EMPLOYEE ABSENTEEISM

CONTENTS

Chapter 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 loss every month can we project in 2011 if same trend of absenteeism continues?

1.2 Data

Our task is to build a time series forecasting model that will forecast the employee absenteeism rate for the year 2011 by using the data for the years 2007, 2008, 2009 and 2010 assuming that the same trend will be followed in the upcoming year. We have been provided a dataset (with 740 rows and 21 columns). Overview of the data is as follows:

Table 1.1 Employee Absenteeism Dataset (Columns 1 to 9)

ID	Reason. for. absence	Month. of. absence	Day.of. the. Week	Seasons	Transportation. expense	Distance. from. residence. to. work	Service. time	Age
11	26	7	3	1	289	36	13	33
36	0	7	3	1	118	13	18	50
3	23	7	4	1	179	51	18	38
7	7	7	5	1	279	5	14	39
11	23	7	5	1	289	36	13	33

Table 1.1 Employee Absenteeism Dataset (Columns 9 to 18)

Work. load. average. per.day	Hit. target	Disciplinary. Failure	Education	Son	Social. drinker	Social. smoker	Pet	Weight
2,39,554	97	0	1	2	1	0	1	90
2,39,554	97	1	1	1	1	0	0	98
2,39,554	97	0	1	0	1	0	0	89
2,39,554	97	0	1	2	1	1	0	68
2,39,554	97	0	1	2	1	0	1	90

Table 1.1 Employee Absenteeism Dataset (Columns 19 to 21)

Height	Body.mass.index	Absenteeism.time.in.hours
172	30	4
178	31	0
170	31	2
168	24	4
172	30	2

The dataset has 20 independent variables that help to determine the value of the response variable “Absenteeism time in hours”.

1.3 Attribute Information

a) ID: Individual identification

b) Reason for absence: Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (1 to 21). While the absences not attested by ICD follow into 7 categories (22 to 28).

c) Month of absence

d) Day of the week: (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

e) Seasons (summer (1), autumn (2), winter (3), spring (4))

f) Transportation expense

g) Distance from Residence to Work: (in kilometres)

h) Service time

i) Age

j) Work load Average/day

k) Hit target

l) Disciplinary failure: (yes=1; no=0)

m) Education: (high school (1), graduate (2), postgraduate (3), master and doctor (4))

n) Son (number of children)

o) Social drinker: (yes=1; no=0)

p) Social smoker: (yes=1; no=0)

q) Pet: (number of pet)

r) Weight

s) Height

t) Body mass index

u) Absenteeism time in hours: Target variable whose value for the future year has to be estimated.

1.4 Data Exploration

Table 1.2 shows the data-type of all the variables in the dataset

Table 1.2 Variable Data types

VARIABLE NAME	DATA TYPE	CATEGORICAL/CONTINUOUS
ID	Object	Categorical
Reason for absence	Object	Categorical
Month of absence	Object	Categorical
Day of the week	Object	Categorical
Seasons	Object	Categorical
Transportation expense	Float	Continuous
Distance from residence to work	Float	Continuous
Service time	Float	Continuous
Age	Float	Continuous
Work load average per day	Float	Continuous
Hit target	Float	Continuous
Disciplinary failure	Object	Categorical
Education	Object	Categorical
Son	Object	Categorical
Social drinker	Object	Categorical
Social smoker	Object	Categorical
Pet	Object	Categorical
Weight	Float	Continuous
Height	Float	Continuous
Body mass index	Float	Continuous
Year	Float	Continuous
Absenteeism time in hours	Float	Continuous

Chapter 2

Methodology

2.1 Data Pre-processing

During this stage the data is explored and cleaned so as to make it fit for modelling. The data is visualized using different graphs to gain insight about it. Exploratory data analysis begins by exploring the class or data type of the different predictor variables. The data is searched for presence of any missing values that can be either ignored (if more than 30% of data is missing) or imputed using different methods like mean, median, KNN (for numeric data) or mode (for categorical data). The variables are visualized to analyse their distribution (e.g. histograms can be used to visualize the distribution of variables). Outliers from the data are removed as they are inconsistent with the rest of the data. Further variables are selected that contribute in target value estimation. Predictors that carry repetitive information are removed. Feature engineering may also be performed to generate new variables that will have a relation with the target variable. The following subsections will describe the pre-processing steps followed.

2.1.1 Feature Engineering

The dataset provided does not contain the **"year"** column which is essential for us to forecast future values. However, the data is arranged in the serial order of years and their months beginning from JULY 2007 to JULY 2010. This helps us to generate a new column "year" in the dataset by assigning the year label (2007, 2008, 2009 and 2010) to the subset of the data. Hence, we divide the dataset as:

R Snippet

```
#create new column year
dataset$year=NA
dataset$year[1:113]=2007
dataset$year[114:358]=2008
dataset$year[359:570]=2009
dataset$year[571:740]=2010
```

2.1.2 MISSING VALUE ANALYSIS

The dataset provided as missing values in various columns which need to be imputed as the total missing values is less than 30% of the whole data. The count of missing values is given in the following table.

Table 2.1 Missing value count

Variable Name	Missing values
Body mass index	31
Absenteeism time in hours	22
Height	14
Work load average per day	10
Education	10
Transportation expense	7
Son	6
Hit target	6
Disciplinary failure	6
Social smoker	4
Age	3
Service time	3
Reason for absence	3
Distance from residence to work	3
Social drinker	3
Pet	2
Weight	1
Month of absence	1
Seasons	0
Day of the week	0
Year	0
ID	0

i) Remove insignificant data: The last 3 rows of the dataset contained insignificant information with 0 values in three columns: “reason for absence”, “month of absence”, “absenteeism time in hours”. Hence, these rows are removed thereby reducing the rows to 737.

ii) Correct typing error if any: The entry (67,3) has a missing value. But since the data is serially arranged as per year and month, we can surely say that this missing value has **month** value of 10.

iii) Missing value imputation:

We have computed missing values by grouping the whole data using the year column. This helps to generate authentic results. The missing values in continuous variables are computed using **MEAN** while the missing values in categorical variables are computed using **MODE** method.

Python Snippet:

```
#categorical variable
cat_var=['ID','Reason_for_absence','Month_of_absence','Day_of_the_week','Seasons','Disciplinary_failure','Education','Social_drinker','Social_smoker','Son','Pet']
#numeric variable
num_columns=dataset.select_dtypes(exclude=['object'])
num_var=num_columns.columns
#impute missing value for categorical variables using mode over data of each year
for i in cat_var:
   dataset[i]=dataset.groupby('year')[i].transform(lambda x: x.fillna(x.mode()[0]))
#impute missing value for numeric variables using mean over data of each year
for i in num_var:
   dataset[i]=dataset.groupby('year')[i].transform(lambda x: x.fillna(x.mean()))
```

2.1.3 OUTLIER ANALYSIS

Outliers are the observations that are inconsistent with the rest of the data. Our dataset has outliers in multiple columns. We will impute the outliers using KNN imputation method. Also, since the outlier analysis can be performed only on numeric data, the categorical variable also have to be converted to numeric data by assigning them numeric labels. However our dataset already has numeric categories. The outliers are replaced by missing values (NA) which are imputed using KNN imputation method.

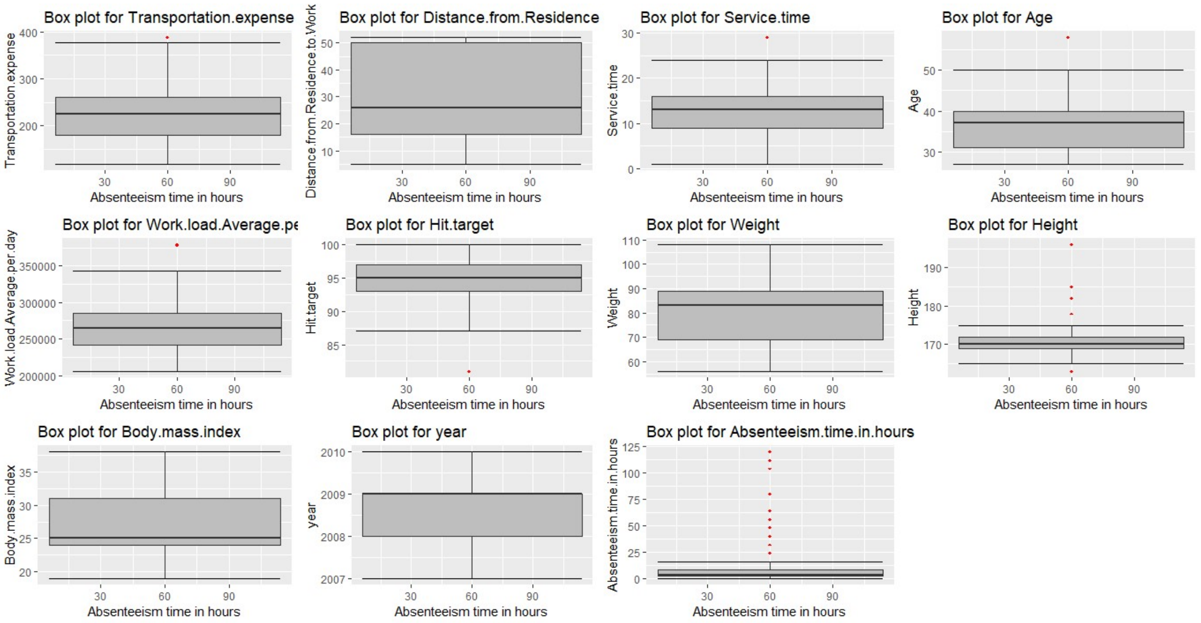


Fig 2.1 Outlier Analysis using Box-plot

2.1.4 FEATURE SELECTION

This process is performed to find variables that contain redundant information. Such variables need to be removed. Also variables that contribute very little towards the target variable have to also be removed. In order to identify the redundant variables we generate a correlation plot as shown in Fig 2.2. Variables that have an absolute value of correlation coefficient greater than 0.95 are said to carry redundant information. In other words such variables are highly correlated with each other. Fig 2.2 shows that our dataset has no redundant variables.

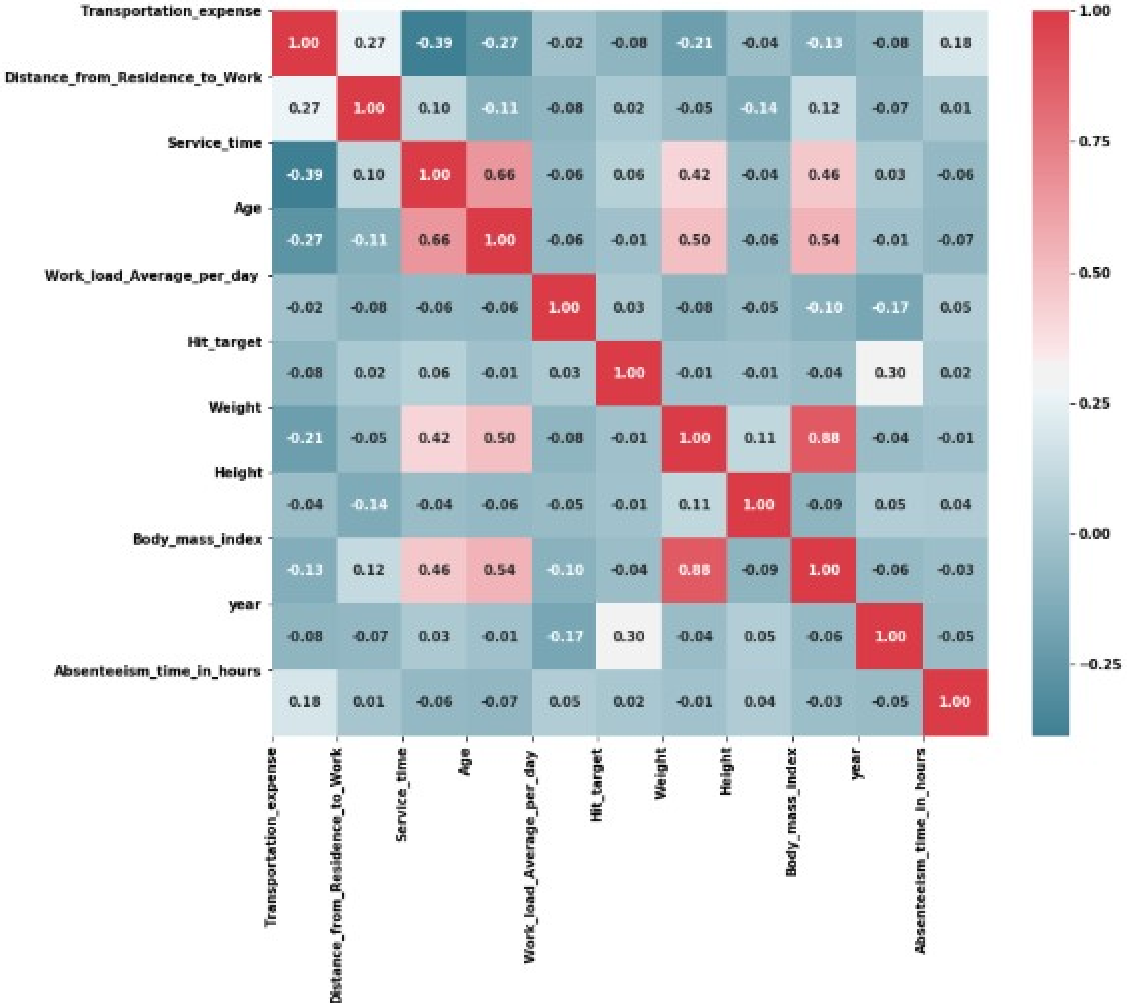


Fig 2.2 Correlation plot

2.1.5 Feature Importance using Random Forest

The Random Forest regressor provides two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy. We have used Mean decrease impurity method for estimating the important variables. Fig 2.3 shows variables arranged in decreasing order of their importance score. Random forest consists of a number of decision trees. Every node in the decision trees is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set. The measure based on which this optimal condition is chosen is called impurity. For classification, it is typically either Gini impurity or information gain/entropy and for regression trees it is variance. Thus when training a tree, it can be computed how much each feature decreases the weighted impurity in a tree. For a forest, the impurity decrease from each feature can be averaged and the features are ranked according to this measure.

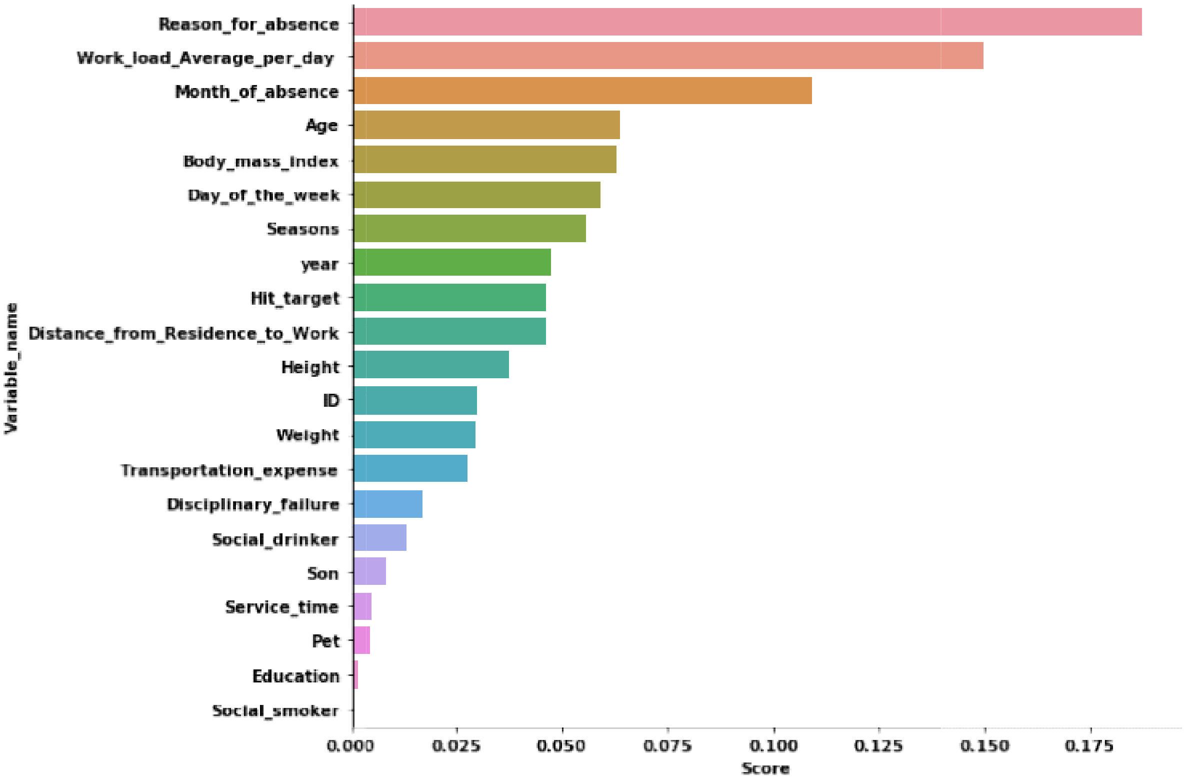


Fig 2.3 Decreasing order of Feature importance

2.2 Data Visualization

Univariate and Bivariate data analysis provides some extra insights of the data. It helps us to easily observe the trends and patterns followed by the data. This will enable us to analyse the pattern that is causing high rate of Absenteeism. Further it will allow us to make necessary changes to lower the future Absenteeism rate.

2.2.1 Univariate Analysis:

Fig 2.4 shows the density plot of the numeric variables.

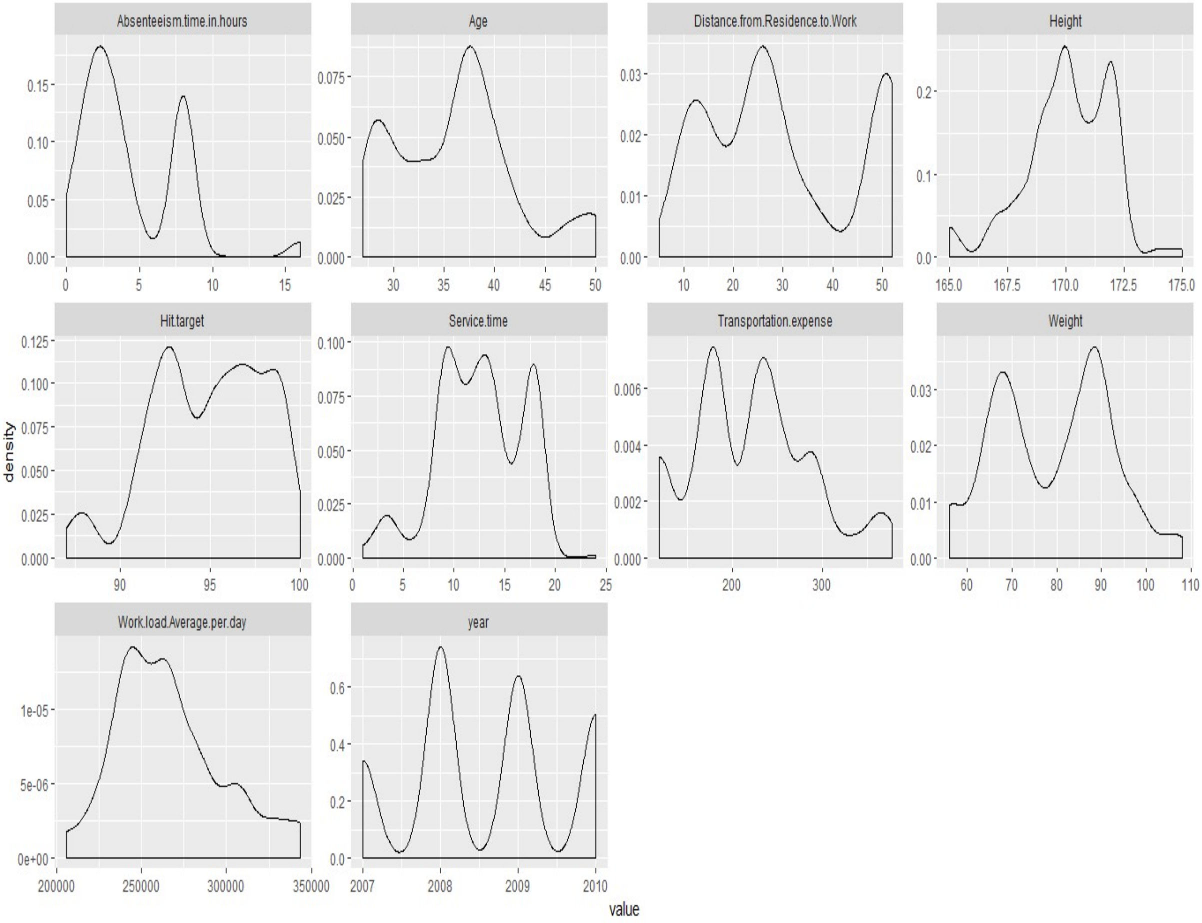
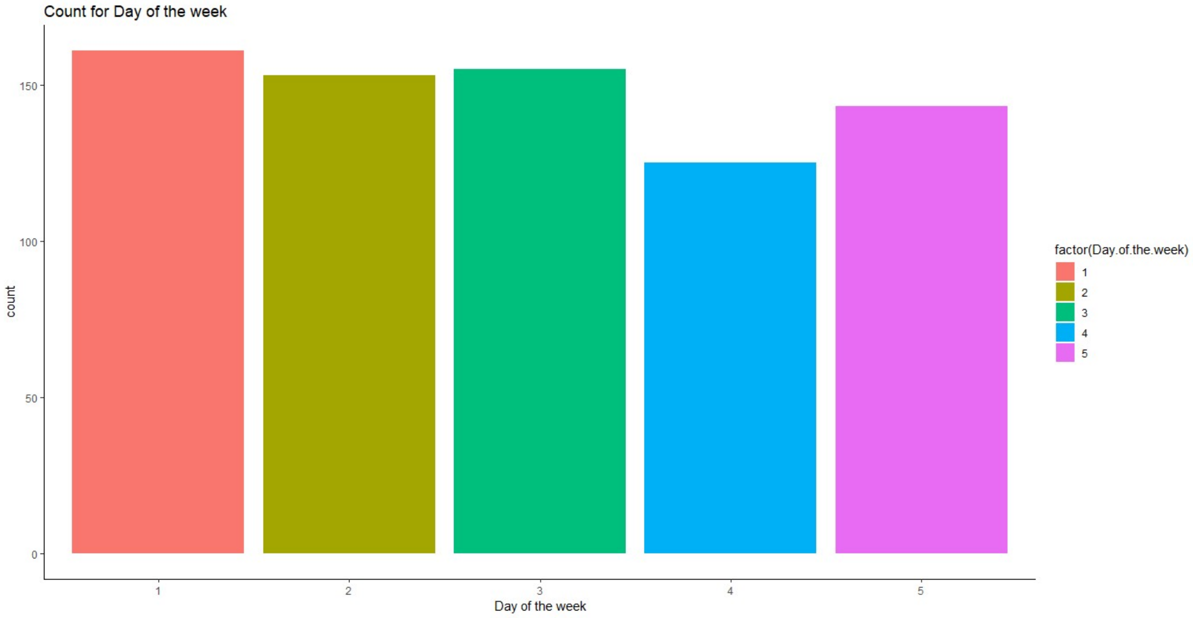
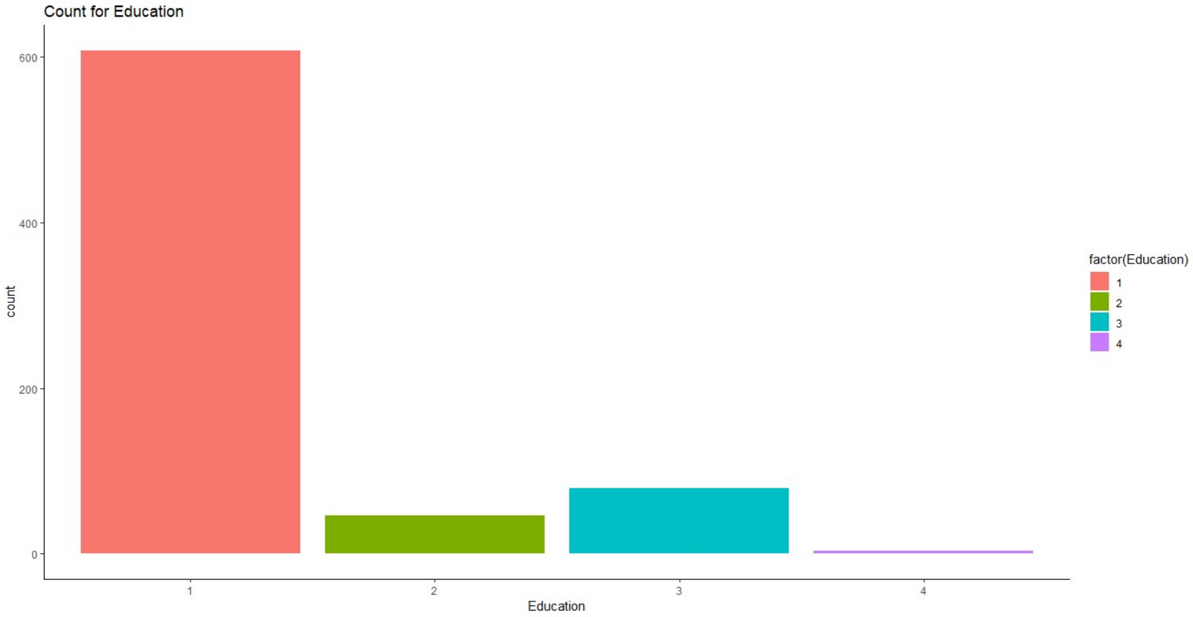
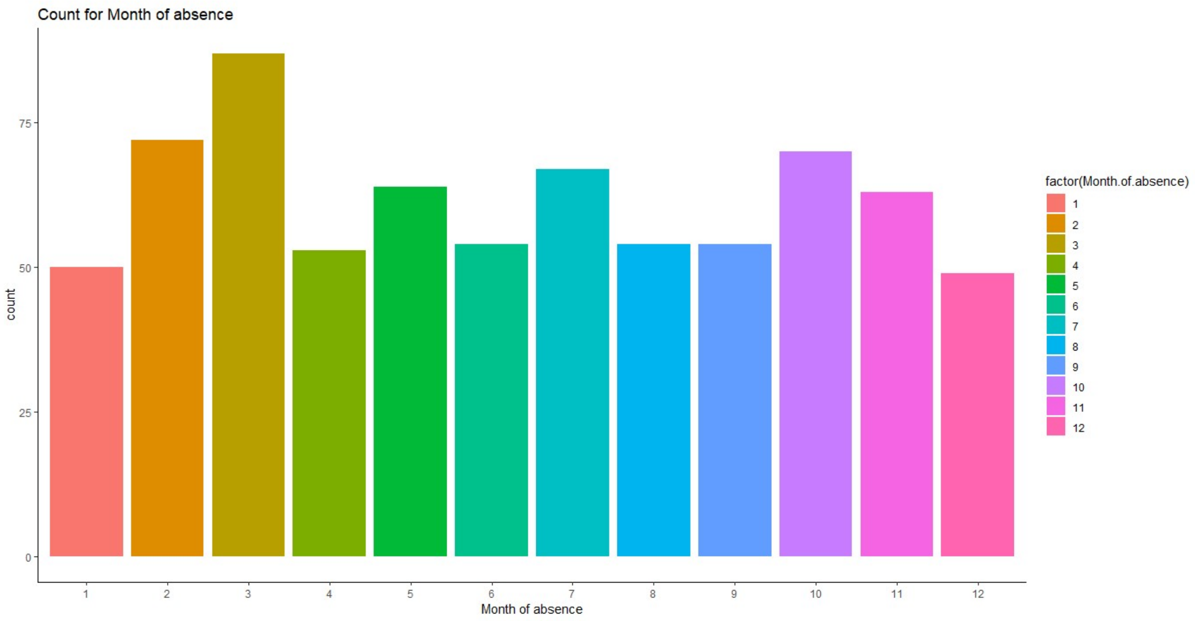


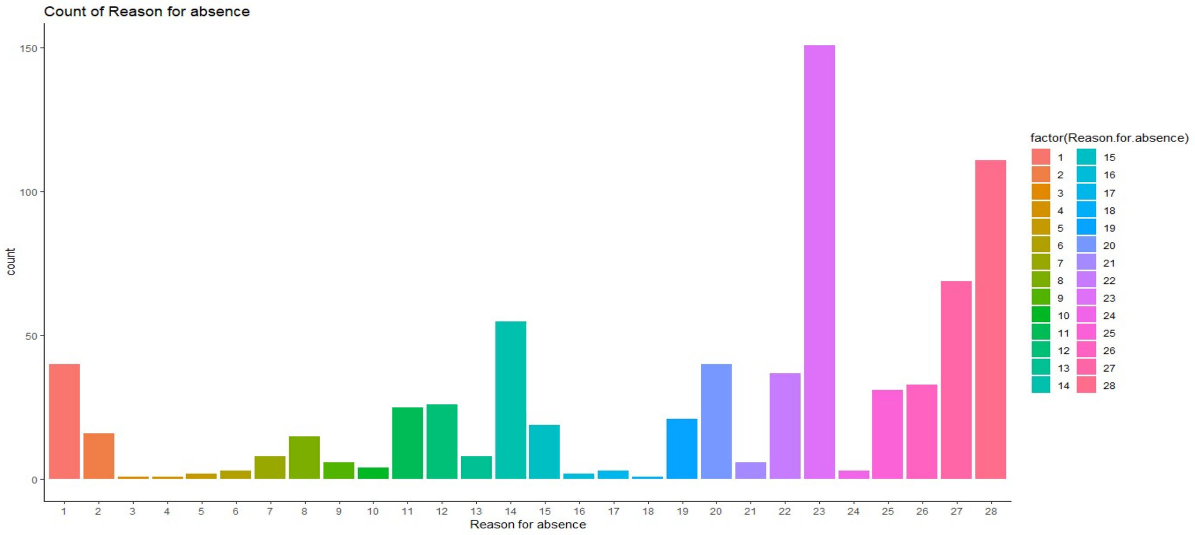
Fig 2.4 Density plots for numeric variables

Fig 2.5 shows the count plot of the categorical variable






Count for Month of absence
count
0
25
50
75
1
2
3
4
5
6
7
8
9
10
11
12
Month of absence
factor(Month.of.absence)
1
2
3
4
5
6
7
8
9
10
11
12


Count of Reason for absence
count
0
50
100
150
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
Reason for absence
factor(Reason.for.absence)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

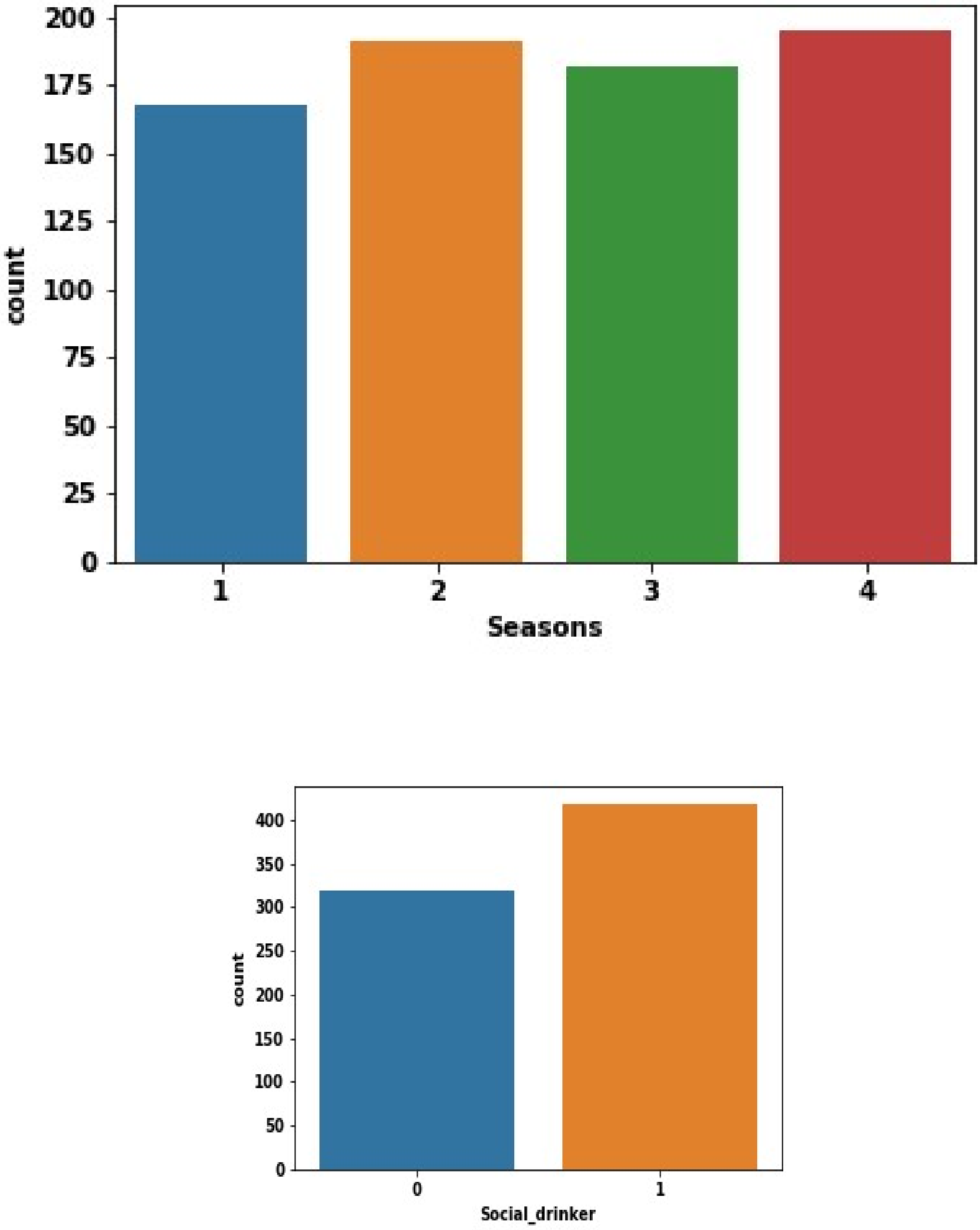


Fig 2.5 Count Plots of categorical variables

The univariate analysis clearly shows the following trends:

a) Maximum absenteeism occurs due to reason Blood Donation (label=23), Dental Consultation (label=28) and Physiotherapy (label=27). It can be observed here that these reasons for absence are not attested by the International Code of diseases (ICD).

b) Maximum absenteeism occurs on Monday (label=2), i.e. the day immediately after Sunday. This means that the employees tend to enjoy a long weekend.

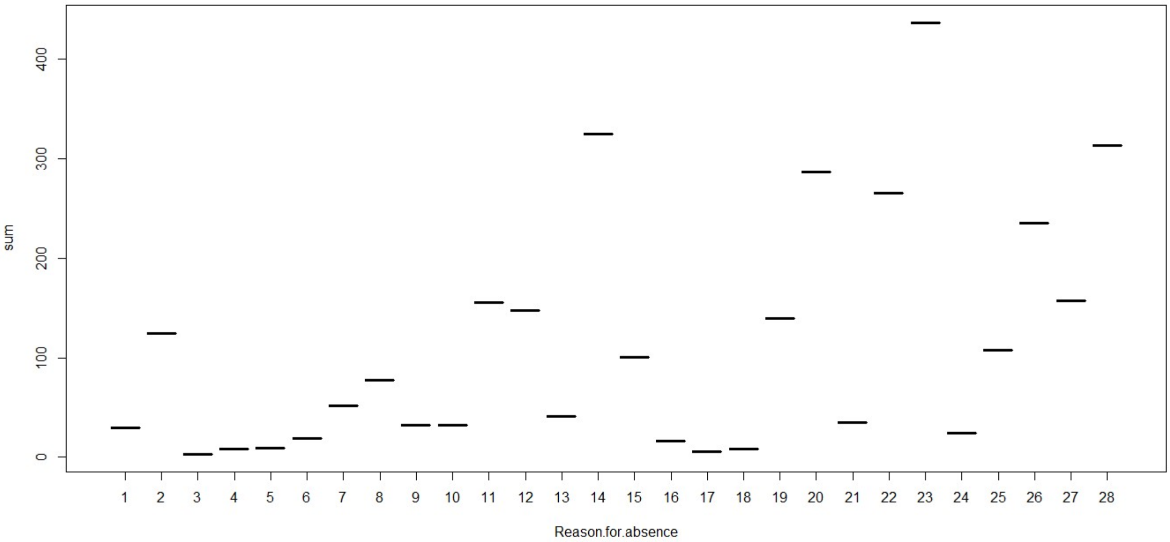
c) Maximum absenteeism occurs in the month of March (label=3). This might due to the fact that the financial year of any company is from 1st April to 31st March and so the employees take their left over holidays in the last month so that they do not lapse.

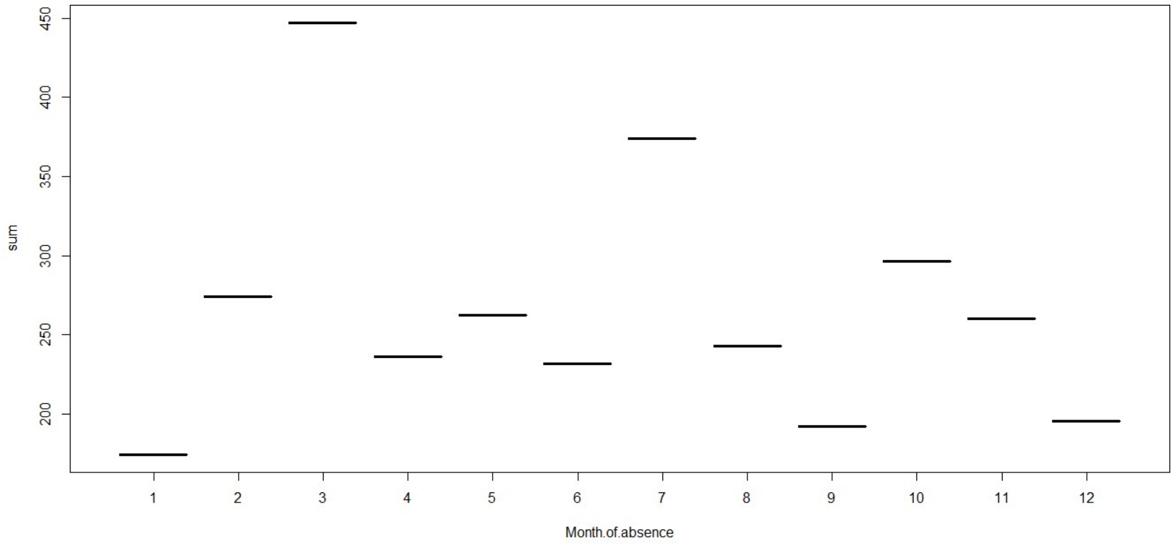
d) Maximum absents are from the employees that have high school degree (label=1). These employees might belong to the class of peons or helpers. As it is seen that employees with higher qualification tend to be present for the maximum time.

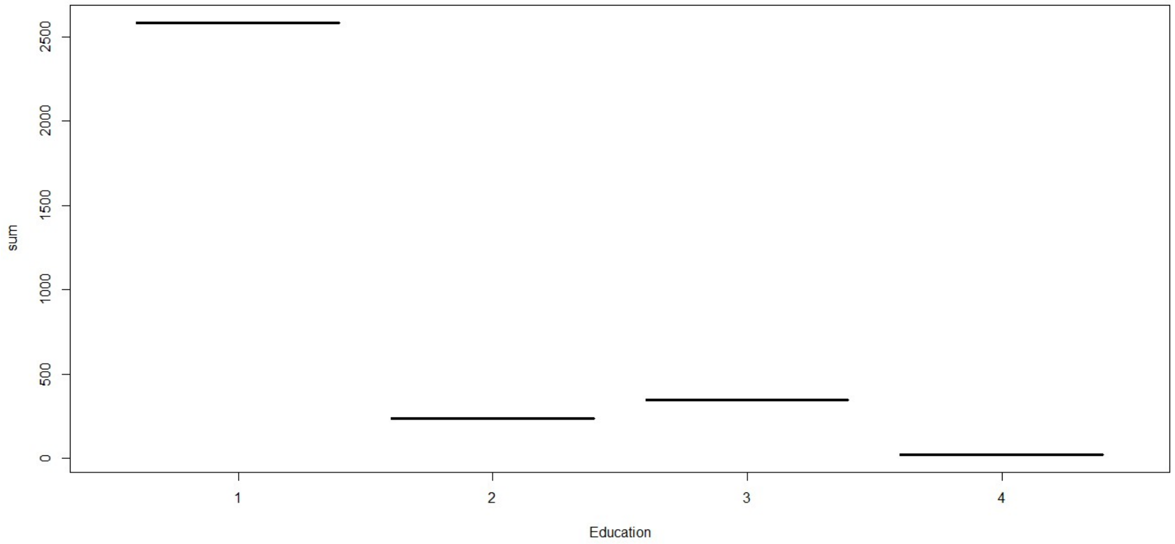
e) Maximum absents are from employees who are social drinkers (label=1). This absenteeism by liquor drinkers occurs due to the obvious reasons.

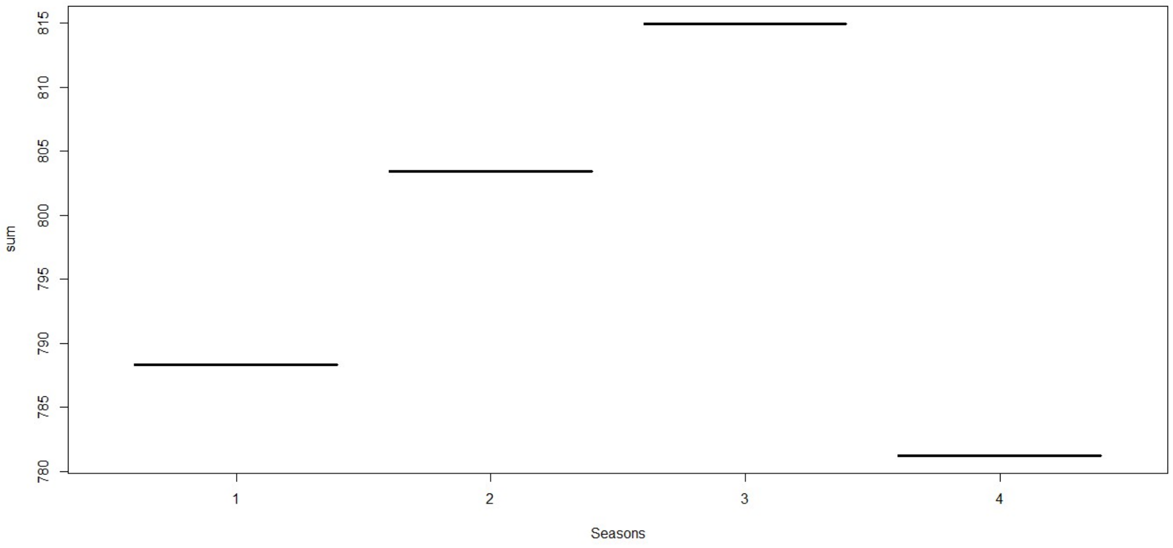
2.2.2 Bivariate Analysis

Fig 2.6 shows the bivariate analysis of categorical variables against target variable- "Absenteeism time in hours"




sum
450
400
350
300
250
200
1
2
3
4
5
6
7
8
9
10
11
12
Month.of.absence


sum
2500
2000
1500
1000
500
0
1
2
3
4
Education



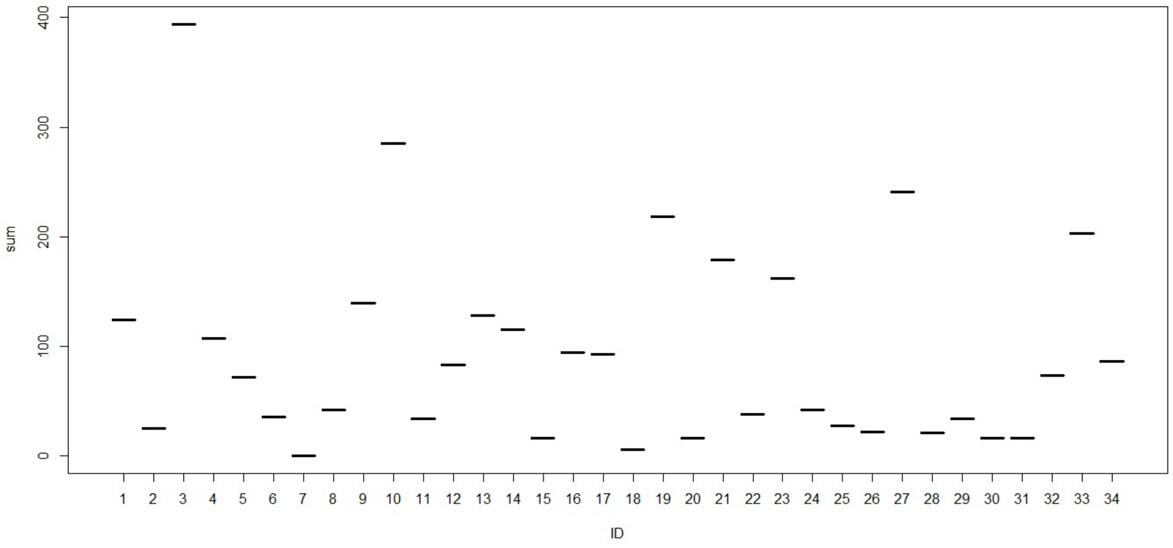


Fig 2.6 Bivariate analysis

The same observations can be drawn from both univariate and bivariate analysis of data.

Conclusion:

a) Maximum absenteeism due to reasons 23,28,27

b) Maximum absenteeism on Monday

c) Maximum absenteeism in the month of March

d) Maximum absenteeism by social drinkers

e) Maximum absenteeism by employees with lowest qualification (high school)

f) Maximum absenteeism in Winter season (season 3)

g) Maximum absenteeism by employees with ID: 3, 11, 28, 20, 34 (in the decreasing order).

SUGGESTIONS:

The company can take the following measures to reduce absenteeism rate in view of the current trend of absenteeism.

a) Leave must be sanctioned only with a valid medical certificate.

b) Keep better facilities for employees during the winter season in the office.

c) Leave on Monday should be granted with prior notice only.

d) Additional perks to employees with high attendance. This will give a sense of competition to other employees as well.

2.3 TIME SERIES ANALYSIS

2.3.1 INTRODUCTION TO THE CONCEPT

The second part of the project expects us to forecast the "Absenteeism time in hours" for the year 2011 using the data of the previous years (2207, 2008, 2009, 2010) made available assuming that the same trends and patterns will be prevalent in the future time. This makes the problem as a "Time Series Forecasting Problem".

2.3.1.1 TIME SERIES:

A sequence of observations spaced over regular intervals of time is called as a Time Series. A Time series comprises of a trend component (upward or downward slope), seasonal component (peaks) and noise component. Decomposing the time series allows us to visualize these components separately for analysis. Combining back these components yields the original series. Time Series Decomposition is mainly of 2 types depending on how these components form the time series (x_t):

a) Additive decomposition

Time Series, x_t = Trend + Seasonality + Random noise

b) Multiplicative decomposition

Time Series, x_t = Trend * Seasonality * Random noise

2.3.1.2 TYPES OF TIME SERIES:

Time series are of the following 2 types:

a) Stationary Time Series: A stationary time series is one in which mean, variance and correlation of i^{th} term with $(i+m)^{th}$ term should not be a function of time. Its behaviour and properties don not depend on the time at which the series is observed.

b) Non-stationary Time Series: A Time series that is not stationary is a non-stationary time series.

2.3.1.3 REASONS FOR TIME SERIES TO BE STATIONARY

1) A time series needs to be stationary because if a time series has a particular behaviour over time, there is a high probability that it will follow the same in the future.

2) Also, theories related to stationary series are more mature and easier to implement than the non-stationary series.

2.3.1.4 CHECKING TIME SERIES STATIONARITY

a) Visual Techniques

Plotting Rolling Statistics:

Visualize the plot of moving average and moving variance to see if it varies with time. This plot is generated by computing the mean/ variance over a window of observations.

b) Statistical Techniques

Dickey-fuller test:

This is one of the statistical tests for checking stationarity. First we consider the null hypothesis: the time series is non- stationary. The result from the rest will contain the test statistic and critical value for different confidence levels. The idea is to have Test statistics value less than critical value; in this case we can reject the null hypothesis and say that this Time series is indeed stationary.

2.3.2 TIME SERIES FORECASTING MODEL GENRATION

1) Creating a Time Series

We create the time series from our dataset by summarizing the value of the "Absenteeism time in hours" over month and year. Fig 2.7 illustrates this aggregation. A Date-time index is attached with this result to generate our time series of interest. The data made available to us is from JULY 2007 to JULY 2010, while we have to make forecast for the year 2011. This means that we have to forecast 12 values in the future. The time series is shown in fig 2.8.

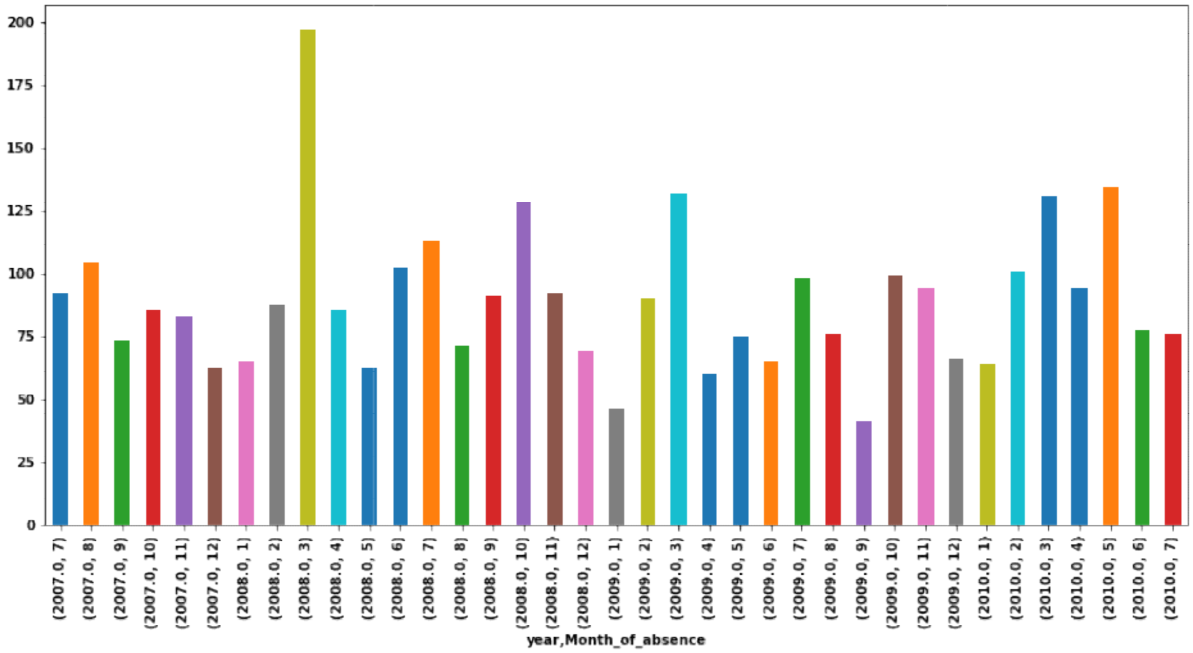


Fig 2.7 Absenteeism hours summarized over year and month

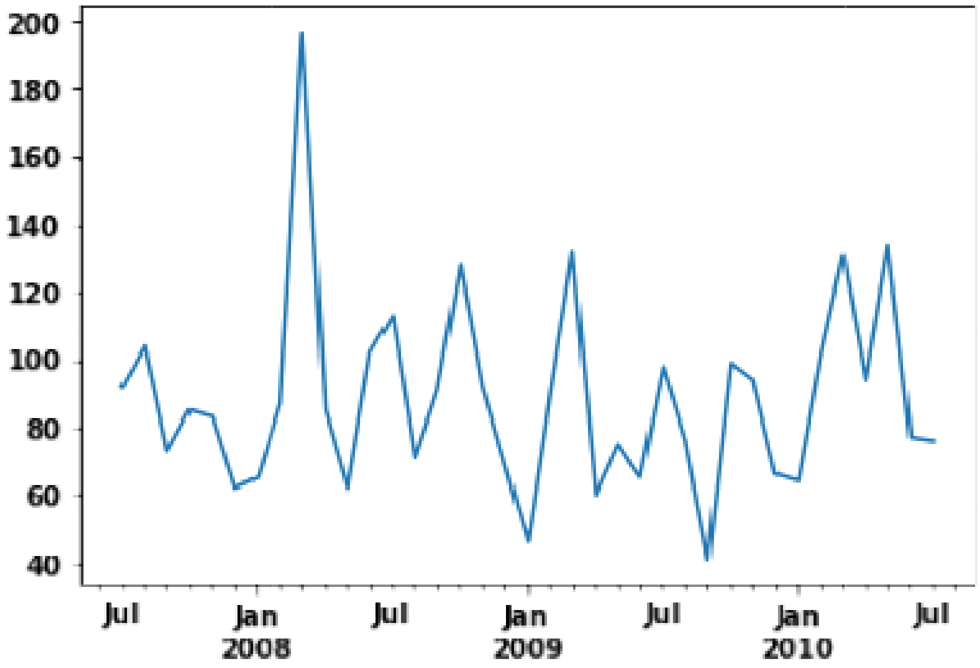


Fig 2.8 Time Series

Python Snippet

```
dataset_ts =
dataset.groupby(['year','Month_of_absence'])['Absenteeism_time_in_hours'].sum()
dataset_ts.index=pd.date_range(start = '2007-07-01',end='2010-08-01', freq = 'M')
```

R Snippet

```
#aggreagte absenteeism hours w.r.t year and months
hours_vs_year_month=dataset%>%
  group_by(year,Month.of.absence)%>%
  summarise(sum=sum(Absenteeism.time.in.hours))
hours_vs_year_month=as.data.frame(hours_vs_year_month)
#rename last variable indicating sum of absenteeism hours
names(hours_vs_year_month)[names(hours_vs_year_month)=="sum"]
="absent_hours"
#create a time series by assigning the time to the absent_hours variable
#each data point is mapped to a month of the year in the series from july 2007 to july
2010
dataset_ts=ts(hours_vs_year_month$absent_hours,frequency =
12,start=c(2007,7),end=c(2010,7))
```

2) Dividing Data into Train and Test subsets:

The dataset is divided into train and test dataset. The time series model is fit on the training data and later applied to forecast values for the time interval belonging to the test subset and further more time intervals. In our dataset, we have observations from JULY 2007 to JULY 2010 totalling to 37 observations. We keep the observations from JULY 2007 to DEC 2009 into the training set (i.e. 30 observations) and observations from JAN 2010 to JULY 2010 into the testing dataset (i.e. 7 observations).

This step helps us to calculate the forecast error. The forecast error is actually calculated using the observed and forecasted values of the test dataset. This can be done by using the fitted model to forecast (7+12) values into the future. Here, 7 values will be the forecasted values of the test data observations and the 12 values will be the forecasted values for the year 2011.

3) Checking Data Stationarity:

I) Plotting Rolling statistics:

The rolling mean and standard deviation plotted for our time series is illustrated in Fig 2.9. Figure shows that both the mean and standard deviation are almost constant over time proving the time series to be stationary.

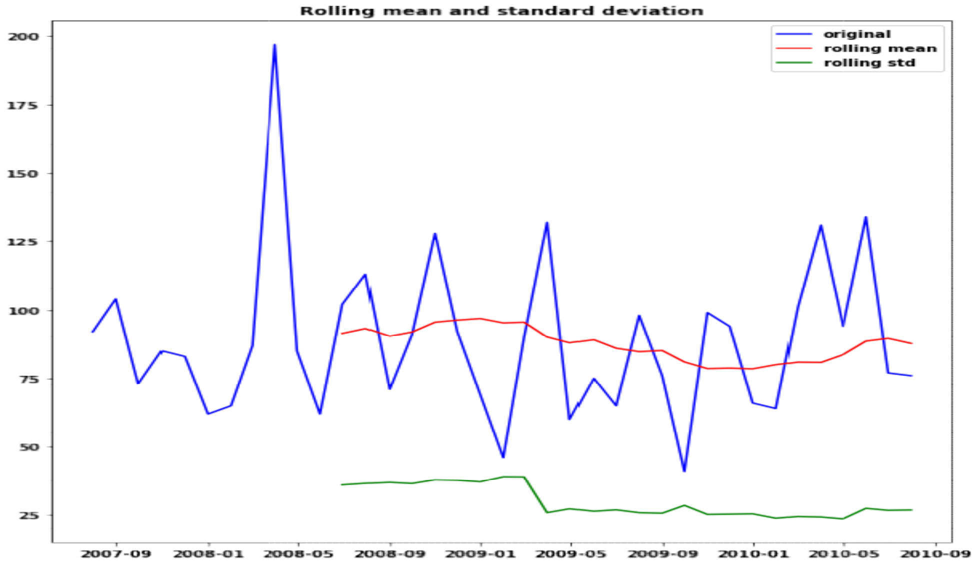


Fig 2.9 Rolling statistics for Time series

II) Dickey Fuller Test:

The result obtained after applying Dickey Fuller test to check whether data is stationary or not is given as:

```
ADF Statistic: -5.665478
p-value: 0.000001
Critical Values:
    1%: -3.633
    5%: -2.949
    10%: -2.613
```

Null Hypothesis, H0: Time series is not stationary

According to the above result,

Test statistic < Critical value

p-value < Significance value

Hence, we reject the null hypothesis, indicating that our time series is stationary.

4) Decomposing Time Series

We have used an additive model for decomposing the time series into its trend component, seasonality component and random fluctuations component. Fig 2.10 shows all the constituent components.

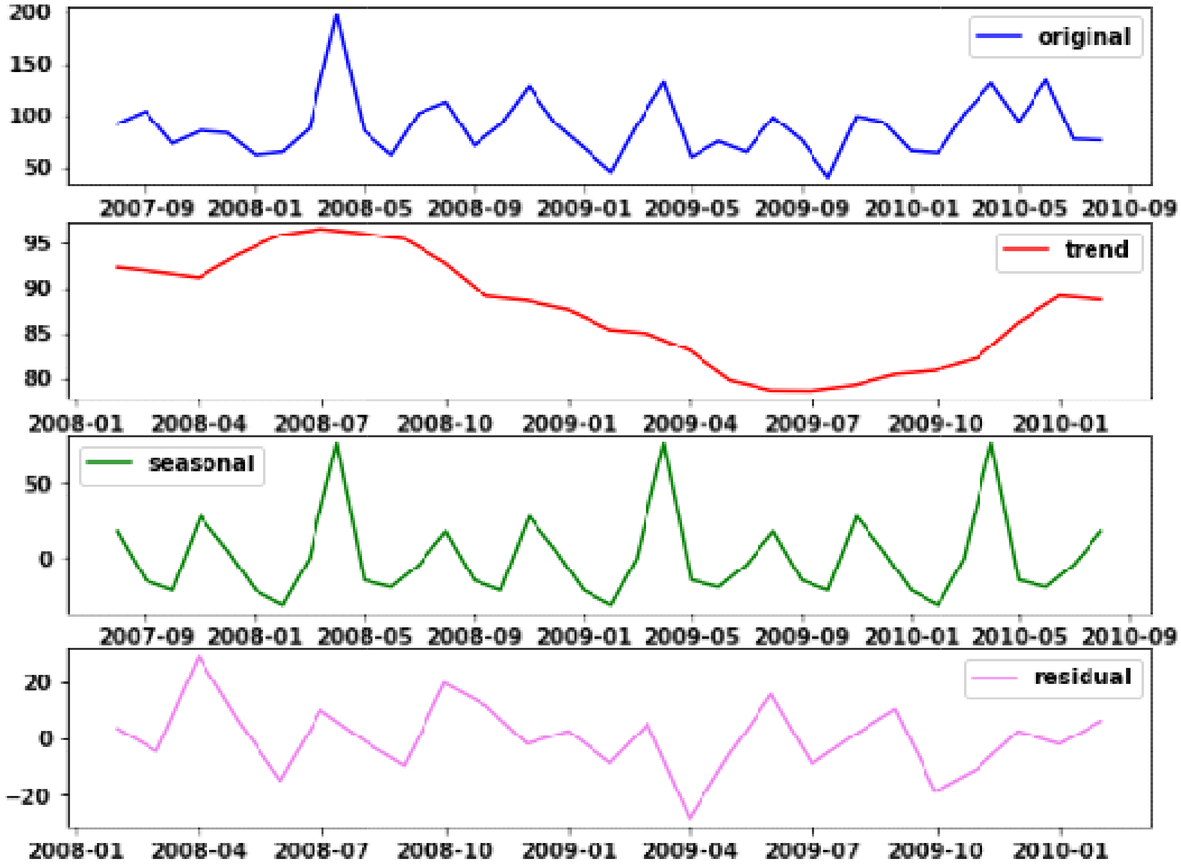


Fig 2.10 Time Series Components

5) Time Series Forecasting Models:

METHOD 1: HOLT WINTERS METHOD

(TRIPLE EXPONENTIAL SMOOTHNING METHOD)

Exponential Smoothing: According to simple exponential smoothing, the expected value can be computed using the current observed value and the previous expected value as per the following formula:

$$y^\wedge x=\alpha\cdot yx+(1-\alpha)\cdot y^\wedge x-1$$

here, α is the smoothing coefficient.

Triple Exponential smoothing is made of three terms: Level, Trend and Seasonality.

Level refers to the expected value.

Trend refers to the slope.

Seasonality refers to the interval after which a series repeats.

The future values are calculated based on the following formulae for an additive model:

$\ell x=\alpha(yx-sx-L)+(1-\alpha)(\ell x-1+bx-1)$ level

$bx=\beta(\ell x-\ell x-1)+(1-\beta)bx-1$ trend

$sx=\gamma(yx-\ell x)+(1-\gamma)sx-L$ seasonal

$y\hat{}x+m=\ell x+mbx+sx-L+1+(m-1)modL$ forecast

OUTPUT:

The output of the forecasting is given in Fig 2.11

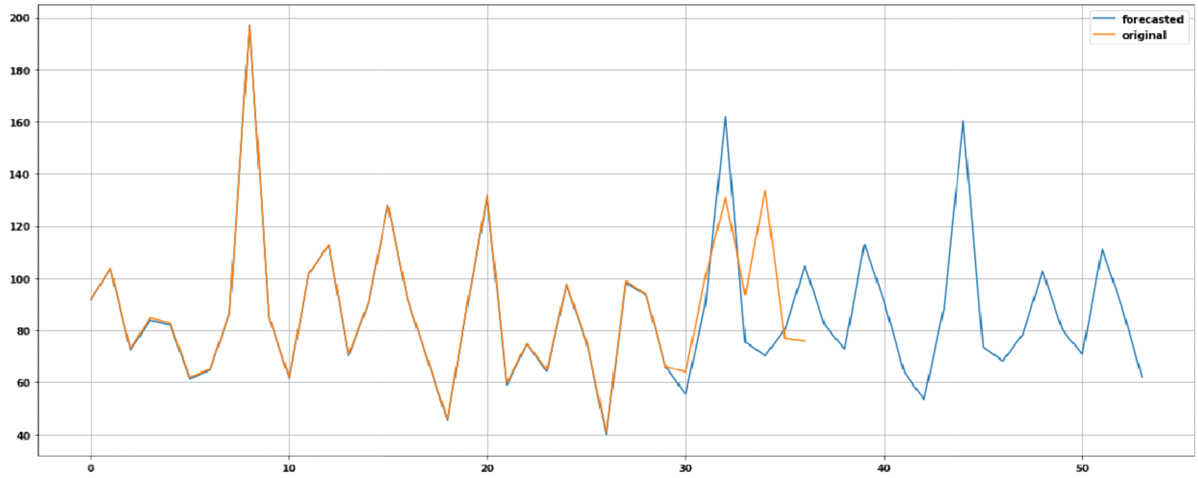


Fig 2.11 Forecasting using Holt Winters model

ERROR METRICS

```
MAPE: 22.47%
RMSE: 30.26%
```

METHOD 2: ARIMA

This method comprises of 2 common processes and integrates them. The ARIMA function uses three parameters p,d,q.

1) Autoregressive process (AR):
   In this process each process is made up of a random error component and a linear combination of prior observations.

   $$x_t = \xi + \phi_1 * x_{(t-1)} + \phi_2 * x_{(t-2)} + \phi_3 * x_{(t-3)} + \ldots + \varepsilon$$

   where,

   ξ is a constant (intercept), and

   $\phi 1$, $\phi 2$, $\phi 3$ are the autoregressive model parameters.

   The number of autoregressive model parameters gives the value of the "p" parameter and can be visualized using a PACF (Partial Auto-correlation) plot.

2) Moving Average process (MA):

Each element in the series can also be affected by the past error that cannot be accounted for by the autoregressive component. Each observation is made up of a random error component and a linear combination of prior random shocks.

$$x_t = \mu + \varepsilon_t - \theta_1 * \varepsilon_{(t-1)} - \theta_2 * \varepsilon_{(t-2)} - \theta_3 * \varepsilon_{(t-3)} - \ldots$$

where,

μ is a constant, and

$\theta 1$, $\theta 2$, $\theta 3$: are the moving average model parameters.

The number of moving average model parameters gives the value of "q" and can be visualized using ACF (Auto correlation function) plot.

We have used SARIMA (Seasonal Auto Regression Integrated Moving Average) model for the time series forecasting.

SARIMA takes the following inputs: SARIMA (p,d,q) * (P,D,Q)S

p=non seasonal AR order

q= non seasonal MA order

d= non seasonal differencing

P=seasonal AR order

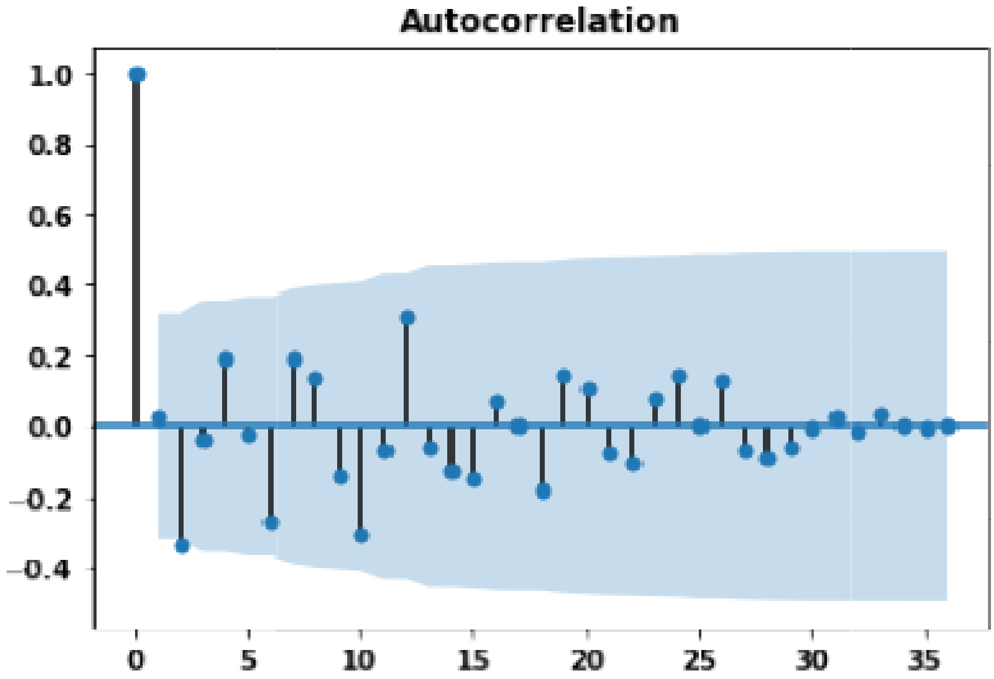
Q=seasonal MA order

D=seasonal differencing

S=time span of repeating seasonal pattern

Fig 2.12 shows the ACF and PACF plots and Fig 2.13 shows the forecast output.

OUTPUT:



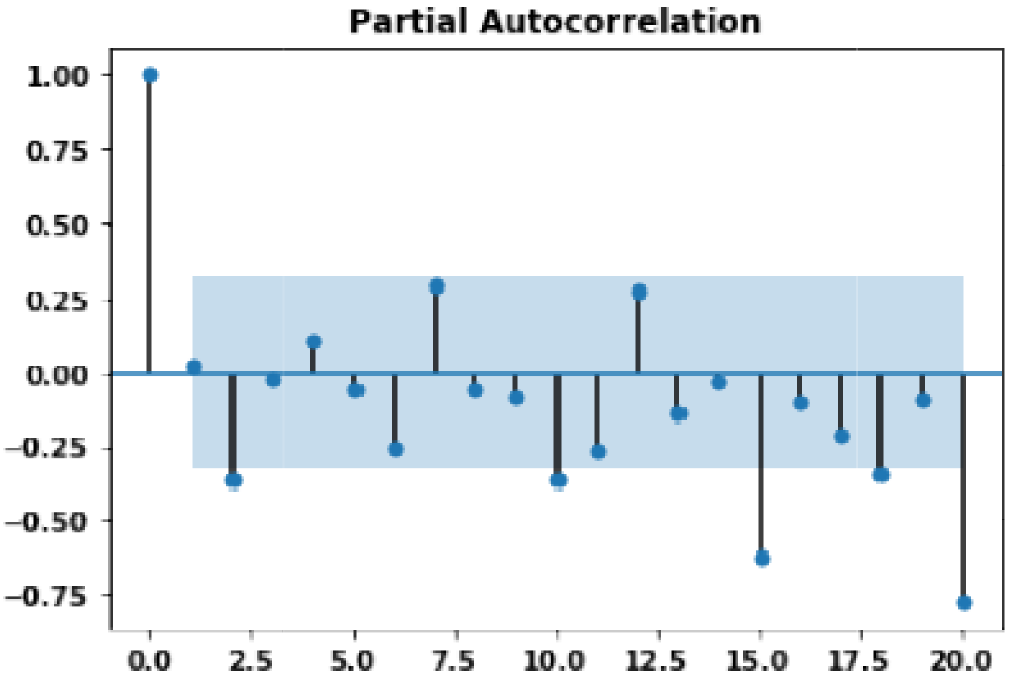

Partial Autocorrelation

Fig 2.12 ACF and PACF plots

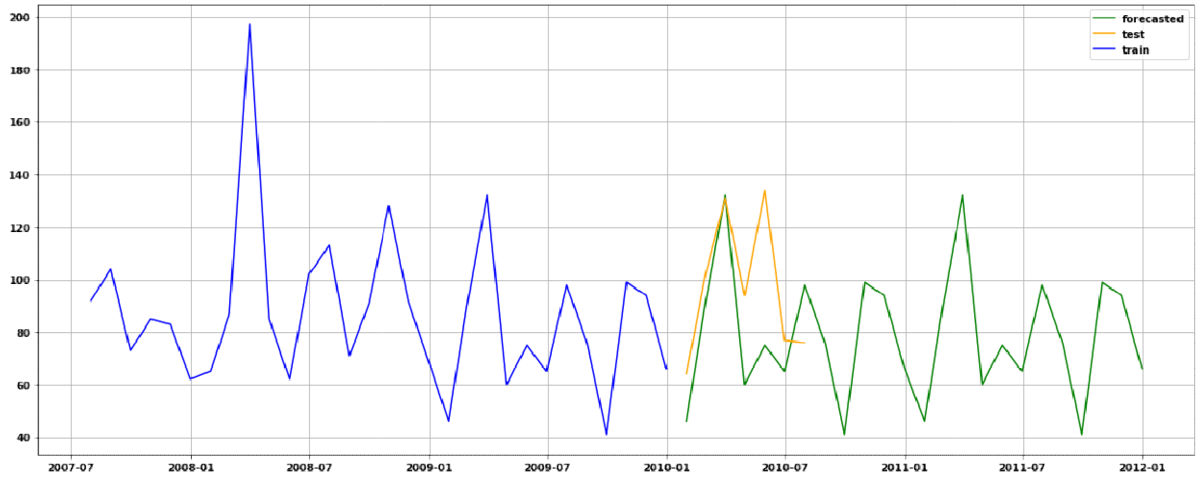


Fig 2.13 forecast using SARIMA model

Error metrics for different SARIMA (p,d,q)*(P,D,Q)S parameters

a) SARIMA(2,1,2)(1,0,0,12):
MAPE: 22.10%
RMSE: 29.08%

b) SARIMA(0,1,0)(1,0,0,12):
MAPE: 24.59%
RMSE: 31.34%

c) SARIMA(0,0,0)(0,1,0,12):
MAPE: 23.50%
RMSE: 31.34%

d) SARIMA(0,0,0)(1,0,0,12):
MAPE: 25.60%
RMSE: 30.43%

e) SARIMA(1,0,1)(0,1,0,12):
MAPE: 25.91%
RMSE: 30.29%

Chapter 3
Conclusion

Among the two techniques- Holt Winters and SARIMA, SARIMA showed slightly better results. So we will use this model for projecting the absenteeism hours for the year 2011.
As per SARIMA, the absenteeism hours every month for the year 2011 are given as:

```
2010-08-31     76.0
2010-09-30     41.0
2010-10-31     99.0
2010-11-30     94.0
2010-12-31     66.0
2011-01-31     46.0
2011-02-28     90.0
2011-03-31    132.0
2011-04-30     60.0
2011-05-31     75.0
2011-06-30     65.0
2011-07-31     98.0
2011-08-31     76.0
2011-09-30     41.0
2011-10-31     99.0
2011-11-30     94.0
2011-12-31     66.0
```