Employee Absenteeism

November 26, 2018

1 Chapter 1

1.1 Introduction

1.1.1 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?

Input Data *Absenteeism_at_work_Project.xls*

Output Data *Solution to given problems, i.e. inferences based on analysis and losses we can project in 2011 if same trend of absenteeism continues*

1.1.2 Data

Our task is to analyse the given data and derive meaningful inferences. Given below is a sample of the data set that we are using:

Absenteeism_at_work_Project Sample Data (1-5 columns)

ID	Reason for absence	Month of absence	Day of the week	Seasons
11	26	7	3	1
36	0	7	3	1
3	23	7	4	1
7	7	7	5	1
11	23	7	5	1

Absenteeism_at_work_Project Sample Data (6-10 columns)

Transportation expense	Distance	Service time	Age	Work load Average/day
289	36	13	33	239,554
118	13	18	50	239,554
179	51	18	38	239,554
279	5	14	39	239,554
289	36	13	33	239,554

Absenteeism_at_work_Project Sample Data (11-15 columns)

Hit target	Disciplinary failure	Education	Son	Social drinker
97	0	1	2	1
97	1	1	1	1
97	0	1	0	1
97	0	1	2	1
97	0	1	2	1

Absenteeism_at_work_Project Sample Data (16-21 columns)

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

Hence, we have below 21 predictors which we will use to analyse the Absenteeism trend with the employees.

Predictors

S. No.	Predictors	Type
1.	ID	Categorical
2.	Reason for absence	Categorical
3.	Month of absence	Categorical
4.	Day of the week	Categorical
5.	Seasons	Categorical
6.	Transportation expense	Continuous
7.	Distance from Residence to Work	Continuous
8.	Service time	Continuous
9.	Age	Continuous
10.	Work load Average/day	Continuous
11.	Hit target	Continuous
12.	Disciplinary failure	Categorical
13.	Education	Categorical
14.	Son	Continuous
15.	Social drinker	Categorical
16.	Social smoker	Categorical
17.	Pet	Continuous
18.	Weight	Continuous
19.	Height	Continuous
20.	Body mass index	Continuous
21.	Absenteeism time in hours	Continuous

2 Chapter 2

2.1 Exploratory Data Analysis

2.1.1 Univariate Analysis

Exploratory data analysis is most important step before we can apply any machine learning model. It specifically cleans up the data for the model, so that the model can work as expected from it, and not get any unexplained predictions. It involves various steps which are explained below.

Similarly for more complex exploratory data analysis, we perform following steps to filter out variables and make them model ready.

Missing Values Analysis

What are missing values? In real world, the data are not always complete. There are numerous times when we get some observations with one or many variables values as missing. These data are not helpful in creating model and doing analysis. Hence, they need to be taken care of at the start of model making, or in the pre-processing stage.

In our data, we found some variables with missing data. It can be observed from below missing values analysis table.

Missing Values count for each variable

Variables	Missing_percentage
Body mass index	4.189189
Absenteeism time in hours	2.972973
Height	1.891892
Work load Average/day	1.351351
Education	1.351351
Transportation expense	0.945946
Son	0.810811
Disciplinary failure	0.810811
Hit target	0.810811
Social smoker	0.540541
Age	0.405405
Reason for absence	0.405405
Service time	0.405405
Distance from Residence to Work	0.405405
Social drinker	0.405405
Pet	0.270270
Weight	0.135135
Month of absence	0.135135
Seasons	0.000000
Day of the week	0.000000
ID	0.000000

Code used to evaluate missing values in the data:

```
#Create dataframe with missing percentage
missing_val = pd.DataFrame(dataset.isnull().sum())
```

```
#Reset index
missing_val = missing_val.reset_index()

#Rename variable
missing_val = missing_val.rename(columns = {'index': 'Variables', 0: 'Missing_percentage'})
```

Hence we need to apply some algorithm to fill the missing variables for the above variables, as we can't drop any variable because the missing values for each variable is less than 30% (which could be threshhold for dropping a variable with missing values in a variable).

After appling various imputation techniques, KNN imputation algorithm gives us the best and closest results. Therefore, we will be using KNN imputation with k=3 to impute missing values in our given data.

Also we observed that observations where *Reason of absence* is 0 (not a code), *Absenteeism time in hours* is also 0 and vice versa. We used this information too to fill many missing values.

Outlier Analysis

What are outliers? Basically when the observations has some inconsitent data with the rest of dataset. It can be caused by number of things like poor data quality or contamination, low quality measurements, malfunction equipment and manual errors. Sometimes they are correct data but some exceptional cases.

It can be detected by lots of ways, some of which are: * Graphical Tools * Box plot * QQ Plot * Scatter Plots * Statistical Techniques * Grubb's Technique (It will only work on those data which are uniformly distributed.)

We are going to use the graphical tools to detect if we have any outliers in our data. In our case we used Box plots to analyze the data and observed below analysis: * For the variable *Absenteeism time in hours*, we observed that for a given day, the number of hours are greater than 24 hours for some observations, which should not possible. These data are 32, 40, 48, 56, 64, 80, 104, 112, 120. Hence we replaced this data with nan and applied KNN imputation to impute these observations.

Box Plots

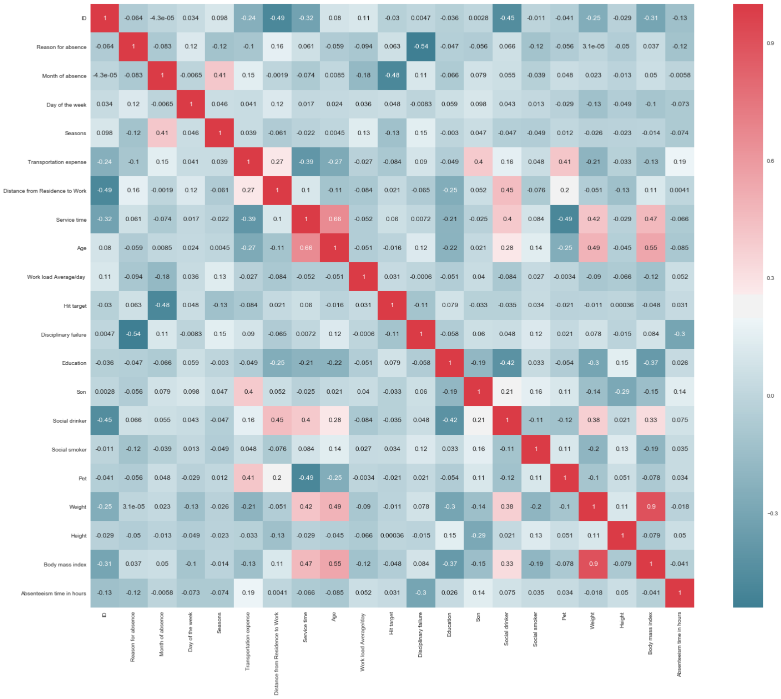
Refer to Appendix 1A (Code and figure)

2.2 Correlation plot

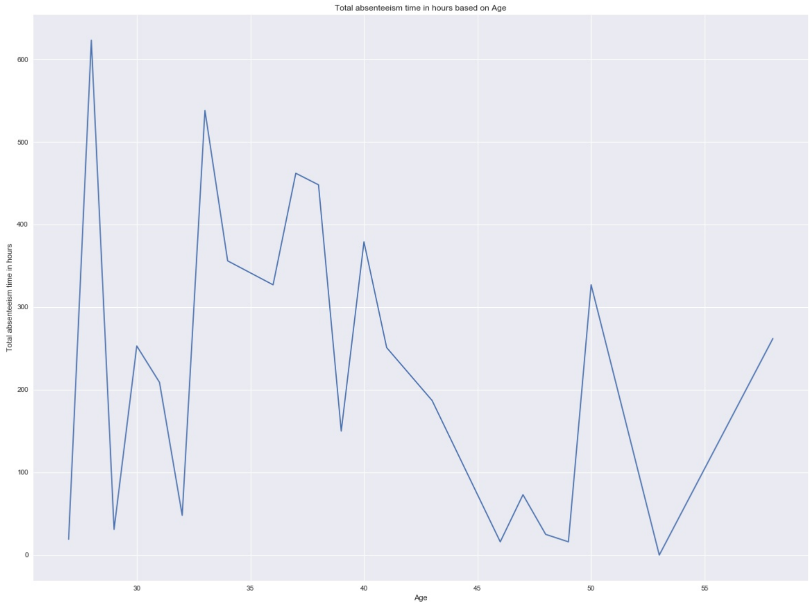
The correlation plot for the given data can be seen as below:

2.2.1 Plots to explain the reason of Absenteeism

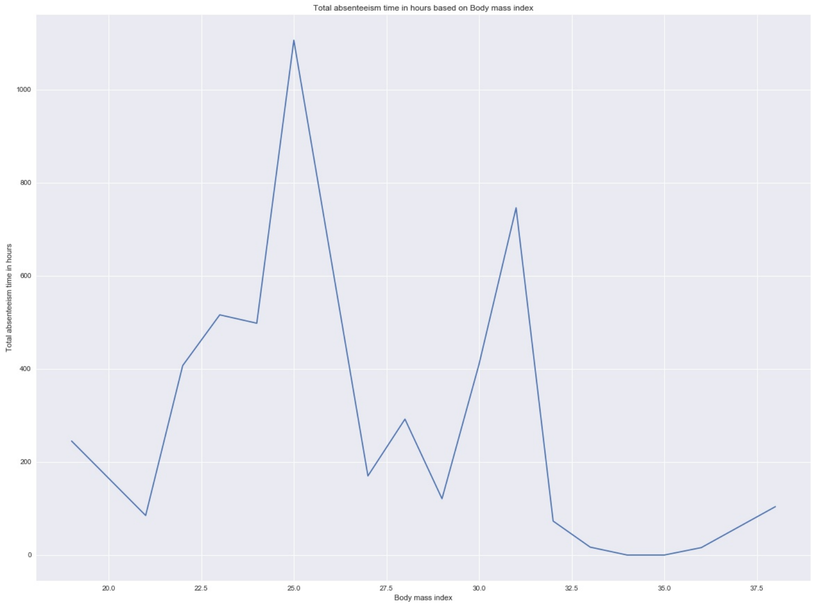
We have gone through various plots to determine the reasons of Absenteeism. These plots include bar graphs and line graphs as follows:



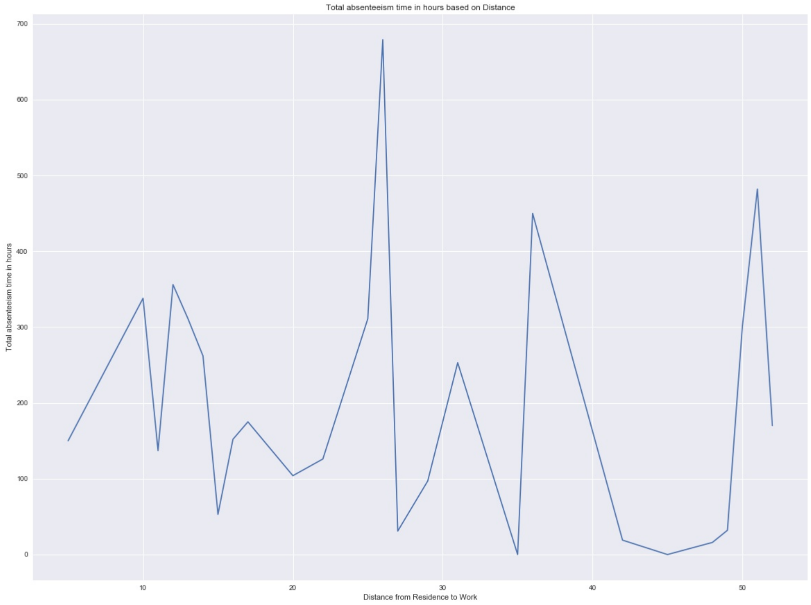
CorrelationAnalysis.jpg

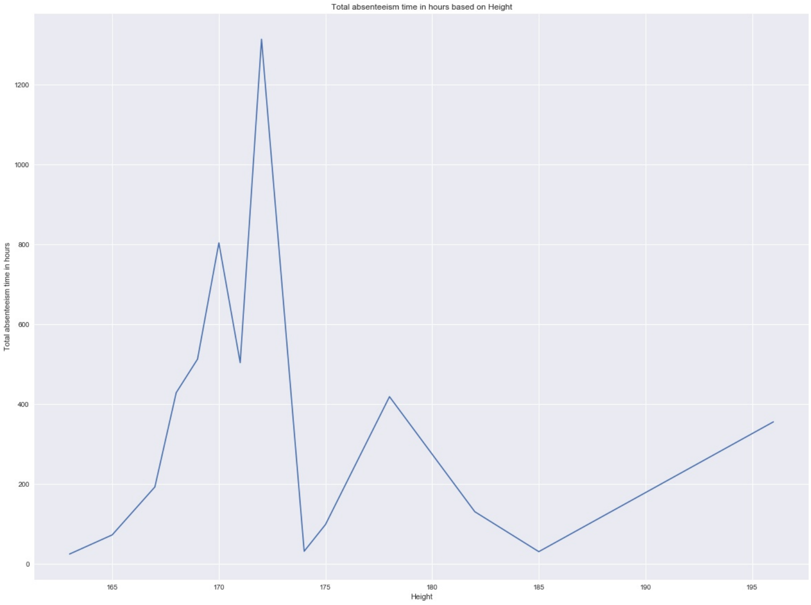


mass

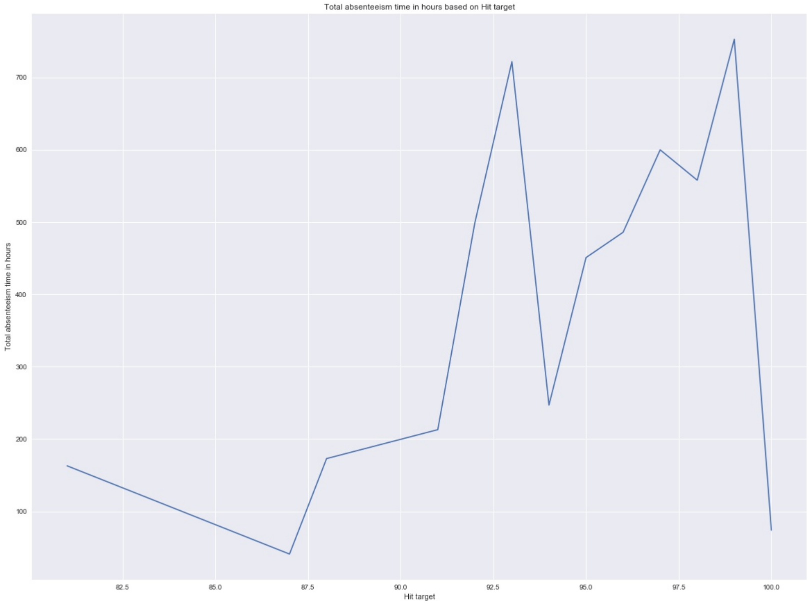

Total absenteeism time in hours based on Body mass index
Total absenteeism time in hours
0
200
400
600
800
1000
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
Body mass index

index.bb

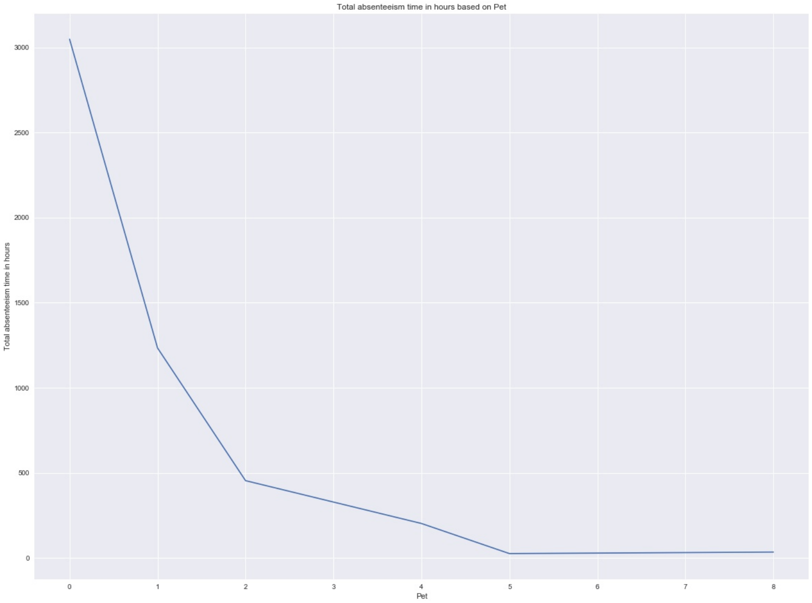

Total absenteeism time in hours based on Distance
Total absenteeism time in hours
0
100
200
300
400
500
600
700
10
20
30
40
50
Distance from Residence to Work


Total absenteeism time in hours based on Height
Total absenteeism time in hours
0
200
400
600
800
1000
1200
165
170
175
180
185
190
195
Height

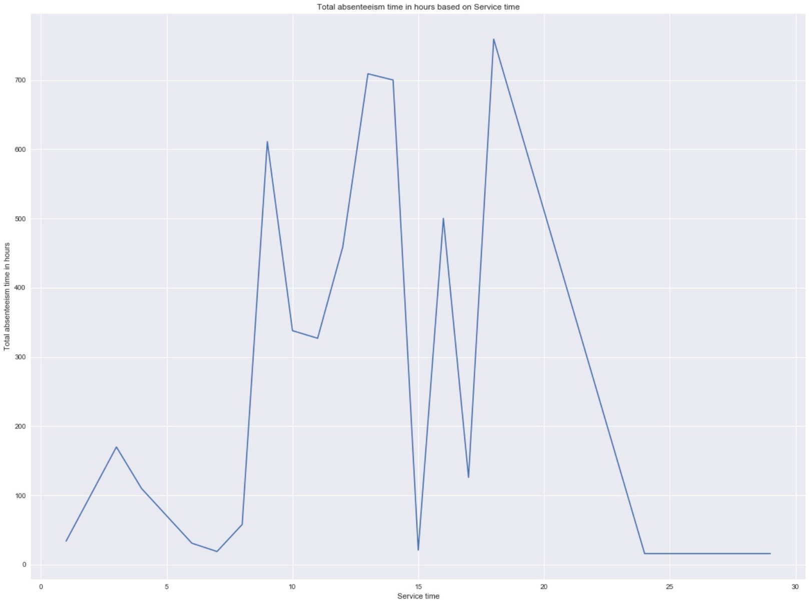
tar-

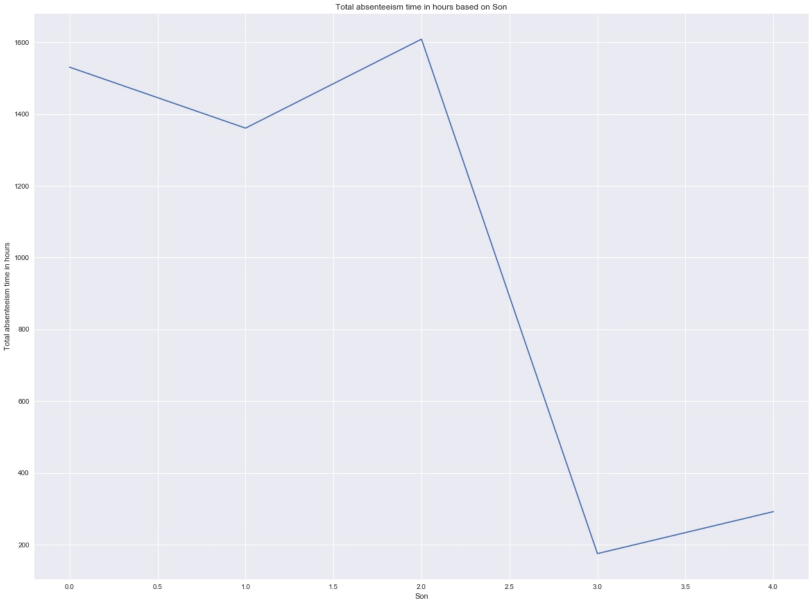

Total absenteeism time in hours based on Hit target
Total absenteeism time in hours
100
200
300
400
500
600
700
82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
Hit target

get.bb

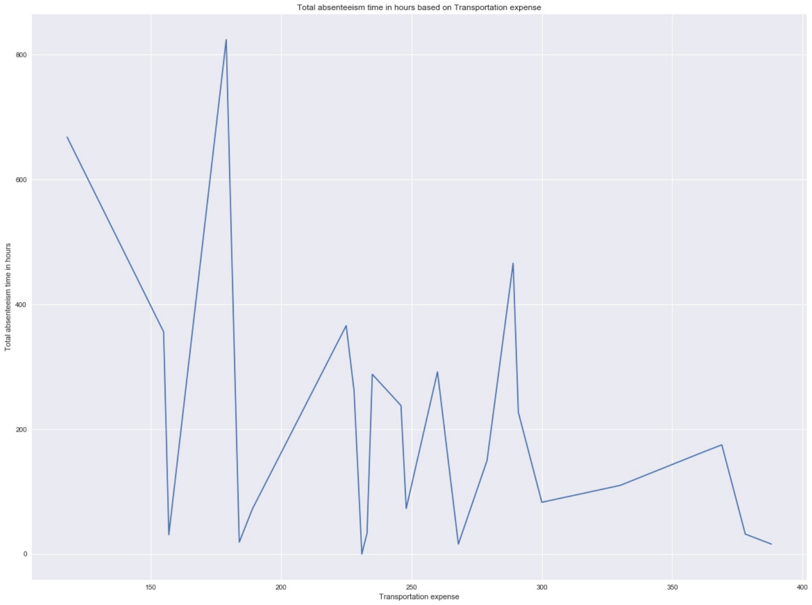

Total absenteeism time in hours based on Pet
Total absenteeism time in hours
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
Pet

time.bb

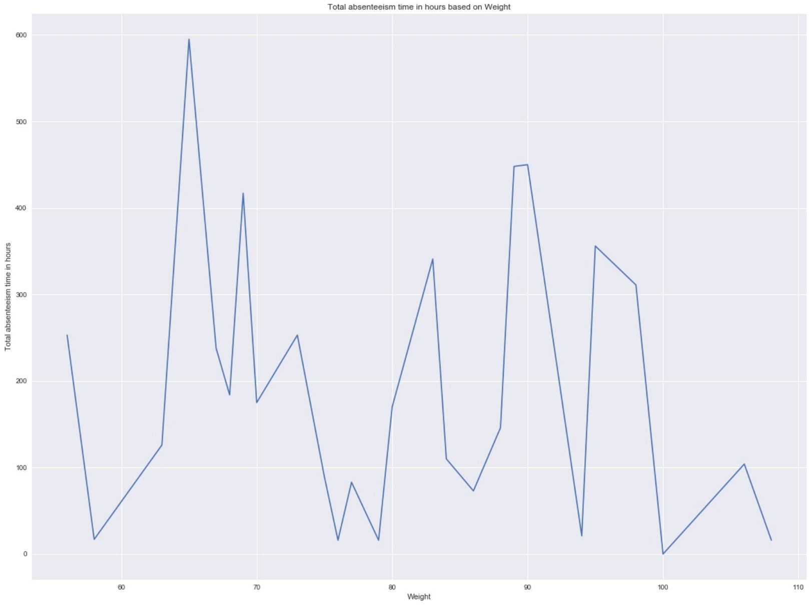

Total absenteeism time in hours based on Service time
Total absenteeism time in hours
700
600
500
400
300
200
100
0
0
5
10
15
20
25
30
Service time


Total absenteeism time in hours based on Son
Total absenteeism time in hours
1600
1400
1200
1000
800
600
400
200
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Son

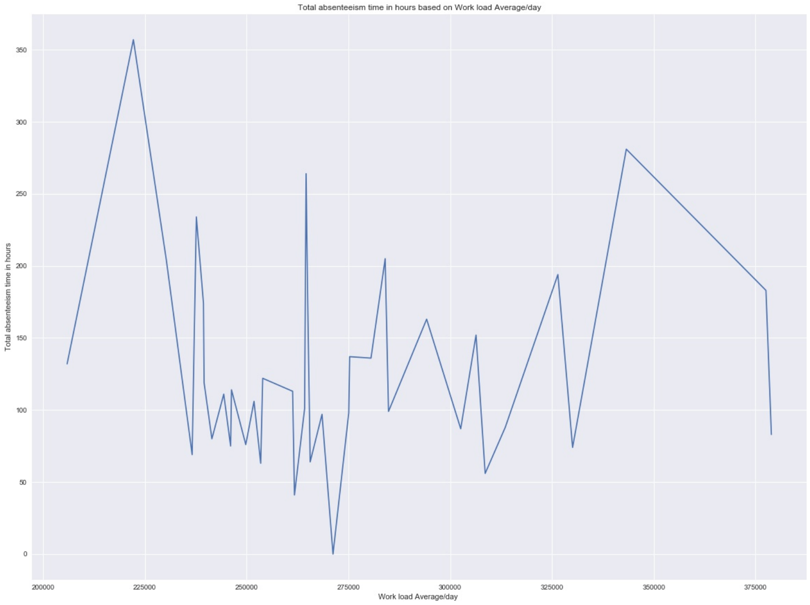
ex-

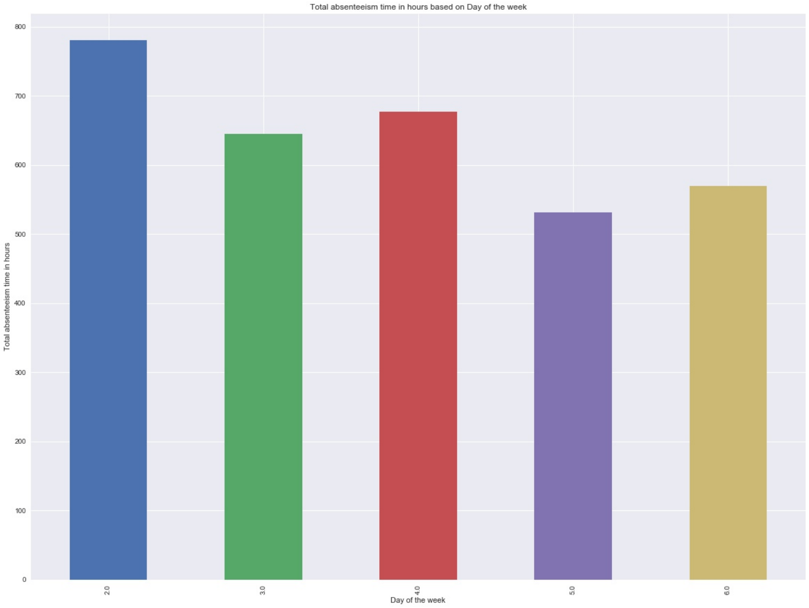

Total absenteeism time in hours based on Transportation expense
Total absenteeism time in hours
800
600
400
200
0
150
200
250
300
350
400
Transportation expense

pense.bb

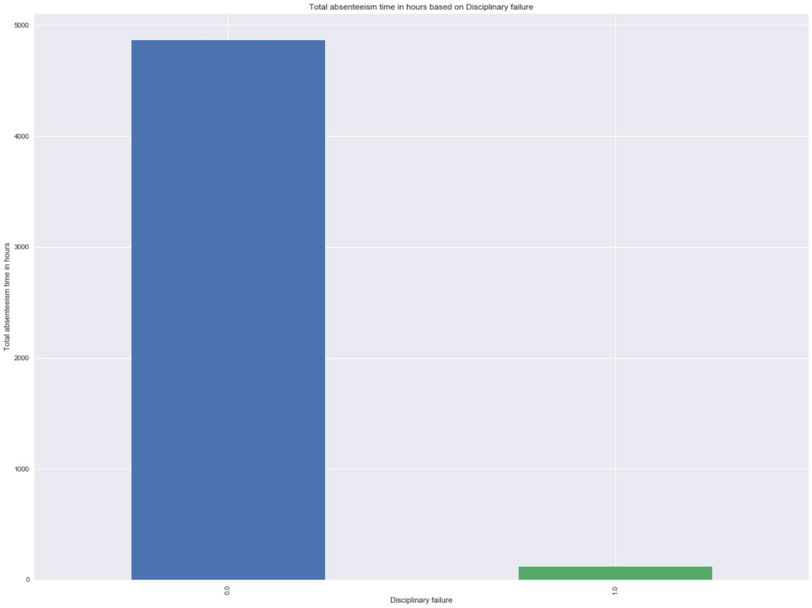

Total absenteeism time in hours based on Weight
Total absenteeism time in hours
0
100
200
300
400
500
600
60
70
80
90
100
110
Weight

load.bb

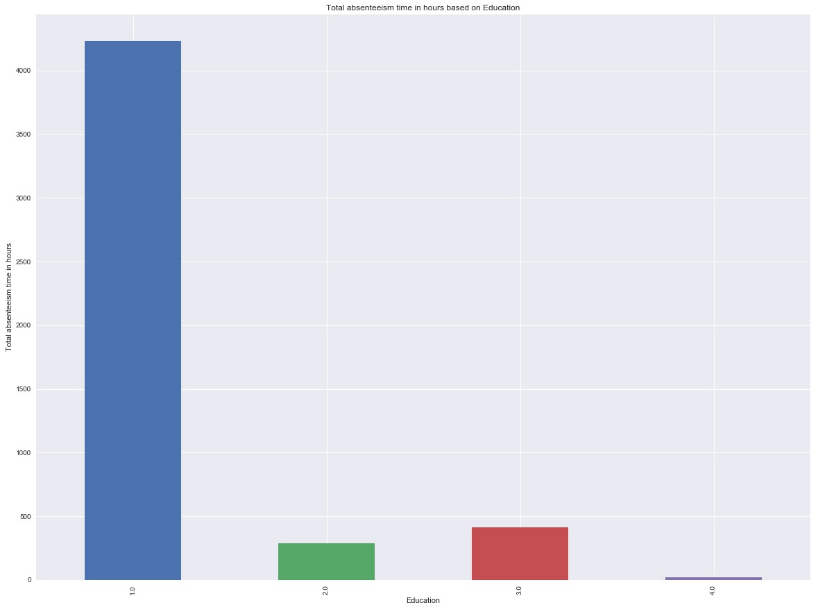

Total absenteeism time in hours based on Work load Average/day
Total absenteeism time in hours
0
50
100
150
200
250
300
350
200000
225000
250000
275000
300000
325000
350000
375000
Work load Average/day

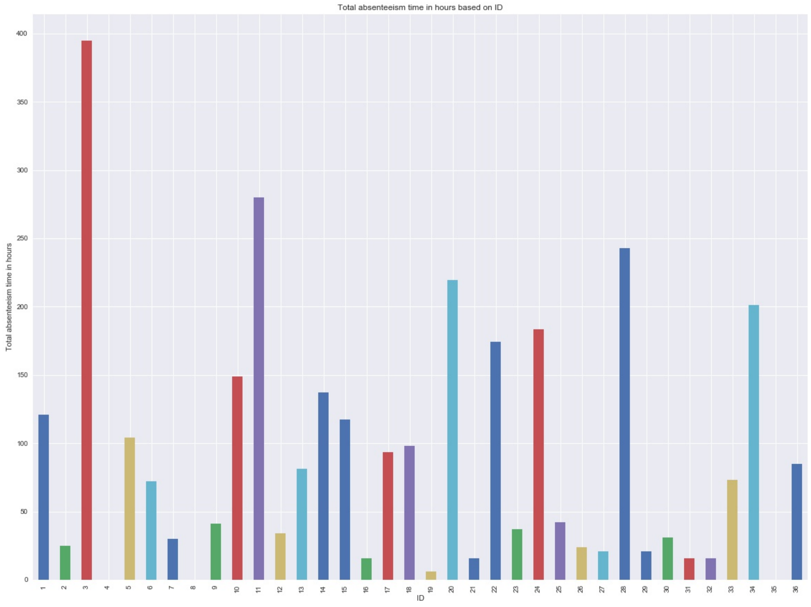

Total absenteeism time in hours based on Day of the week
800
700
600
500
400
300
200
100
0
Total absenteeism time in hours
2.0
3.0
4.0
5.0
6.0
Day of the week

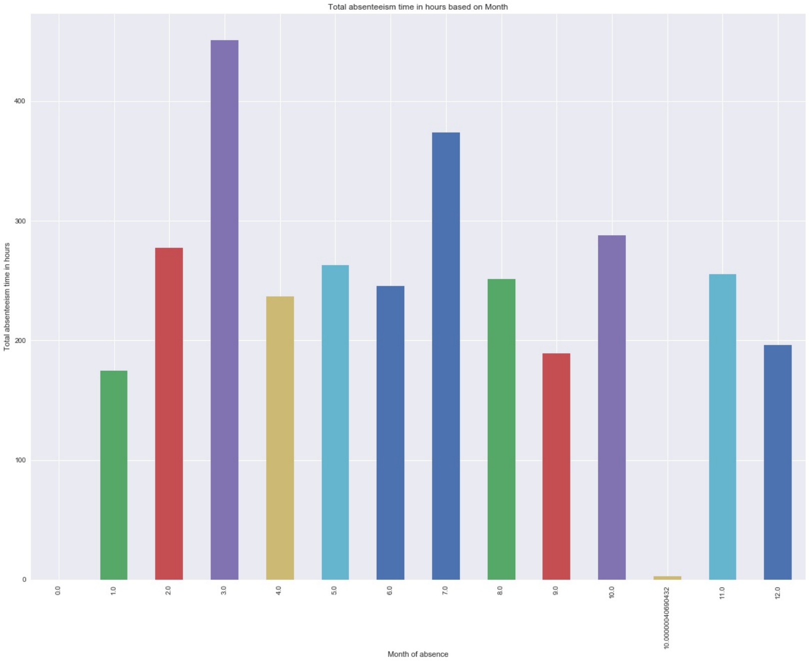
fail-

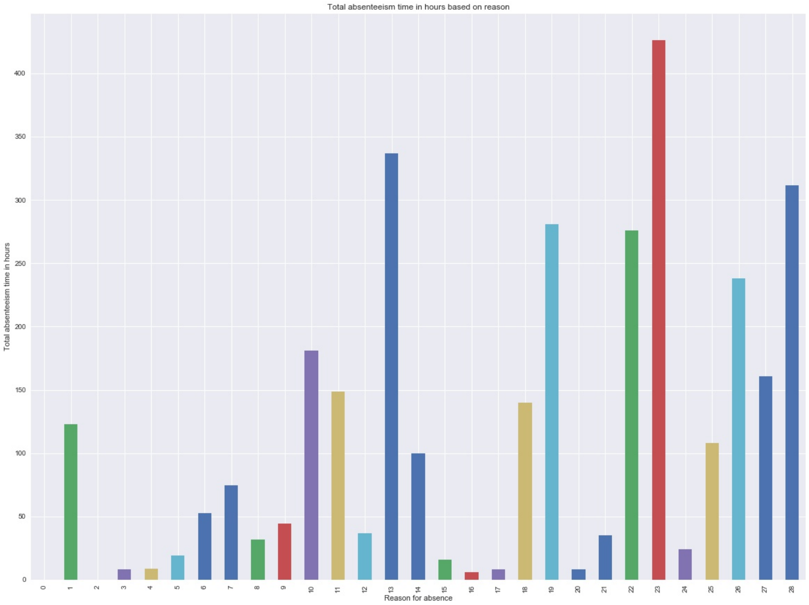

Total absenteeism time in hours based on Disciplinary failure
5000
4000
3000
2000
1000
0
Total absenteeism time in hours
0.0
1.0
Disciplinary failure

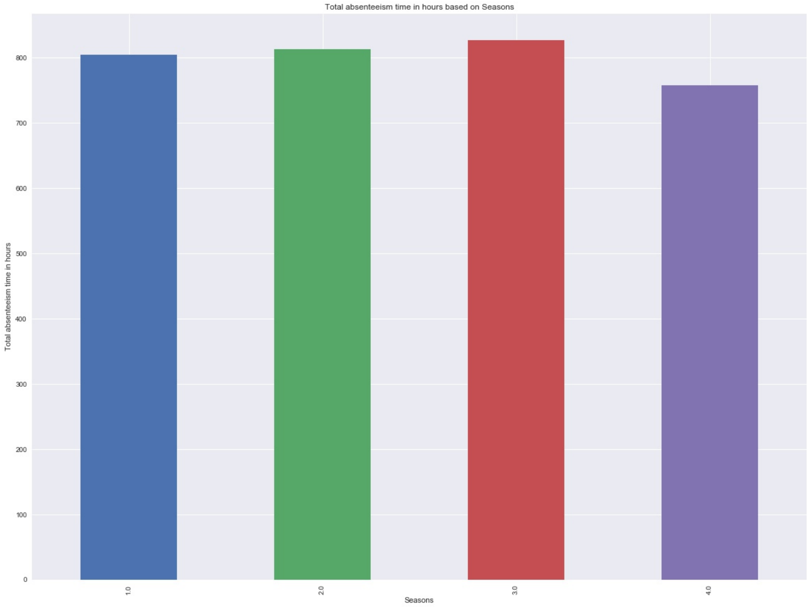
ure.bb


Total absenteeism time in hours based on Education
Total absenteeism time in hours
0
500
1000
1500
2000
2500
3000
3500
4000
1.0
2.0
3.0
4.0
Education

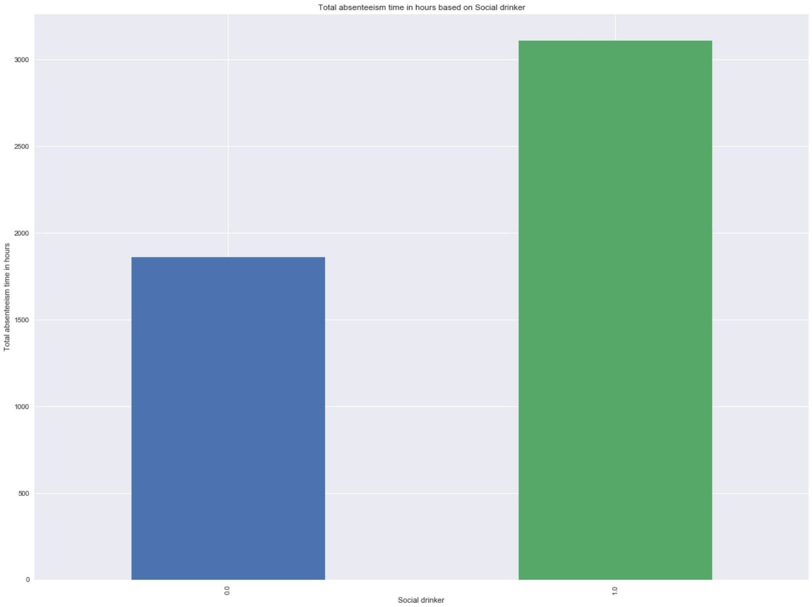

Total absenteeism time in hours based on ID
Total absenteeism time in hours
0
50
100
150
200
250
300
350
400
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
ID


Total absenteeism time in hours based on Month
Total absenteeism time in hours
0
100
200
300
400
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
10.000000406690432
11.0
12.0
Month of absence

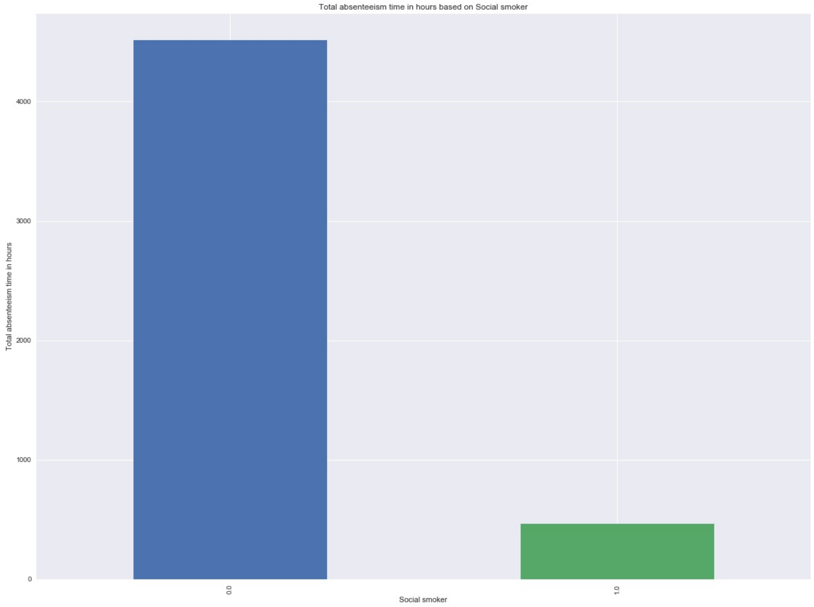

Total absenteeism time in hours based on reason
Total absenteeism time in hours
0
50
100
150
200
250
300
350
400
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
Reason for absence



drinker.bb



smoker.bb



From all the above given graphs, following inferences can be made to get a clear picture for getting the reasons for absenteeism: * On start of the week, i.e. Mondays, most people tend to take leaves.

- People, who don't properly face Disciplinary actions tend to take more leaves.
- Employees who are High School graduates take more leaves compare to other employees with higher education.
- In March, the number of employees taking leaves are more compared to other months.
- Number of leaves decrease with owning more pets.
- Employees take more leaves for medical consulation than any other ICD code, i.e. code 23.
- Social drinker are more prone to be absent, than those who don't drink.
- Social smoker are better able to come to work than those who don't smoke.

With above inferences, following solutions can be provided to cope with Absenteeism. Each of the below solutions corresponds to the above given inference: * Since it's the day after weekends, Mondays are always blue. This is major issue faced by the industry. It can be solved by reducing the number of hours reduces on Mondays and compensating them in middle days of the week. This will encourage the employees to come to the office.

- The company needs to be stricter, while implementing Disciplinary actions against the employees who are defaulting. This will create positive drive among the workers to book target accomplishments.

- High School graduates are not professionals as such, i.e. they don't have industry experience, since they just got out of school environment. The company may provide 2-3 months training programmes to inculcate them with professional behaviour. Then they may have better understanding of the work culture.

- In most of the countries, March is the end of financial year. This involves documenting various tax and finance related files and long queues of tax filings. This can be reduced if company can provide tax and finance consultations in the company itself, so that employees don't have to go out of the office for consultations.

- In many research studies, it has been implied that pets are great stress reliever. Hence owning more pets, may help employees relax properly at home, so they can perform with full efficiency at work. Hence company can encourage the employees to own a pet, or may involve employees to get into NGO where they take care of abandoned animals. This can help the employees to relax more and work efficiently.

- For reasons like medical consultations (Code 23), employees may not have to provide official consultations slips to apply for leave. This can create havoc as company have no way to track all these leaves for their authenticity. This can be solved if the companies can provide free medical consultations to all it's employees every quater or every two quaters.

- Drinking creates lots of health risk, specially after the day people consume drinks, like hangovers and head aches. This compel the employees to take unplanned extra leaves, as they usually don't find themselves able to go to office the next day. Company can start programs to inform it's employess of the ill effects of consumption of liquor and may provide with activities to reduce the stress levels through other means.

- According to research, cigarettes help to reduce stress levels of persons, as it has nicotine that reduces brain stress. This indicates that employees who are less stressed tend to come to company more. Company can provide stress reliver activities to reduce tention among it's employees and may encourage it's employees to involve in such activities rather than having harmful and ill effects of tobacco.

2.3 Data Preparation

Now for the second problem, we need to estimate Absenteeism hours for the year of 2011. For that, we have *Month* data from July 2007 to July 2010. However, to apply it for Linear Regression we need to convert the Month data to numerical data, as we can't fit the model with Date time data.

Hence we have added a column of *Year* in the dataframe and will take out sum of each month by grouping *Month* & *Year* data and create a seperate dataframe with the data.

Now we will apply different models to forcast the data for the year 2011.

3 Chapter 3

3.1 Model Selection

In exploratory analysis stage, we arrived at the conclusion that the problem requires **Regression Model** and **Time series Model** to predict our dependent variable, i.e. *Absenteeism.*

We have applied multiple classification model which will be explained in below sections. For choosing a regression model, we are going to use metrics like MSE (mean squared error), MAE (mean absolute error), RMSE (root mean square error).

3.1.1 Linear Regression

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

In our case, we have applied linear regression with one independent variable, i.e. Months and one dependent variable i.e. Absenteeism in hours.

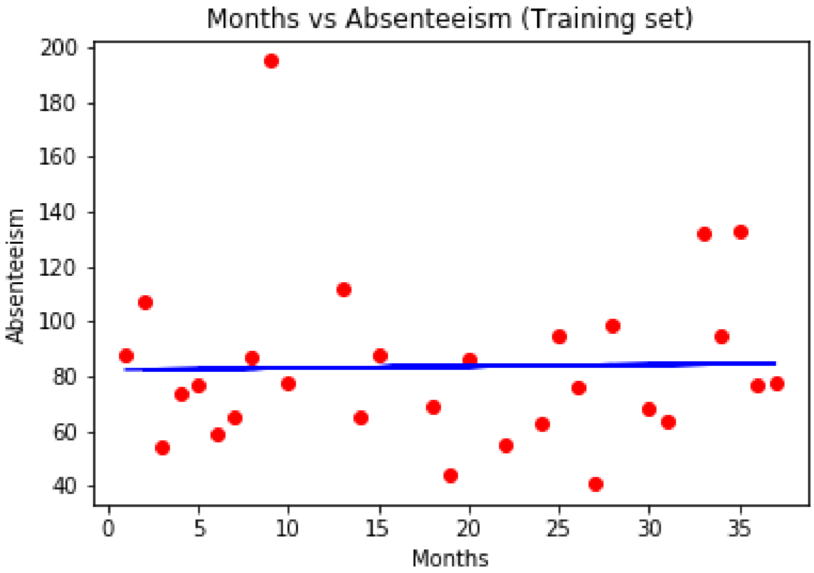
After applying the model in our cleaned data, we can calculate below metrics for the model:

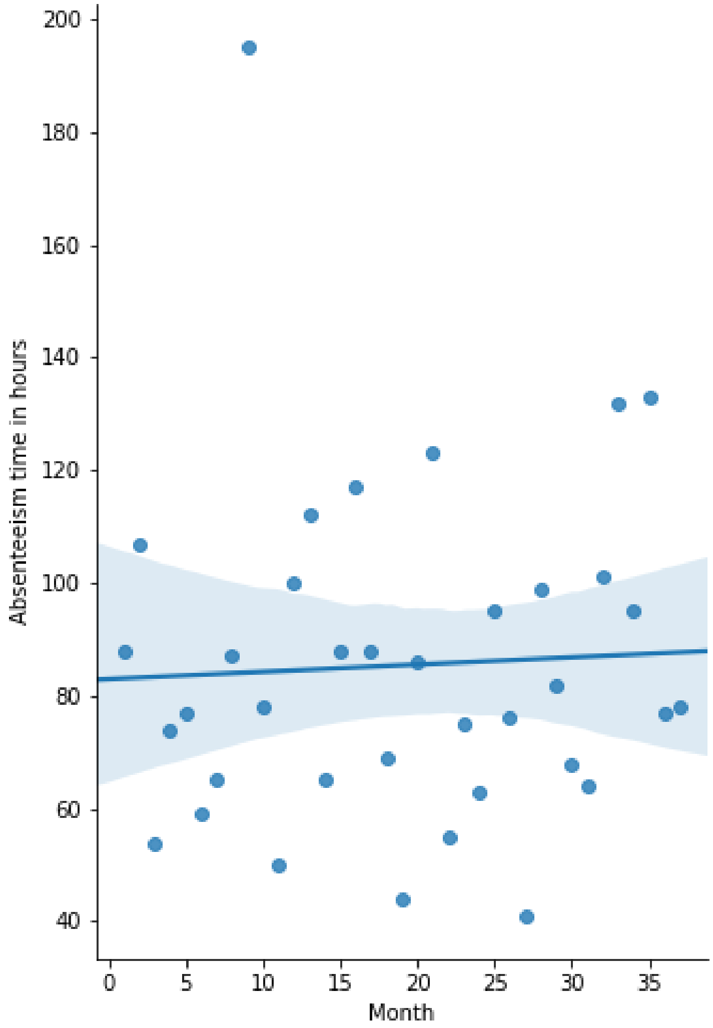
ModelName	MSEˆ	MAEˆ	RMSEˆ
Linear Regression	553.434	19.37	23.52

ˆThe meaning of each metrics are explained in Conclusion. Please refer to it for more details

These metrics tell us that model did not performed well on the data. Let's look into how other models perform on the given data.

The performance of the model can be viewed with below plots.





3.1.2 Polynomial Regression

In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x.

For our data, the best performance is given when n=5, that is 5th degree polynomial equation.

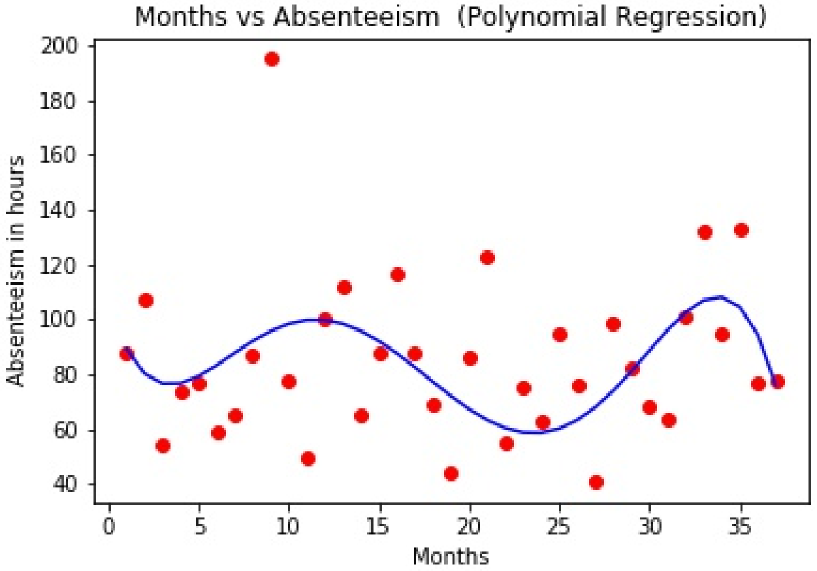
After applying the model in our cleaned data, we can calculate below metrics for the model:

ModelName	MSEˆ	MAEˆ	RMSEˆ
Polynomial Regression	898.78	20.36	29.97

ˆThe meaning of each metrics are explained in Conclusion. Please refer to it for more details

These metrics tell us that this model did not performed well on the data. Let's look into how other time series models perform on the given data.

The performance of the model can be viewed with below plots.



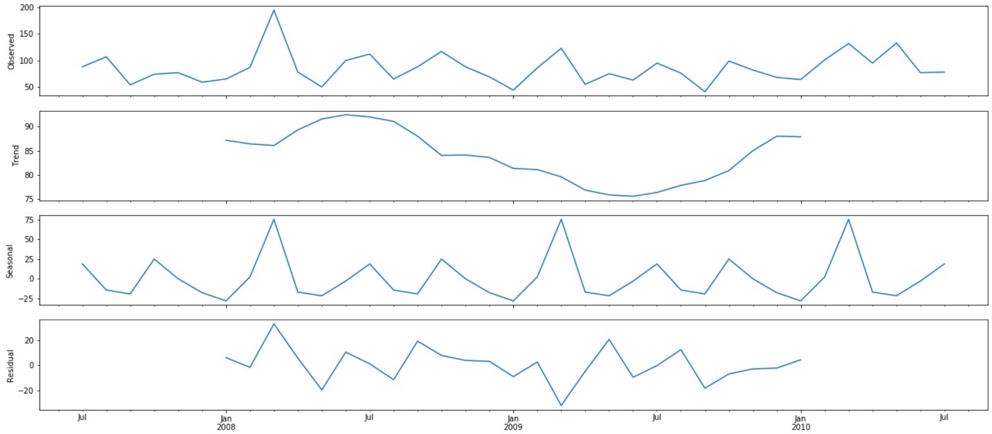
graph3.jpg

3.1.3 Time series analysis - ARIMA

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

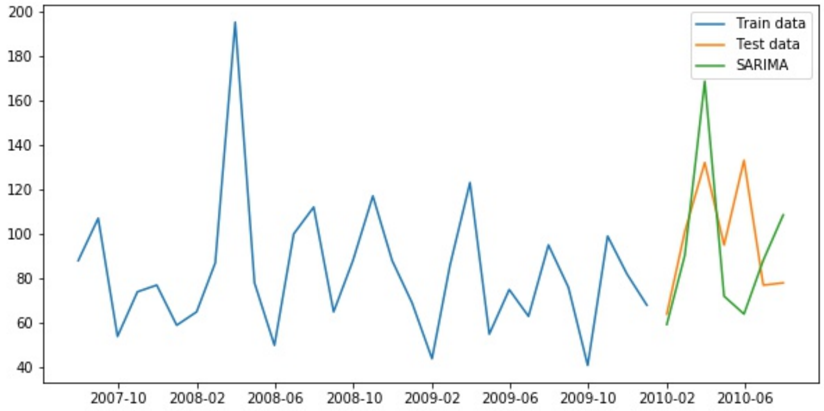
For this, we did the indexing of data using time series data, starting from 2007-07-31 to 2010-07-31, as the total number of absent hours will be calculated at the end of each month from July 2007 - July 2010.

We can also visualize our data using a method called time-series decomposition that allows us to decompose our time series into three distinct components: trend, seasonality, and noise. This can be seen using below plot:

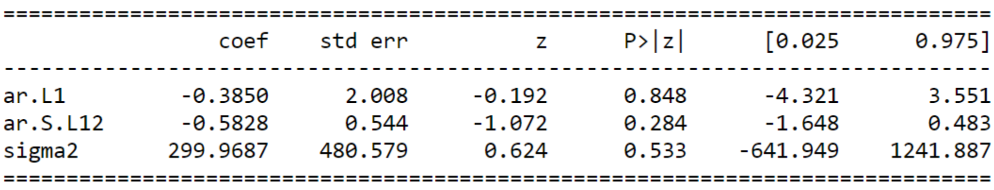


We are going to apply one of the most commonly used method for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average.

ARIMA models are denoted with the notation ARIMA(p, d, q). These three parameters account for seasonality, trend, and noise in data.



arima.jpg



```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3850      2.008     -0.192      0.848      -4.321       3.551
ar.S.L12      -0.5828      0.544     -1.072      0.284      -1.648       0.483
sigma2       299.9687    480.579      0.624      0.533    -641.949    1241.887
==============================================================================
```

Capture.jpg

After applying various combinations p,d,q and observing AIC (Akaike information criterion), we set to choose this combination for best fit for the model:

ARIMA(1, 1, 0)x(1, 1, 0, 12)12 - AIC:40.166185689188175

With this combination, we created ARIMA model. The results can be seen in below plot:

The summary of the model can be seen as below:

After applying the model in our cleaned data, we can calculate below metrics for the model:

ModelName	MSEˆ	MAEˆ	RMSEˆ
ARIMA	1115.38	26.48	33.39

4 Conclusion

4.1 Model Evaluation

As we have seen in previous section, as we tried on different models, we came across different results. To evaluate a regression model, following metrics are used.

- **Mean Square Error**: The mean squared error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated. This is given by,
- **Mean Absolute Error**: The mean absolute error (MAE) is a measure of difference between two continuous variable. This can be calculated by,

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - x\right|$$

ModelName	MSEˆ	MAEˆ	RMSEˆ
Linear Regression	553.434	19.37	23.52
Polynomial Regression	898.78	20.36	29.97
ARIMA	1115.38	26.48	33.39

4.2 Model Selection

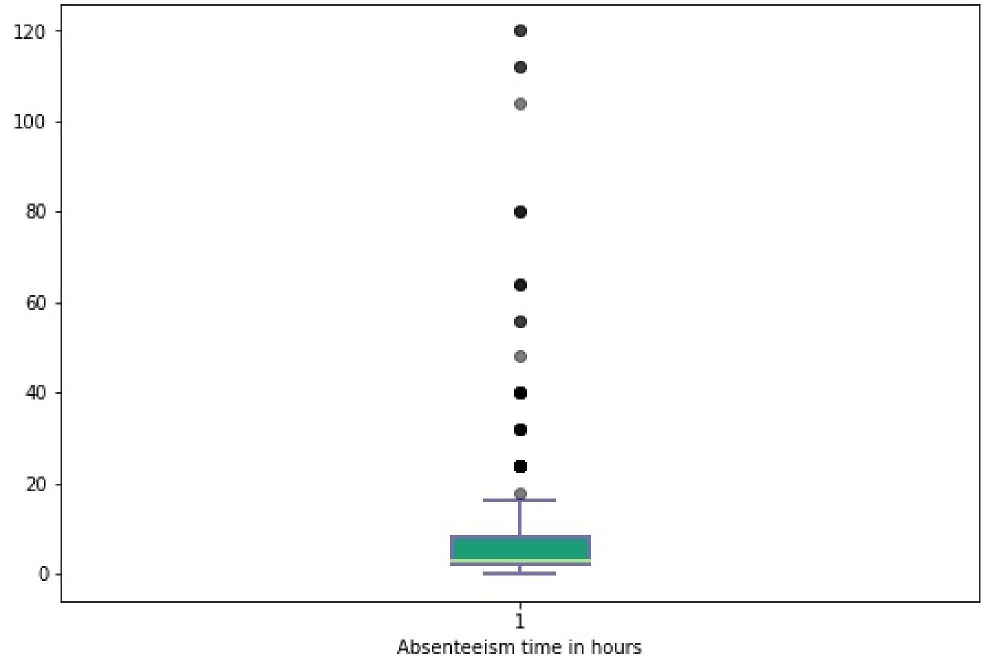
From the above table, we can clearly see that **Linear Regression** model works better than the rest of the n=models. Hence we will go with Linear Regression to predict the data.

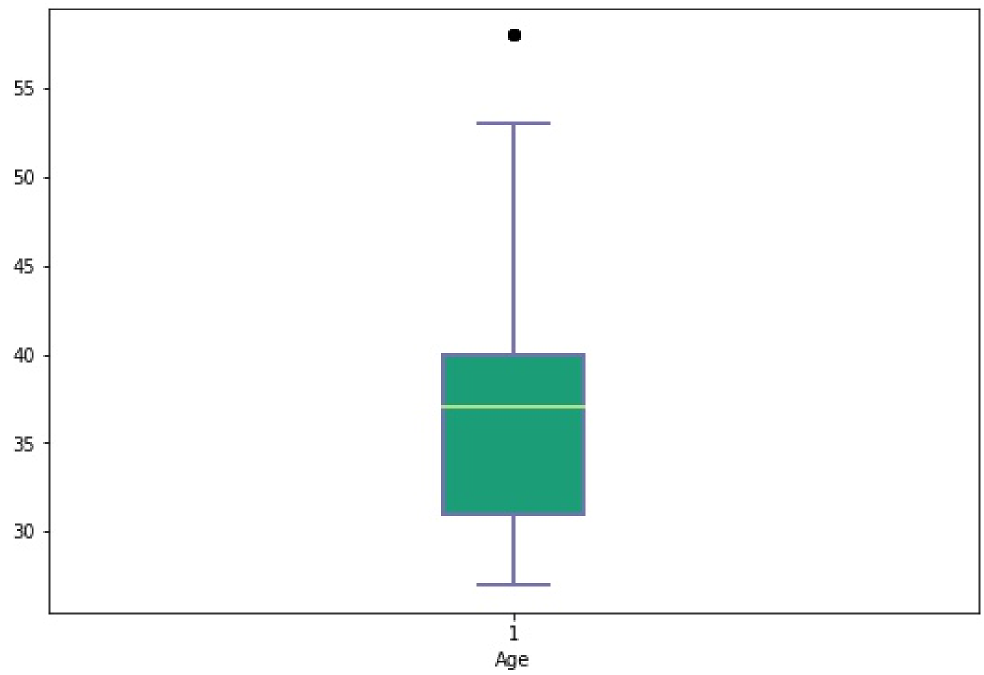
5 Appendix 1 - Univariate Analysis Graphs

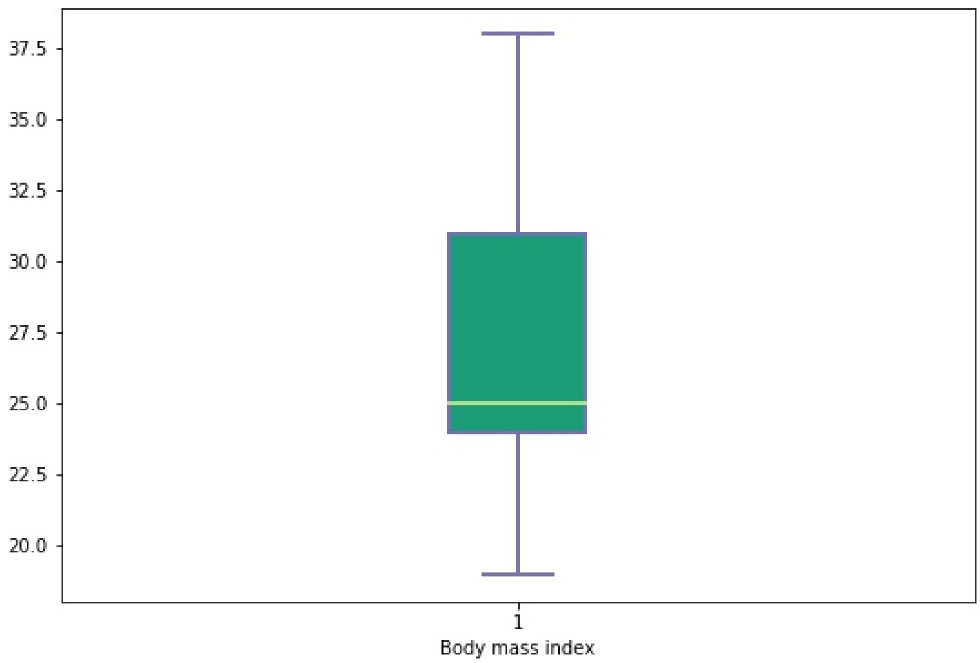
5.1 A) Box Plots

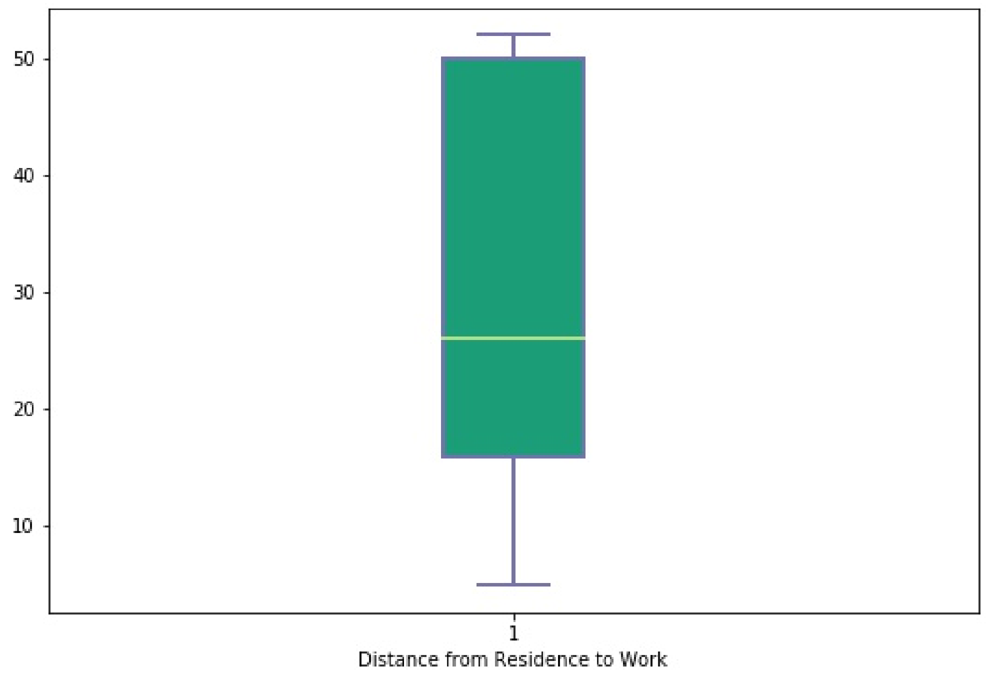
What are Box plots? A box plot is a method for graphically depicting groups of numerical data through their quartiles.

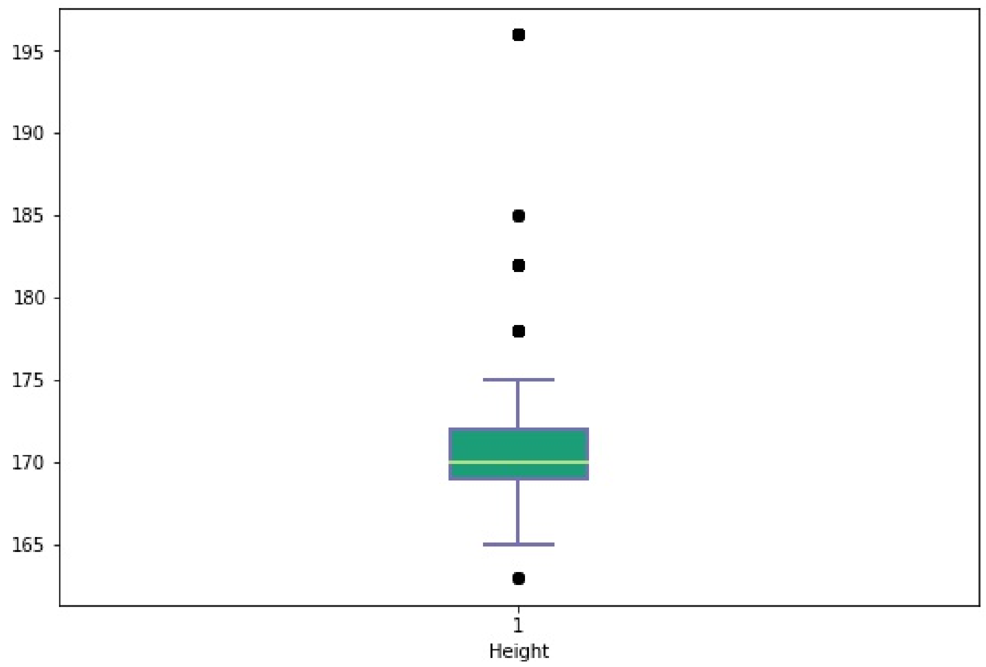
Box plots for each numerical data are shown below:

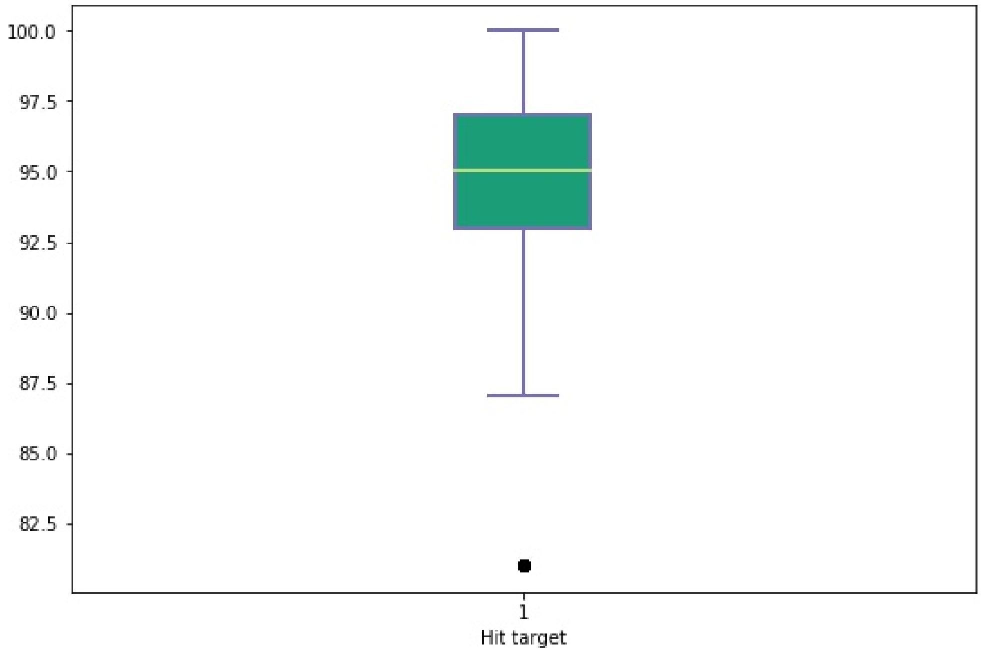


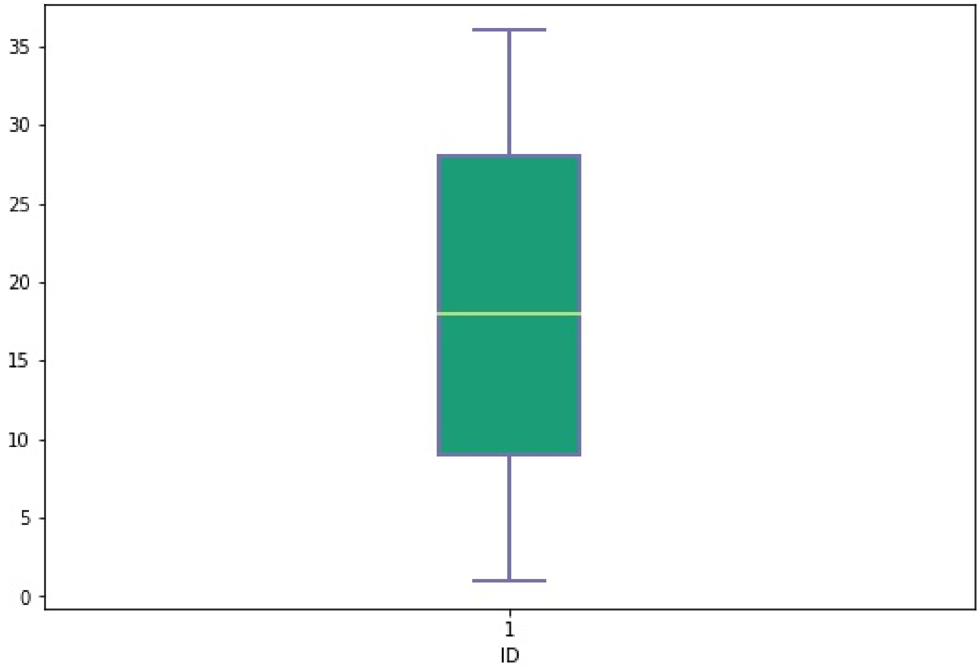

55
50
45
40
35
30
1
Age

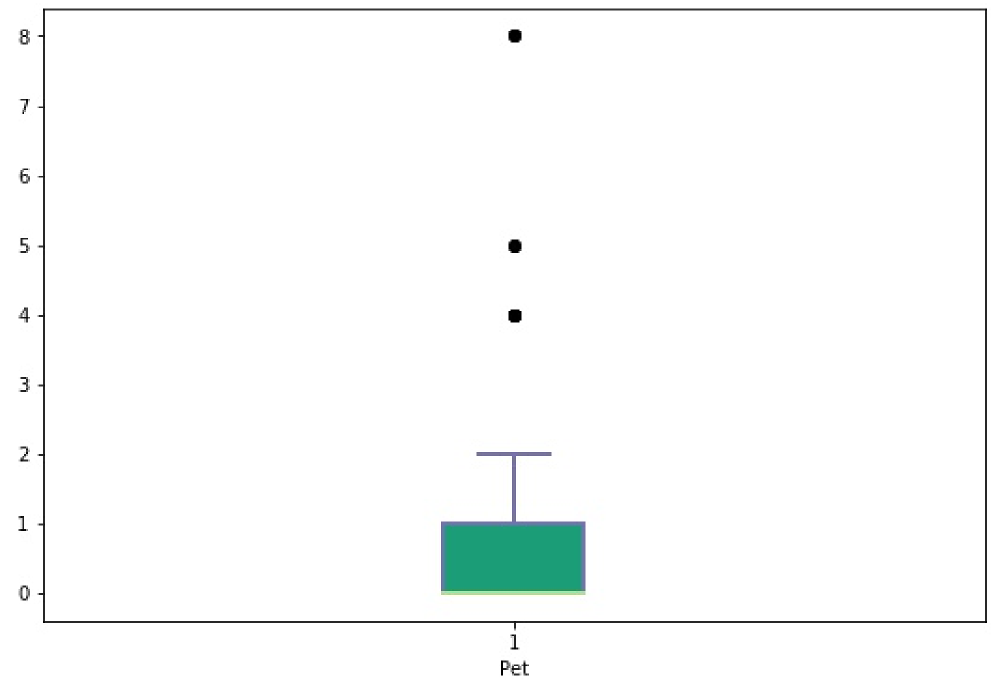

37.5
35.0
32.5
30.0
27.5
25.0
22.5
20.0
1
Body mass index

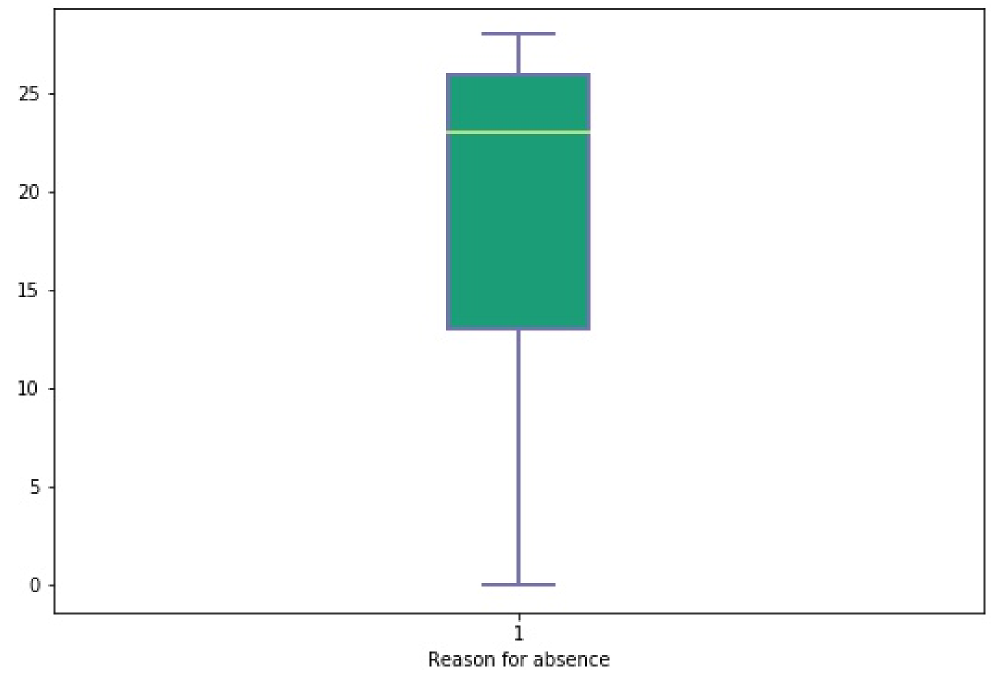

50
40
30
20
10
1
Distance from Residence to Work

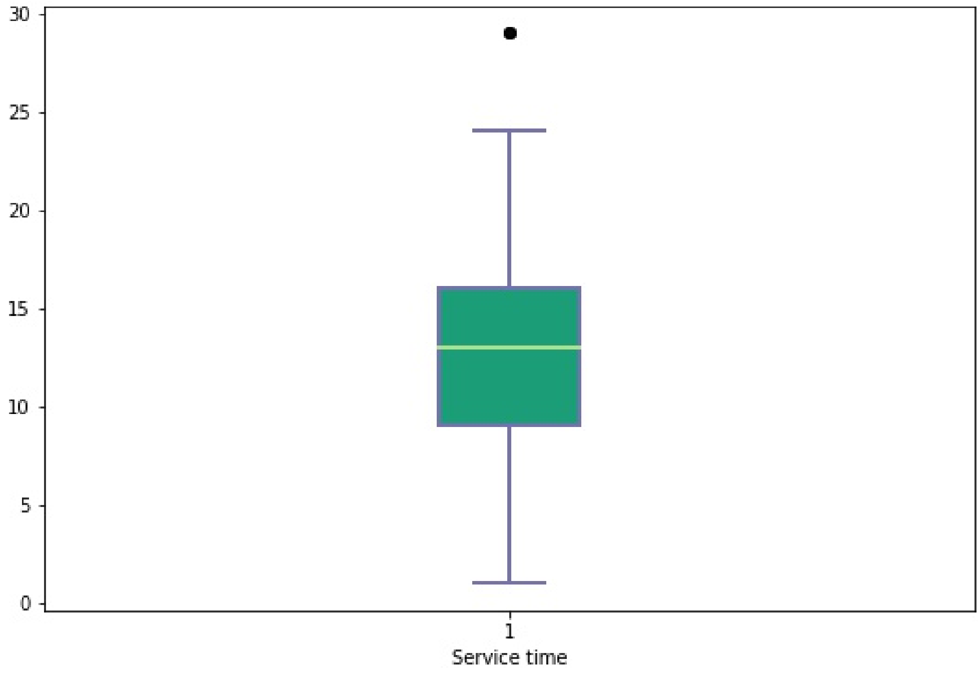

195
190
185
180
175
170
165
1
Height

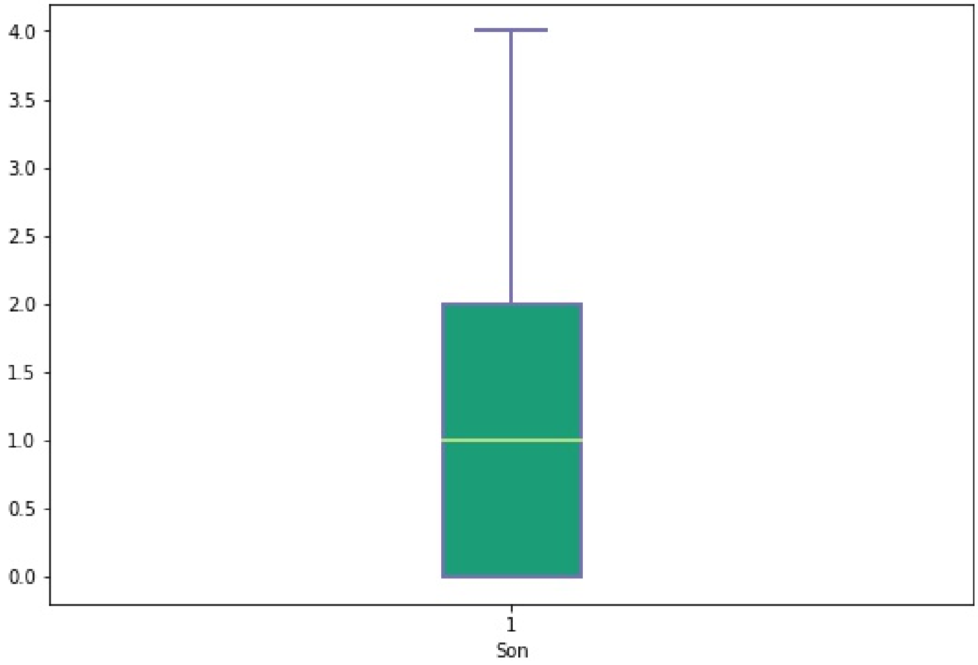

100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
1
Hit target

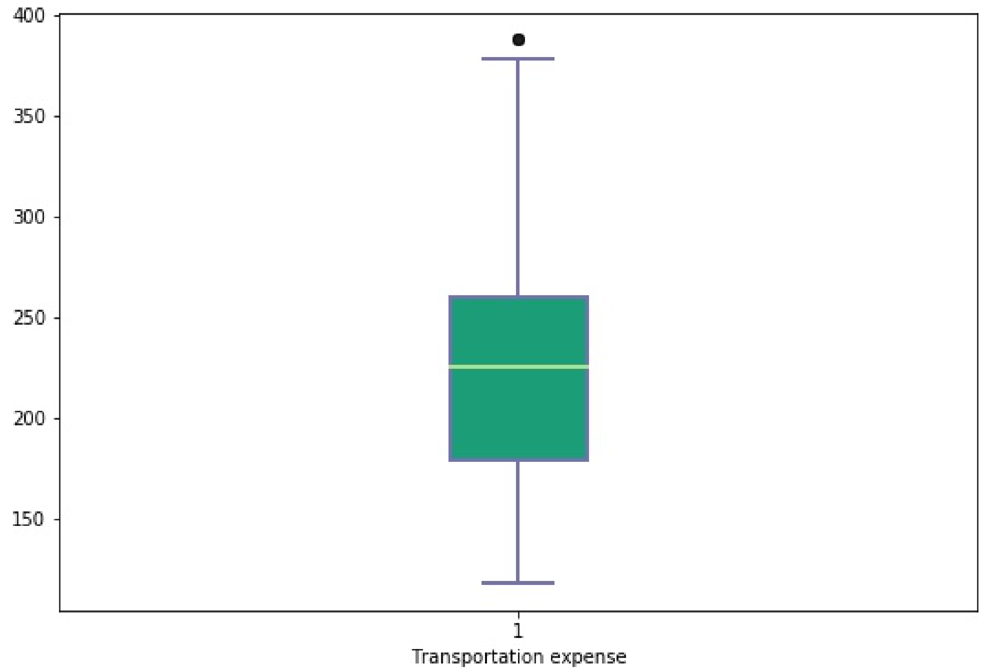

35
30
25
20
15
10
5
0
1
ID

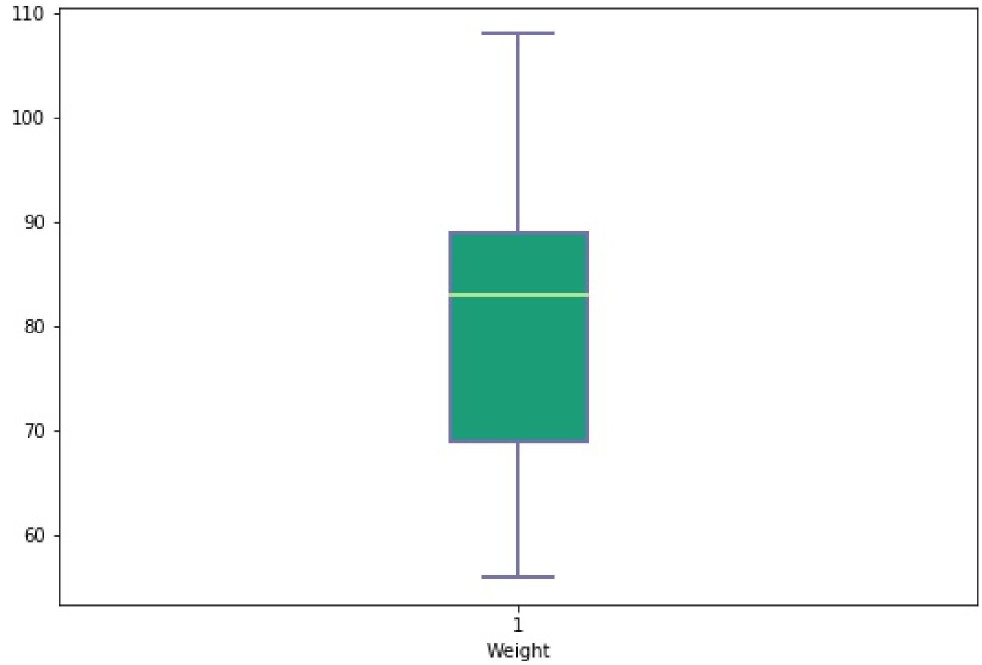

8
7
6
5
4
3
2
1
0
1
Pet

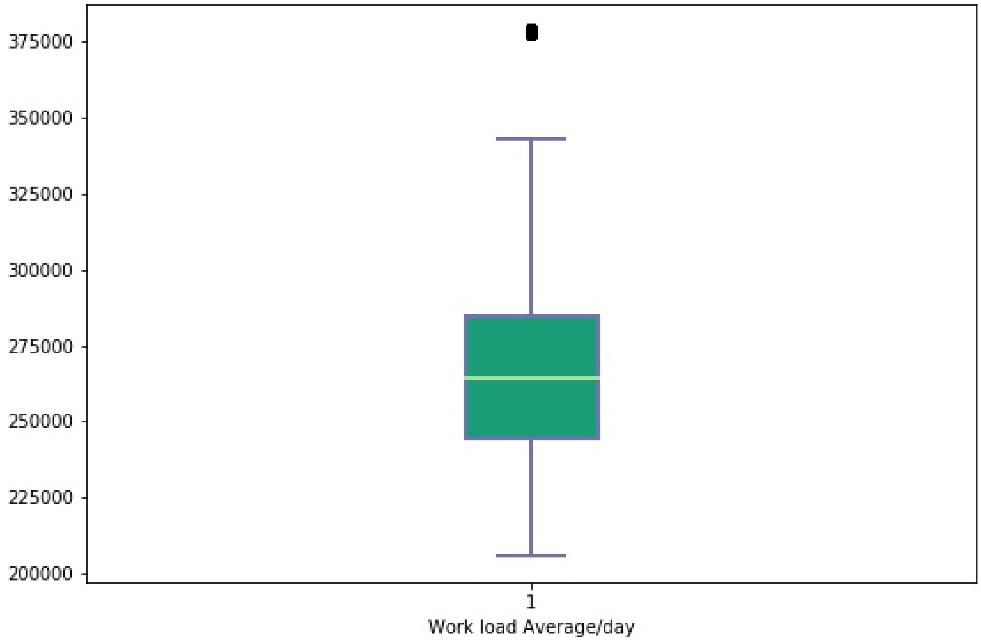

25
20
15
10
5
0
1
Reason for absence


30
25
20
15
10
5
0
1
Service time


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
Son


400
350
300
250
200
150
1
Transportation expense


110
100
90
80
70
60
1
Weight



Code used to create boxplots:

```
def CreateBoxPlot(dataset, columnNames):
    fig = plt.figure(1, figsize=(9, 6))
    #ax = fig.add_subplot(111)
    bp = plt.boxplot(dataset[columnNames], patch_artist=True)

    for box in bp['boxes']:
        # change outline color
        box.set( color='#7570b3', linewidth=2)
        # change fill color
        box.set( facecolor = '#1b9e77' )

    for whisker in bp['whiskers']:
        whisker.set(color='#7570b3', linewidth=2)

    for cap in bp['caps']:
        cap.set(color='#7570b3', linewidth=2)

    for median in bp['medians']:
        median.set(color='#b2df8a', linewidth=2)

    for flier in bp['fliers']:
        flier.set(marker='o', color='#e7298a', alpha=0.5)

    plt.xlabel(columnNames)

    filename = "D:\\edWisor\\Project-I\\Boxplot Figures\\"+columnNames+' boxplot.png'
    fig.savefig(filename, bbox_inches='tight')
    return filename
```