Project

Employee Absenteeism

**Aditya Tiwari**

**Adi138250@gmail.com**

|  |
| --- |
| **Contents** |
| 1. Introduction    1. Problem Statement…………………………………………………………………………………………….1    2. Data…………………………………………………………………………………………………………………1-3 |
| 1. Methodology…………………………………………………………………………………………………………….4    1. Pre-Processing…………………………………………………………………………………………………….4   2.1.2 Missing Value Analysis………………………………………………………………………………………4  2.1.2 Outlier Analysis……………………………………………………………………………………………5-13  2.1.3 Feature Selection…………………………………………………………………………………………….14 |
| 1. Data Distribution Analysis …………………………………………………………………………………15-17 2. Dummies Creation…………………………………………………………………………………..................18 |
| 1. Modeling……………………………………………………………………………………………………….....…….19    1. Model Selection………………………………………………………………………………………………………19       1. Principal Component Analysis (PCA)…………………………………………………… 19       2. Decision Tree …………………………………………………….………………………………… 20       3. Random Forest …………………………………………………………………………………20-21       4. Linear Regression ……………………………………………………….……………………21-22 2. Conclusion………………………………………………….……………………………………………………………22   Appendix – A – Ans 1 …………………...........................………………………………………............…23  Ans 2……………………………….……………………………………………………………..24-29  Appendix – B – Python Code……………………………………………………………………………………30-44 |
| 1. References…………..………………………………………………………………………………………………....45 |

1. **Introduction**
2. **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. **Data**

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+ Fog

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)

Table : Employee Absenteeism Sample Data (Columns: 1-8)



Table : Employee Absenteeism Sample Data (Columns: 9-17)



Table : Employee Absenteeism Sample Data (Columns: 17-21)



Below are the variables present in Employee Absenteeism dataset

Table: Employee Absenteeism

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

**2. Methodology**

1. **Pre-Processing**

Any predictive modeling requires that we look at the data before we start modeling. We decided to simply remove few variables after loading data set but here we have dropped few variables after correlation test for continuous variables and anova test for categorical variables. Since our target variable is continuous and few variables are categorical hence we have applied anova test.

However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process, we will first check the presence of missing values in data set.

**2.1.1 Missing Value Analysis**

Missing values Analysis is required to be done so that we can check if there is any missing data. In case data is missing at few places we will impute those missing values by different methods in order to generate appropiate results. In our case we have missing values we will proceed to impute missing values. Below table illustrate missing values present in variables of the dataset.

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Variables** | **Number of Missing Values** |
| **1** | ID | 0 |
| **2** | Reason for absence | 3 |
| **3** | Month of absence | 1 |
| **4** | Day of the week | 0 |
| **5** | Seasons | 0 |
| **6** | Transportation expense | 7 |
| **7** | Distance from Residence to Work | 3 |
| **8** | Service time | 3 |
| **9** | Age | 3 |
| **10** | Work load Average/day | 10 |
| **11** | Hit target | 6 |
| **12** | Disciplinary failure | 6 |
| **13** | Education | 10 |
| **14** | Son | 6 |
| **15** | Social drinker | 3 |
| **16** | Social smoker | 4 |
| **17** | Pet | 2 |
| **18** | Weight | 1 |
| **19** | Height | 14 |
| **20** | Body mass index | 31 |
| **21** | Absenteeism time in hours | 22 |

In above table we can clearly see the number of missing values present in each variable. So, we have imputed these values. We applied mean, median and Knn method to impute , where we found that knn shows best values to impute, hence we made knn method in use to replace missing values with NA and later with knn.

**2.1.2 Outlier Analysis**

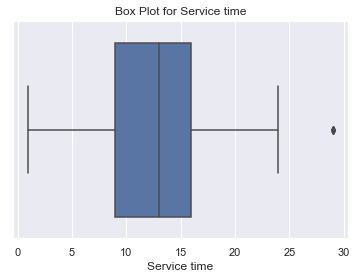
The Other steps of Preprocessing Technique is Outliers analysis, an outlier is an observation point that is distant from other observations. Outliers in data can be good and it can be bad as well. Here in our case we don’t want outliers the reason for removing these outliers instead of substituting them with other balancing values (such as mean, median or knn method) because we expect them to be relatively random values and replacing them with set values may cause inaccuracy in analysis later.

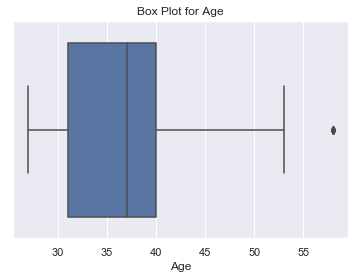
The outlier analysis is done by plotting the box plot. Boxplot is a method for graphically depicting groups of numerical data through their [quartiles](https://en.wikipedia.org/wiki/Quartile). Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles.

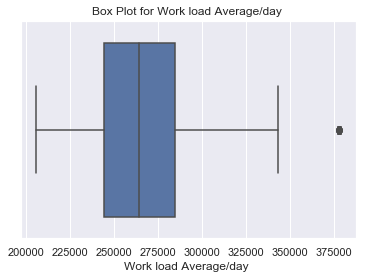
Fig. 1.0

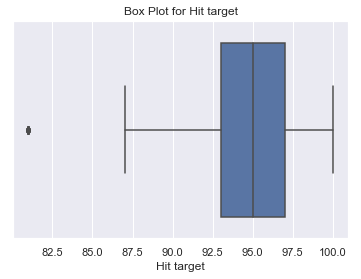


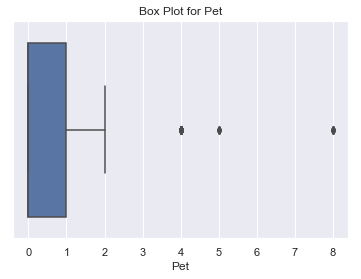


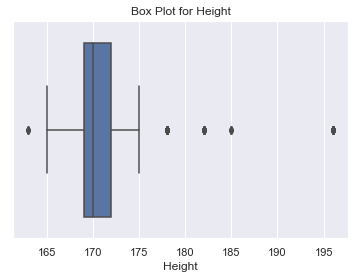


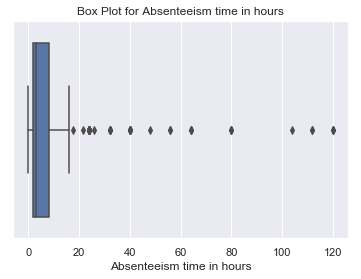






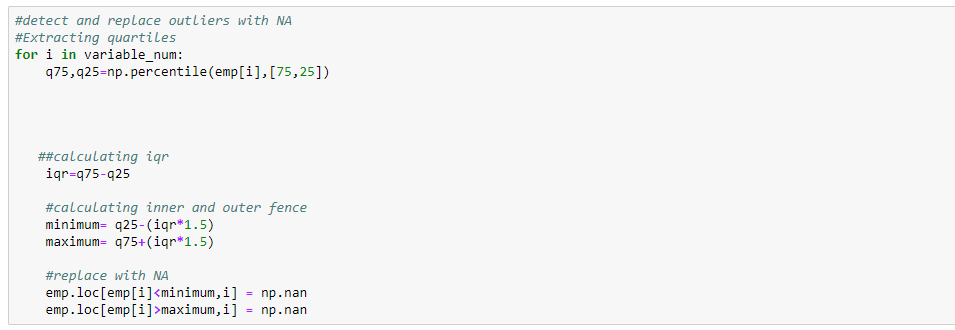


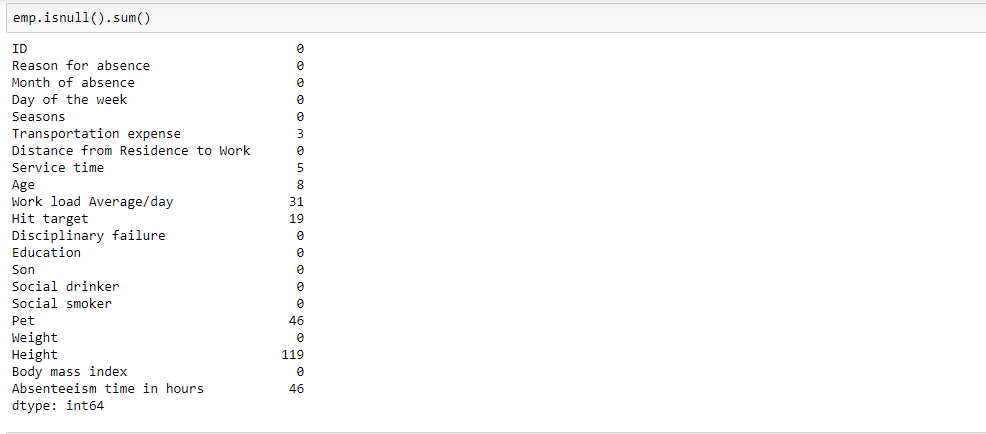


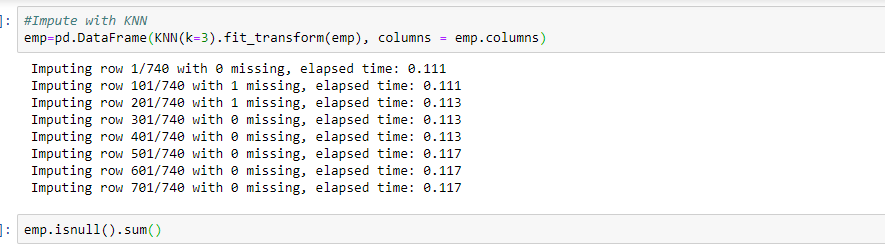


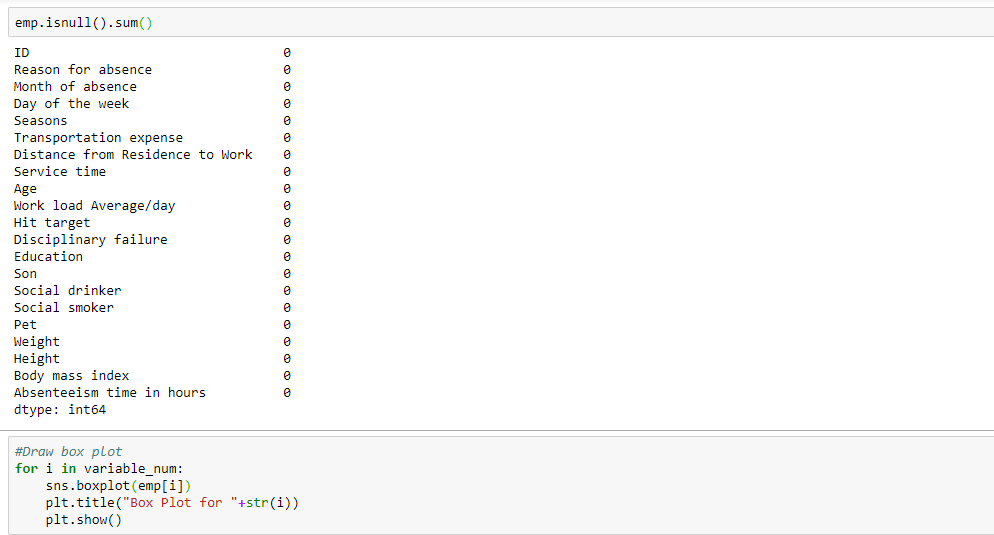
The above boxplots clearly shows the outliers present in the variables. So to treat the outlier we calculated upper and lower quartile then the interquartile range and then we created nan value in place of outliers which are later replaced by knn imputation method.

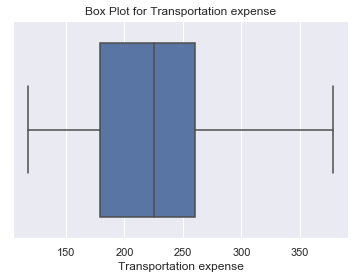
Boxplots after outlier treatment

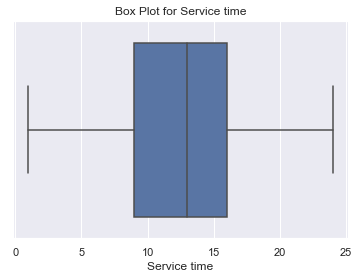


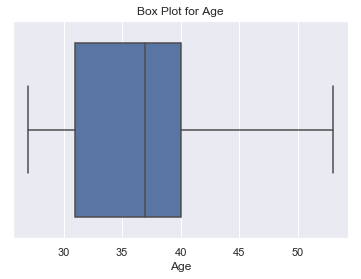


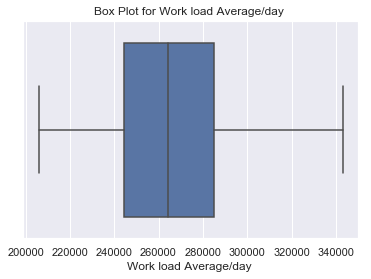


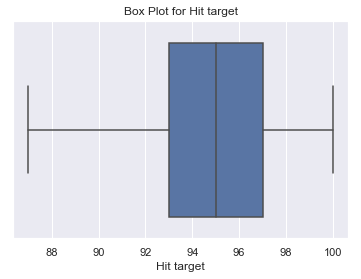


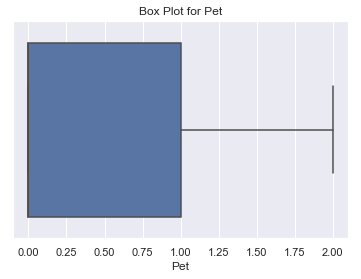


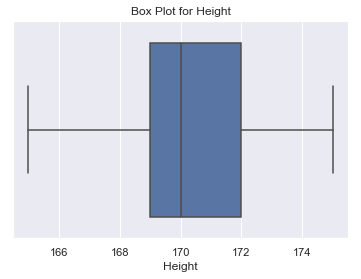












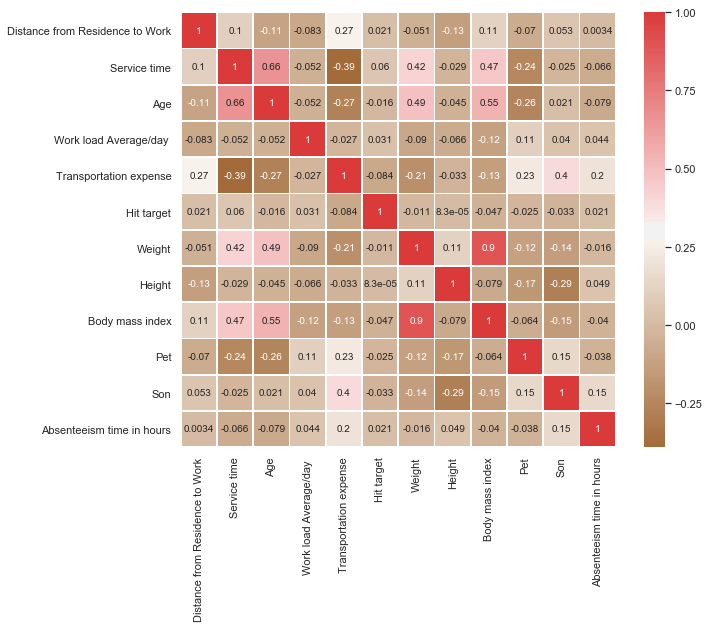
**2.1.3 Feature Selection**

Machine learning works on a simple rule of GIGO i.e. Garbage In Garbage Out. Here garbage refers to the noise or redundant values.

This becomes even more important when the number of features are very large. We need not use every feature at our disposal for creating an algorithm. We can assist our algorithm by feeding in only those features that are important. Feature subsets gives better results than complete set of features for the same algorithm or “Sometimes, less is better!”.

We should consider the selection of feature for model keeping in mind that there should be low correlation between two independent variables otherwise there will be problem of multicollinearity.

Fig. 3.0



From above correlation plot we can clearly see that variable Weight and Body mass index have high correlation variable Age and Service Time also have high correlation. It means that we must drop one variable out of two having high correlation. So in our study here we will drop variables Weight and Age.

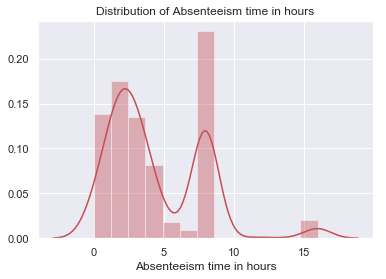
Color dark Red indicates there is strong positive correlation and if dark brown indicates negative correlation.

1. **Data Distribution**

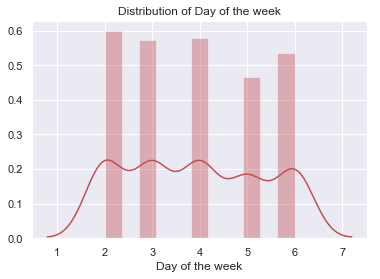
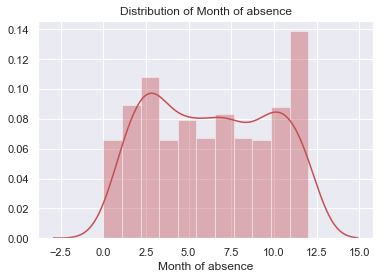
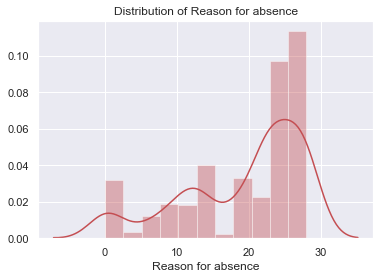
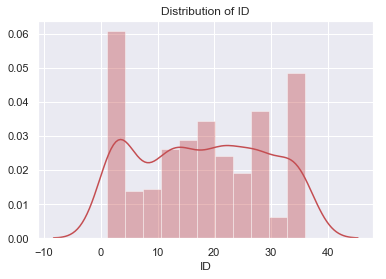
**Checking distribution of variables with help of distribution plot**

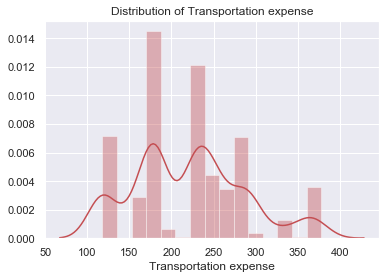
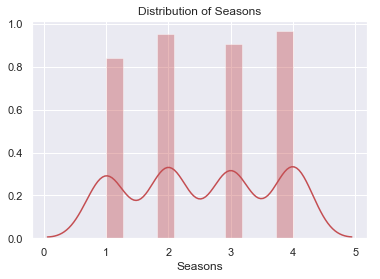
Distribution of target variable (Absenteeism time in hours)

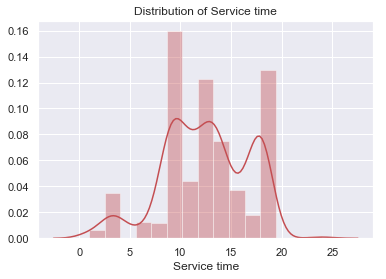
Fig. 4.0

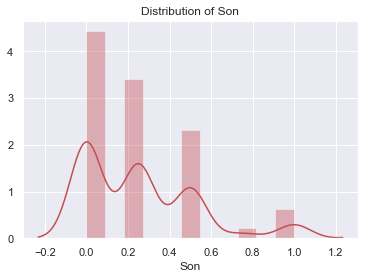
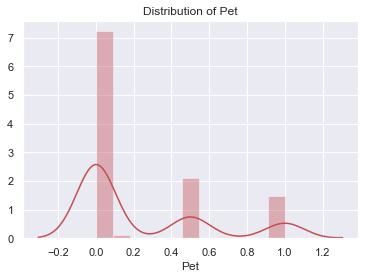
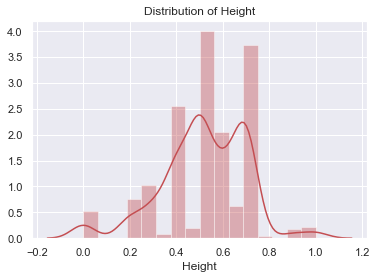
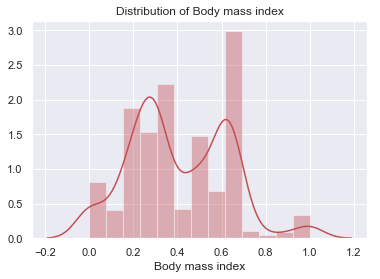
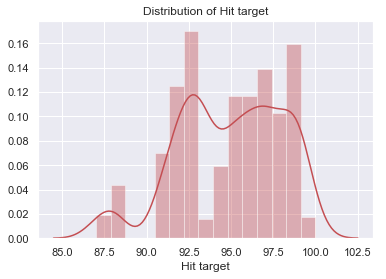
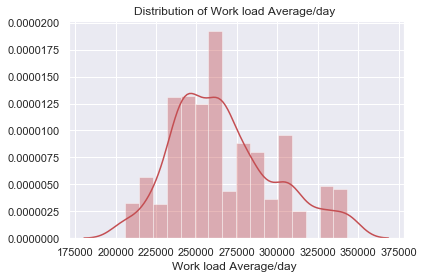


From above grapgh we can see that our target variable Absenteeism time in hours is not normally distributed.

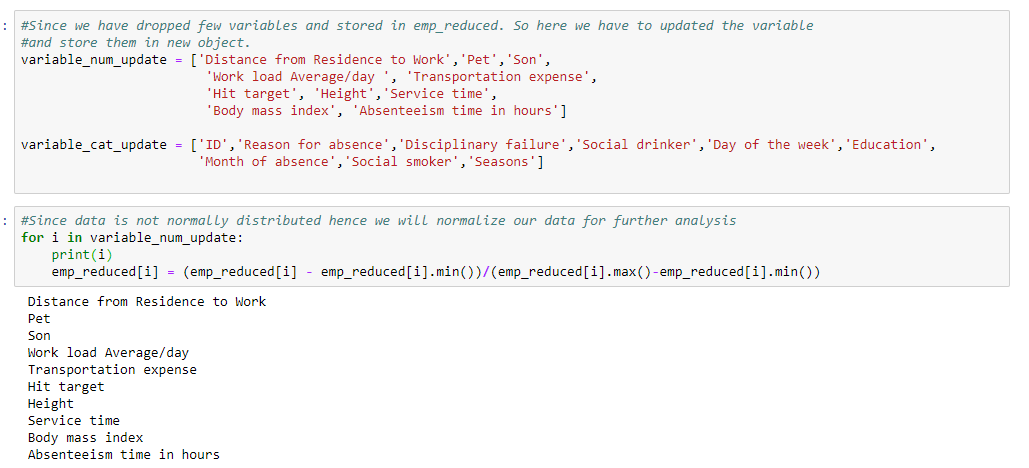






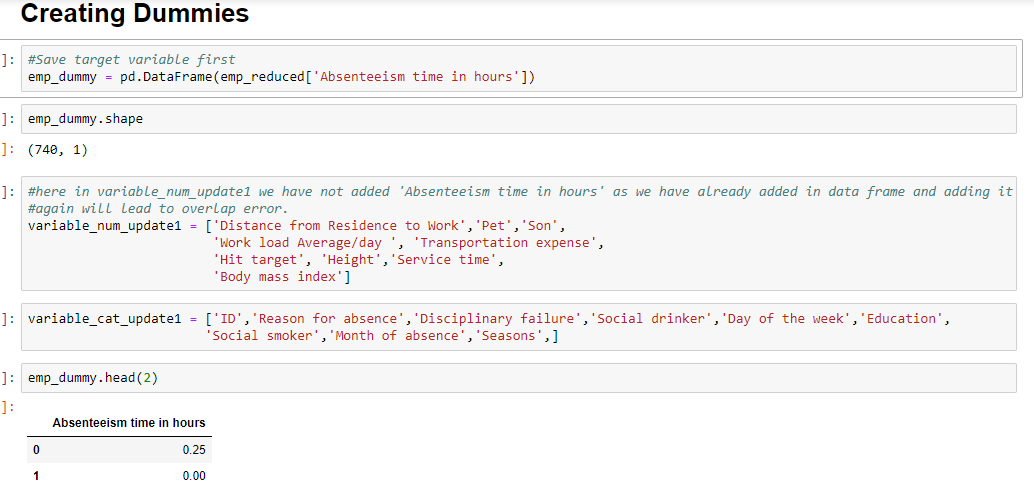


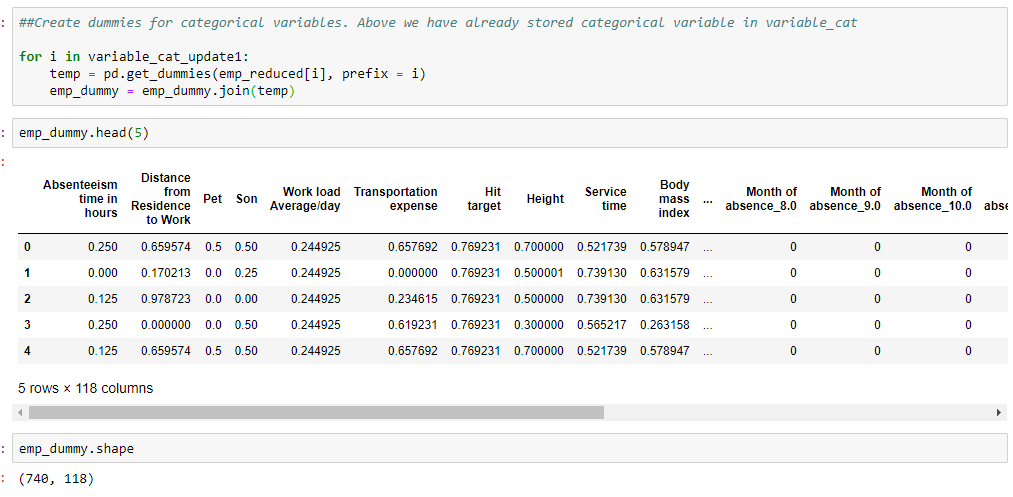
From above distribution plots we can clearly see that data is not normaly distributed. So, we will normalize data i.e. we will bring in range of 0 to 1 So that data gets normally distributed.



1. **Dummies Creation**

Since we have categorical variable also in our data set and we can’t make calculation on data having categorical variable. So, we have made use of dummy variable for categories which helps us to make calculations and build suitable model.



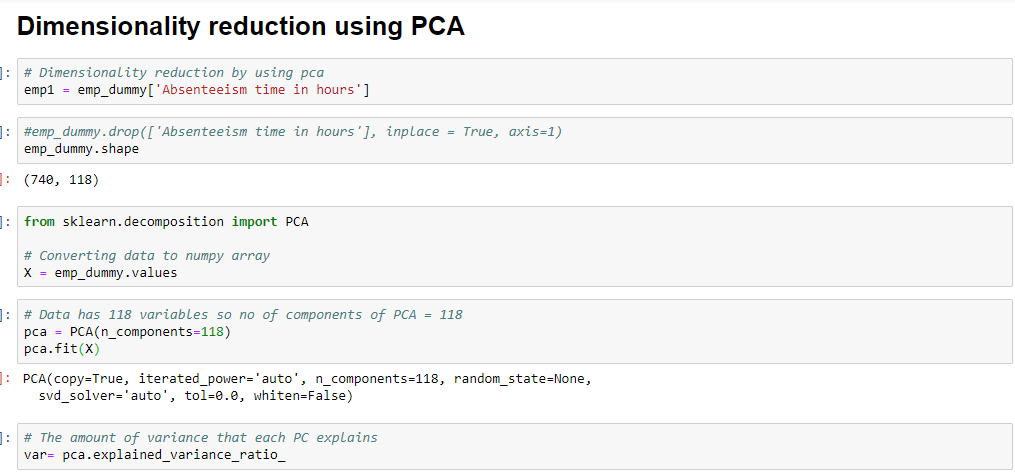


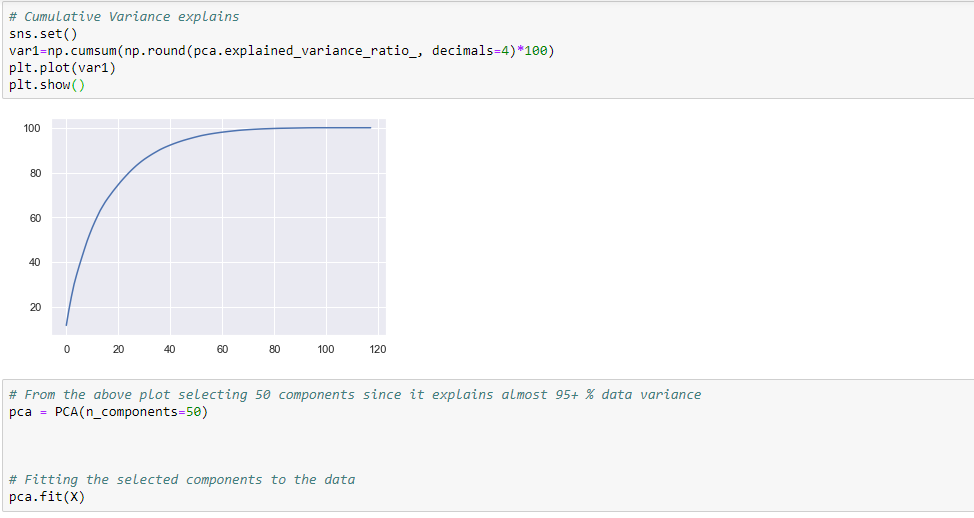
1. **Modelling**

**5.1 Model Selection**

Model Selection is a process of selecting the model which have better accuracy and can work on train and test data. We must select a model where algorithm works well and shows low error rate. We made Decision Tree, Random Forest and Linear regression model. Previously when we ran all three models the results were not so good. So we made use Principal component analysis which helped us to select the variables that explained variance of data. It happens that at a certain point the variables don’t show much of variance in data and can lead to complexity so in such case the PCA helps to select variables that actually show the variance.

**5.1.1 Principal Component Analysis (PCA)**

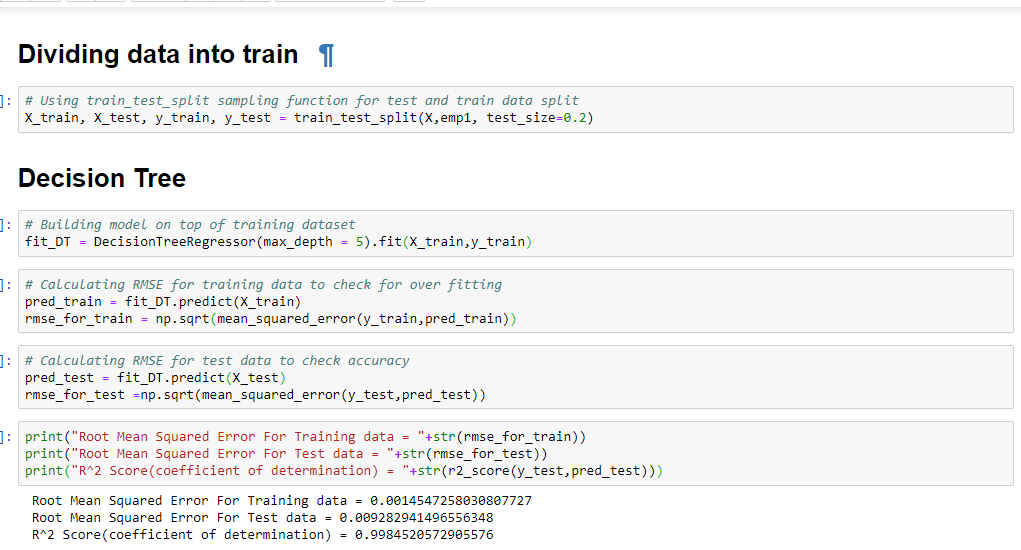


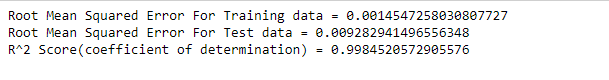


**5.1.2 Decision Tree**

A tree has many analogies in real life and turns out that it has influenced a wide area of **machine learning**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.

Decision Tree Algorithm





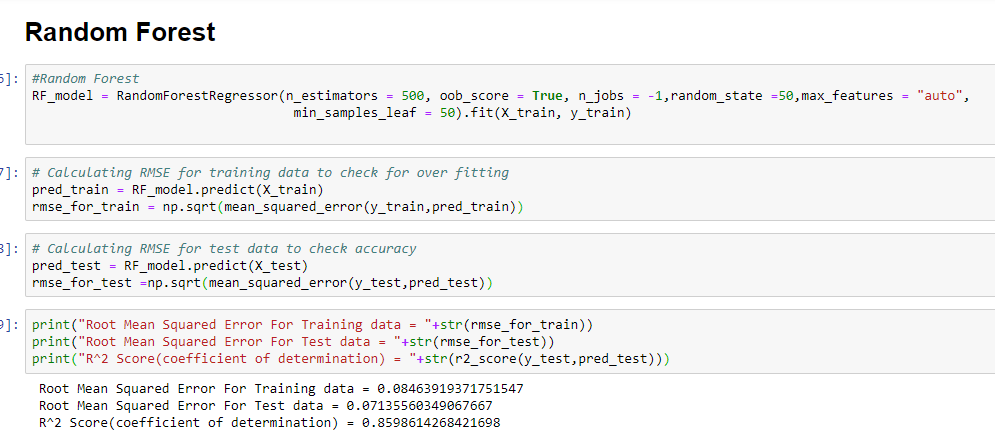
**5.1.3 Random Forest**

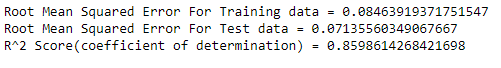
Random forests is an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis)  that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

Random forest functions in following way

* Draws a bootstrap sample from training data.
* For each sample grow a decision tree and at each node of the tree

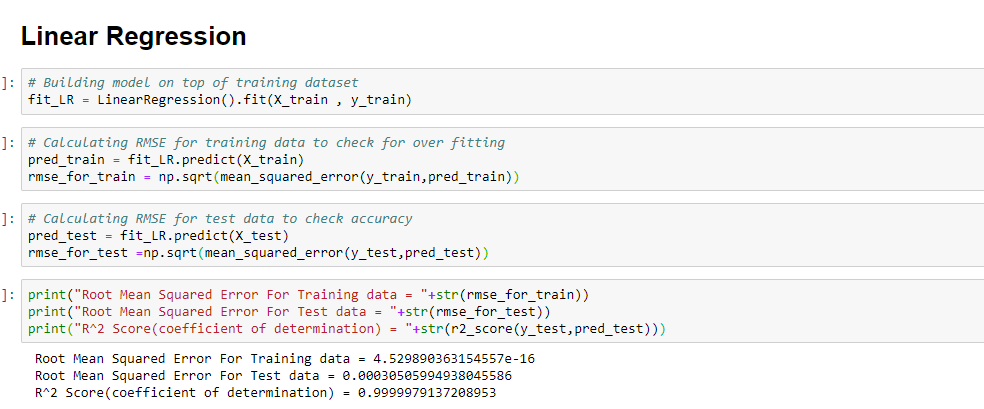
1. Ramdomly draws a subset of mtry variable and p total of features that are available
2. Picks the best variable and best split from the subset of mtry variable
3. Continues until the tree is fully grown.

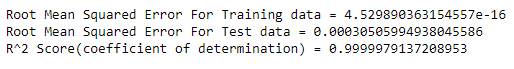
Random Forest Implementation 



**5.1.4 Linear Regression**

[Multiple linear regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-multiple-linear-regression/) is the most common form of linear regression analysis.  As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.  The independent variables can be continuous or categorical.





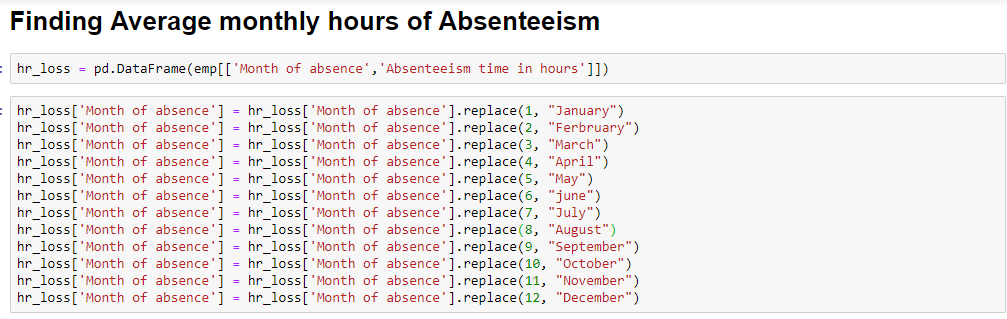
**R**-**squared** is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. 0% indicates that the model explains none of the variability of the response data around its mean.

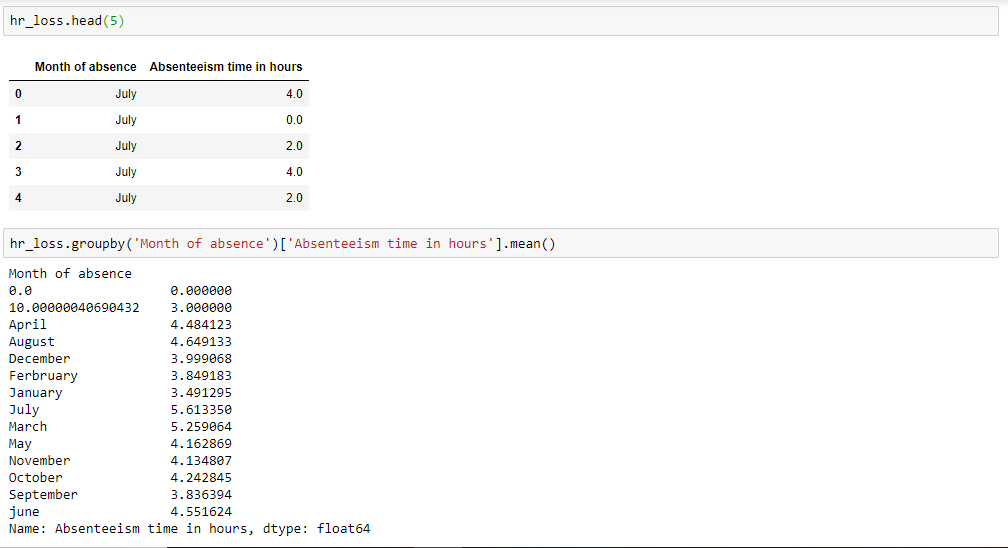
As R-Square value is 0.9999 and root mean squared error is also low both for training and test data hence linear regression model is well suited.

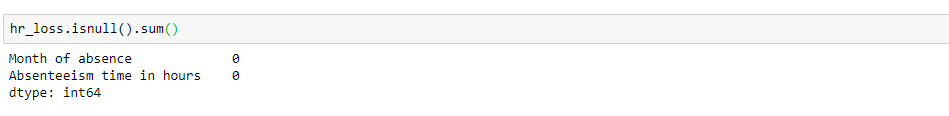
**6. Conclusion**: - For the Employee Absenteeism Linear regression Model is best model to predict the count.

**Appendix - A**

**Answer to first question monthly absenteeism in hours**

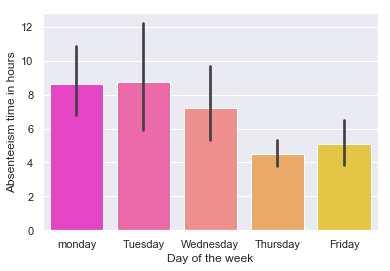


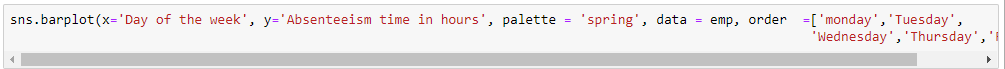




**Answer to second question where company must take action**

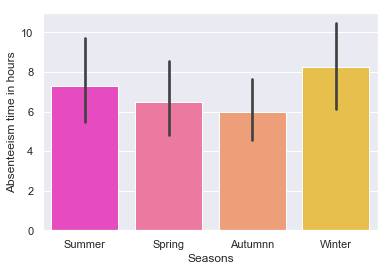
**Bar plot between Day of week and Absenteeism in hours**





From above bar plot we can see that on Tuesdays the absenteeism hours are highest. The company should see that why on Tuesday’s absenteeism time is so high.

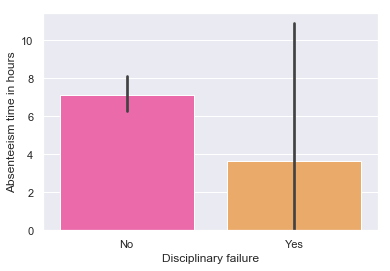
**Bar plot between Seasons and Absenteeism time in hours**





As we can see from above bar plot Absenteeism time in hours is more in winter season.

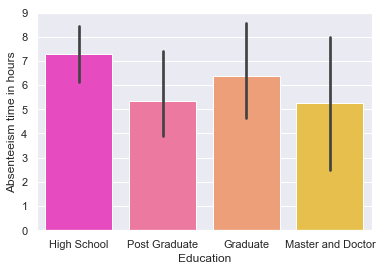
**Bar plot between Disciplinary failure and Absenteeism time in hours**





From above bar plot we can see that Disciplinary failure is not the reason for Absenteeism in hours.

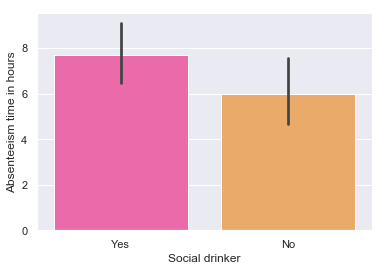
**Bar plot between Education and Absenteeism time in hours**





We can see from above bar plot that those who have education of high school have highest absenteeism time in hours which means company should watch them and should take action against them.

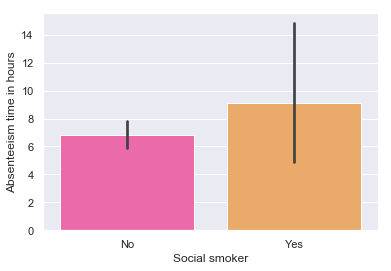
**Bar plot between Social drinker and Absenteeism time in hours**





From above plot we can clearly see that those who are social drinker are having a greater number of absenteeism hours which means the company should warn them for drinking habit that should not affect the company’s work and performance.

**Bar plot between Social smoker and Absenteeism time in hours**

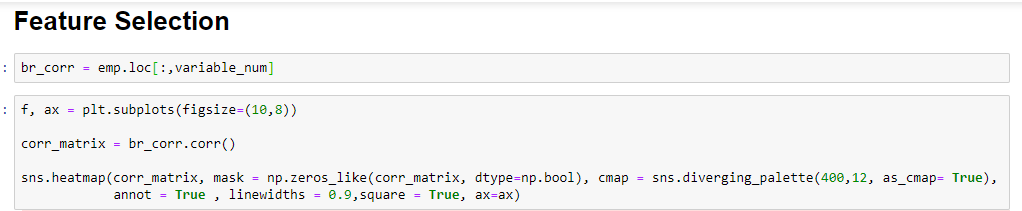




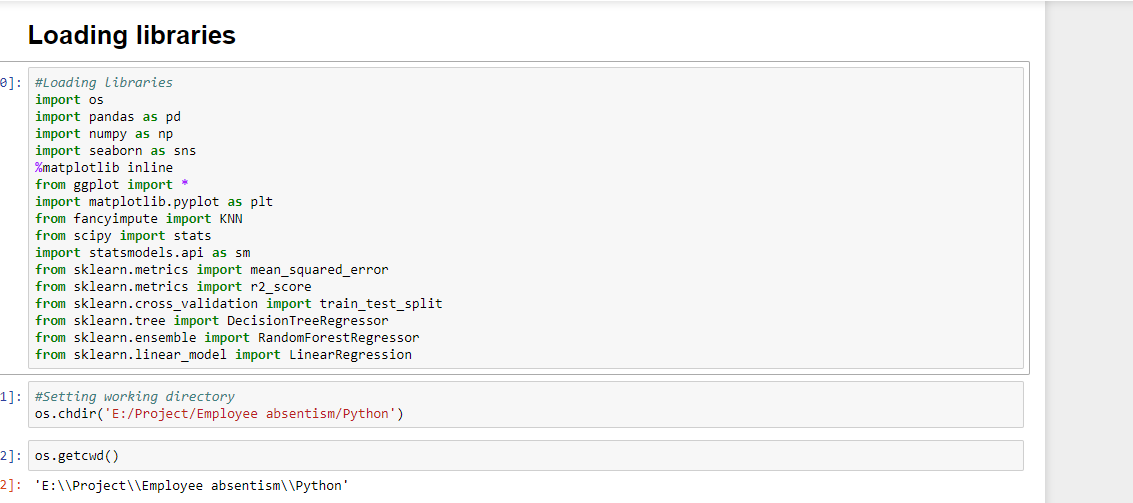
From above plot we can clearly see that those who are social smoker are having a greater number of absenteeism hours which means the company should warn them for smoking habit that should not affect the company’s work and performance.

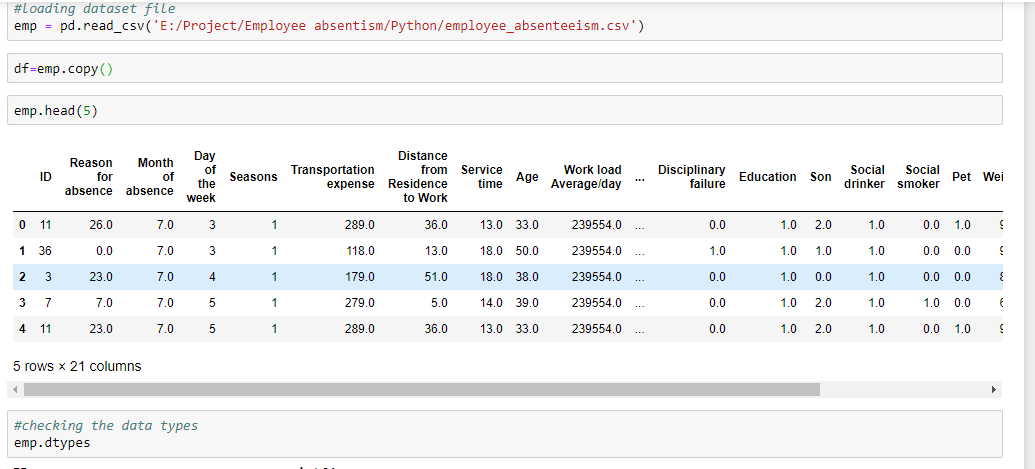
**Appendix- B - Python Code**

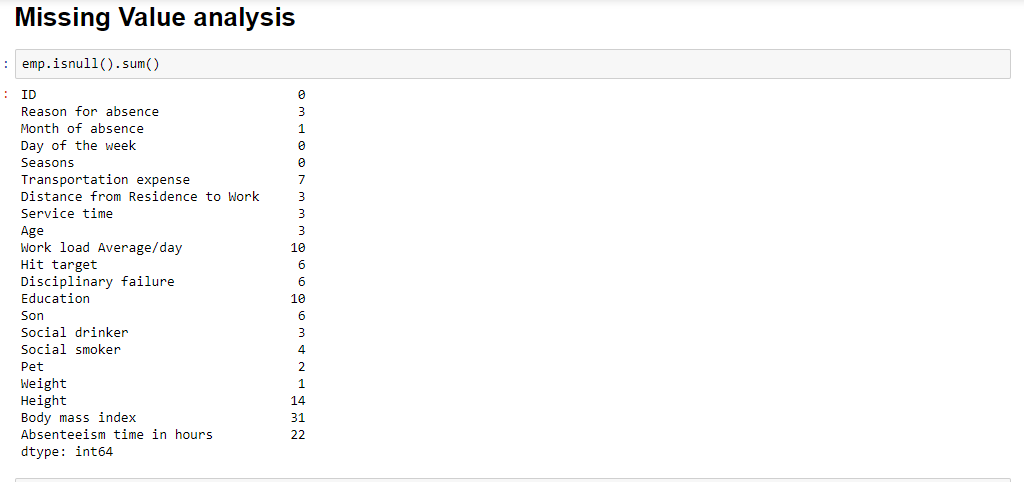
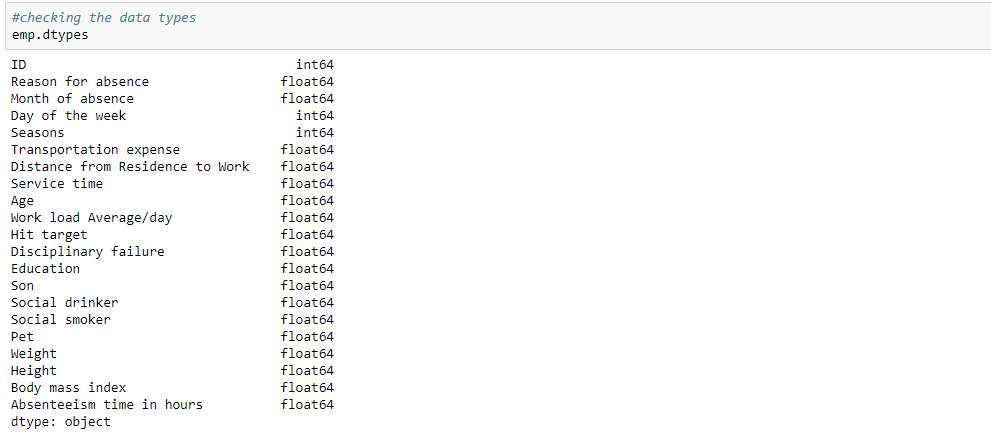
**Fig 3.0 Python Code**

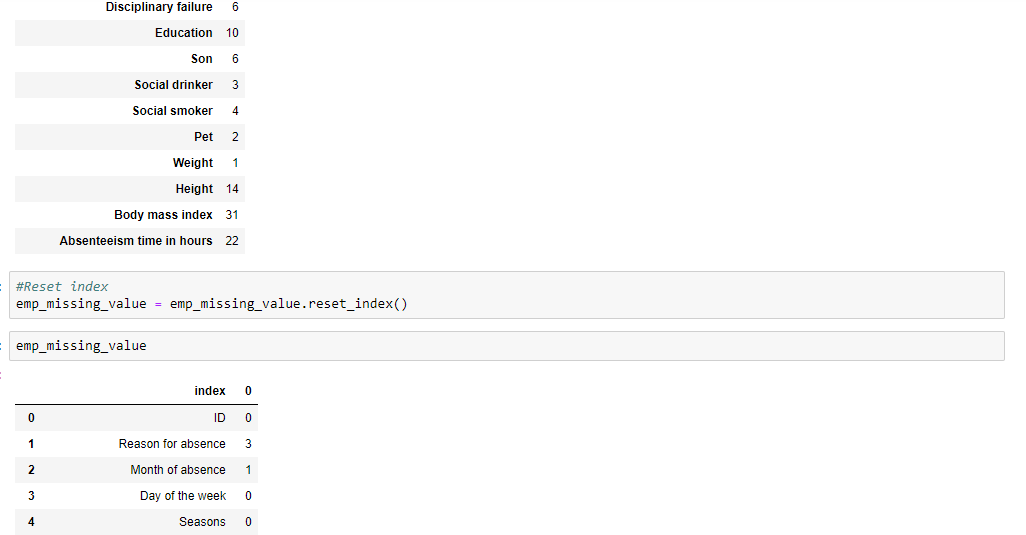
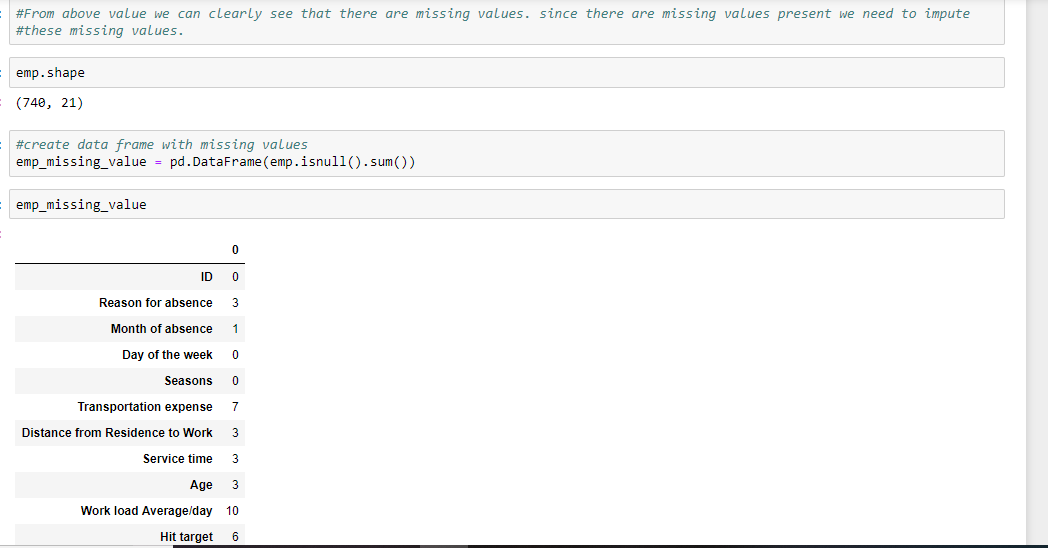


**Complete Python File**

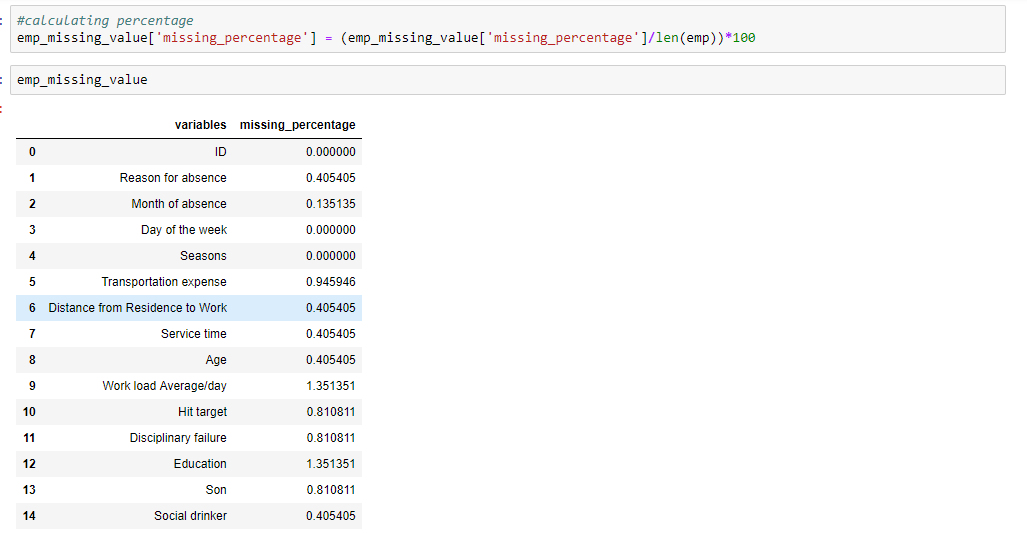




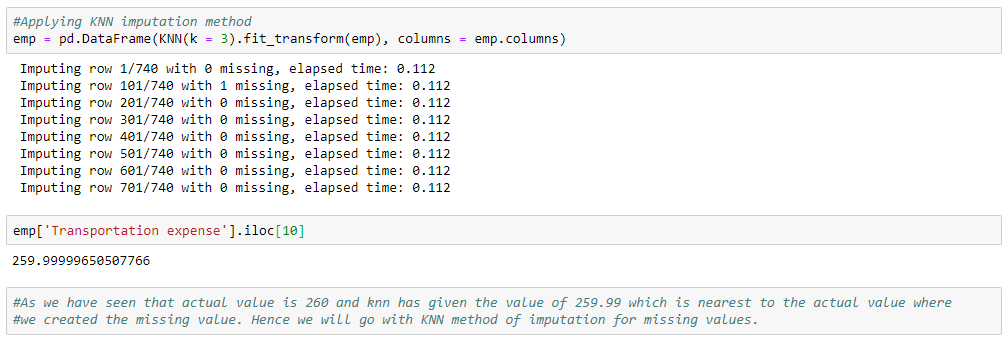




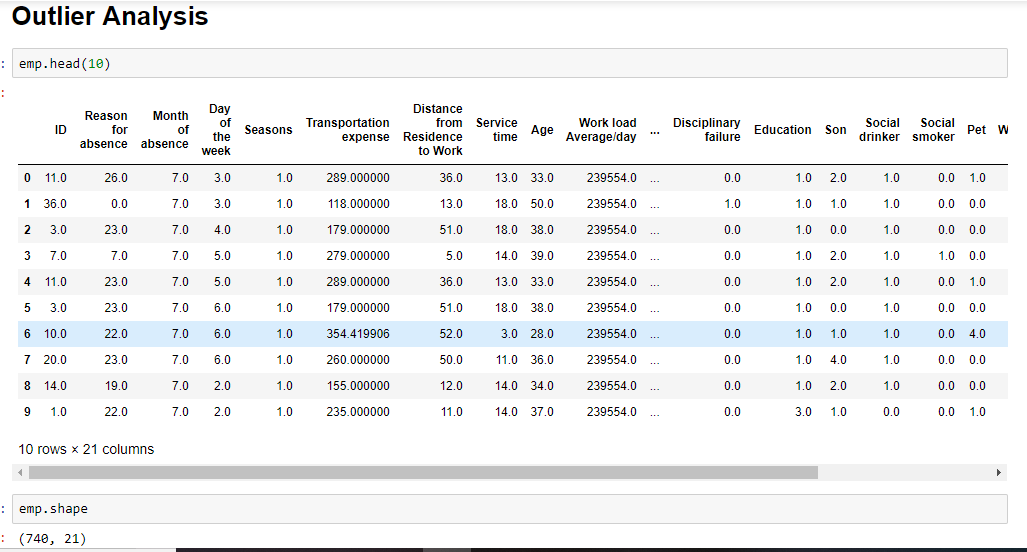


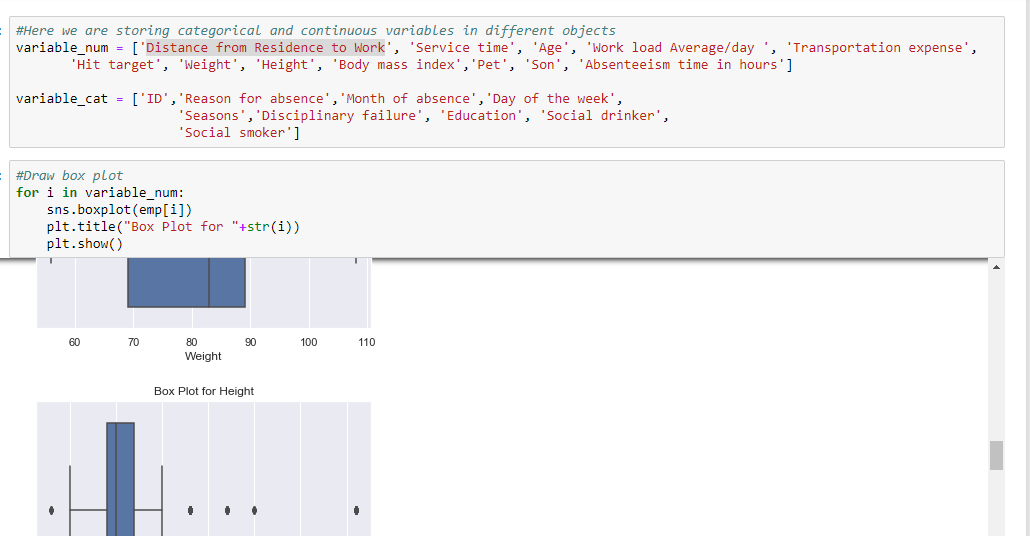


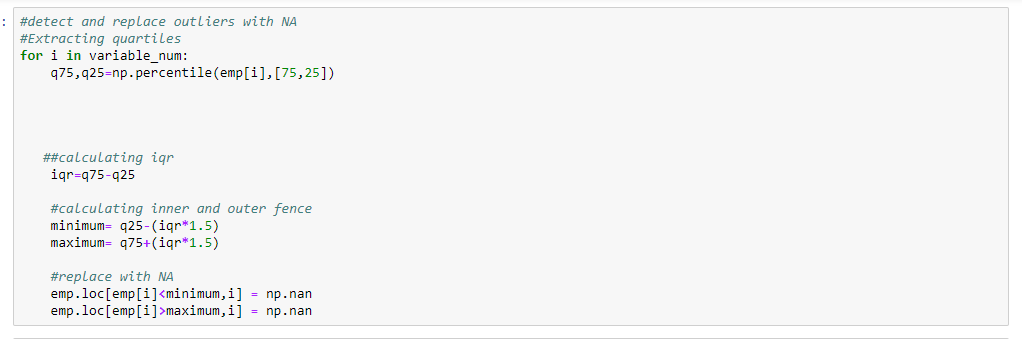


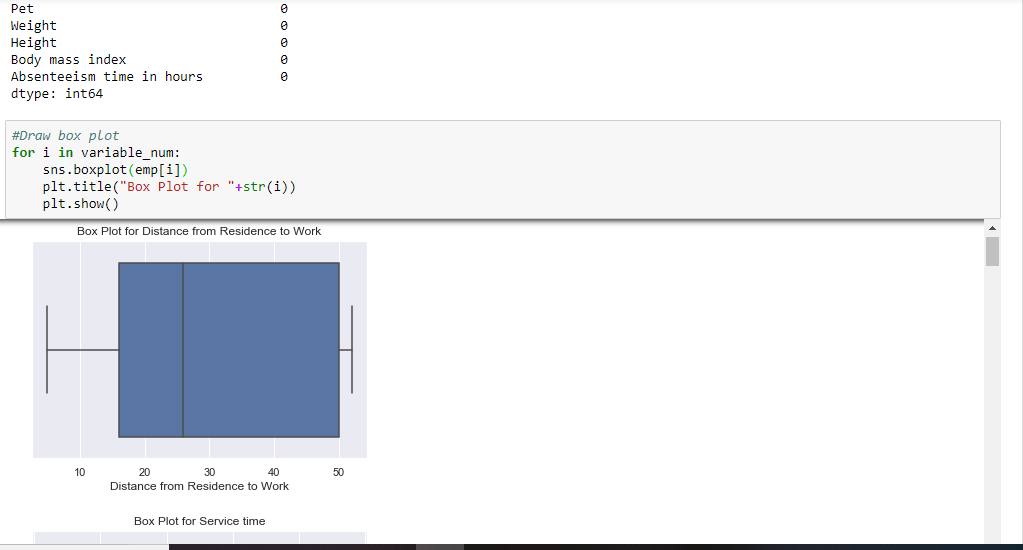
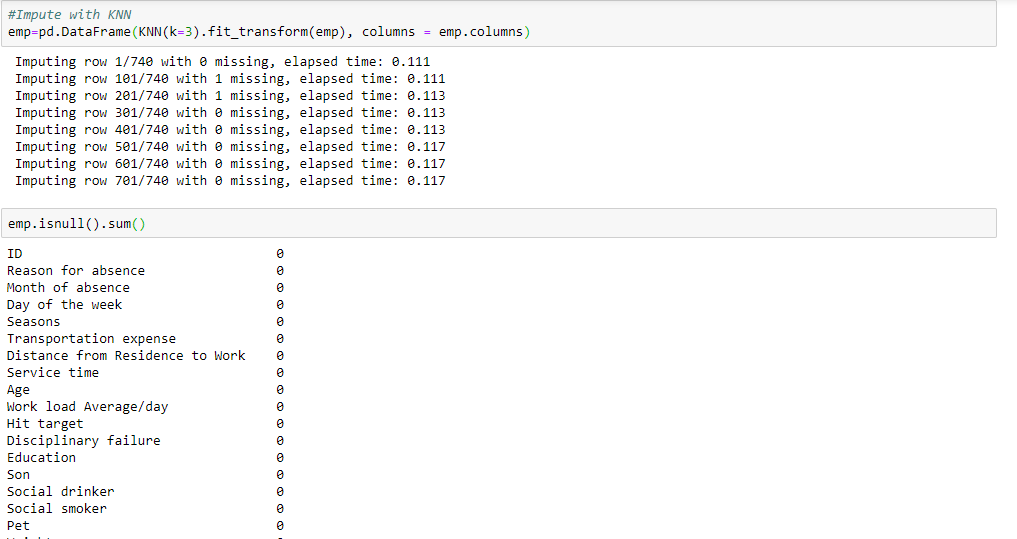
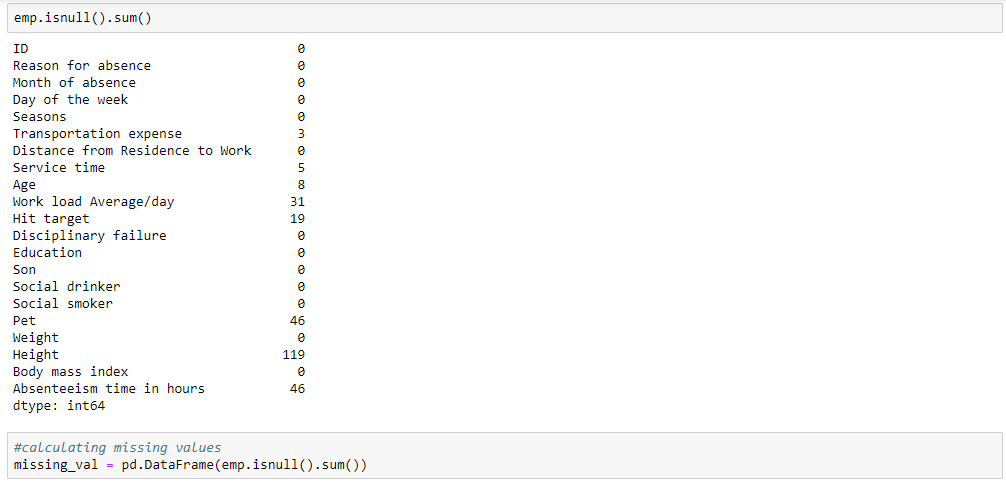


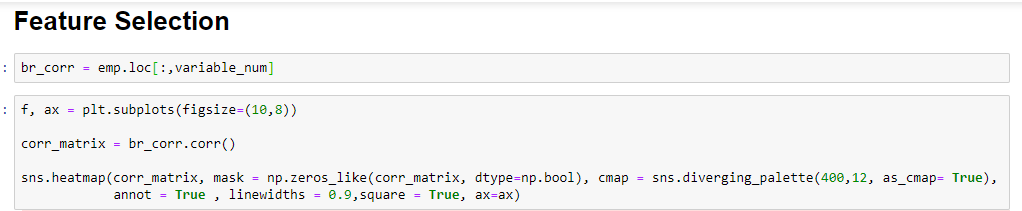


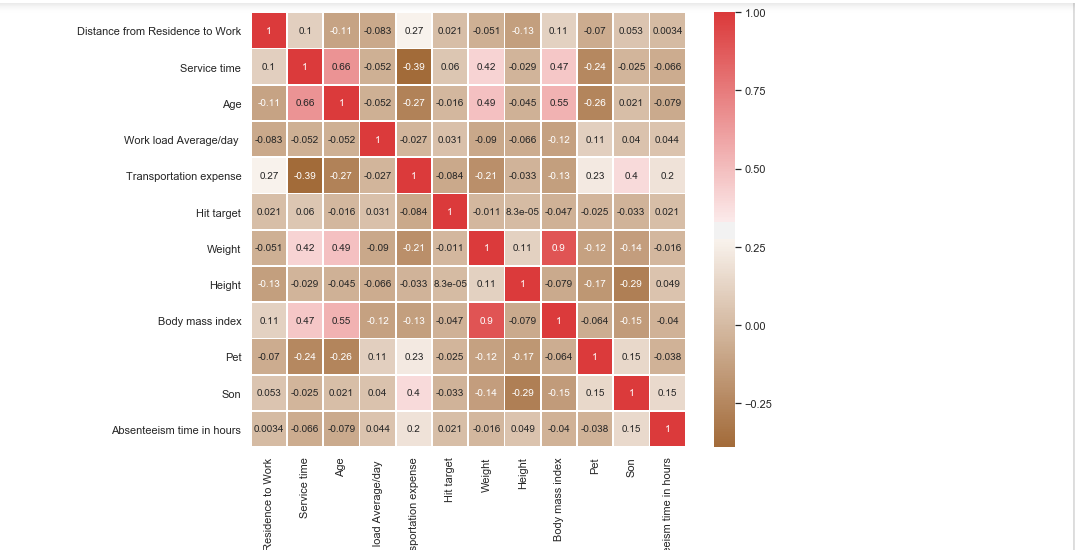


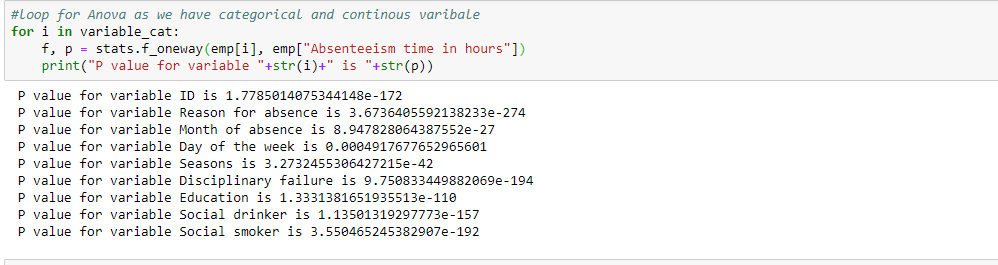


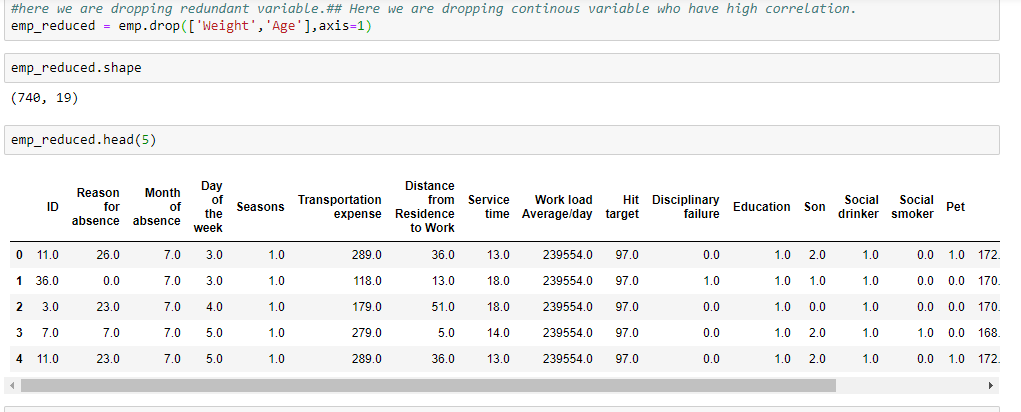


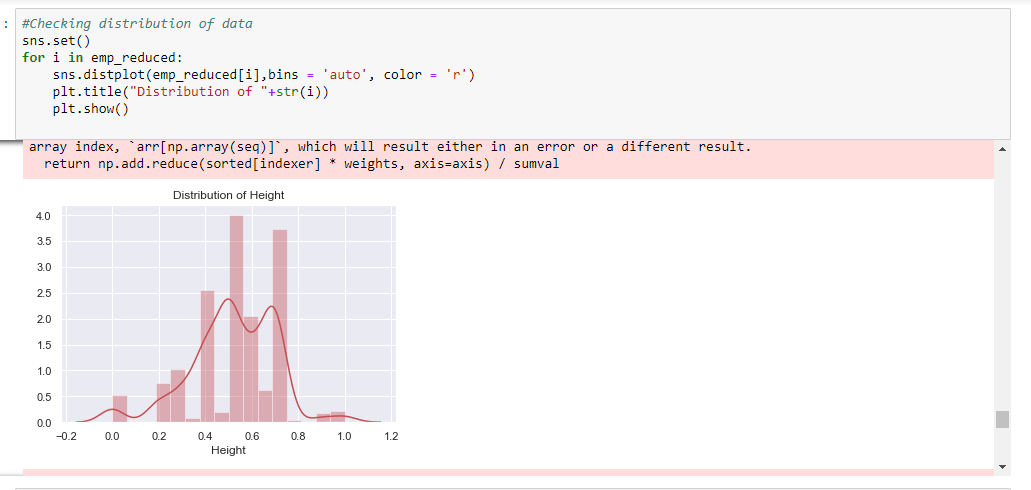


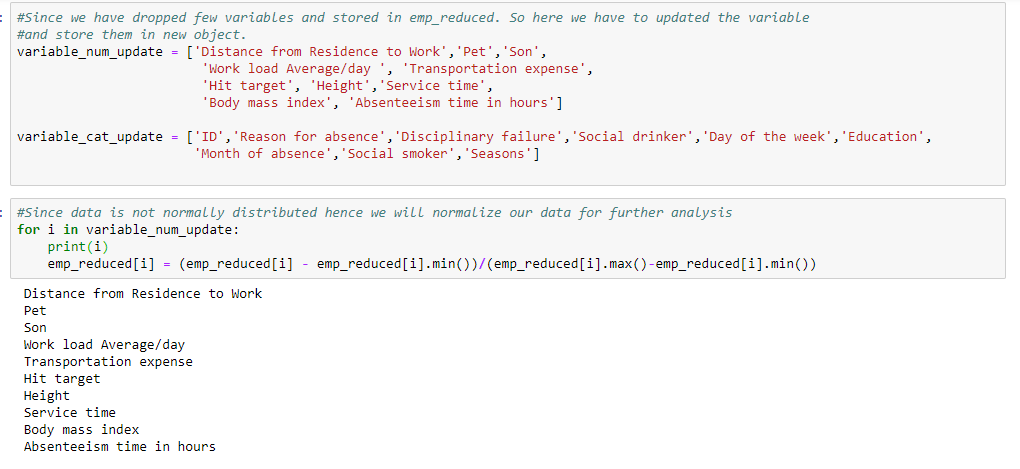


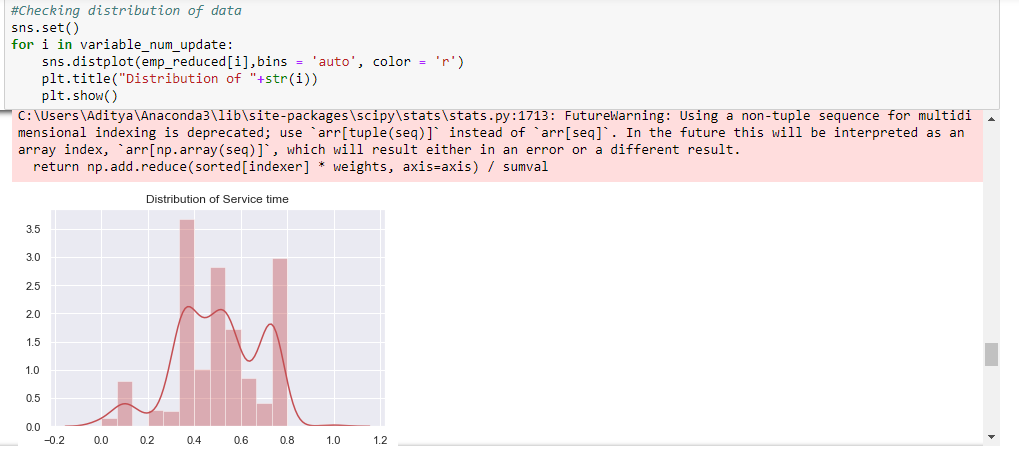


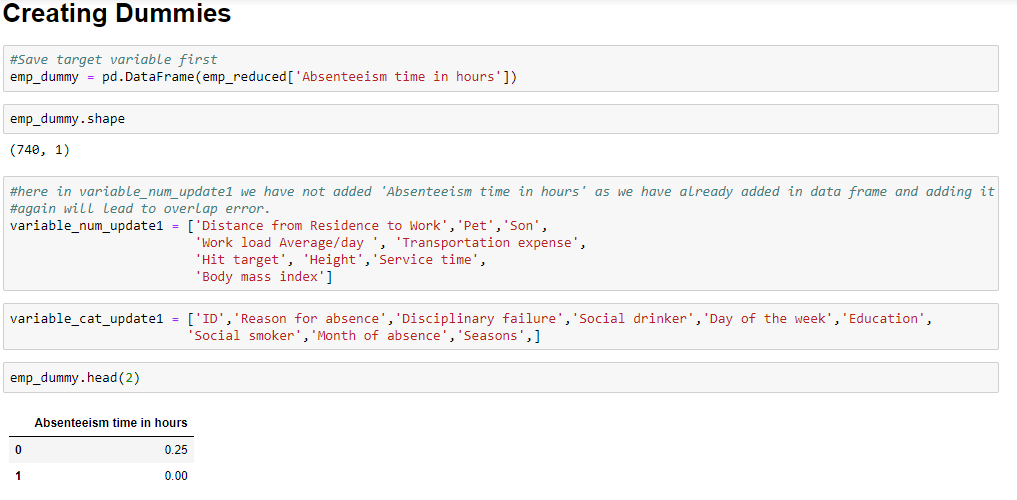


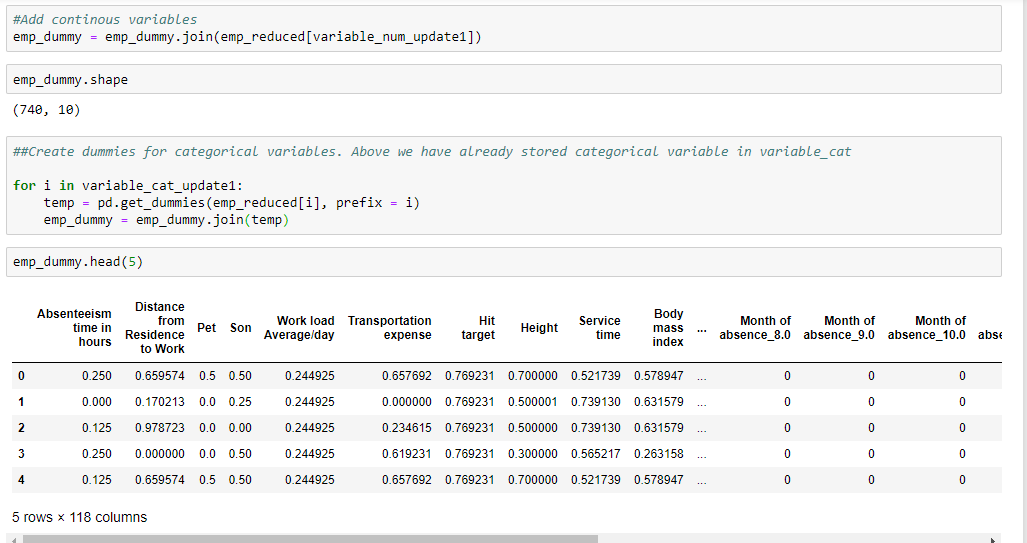


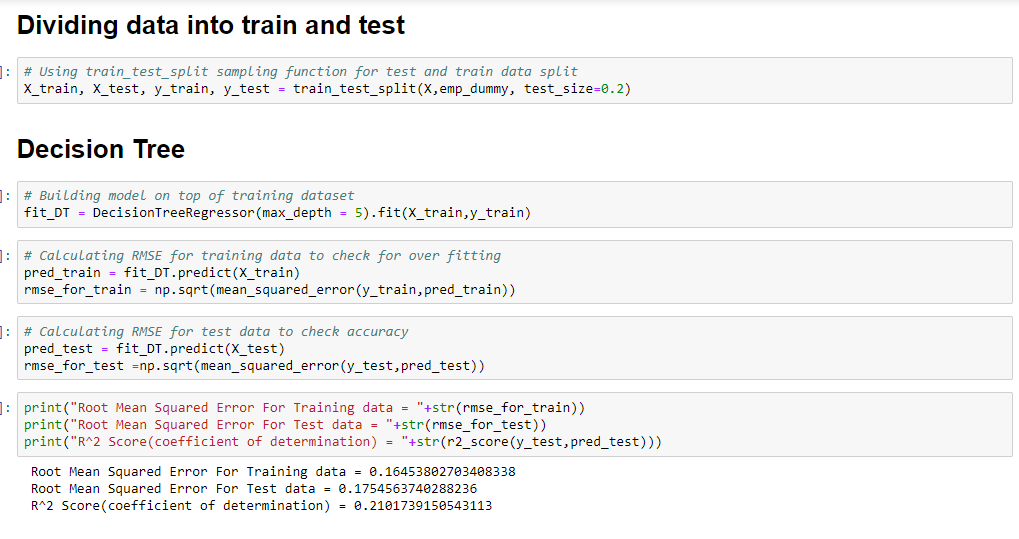


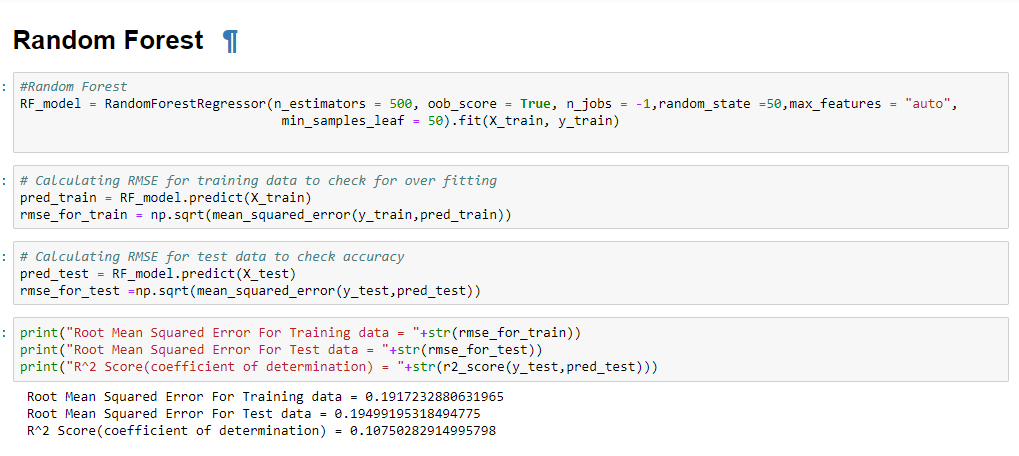




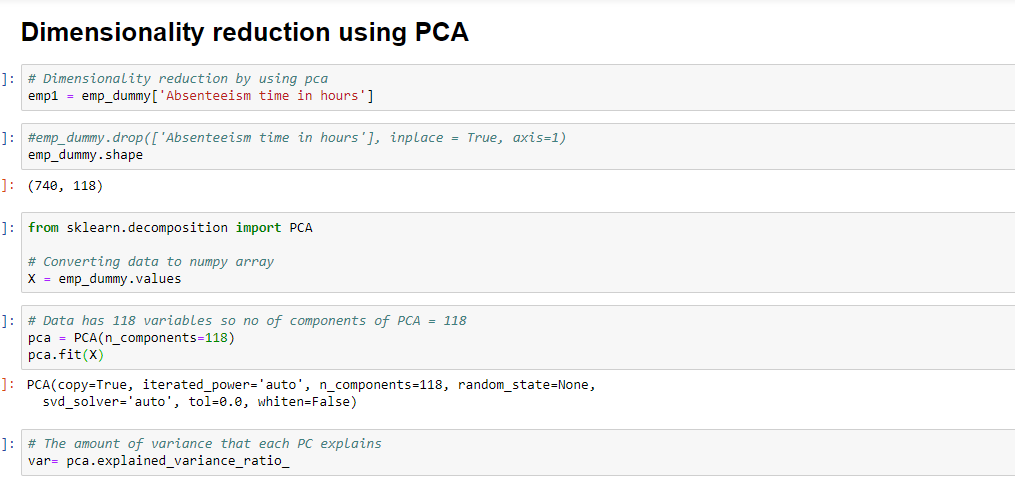




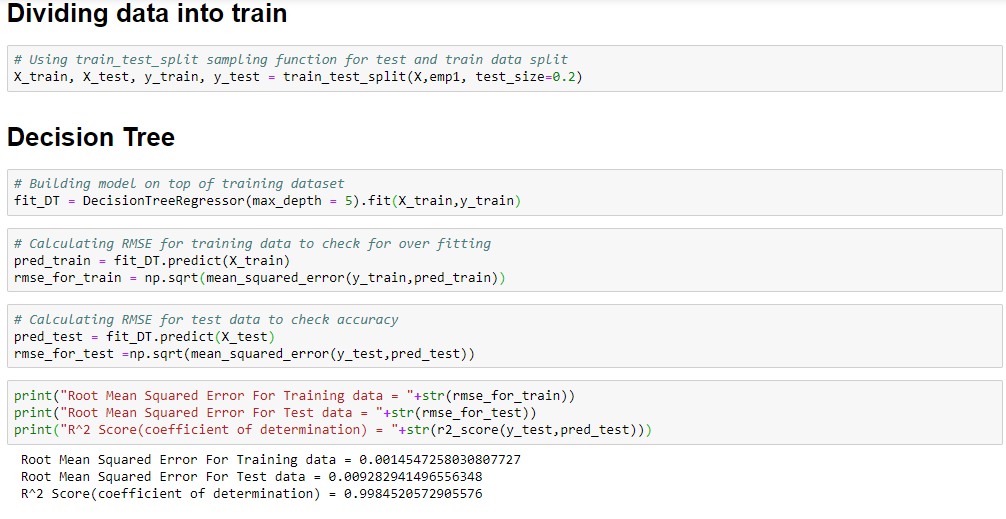


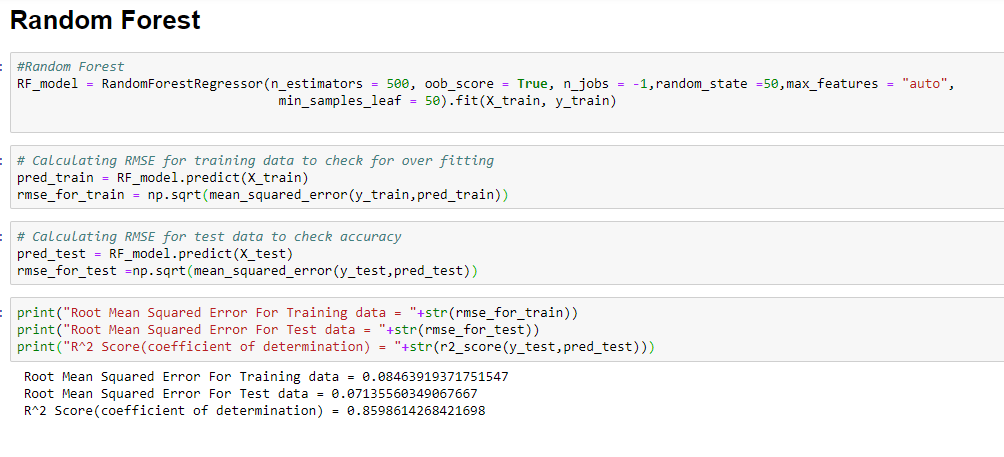


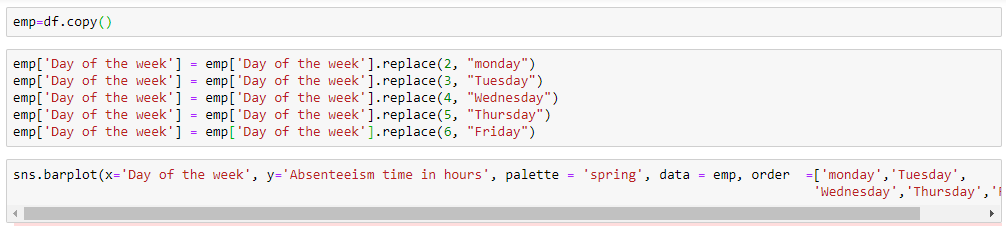
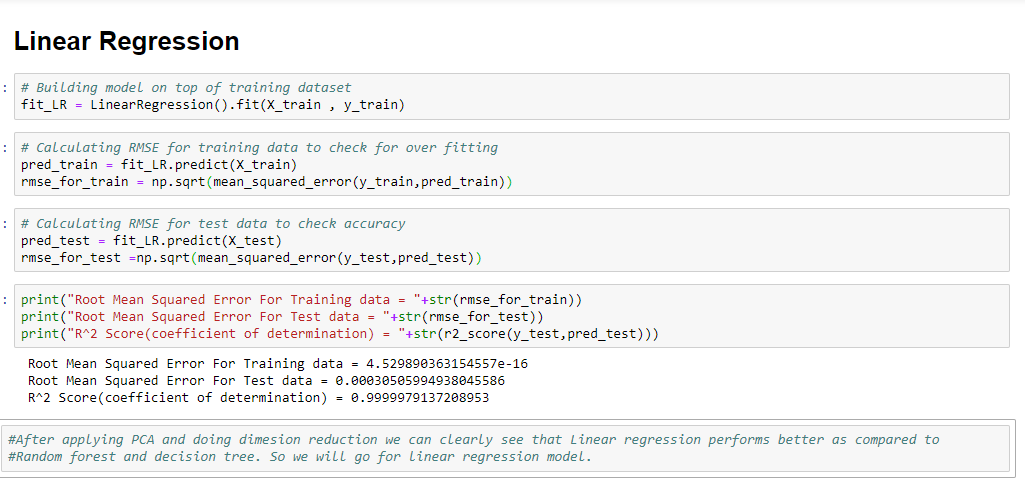


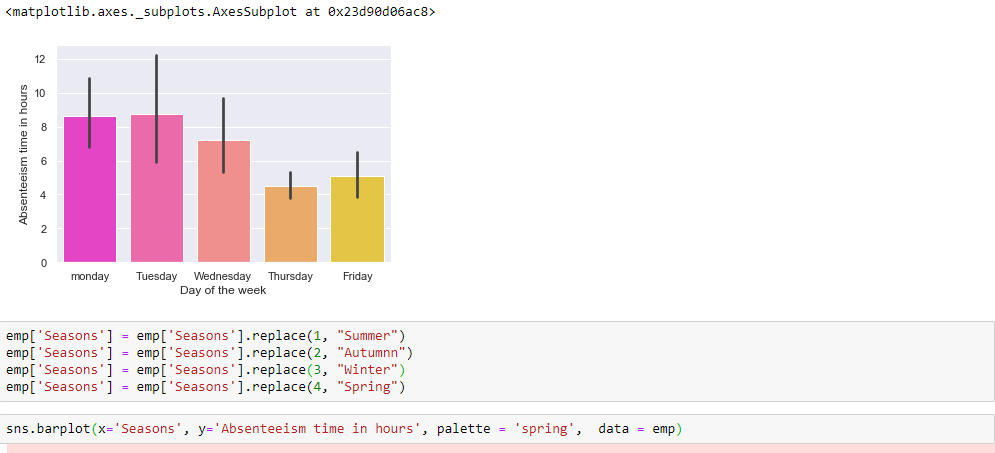


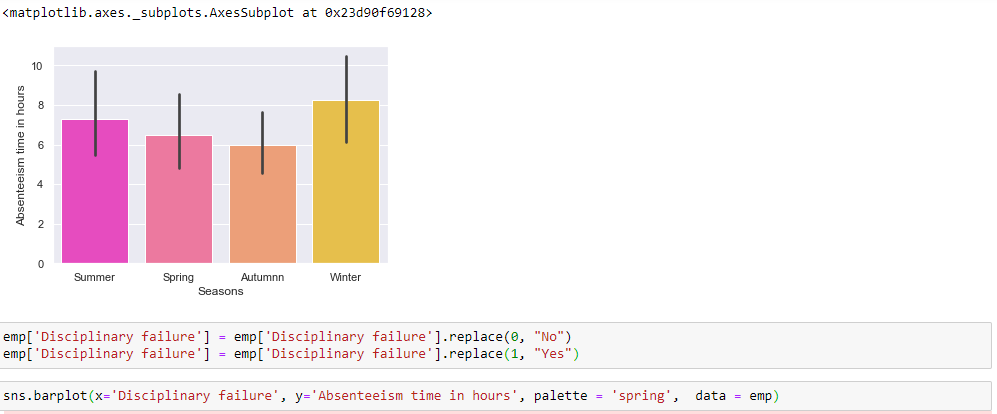




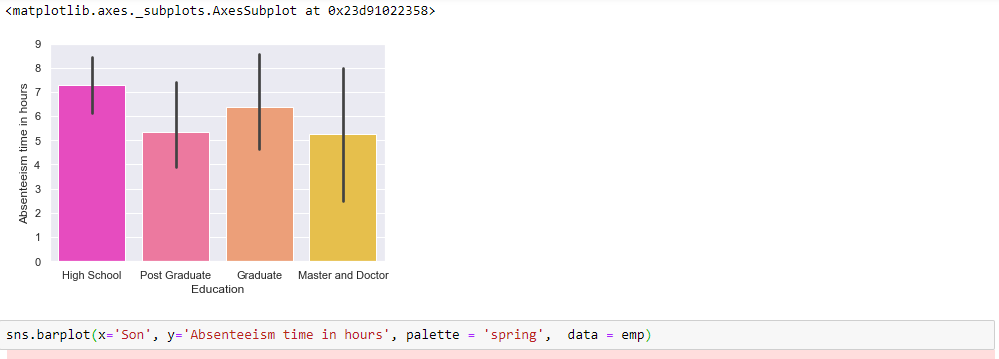


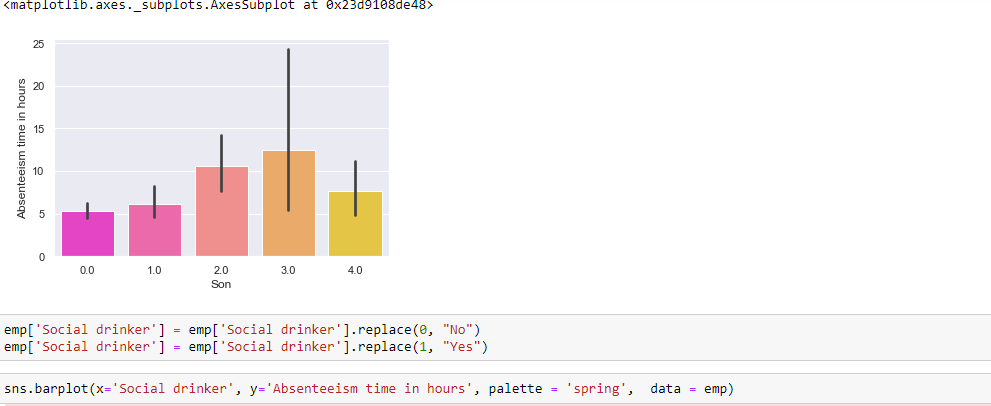


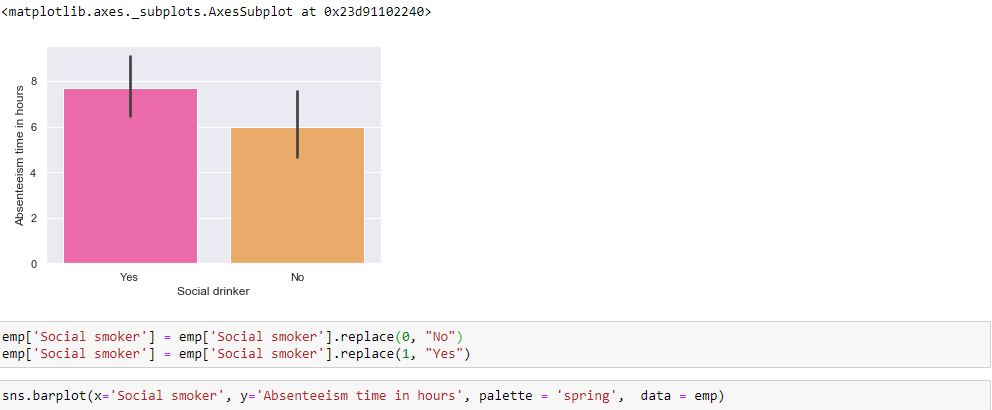


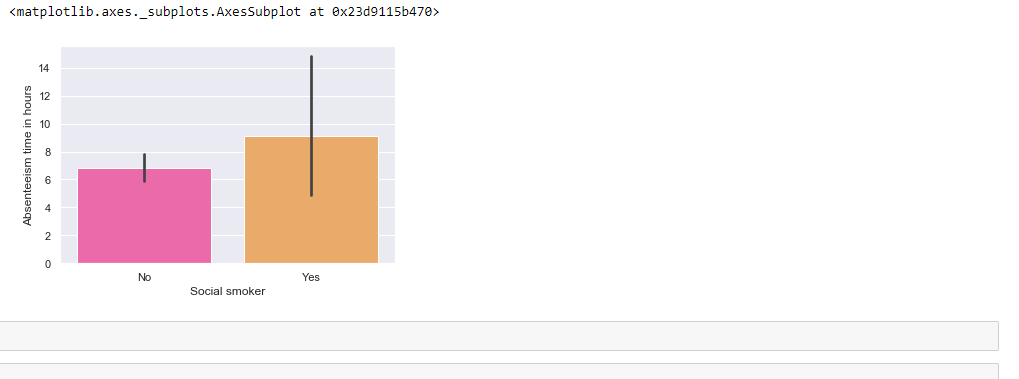












**References**

1. For Data pre-processing and Model Development -

<https://learning.edwisor.com/>

1. For PCA -

<https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/>

<https://www.youtube.com/watch?v=_UVHneBUBW0>

1. For Visualization –

<https://www.youtube.com/channel/UCQTQ0AbOupKNxKKY-_x46OQ>

<https://learning.edwisor.com/>