**PROJECT 2**

**Employee Absenteeism**

**Rishabh Jain**

**Data Analyst, Home Credit India Pvt. Ltd.**

**Content**

1. Introduction
   1. Problem Statement
   2. Business Problem
   3. Data
2. Methodology
   1. Data Checks and Modification
   2. Missing and Invalid Value Computation
   3. Feature Engineering
   4. Data Distribution
   5. Outlier Analysis
   6. Relationship between Independent and Dependent Variable
   7. Final Input DataFrame
3. Modelling (ARIMA)
4. Conclusion with Answers
5. **Introduction**
   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?

Dataset Details:

Dataset Characteristics: Timeseries Multivariant

Number of Attributes: 21

Missing Values : Yes

Attribute Information:

1. Individual identification (ID)

2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) stratified into 21

categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the

immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioural disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system

XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities

XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere

classified

XIX Injury, poisoning and certain other consequences of external causes

XX External causes of morbidity and mortality

XXI Factors influencing health status and contact with health services.

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood

donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27),

dental consultation (28).

3. Month of absence

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

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

6. Transportation expense

7. Distance from Residence to Work (kilometers)

8. Service time

9. Age

10. Work load Average/day

11. Hit target

12. Disciplinary failure (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

14. Son (number of children)

15. Social drinker (yes=1; no=0)

16. Social smoker (yes=1; no=0)

17. Pet (number of pet)

18. Weight

19. Height

20. Body mass index

21. Absenteeism time in hours (target)

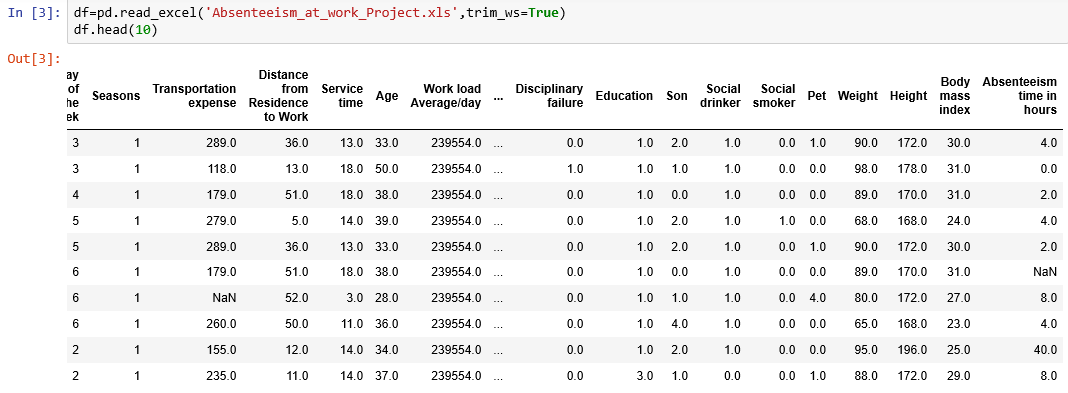
* 1. Business Problem

Most of the organizations race to reduce expenses, increase productivity, catch the opportunity and meet customer satisfaction. The main factor that has a direct effect on the organization’s performance is the human resource. The presence of employees means the planned workflow is performed as expected. It also means a reduction in the costs of management and supervision, overtime, additional workers and the penalty clause. By contrast, the absence of an employee leads to the exact opposite. The organization gains a

highly competitive advantage tool that could be used to address the consequences of the employees’ absence and help human resources management to improve the process of recruitment and crisis management. So, based on the data understanding it can be predicted that how much loss would a company will bear in coming future and how it can be reduced to increase productivity.

* 1. Data

The ‘Absenteeism\_at\_work\_Project.xls’ file is used for the analysis and modelling. The glimpse of few rows of data is given below



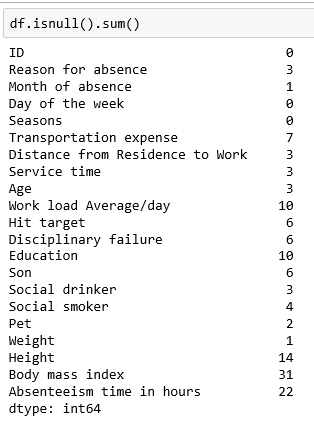
The data contain 740 rows and 21 variables or attributes as mentioned in the problem statement. Out of these variables, “Absenteeism time in hours” variable is target variable.

1. **Methodology**
   1. Data Checks and Modification

To initialize the process of data modelling, certain checks are performed so that the model doesn’t get biased and predict wrong value of target variable for the certain input variables.

First check is check for NULL values in the dataset. If so, then certain procedures are to be used to fill that NULL values, for example

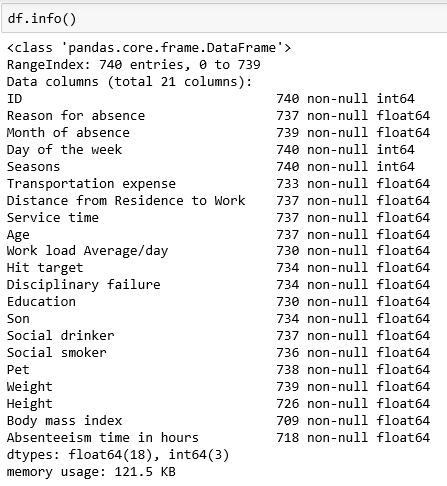
1. Removing NULL value containing row
2. Mean method i.e., filling NULL value with the mean for the values in that column
3. Median method i.e., fiulling NULL values with median value
4. KNN Imputation (K-Nearest Neighbour mehtod)



‘Absenteeism time in hours’ is our target variable so NA value in the dataset will be removed as the information will not contribute to build model to predict year 2011 losses.

The above is used in python to check for NULL values in each column of the dataframe.

Second check is to check for the datatypes of the variables in the loaded dataframe.



* 1. Missing and Invalid Value Computation

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If a column has more than 30% of data as missing value either we ignore the entire column or we ignore those observations.

To calculate missing value percentage for each column the following function is used in Python:

*def missing\_perc(data):*

*Missing\_Value = pd.DataFrame((data.isnull().sum()/len(data)\*100))*

*Missing\_Value = Missing\_Value.rename(columns = {0: 'Missing\_percentage'})*

*#Arranging Missing Values in Decreasing Order*

*Missing\_Value = Missing\_Value.sort\_values('Missing\_percentage', ascending = False)*

*print(Missing\_Value)*

in R:

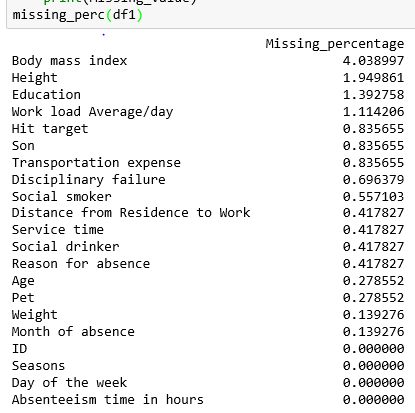
*missing\_perc<-function(data){*

*missing\_value = (as.data.frame(colSums(is.na(data)))\*100/nrow(data))*

*colnames(missing\_value) <- c("Missng Value Percentage")*

*View(missing\_value)*

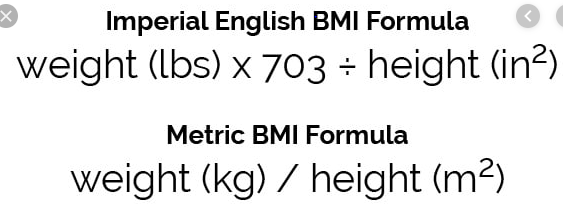
*}*



In the given data the maximum percentage of missing value is 4.189% for body mass index column. So, we will compute missing values. Also, datapoints with more than one null value i.e., 16 observations have been dropped to easily compute the rest of the missing data.

1. Body Mass Index, Height & Weight

Body mass index (BMI) is an estimate of body fat based on height and weight. It doesn’t measure body fat directly, but instead uses an equation to make an approximation. BMI can help determine whether a person is at an unhealthy or healthy weight. It is calculated as below:



From the observations it is noted that the height is given in centimetres and weight is given in kg. So, metric formula is used to calculate missing value.





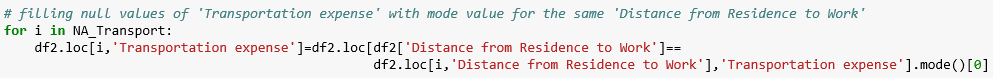


1. Transportation expense & Distance from Residence to Work

It is noted from the two columns that both are directly proportional to each other. So compute Transportation expense following strategy is applied:

1. Indexes of such Null values are stored in an array
2. Whatever value of ‘Distance from Residence to Work’ is present at that index, for the same value corresponding mode for ‘Transportation expense’ is calculated.
3. That mode value is then assigned to that Null value.

The same strategy is performed for Null values in ‘Distance from Residence to Work’ as both are directly proportional to each other.

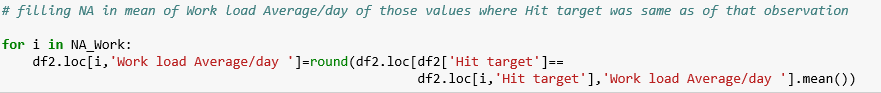




1. Hit Target & Work load Average/day

Hit Target and Work load Average/day were both proportional to each other as to achieve higher targets the employee has to do more work. So, the same strategy is applied for null Hit target values i.e., filling it with the mode value of Hit target for same Work load Average/day in the dataset.

But for computation of Null Work load Average/day rather than mode, mean value is filled.



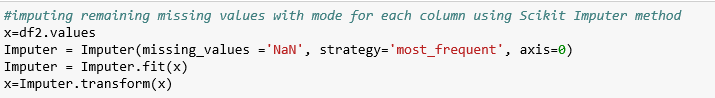


1. Month of absence

It is observed that for the months, the data is given in sequence with the index values as from 56-68, 261-286, 473-497 all indexes have same value for "Month of absence", so it will be logical to fill the null value in month of absence with 10 as it is in sequence with month 10.

1. For the rest of variables

To compute the rest of null values Scikit imputer method is applied, filling null values with mode value for each respective column.



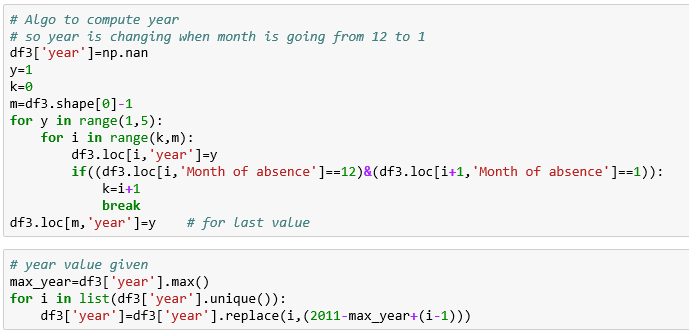
For invalid value of “Reason for absence” i.e., 0 the value is replaced by 23 which means unjustified absence.

For invalid value “Month of absence” i.e., 0 the last two observations out of 3 such observations were dropped off as both has ‘Absenteeism time in hours’ equals to 0 and season is 2 or 3. So, these 2 observations won’t contribute much. But for the third observation the ‘Month of absence’ is replaced with 7 as the season is 1 and the dataset is following the sequence for months.

* 1. Feature Engineering

The dataset contains monthly data for employees, and seeing the pattern for the ‘Month of absence’ it seems like the data given is from June, 2007 to June, 2010. So, using this information new variables are added i.e., year & Period. These features will help in the building of machine learning model.

To compute ‘year’



To compute ‘Period’

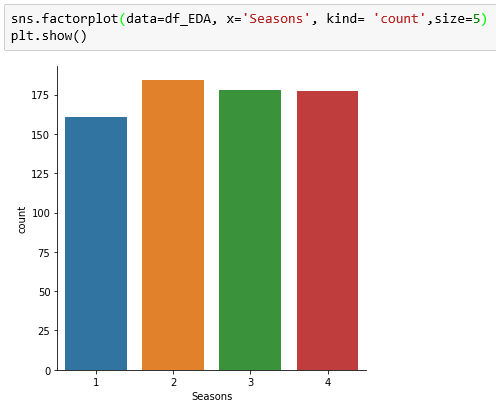


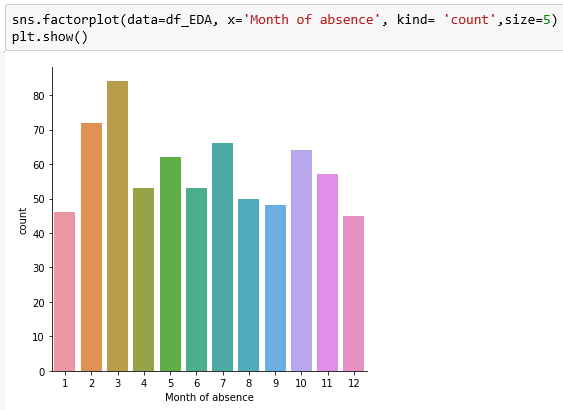
* 1. Data Distribution

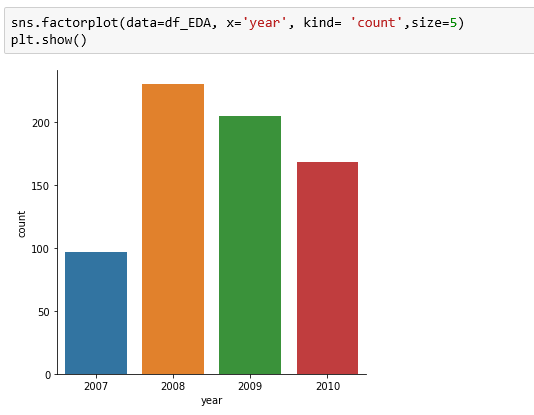
The model is greatly impacted by the data distribution because on the basis of that the model makes rules to predict the target variable. The data distribution for continuous variable should be normalised to develop a better model. So, to have a look at the distribution of the data, Histograms are used with density curves to know how the data is distributed among the dataframe.

To predict next year losses the data must contain comparable amount of datapoints for season, Month of absence and year so as to build a time series model.

Below are the graph and plot produced for visualization:

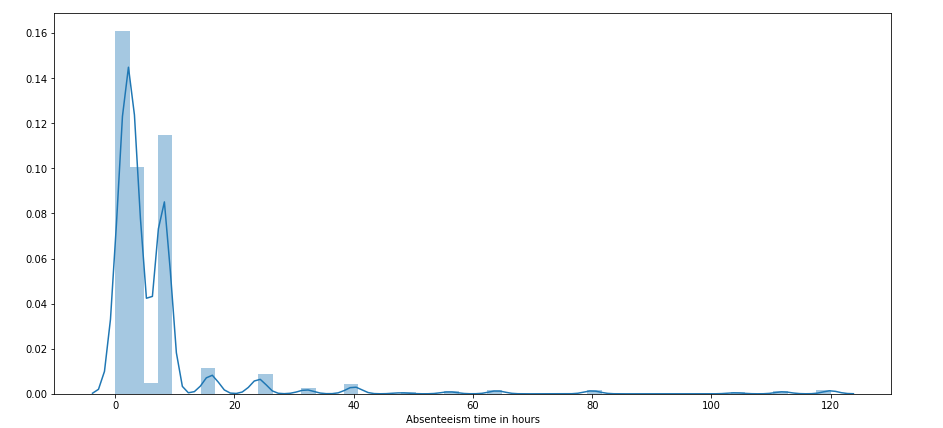




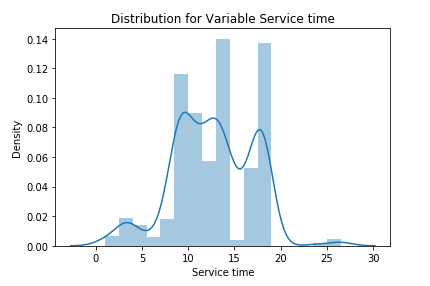
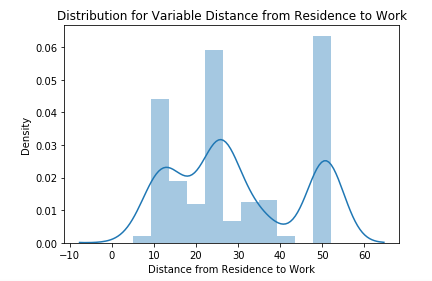


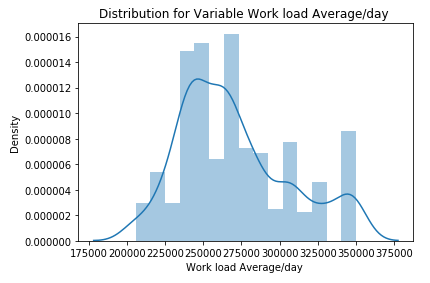
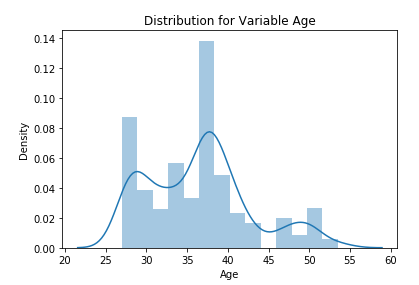
From the above graphs, it is clear that over the month, year and season the data is equally distributed.

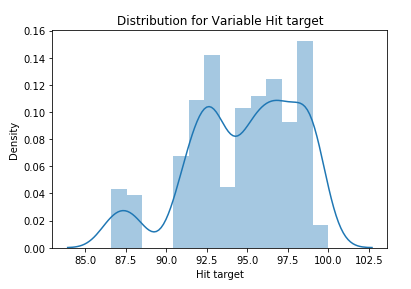
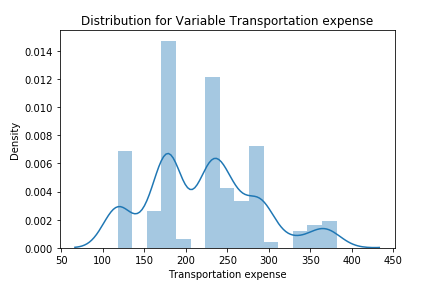
Density plot for ‘Absenteeism time in hours’ clearly depicts that data distribution is almost normalised for the values ranging from 0 to 20 hours.

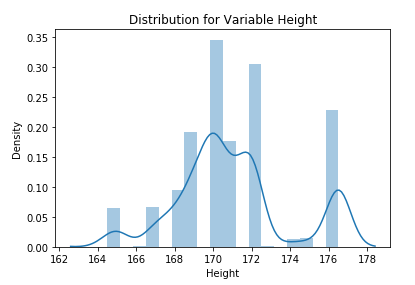
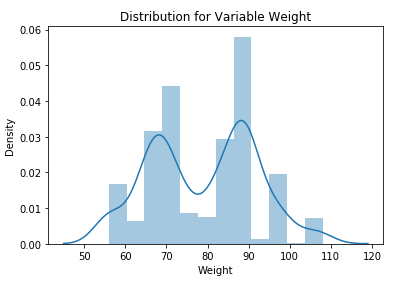


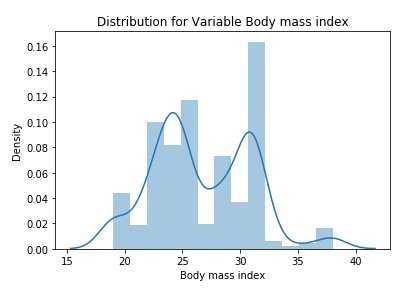
Checking for other continuous variables,







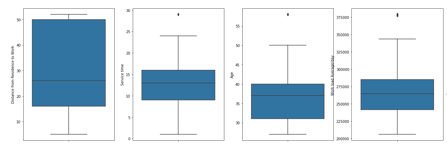


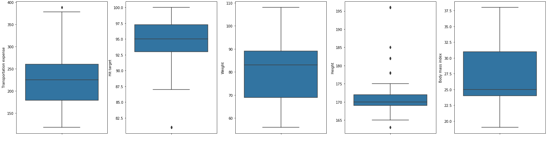


It is noted that all of the continuous variables are little skewed.

* 1. Outlier Analysis

It is observed from these probability distributions that most of the variables are skewed. The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data. One of the other steps of pre-processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers. Plots for the outliers using boxplots:





To treat these outliers, Capping method is used as because of it no datapoint will be removed and the outliers will be replaced by the value equals to the upper fence and lower fence for each boxplot of variable.

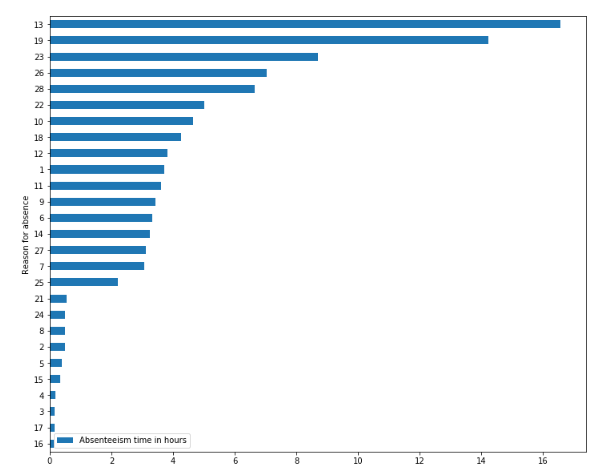
Below is given the code used in python where minimum and maximum represents lower and upper fence value respectively.



* 1. Relationship between Independent and Dependent Variable

The relationship of the independent variables vs dependent variable is our main purpose for the analysis as this will tell what is the trend of Employee absenteeism time over these parameters.

Reason for absence



Top 3 categories in order of Absenteeism time in hours are:

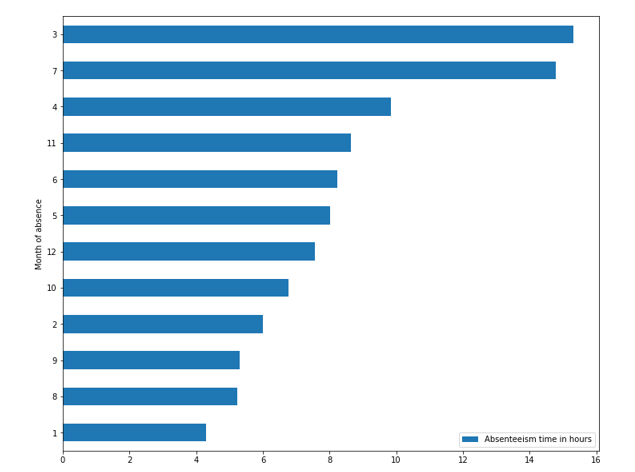
Category 13: Diseases of the musculoskeletal system and connective tissue - 16.55 % of total time

Category 19: Injury, poisoning and certain other consequences of external causes – 14.22 % of total time

Category 23: medical consultation – 8.7 % of total time

Category 26 & 28: unjustified absence & dental consultation - 7.02 % & 6.64 % 0f total time respectively

Month of absence



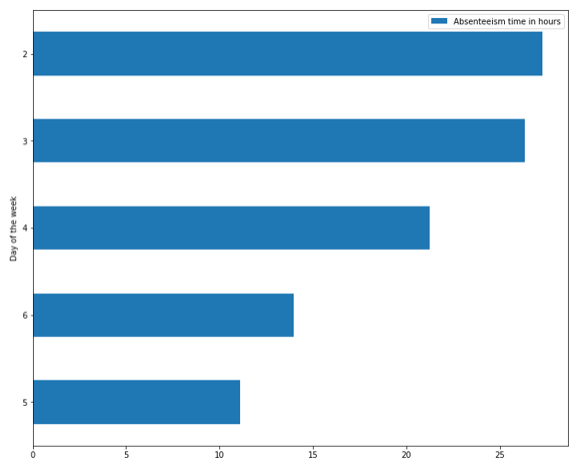
Top 3 months in order of Absenteeism time in hours are:

Month 3: March - 15.30 % of total time

Month 7: July – 14.79 % of total time

Month 4: April – 9.84 % of total time

Day of the week



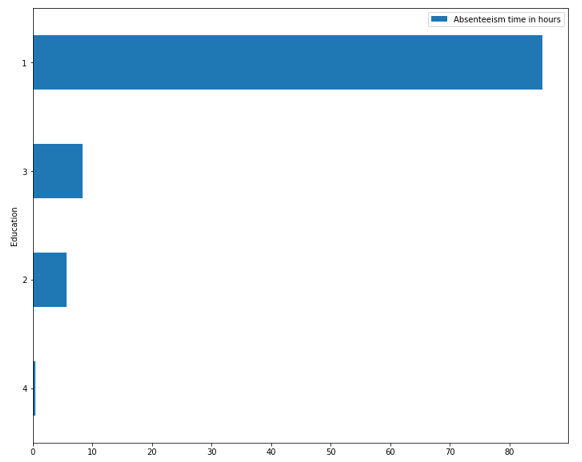
Top 3 days in order of Absenteeism time in hours are:

Day 2: Monday - 27.29 % of total time

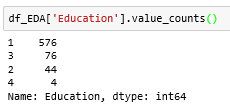
Day 3: Tuesday – 26.33 % of total time

Day 4: Wednesday - 21.27 % of total time

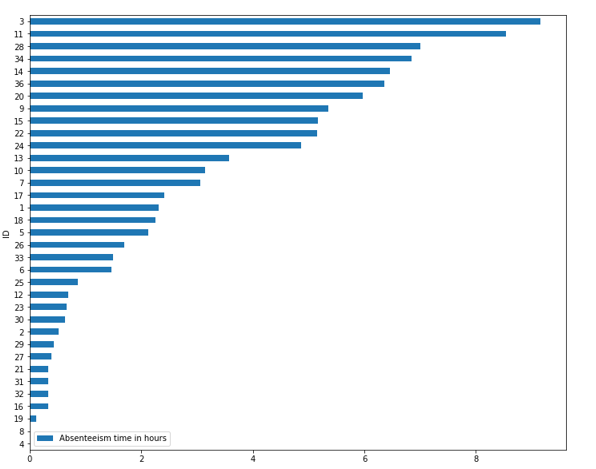
Education



85.57 % of Absenteeism time in hours is contributed by people having high school education. This may be due to majority of people having high school education (as shown below). No conclusion may be drawn from this graph.

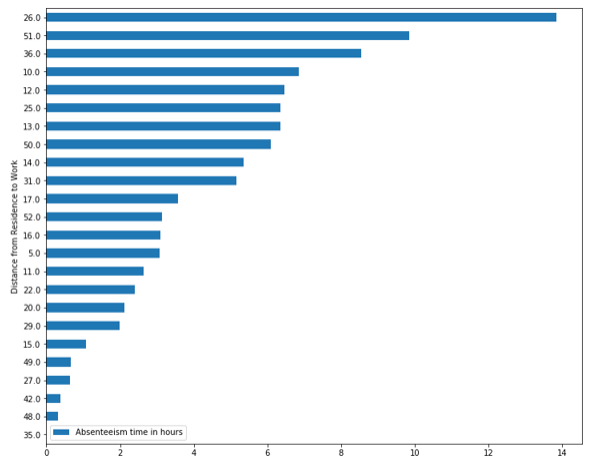


ID



Employees with ID 3, 11 & 28 are absent most of the time. Some action must be taken against them so that they won’t do it again in future.

Distance from Residence to Work

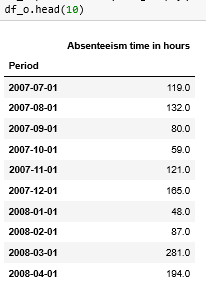


People at distance 26 km from the Workplace are more absent then those who are more than 30 km away from the workplace.

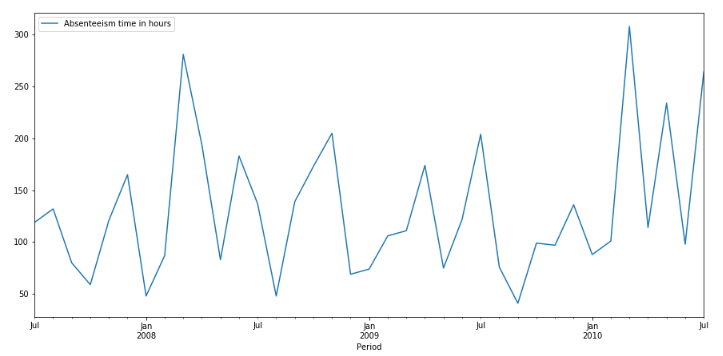
* 1. Final Input DataFrame

The main purpose of this modelling is to predict losses that the company will incur in the next year 2011.   
So, the data is being aggregated at Period level with Absenteeism time in hours to produce a time series data and matching that trend the model will be built upon that.

Now, the dataframe has the required predictor variable ‘Period’. The final dataframe “df\_o” is of size 37x1.



The dataframe contains one outlier which is treated with outlier capping method to be in the bounds and don’t produce abnormal results.



1. **Modelling**

The type of model needs to be implemented mainly depends upon the target variable data type, predictor variable’s data type and type of relationship they carry with each other.

The target variable is continuous variable and the predictor variable is of timestamp type. As the target variable is a timestamp variable discussed below machine learning model would be a better prediction model for this problem.

**ARIMA (Auto-Regressive Integrated Moving Averages)**

ARIMA is a very popular statistical method for time series forecasting. ARIMA stands for Auto-Regressive Integrated Moving Averages. ARIMA models work on the following assumptions –

1. The data series is stationary, which means that the mean and variance should not vary with time. A series can be made stationary by using log transformation or differencing the series.
2. The data provided as input must be a univariate series, since arima uses the past values to predict the future values.

ARIMA has three components – AR (autoregressive term), I (differencing term) and MA (moving average term). Let us understand each of these components –

AR term refers to the past values used for forecasting the next value. The AR term is defined by the parameter ‘p’ in arima. The value of ‘p’ is determined using the PACF plot.

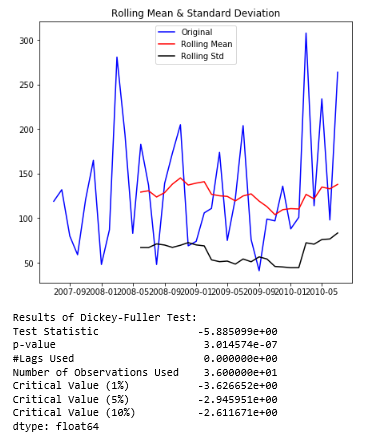
MA term is used to defines number of past forecast errors used to predict the future values. The parameter ‘q’ in arima represents the MA term. ACF plot is used to identify the correct ‘q’ value.

Order of differencing specifies the number of times the differencing operation is performed on series to make it stationary. Test like ADF and KPSS can be used to determine whether the series is stationary and help in identifying the d value.

Here to forecast the losses that will going to be happen in year 2011, this model will be implemented as it is one of the best model to forecast the coming future.

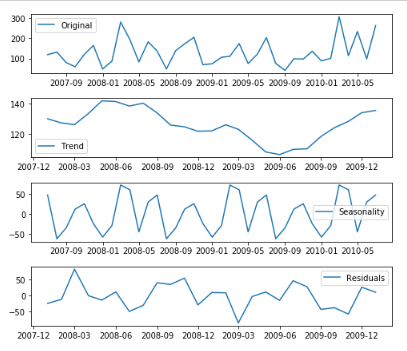
To proceed with the modelling the very first step would be checking the data is stationary or not. Stationary data doesn’t increase or decrease with respect to time over a longer period and it should have some sort of periodicity with time. To check the stationarity of the data Augmented Dickey Fuller test is performed.

Null Hypothesis for this test is that the data is non-stationary.



As it can be seen that ADF or Dickey Fuller Test statistics (i.e., -5.8) which is less than Critical value 10%, 5% as well as 1%. Hence null hypothesis can be rejected stating that the dataset is stationary in nature and ARIMA model can be applied to it without performing any extra calculations.

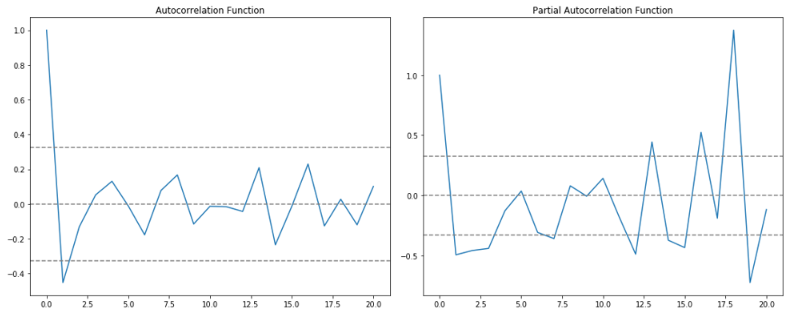
Below is the plot for trend, seasonality and residuals of the data over time period of June, 2007 to June 2010.



ARIMA stands for Auto-Regressive Integrated Moving Averages. The ARIMA forecasting for a stationary time series is nothing but a linear (like a linear regression) equation. The predictors depend on the parameters (p,d,q) of the ARIMA model:

1. Number of AR (Auto-Regressive) terms (p): AR terms are just lags of dependent variable. For instance if p is 5, the predictors for x(t) will be x(t-1)….x(t-5).
2. Number of MA (Moving Average) terms (q): MA terms are lagged forecast errors in prediction equation. For instance if q is 5, the predictors for x(t) will be e(t-1)….e(t-5) where e(i) is the difference between the moving average at ith instant and actual value.
3. Number of Differences (d): These are the number of nonseasonal differences, i.e. in this case we took the first order difference. So either we can pass that variable and put d=0 or pass the original variable and put d=1. Both will generate same results.

To determine p and q, PACF(Partial Auto Correlation Function) and ACF(Auto Correlation Function) plots are observed respectively and whatever value the line is crossing the upper confidence interval will be the values for p and q.

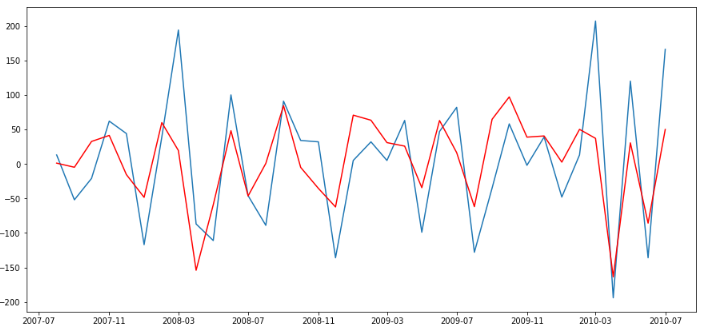


It can be seen that the PACF curve is crossing the upper confidence interval at 1 and ACF curve is also crossing at 1. Therefore, p and q both will be assigned value equals to 1.

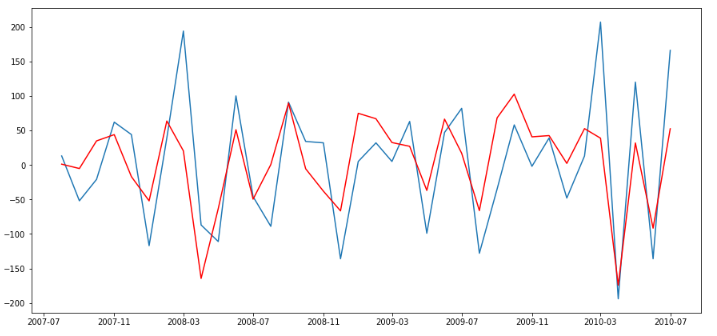
Plotting by building AR model only on the data:



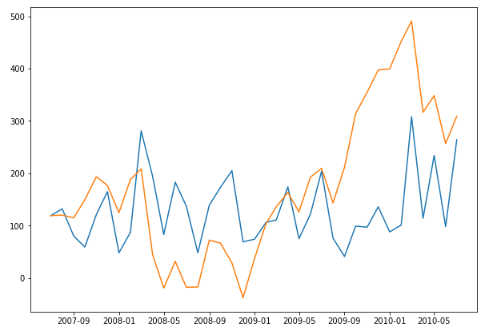
Plotting by building MA model only on the data:



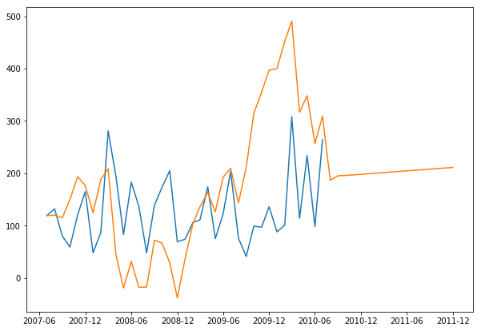
Plotting the ARIMA model with results:



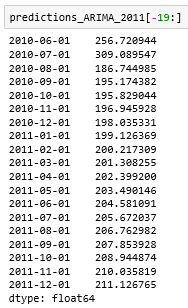
Predicted values the with ARIMA model with original value plot (orange-Predicted, blue -Original):



Plot for Forecasting ‘Absenteeism time in hours’ for the year 2011:



Values for ‘Absenteeism time in hours’ are given as below:



1. **Conclusion**

ARIMA model is one of the best Time series models that is used in Machine learning for forecasting the future as the question being asked in the problem was to predict the losses that are going to happen in the coming future, year 2011 and the data provided was till June, 2010. From the Prediction plot it can be seen that the predicted curve is following the same trend of increase in Absenteeism time in hours. And also, the predicted data pretty much align with the original data provided.

**Answers to the asked Questions**

**Q1. What changes company should bring to reduce the number of absenteeism?**

Ans. When the reason for absence is analysed it was found out that most of the employees are being medically unable to come to work place. Some major solution and remedies to that would be:

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

When closely observed with particular employees it was found out that employee with ID 3, 11 & 28 are taking most of the leaves, so regarding to that they should be consulted.

Also, employees who have distance from residence to work equals to 26 km are taking more leaves than those who are coming from more than 30 km. This might be due to the route from they are coming from. So, they need to be advised to relocate such that they would come to office daily.

**Q2. How much losses every month can we project in 2011 if same trend of absenteeism continues?**

Ans. With the ARIMA model forecasts the company can expect Absenteeism time in hours equals to 2461 hrs.