PROJECT EMPLOYEE ABSENTEEISM

Done by,
Ranjith P

Index:

Chapter 1

Introduction

1.1 Problem Description & 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?

1.2 Business Understanding:

Being an Analyst, I understood that the XYZ Courier company would like to find the reasons for employee absenteeism in order to derive new HR strategies in the company. The insights from the data will help the XYZ Courier company to know what changes that the company has to bring to reduce the number of absenteeism. Also, it needs to forecast the employee absenteeism if 2011

The result of this project will help the XYZ Courier company to take proactive actions in order to reduce employee absenteeism and increase benefits mutually.

1.3 Dataset Details:

Dataset Characteristics: Timeseries Multivariant Number of Attributes: 21

Missing Values: Yes

There are 21 variables in our data in which 20 are independent variables and 1 (Absenteeism time in hours) is dependent variable. Since our target variable is continuous in nature, this is a regression problem.

Variables 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

3. Month of absence
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
5. Seasons (summer (1), autumn (2), winter (3), spring (4))
6. Transportation expense
7. Distance from Residence to Work (kilometres)
8. Service time
9. Age
10. Work load Average/day
11. Hit target
12. Disciplinary failure (yes=1; no=0)
13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
14. Son (number of children)
15. Social drinker (yes=1; no=0)
16. Social smoker (yes=1; no=0)
17. Pet (number of pet)
18. Weight
19. Height
20. Body mass index
21. Absenteeism time in hours (target)

Chapter 2

Exploratory Data Analysis or Data Pre-processing

Exploratory Data Analysis

Data Pre-processing:

Exploratory Data Analysis helps us to understand the data better. It will also help us to do necessary data cleaning and data preparations in the pre-processing stage. Most probably the exploratory data analysis and data pre-processing are done together in order to sanitize the data for modelling. The nature of data will help us to decide the methodology of dealing with the data and perform the necessary algorithms.

To start this process, we look at the summary, structure and dimension of data to have a basic understanding on the data. We visualize the data to know the distribution of each features to check the normality of the data. Also, multivariate visualizations can be done with the features to know the relationship of features. Generally, Data Explorations and Pre-processing includes understanding the data, cleaning the data and visualizing the data as well.

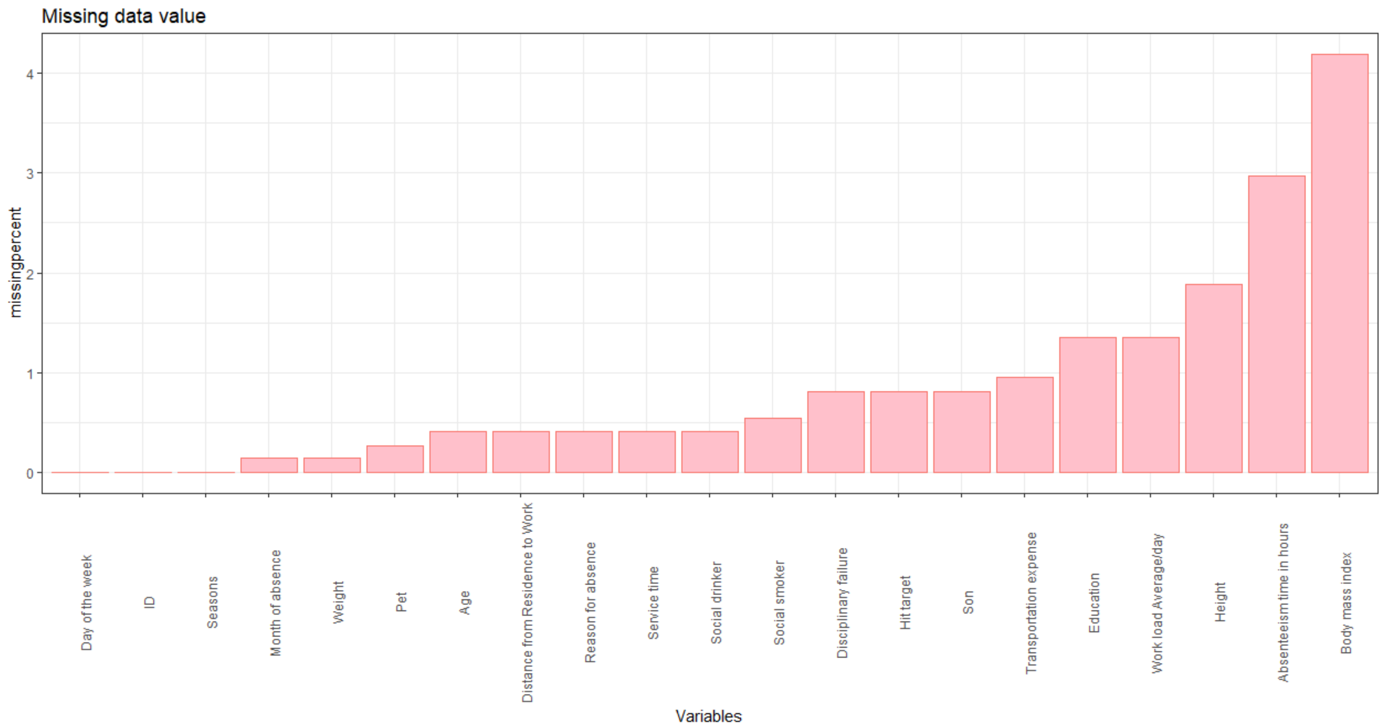
Data Summary

```
> summary(emp)
       ID        Reason for absence Month of absence Day of the week    Seasons      Transportation expense
 Min.   : 1.00   Min.   : 0.00      Min.   : 0.000   Min.   :2.000   Min.   :1.000   Min.   :118
 1st Qu.: 9.00   1st Qu.:13.00      1st Qu.: 3.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:179
 Median :18.00   Median :23.00      Median : 6.000   Median :4.000   Median :3.000   Median :225
 Mean   :18.02   Mean   :19.19      Mean   : 6.319   Mean   :3.915   Mean   :2.545   Mean   :221
 3rd Qu.:28.00   3rd Qu.:26.00      3rd Qu.: 9.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:260
 Max.   :36.00   Max.   :28.00      Max.   :12.000   Max.   :6.000   Max.   :4.000   Max.   :388
                 NA's   :3          NA's   :1                                        NA's   :7
 Distance from Residence to Work  Service time        Age        Work load Average/day   Hit target
 Min.   : 5.00                   Min.   : 1.00   Min.   :27.00   Min.   :205917        Min.   : 81.00
 1st Qu.:16.00                   1st Qu.: 9.00   1st Qu.:31.00   1st Qu.:244387        1st Qu.: 93.00
 Median :26.00                   Median :13.00   Median :37.00   Median :264249        Median : 95.00
 Mean   :29.67                   Mean   :12.57   Mean   :36.45   Mean   :271189        Mean   : 94.59
 3rd Qu.:50.00                   3rd Qu.:16.00   3rd Qu.:40.00   3rd Qu.:284853        3rd Qu.: 97.00
 Max.   :52.00                   Max.   :29.00   Max.   :58.00   Max.   :378884        Max.   :100.00
 NA's   :3                       NA's   :3       NA's   :3       NA's   :10            NA's   :6
 Disciplinary failure   Education          Son        Social drinker   Social smoker         Pet
 Min.   :0.00000      Min.   :1.000   Min.   :0.000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.00000      1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.00000      Median :1.000   Median :1.000   Median :1.0000   Median :0.00000   Median :0.0000
 Mean   :0.05313      Mean   :1.296   Mean   :1.018   Mean   :0.5672   Mean   :0.07337   Mean   :0.7466
 3rd Qu.:0.00000      3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
 Max.   :1.00000      Max.   :4.000   Max.   :4.000   Max.   :1.0000   Max.   :1.00000   Max.   :8.0000
 NA's   :6            NA's   :10      NA's   :6       NA's   :3        NA's   :4         NA's   :2
     Weight           Height      Body mass index Absenteeism time in hours
 Min.   : 56.00   Min.   :163.0   Min.   :19.00   Min.   :  0.000
 1st Qu.: 69.00   1st Qu.:169.0   1st Qu.:24.00   1st Qu.:  2.000
 Median : 83.00   Median :170.0   Median :25.00   Median :  3.000
 Mean   : 79.06   Mean   :172.2   Mean   :26.68   Mean   :  6.978
 3rd Qu.: 89.00   3rd Qu.:172.0   3rd Qu.:31.00   3rd Qu.:  8.000
 Max.   :108.00   Max.   :196.0   Max.   :38.00   Max.   :120.000
 NA's   :1        NA's   :14      NA's   :31      NA's   :22
> 
```

2.1 Missing Value Analysis:

The first step in cleaning the data is detecting the missing values in the data and removing or imputing it. The problem of missing value is common. Missing values in the data can complicate our analysis by creating bias or reducing statistical efficiency. So, we detect the missing values and treat it. Either we remove the missing values or we will impute it through various techniques. In our data, there is no missing value.

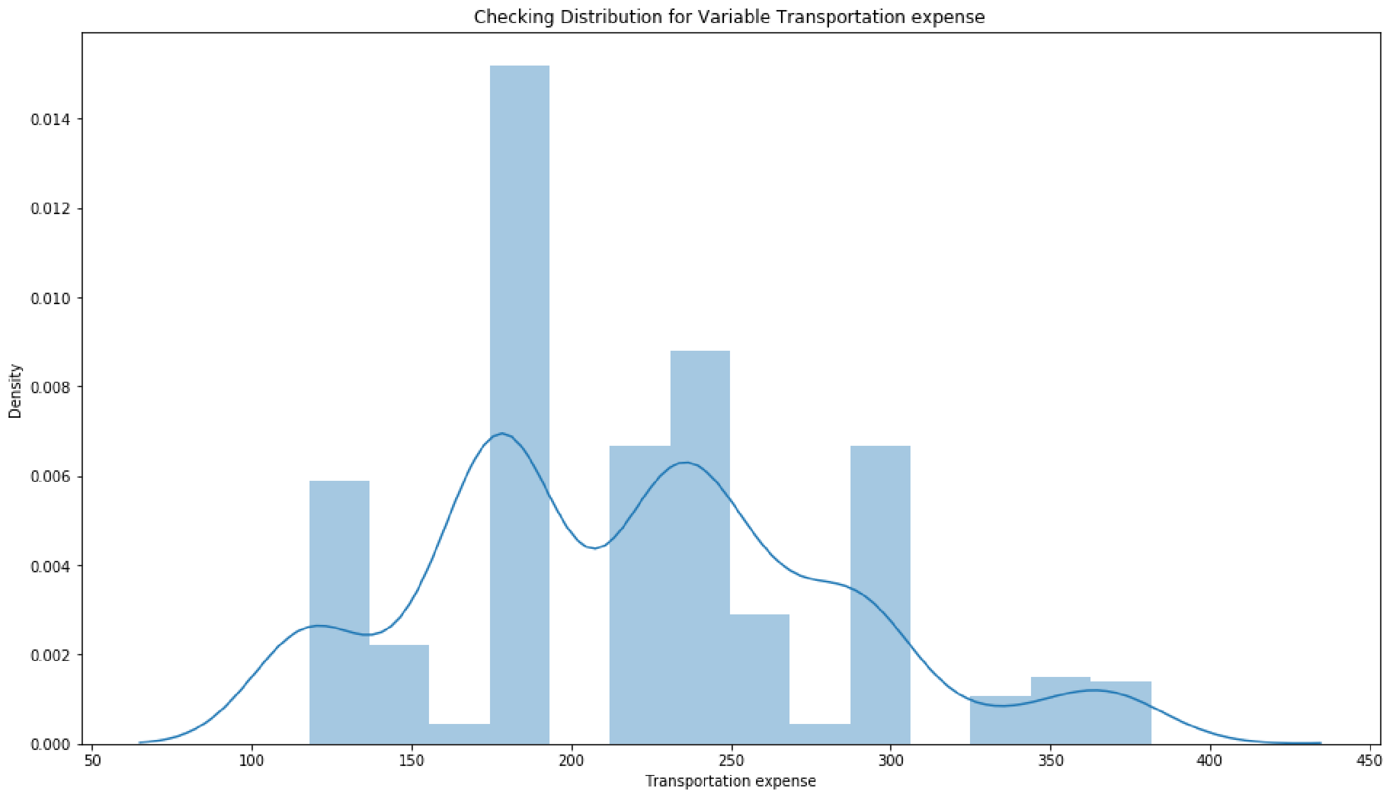
ID	0
Reason for absence	3
Month of absence	1
Day of the week	0
Seasons	0
Transportation expense	7
Distance from Residence to Work	3
Service time	3
Age	3
Work load Average/day	10
Hit target	6
Disciplinary failure	6
Education	10
Son	6
Social drinker	3
Social smoker	4
Pet	2
Weight	1
Height	13
Body mass index	29
Absenteeism time in hours	22

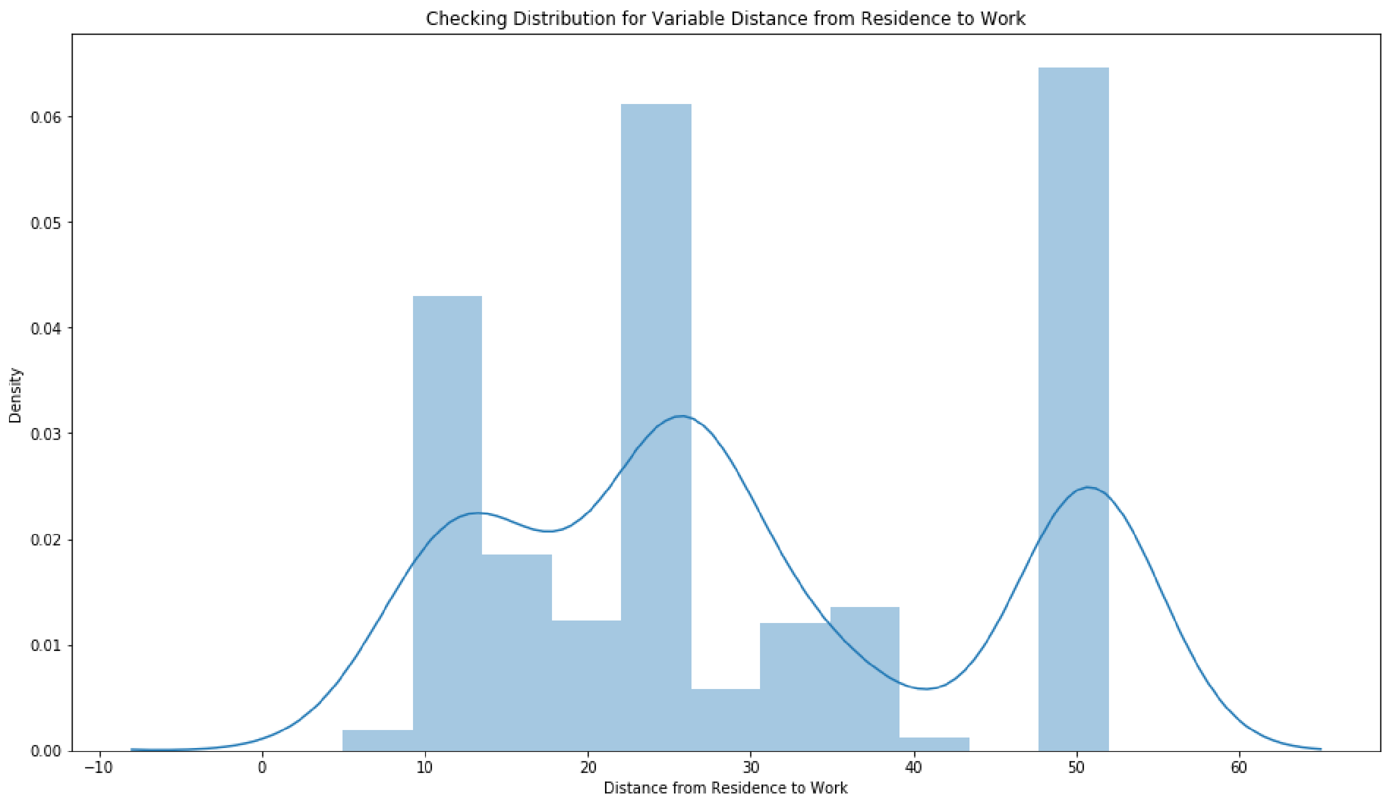


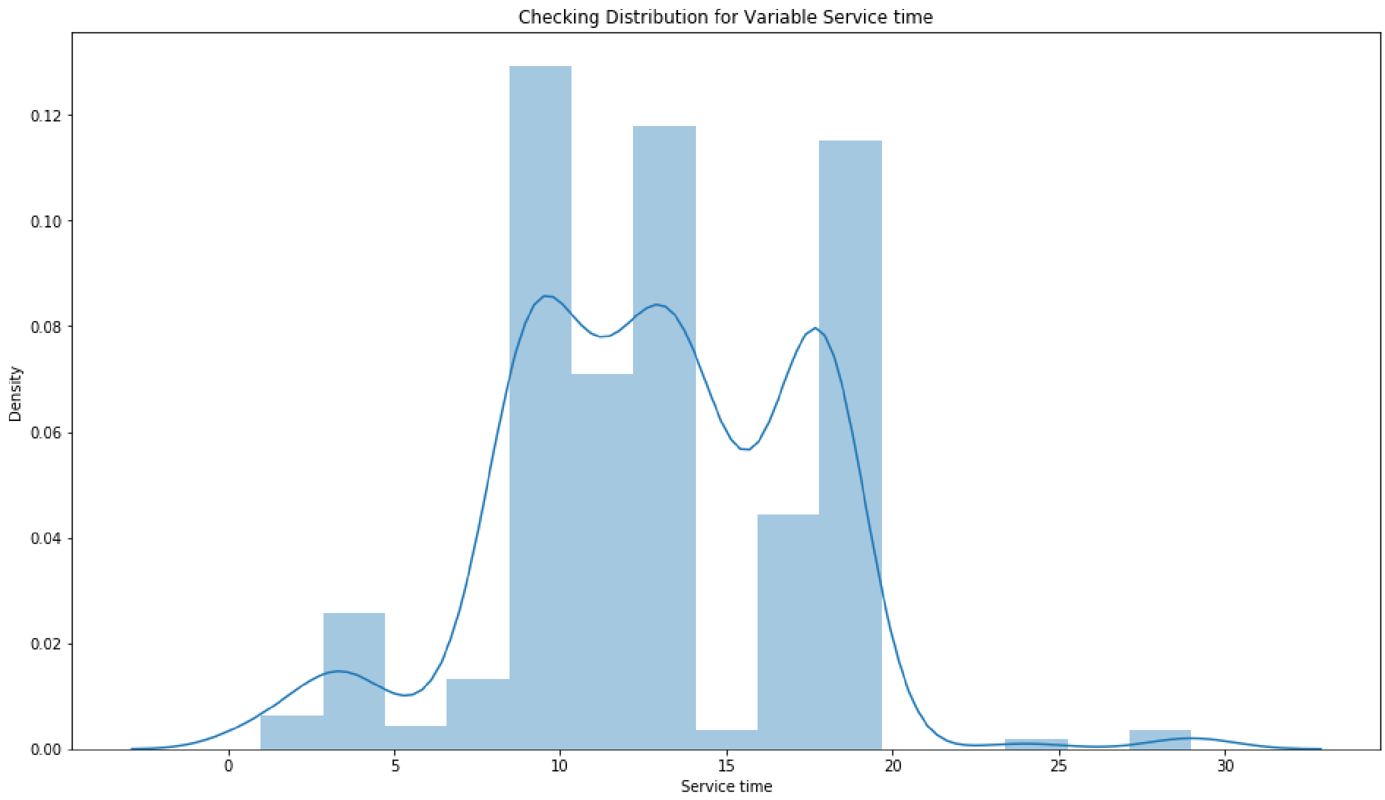
- The above plot shows that many variables have missing values but not more than 5%
- It's also seen that the target variable Absenteeism time in hours has 3% of missing values. We should not impute anything in target variable. So, I'm removing the observations which has missing value in target variable.
- Rest of the missing values are imputed using K Nearest Neighbours. If we have more percent of missing values in any columns, then we will remove that columns from our analysis. Since, we have only less than 5% of missing values is the dataset, we are imputing it.

2.2 Data Understanding:

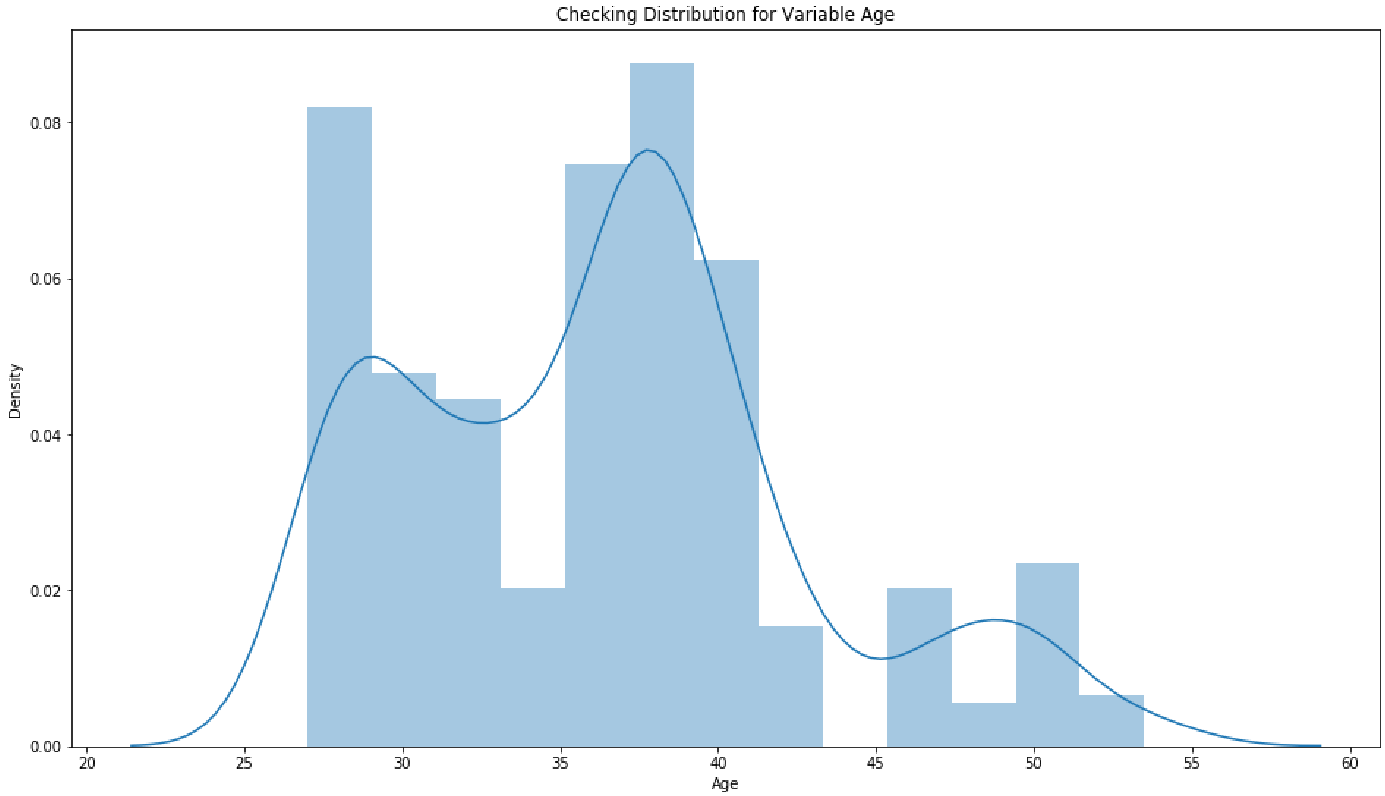
Data Visualizations of Numerical variables to understand the distribution:



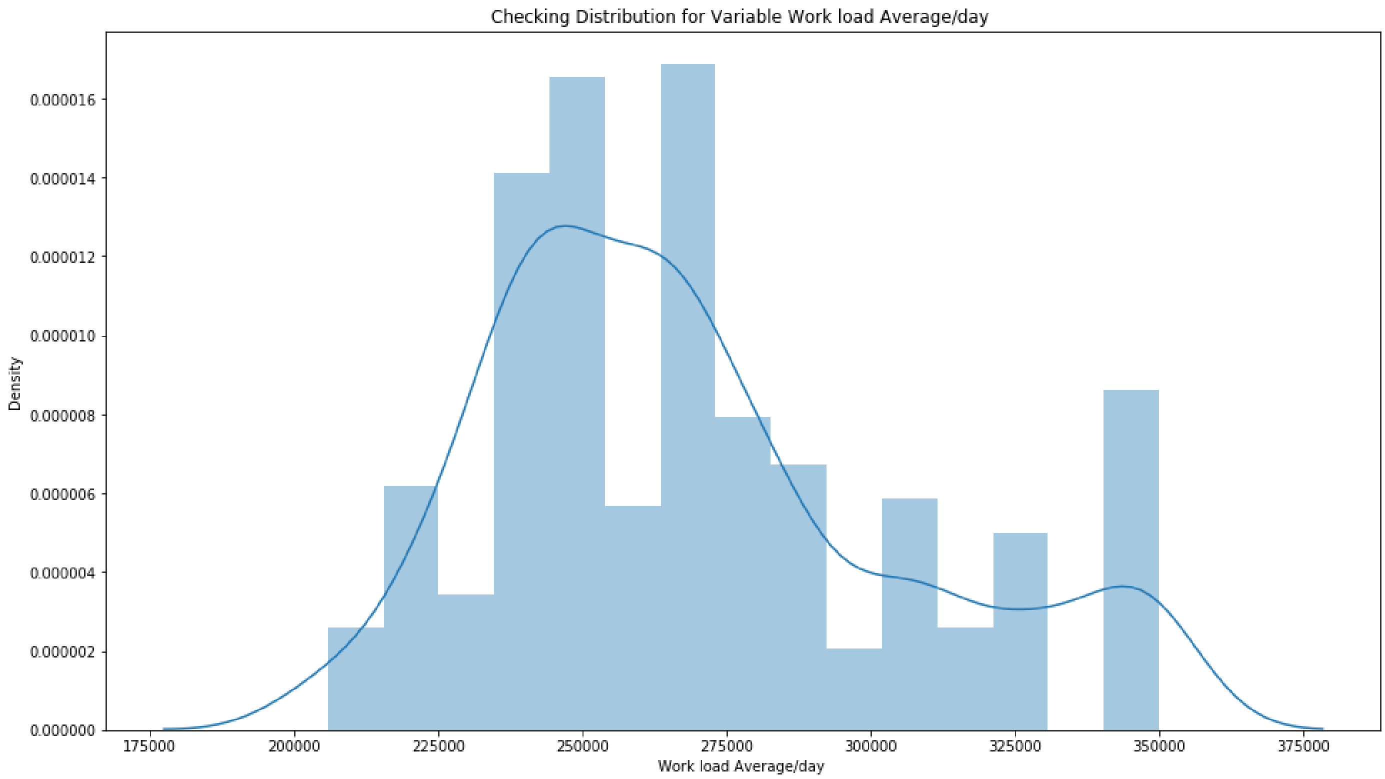


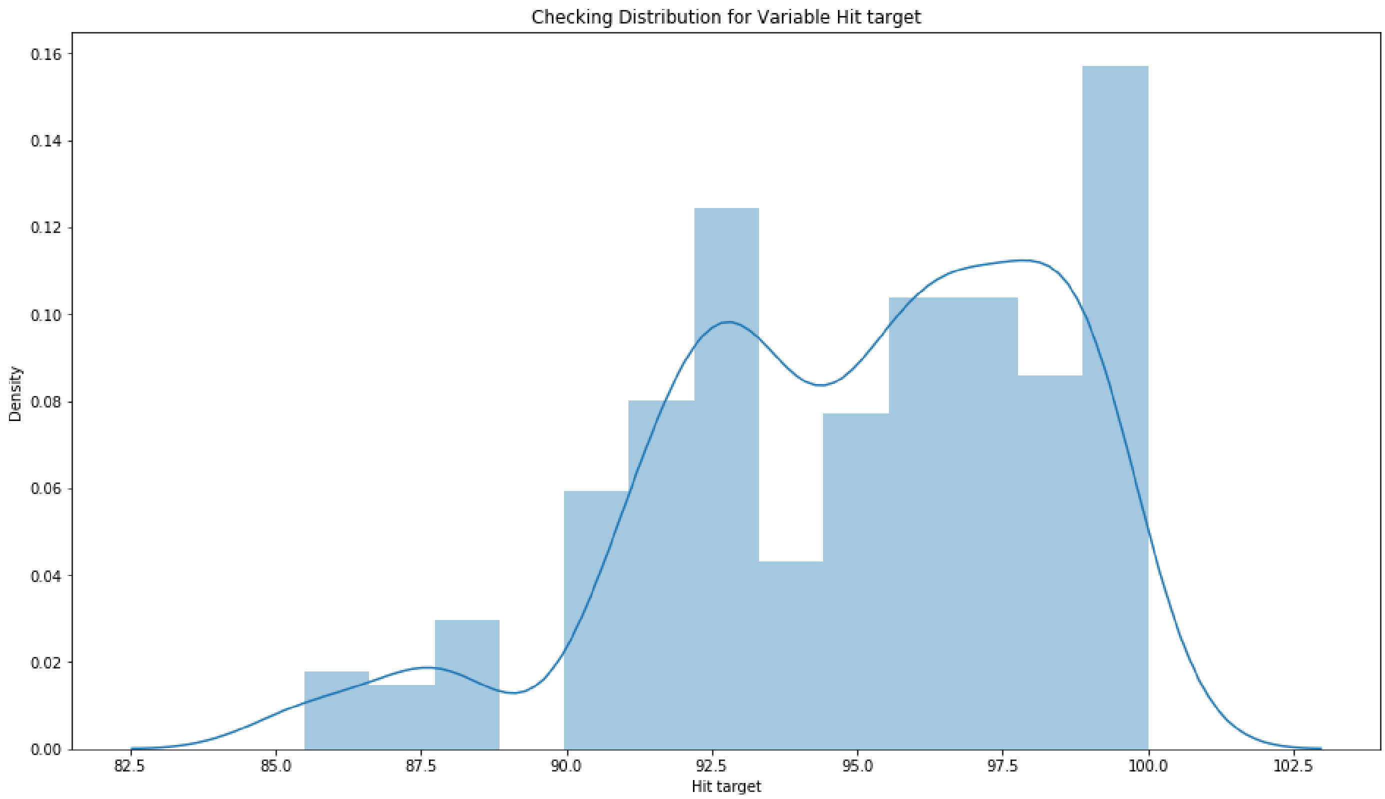


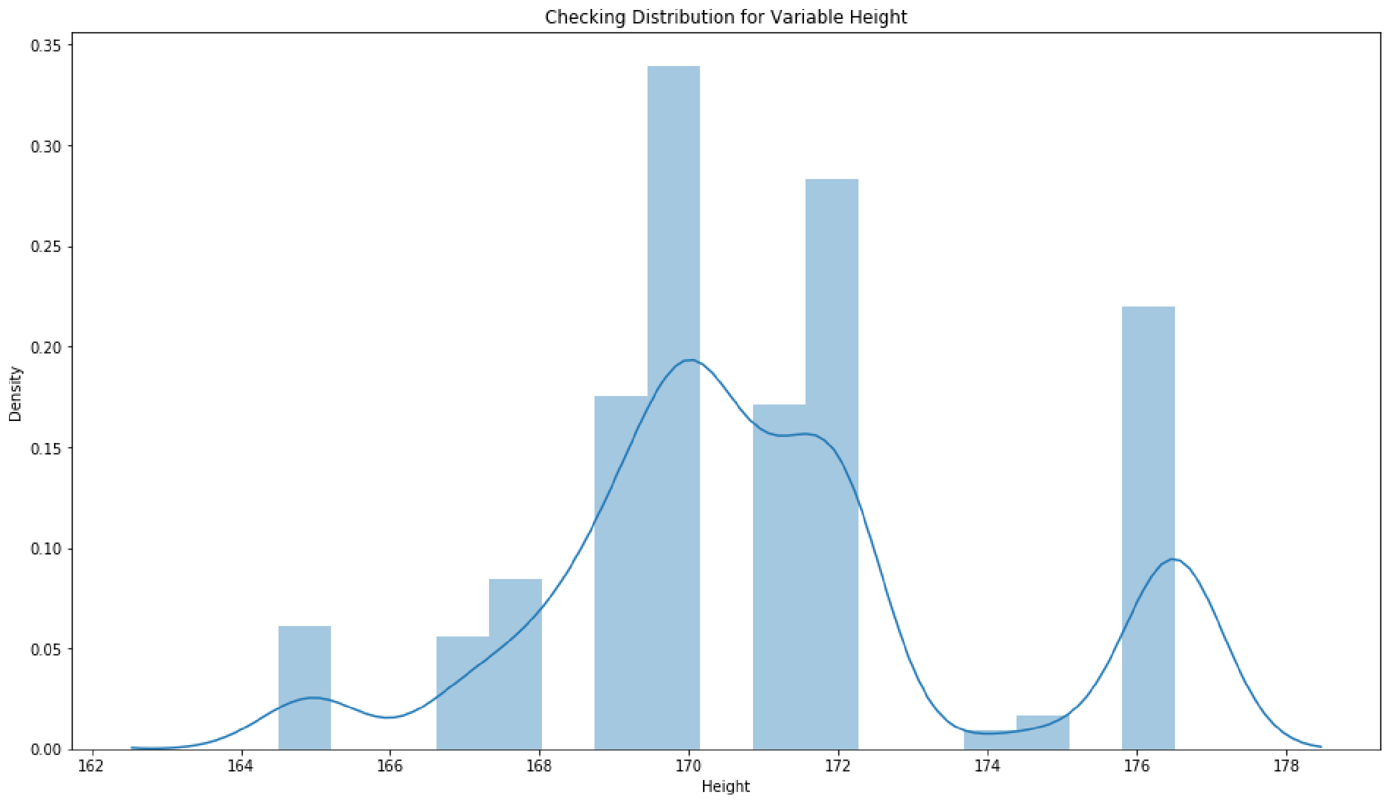
Checking Distribution for Variable Service time
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
5
10
15
20
25
30
Service time

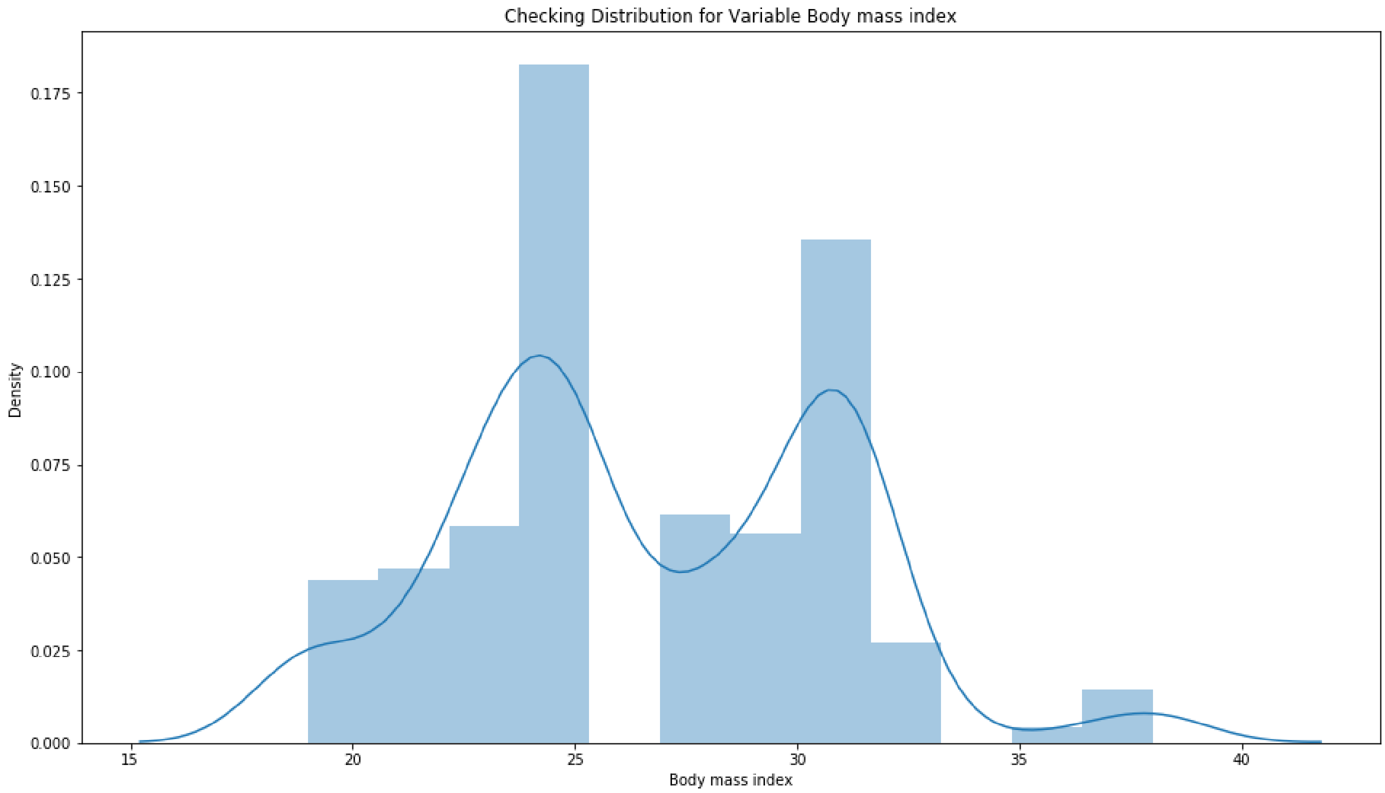


Checking Distribution for Variable Age
Density
0.00
0.02
0.04
0.06
0.08
20
25
30
35
40
45
50
55
60
Age


Checking Distribution for Variable Work load Average/day
Density
0.000016
0.000014
0.000012
0.000010
0.000008
0.000006
0.000004
0.000002
0.000000
175000
200000
225000
250000
275000
300000
325000
350000
375000
Work load Average/day


Checking Distribution for Variable Hit target
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
82.5
85.0
87.5
90.0
92.5
95.0
97.5
100.0
102.5
Hit target





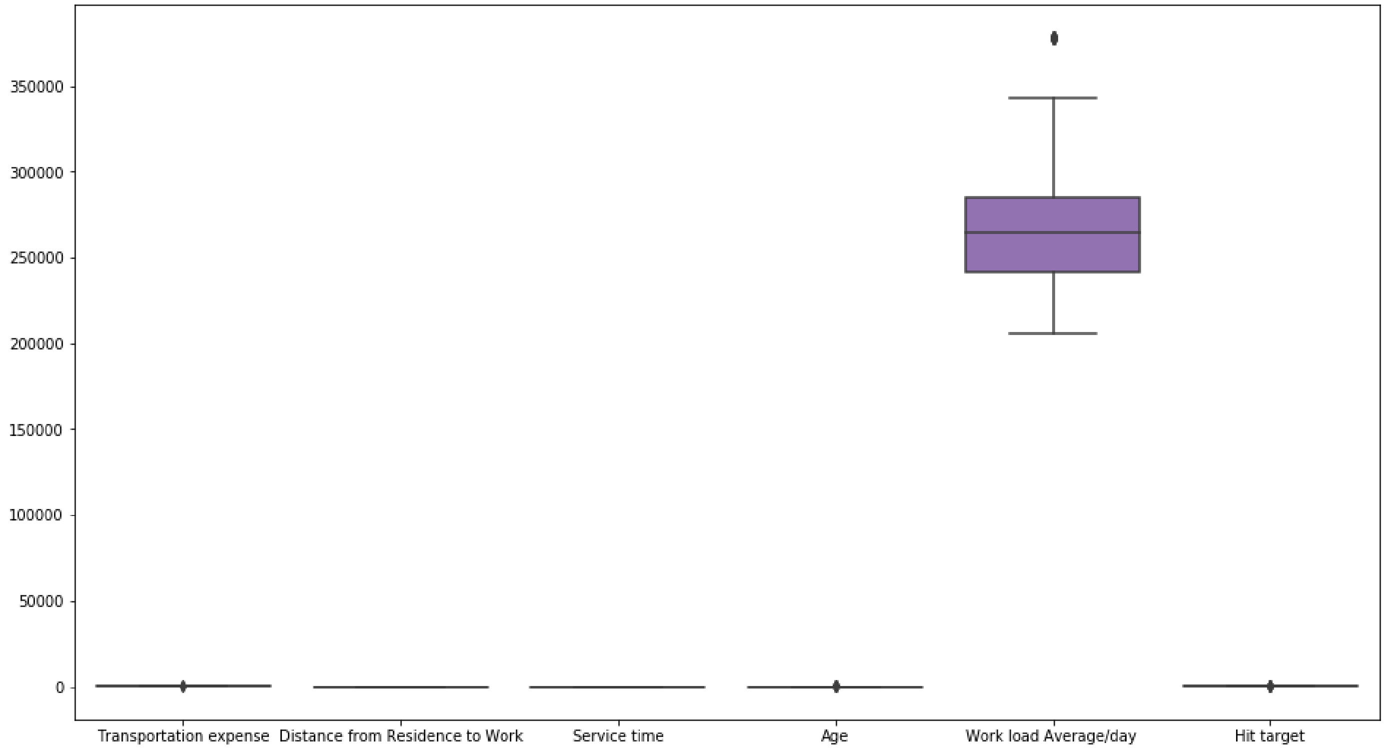
- All the numerical variables in our dataset are not normal.
- The above plots help us to know the range of the values in the numerical variables.
- This will help us to perform normalisation during feature scaling.

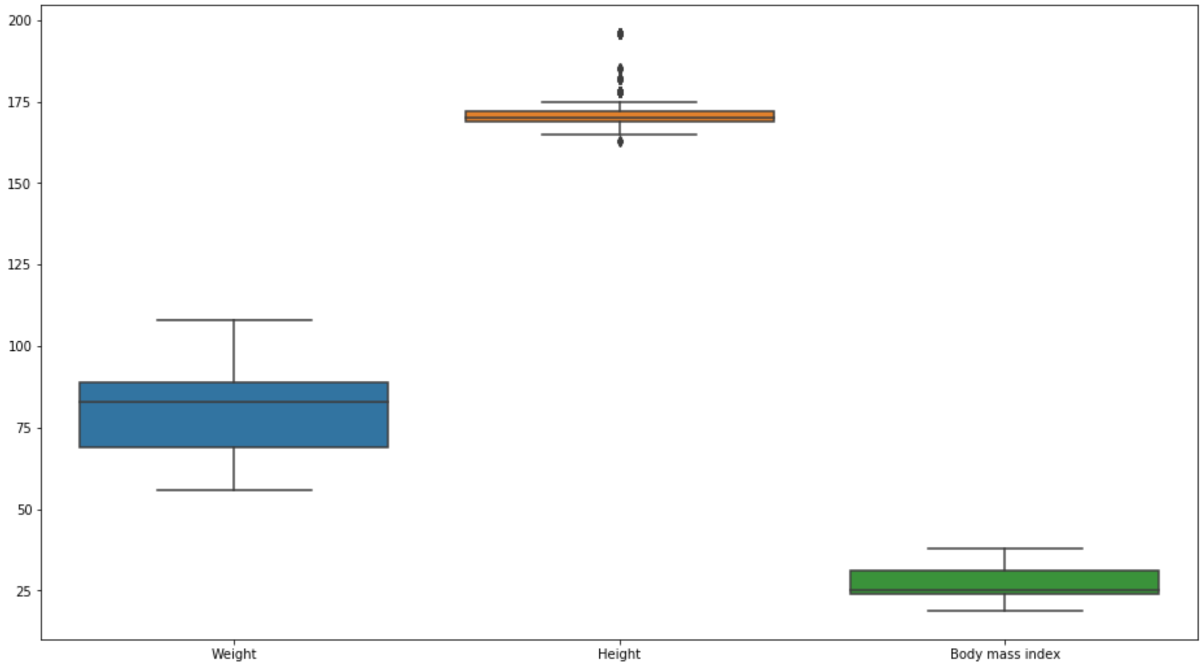
2.3 Outlier Detection & Treatment:

Outliers are the extreme values which may skew the data and creates bias in our analysis. Outliers will affect the assumptions and results of our analysis. So, it's better to remove or impute the outliers. We can visualize the outliers using the box plot. Generally, outliers are considered as the values above q75+(1.5*iqr) or values below q25-(1.5*iqr).

I thought of imputing the outliers instead if removing it. Because, already there are only 3333 observations in my train data. If I remove outliers, then the observations will be reduced. So, I converted the outliers to NA's and imputed using K Nearest Neighbours.

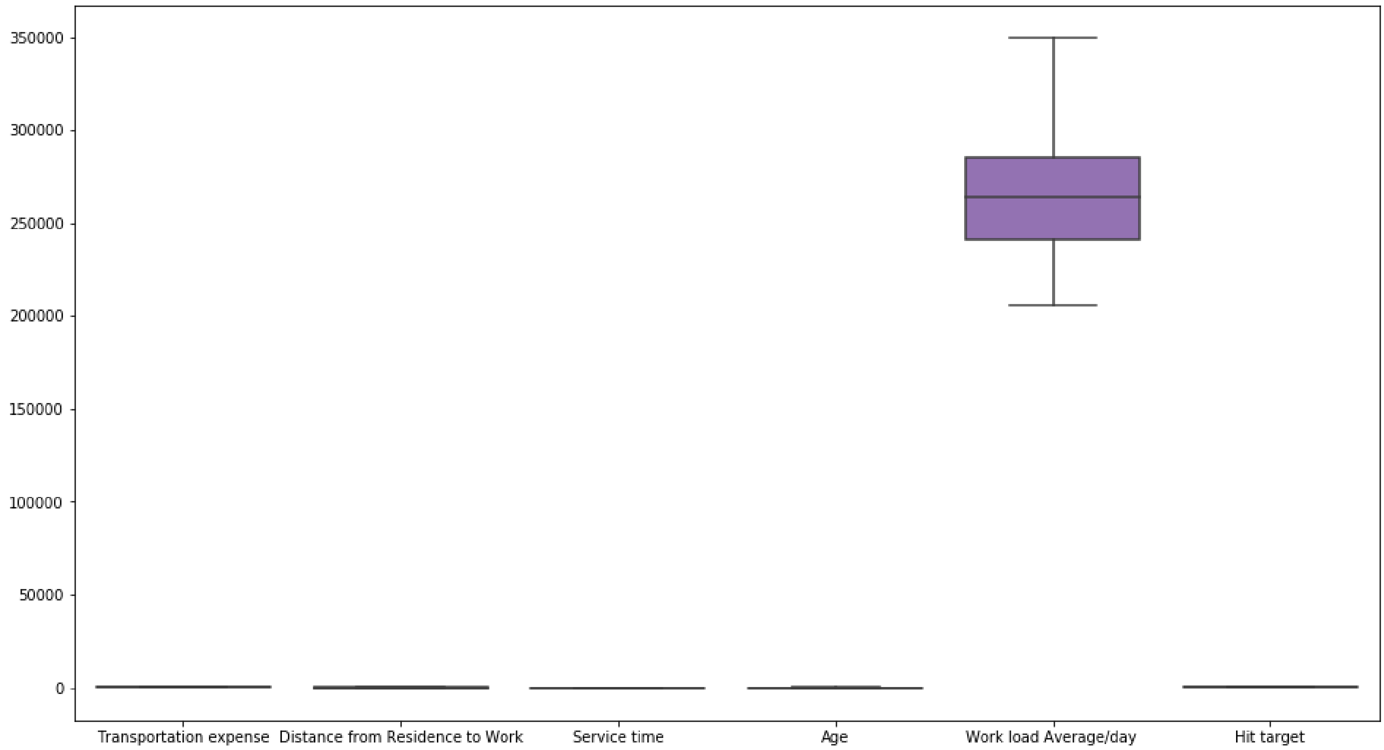
Boxplot Visualization of Numerical Variables to see the presence of Outliers.

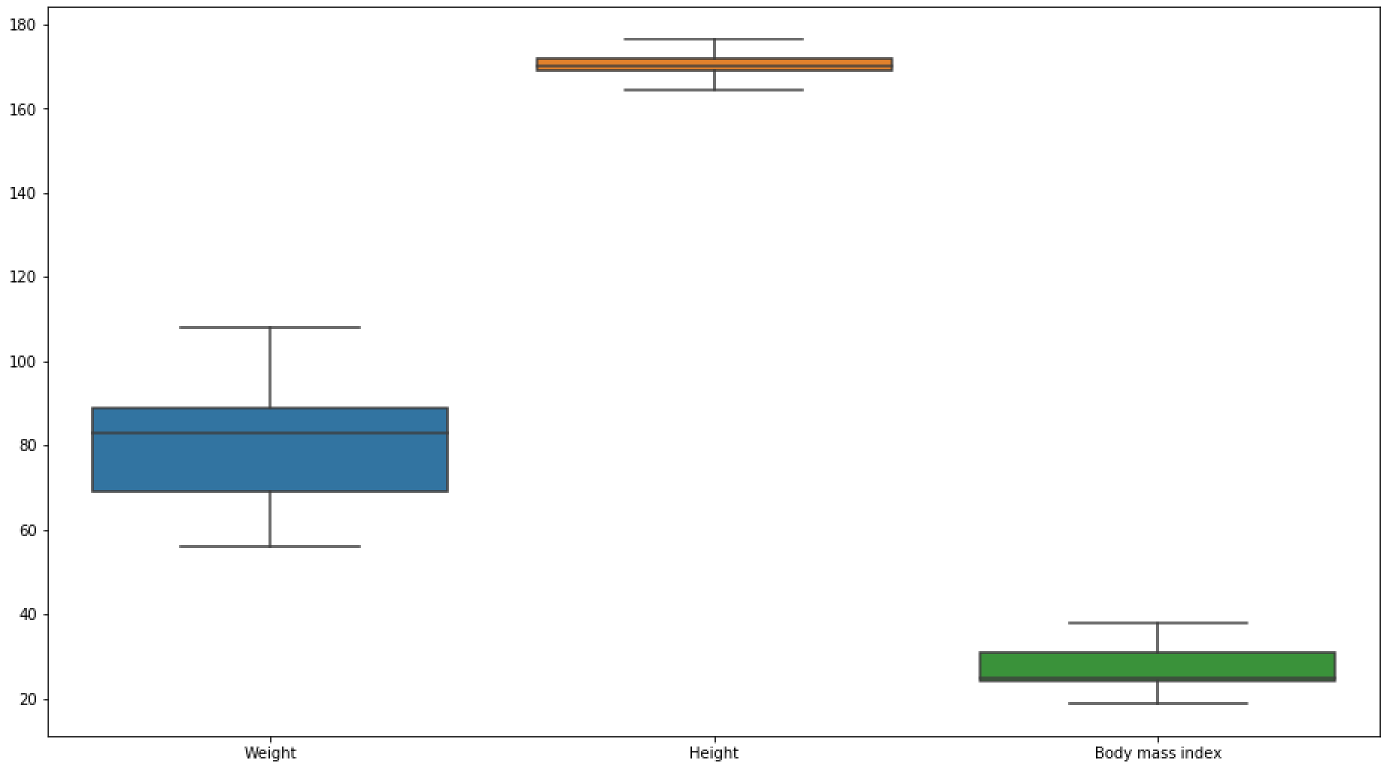




- There are some outliers present in the dataset. I used **Winsorization** (capping) technique to impute the outliers.

After Winsorization;





2.4 Feature Selection:

Feature Selection is a process used to select the important features among the predictors for the model building. The general rule is that the target should be dependent on predictors and the predictors (either numeric or categorical) should be independent to each other. If two or more predictors are dependent on each other, then there exists the problem of multicollinearity. Then we remove the multicollinear features to get rid of multicollinearity issue. It is selecting relevant features from dataset to use in model. It is also called as Dimensionality reduction. For numerical features we perform correlation and for categorical features we perform Chi-Square test.

In our data, most of the features are numerical and the few categorical variables are converted to numeric by converting levels to numbers. Then Correlation plot is found to know the multicollinearity affected features.

Feature selection is another pre-processing technique which decreases the load over machine learning algorithm checking the correlation between other feature and check which feature is highly correlated to another feature. Feature Selection reduces the complexity of a model and makes it easier to interpret. It also reduces over fitting. Features are selected based on their scores in various statistical tests for their correlation with the outcome variable. Correlation plot is used to find out if there is any multi-collinear between variables. The highly collinear variables are dropped and then the model is executed.

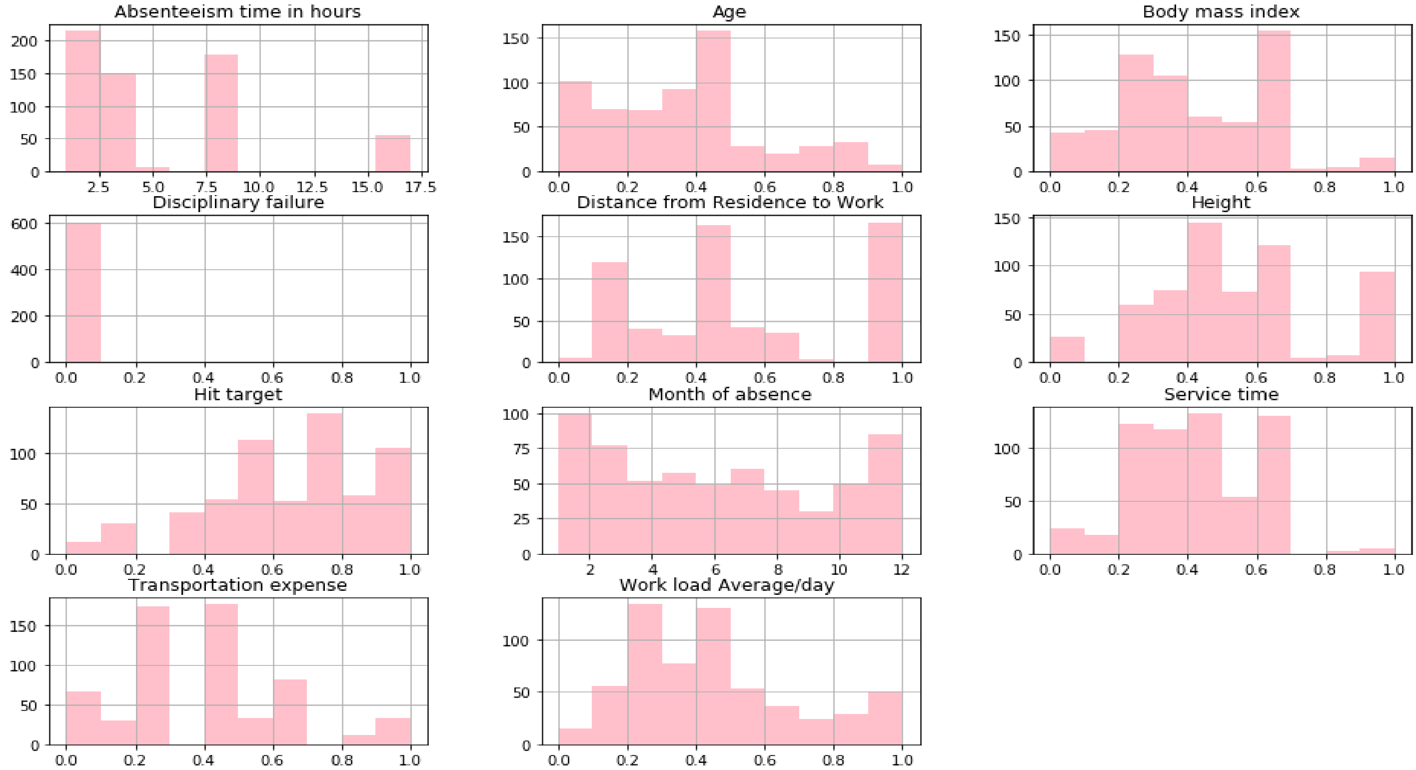
There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi-collinear. In this project we have selected Correlation Analysis for numerical variable and ANOVA (Analysis of variance) for categorical variable.

	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Weight	Height	Body mass index	Absenteeism time in hours
Transportation expense	1	0.216423	-0.358485	-0.216362	-0.000419361	-0.0639974	-0.187641	-0.161371	-0.111822	0.184256
Distance from Residence to Work	0.216423	1	0.163002	-0.13077	-0.0630917	-0.0146716	0.0062089	-0.333643	0.174845	-0.0821255
Service time	-0.358485	0.163002	1	0.678092	0.00817686	0.00984036	0.47029	-0.103086	0.519562	-0.0525903
Age	-0.216362	-0.13077	0.678092	1	-0.0537277	-0.0185365	0.422135	-0.0127698	0.480798	-0.0133345
Work load Average/day	-0.000419361	-0.0630917	0.00817686	-0.0537277	1	-0.073497	-0.030017	0.0395006	-0.0848966	0.12922
Hit target	-0.0639974	-0.0146716	0.00984036	-0.0185365	-0.073497	1	-0.0254192	0.0688185	-0.0638012	0.000893046
Weight	-0.187641	0.0062089	0.47029	0.422135	-0.030017	-0.0254192	1	0.252202	0.901747	0.0239127
Height	-0.161371	-0.333643	-0.103086	-0.0127698	0.0395006	0.0688185	0.252202	1	-0.123746	0.094528
Body mass index	-0.111822	0.174845	0.519562	0.480798	-0.0848966	-0.0638012	0.901747	-0.123746	1	-0.030359
Absenteeism time in hours	0.184256	-0.0821255	-0.0525903	-0.0133345	0.12922	0.000893046	0.0239127	0.094528	-0.030359	1

2.5 Feature Scaling:

Features will be in different ranges. Feature Scaling is a technique used to limit the range of features.

First, I'm plotting the data to see the shape of each feature. I'm performing normalization on the skewed features. Then I'm performing standardization on the normally (symmetry) distributed features. Thus, the values in the features are scaled down. Now the train & test data are cleaned properly and kept ready for modelling.



Chapter 3

Modelling

3.1 Linear Regression: (optional) To find the significant and impacting features.

```
                                                                                           t value Pr(>|t|)
(Intercept)                                                                                  2.616  0.00909 **
Transportation.expense                                                                       3.165  0.00162 **
Distance.from.Residence.to.Work                                                             -4.018 6.51e-05 ***
Service.time                                                                                 3.166  0.00161 **
Body.mass.index                                                                             -2.186  0.02918 *
Reason.for.absence_unknown                                                                  -2.156  0.03146 *
`Reason.for.absence_medical consultation`                                                   -3.007  0.00274 **
`Reason.for.absence_Injury, poisoning and certain other consequences of external causes`    5.667 2.12e-08 ***
`Reason.for.absence_Diseases of the musculoskeletal system and connective tissue`           4.956 9.04e-07 ***
`Reason.for.absence_dental consultation`                                                    -2.824  0.00488 **
`Reason.for.absence_Diseases of the nervous system`                                          3.280  0.00109 **
`Reason.for.absence_Diseases of the skin and subcutaneous tissue`                            3.929 9.36e-05 ***
`Reason.for.absence_Diseases of the circulatory system`                                      5.371 1.07e-07 ***
Day.of.the.week_thurs                                                                       -2.589  0.00984 **
Education_postgraduate                                                                      -2.482  0.01331 *
---
```

The above model may help us t0 understand the significant features that impact the absenteeism.

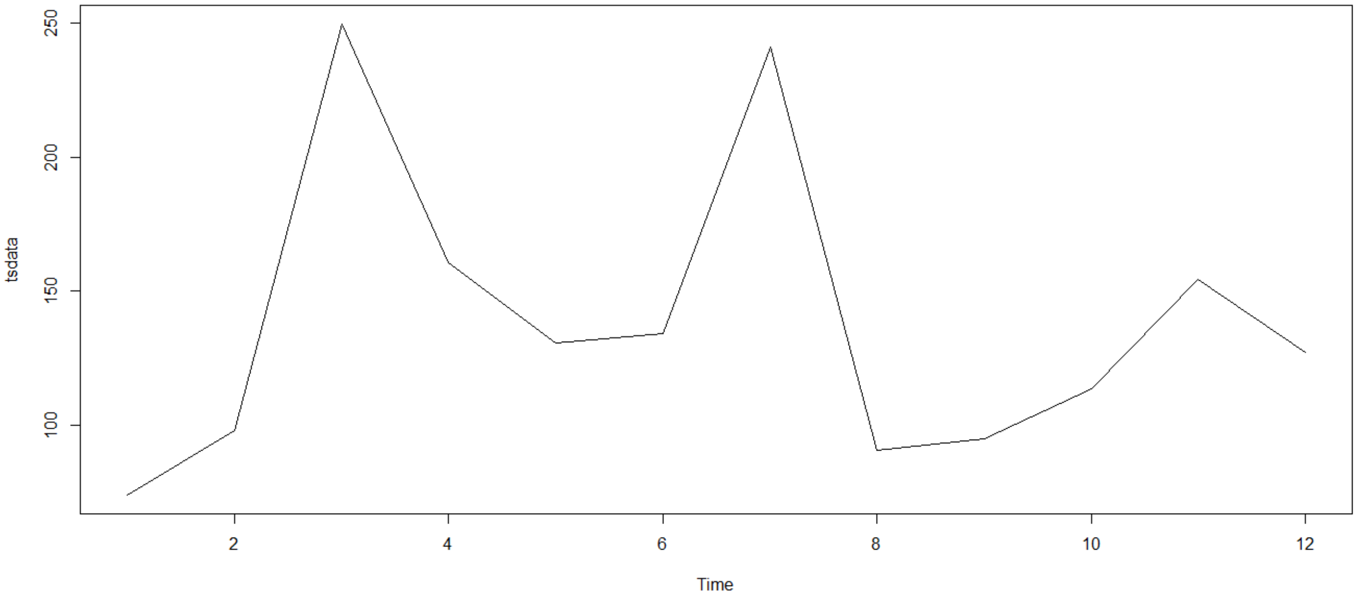
The following is the variable importance table.

Overall	features
5.666725	`Reason.for.absence_Injury, poisoning and certain other...
5.370600	`Reason.for.absence_Diseases of the circulatory system`
4.955626	`Reason.for.absence_Diseases of the musculoskeletal sys...
4.017863	Distance.from.Residence.to.Work
3.929437	`Reason.for.absence_Diseases of the skin and subcutane...
3.279734	`Reason.for.absence_Diseases of the nervous system`
3.165657	Service.time
3.165337	Transportation.expense
3.006642	`Reason.for.absence_medical consultation`
2.823613	`Reason.for.absence_dental consultation`
2.588605	Day.of.the.week_thurs
2.481728	Education_postgraduate
2.185522	Body.mass.index

- It shows that more no. of absent happens due to health reasons.
- Also, if the employee’s residence is far away from workplace, then it leads to more absenteeism.
- But we can’t conclude with this alone.

3.2 Time Series Analysis: Time Series Visualization:

Time Series Plots from R:

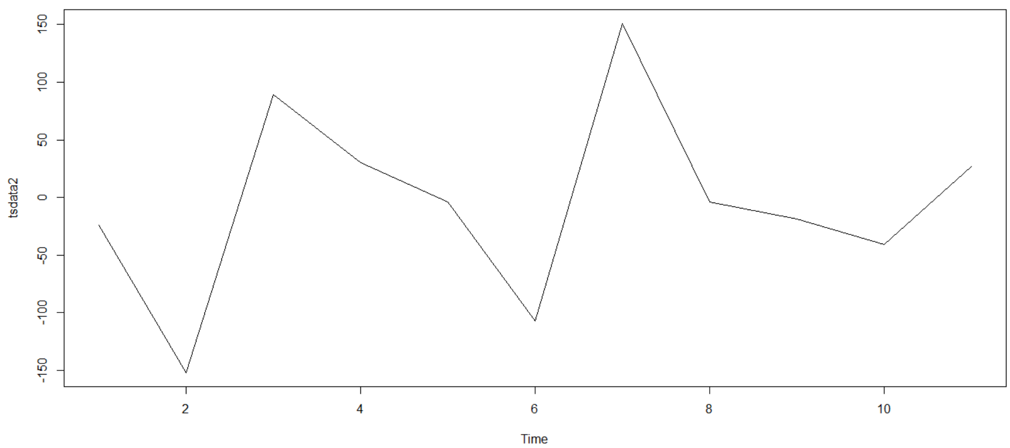


Stationarity test:

```
        Augmented Dickey-Fuller Test

data:  tsdata
Dickey-Fuller = -3.3957, Lag order = 0, p-value = 0.07838
alternative hypothesis: stationary
```

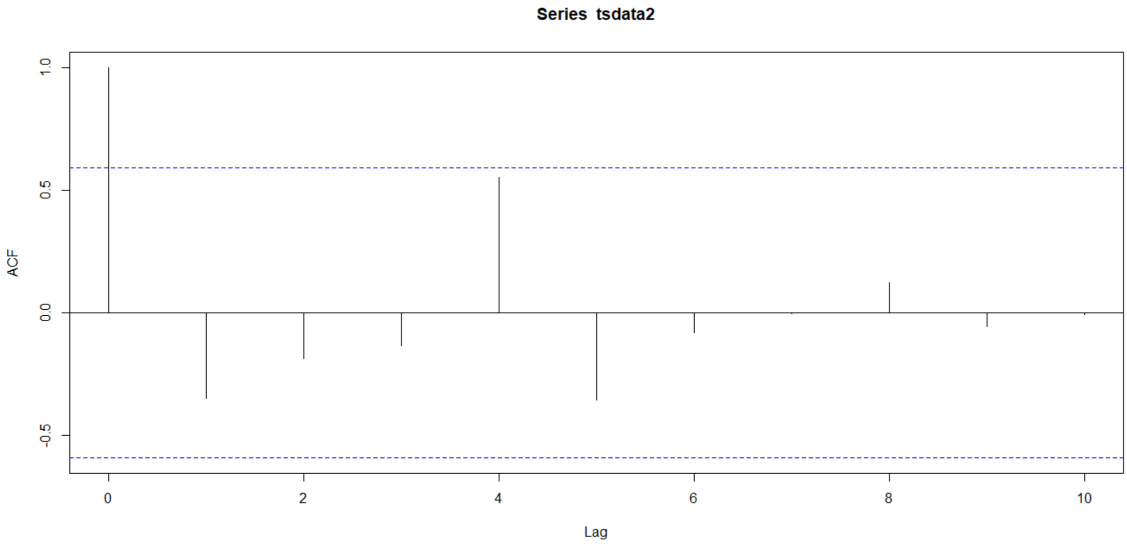
After taking log

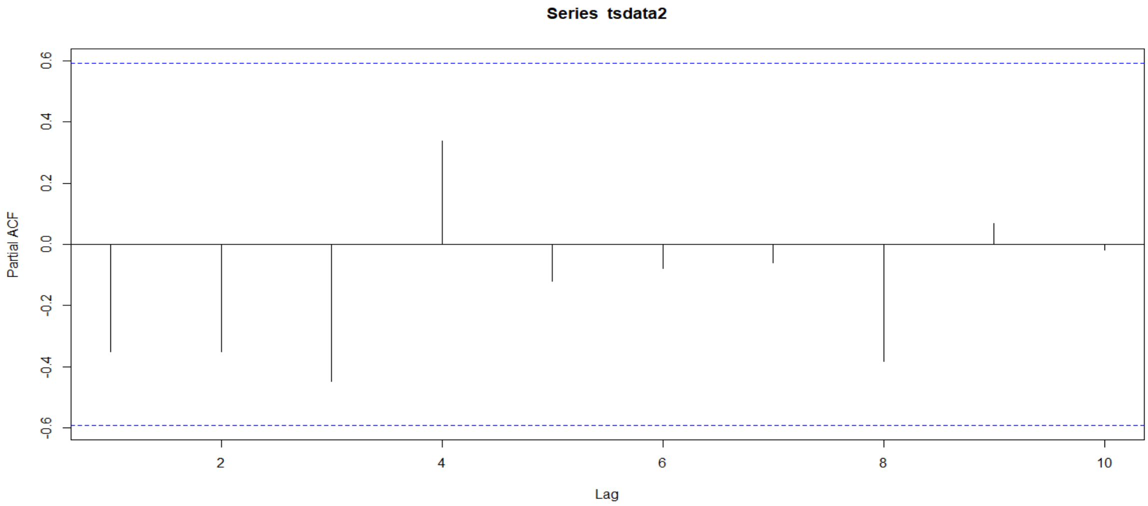
```
        Augmented Dickey-Fuller Test

data:  tsdata2
Dickey-Fuller = -3.9356, Lag order = 0, p-value = 0.02603
alternative hypothesis: stationary
```



ACF & PACF Plots:

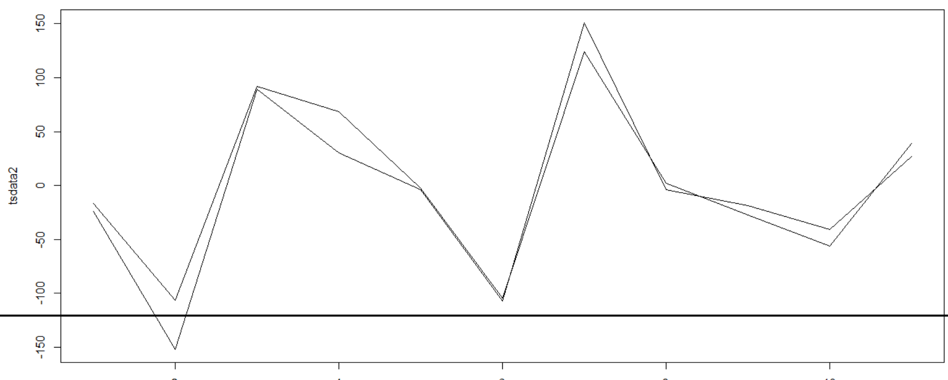
Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) were plotted to find the values of p (AR order) and q (MA order).





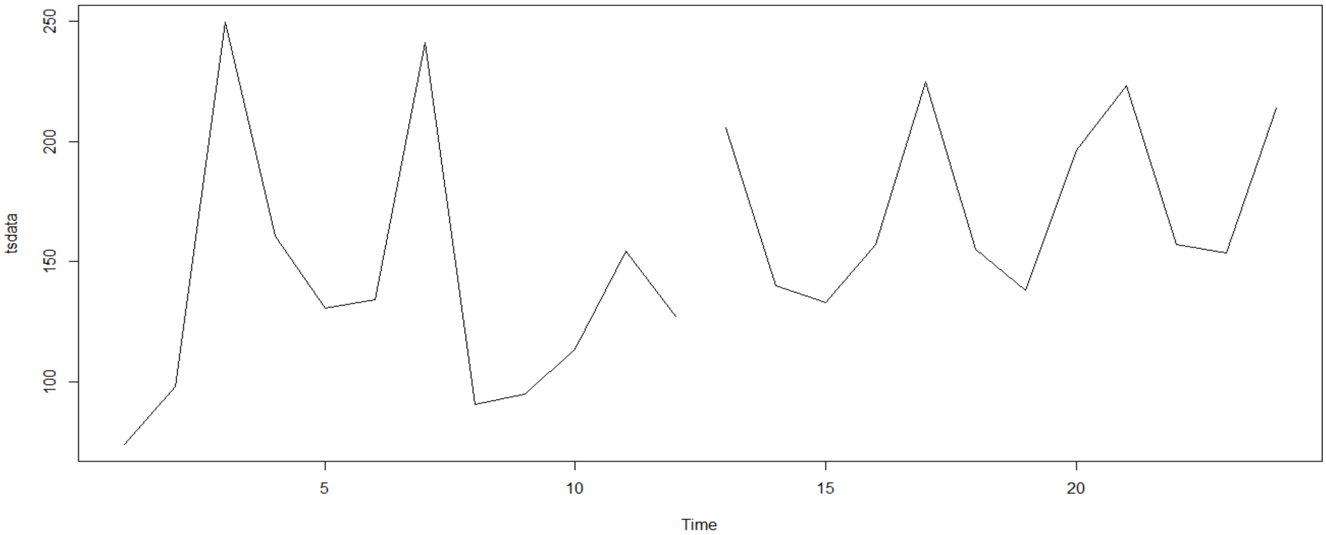
- Value of q (MA order) should be 0.
- We will have to check for several combinations of p & q to decide which model is best for forecasting.
- ARIMA model was applied for several combinations of p, d, q and Residual Sum of Squares (RSS) was calculated to check which combination gives lowest RSS.
- ARIMA with order=(4,0,9) gives us lowest RSS of 2222.32 so order=(4,0,9).

Plot of Time series and fitted values:

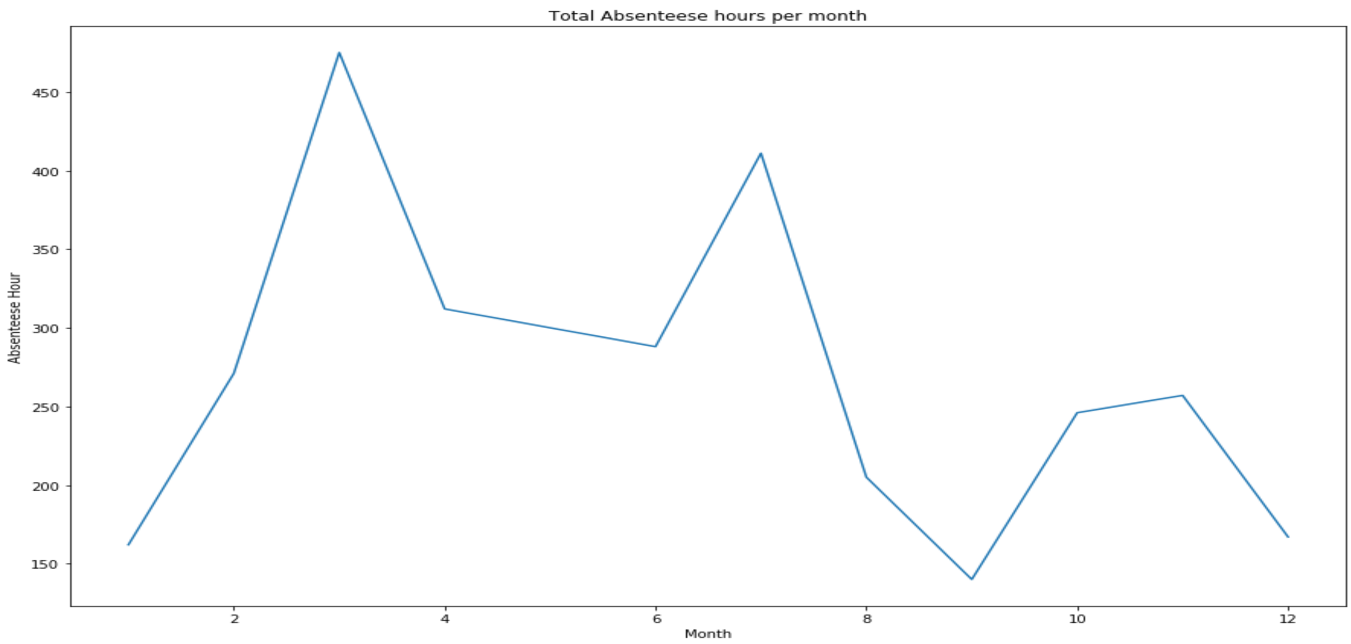


Time Series Forecasting for months in 2011:

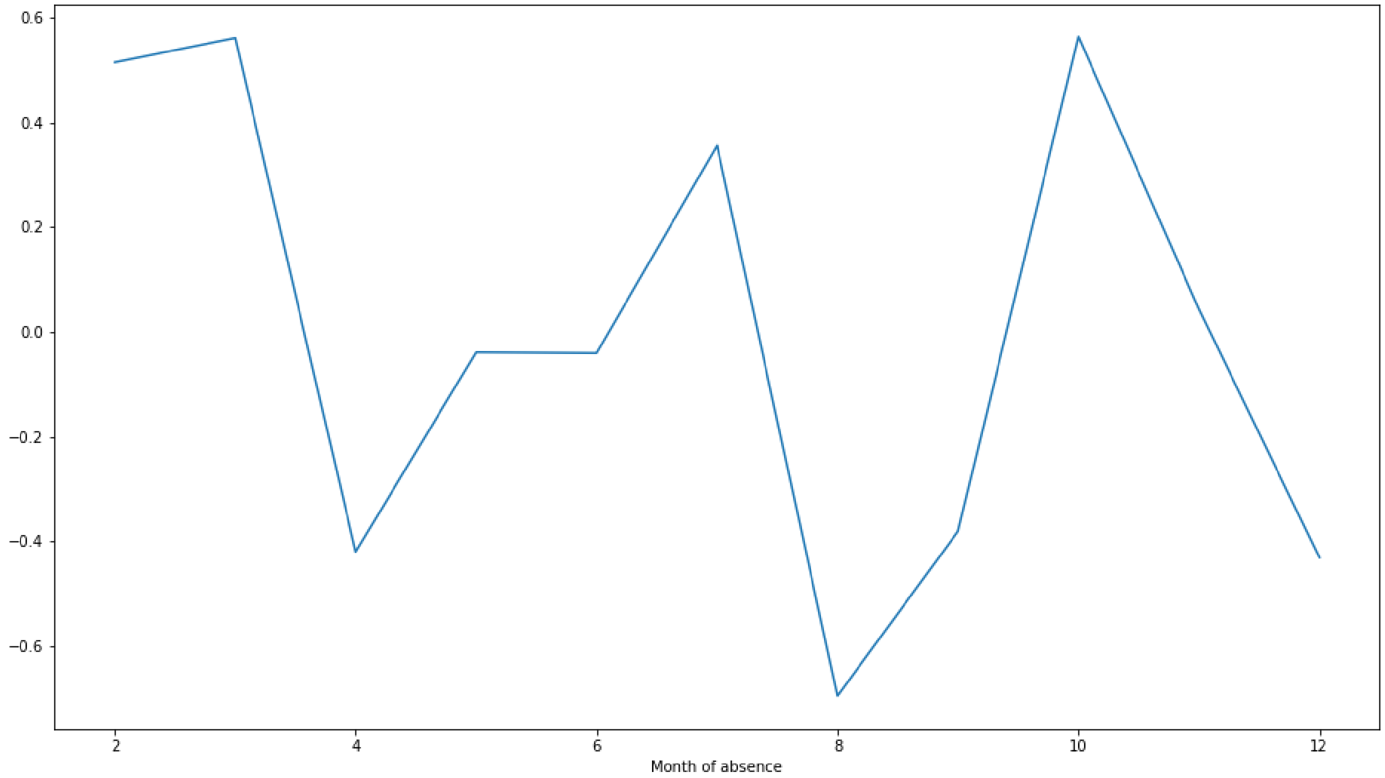
```
Time Series:
Start = 13
End = 24
Frequency = 1
 [1] 205.8994 139.8678 132.8639 156.9400 225.0463 155.1189 138.1218 196.3132 223.2284 157.1628 153.5244 213.9632
```



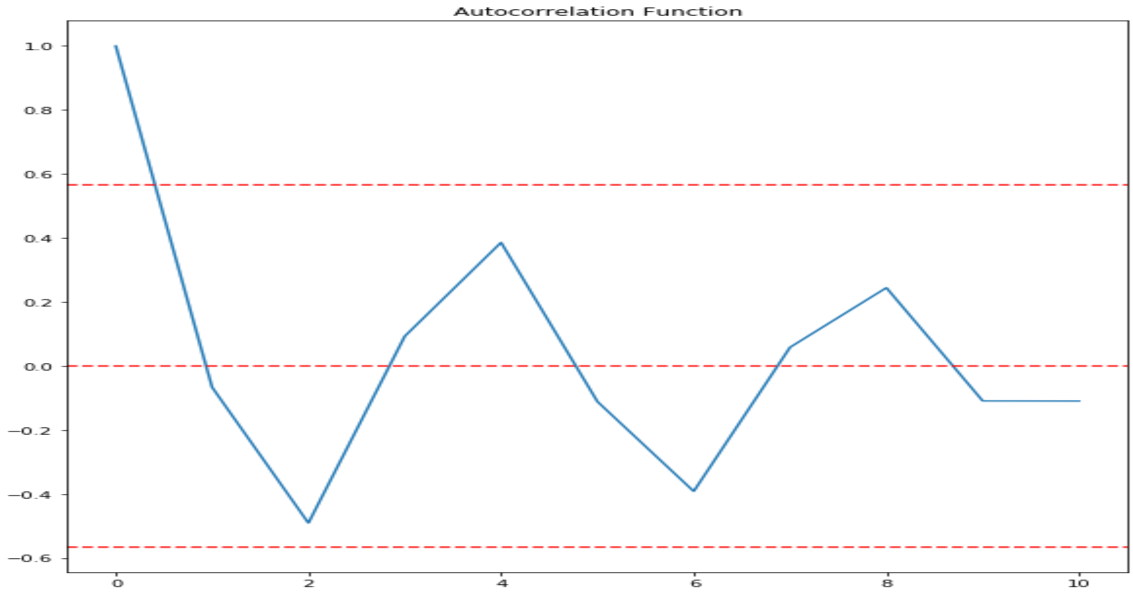
Time Series Plots from Python:

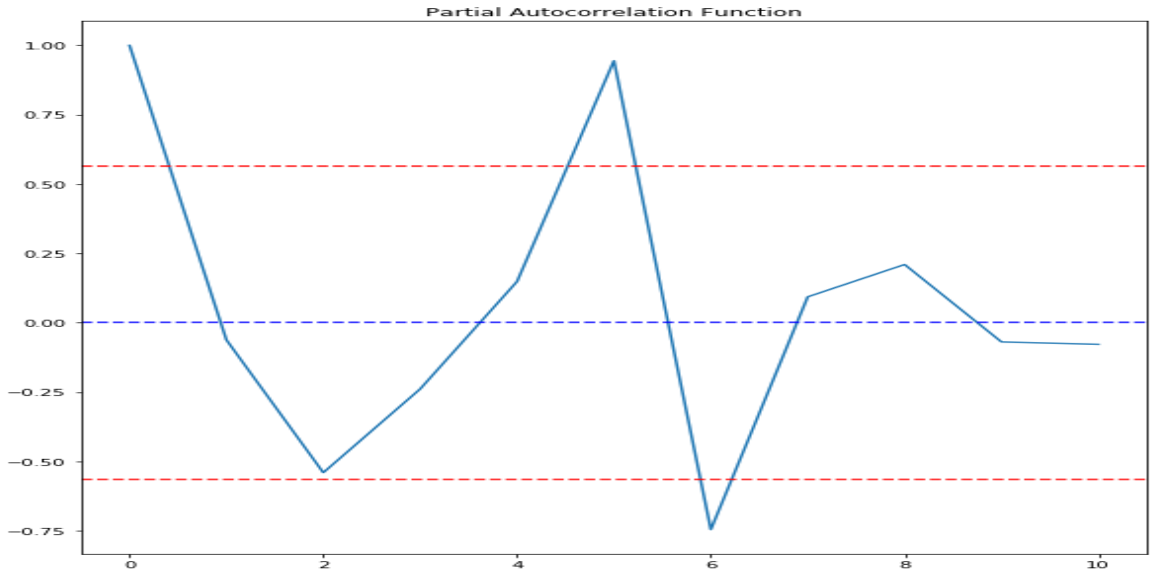


After taking log

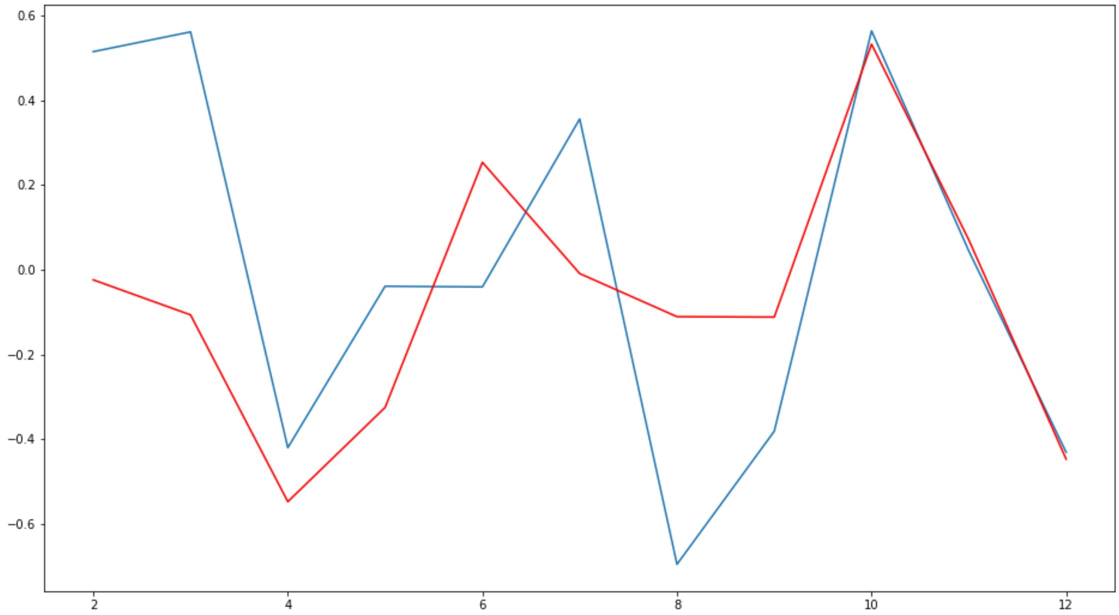


ACF & PACF Plots:

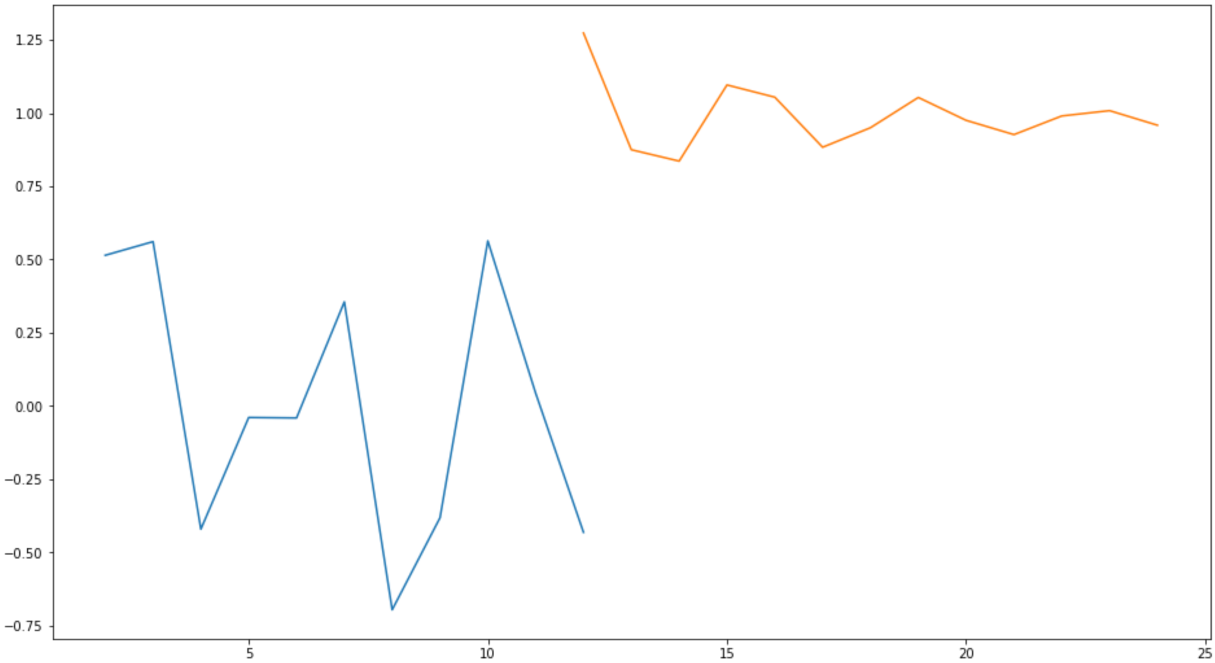




Plot of Time series and fitted values:



Time Series Forecasting for months in 2011:f



Forecasted values:

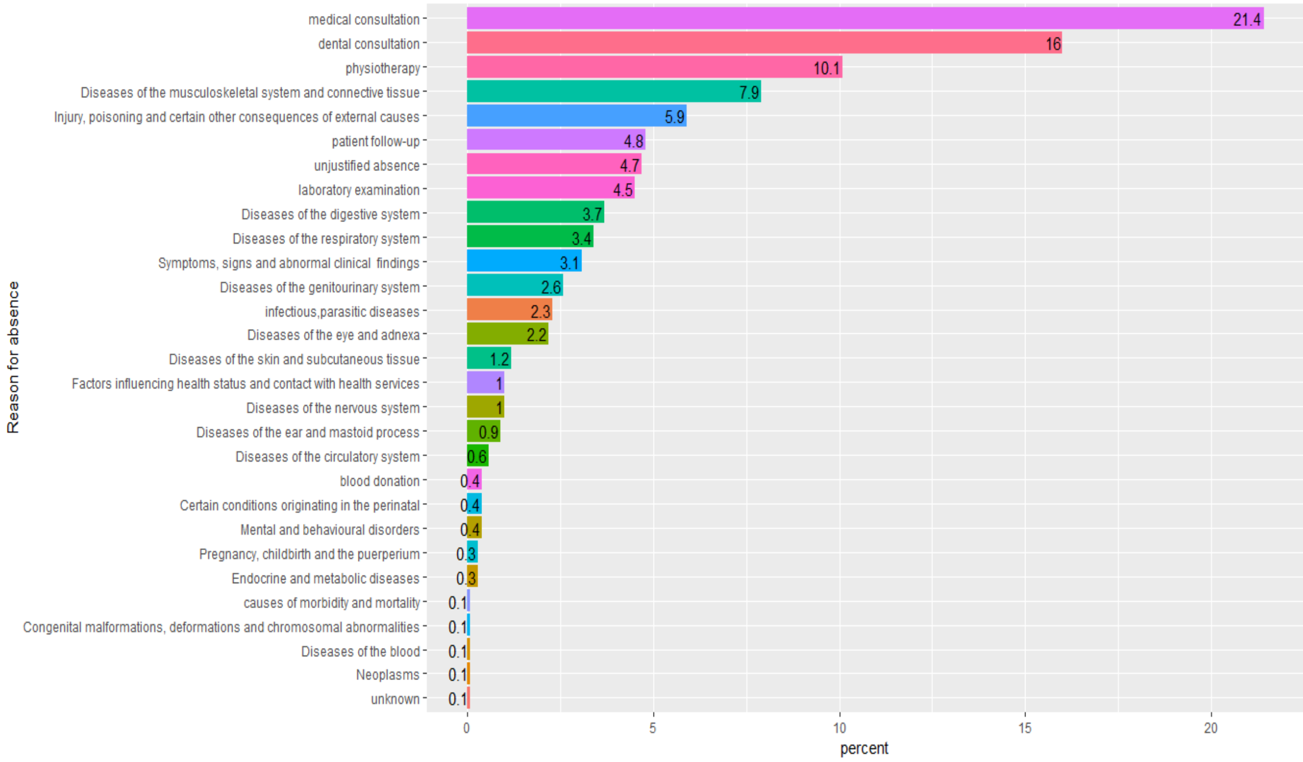
```
13    135.368574
14    113.173529
15    124.047176
16    130.779693
17    115.527548
18    109.753864
19    115.605487
20    112.745599
21    104.461132
22    103.435529
23    104.295701
24     99.975984
dtype: float64
```

Chapter 4

Visualization Insights:

Independent Variable: Reason for Absence

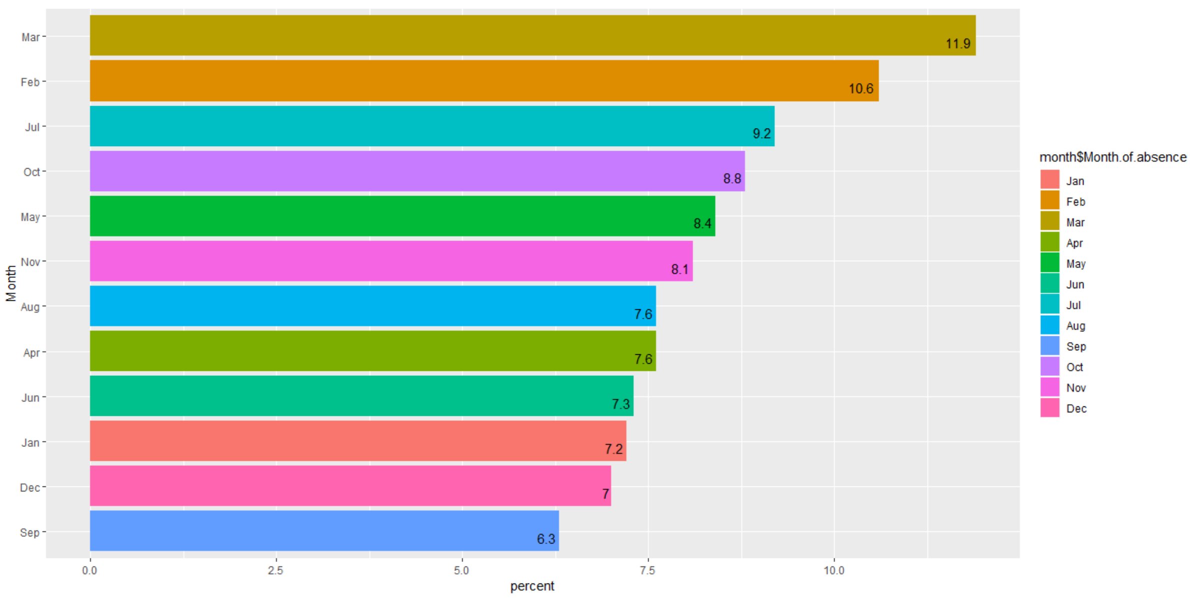
The below plots show the proportion of each categorical predictor to the target variable.



Insights:

- Medical Consultation & Dental Consultation are the major reasons for high absent rate.
- The XYZ company can organize the medical camp for the wellness of the employees. This may reduce the absent rate.

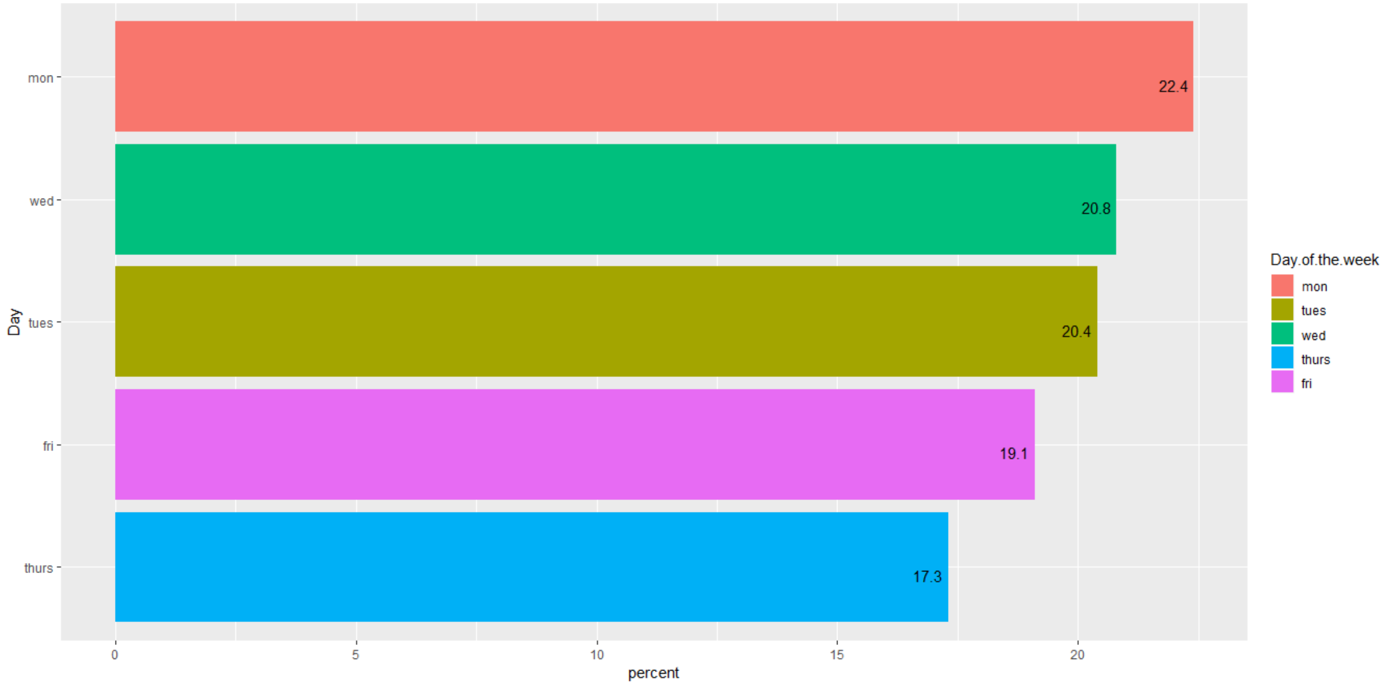
Independent Variable: Month of Absence



Insights:

- March month has high absent rate of 12%
- Also, February month has second highest absent rate of 11%
- It seems that February & March has highest rate of absenteeism, It's important to know the reason for high absent rate in these months.

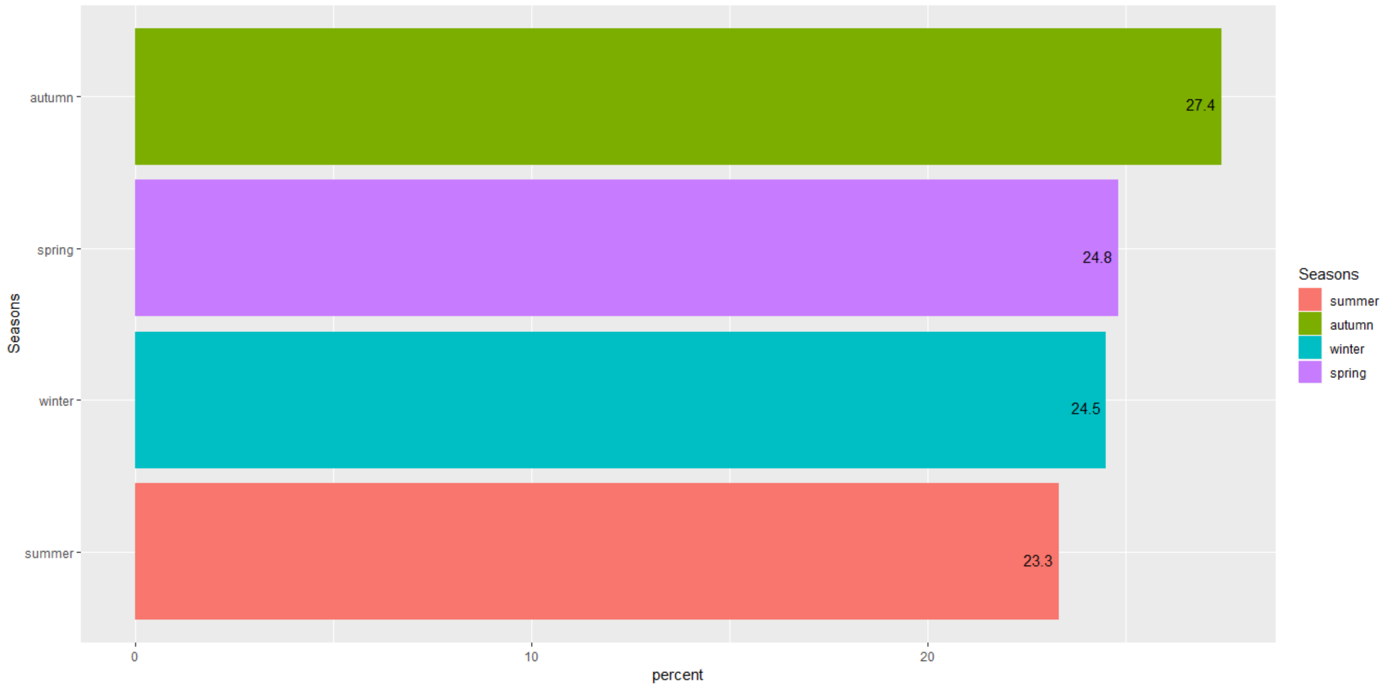
Independent Variable: Day of week



Insights:

- Monday has high absent rate of 22.4%
- Then Wednesday has 21 % of absent rate.
- Less absents happen in Thursdays.

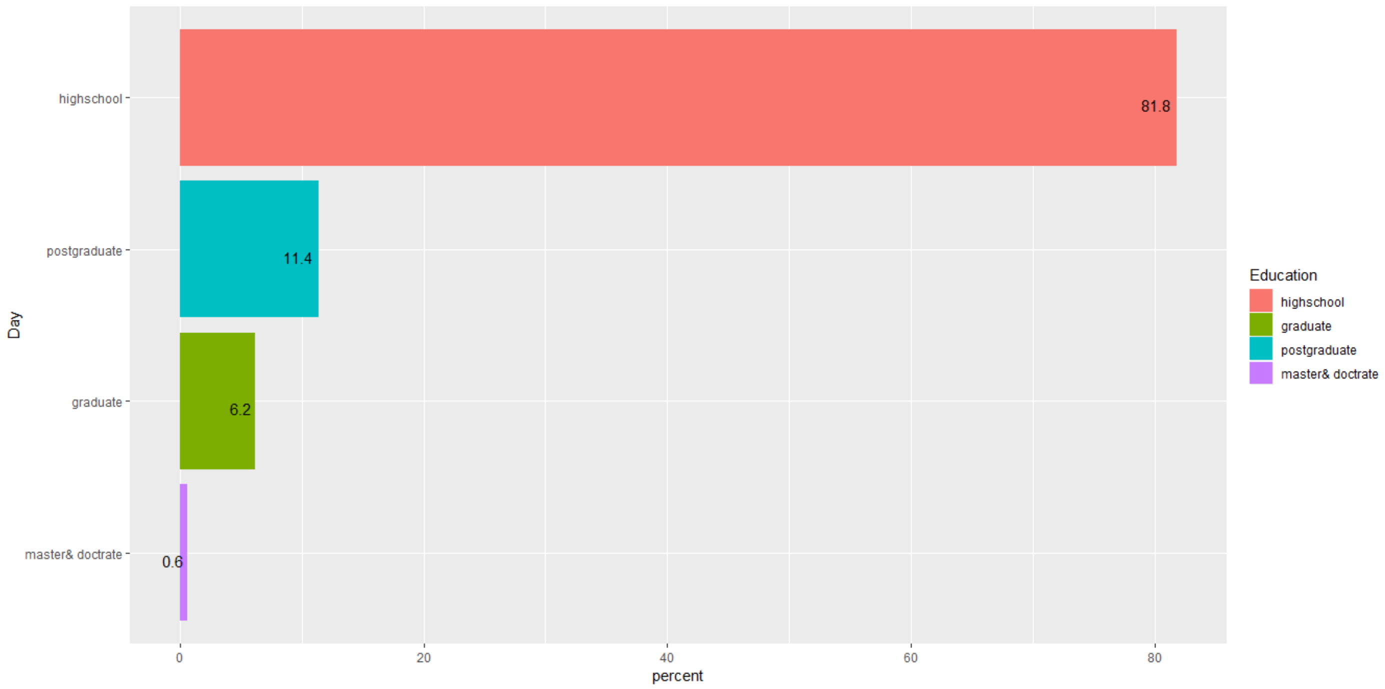
Independent Variable: Seasons



Insights:

- Autumn season has highest absent rate of 27.4%. After that the absent rate reduces in Spring, then in winter. Absent rate is very less in Summer.

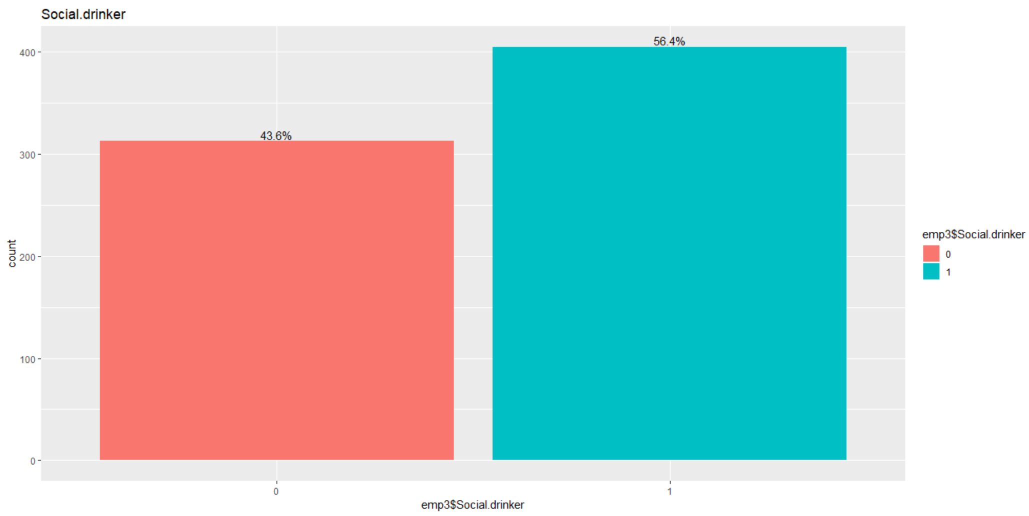
Independent Variable: Education



Insights:

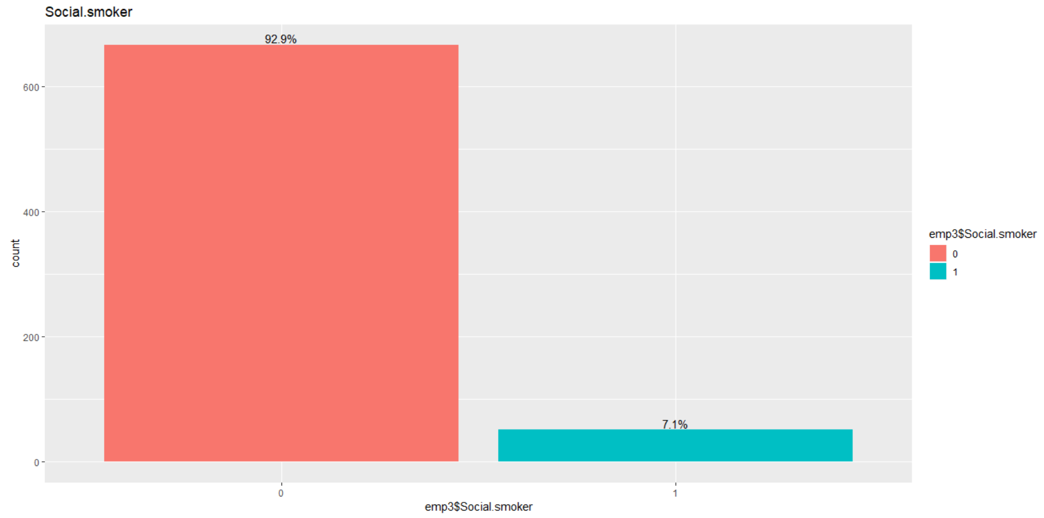
- Employees with high-school education has high absent rate. It's because the majority of the employees in the company are high school literates.

Independent Variable: Social drinker



- Majority of the employees (56%) in this company is Social drinker.

Independent Variable: Social Smoker



- 93% of employees in the company are not social smokers.

Chapter 5

Findings & Conclusion

Findings:

- It seems that the employee absenteeism will increase in the forthcoming year. Our forcasting results shows that employee absenteeism will increase in 2011.
- It's important to take proactive actions to reduce employee absenteeism in order to increase revenue to the business.
- At the same time benefits should be given to employees.
- It seems that more no. of absents occur due to health issues. So, the comoany has to organize health camp regularly for the well being of employees.
- Also, if the residence of employees are far away from the company, then it leads to absenteeism. So, the company can hire people whose residence is near to the office.
- **Based on the insights got from exploratory data analysis, we can derive many actions o reduce employee absenteeism.**

Conclusion:

At last, the values to be predicted by using a past data driven by ARIMA model according to the time based analysis.

The Time Series Analysis helped the XYZ Courier company to forecast the employee absent data for the forthcoming year 2011. An also found various features that impact absenteeiam. This helps to make proactive strategies to reduce employee abseteeism.

Attachments:

R file

Python file

Thank you…