Project: EMPLOYEE ABSENTEEISM

Name: JOEL BRAGANZA

Email: [joelbraganza@yahoo.in](mailto:joelbraganza@yahoo.in)

(Project Report)

13th February 2019

INDEX:

Problem Statement

Data

Exploratory Data Analysis

Pre Processing

Missing Value Analysis

Outlier Analysis

Feature Selection

Feature Scaling (normalization)

Principal Component Analysis

Modelling

Decision Tree

Random Forest

Linear Regression

**Conclusion**

Model Evaluation

Suitable Model Selection

Solutions/Answers of asked questions

**Project description**

**Project’s 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?

**Data**

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 behavioral 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)

**Exploratory Data - Analysis**

Exploratory Data-Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics. In the given data set there are 21 variables and data types of all variables are either float64 or int64. There are 740 observations and 21 columns in our data set. There are null (missing) values present also

**Columns/variables and their number of unique values** -

ID 36 (categorical)

Reason for absence 28 (categorical)

Month of absence 13 (categorical)

Day of the week 5 (categorical)

Seasons 4 (categorical)

Transportation expense 24 (continuous)

Distance from Residence to Work 25 (continuous)

Service time 18 (continuous)

Age 22 (continuous)

Work load Average/day 38 (continuous)

Hit target 13 (continuous)

Disciplinary failure 2 (categorical)

Education 4 (categorical)

Son 5 (categorical)

Social drinker 2 (categorical)

Social smoker 2 (categorical)

Pet 6 (categorical)

Weight 26 (continuous)

Height 14 (continuous)

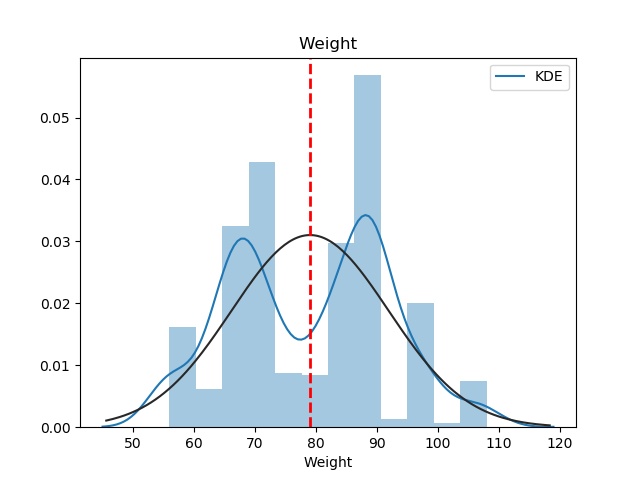
Body mass index 17 (continuous)

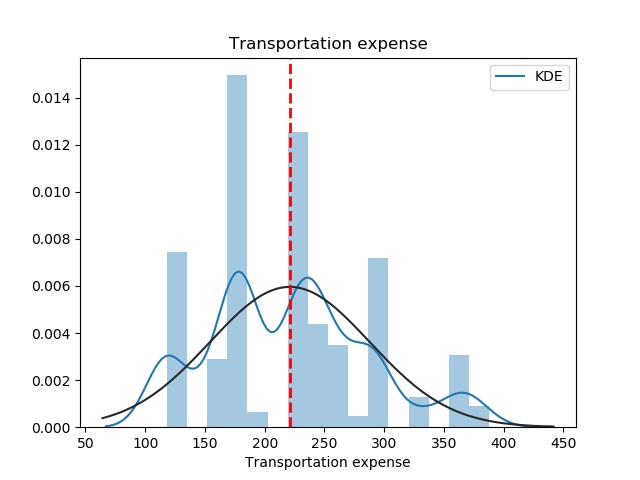
Absenteeism time in hours 19 (continuous)

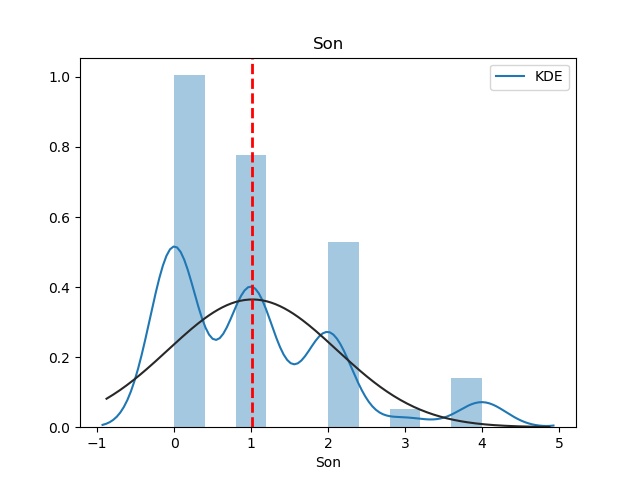
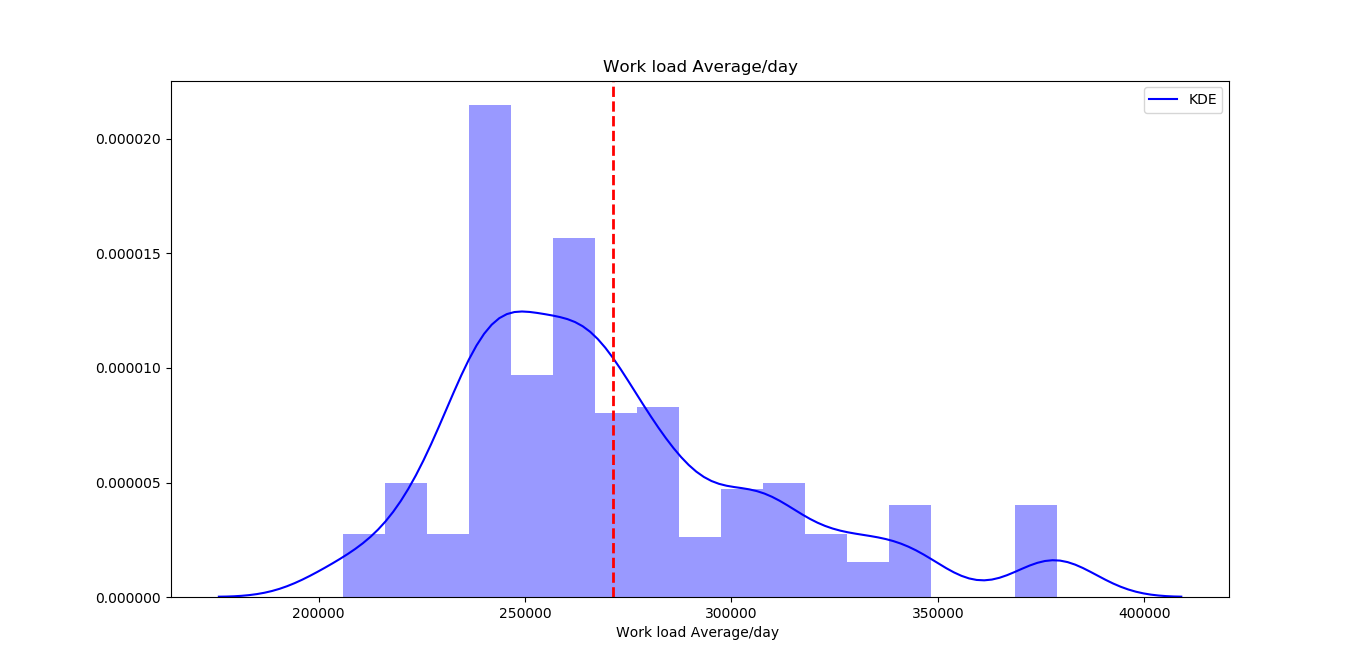
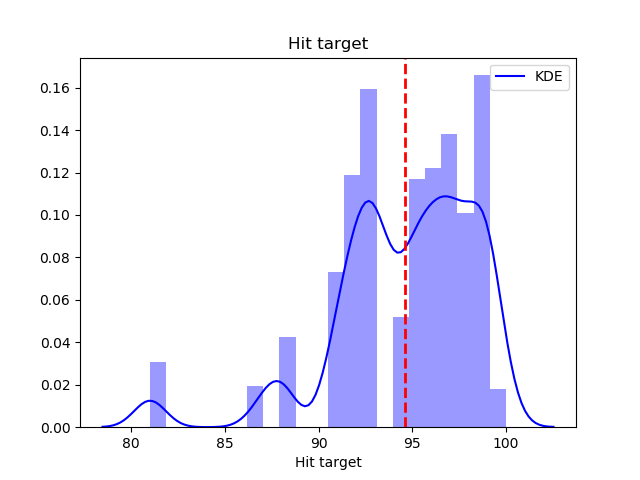
**After exploring the data we know there are 10 continuous-variables and 11 categorical variables in nature.**

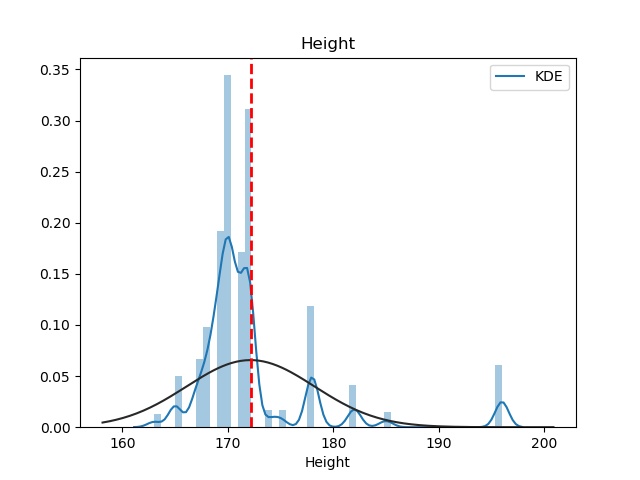
**Pre Processing**

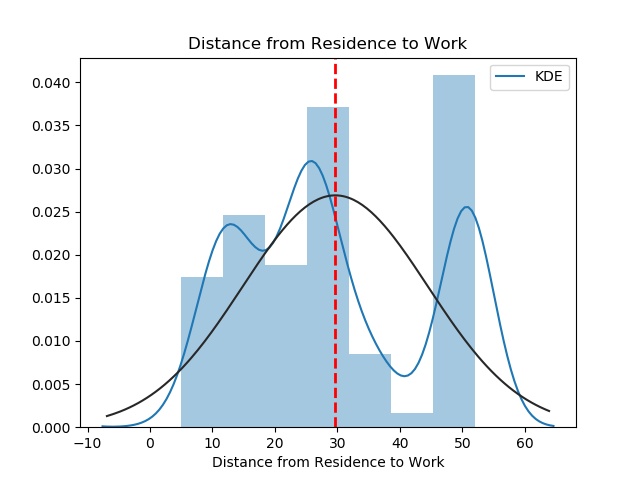
Preprocessing is to look at the dataset and observe the various characteristics(maybe even with intuition) to explore, visualize, clean the data, before we go about using any models to make predictions or draw conclusions from the data.What we have done is, we have made graphs to look at the probability distributions of the variables. As the target variable is continuous, we would like to see if the data is normally distributed or not. We can see that looking at the probability distributions or probability density functions of the continuous- variables and also their KDE(Kernel-Density-Estimations).

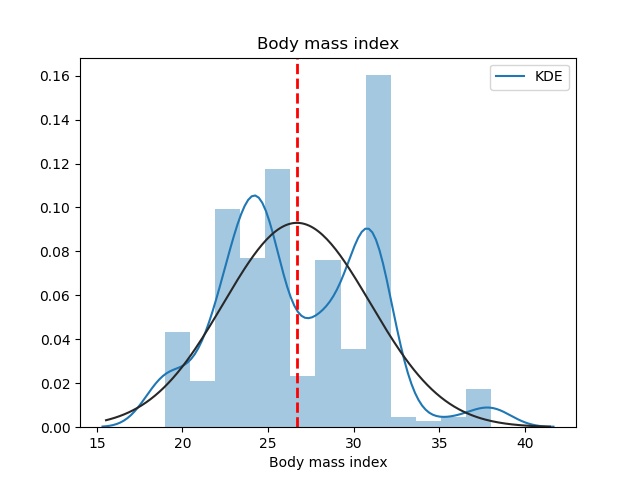


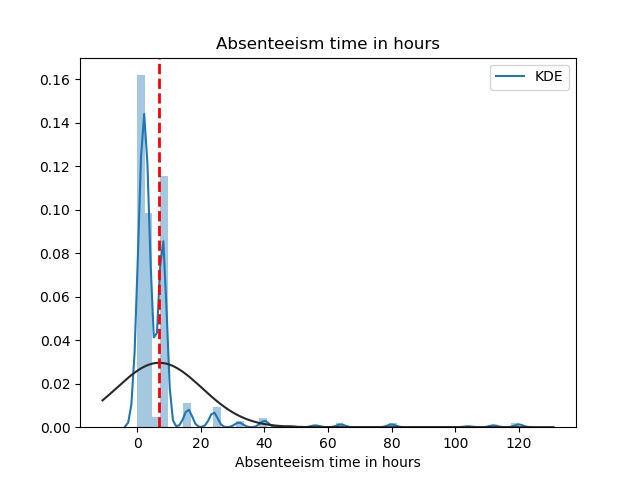










Graph analysis:

As we can see in the graph,

Blue-line: represents the KDE (Kernel-density estimations)

Black-line: represents the probability distributions of the variables.

Red-line: represents the mean-value of the variables.

The unequal distribution of data around the mean-line shows skewness in the dataset.

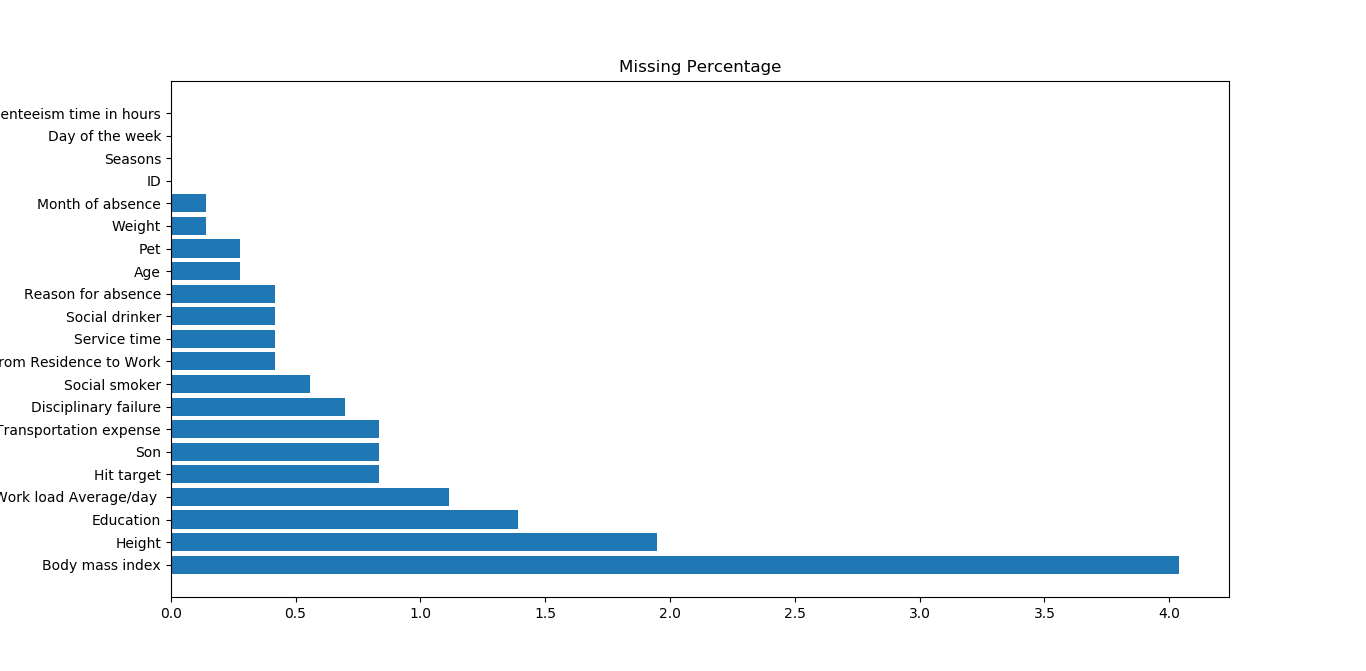
Some variables seem to be fine, others are pretty much left or right skewed. It is these variables on which we have to work more on like missing-values imputation, outlier-analyis and so on.

**Missing Value Analysis**

Below is the graph of our missing-values of each column/variable. From looking at that we see that the maximum percentage of missing-data is 4.1%, so we don’t have to remove any of the variables from the dataset (on >=30%, we must remove).

Both in R and Python, one question in my mind about imputation was that should I remove the outliers first and then impute the total missing-null-values OR should I do the imputation first and then do outlier-removal and again go impute the removed-variables.

After asking the query in sessions and reading on some forums, I realized it depends on the end-goal(purpose) of analyzing the model. So finally, not to base my analysis solely on the present values and not considering the possible-missing-values for crating upper-lower fence of the normal-range data, I first imputed the data( custom imputation for each variable mean/median/knn best suitable) and then did the outlier analysis and removal. Finally ending with the knn-imputation for the removed-outliers.



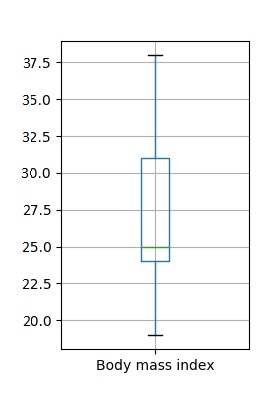
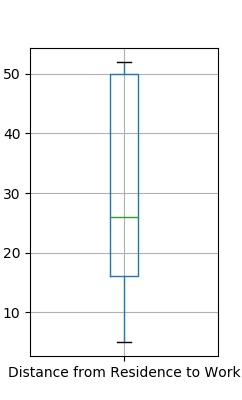
After the missing-value analysis, we have removed the observations/rows from the dataset with null-values in the target-variable ‘Absenteeism time in hours’.

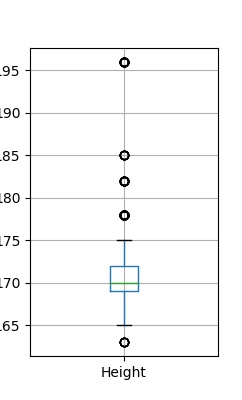
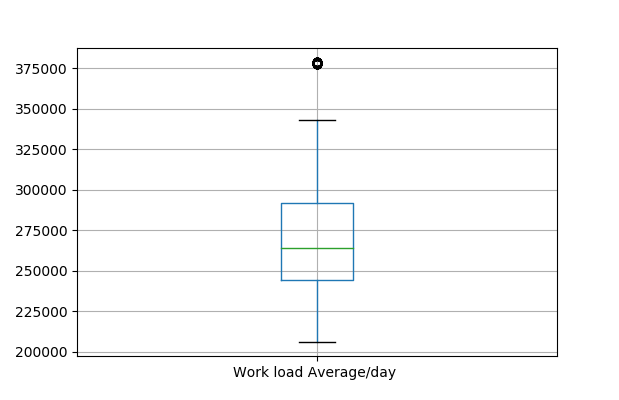
As there was no point in looking for outliers in it, neither in imputing the values for this column.

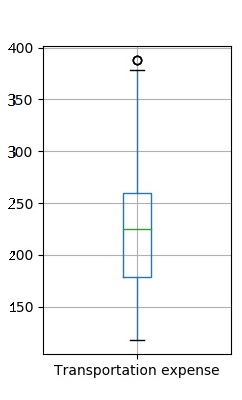
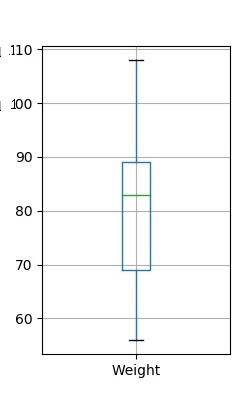
**Outlier Analysis**

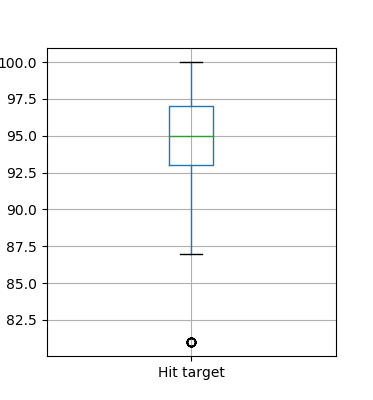
After seeing the probability distributions graph that many of the variables are skewed. The skewness maybe because of presence of outliers and/or extreme values in the data .As taught in the sessions using the inter-quartile range (boxplot method) we are removing outliers. We visualize the outliers using boxplots below.

Below we have plotted the boxplots of the 11 continuous variables. A lot of useful inferences can be made from these plots. You can see which all variables have outliers and extreme values in each of the data set and how much of these.



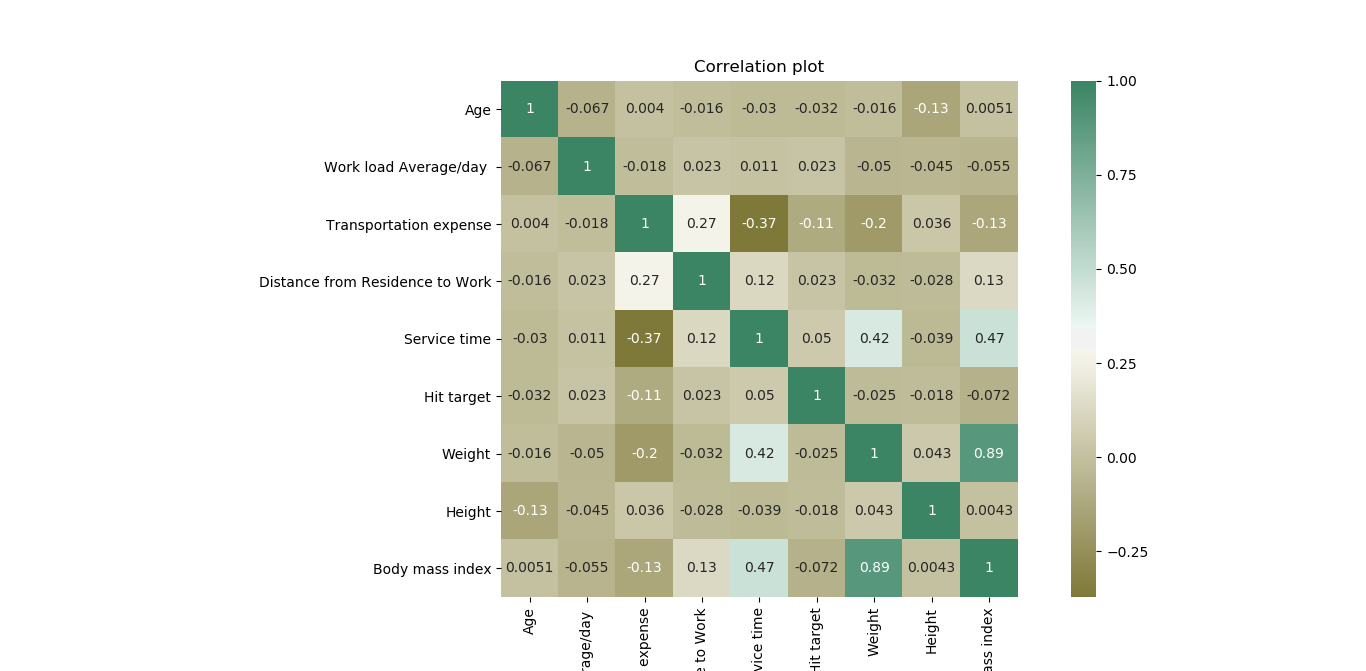
After looking at these boxplots, we know that all the variables **except “Distance from residence to work”, “Weight” and “Body mass index”** consists of outliers.

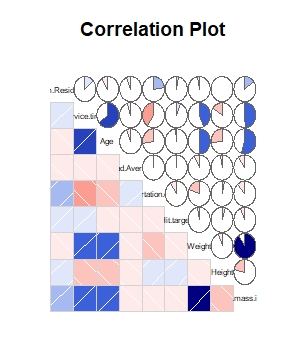
After removing the outliers, we have imputed them using the knn method.

FrEv

**Feature Selection**

Feature selection is the selection of only the necessary (or subset) of variables/columns for then feeding to the ML models for predictions. Before even doing our predictive modeling, we have to calculate the value/importance of each predictor-variable, see if there is any correlation to avoid multicollinearity to use only the important (meaningful) variables , otherwise we will provide giving similar information via multiple variables which will unnecessarily use more time and resources (overhead). Here we have used **Correlation Analysis** for continuous variables and **ANOVA** (Analysis of variance) for categorical variable.



Even

We can see in both R and Python correlation plots, that ‘Body mass index’ and ‘Weight’ have a correlation of 0.89 (almost 0.9) which is > 0.7, so we remove ‘Weight’ variable from the dataset.

ANOVA test:

The ANOVA tests speak very differently in both PYTHON and R even after the same kind of pre-processing (the imputation methods and outlier analysis are done in the same order)

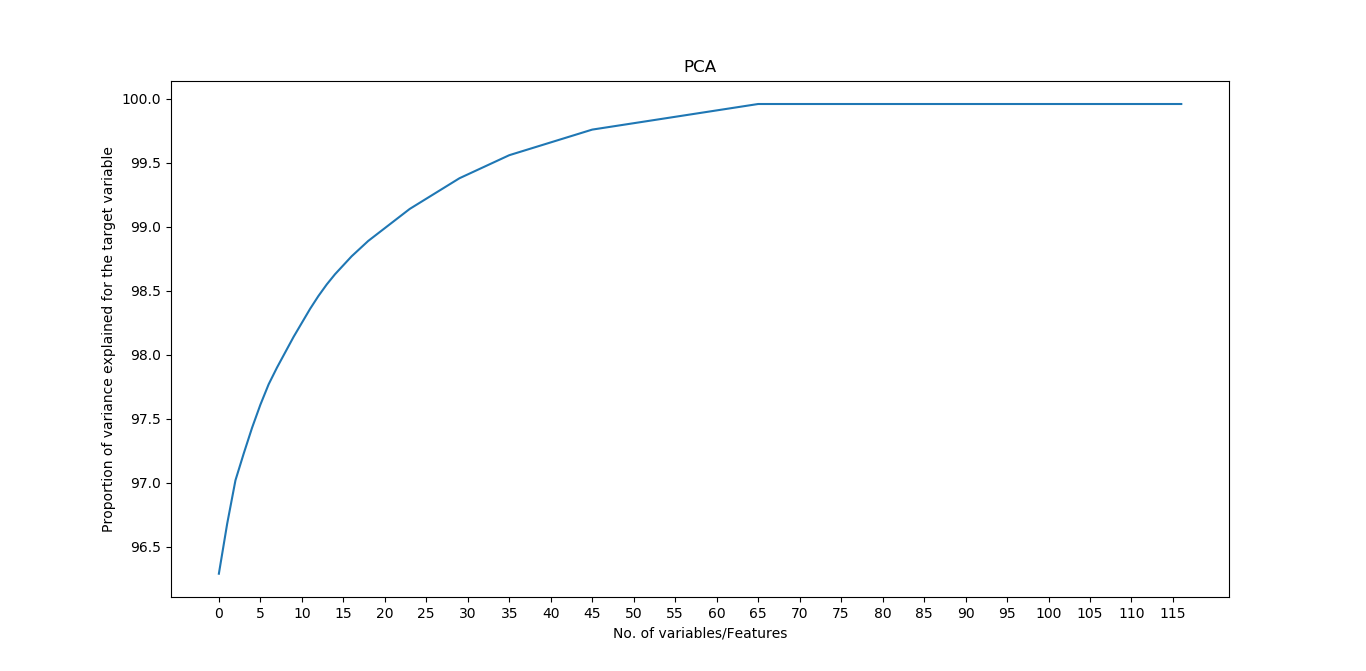
However, for now we will keep these Categorical variables in our dataset and view their influence(contribution) to the target variable using Principal-Component Analysis and later on using histograms for these variables.

NORMALIZATION:

We now normalize the data, to bring it in a range of 0 to 1, so that there is no influence of one variable over the other because of wide difference in their quantity range. We do this for all variables, except the target variable’ Absenteeism time in hours’. After this we start developing models. Both, with and without PCA to see the differences.

**Principal Component Analysis**

Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data. After creating dummy variable of categorical variables the shape of our data became 115 columns and 714 observations, this high number of columns leads to bad accuracy.



We have applied PCA algorithm on our data and from the above graph we have concluded that 45 variables out of 115 explains more than 95% of data. So we have selected only those 45 variables to feed our models. We have received the same conclusion of keeping 45 variables in both R and Python.

**Regression Modeling**

We will split the data into training (80%) and test (20%) with random stratified techniques.

We will now be using the regression models taught in our sessions on our processed data to predict the testing-data of target variable. Following are the models which we have built –

We have applied the models both without feature-reduction ( dummy-categorizing variables and PCA) and with too.

**Decision Tree**

The decision tree algorithm accept continuous and categorical variables as independent variables .A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utilityThe branches connect nodes with “and” and multiple branches are connected by “or”. It can be used for classification and regression. It is a supervised machine learning algorithm. Extremely easy to understand by the business users. Split of decision tree is seen in the below tree. Here we have applied the decision-tree both before and after the dummy-categorizing and PCA. The RMSE value and R^2 value for our project in R and Python are –

Pre- dummy categorizing and PCA

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **R** | **PYTHON** |
| **RMSE Train** | 10.7227598 | 0.529 |
| **RMSE Test** | 14.90178731 | 0.834 |
| **R^2 Test** | 0.06795472 | 0.986 |

Post- dummy categorizing and PCA

|  |  |  |
| --- | --- | --- |
| **Decision Tree** | **R** | **PYTHON** |
| **RMSE Train** | 3.61 | 0.526 |
| **RMSE Test** | 3.19 | 1.07 |
| **R^2 Test** | 0.94 | 0.992 |

**Random Forest**

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree. It means to build each decision tree on random forest we are not going to use the same data. The RMSE value and R^2 value for our project in R and Python are –

Pre- dummy categorizing and PCA

|  |  |  |
| --- | --- | --- |
| **Random Forest** | **R** | **PYTHON** |
| **RMSE Train** | 7.15 | 0.54 |
| **RMSE Test** | 13.9 | 0.82 |
| **R^2 Test** | 0.17 | 0.987 |
|  |  |  |

Post - dummy categorizing and PCA

|  |  |  |
| --- | --- | --- |
| **Random Forest** | **R** | **PYTHON** |
| **RMSE Train** | 2.45 | 0.033 |
| **RMSE Test** | 2.50 | 0.029 |
| **R^2 Test** | 0.927 | 0.99 |

**2.2.3 Linear Regression**

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. To build any model we have some assumptions to put on data and model. Here are the assumptions to the linear regression model.

Pre- dummy categorizing and PCA

|  |  |  |
| --- | --- | --- |
| **Linear Regression** | **R** | **PYTHON** |
| **RMSE Train** | 12.30 | 9.53e-15 |
| **RMSE Test** | 15.09 | 7.19e-15 |
| **R^2 Test** | 0.04 | 1 |

Post - dummy categorizing and PCA

|  |  |  |
| --- | --- | --- |
| **Linear Regression** | **R** | **PYTHON** |
| **RMSE Train** | 3.61 | 1.01e-14 |
| **RMSE Test** | 3.19 | 8.045e-15 |
| **R^2 Test** | 0.94 | 1 |

**Model Evaluation**

In the previous chapter we have seen the **Root Mean Square Error** (RMSE) and **R-Squared** Value of different models. **Root Mean Square Error** (RMSE) is the standard deviation of the residuals (prediction **errors**). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Whereas **R**-**squared** is a relative measure of fit, **RMSE** is an absolute measure of fit. As the square root of a variance, **RMSE** can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of **RMSE** and higher value of **R-Squared Value** indicate better fit.

**Model Selection**

From the observation of all **RMSE Value** and **R-Squared** Value in both pre and post-dummy categorizing and PCA, we have concluded that **Linear Regression Model**  has minimum value of RMSE and it’s **R-Squared** Value is also maximum (i.e. 1).

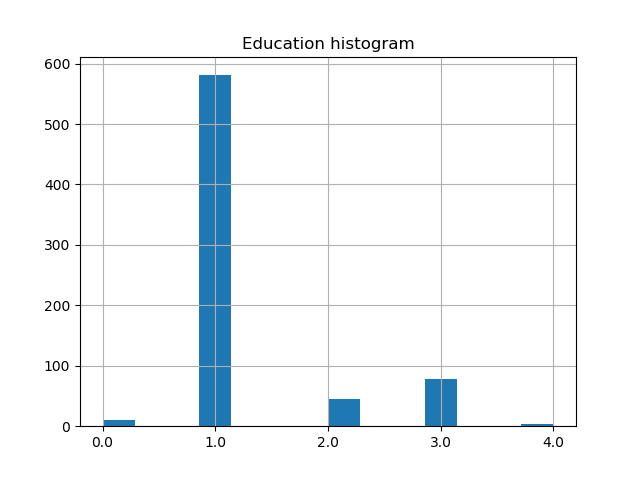
The RMSE value of Testing data and Training does not differs a lot this implies that it is not the case of overfitting (basing the conclusion with preference more to Python observations).

**Answers of asked questions**

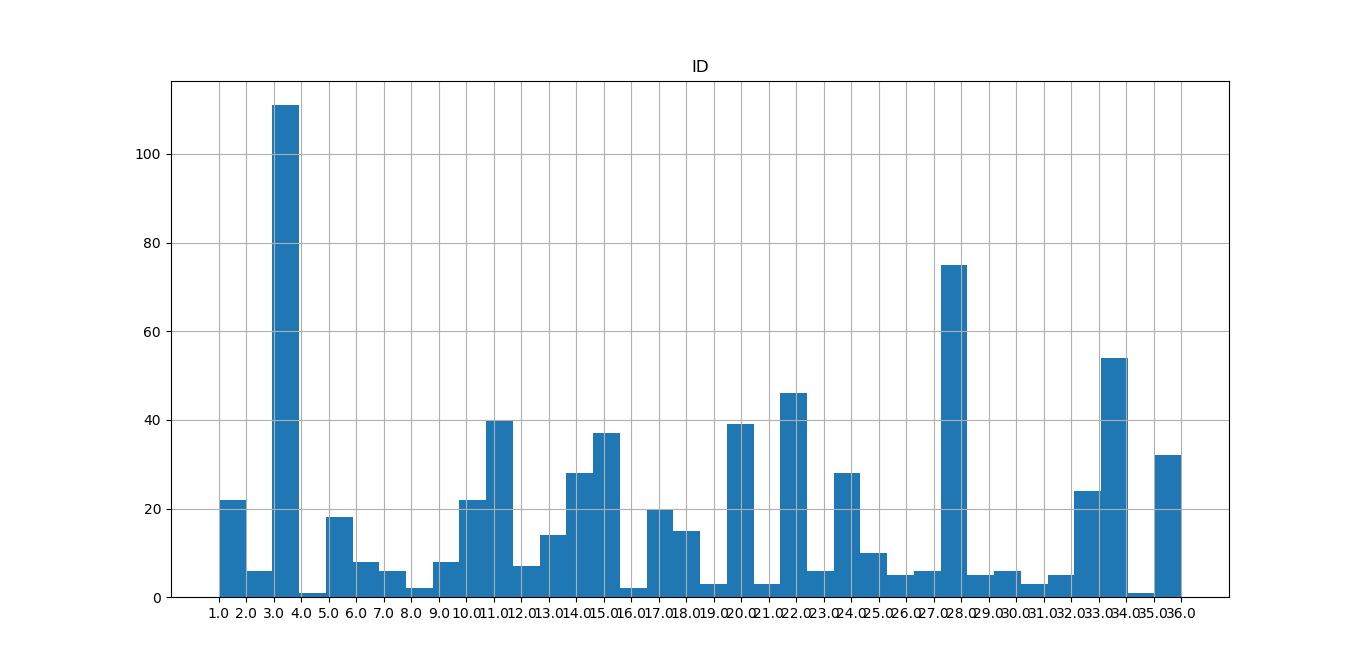
**The Changes which company should bring to reduce the number of absenteeism –**

1. It is observed that employee with low education have maximum absentee time.

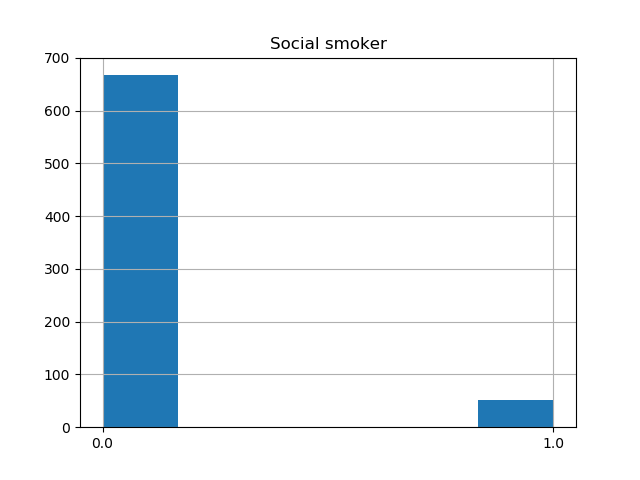
The employees with education –category 1 tend to be more absent.



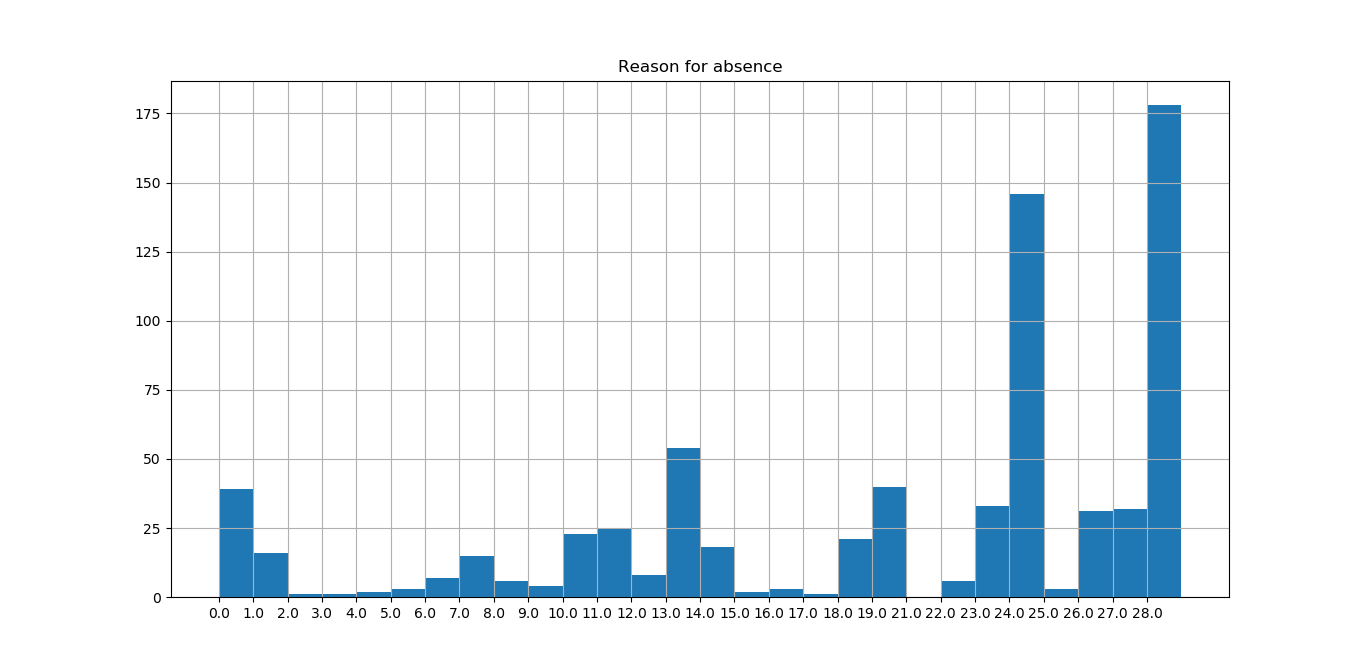
2 . Some employee with **ID 3, 28, 34** are often absent from work, company should speak with them, asking for genuine reason for their past absence and tell them to take actions as not to be more absent like in the past .



3.Employees who are social smoker have more absentee hour than who are not social smoker.



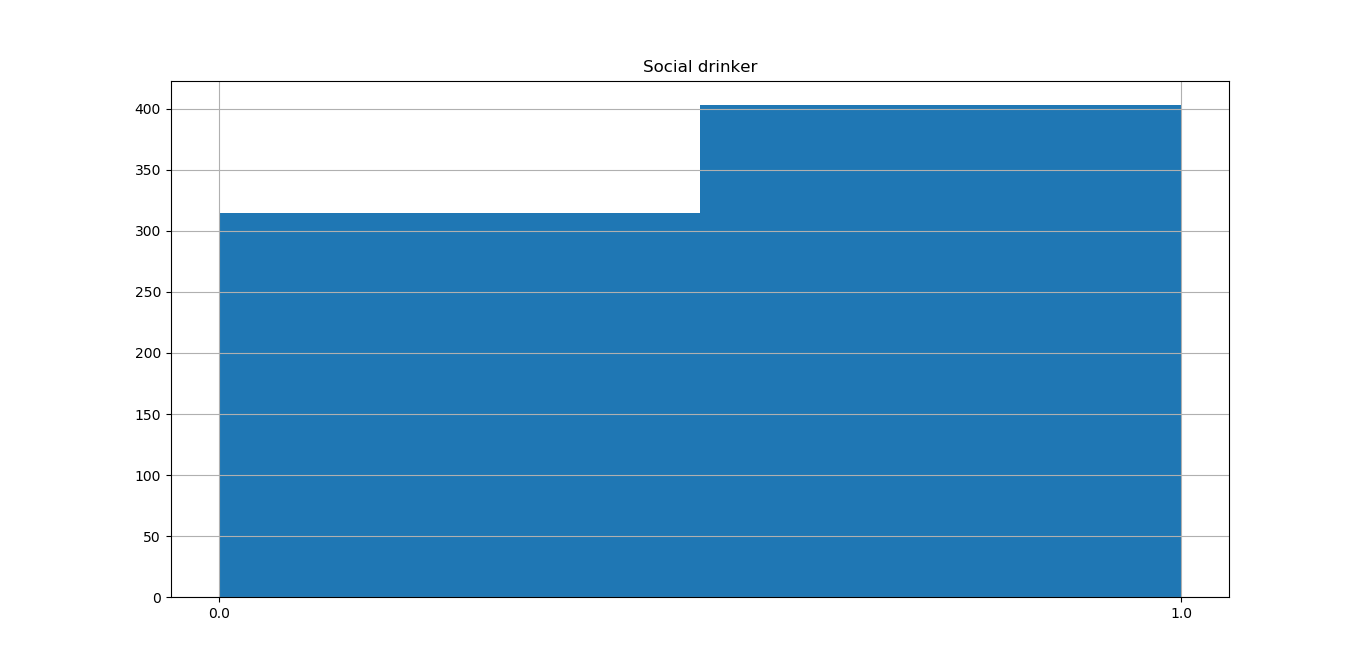
4.Most often Reason for absence are medical consultation and dental consultation, company should take care of it.

****

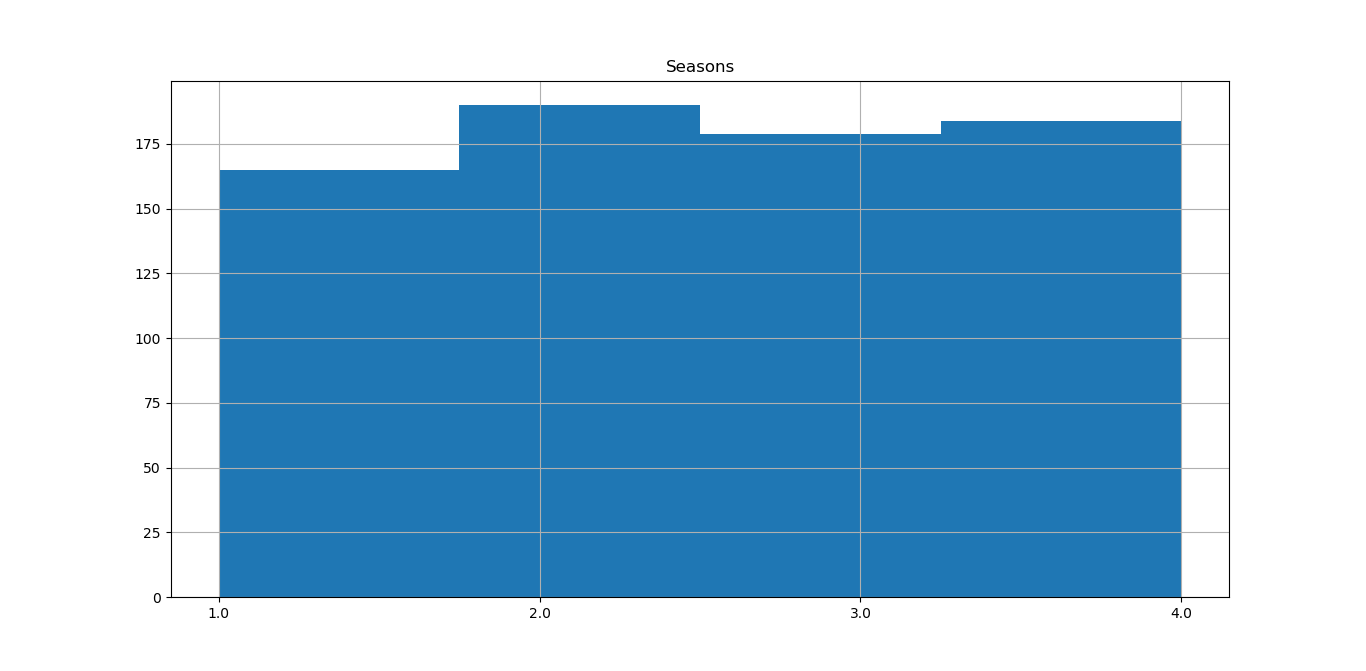
The rest of the variables have a distribution, which is pretty much equal for all the unique values, so not much of changes could be required to put based on those variables.

Below are the graphs of categorical-variables which showed some relationship with the target-‘Absenteeism time in hours’ in the ANOVA test in both R and Python.

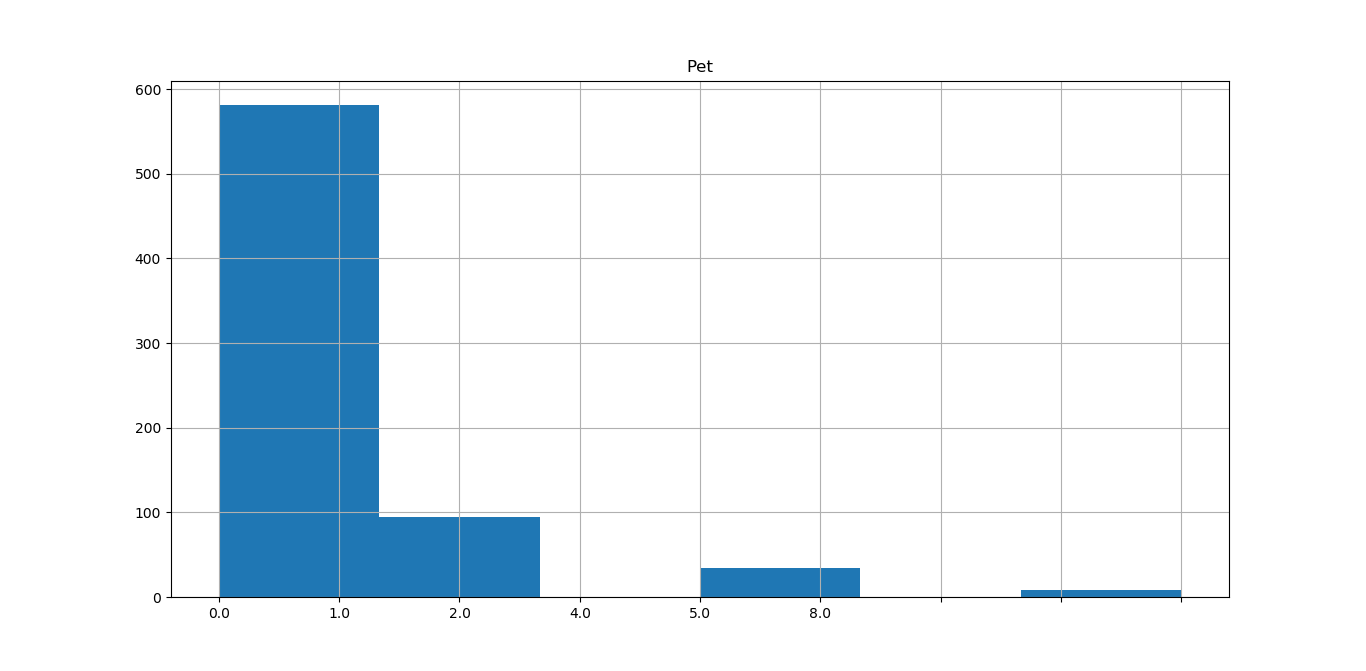
Here are the histograms showing their distributions.



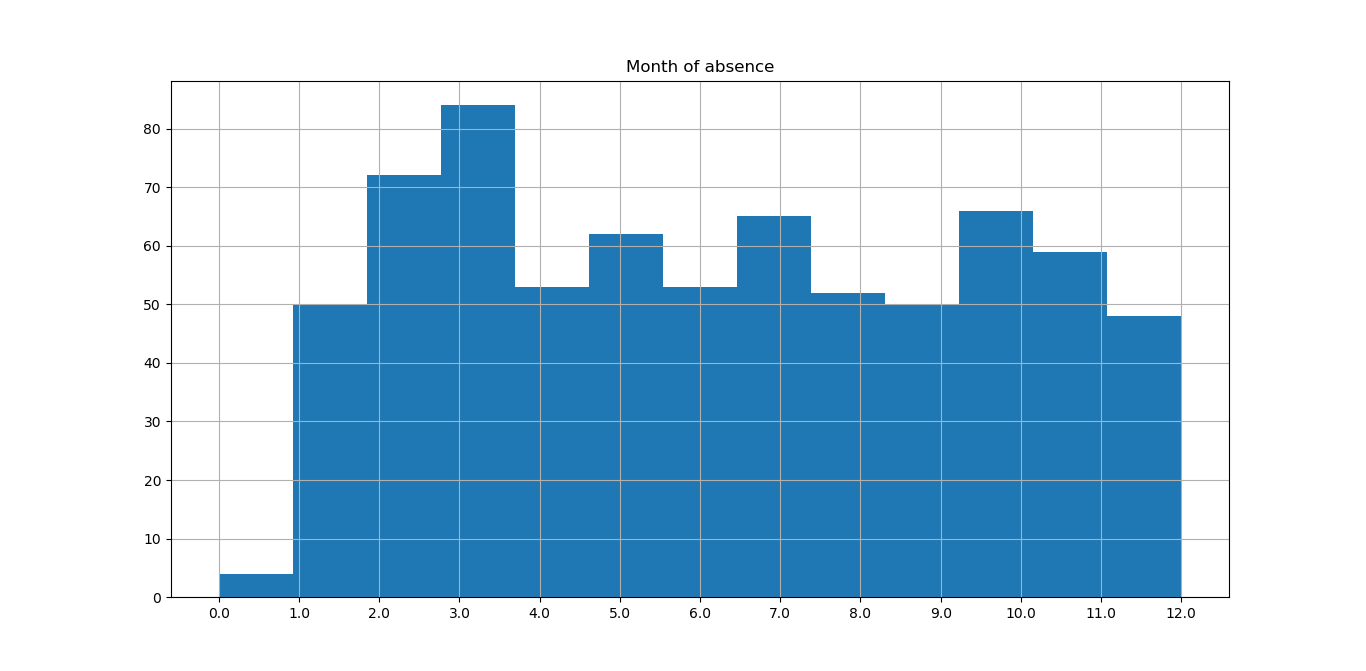
Social smoker have a slight more count than the ones who don’t smoke.



Seasons has a pretty much equal distribution.



Here, the ones who don’t have pets have a higher frequency of absenteeism. So this variable doesn’t much contribute in a sense that having a pet may or may not reduce their absenteeism frequency.



The month of absence variable also has more or less the same distribution for every value, so this too doesn’t contribute much.