**PROJECT REPORT**

**EMPLOYEE ABSENTEEISM**

**BY**

**DIBYANSHU KUMAR**

**24-03-2019**

CONTENTS:

1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

3 EDA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

4 Data Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

5 Missing Value Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

6 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

7 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

8 Feature Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

9 Univariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

10 Bivariate Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

11 Inferences from Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

12 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

**INTRODUCTION**

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

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)

EDA

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

For the given data:

Information of data:

RangeIndex: 740 entries, 0 to 739

Data columns (total 21 columns):

ID 740 non-null int64

Reason for absence 737 non-null float64

Month of absence 739 non-null float64

Day of the week 740 non-null int64

Seasons 740 non-null int64

Transportation expense 733 non-null float64

Distance from Residence to Work 737 non-null float64

Service time 737 non-null float64

Age 737 non-null float64

Work load Average/day 730 non-null float64

Hit target 734 non-null float64

Disciplinary failure 734 non-null float64

Education 730 non-null float64

Son 734 non-null float64

Social drinker 737 non-null float64

Social smoker 736 non-null float64

Pet 738 non-null float64

Weight 739 non-null float64

Height 726 non-null float64

Body mass index 709 non-null float64

Absenteeism time in hours 718 non-null float64

dtypes: float64(18), int64(3)

memory usage: 121.5 KB

Type of Problem:

The data consists of 10 continuous variables and 11 categorical variables . Out of the 10

continuous variable **‘Absenteeism time in hours’** is our target variable. Since this variable is

continuous therefore it is a **regression** problem.

The continuous and categorical data’s are as follows:

Continuous= ['Distance from Residence to Work', 'Service time', 'Age', 'Work load Average/day ',

'Transportation expense', 'Hit target', 'Weight', 'Height', 'Body mass index',

'Absenteeism time in hours']

Categorical = ['ID', 'Reason for absence', 'Month of absence', 'Day of the week',

'Seasons', 'Disciplinary failure', 'Education', 'Social drinker',

'Social smoker', 'Pet', 'Son']

DATA PRE PROCESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in

certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a

proven method of resolving such issues.

In this project we would use preprocessing technique such as missing value analysis, outlier

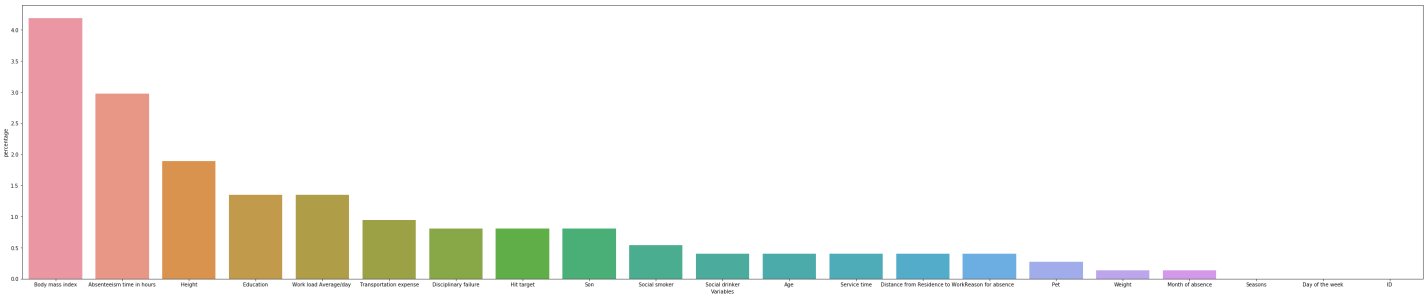
analysis, feature selection and feature scaling.

MISSING VALUE ANALYSIS:

We have seen in the data info that this data consists of missing values. We have to impute this

missing values in a right way so that our model predictions are good.

Here is the plot for percentage of missing data:



|  | **no of nulls** | **percentage** |
| --- | --- | --- |
| **Variables** |  |  |
| **Body mass index** | 31 | 4.189189 |
| **Absenteeism time in hours** | 22 | 2.972973 |
| **Height** | 14 | 1.891892 |
| **Education** | 10 | 1.351351 |
| **Work load Average/day** | 10 | 1.351351 |
| **Transportation expense** | 7 | 0.945946 |
| **Disciplinary failure** | 6 | 0.810811 |
| **Hit target** | 6 | 0.810811 |
| **Son** | 6 | 0.810811 |
| **Social smoker** | 4 | 0.540541 |
| **Social drinker** | 3 | 0.405405 |
| **Age** | 3 | 0.405405 |
| **Service time** | 3 | 0.405405 |
| **Distance from Residence to Work** | 3 | 0.405405 |
| **Reason for absence** | 3 | 0.405405 |
| **Pet** | 2 | 0.270270 |
| **Weight** | 1 | 0.135135 |
| **Month of absence** | 1 | 0.135135 |
| **Seasons** | 0 | 0.000000 |
| **Day of the week** | 0 | 0.000000 |
| **ID** | 0 | 0.000000 |

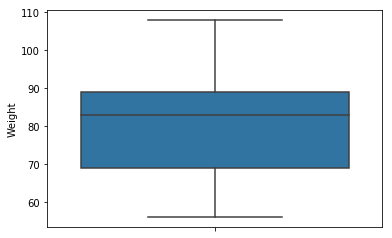
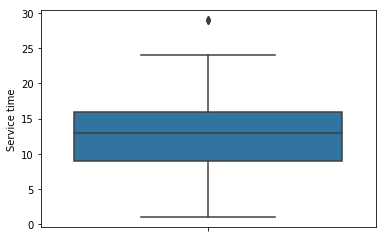
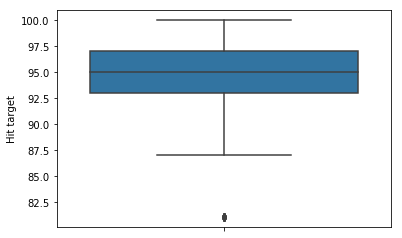
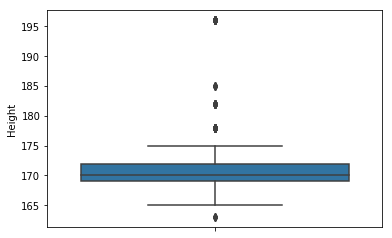
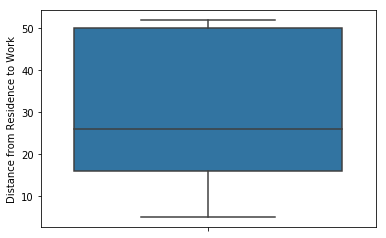
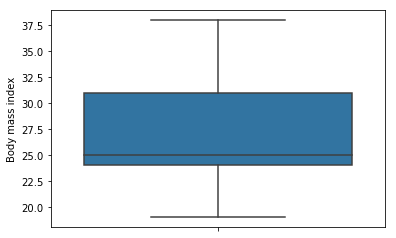
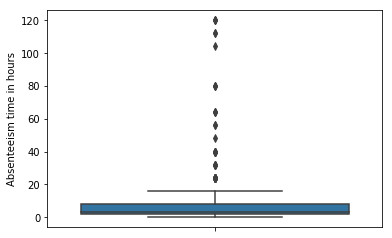
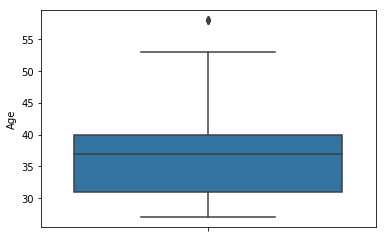
We can see that ‘Body mass index’ contains the most number of null. For imputing null values we need to check which method is the best. For this we will create an arbitrary null value in the ‘Body mass index’ and then would use all the methods to impute the null values. The best method would be chosen on the basis of the closeness of the imputed arbitrary value to the original value and also the change in standard deviation after the imputation.

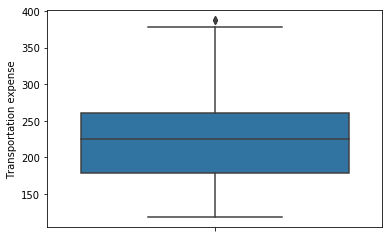
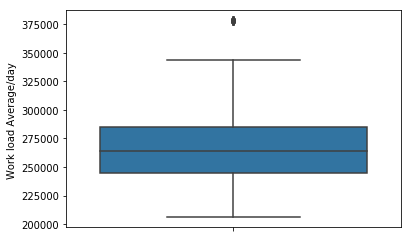
In this problem we have used three methods of imputation i.e. imputation by mean, imputation by median and imputation using K nearest neighbors. Here K nearest neighbors gave the closest value and least change in standard deviation. So we have chosen KNN method for imputation of missing values.

OUTLIER ANALYSIS

In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

**Box Plot** is the best method to find the outliers. For this data set we have plotted some boxplot for the continuous data as shown:





After checking for outliers in continuous data we got:

Service time 5

Age 8

Work load Average/day 29

Transportation expense 3

Hit target 19

Height 114

Absenteeism time in hours 43

These are the count of outliers for respective variables in the given data.

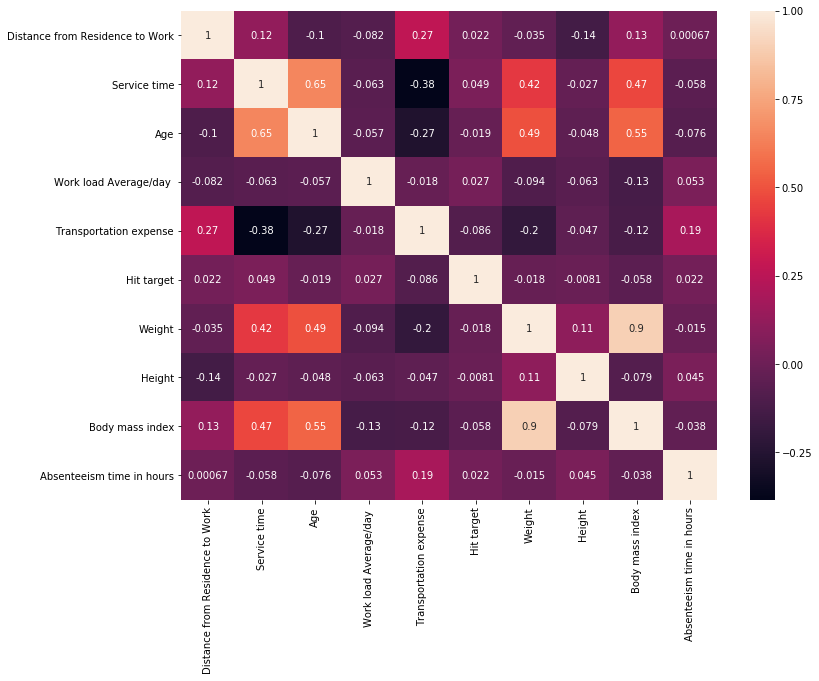
Removing outliers would be a bad option over here because we don’t have a very large dataset to train our model. So we would first change the outliers to null values and then impute those values using KNN imputation technique.

FEATURE SELECTION

Feature selection is a process where we select those features which contribute most to our prediction variable or output in which we are interested in. Having irrelevant features in our data can decrease the accuracy of the models and make our model learn based on irrelevant features.

For **continuous variable**:

Here we have used seaborn heatmap to get the correlation between variables in visual form.



In the heatmap we can see that weight and body mass index are quite correlated with each other hence out of these two we are taking one of the two variables. We would drop the weight column from the data.

For **Categorical variables**:

In case of categorical variables we would use ANOVA Test (Analysis of variance) since our target variable is continuous. In this test the mean of target is calculated for each category of a categorical variable and is compared. On the basis of this a p-value is calculated. If p-value is less than 0.05 then we reject the null hypothesis.

The **null hypothesis** for **ANOVA** is that the mean (average value of the dependent variable) is the same for all groups. The alternative or research **hypothesis** is that the average is not the same for all groups. The **ANOVA** test procedure produces an F-statistic, which is used to calculate the p-value.

For the variable to be in the model, the p-value must be less than 0.05.

For this data all categorical variables had p-value less than 0.05.

P value of ID is 8.449881295013552e-167

P value of Reason for absence is 9.770767089088417e-277

P value of Month of absence is 3.3124782278857673e-25

P value of Day of the week is 0.0008188161594849071

P value of Seasons is 3.127506937786291e-40

P value of Disciplinary failure is 1.2189432024253421e-185

P value of Education is 8.375003325123203e-105

P value of Social drinker is 1.2794395762714786e-150

P value of Social smoker is 9.117849965003895e-184

P value of Pet is 5.325984030592952e-127

P value of Son is 9.45269711512623e-116

FEATURE SCALING

**Feature scaling** is a method used to standardize the range of independent variables or

features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use **Normalization** as Feature Scaling Method.

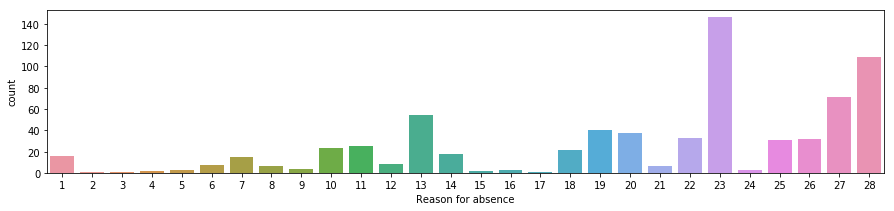
Normalization = (ith data – min data)/(max data – min data)

UNIVARIATE ANALYSIS

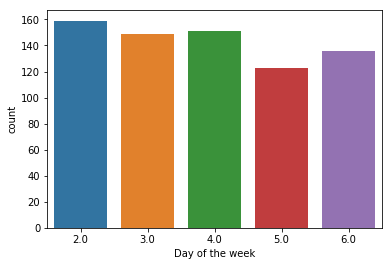
Here we will be analyzing all the variables individually.

For the categorical variables we will see the count plots of the respective variables:

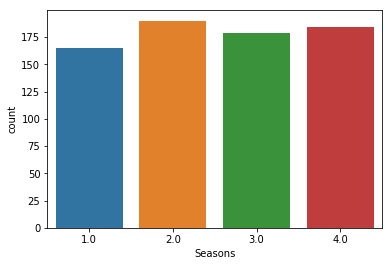
1) Reason for absence



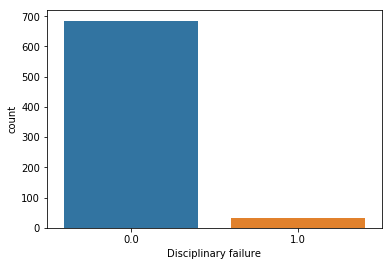
2) Day of the week



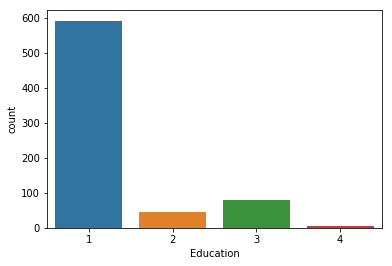
3) Seasons



4) Disciplinary Failure



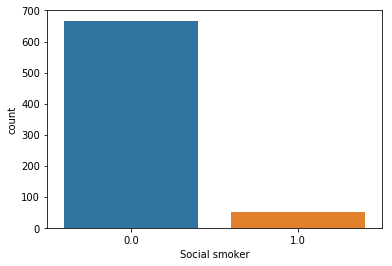
5) Education



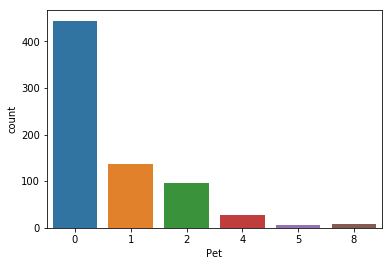
6) Social drinker



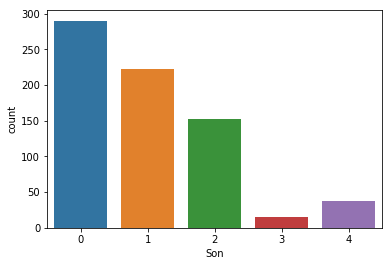
7) Social smoker



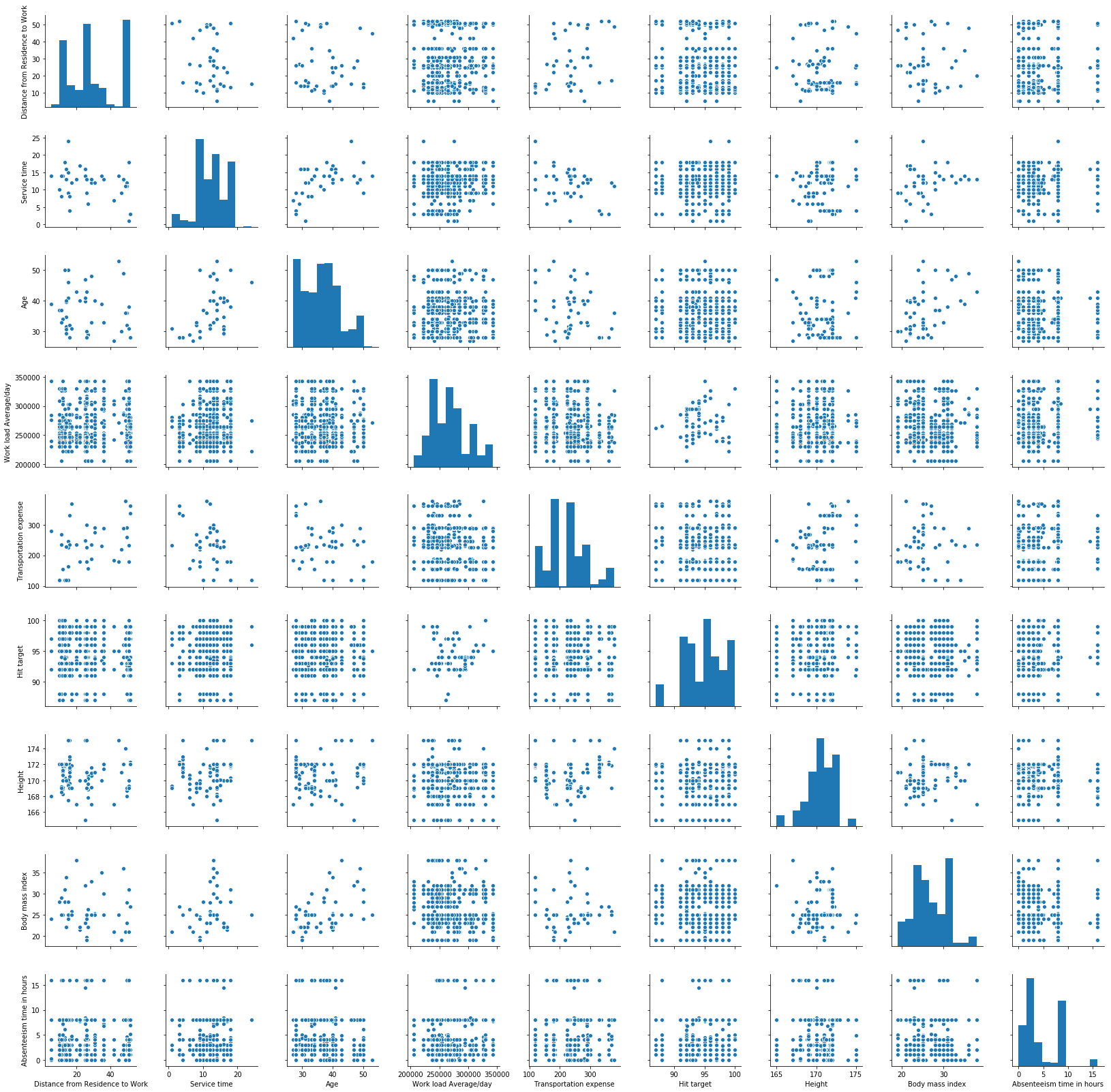
8) Pet



9) Son



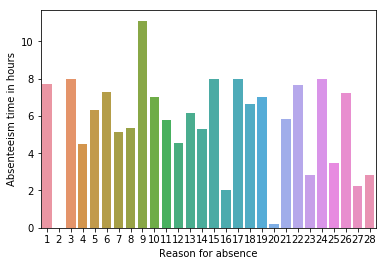
For continuous variables we have plotted a pair plot.



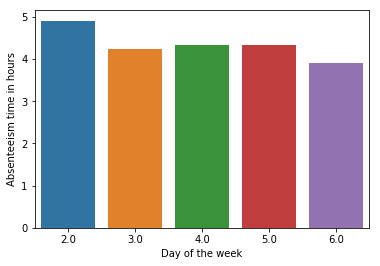
BIVARIATE ANALYSIS

Now we would see how the variables are dependent on our target variable (Absenteeism time in hours).

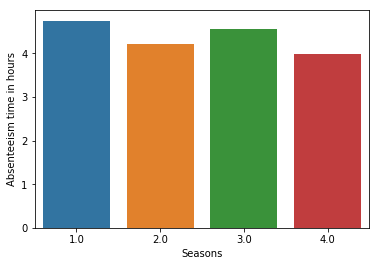
1) ‘Reason of absence’ and ‘Absenteeism time in hours’



2) ‘Day of week’ and ‘Absenteeism time in hours’



3) ‘Seasons’ and ‘Absenteeism time in hours’



4) ‘Disciplinary Failure’ and ‘Absenteeism time in hours’



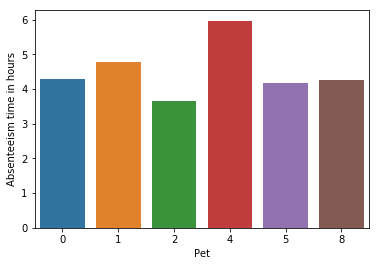
5) ‘Social drinker’ and ‘Absenteeism time in hours’



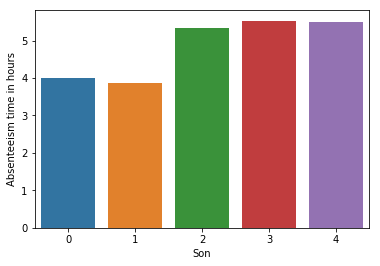
6) ‘Social smoker’ and ‘Absenteeism time in hours’



7) ‘Pet’ and ‘Absenteeism time in hours’



8) ‘Son’ and ‘Absenteeism time in hours’



INFERENCES FROM THE ANALYSIS:

* **‘Reason for absence’** has 28 different categories. Although reason no 23, 27 and 28 are more frequent (as we can see in the count plot) but reason no 9 has maximum hours of absence.

Since most reasons are ‘medical consultation’, a regular health checkup in the company itself can decrease the absence ratio. Also diseases related to musculoskeletal system are more and people are also going for physiotherapy which means work load per employee is a little more. A blood donation camp organized in the company could decrease the absence ratio more.

* **‘Day of the week’** and **‘Seasons’** does not have much dependency on ‘Absenteeism time in hours’
* ‘**Education**’ also does not depend much on target but people with ‘high school’ qualification are absent more frequently.
* People who often **drink** and **smoke** are absent for more hours.
* Absenteeism time in hours increases as the no of children of the employees increases. Although the frequency of absence decreases.
* **ID** no 3,11, 28 and 20 have most no of absenteeism hours.
* Most of the continuous data are not distributed normally.

MODELLING

This is the final phase of our project where we would build some machine learning models and will train our model on the data for future predictions. We will also apply Principal Component Analysis to reduce the dimension of our data. We would consider different machine learning algorithms to check which gives the best result.

**Algorithms used**: In this project we have used three ML algorithms i.e. Random Forest, XGboost and Linear Regression.

**Metrics used for validation:** Root mean square error, R^2 score and K-fold cross validation.

**Results before applying PCA:**

**Random Forest:**

Root Mean Squared Error For Test data = 2.9298663006816956

R^2 Score = 0.3256183743889529

mean of accuracies of 10 folds = 0.34341484170008874

standard deviation of accuracies of 10 folds = 0.1950853956136966

**XGboost:**

Root Mean Squared Error For Test data = 3.0027673996569515

R^2 Score = 0.2916408499482207

mean of accuracies of 10 folds = 0.3747530814867582

standard deviation of accuracies of 10 folds = 0.17696861141034242

**Linear Regression:**

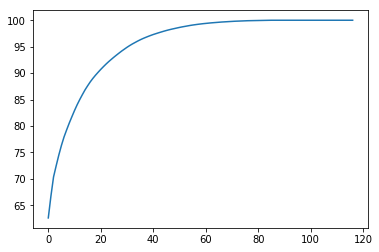
Root Mean Squared Error For Test data = 4113093289331.5054

R^2 Score = -1.329068128365188e+24

mean of accuracies of 10 folds = -2.436559831584439e+25

standard deviation of accuracies of 10 folds = 3.7804920679299655e+25

Principal Component Analysis (PCA):



From this graph of PCA we get that with approx. 45 variables we could explain our data so that it neither overfits nor underfits the model.

So after taking 45 majority variables which tells about the data the most, we got quite better results.

**Results after applying PCA (dimensionality reduction):**

**Random Forest:**

Root Mean Squared Error For Test data = 0.027997411883770484

R^2 Score = 0.9999294303713552

mean of accuracies of 10 folds = 0.9992887285308395

standard deviation of accuracies of 10 folds = 0.0010039220505104827

**XGboost:**

Root Mean Squared Error For Test data = 0.026824201363644235

R^2 Score = 0.9999352207880676

mean of accuracies of 10 folds = 0.9995829826757993

standard deviation of accuracies of 10 folds = 0.001026387070994959

**Linear Regression:**

Root Mean Squared Error For Test data = 1.1888939426194731e-05

R^2 Score = 0.9999999999872747

mean of accuracies of 10 folds = 0.9999999301846246

standard deviation of accuracies of 10 folds = 1.3952101035353957e-07

Therefore we can see that linear regression gives the best set of results thus we would go for this model for our predictions.