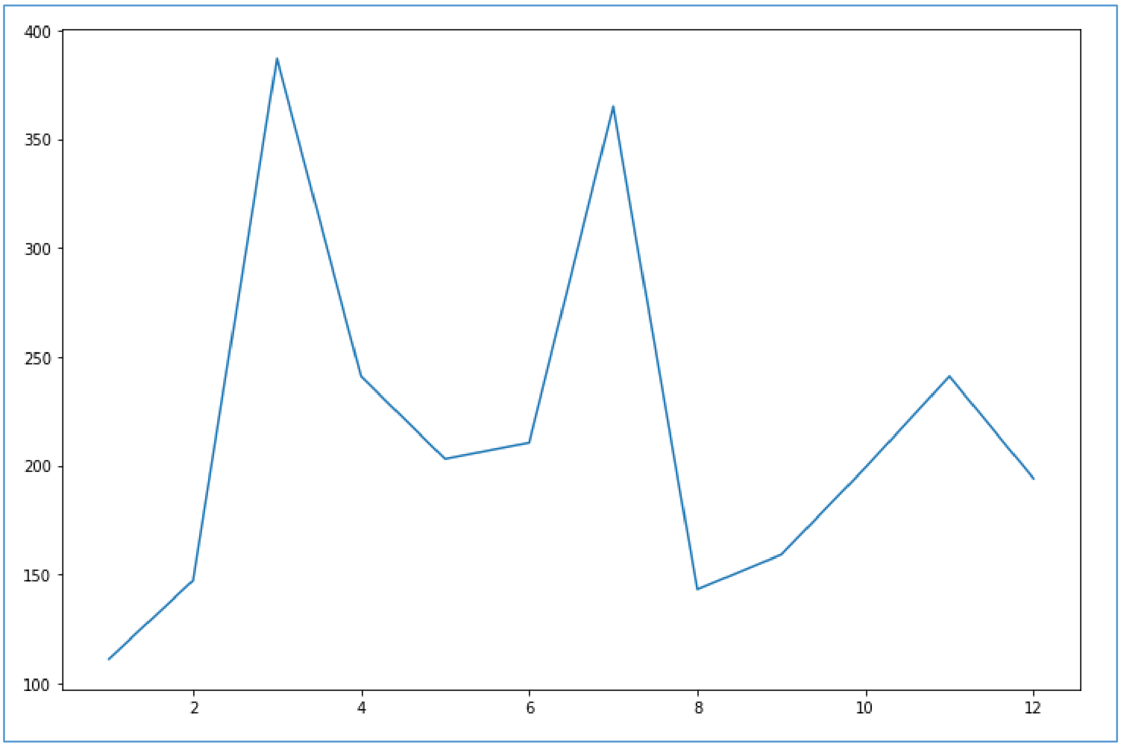
As timeframe of dataset isn't provided, proper course of action would be to assume a timeframe and proceed with the modelling. The minimum timeframe appropriate for the modelling would be **two** years, since the longer the timeframe, the more accurate the predictions.

Dividing the target variables by 2, to help with the assumption:

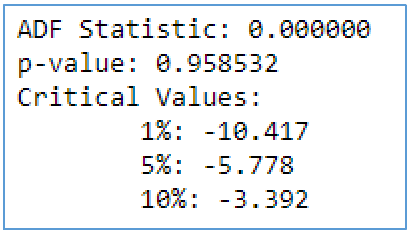


	AbsenteeismHoursInMonth
MonthOfAbsence	
1	111.0
2	147.0
3	387.0
4	241.0
5	203.0
6	210.5
7	365.0
8	143.0
9	159.0
10	199.0
11	241.0
12	194.0

Plotting the time series:

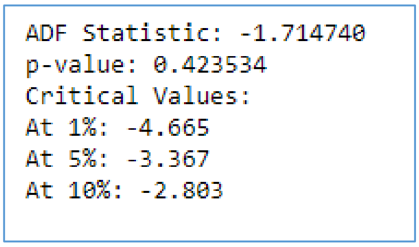


The next step is to perform the Augmented Dickey-Fuller test on the series, to check the stationarity of the series.
The test statistics after performing the ADF test on the series:



```
ADF Statistic: 0.000000
p-value: 0.958532
Critical Values:
        1%: -10.417
        5%: -5.778
        10%: -3.392
```

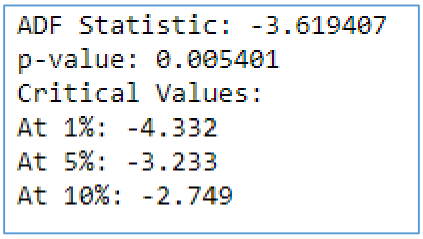
Since the ADF statistic is greater than the critical values, the series is not stationary. The next step is to get log values of the series, calculate the 'first difference' and redo the test.
Test statistics after performing the ADF test on the first difference:



```
ADF Statistic: -1.714740
p-value: 0.423534
Critical Values:
At 1%: -4.665
At 5%: -3.367
At 10%: -2.803
```

Here, the ADF statistic has decreased, but is still greater than the critical values, implying the series is not stationary. That means the next step should be calculating the 'second difference' and redoing the test.

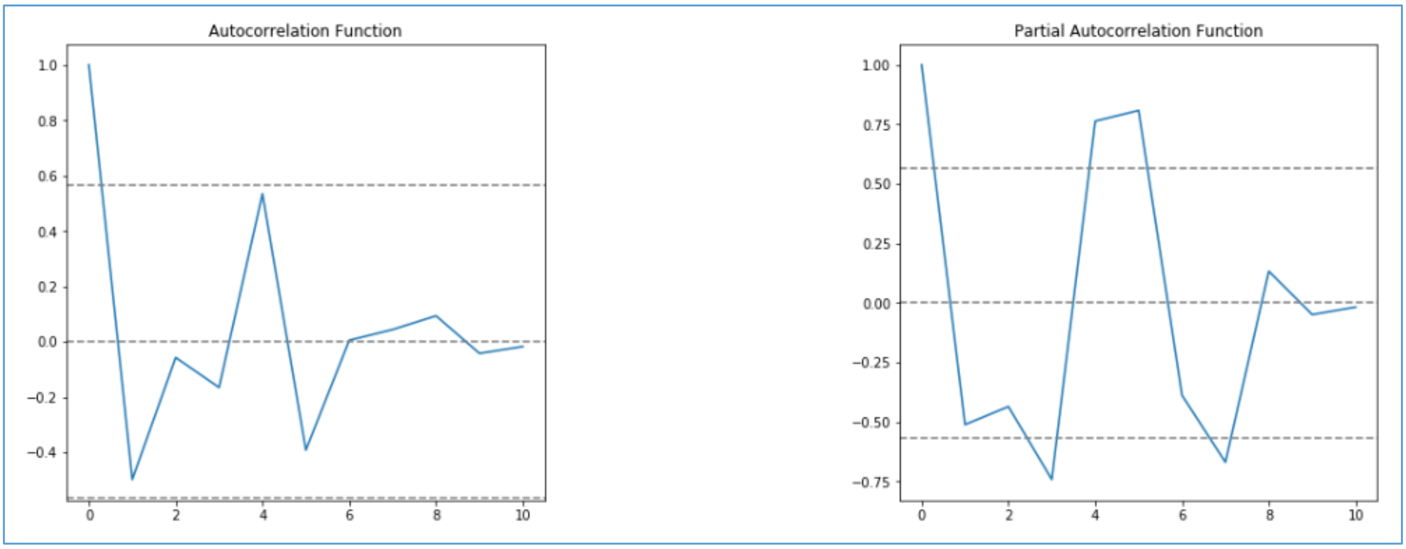
Test statistics after performing the ADF test on the second difference:



```
ADF Statistic: -3.619407
p-value: 0.005401
Critical Values:
At 1%: -4.332
At 5%: -3.233
At 10%: -2.749
```

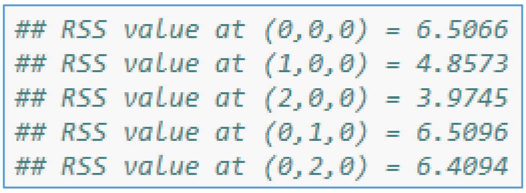
Here, the ADF statistic is lower than the critical value at 10%, which means now the series is stationary, and we can proceed further to the Autocorrelation and Partial Autocorrelation functions (ACF & PACF).

Creating the ACF and PACF plots:



Since the plots don’t give any clear values of the metrics *p* and *q*, the model ARIMA(0,0,0) is taken as the initial candidate, with the other models to be fitted being ARIMA(1,0,0), ARIMA(0,1,0), ARIMA(2,0,0), and ARIMA(0,2,0). The models are fitted over the time series, and the Residual Sum of Squares (RSS) is extracted as a parameter for comparison.

The RSS values recorded after the models were fitted are:

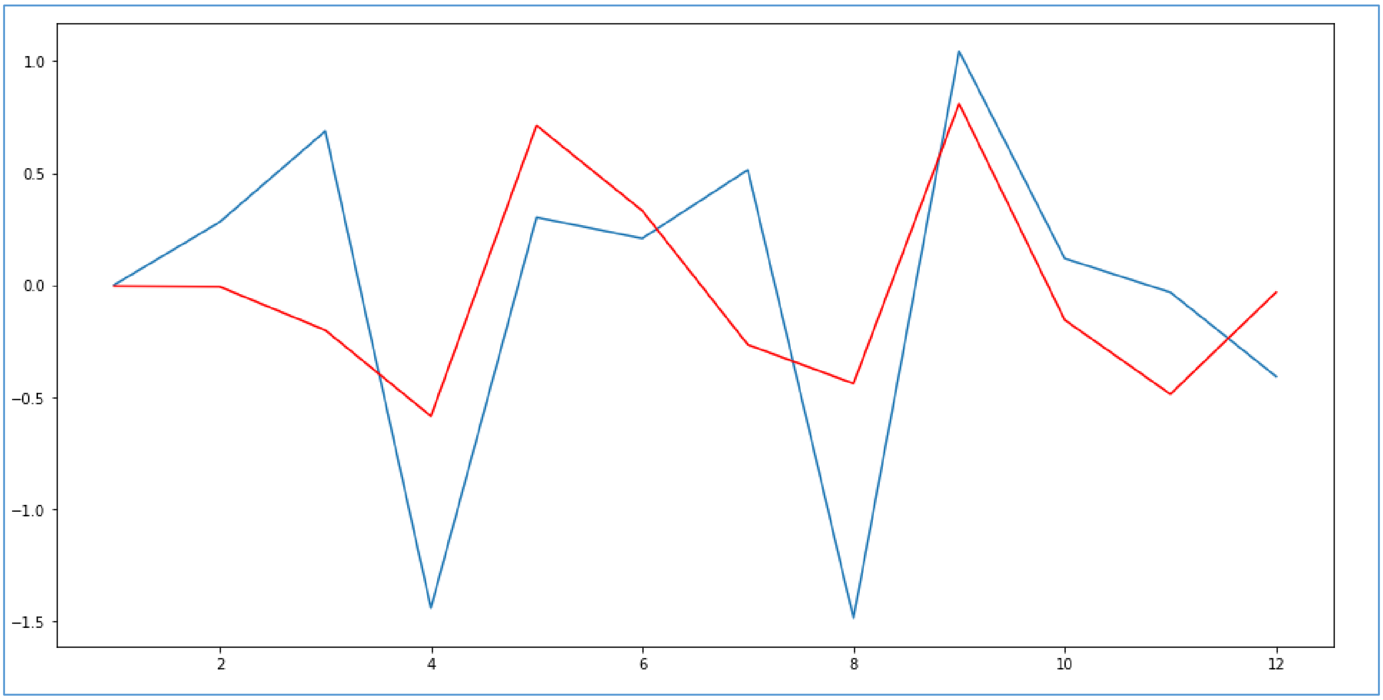


```
## RSS value at (0,0,0) = 6.5066
## RSS value at (1,0,0) = 4.8573
## RSS value at (2,0,0) = 3.9745
## RSS value at (0,1,0) = 6.5096
## RSS value at (0,2,0) = 6.4094
```

Since the model ARIMA(2,0,0) gives the lowest RSS value of 3.9745, it will be used for the forecasting of the time series.

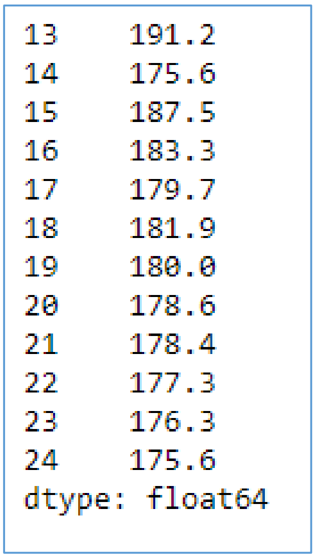
ARIMA(2,0,0) is fitted over the time series, and the fitted values are acquired and compared with the time series (second difference).

Plotting the fitted values and the second difference together:



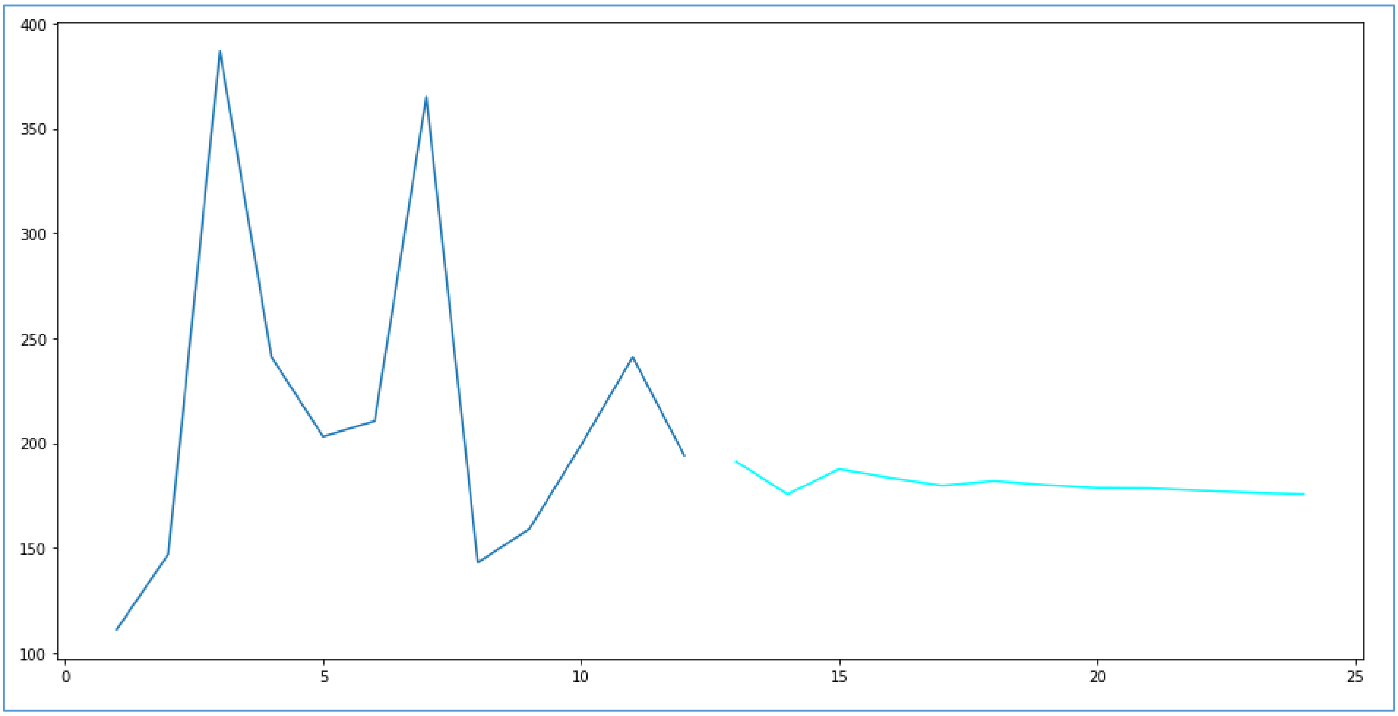
The predictions are made over the fit of the chosen model, and then added to a time series for the next 12 months (for the next year). The series is then treated to an exponential function (in correspondence to the log function used earlier), and that results in the time series predicting the absenteeism hours for the next 12 months.

The predicted absenteeism hours for the next 12 months are as follows:



```
13     191.2
14     175.6
15     187.5
16     183.3
17     179.7
18     181.9
19     180.0
20     178.6
21     178.4
22     177.3
23     176.3
24     175.6
dtype: float64
```

Plotting the provided values (of absenteeism hours) and the forecasted values together:



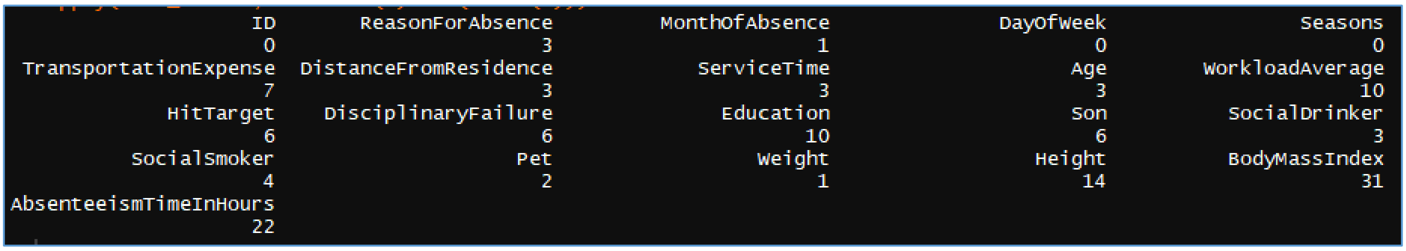
The predictions appear to be lacking the irregularity of the provided data, but that's because of the assumption that the provided data is of the past two years (2009 and 2010). Had there been a longer timeframe of the data, the irregularity would have been much lower. Not to forget that the data was simply halved instead of having separate sets of observations for the two years in question.

Chapter 3

Modelling In R

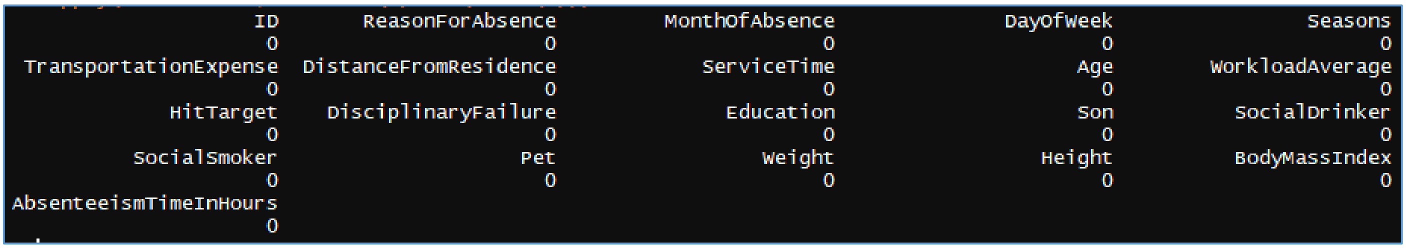
3.1 Missing Value Analysis

After loading the dataset and assessing it, the first order of business is to manage all the missing values in the dataset.



```
                    ID       ReasonForAbsence         MonthOfAbsence              DayOfWeek                Seasons
                     0                      3                      1                      0                      0
 TransportationExpense  DistanceFromResidence            ServiceTime                    Age        WorkloadAverage
                     7                      3                      3                      3                     10
             HitTarget    DisciplinaryFailure              Education                    Son           SocialDrinker
                     6                      6                     10                      6                      3
           SocialSmoker                    Pet                 Weight                 Height           BodyMassIndex
                     4                      2                      1                     14                     31
AbsenteeismTimeInHours
                    22
```

The major imputation methods (mean, median, KNN) are tested for accuracy, and the method with the highest accuracy is chosen to impute the missing values. In this scenario, KNN (k-nearest neighbour) method had the most accuracy, and is used to impute the missing values.

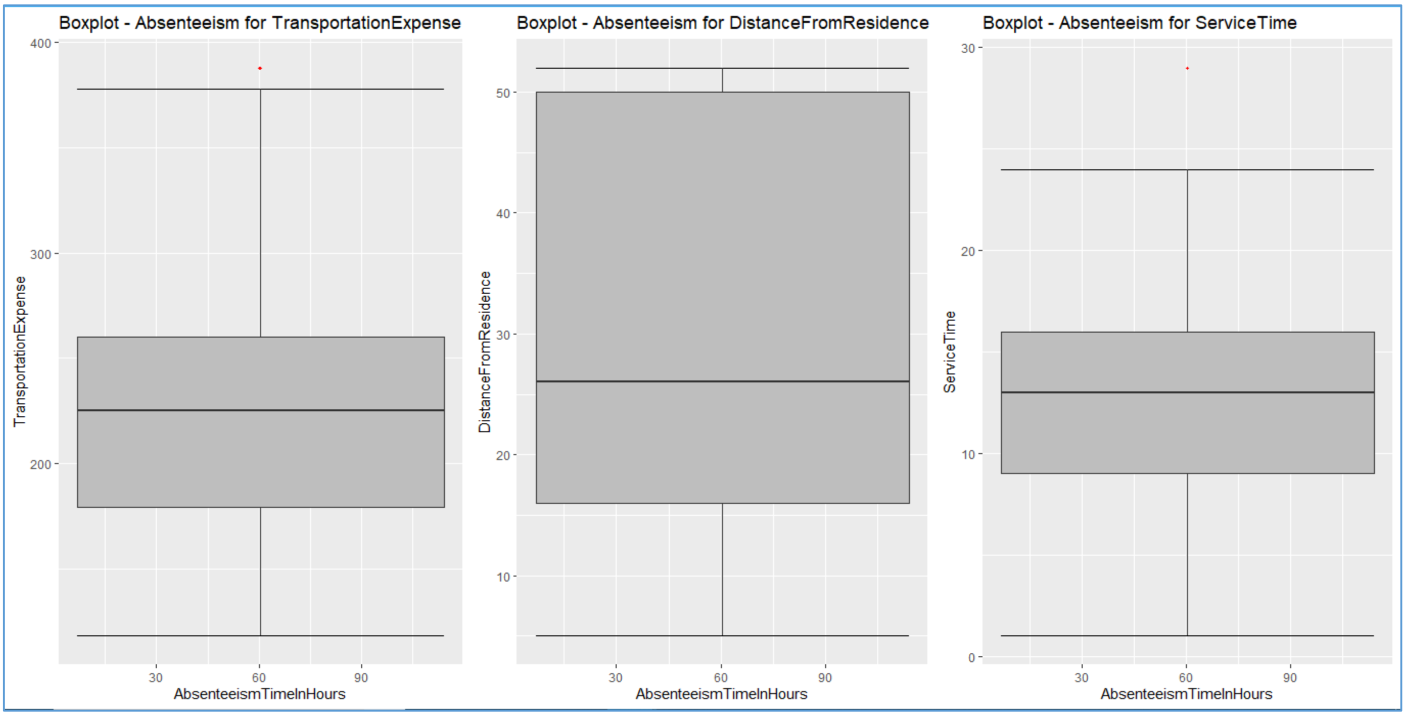


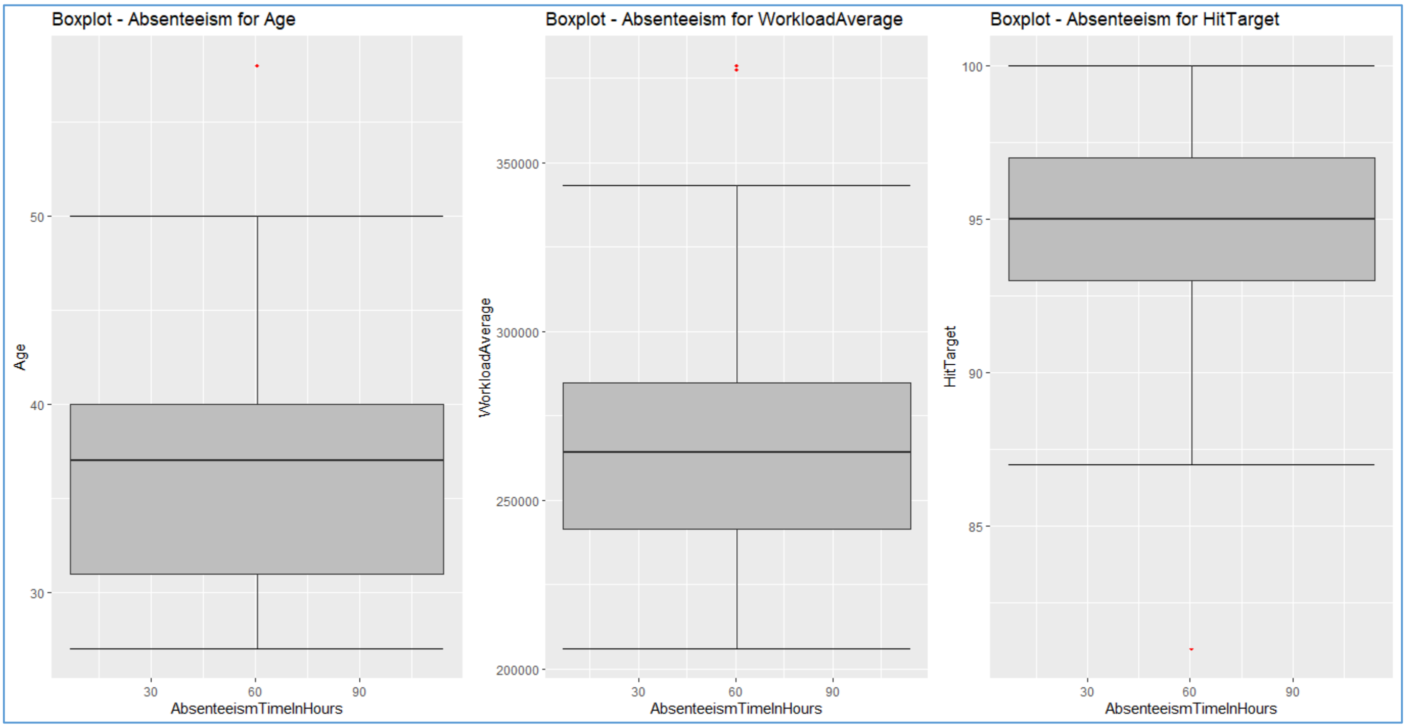
```
                    ID       ReasonForAbsence         MonthOfAbsence              DayOfWeek                Seasons
                     0                      0                      0                      0                      0
 TransportationExpense  DistanceFromResidence            ServiceTime                    Age        WorkloadAverage
                     0                      0                      0                      0                      0
             HitTarget    DisciplinaryFailure              Education                    Son           SocialDrinker
                     0                      0                      0                      0                      0
           SocialSmoker                    Pet                 Weight                 Height           BodyMassIndex
                     0                      0                      0                      0                      0
AbsenteeismTimeInHours
                     0
```

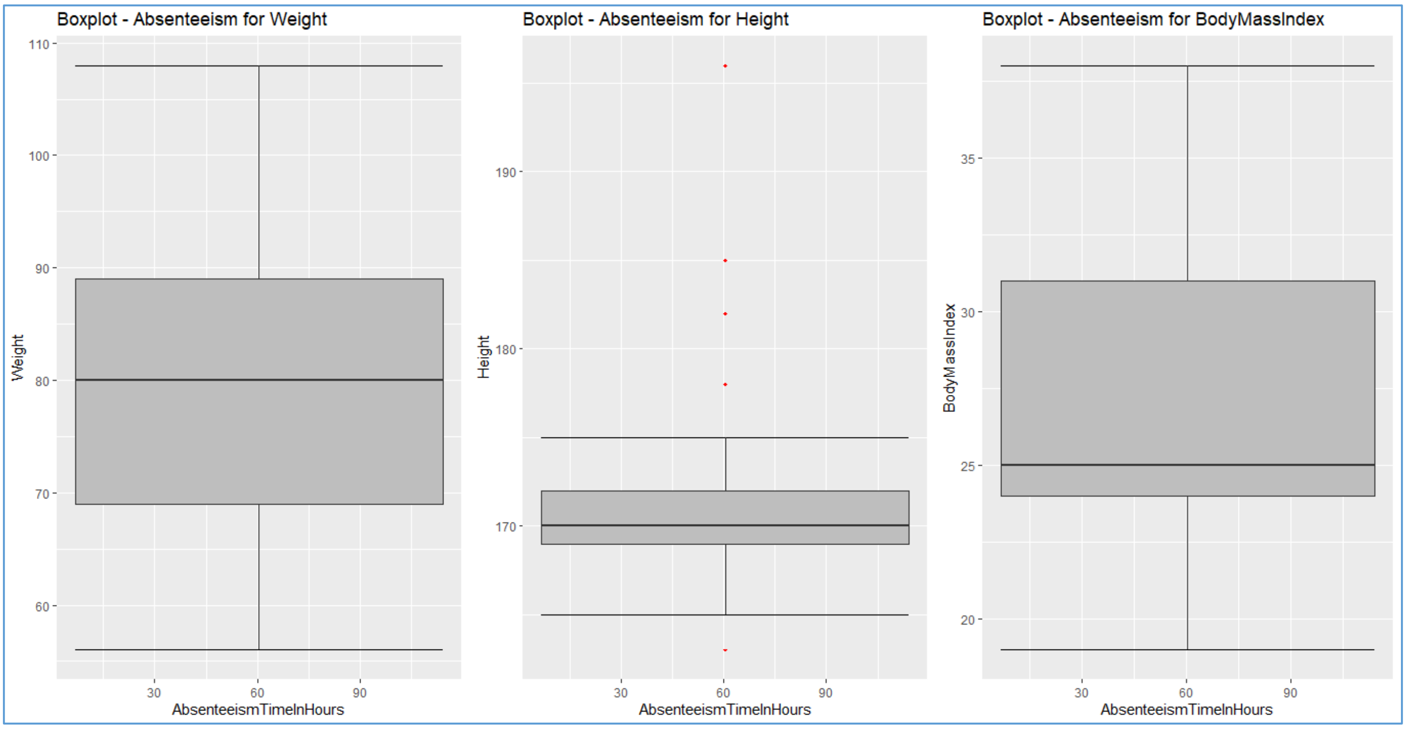
3.2 Outlier Analysis

The next task to handle is to check for outliers in the dataset, and replace those values to get rid of the skew in distribution. The detection of outliers is done by the boxplot graphical method.

Boxplots of all continuous variables (except target variable) are plotted to detect outliers.





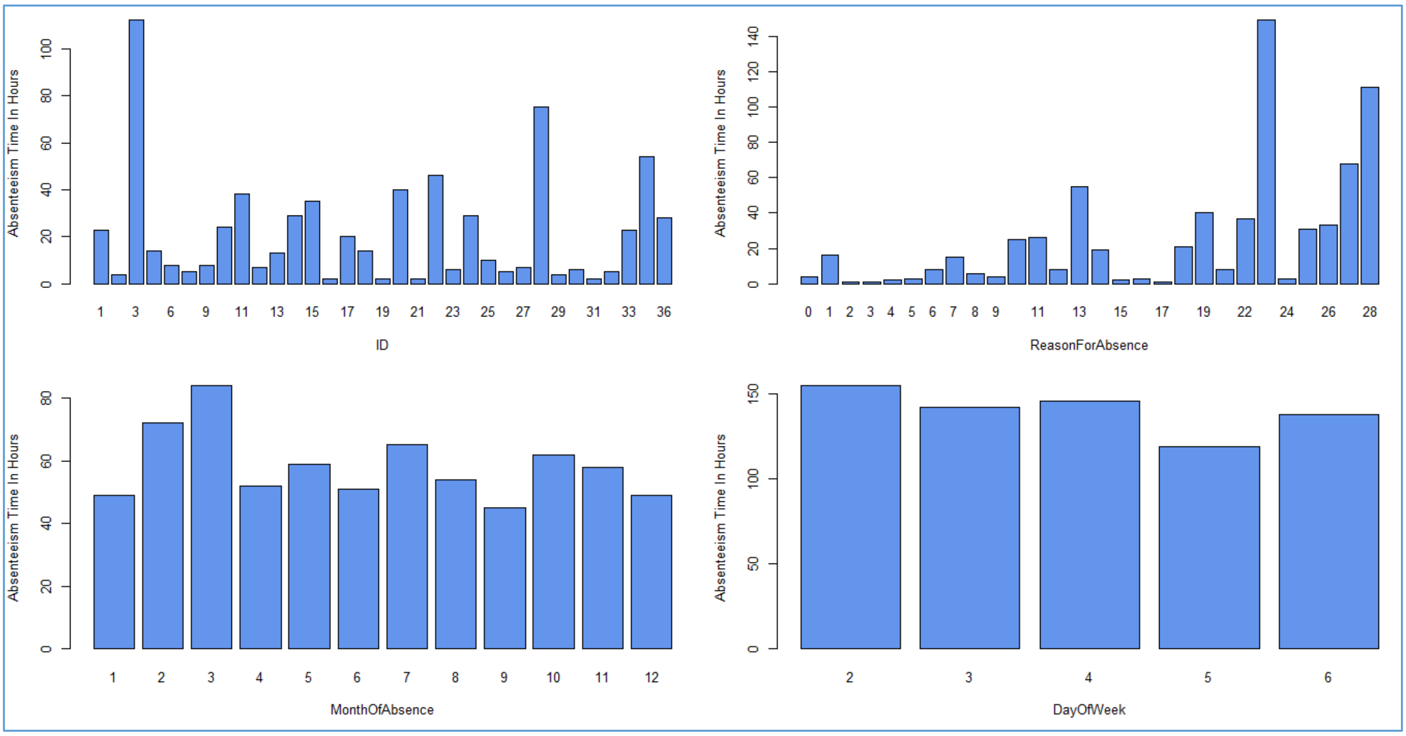


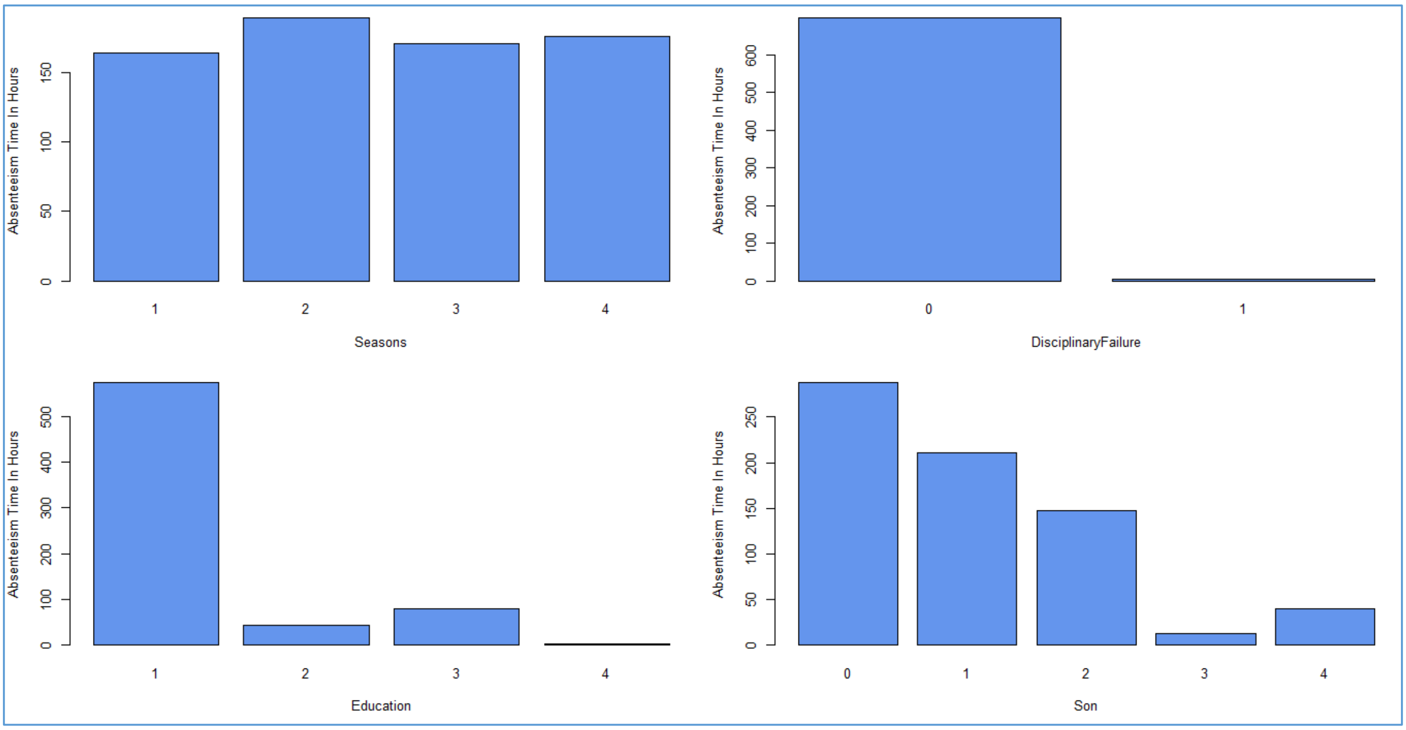
The outlying values are changed to NA to classify them as missing values, and then the missing values are imputed using KNN imputation.

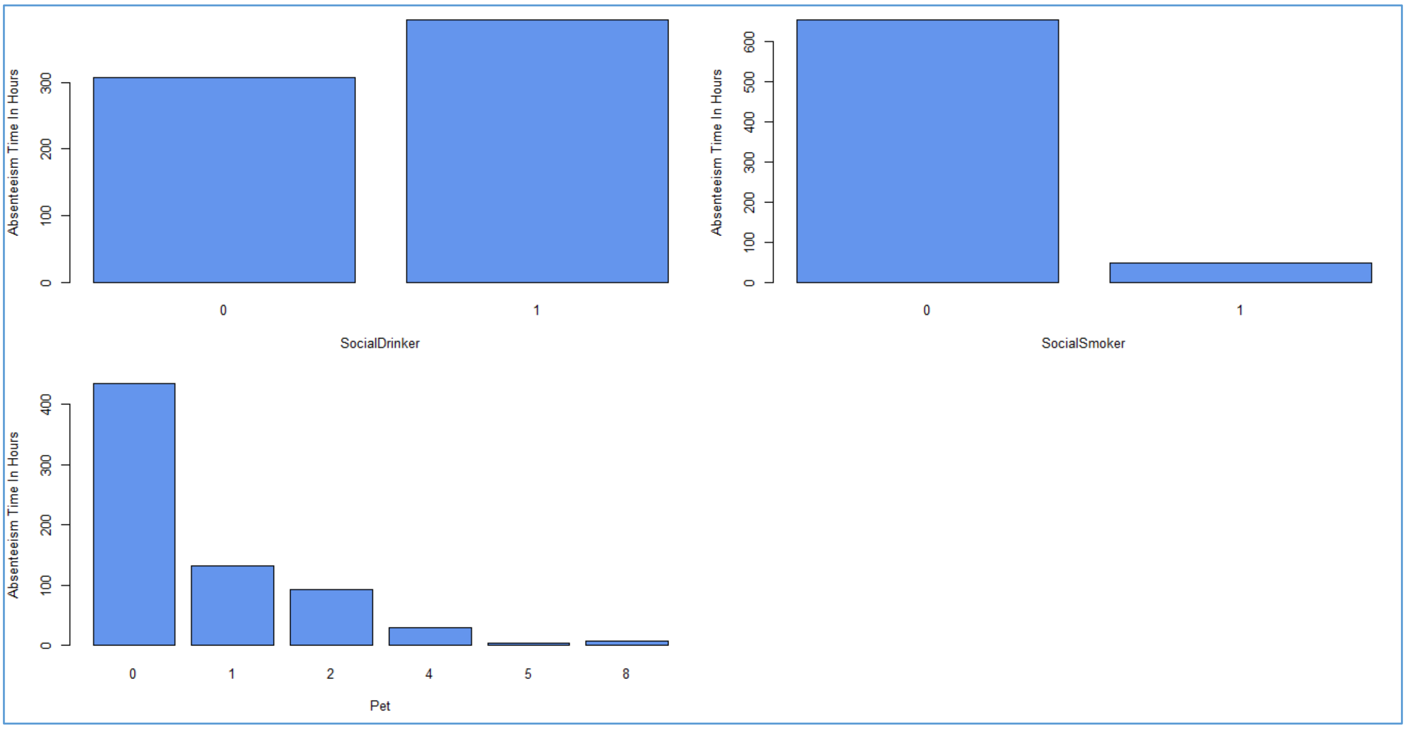
3.3 Data Analysis

The next step is to try and draw conclusions from the dataset. This can be performed by creating various plots of the variables.

Bar graphs of the categorical variables –







From these graphs some *primitive conclusions* can be drawn, such as:

# Employee with ID = 3 has the most entries for absenteeism.

# The top 5 reasons for absenteeism (in number of entries) are 23 (*Medical Consultation*), 28 (*Dental Consultation*), 27 (*Physiotherapy*), 13 (*Musculoskeletal Diseases*), and 19 (*Injury, poisoning or other external causes*).

# Most entries for absenteeism is in month number 3 (March).

# Most entries for absenteeism are for the second day of the week (Monday).

# Disciplinary Failures are rare.

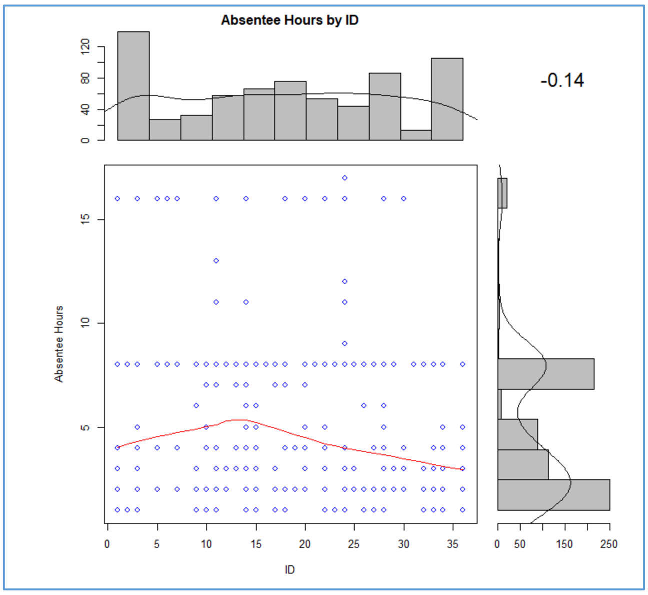
# A sizeable number of employees don't have kids and/or pets.

# There is a moderate number of social drinkers and a miniscule number of social smokers.

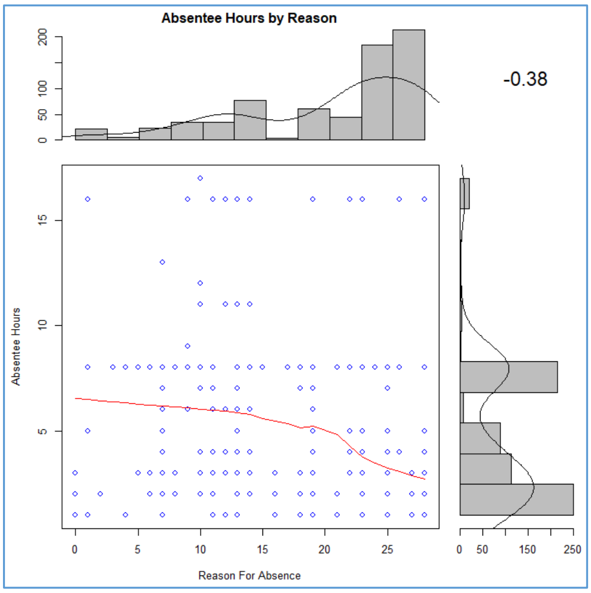
In order to understand the relation and draw conclusions in a better way, plots must be created which show the variables accounting for the highest counts in the target variable (Absenteeism time in hours).

Following are the joint scatter-histogram plots that show the correlation between the independent variables and the target variable:

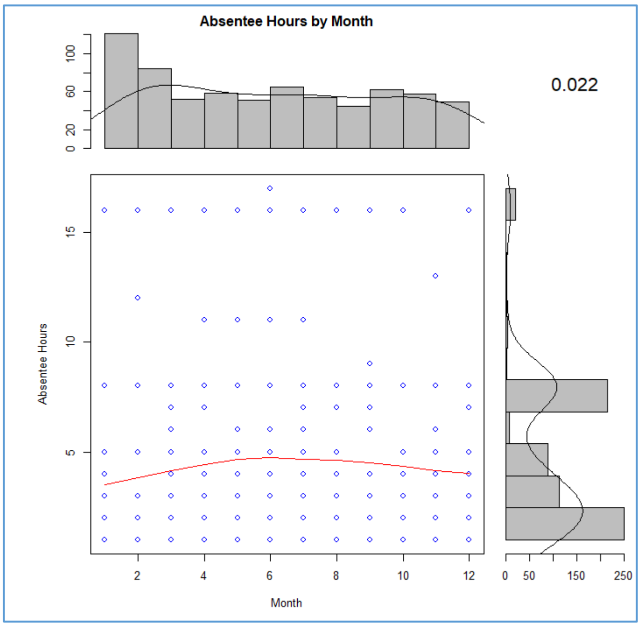
Categorical Variables



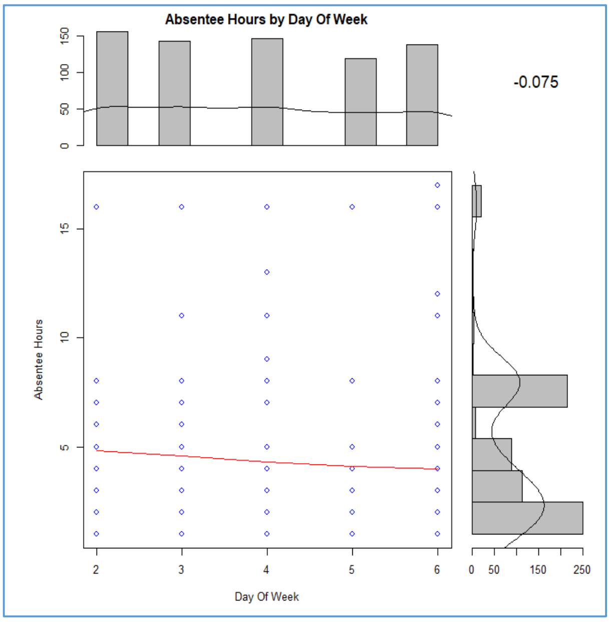
Absenteeism Hours by ID



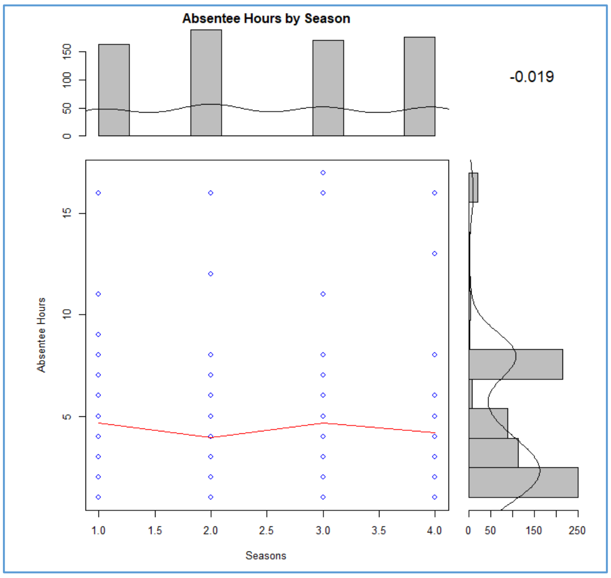
Absenteeism Hours by Reason



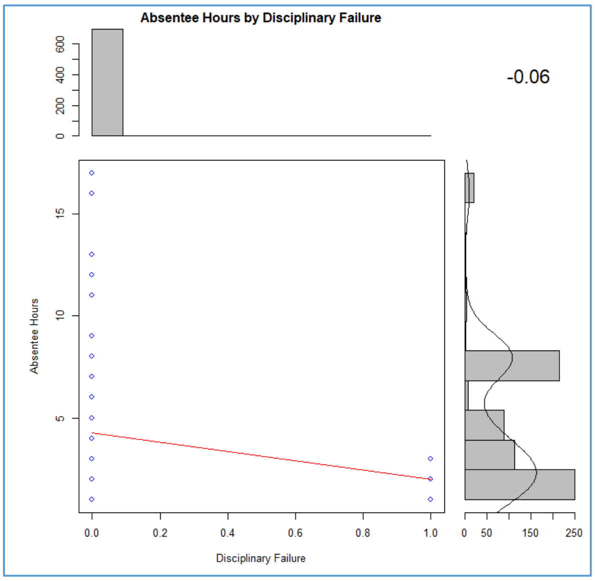
Absenteeism Hours by Month



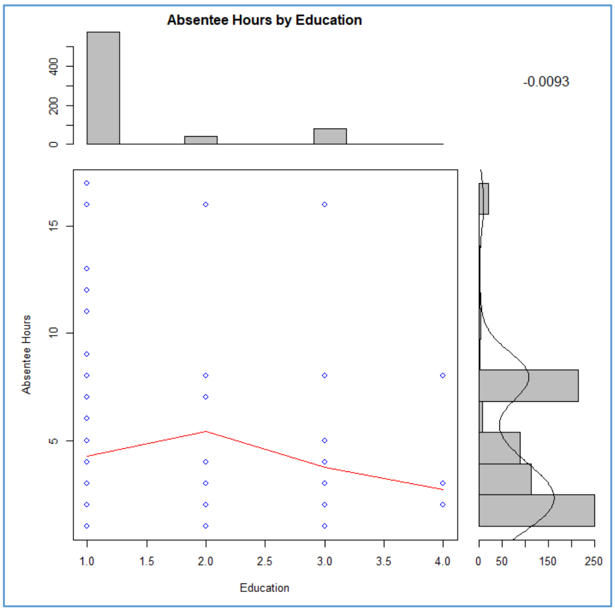
Absenteeism Hours by Day Of Week



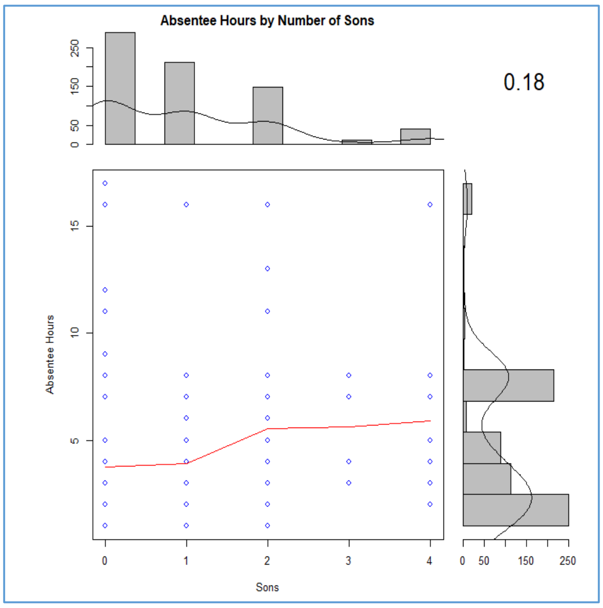
Absenteeism Hours by Season



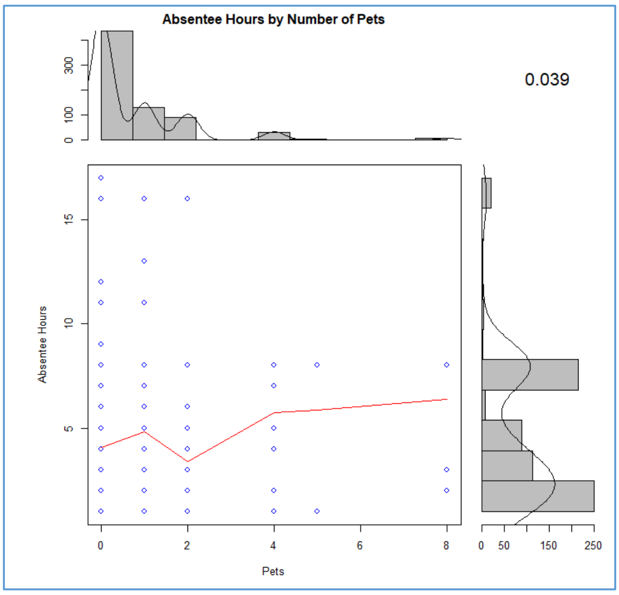
Absenteeism Hours by Disciplinary Failure



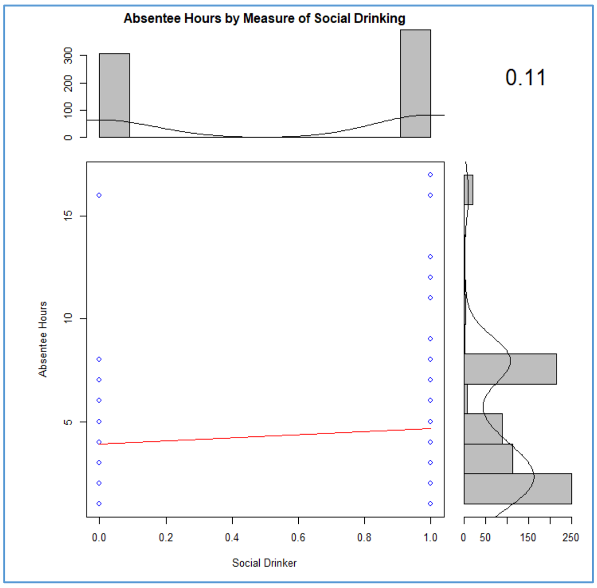
Absenteeism Hours by Education



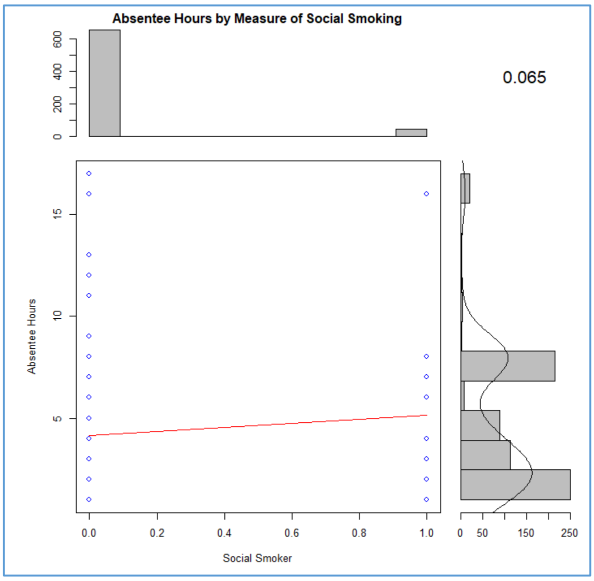
Absenteeism Hours by No. Of Sons



Absenteeism Hours by No. Of Pets

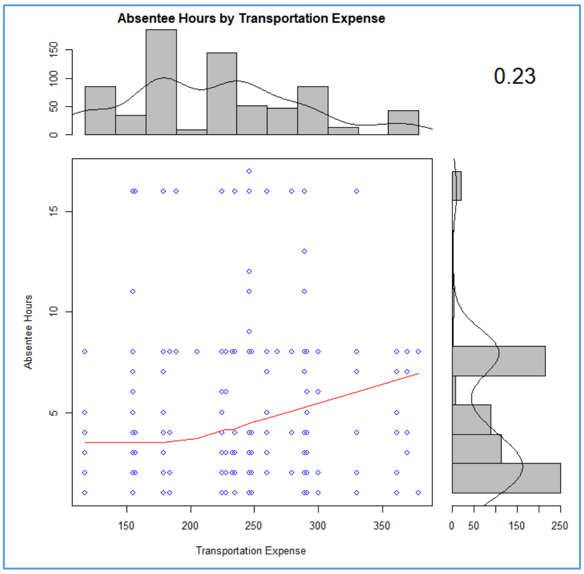


Absenteeism Hours by No. Of Social Drinkers

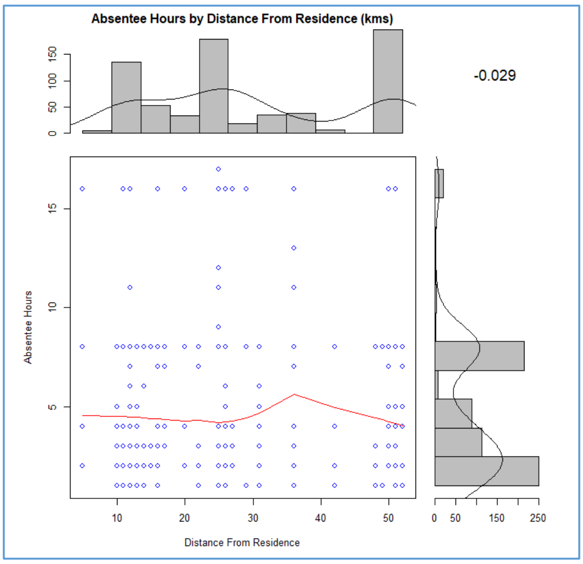


Absenteeism Hours by No. Of Social Smokers

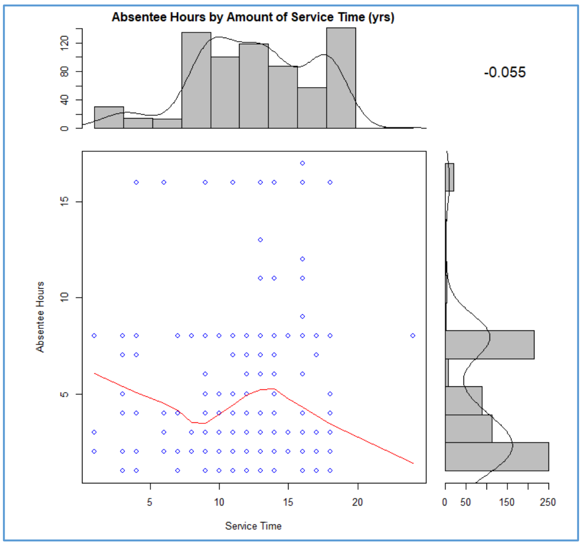
Continuous Variables



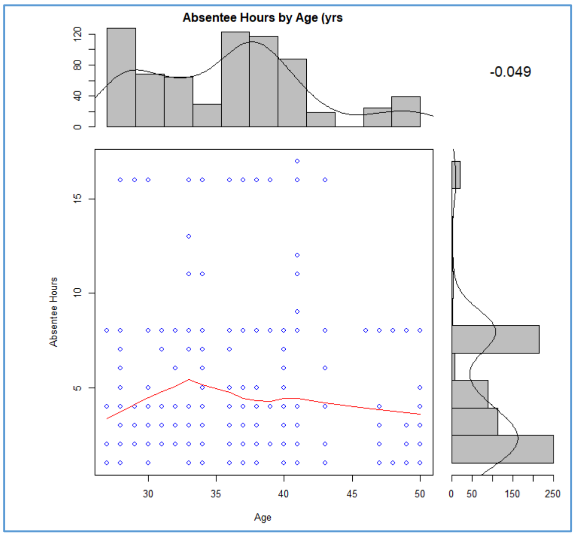
Absenteeism Hours by Transportation Expense



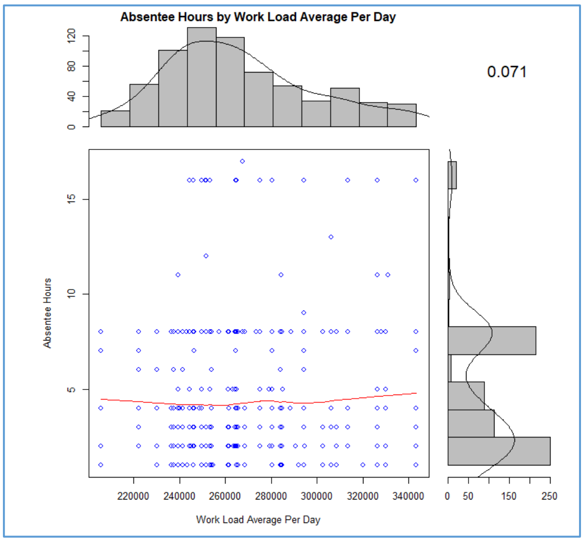
Absenteeism Hours by Distance From Residence



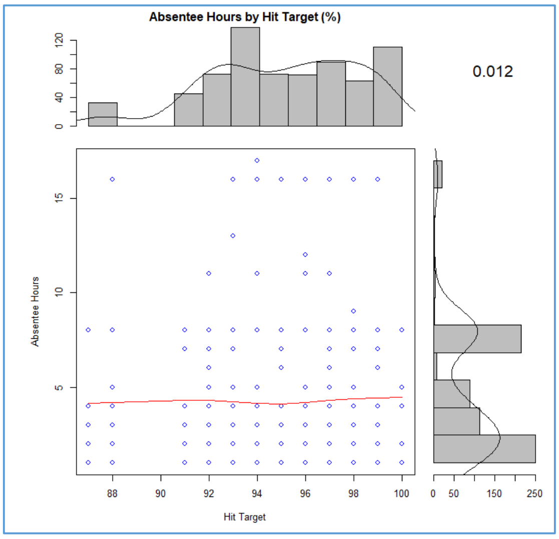
Absenteeism Hours by Service Time



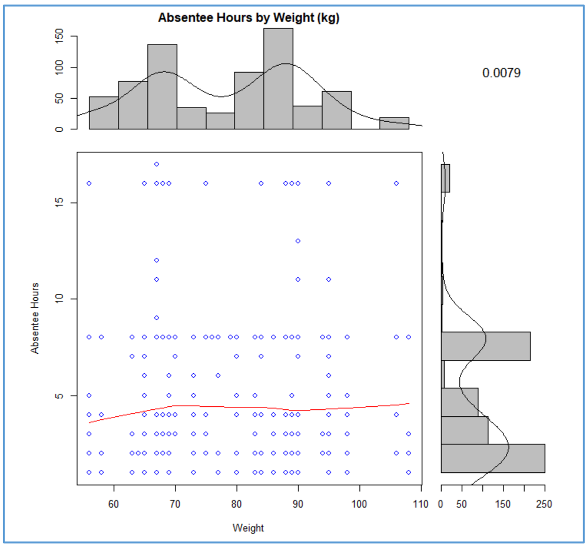
Absenteeism Hours by Age



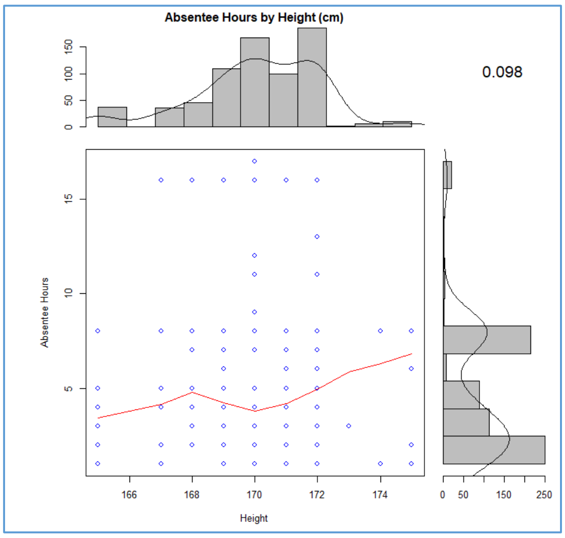
Absenteeism Hours by Workload Average / Day



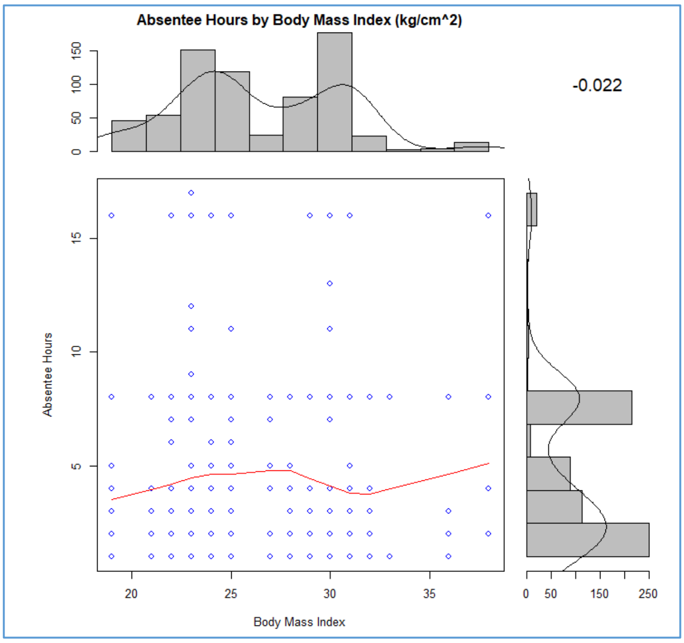
Absenteeism Hours by Target Hit (%)



Absenteeism Hours by Weight



Absenteeism Hours by Height



Absenteeism Hours by Body Mass Index

From these graphs, some *proper informed conclusions* can be drawn, such as:

The company lost most hours to absenteeism due to employees having *Physiotherapy* and *Medical Consultations*.

The company lost most hours to absenteeism in the months of *February*, *July* and *December*.

The company lost significant hours to absenteeism from employees who have residences *farthest from work*.

The company lost significant hours to absenteeism from employees who are nearing *10* or *20* years of service.

The company lost most hours to absenteeism from employees in the *36-42* age group, and lost significant hours to absenteeism from employees in the *28-29* age group.

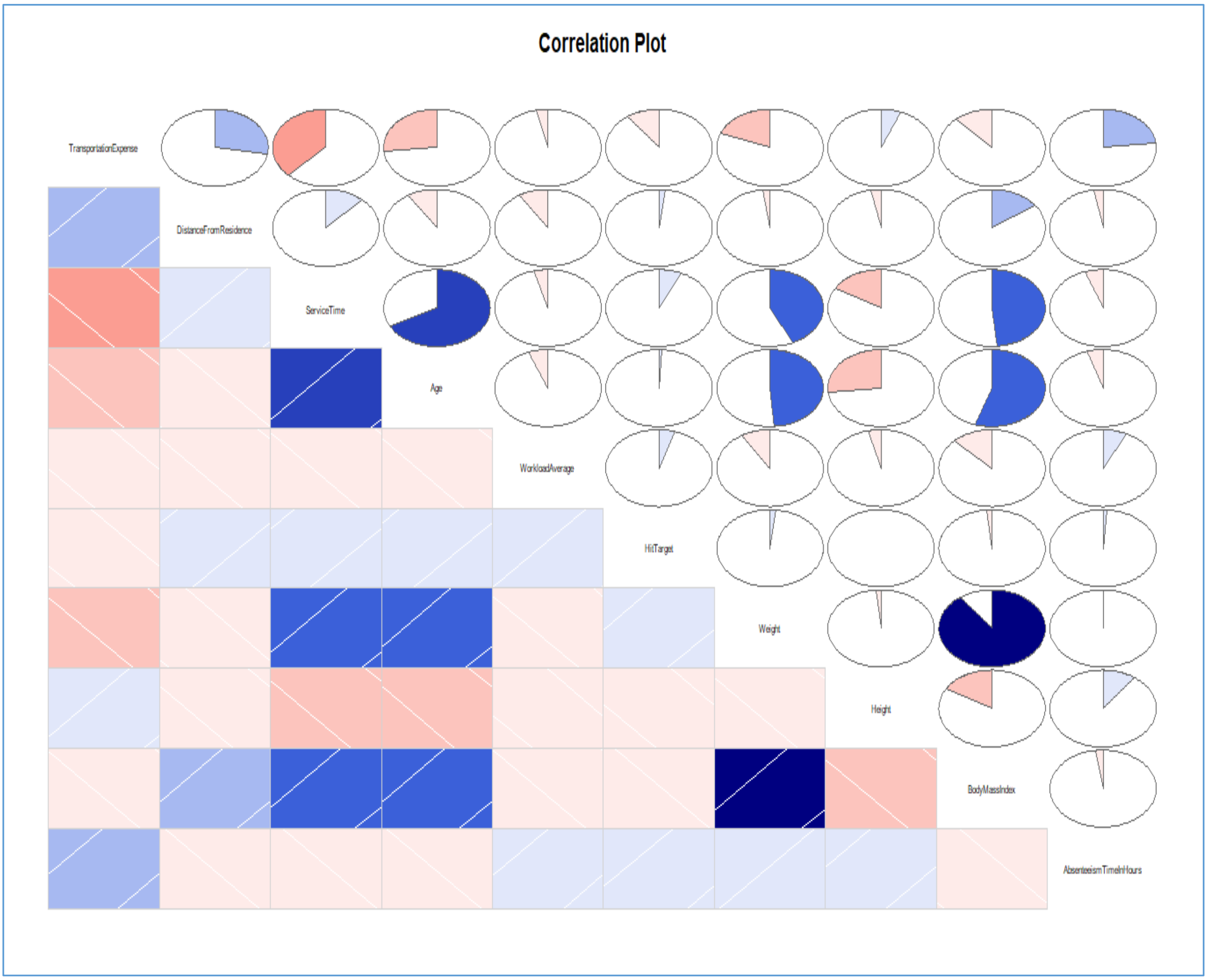
The company lost most hours to absenteeism from employees with average daily workload of *270000* units, and lost significant hours to absenteeism from employees with average daily workload of *240000* units.

The company lost most hours to absenteeism from employees who hit a target of *93%*, and lost significant hours to absenteeism from employees who hit the target range of *96-99%*.

3.4 Feature Selection

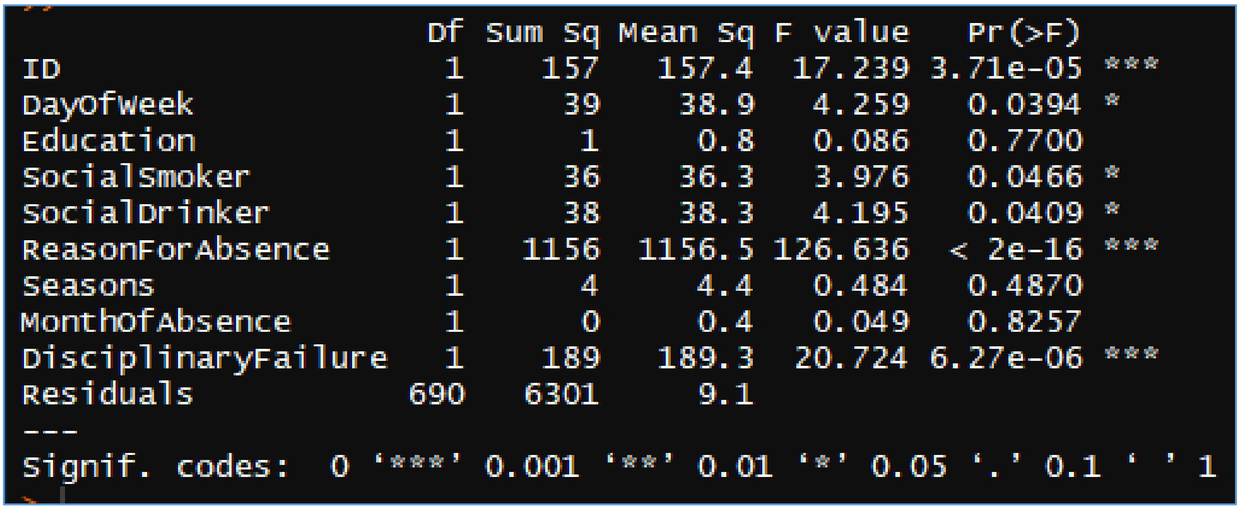
The variables are then tested for correlation so as to reduce their number for the ease of handling the dataset. Variables which are deemed to be heavily correlated to other variable/s are then removed form the dataset.

The correlation among continuous variables is checked by created a correlation matrix and depicting it using a heatmap:



Here, it is apparent from the heatmap that the column BodyMassIndex is very highly correlated to the Weight column. Therefore, the BodyMassIndex column is dropped since it is the derived variable (BodyMassIndex = Weight/Height^2).

To check correlation for categorical variables, the ANOVA test (Analysis Of Variance) is performed:



```
                     Df Sum Sq Mean Sq F value   Pr(>F)    
ID                    1    157   157.4  17.239 3.71e-05 ***
DayOfWeek             1     39    38.9   4.259   0.0394 *  
Education             1      1     0.8   0.086   0.7700    
SocialSmoker          1     36    36.3   3.976   0.0466 *  
SocialDrinker         1     38    38.3   4.195   0.0409 *  
ReasonForAbsence      1   1156  1156.5 126.636  < 2e-16 ***
Seasons               1      4     4.4   0.484   0.4870    
MonthOfAbsence        1      0     0.4   0.049   0.8257    
DisciplinaryFailure   1    189   189.3  20.724 6.27e-06 ***
Residuals           690   6301     9.1                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

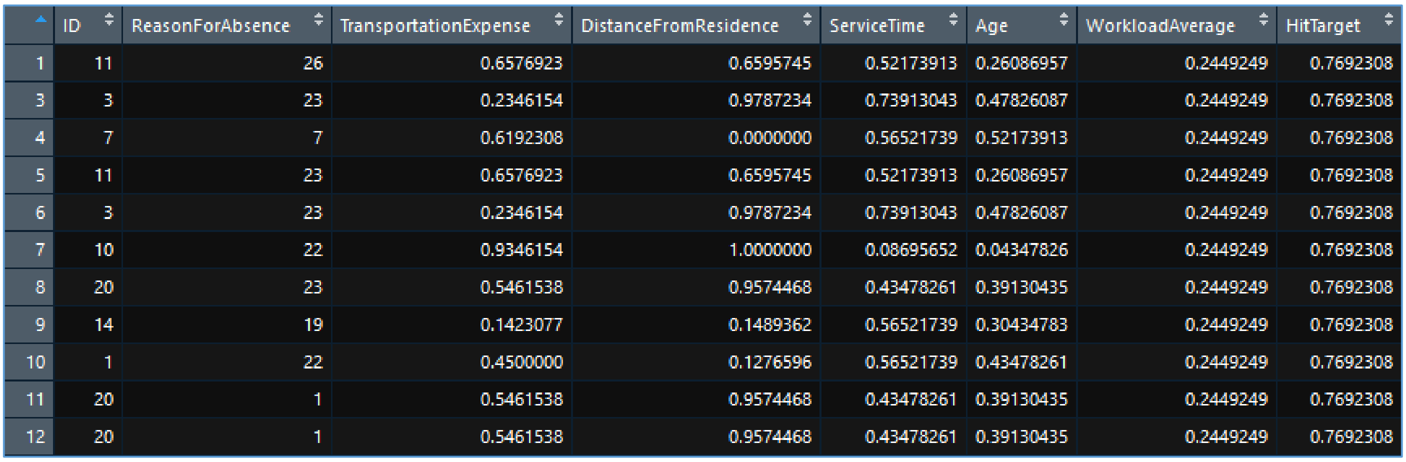
It is instantly observable that all columns except ID and ReasonForAbsence have p-values more than 0.05. Therefore, all categorical variables except ReasonForAbsence are dropped from the dataset.

3.5 Feature Scaling

To get rid of the unwanted variation within or between variables, the observations need to be scaled to conform to uniform distribution. Of the two available options, Normalization is used for the continuous variables in the dataset (since there is only one categorical variable left after the Feature Selection process).

$$\text{Normalized value} = \frac{\text{Individual Observation} - \text{Min(Observations)}}{\text{Max(Observations)} - \text{Min(Observations)}}$$

After Feature Scaling is performed, the dataset looks like this:



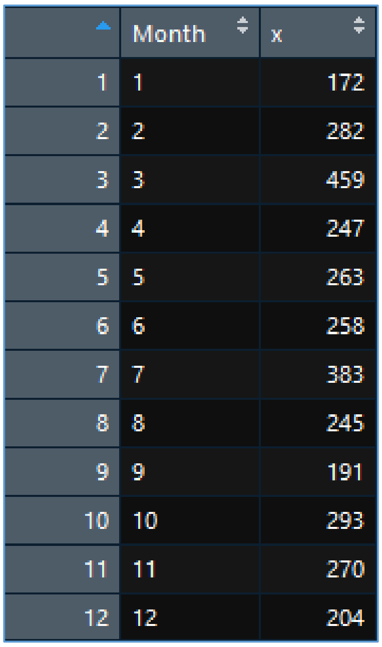
	ID	ReasonForAbsence	TransportationExpense	DistanceFromResidence	ServiceTime	Age	WorkloadAverage	HitTarget
1	11	26	0.6576923	0.6595745	0.52173913	0.26086957	0.2449249	0.7692308
3	3	23	0.2346154	0.9787234	0.73913043	0.47826087	0.2449249	0.7692308
4	7	7	0.6192308	0.0000000	0.56521739	0.52173913	0.2449249	0.7692308
5	11	23	0.6576923	0.6595745	0.52173913	0.26086957	0.2449249	0.7692308
6	3	23	0.2346154	0.9787234	0.73913043	0.47826087	0.2449249	0.7692308
7	10	22	0.9346154	1.0000000	0.08695652	0.04347826	0.2449249	0.7692308
8	20	23	0.5461538	0.9574468	0.43478261	0.39130435	0.2449249	0.7692308
9	14	19	0.1423077	0.1489362	0.56521739	0.30434783	0.2449249	0.7692308
10	1	22	0.4500000	0.1276596	0.56521739	0.43478261	0.2449249	0.7692308
11	20	1	0.5461538	0.9574468	0.43478261	0.39130435	0.2449249	0.7692308
12	20	1	0.5461538	0.9574468	0.43478261	0.39130435	0.2449249	0.7692308

All values of the continuous variables are now situated between 0 and 1.

3.6 Time Series Forecasting

The problem statement specifies that the company has requested for forecasts regarding the absenteeism hours for each month of the year 2011. Time series modelling will be used to forecast absenteeism hours for each month.

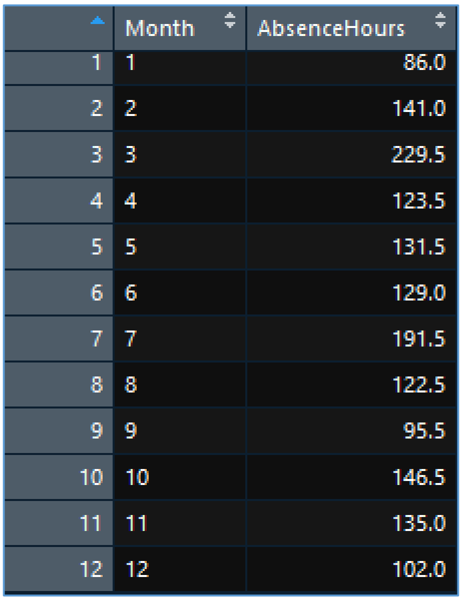
Absenteeism hours aggregated by month gives the following dataset:



	Month	x
1	1	172
2	2	282
3	3	459
4	4	247
5	5	263
6	6	258
7	7	383
8	8	245
9	9	191
10	10	293
11	11	270
12	12	204

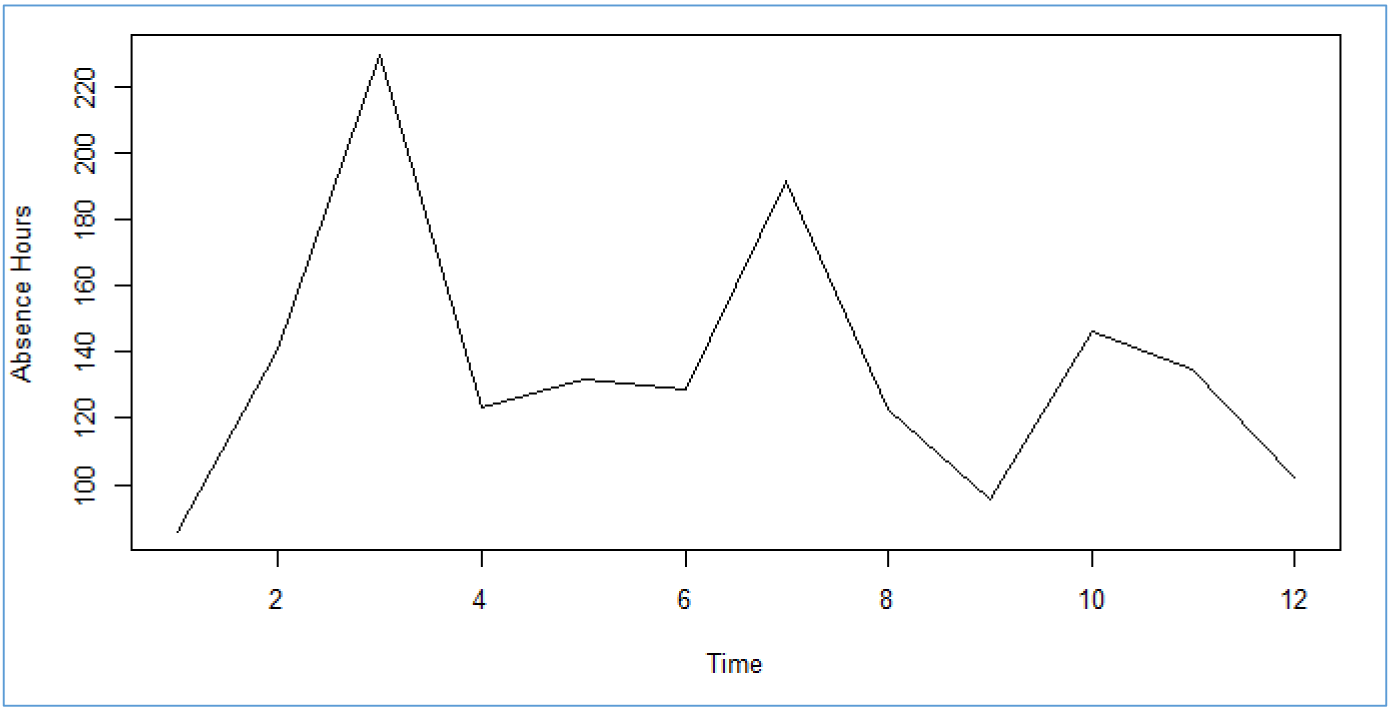
As timeframe of dataset isn't provided, proper course of action would be to assume a timeframe and proceed with the modelling. The minimum timeframe appropriate for the modelling would be two years, since the longer the timeframe, the more accurate the predictions.

Dividing the target variables by 2, to help with the assumption:

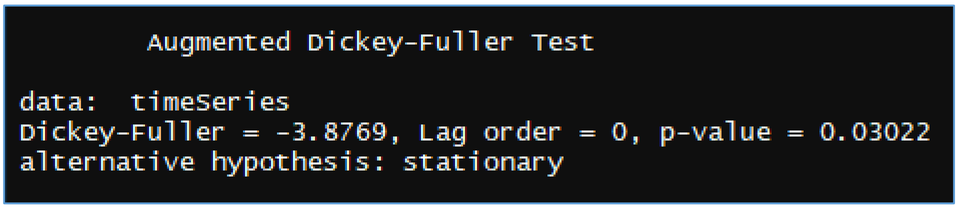


	Month	AbsenceHours
1	1	86.0
2	2	141.0
3	3	229.5
4	4	123.5
5	5	131.5
6	6	129.0
7	7	191.5
8	8	122.5
9	9	95.5
10	10	146.5
11	11	135.0
12	12	102.0

Plotting the time series:



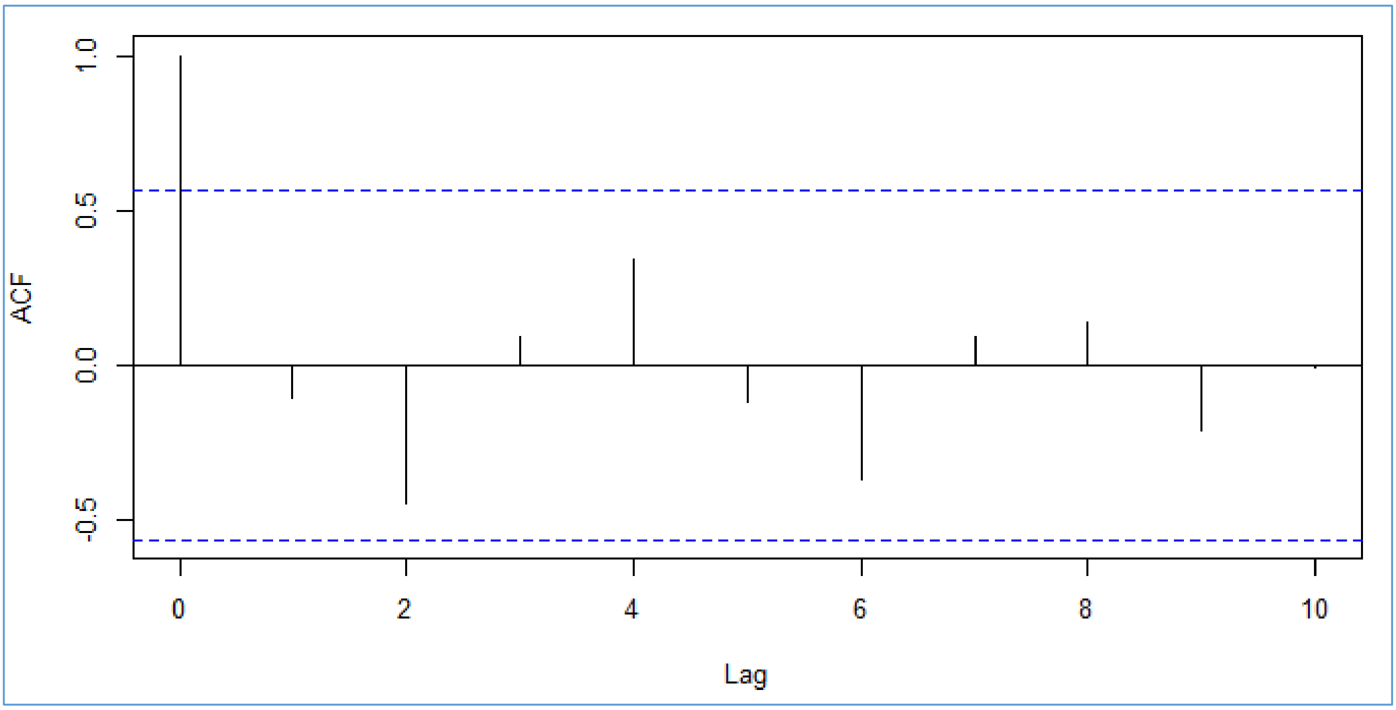
The next step is to perform the Augmented Dickey-Fuller test on the series, to check the stationarity of the series.
The test statistics after performing the ADF test on the series:



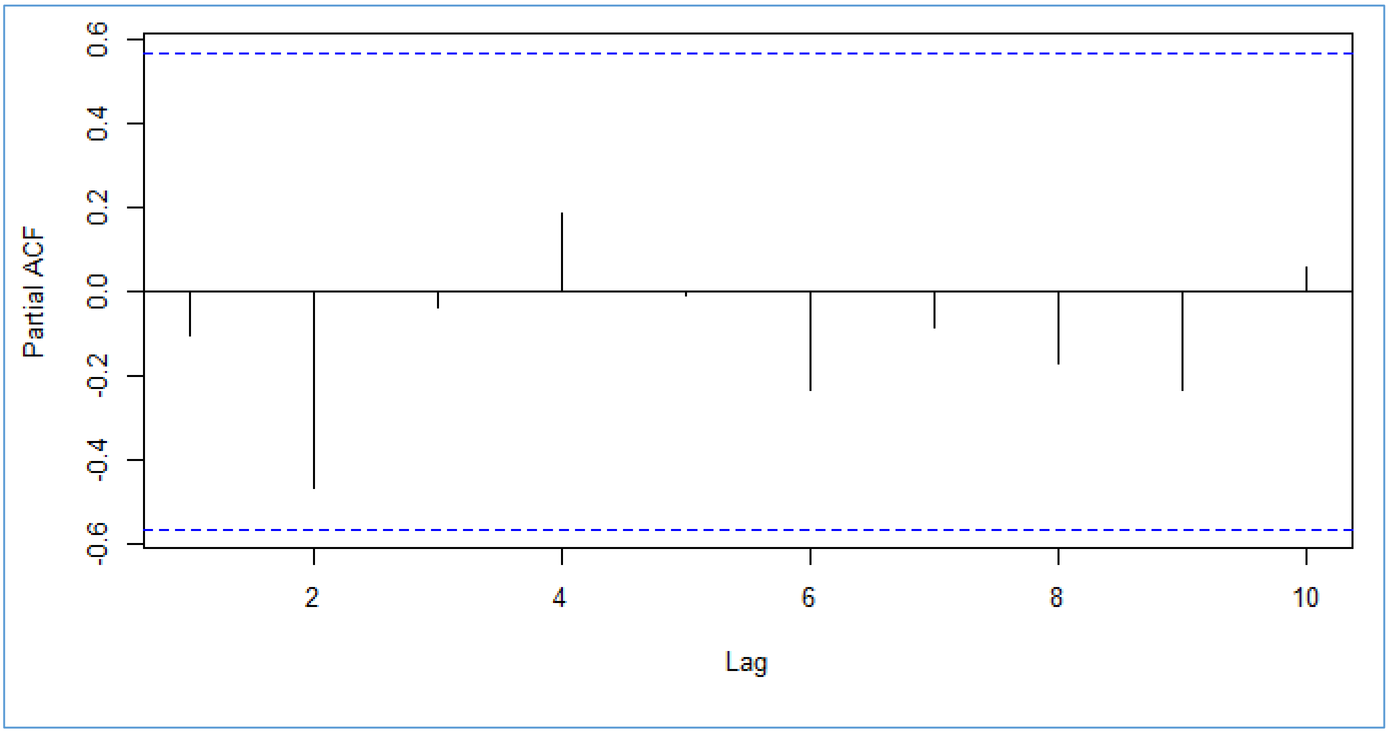
```
        Augmented Dickey-Fuller Test

data:  timeSeries
Dickey-Fuller = -3.8769, Lag order = 0, p-value = 0.03022
alternative hypothesis: stationary
```

Since the p-value is less than 0.05, the null hypothesis can be rejected, the series can be classified as stationary, and we can proceed further to the Autocorrelation and Partial Autocorrelation functions (ACF & PACF).

Creating the ACF and PACF plots:



Autocorrelation Function Plot

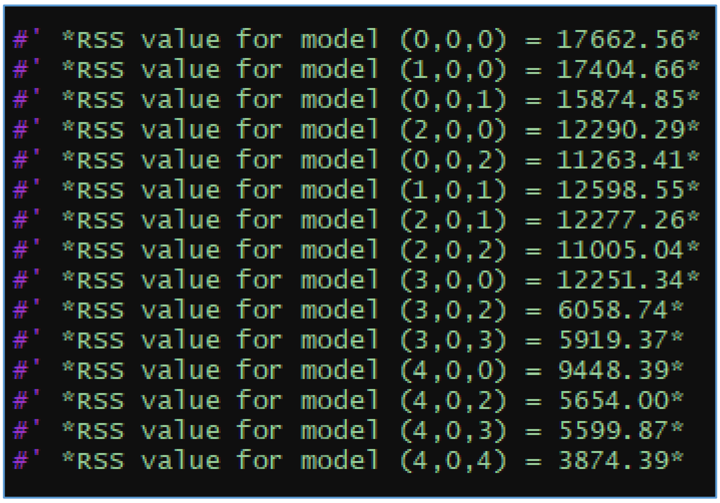


Partial Autocorrelation Plot

Since the plots don’t give any clear values of the metrics p and q, the model ARIMA(0,0,0) is taken as the initial candidate, with the other models to be fitted being ARIMA(1,0,0), ARIMA(0,1,0), ARIMA(2,0,0), and ARIMA(0,2,0); with ARIMA(3,0,0), ARIMA(3,0,2), ARIMA(3,0,3), ARIMA(4,0,0), ARIMA(4,0,2), ARIMA(4,0,3) and ARIMA(4,0,4) also being fitted.

The models are fitted over the time series, and the Residual Sum of Squares (RSS) is extracted as a parameter for comparison.

The RSS values recorded after the models were fitted are:

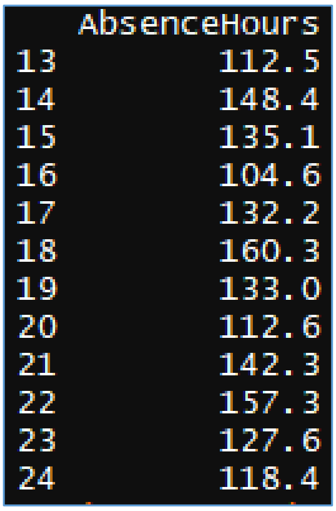


```
#' *RSS value for model (0,0,0) = 17662.56*
#' *RSS value for model (1,0,0) = 17404.66*
#' *RSS value for model (0,0,1) = 15874.85*
#' *RSS value for model (2,0,0) = 12290.29*
#' *RSS value for model (0,0,2) = 11263.41*
#' *RSS value for model (1,0,1) = 12598.55*
#' *RSS value for model (2,0,1) = 12277.26*
#' *RSS value for model (2,0,2) = 11005.04*
#' *RSS value for model (3,0,0) = 12251.34*
#' *RSS value for model (3,0,2) = 6058.74*
#' *RSS value for model (3,0,3) = 5919.37*
#' *RSS value for model (4,0,0) = 9448.39*
#' *RSS value for model (4,0,2) = 5654.00*
#' *RSS value for model (4,0,3) = 5599.87*
#' *RSS value for model (4,0,4) = 3874.39*
```

Since the model ARIMA(4,0,4) gives the lowest RSS value of 3874.39, it will be used for the forecasting of the time series.

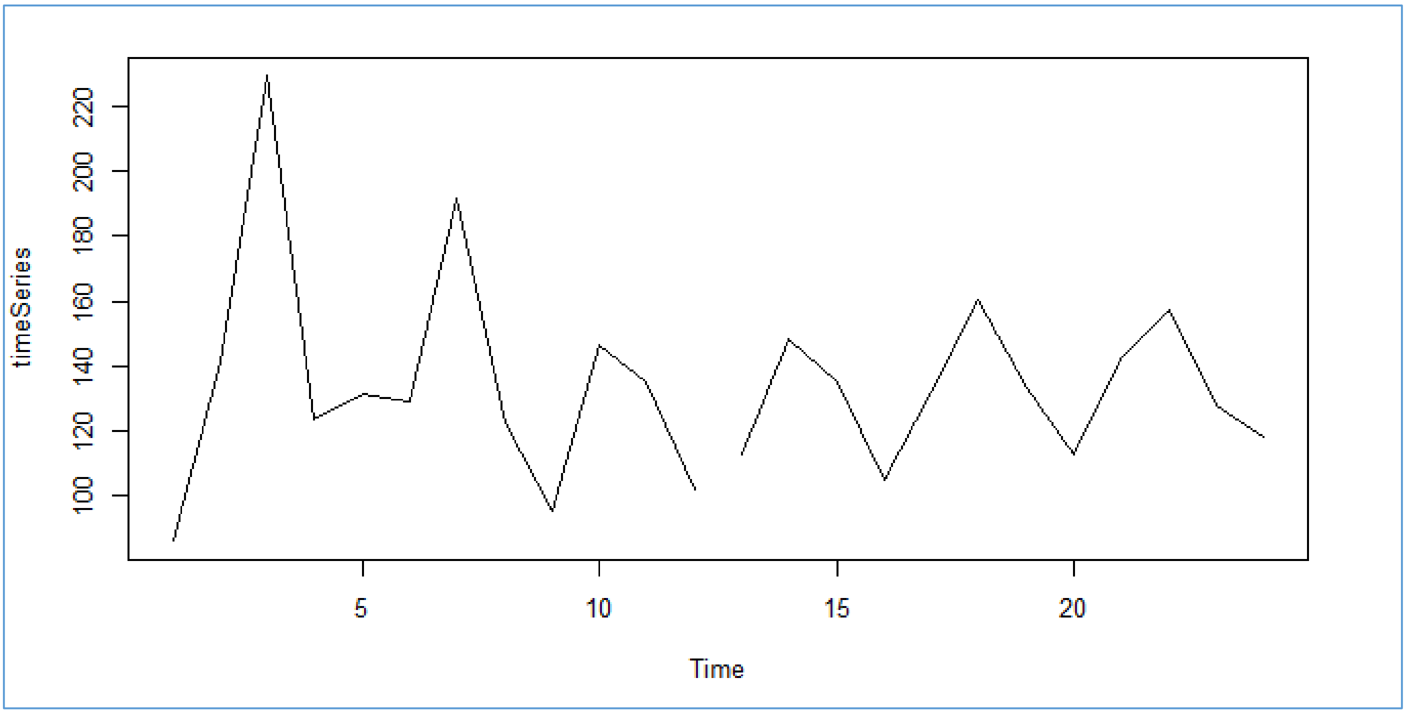
ARIMA(4,0,4) is fitted over the time series, and the fitted values are acquired and compared with the time series.

The predictions are made over the fit of the chosen model, and then added to a time series for the next 12 months (for the next year).
The predicted absenteeism hours for the next 12 months are as follows:



```
   AbsenceHours
13        112.5
14        148.4
15        135.1
16        104.6
17        132.2
18        160.3
19        133.0
20        112.6
21        142.3
22        157.3
23        127.6
24        118.4
```

Plotting the provided values (of absenteeism hours) and the forecasted values together:



The predictions appear to be a little less irregular than the provided data, but that's because of the assumption that the provided data is of the past two years (2009 and 2010). Had there been a longer timeframe of the data, the irregularity would have been much lower. Not to forget that the data was simply halved instead of having separate sets of observations for the two years in question.

Chapter 4

Conclusion

4.1 Changes to reduce number of absenteeism hours

Two of the top three reasons contributing to highest number of absenteeism hours are discovered to be **Physiotherapy** and **Diseases of the Musculoskeletal system and connective tissue**. It becomes clear that problems caused by physical stress is resulting in employees being absent. The company could work towards reducing/redistributing the amount of physical tasks among the employees.

The employees with the highest number of absenteeism hours are employees with IDs 3, 28 and 34. The company could investigate those employees to get a better idea about the underlying reasons, and assess if those employees need assistance/support in any way.

The highest number of absenteeism hours have been accumulated in the months of February, July and December. The company could examine what could be the reason behind employees being absent during those particular months, and work towards rectifying the problem, if there is any.

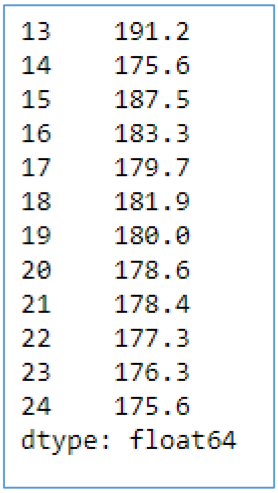
The highest number of absenteeism hours have been accumulated by employees who have hit over 92% of their target. The company can look into the issue and determine if that group of employees requires guidance or support of any sort.

The employees who live the farthest from the job have amassed the highest number of absenteeism hours. The company could try and arrange for special transportation, which might make it easier for those employees to report to and from the job.

4.2 Absenteeism hours/month in 2011

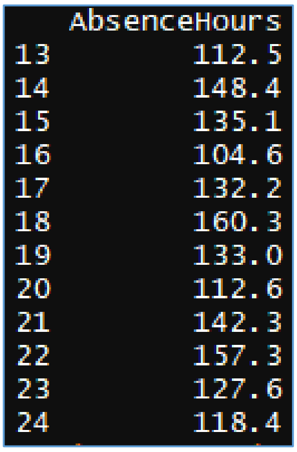
Time series modelling performed on the provided data yielded projections for the absenteeism hours per month for the year of 2011. The projections are as follows:

Modelling In **Python**



```
13     191.2
14     175.6
15     187.5
16     183.3
17     179.7
18     181.9
19     180.0
20     178.6
21     178.4
22     177.3
23     176.3
24     175.6
dtype: float64
```

Modelling In **R**



```
   AbsenceHours
13        112.5
14        148.4
15        135.1
16        104.6
17        132.2
18        160.3
19        133.0
20        112.6
21        142.3
22        157.3
23        127.6
24        118.4
```

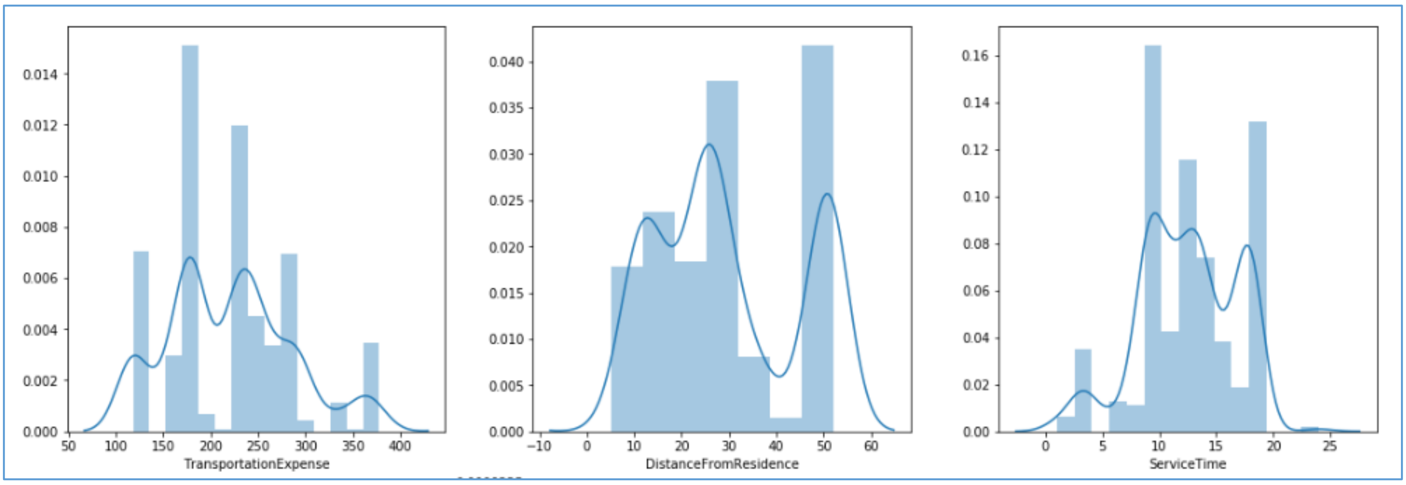
The two sets of projections are fairly distinct, and provide indications of what the company should expect in the twelve months of 2011.

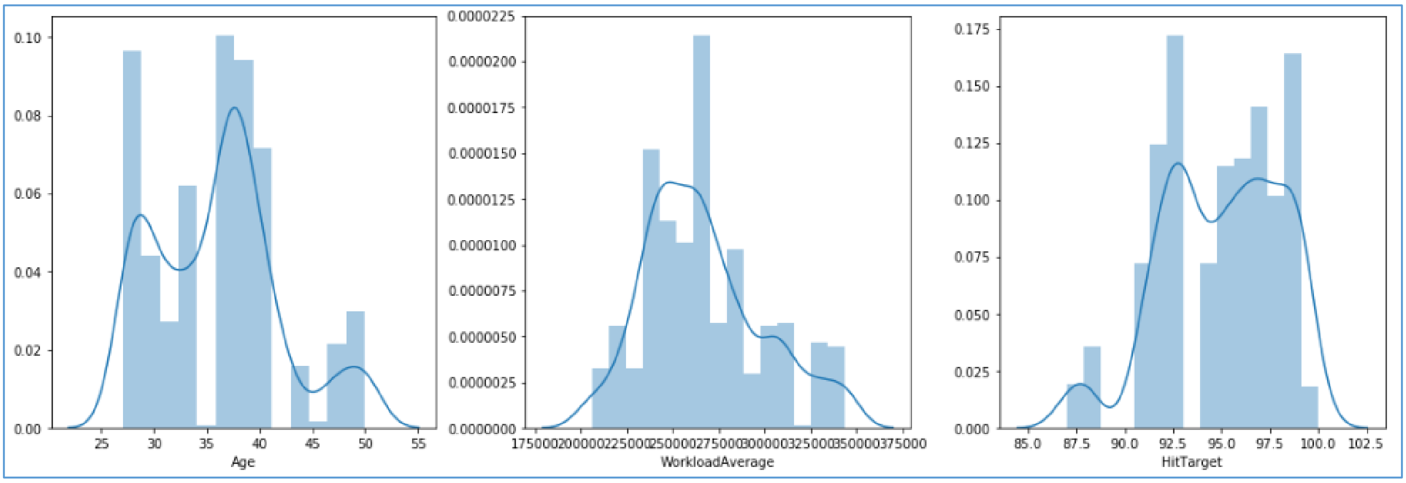
Chapter 5

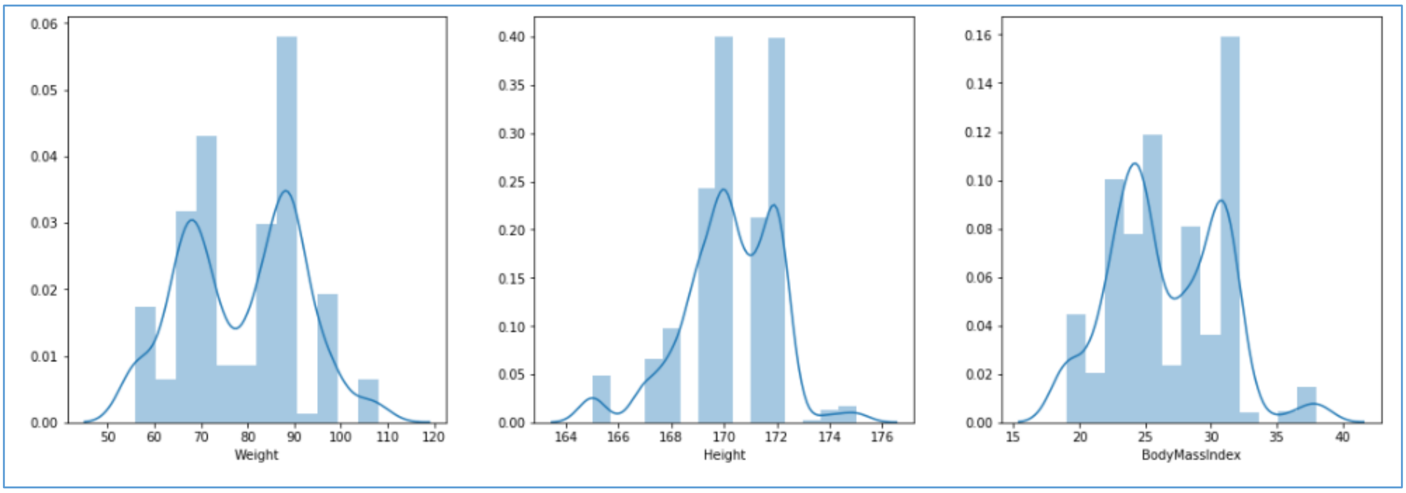
Appendix

5.1 Additional plots

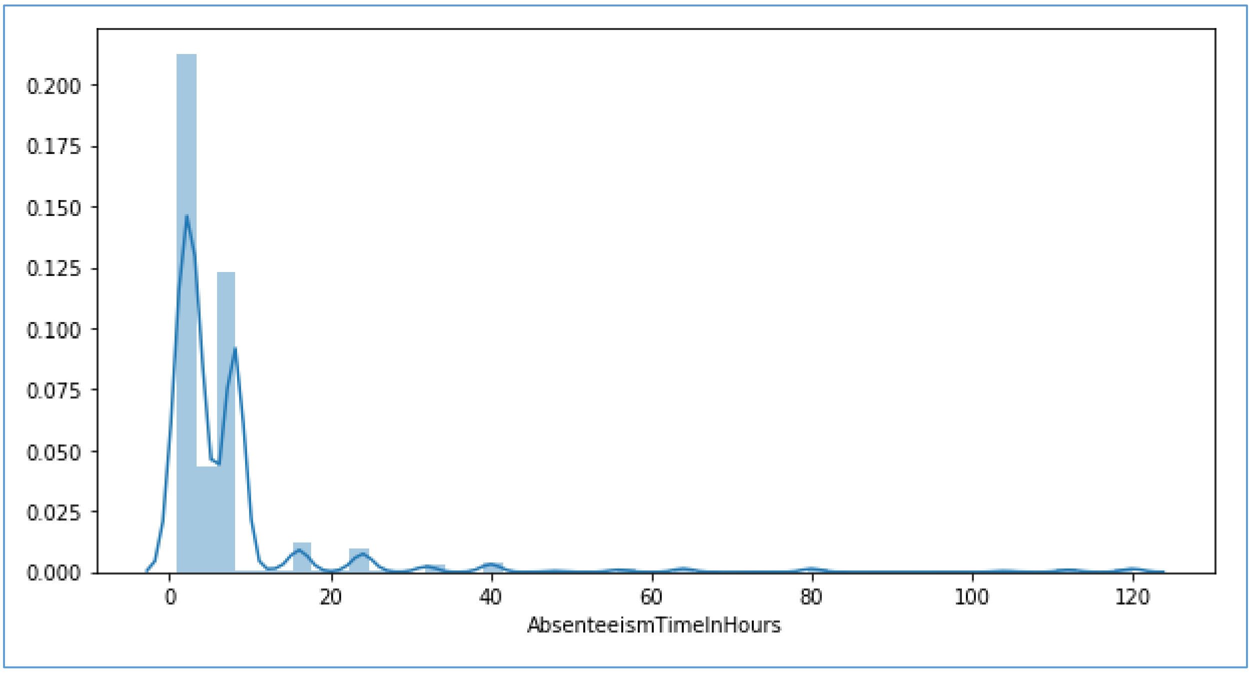
Distribution plots of the continuous variables:







Distribution plot of the target variable:



Chapter 6

References

6.1 Websites used for subject matter

# seaborn.distplot — seaborn 0.10.1 documentation

# seaborn.jointplot — seaborn 0.10.1 documentation

# Patterns in Time Series Analysis - dummies

# Time Series Analysis For Beginners

# Stationarity and differencing | Forecasting: Principles and Practice

# statsmodels.tsa.stattools.adfuller

# Detecting stationarity in time series data

# Identifying the numbers of AR or MA terms in an ARIMA model

# ARIMA Modelling in R | Forecasting: Principles and Practice

# scatter.hist: Draw a scatter plot with associated X and Y histograms

# ts function | R Documentation