7/15/2018

Employee Absenteeism

Analysis and Model Generation in Python and R

Lakshveer Singh

Table of Contents

1. Introduction

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 genuine issue of Absenteeism. The company has shared it dataset and requested to have an answer on the following areas:

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

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

2. Data Pre-processing

2.1. Data

Our first objective is to find the patterns which leads to number of absenteeism and how the changes can help the company reduce that. Given below is a sample of the data set that we are using to find the trend in the absenteeism.

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

2.2. Data Size and Structure:

Size: 740 obs. of 21 variables:
Raw Structure:

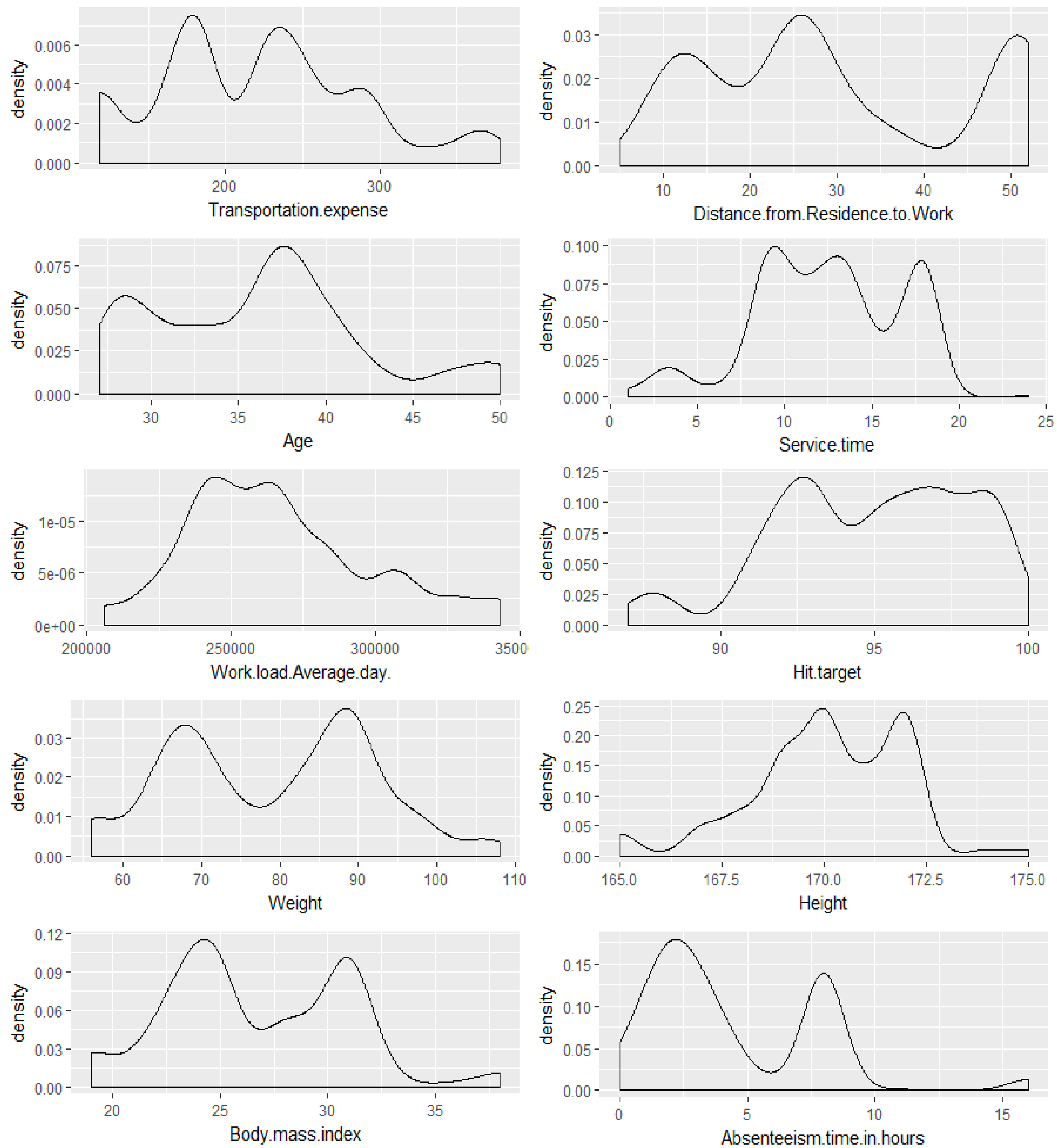
VARIABLES	*DATA TYPE*
ID	num 11 36 3 7 11 3 10 20 14 1 ...
Reason.for.absence	num 26 0 23 7 23 23 22 23 19 22 ...
Month.of.absence	num 7 7 7 7 7 7 7 7 7 7 ...
Day.of.the.week	num 3 3 4 5 5 6 6 6 2 2 ...
Seasons	num 1 1 1 1 1 1 1 1 1 1 ...
Transportation.expense	num 289 118 179 279 289 179 NA 260 155
Distance.from.Residence.to.Work	num 36 13 51 5 36 51 52 50 12 11 ...
Service.time	num 13 18 18 14 13 18 3 11 14 14 ...
Age	num 33 50 38 39 33 38 28 36 34 37 ...
Work.load.Average.day.	num 239554 239554 239554 239554 239554 ..
Hit.target	num 97 97 97 97 97 97 97 97 97 97 ...
Disciplinary.failure	num 0 1 0 0 0 0 0 0 0 0 ...
Education	num 1 1 1 1 1 1 1 1 1 3 ...
Son	num 2 1 0 2 2 0 1 4 2 1 ...
Social.drinker	num 1 1 1 1 1 1 1 1 1 0 ...
Social.smoker	num 0 0 0 1 0 0 0 0 0 0 ...
Pet	num 1 0 0 0 1 0 4 0 0 1 ...
Weight	num 90 98 89 68 90 89 80 65 95 88 ...
Height	num 172 178 170 168 172 170 172 168 196
Body.mass.index	num 30 31 31 24 30 31 27 23 25 29 ...
Absenteeism.time.in.hours	num 4 0 2 4 2 NA 8 4 40 8 ...

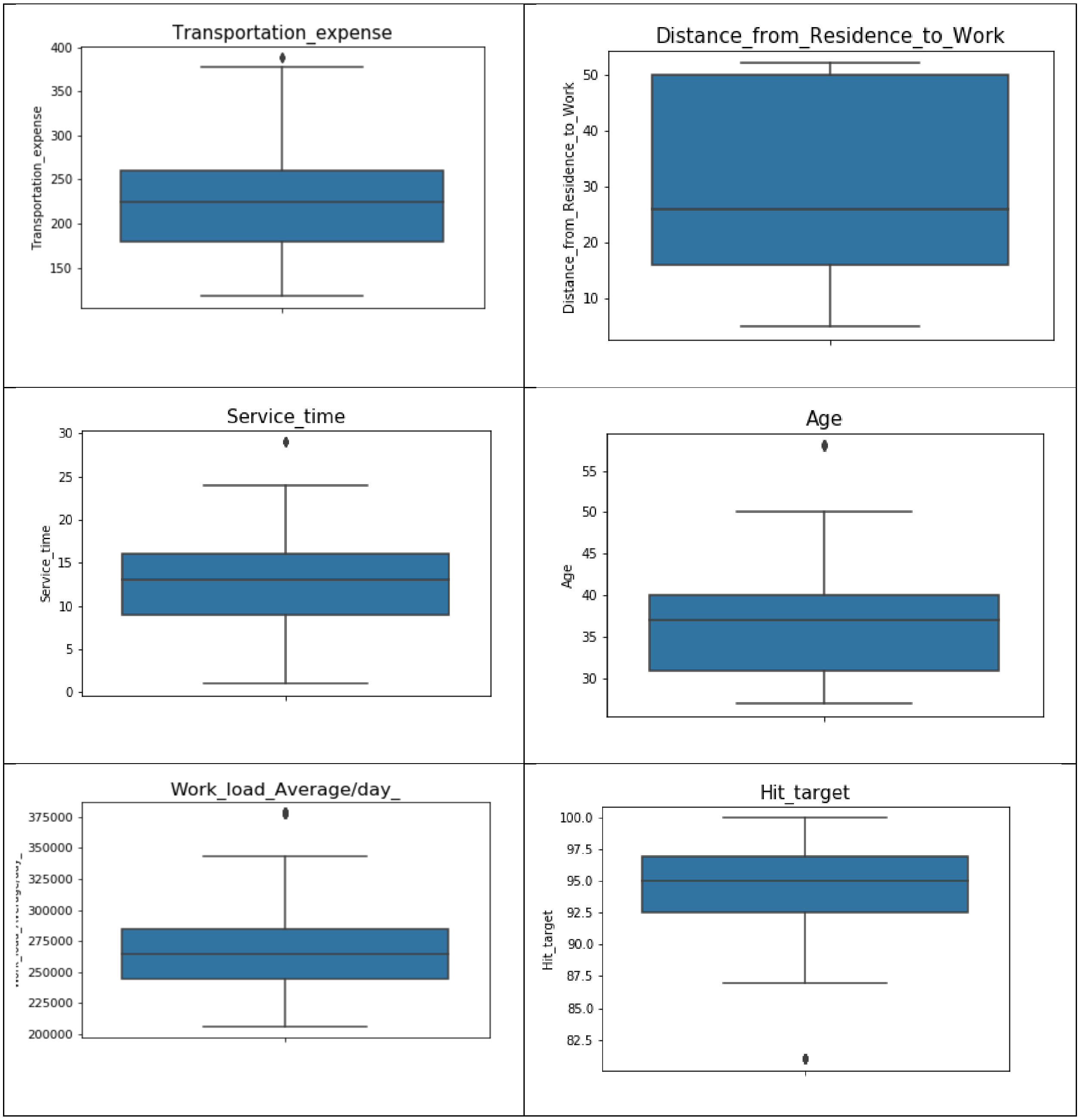
2.3. Completeness of data:

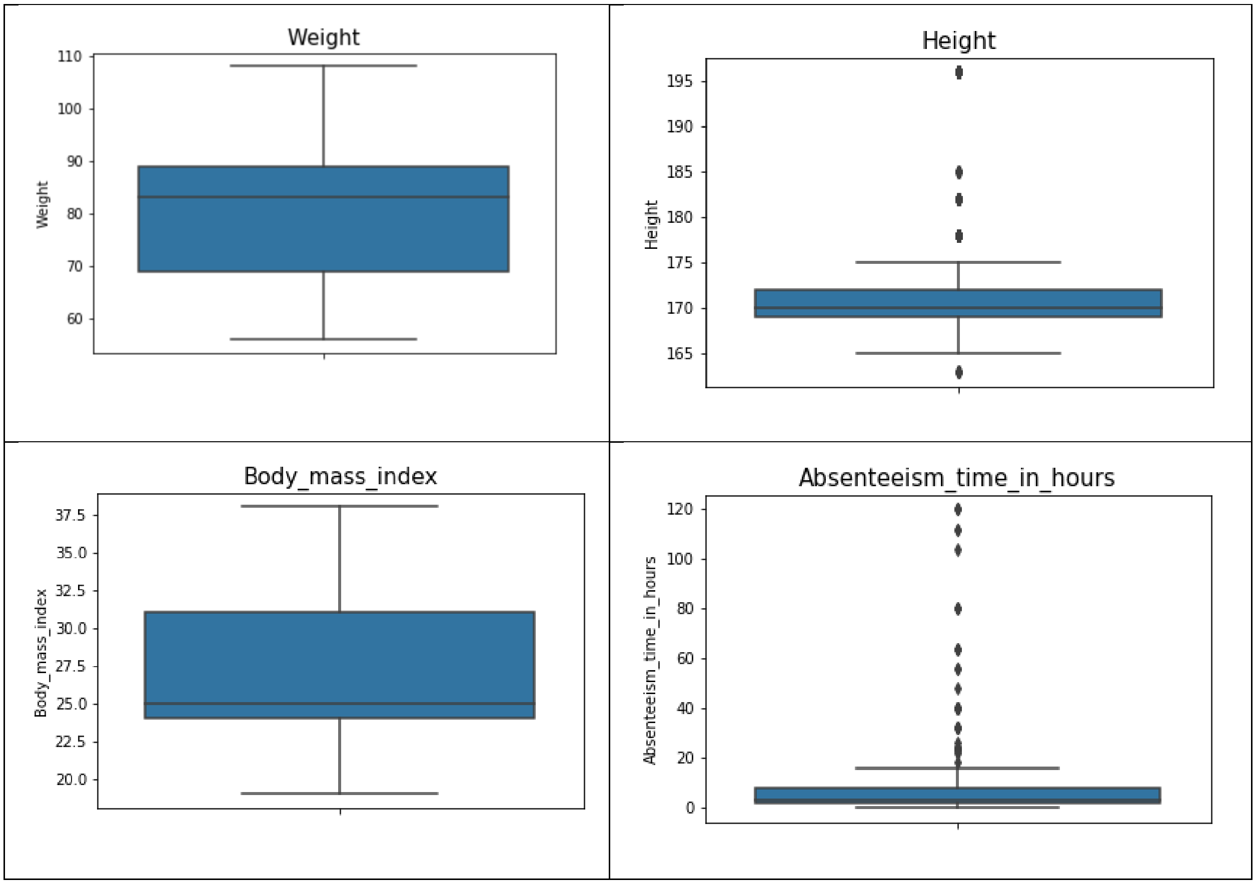
VARIABLES	*MISSING COUNT*
ID	0
Reason.for.absence	3
Month.of.absence	0
Year	0
Day.of.the.week	0
Seasons	0
Transportation.expense	7
Distance.from.Residence.to.Work	3
Service.time	3
Age	3
Work.load.Average.day.	10
Hit.target	6
Disciplinary.failure	6
Education	10
Son	6
Social.drinker	3
Social.smoker	4
Pet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

2.4.Outlier Analysis

By looking at the below probability distribution we can clearly see that the most of the variables are skewed. The skew in the distribution can be most likely explained by the presence of outliers in the data. Let's plot a box to check this.



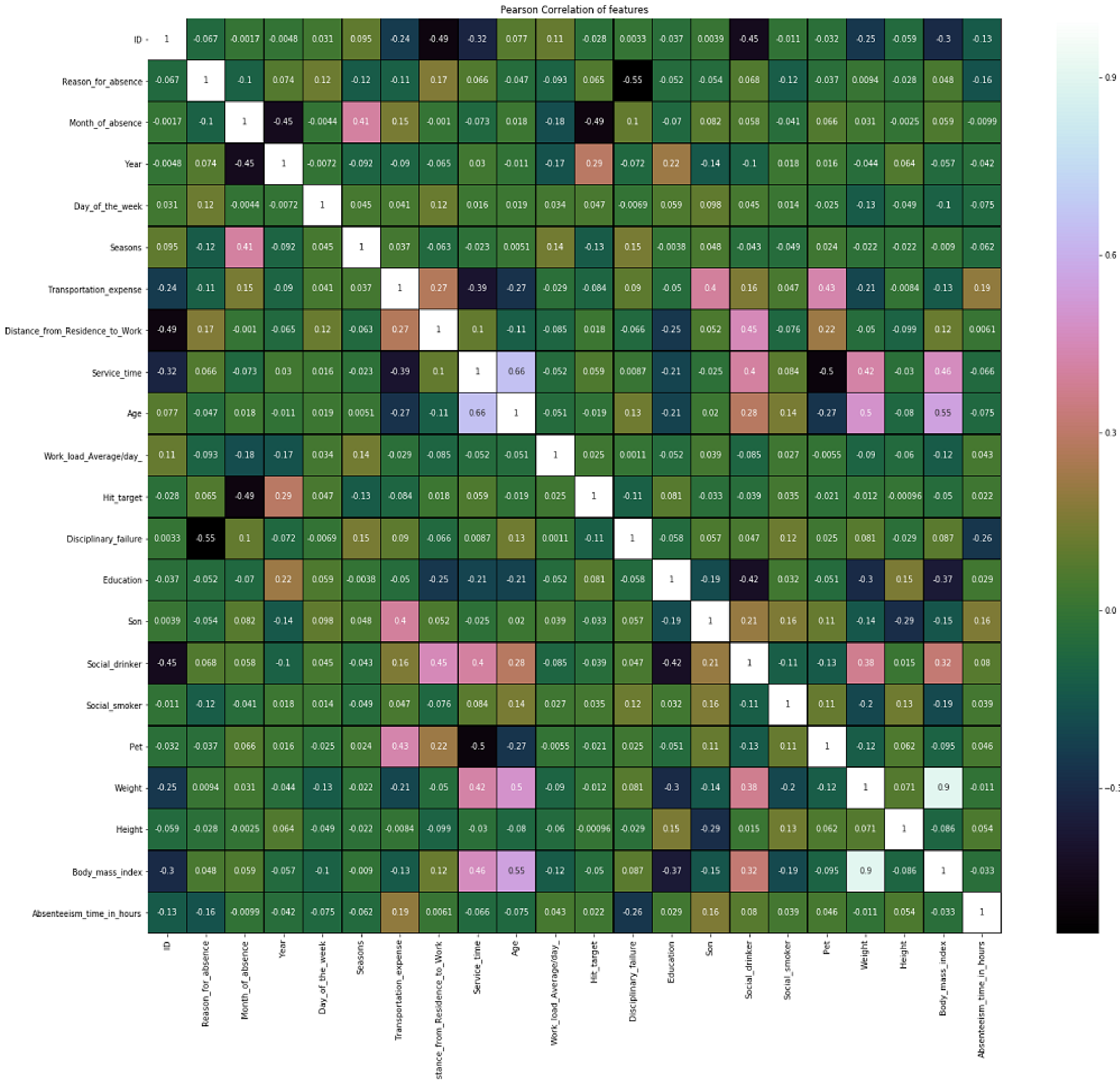

Transportation_expense
Transportation_expense
400
350
300
250
200
150
Distance_from_Residence_to_Work
Distance_from_Residence_to_Work
50
40
30
20
10
Service_time
Service_time
30
25
20
15
10
5
0
Age
Age
55
50
45
40
35
30
Work_load_Average/day_
375000
350000
325000
300000
275000
250000
225000
200000
Hit_target
Hit_target
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5



There seems to be many outlier in our target variable Absenteeism_time_in_hours. There are value of 120, 100, 80, 60 which is not possible.

Reason: The data set is a daily data set with no of absent hour per day. A day has max 24 hours, so all these values seems redundant and we need to eliminate these out. Logically the absenteeism hours should be less than the service time of that employee. *We will use KNN imputation to impute these outliers*.

2.5. Correlation Plot

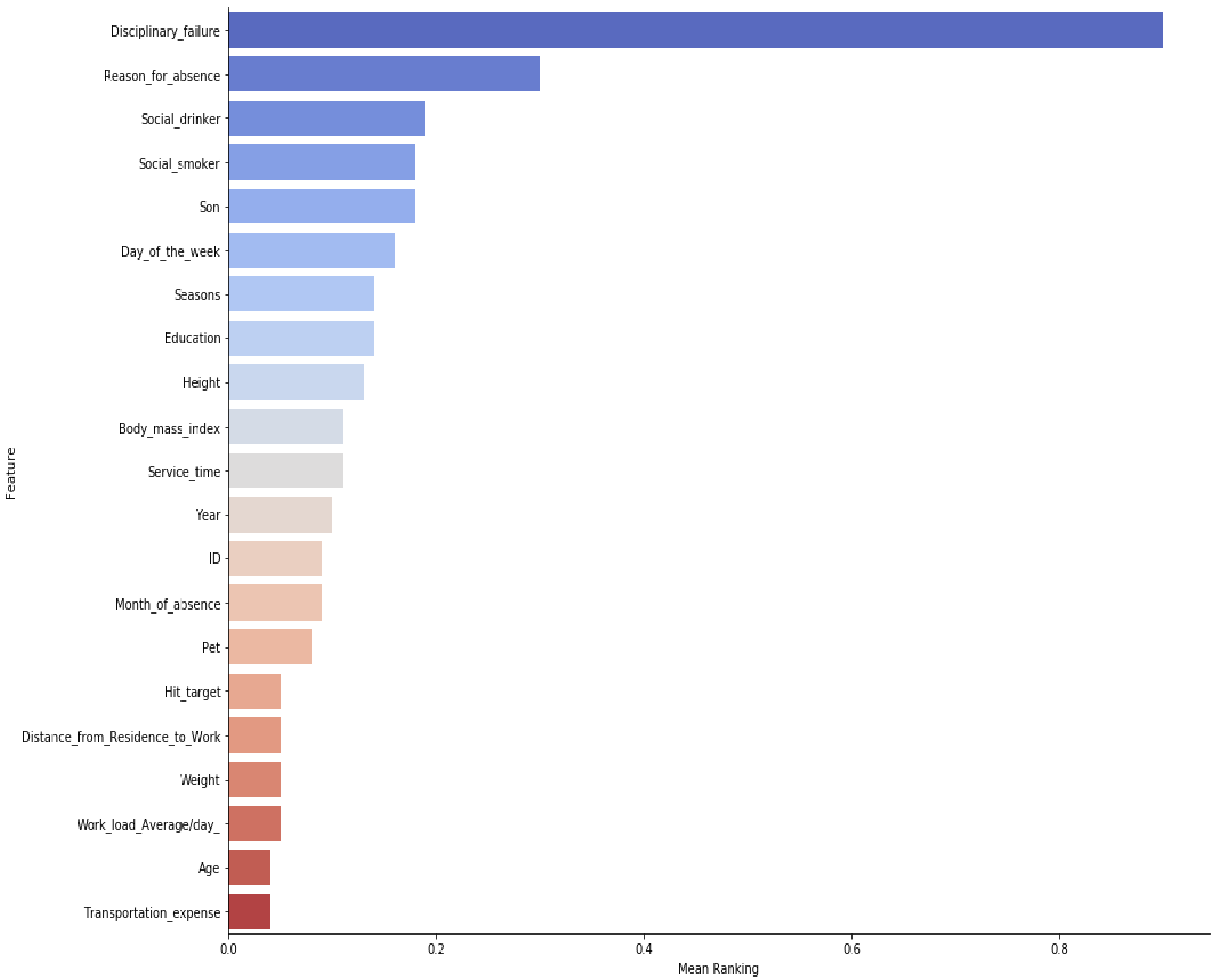


Looking at the above correlation plot looks like no feature is related much with our target variable.

2.6. Feature Importance Ranking:

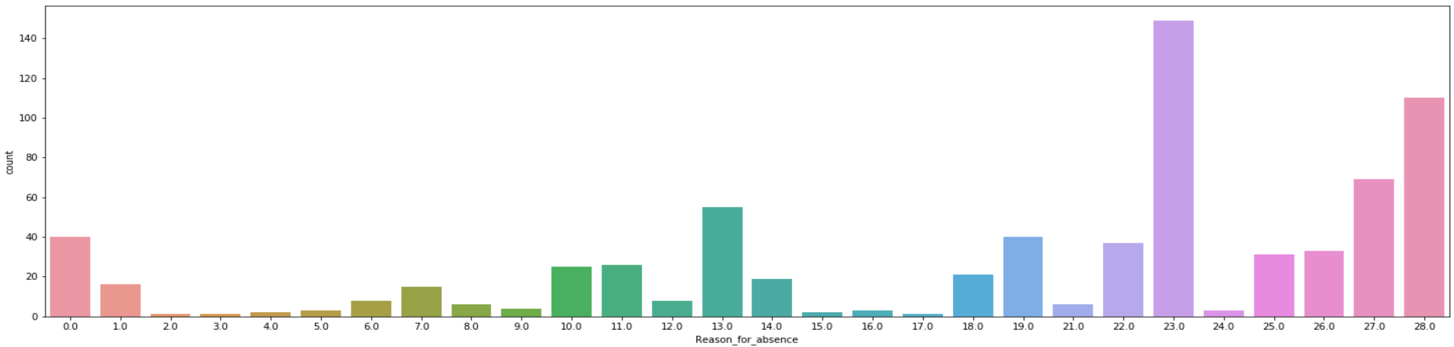
I've used the scores of stability selection via Randomized Lasso method, Recursive feature elimination, linear model feature coefficients(Linear Regression, Lasso and Ridge) and random forest feature selection to come up with the below ranking. The mean of these scores is used to rank the features.

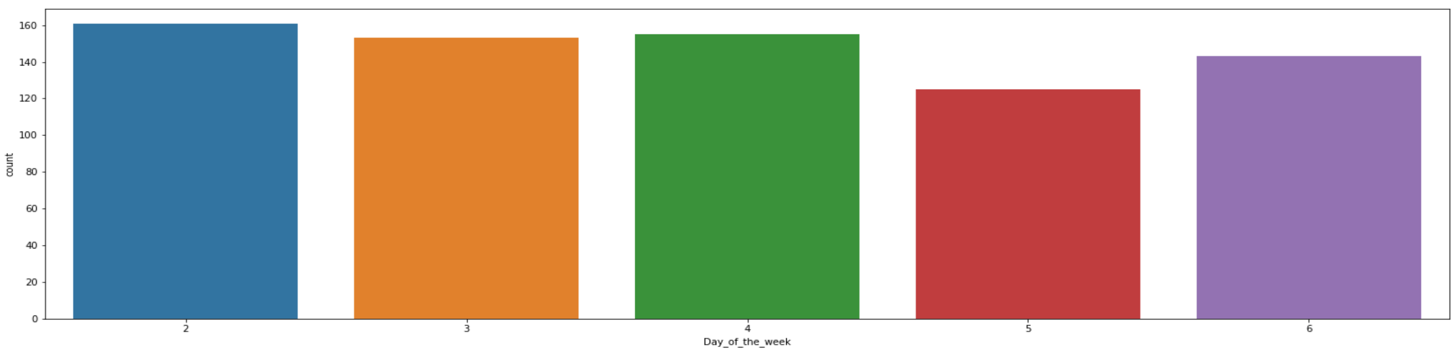
Our top 5 feature are Disciplinary failure, Reason of Absence, Social Drinker, Social Smoker and Son.

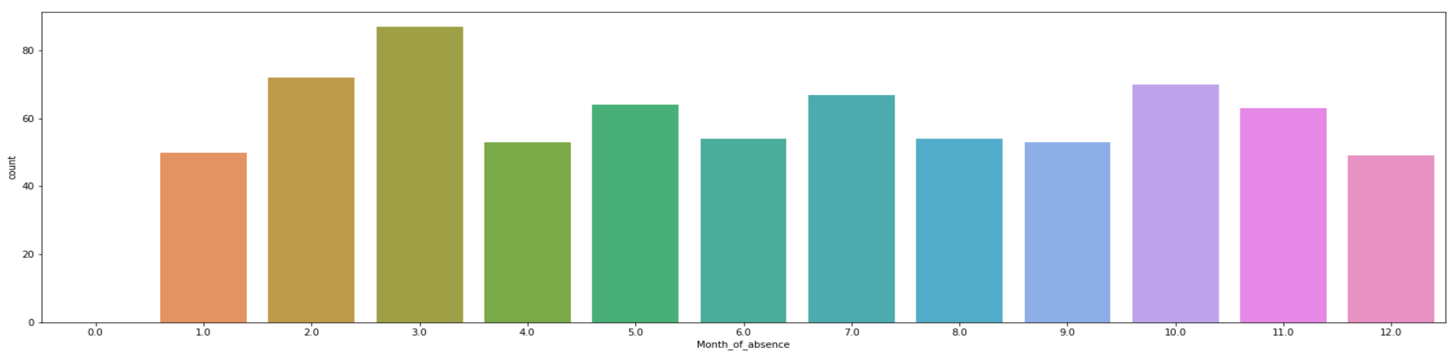


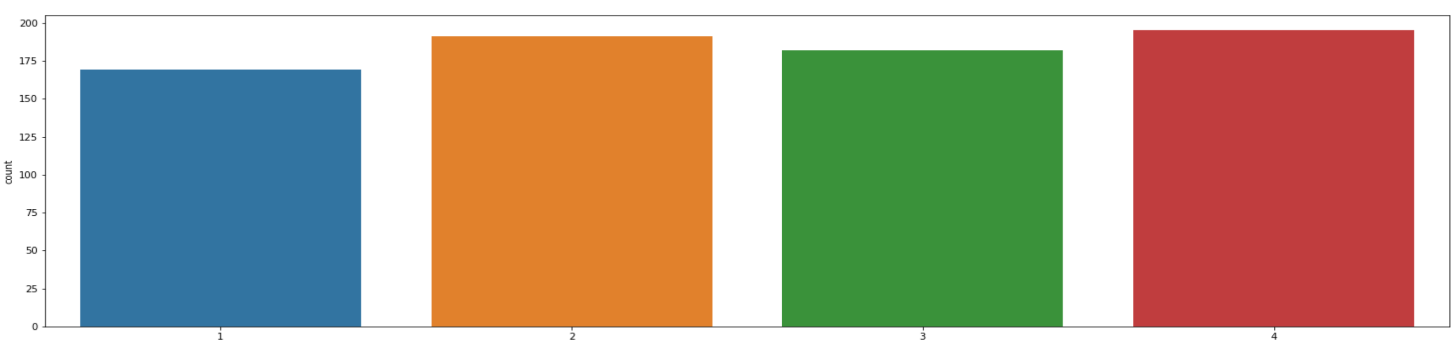
3. Exploring some of the most important variables

3.1. Univariate Analysis based on number of absent hours on a new day









Quick Observation:

The above count plots are based on the employee count. There doesn't seem to be of a much difference in the employee absenteeism hours based on **new day**, month and season. There is no specific day or month or season where the employees are absent. It is distributed uniformly.

But there is something we can see in "reason of absence" plot, the reason of absence on a new day seems to be max for 23 followed by 28, 27, and 13 and so on. (Descending order)

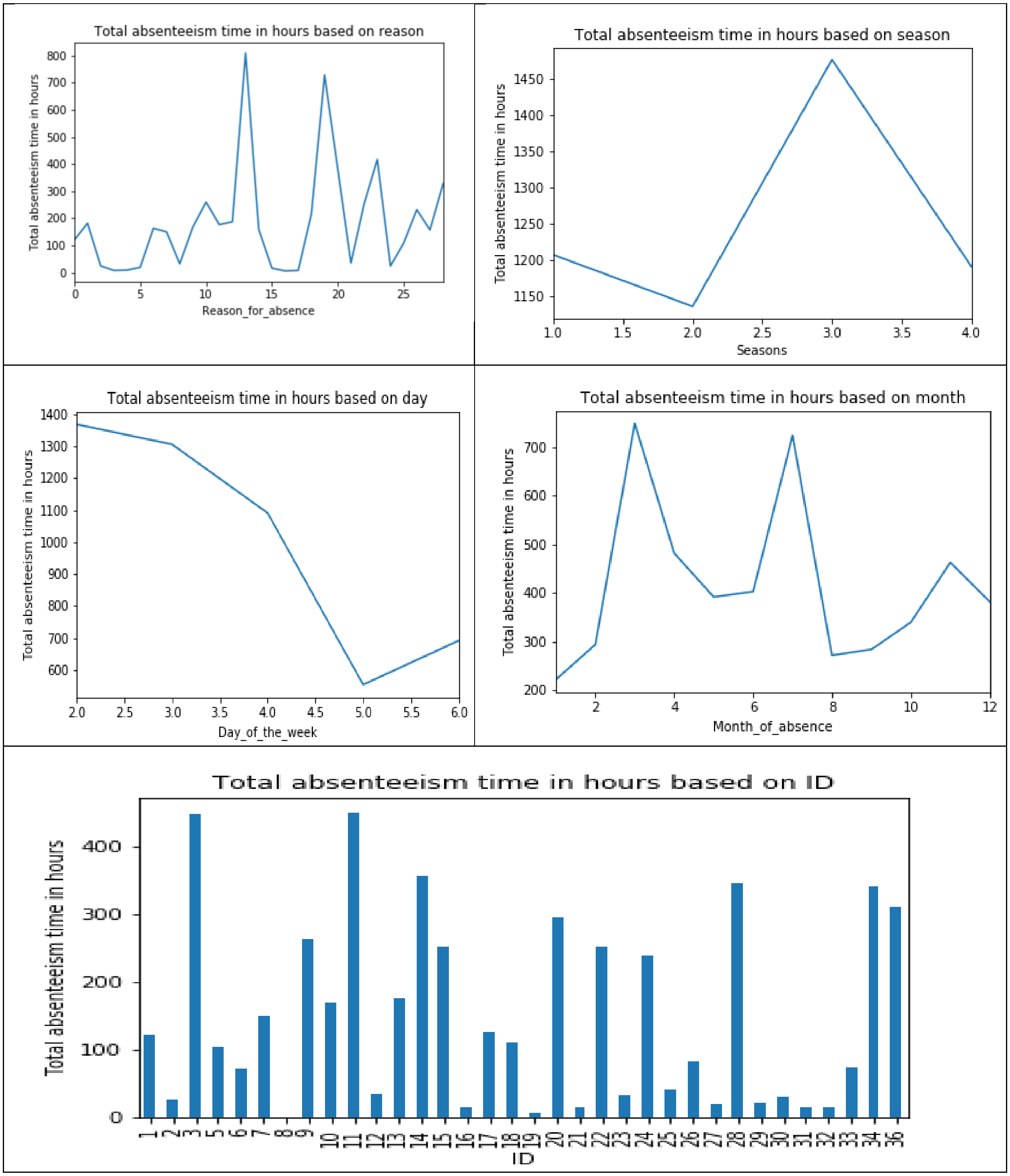
23: Medical Consultation. 28: Dental consultation 27: Physiotherapy 13: Disease of musculoskeletal system and connective tissue.

The point to be noted here is the reason of absence from 1 to 21 are absences attested by the International code of diseases. Reason of absence for 22 to 28 have no attested medical proof.

Out univariate analysis shows that the no attested are the top 3 reason of absence given by the employee on a new day.

3.2. Bivariate Analysis with Absenteeism time

Grouping based on total absenteeism time in hours for all the years



<u>Quick Insight</u>:

Based on Season Plot: We can see that season 3 has the max absentees hours and season 2 lowest but the difference is not much.

Based on Day Plot: Monday has the highest absentees' s hours followed by Tuesday and Wednesday. Thursday and Friday being the lowest. The difference here is quite a lot. **Monday has 1390 hours whereas Thursday and Friday has 550 and 700. Looks like people don't want to go to work on time after a good weekend**.

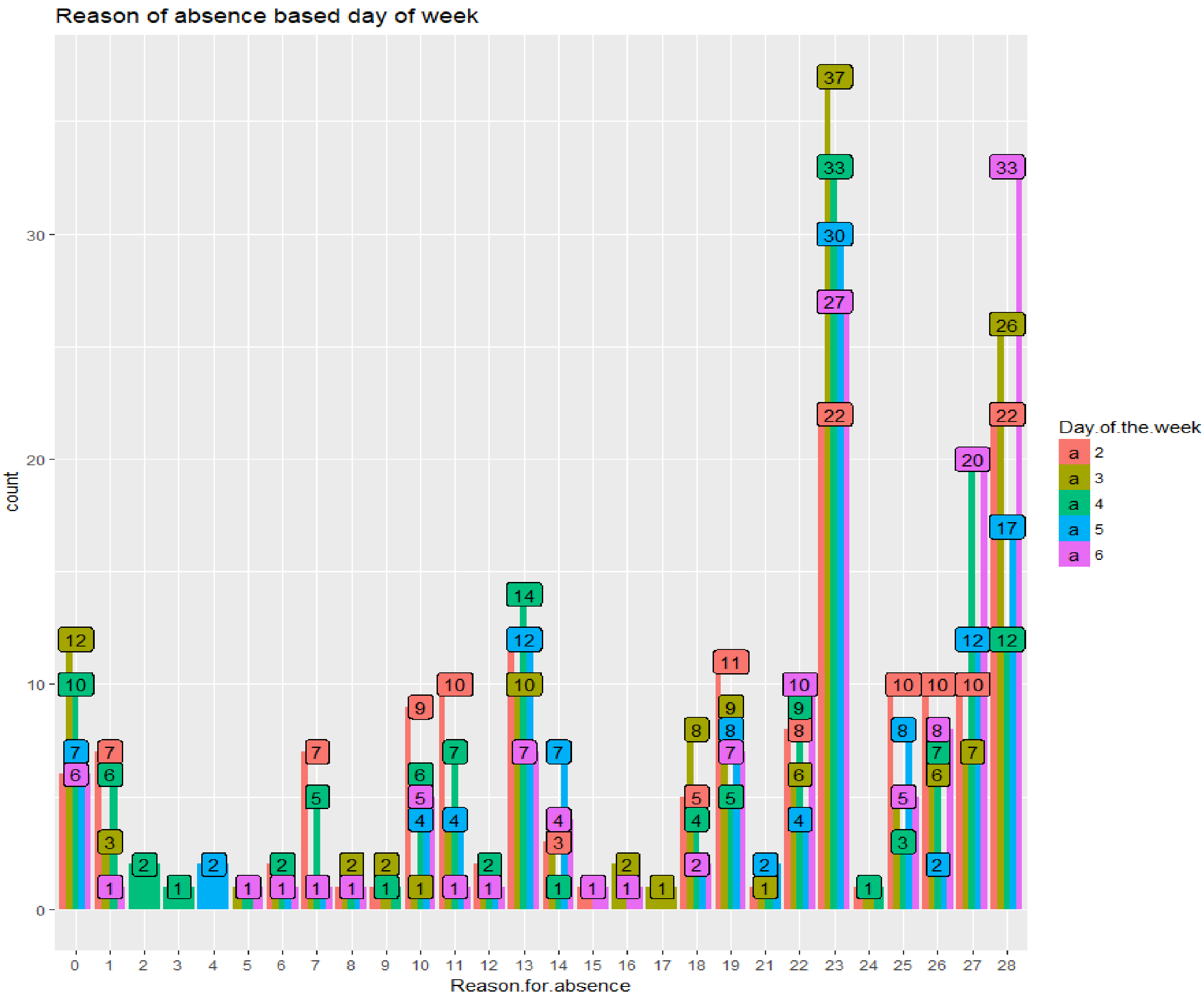
Based on Month of Absence: March and August sees the highest absentee's hours.

Based on ID: There are 11 employees with more than 200 hours of absence in 3 years. The highest seems to be above 400 by employee with ID 3.

3.3. Grouping the variables based on total absenteeism time

```
Disciplinary.failure Absenteeism.time.in.hours
                   0                3198.25398
                   1                  22.91581
aggre.absent.son
Son Absenteeism.time.in.hours
  0                 1202.3555
  1                  875.3949
  2                  839.9455
  3                   87.5853
  4                  215.8886
aggre.absent.drinker
Social.drinker Absenteeism.time.in.hours
             0                  1312.857
             1                  1908.313
aggre.absent.smoker
Social.smoker Absenteeism.time.in.hours
            0                  2943.729
            1                   277.441
```

Count of Reason of absence based on day of week



Observation: There doesn’t seem to be a specific pattern based on the reason and day of week. Monday and Tuesday seems to be favorite days for the employee to come late/ miss office hours

3.4. Inference

3.4.1. Current Trend of Absentee's:

- The maximum people taking the absent hours are from category *23 followed by 28 and 27*. These category are not attested by doctors.

  23: Medical Consultation. 28: Dental consultation 27: Physiotherapy

- These reasons seems to be absurd as these are consultation which people might give as an excuse as they don't have a medical certificate to show.
- Monday, Tuesday and Wednesday have the highest absentees' hours (In Descending Order), which seems obvious after weekend.
- The employee who doesn't have any kids seems to be highest in term of total absentee's hours.
- When there is a disciplinary action, the absentees hours are very low almost negligible. Looks like people become serious once they have been warned for disciplinary issues.

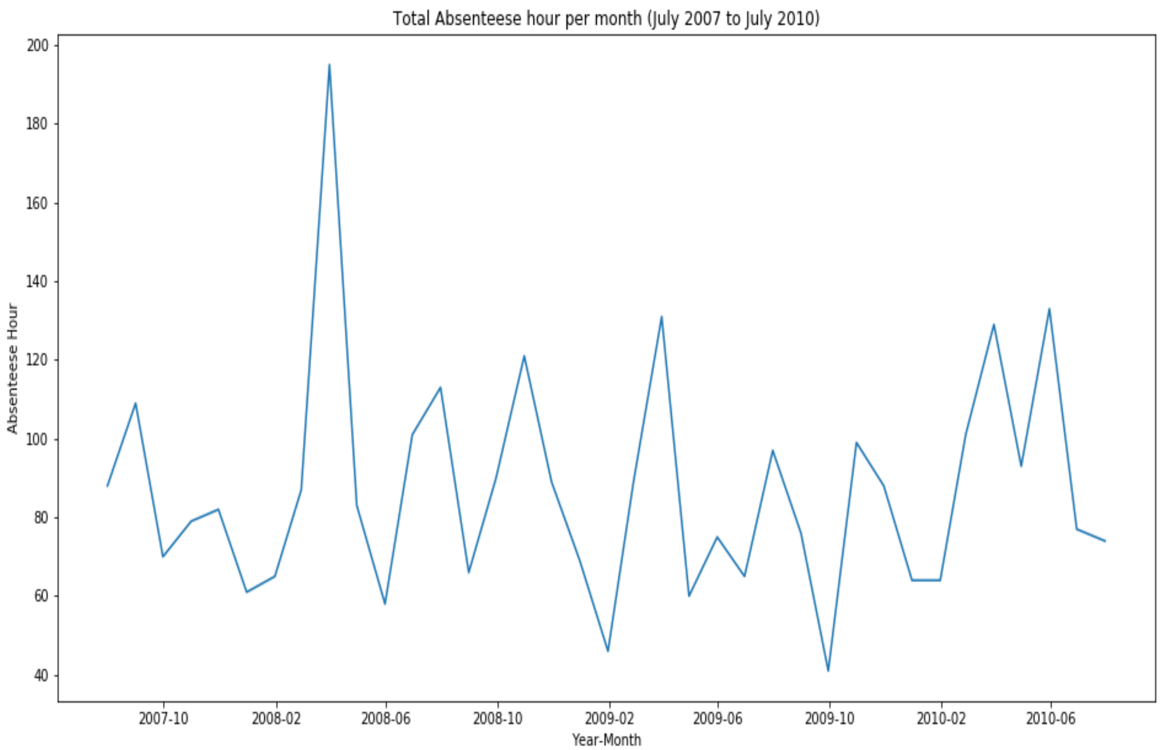
3.4.2.Measures that company can implement:

- As the majority of the reason are *consultation*, company can organize a free health checkup once in 6 months to keep the track of the medical history of employee. This will also keep a good company environment for the employees and an added perk which can help the company loses in the important business hours.
- As people take disciplinary actions seriously, they can implement a rule where a person being absent for more than 15 hours quarterly will be given a warning. After three warnings employer has the right to fire that employee based on professional ethics.
- As Monday have the highest hours, may be company can extend the service hours for Friday and Thursday and decrease a bit on Monday by opening the office 1 or 2 hours later than usual.
- Company can also introduce a policy where in the Top 5 disciplined employees, holding the least Absentee's hours will be rewarded. This Reward can be in form of some Reward points that they can redeem later or some gift voucher. This action will encourage employees to strive for excellence in Discipline.

4. Model Generation

We have grouped the total absentee's hours based on month of the year, as we need to forecast the total absentee's hours for 2011.

The line chart shows the total absentee's hours per month from July 2007 to July 2010.



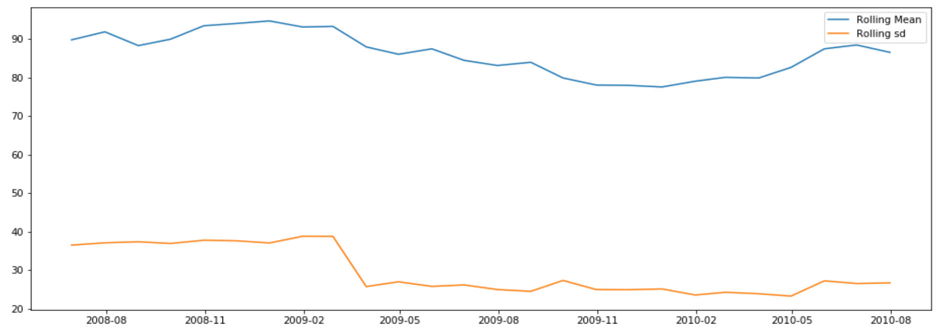
We will build three models for forecasting and will pick the model with better AIC score. We have converted the grouped data based on month of the year into a time series data. We will train the model using this data.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007							91.60233	107.37544	62.04002	75.14784	81.35927	65.14982
2008	60.00000	87.00000	193.65489	83.66371	62.17858	99.46827	113.00000	66.20641	82.20104	122.10755	96.66075	69.00000
2009	43.61238	86.34231	129.38262	57.88039	75.00000	63.00223	96.55055	76.00000	41.00000	99.00000	87.56481	59.90210
2010	71.00000	101.00000	135.14268	95.51763	131.27025	78.98077	75.20516					

4.1. Trend and Seasonality

Rolling Mean

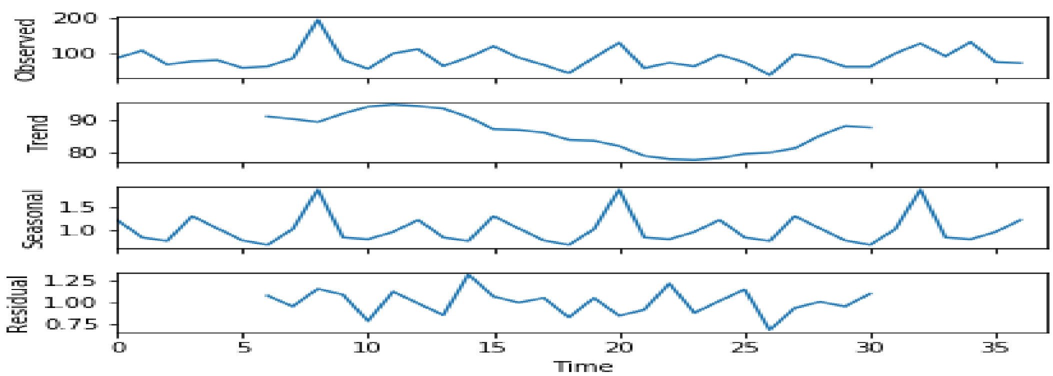
As we can see in the below plot that our means remains constant with not much variation, there seems to be no trend in the data.



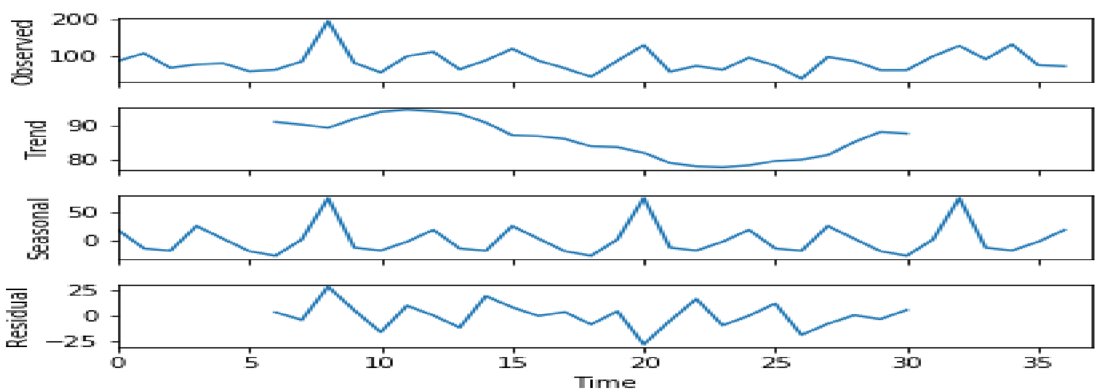
4.1.1. Decomposition

The below plot shows that there is no trend and seasonality also doesn't seems to be there. We will confirm the seasonality in our statistical test

Multiplicative model



Additive Model



As we can see in rolling mean plot, the mean seems to be constant with time which means there is not much trend in the data i.e. time series seems to be stationary.

In the decomposition plot we can see the trend and seasonality, trend doesn't seem to be there. The seasonal plot show 3 high peaks in the model. We will confirm about seasonality in are stationarity test.

4.2. Stationarity

We will perform Dicker Full Test to check the stationarity in the dataset. This is one of the statistical tests for checking stationarity. Here the null hypothesis is that the TS is non-stationary. The test results comprise of a Test Statistic and some Critical Values for difference confidence levels. If the 'Test Statistic' is less than the 'Critical Value', we can reject the null hypothesis and say that the series is stationary

```
Results of Dickey-Fuller Test:
Test Statistic                -5.702235e+00
p-value                        7.629028e-07
#Lags Used                     1.000000e+00
Number of Observations Used    3.500000e+01
Critical Value (1%)           -3.632743e+00
Critical Value (5%)           -2.948510e+00
Critical Value (10%)          -2.613017e+00
dtype: float64
```

As our p-value is less than 0.05 we can reject our null hypothesis and accept the alternate which says the time series is stationary i.e there is not much trend and seasonality in the data set.

4.3. Linear Regression with Trend/TSLM

NOTE: Model building and forecasting in all the model is done using complete data. Accuracy has been measured by dividing into train and test

To forecast using linear regression, I've used different linear regression models in R and Python.

Python: I've used simple linear regression with trend to forecast the values:

Approach:

- Converted the months into a sequence of number till 37(As we have 37 months data to forecast.
- Calculated the trend series where the value at the current time step is calculated as the difference between the original observation and the observation at the previous time step.
- Forecasting the trend value till 2011
- Then I've build the model using trend and month as predictor and absentees hours as target.
- Forecasted the absentees hours values using the forecasted trend and month variables for the year 2011

R: I've used time series linear regression with trend to build a model.

Below is the summary using tslm in R.

```
Call:
tslm(formula = ts_complete_data ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.544 -14.592  -0.774  10.793  43.027 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  60.1675    13.8987   4.329 0.000229 ***
trend        -0.1033     0.3414  -0.303 0.764771    
season2      33.3466    17.3866   1.918 0.067091 .  
season3      94.7293    17.3967   5.445 1.35e-05 ***
season4      21.1265    17.4134   1.213 0.236848    
season5      31.6922    17.4368   1.818 0.081641 .  
season6      22.7963    17.4669   1.305 0.204228    
season7      35.8854    16.2605   2.207 0.037132 *  
season8      24.4732    17.4669   1.401 0.173977    
season9       3.1296    17.4368   0.179 0.859068    
season10     40.2377    17.4134   2.311 0.029754 *  
season11     30.1175    17.3967   1.731 0.096250 .  
season12      6.3765    17.3866   0.367 0.717020    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.29 on 24 degrees of freedom
Multiple R-squared:  0.6472,	Adjusted R-squared:  0.4708 
F-statistic: 3.669 on 12 and 24 DF,  p-value: 0.003253
```

The estimated trend value is in negative, which means there is no trend. We have observed that in our stationarity test as well. As you can see the adjusted r-square value, we can explain 47% data using time series linear regression.

In python score comes around 54, as we have forecasted trend and then forecasted the absentee hours.

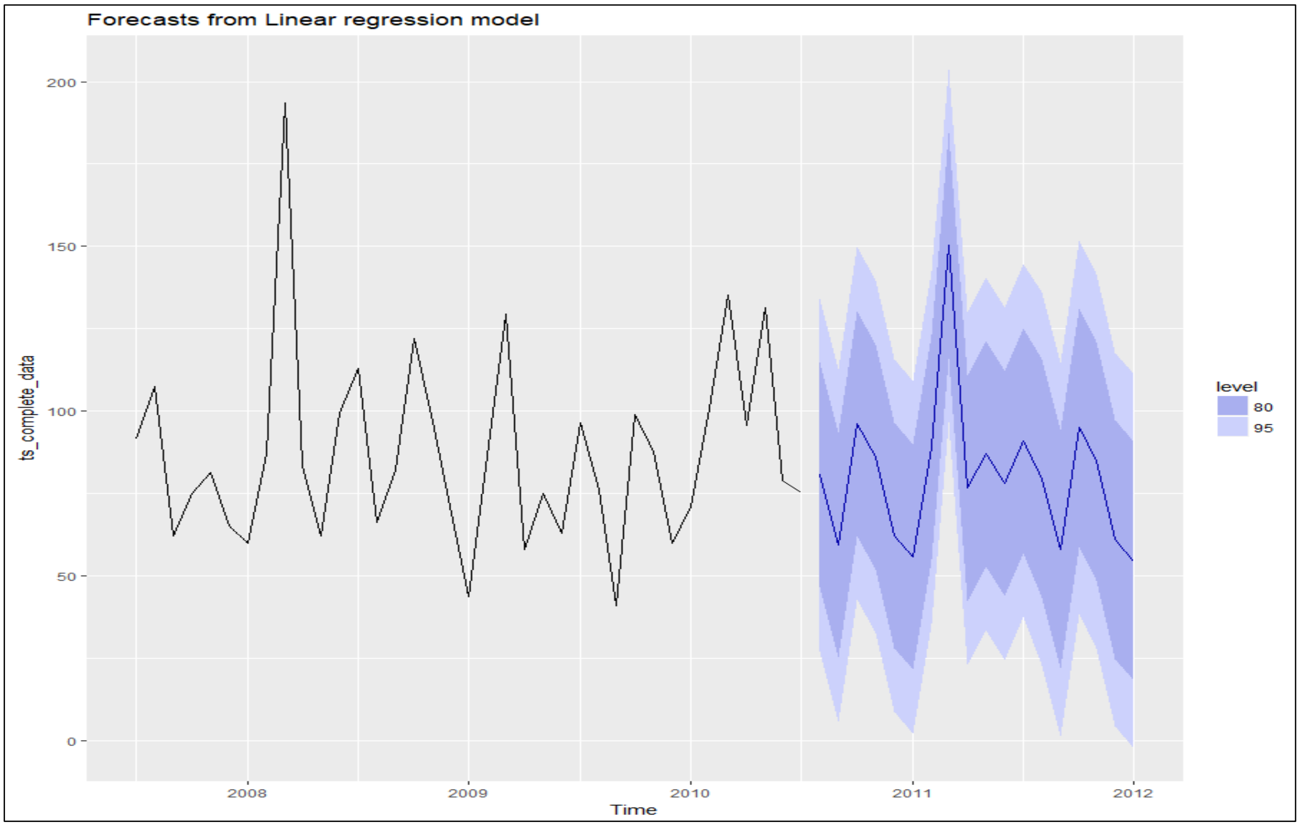
The Akaike Information Criteria (AIC) is 343.2952. We will use this information to compare the robustne ss of the models.

Python:

```
Summary:
Estimated intercept coefficient 87.9454542938
Number of coefficients 2
Coefficients Values
                                  Month          Trend
                               [0.01493117     0.49195177]
AIC of Model:  265.86
```

4.3.1 Forecasting in R using TSLM

Below is the forecast line plot for 2011. Forecasting in all the model is done by training the model on the **complete time series data**.



Forecasted values

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Aug 2010	80.71395	46.56432	114.86357	27.231295	134.1966
Sep 2010	59.26701	25.11739	93.41663	5.784360	112.7497
Oct 2010	96.27179	62.12217	130.42141	42.789140	149.7544
Nov 2010	86.04827	51.89865	120.19789	32.565620	139.5309
Dec 2010	62.20397	28.05434	96.35359	8.721315	115.6866
Jan 2011	55.72412	21.57450	89.87374	2.241468	109.2068
Feb 2011	88.96743	54.81780	123.11705	35.484778	142.4501
Mar 2011	150.24673	116.09710	184.39635	96.764075	203.7294
Apr 2011	76.54057	42.39095	110.69020	23.057921	130.0232
May 2011	87.00294	52.85331	121.15256	33.520285	140.4856
Jun 2011	78.00375	43.85412	112.15337	24.521096	131.4864
Jul 2011	90.98950	56.83988	125.13912	37.506850	144.4722
Aug 2011	79.47394	43.25280	115.69509	22.747024	136.2009
Sep 2011	58.02701	21.80586	94.24815	1.300090	114.7539
Oct 2011	95.03179	58.81064	131.25293	38.304869	151.7587
Nov 2011	84.80827	48.58712	121.02941	28.081350	141.5352
Dec 2011	60.96396	24.74282	97.18511	4.237045	117.6909
Jan 2012	54.48412	18.26297	90.70526	-2.242802	111.2110

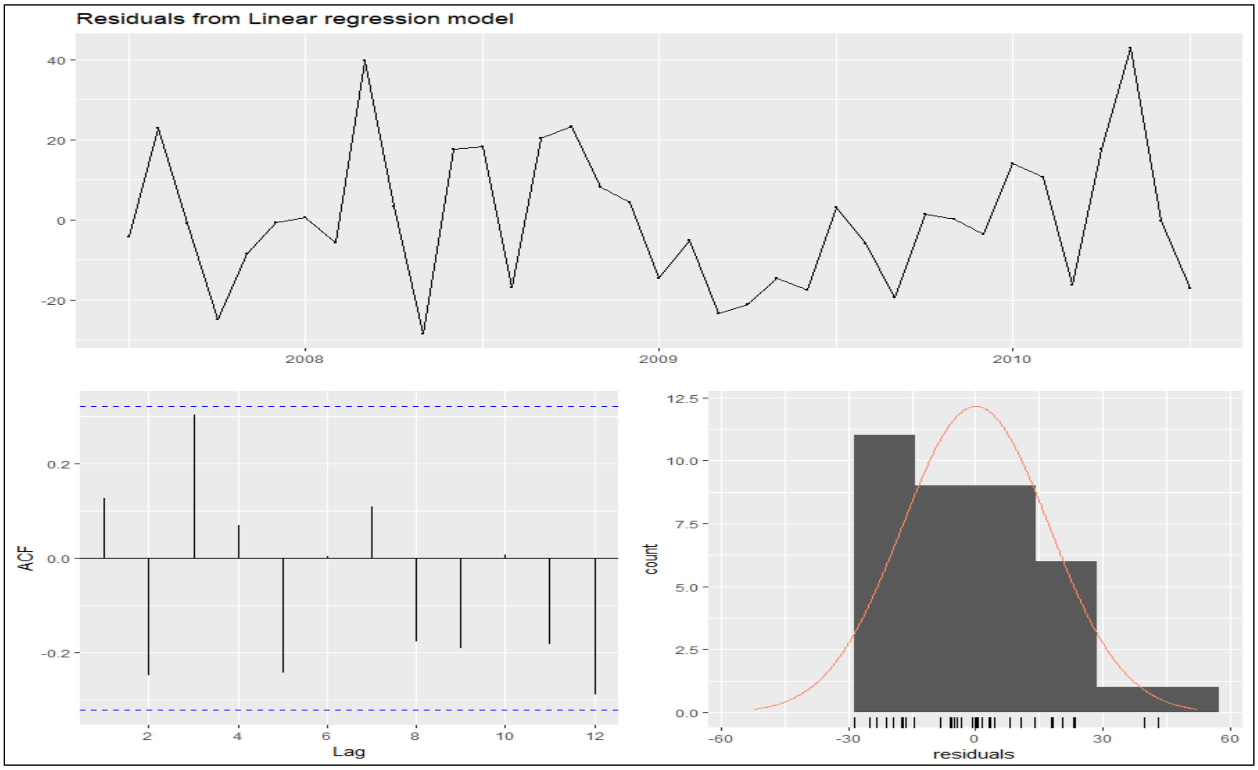
4.3.2. Forecasting in Python using Linear Regression with manually forecasted Trend

Forecasted Values:

Jan 2011	87.68970383
Feb 2011	87.67519849
March 2011	87.66069316
April 2011	85.88377786
May 2011	85.82916699
June 2011	85.77455612
July 2011	85.71994525
August 2011	85.66533438
September 2011	85.61072352
October 2011	85.55611265
November 2011	85.50150178
December 2011	85.44689091

4.3.3. Residuals

We will test the residual as it's an important measure for the performance of the measure. There should be no pattern in the residuals. As we can see no lag is above the threshold level and the residual seems to be unrelated to each other which is what we wanted.



4.4. Auto Regressive Integrated Moving Average (ARIMA)

ARIMA model there are 3 parameters that are used to help model the major aspects of a times series: seasonality, trend, and noise. These parameters are labeled **p, d,** and **q.**

AR: Autoregressive part: Summation of lags, p

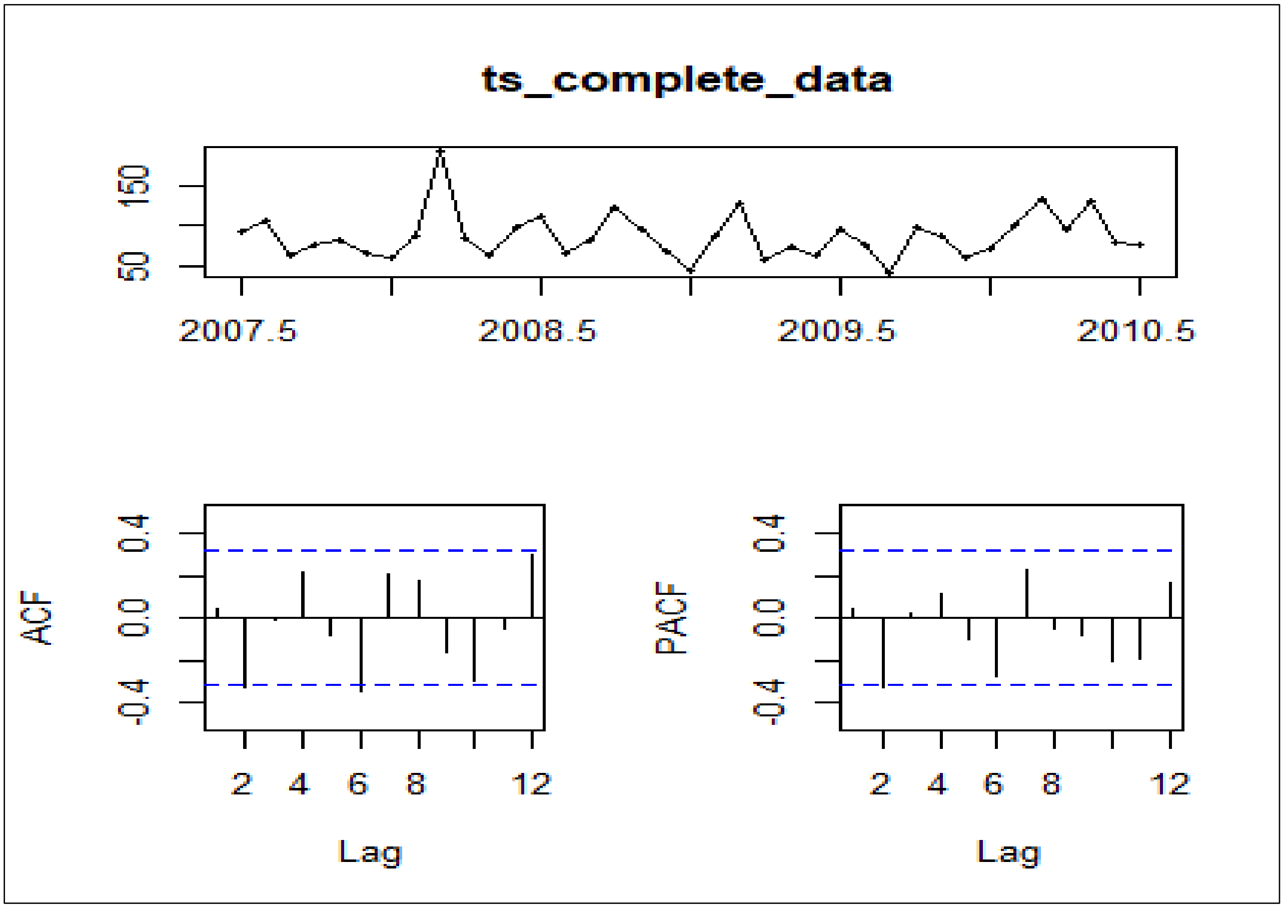
I: Integration, degree of differencing: d

MA: Moving Average: Summation of forecasting errors, q

4.4.1. ACF And PACF Plot

ACF plot tells about q: Moving average part

PACF plot tells about p; auto regression part



The ACF Plot shows two significant lags and PACF Plot show only one significant lag. We can start with ARIMA (1, 0, 1) to start with and then change the parameters and select the one with least AIC score.

We are using auto.arima which will automatically take the best model by comparing the different AIC values at different level of p, d and q.

```
Summary:

 ARIMA(2,0,2)(1,1,1)[12] with drift         : Inf
 ARIMA(0,0,0)(0,1,0)[12] with drift         : 243.3661
 ARIMA(1,0,0)(1,1,0)[12] with drift         : 245.8594
 ARIMA(0,0,1)(0,1,1)[12] with drift         : 243.8133
 ARIMA(0,0,0)(0,1,0)[12]                    : 241.0087
 ARIMA(0,0,0)(1,1,0)[12] with drift         : 243.6941
 ARIMA(0,0,0)(0,1,1)[12] with drift         : 243.4169
 ARIMA(0,0,0)(1,1,1)[12] with drift         : Inf
 ARIMA(1,0,0)(0,1,0)[12] with drift         : 244.2708
 ARIMA(0,0,1)(0,1,0)[12] with drift         : 243.3077
 ARIMA(1,0,1)(0,1,0)[12] with drift         : 243.4347

 Best model: ARIMA(0,0,0)(0,1,0)

Series: ts_complete_data
ARIMA(0,0,0)(0,1,0)[12]

sigma^2 estimated as 825.1:  log likelihood=-119.42
AIC=240.83   AICc=241.01   BIC=242.05
```

ARIMA (0,0,0)(0,1,0))[12] , it's a special case and is known as Seasonal Random Walk. (0, 0, 0) is the non-seasonal part of the model and (0, 1, 0) is the seasonal part of the model. (0, 1, 0) show a seasonal difference.

A seasonal random walk model is a special case of an ARIMA model in which there is *one* order of seasonal differencing, a *constant* term, and *no* other parameters--i.e., an "ARIMA(0,0,0)x(0,1,0) model with constant.

As we have a monthly data, whose seasonal period is 12, the seasonal difference at period t is Y_tY_{t-12}. Applying the mean model to this series yields the equation:

$$\hat{\mathbf{Y}}_t - \mathbf{Y}_{t-12} = \boldsymbol{\mu}$$

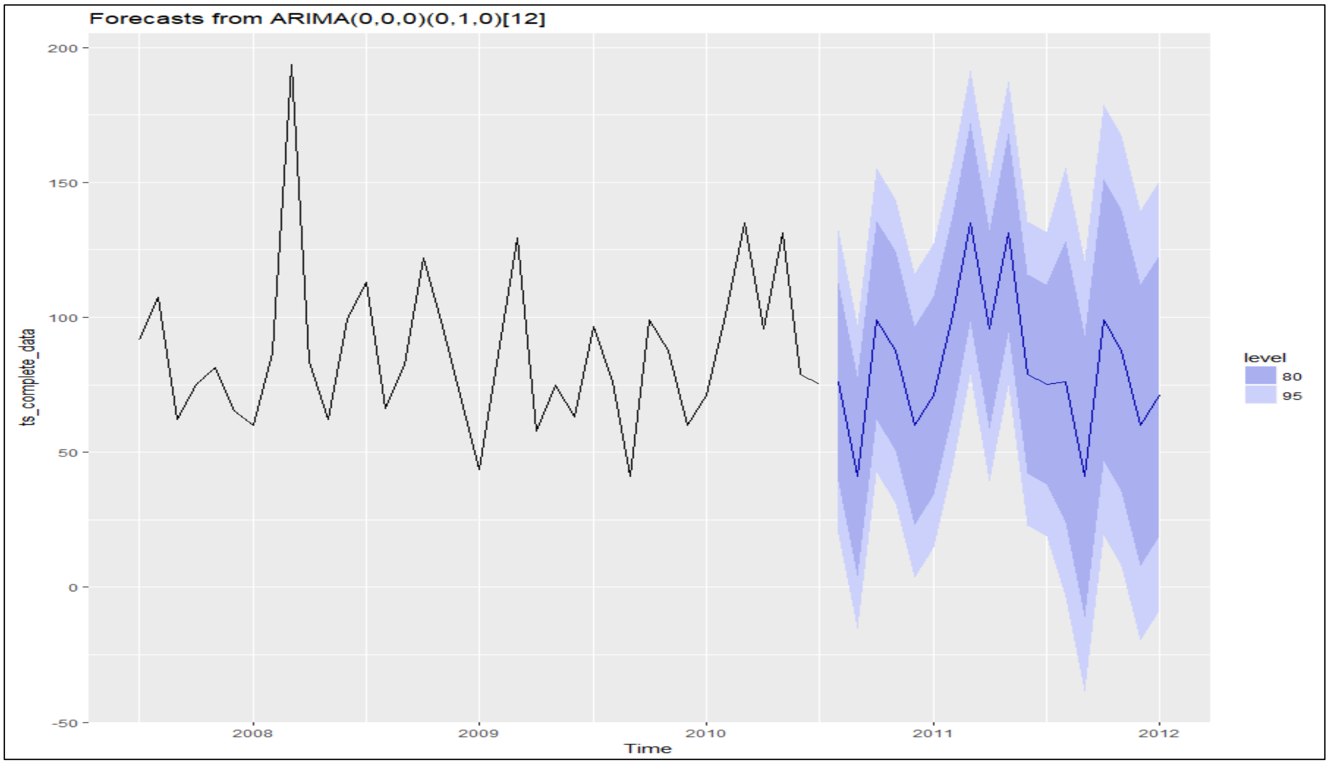
This forecasting model will be called the *seasonal random walk* model, because it assumes that each season's values form an independent random walk. For example, the model assumes that September's value this year is a random step away from September's value last year, October's value this year is a random step away from October's value last year, etc., and the mean value of every step is equal to μ:

$\hat{Y}_{Sep2010}$ = YSep2009 + μ

$\hat{Y}_{Oct2010}$ = YOct2009 + μ

The AIC of ARIMA model is better than the Time series model.

4.4.2. Forecasting

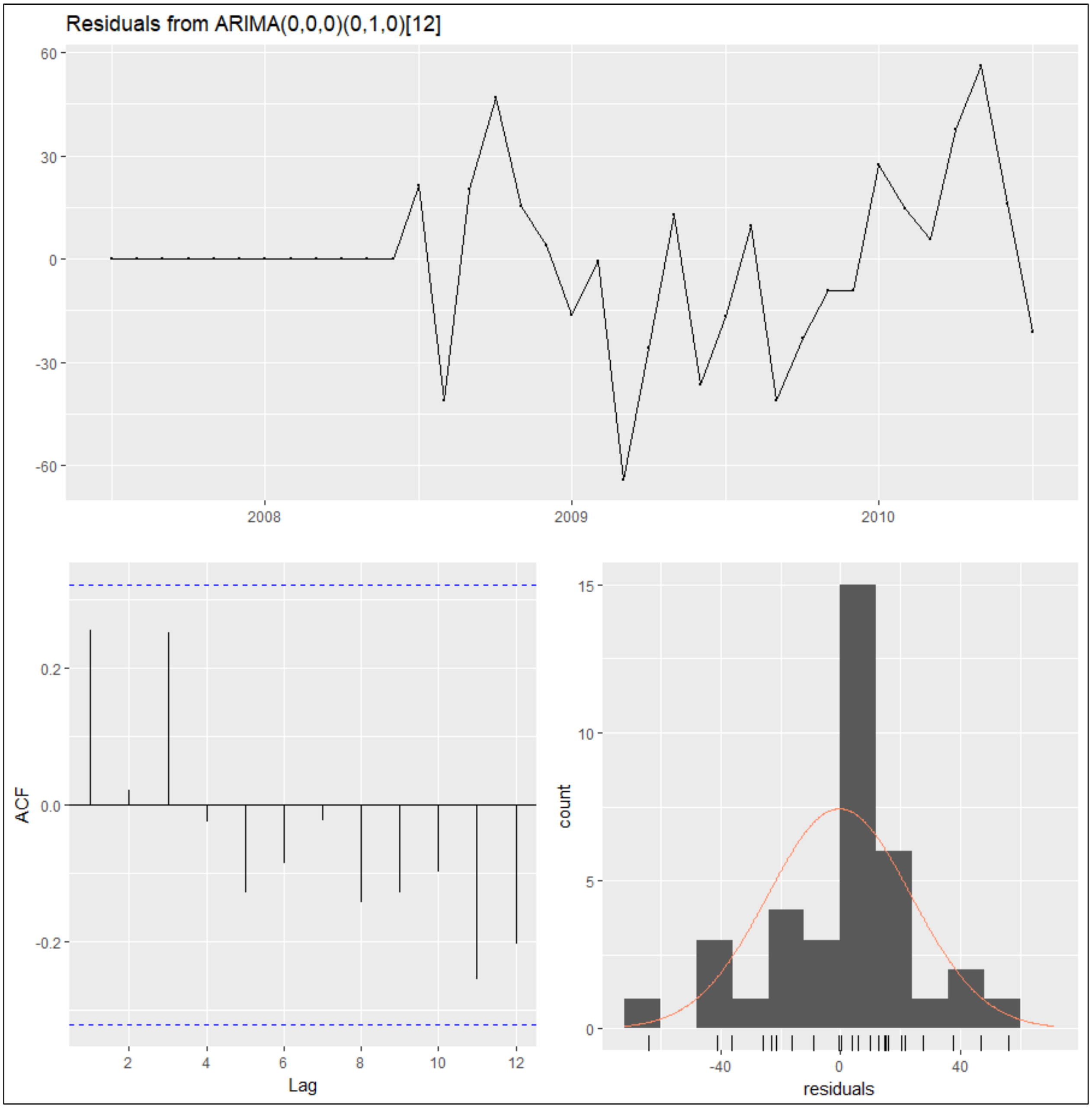


Forecasted values

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Aug 2010	76.00000	39.187742	112.81226	19.700509	132.29949
Sep 2010	41.00000	4.187742	77.81226	-15.299491	97.29949
Oct 2010	99.00000	62.187742	135.81226	42.700509	155.29949
Nov 2010	87.56481	50.752550	124.37707	31.265317	143.86430
Dec 2010	59.90210	23.089839	96.71436	3.602606	116.20159
Jan 2011	71.00000	34.187742	107.81226	14.700509	127.29949
Feb 2011	101.00000	64.187742	137.81226	44.700509	157.29949
Mar 2011	135.14268	98.330421	171.95494	78.843188	191.44217
Apr 2011	95.51763	58.705374	132.32989	39.218141	151.81712
May 2011	131.27025	94.457992	168.08251	74.970760	187.56974
Jun 2011	78.98077	42.168509	115.79303	22.681276	135.28026
Jul 2011	75.20516	38.392900	112.01742	18.905668	131.50465
Aug 2011	76.00000	23.939605	128.06039	-3.619503	155.61950
Sep 2011	41.00000	-11.060395	93.06039	-38.619503	120.61950
Oct 2011	99.00000	46.939605	151.06039	19.380497	178.61950
Nov 2011	87.56481	35.504413	139.62520	7.945305	167.18431
Dec 2011	59.90210	7.841702	111.96249	-19.717406	139.52160
Jan 2012	71.00000	18.939605	123.06039	-8.619503	150.61950

4.4.3. Residuals Check

There is no significant lag and the residuals seems to be normally distributed, there is no pattern in the residuals which we wanted.



4.5. ETS (Exponential Smoothing)

As we saw a seasonal parameter ‘D’ in arima model, let’s forecast using ETS model as well as it works best with seasonal parameter even though there is no seasonality in the data. Let’s check what ETS gives.

This model describes the time series with three parameters.

*Error - additive, multiplicative(x>0)

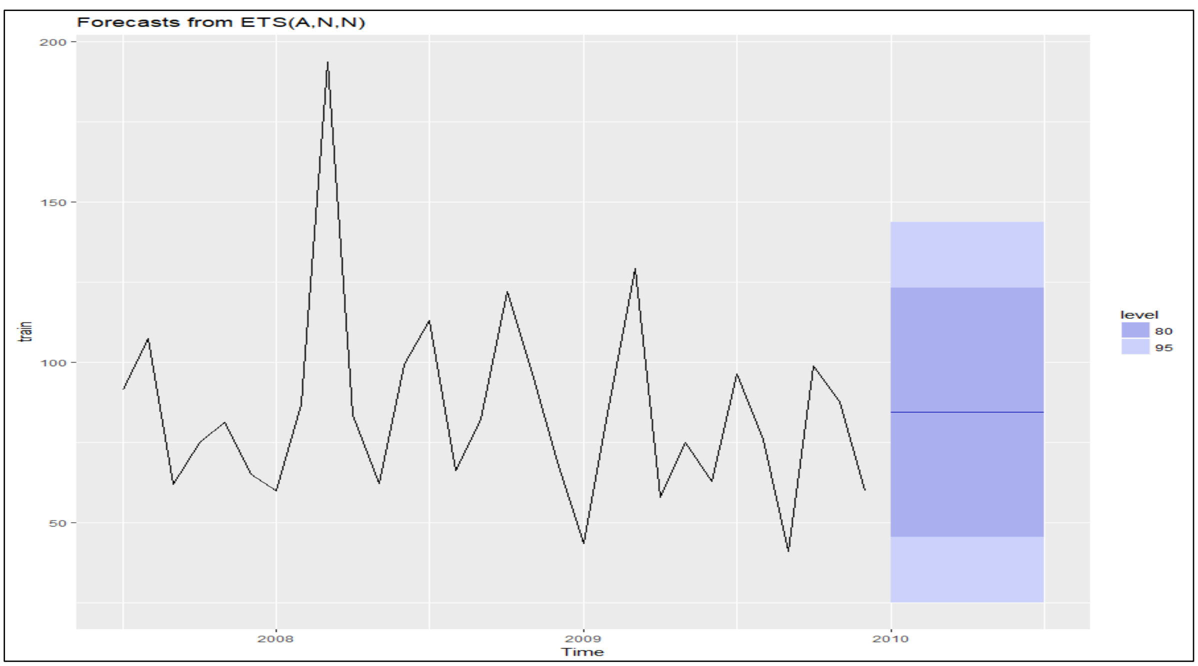
*Trend - non present, additive, multiplicative.

*Seasonality - non present, additive, multiplicative.

```
ETS(A,N,N)

Call:
 ets(y = train)

  Smoothing parameters:
    alpha = 1e-04

  Initial states:
    l = 84.4455

  sigma:  30.2586

     AIC     AICc      BIC
310.5530 311.4761 314.7566
```

ETS gives (A, N, N) model which means there is no trend and no seasonality in the data. A means we have an additive error.

4.5.1. Forecasting

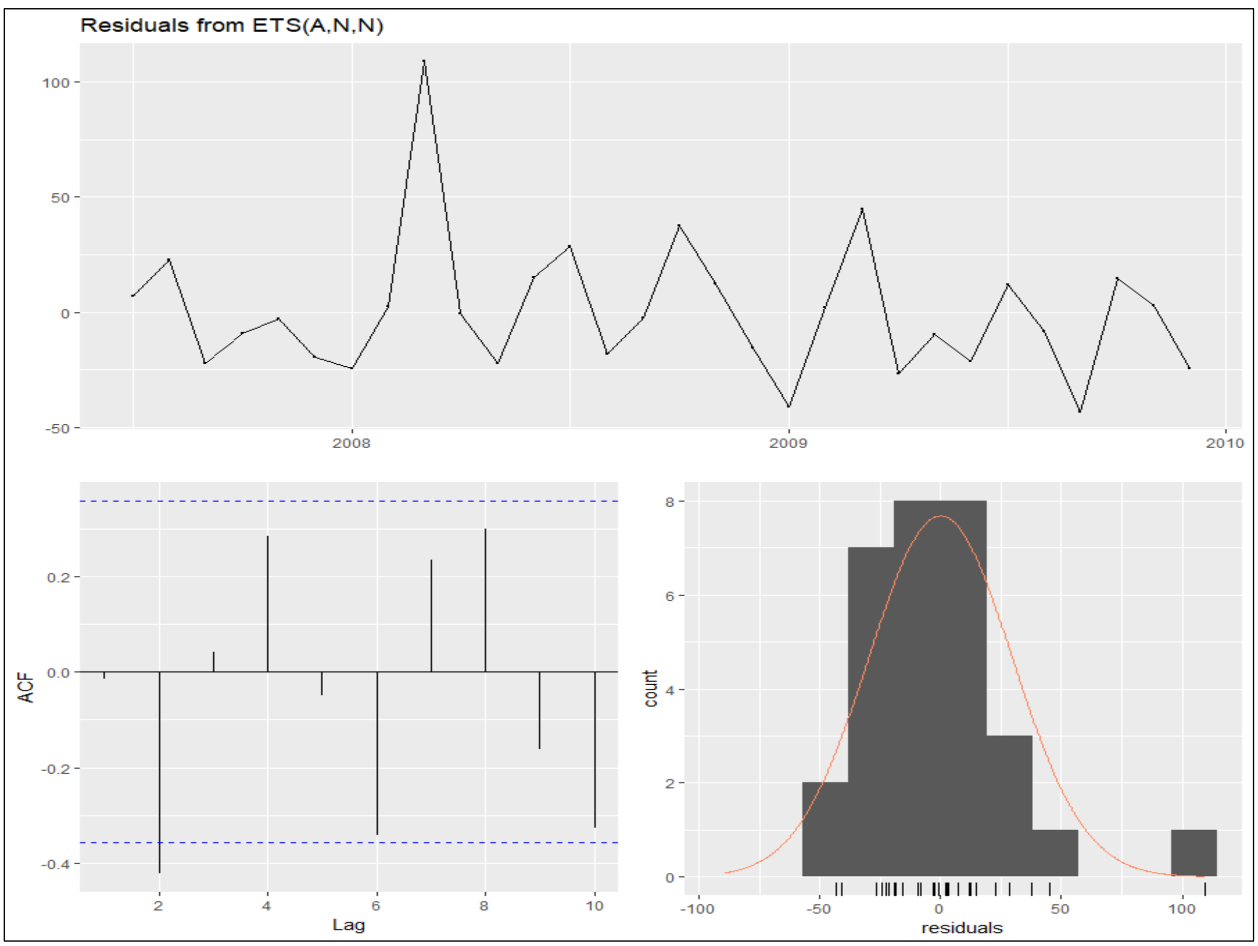


Points Forecasted

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Aug 2010	87.05682	49.01538	125.0983	28.87745	145.2362
Sep 2010	87.05682	49.01538	125.0983	28.87745	145.2362
Oct 2010	87.05682	49.01538	125.0983	28.87745	145.2362
Nov 2010	87.05682	49.01538	125.0983	28.87745	145.2362
Dec 2010	87.05682	49.01538	125.0983	28.87745	145.2362
Jan 2011	87.05682	49.01538	125.0983	28.87745	145.2362
Feb 2011	87.05682	49.01538	125.0983	28.87745	145.2362
Mar 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Apr 2011	87.05682	49.01537	125.0983	28.87745	145.2362
May 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Jun 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Jul 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Aug 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Sep 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Oct 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Nov 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Dec 2011	87.05682	49.01537	125.0983	28.87745	145.2362
Jan 2012	87.05682	49.01537	125.0983	28.87745	145.2362

4.5.2. Residuals

There is one significant lag in the ACF which makes the model weak as compared to the other once we have used.



5. Accuracy of three models

We divided the time series on train and test 80:20, keeping recent month's i.e. from Jan 2010 to July 2010 into test and training the model on the previous data. We are using RMSE and MAPE to determine the model performance.

```
> accuracy(forecast_tslm, test)
                        ME     RMSE      MAE       MPE     MAPE      MASE        ACF1 Theil's U
Training set -9.475855e-16 14.20167 11.53323 -2.571463 13.85181 0.5014775  0.06472768        NA
Test set      1.881355e+01 33.98896 27.67494 18.475412 27.29737 1.2033371 -0.02280608   1.07769
> #Accuracy
> accuracy(arimafore, test)
                    ME     RMSE      MAE      MPE     MAPE      MASE      ACF1 Theil's U
Training set -5.077805 21.86004 13.83472 -9.48517 18.45262 0.6015489 0.1025744        NA
Test set     19.478002 29.98083 25.57669 18.78091 26.89030 1.1121026 0.1268127 0.9062453
> #Accuracy
> accuracy(ets_forecast, test)
                      ME     RMSE      MAE        MPE     MAPE      MASE        ACF1 Theil's U
Training set -0.01455468 29.23262 20.80664 -10.545035 26.54799 0.9046957 -0.01526695        NA
Test set     13.85686556 27.91680 21.89988   9.003352 19.90141 0.9522309 -0.08741275 0.9146704
```

Linear Regression with trends in python with the manual approach of predicting the trend and forecasting has RMSE 18.5194 and MAPE 19.2891 with AIC value 265.0752

It is not correct to select the random data for checking the accuracy in time series model. **There was a drastic increase in absentee's hours in the month of May and a bit in March and April which we have taken into test set. We trained the data which have not seen this pattern and was unable to predict such as high increase which resulted in less RMSE and MAPE value then the Linear Regression with trend using python.**

Seeing the patterns and doing some exhaustive testing on data, ARIMA models looks good with the best AIC score among else. Results of exhaustive testing using different set of train and test with the three model are placed in the Results folder.

The best approach if we want to give estimates to the client will be to train the data on the complete set and then predict for next month as there is no seasonality and trend in employees absentees behaviors.

6. References

- Forecasting principles and practices by Rob J Hyndman
- Blog article by Ando Sabaas on feature selection