**Absenteeism of Employees Project**

**Arnav Ashank**

27 November 2019

## Contents

## Introduction---------------------------------------------------------------------------------------3

Problem Statement------------------------------------------------------------------------------------------------3

Data Overview-------------------------------------------------------------------------------------------------------3

Variables--------------------------------------------------------------------------------------------------------------5

## Methodology---------------------------------------------------------------------------------------7

Pre Processing-------------------------------------------------------------------------------------------------------7

Missing Value Analysis---------------------------------------------------------------------------------------------7

Exploratory Data Analysis-----------------------------------------------------------------------------------------8

Outlier Analysis------------------------------------------------------------------------------------------------------15

## Assessing the data--------------------------------------------------------------------------------18

Exploratory Data Analysis----------------------------------------------------------------------------------------18

Feature Scaling-----------------------------------------------------------------------------------------------------24

## Modeling the data--------------------------------------------------------------------------------28

Building the model--------------------------------------------------------------------------------------------------28

R-------------------------------------------------------------------------------------------------------------------------28

Decision Tree---------------------------------------------------------------------------------------------------------28

Random Forest-------------------------------------------------------------------------------------------------------29

Linear Regression----------------------------------------------------------------------------------------------------29

XGBoost Regression-------------------------------------------------------------------------------------------------30

Support Vector Machines------------------------------------------------------------------------------------------30

Python------------------------------------------------------------------------------------------------------------------32

## Conclusion------------------------------------------------------------------------------------------36

Model Evaluation----------------------------------------------------------------------------------------------------36

Solution to problem statements---------------------------------------------------------------------------------36

## 

**Chapter 1**

**Introduction**

## 1.1 Problem Statement

XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The aim of the project is to find the changes the company can bring so as to reduce number of absenteeism. Further based on the given dataset provided by the company we have to predict the losses every month if same trend of absenteeism continues. Losses here can be considered in terms of absenteeism hours of employees.

## 1.2 Data Overview

Our task is to build a regression model which will predict the losses if the same trend of absenteeism among employees continues and also find as to which factor plays a major role in it . Given below is a sample of the data set that we are using to predict the losses:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Reason for absence | Month of absence | Day of the week | Seasons | Transportation expense |
| 11 | 26 | 7 | 3 | 1 | 289 |
| 36 | 0 | 7 | 3 | 1 | 118 |
| 3 | 23 | 7 | 4 | 1 | 179 |
| 7 | 7 | 7 | 5 | 1 | 279 |
| 11 | 23 | 7 | 5 | 1 | 289 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Distance from Residence to Work | Service time | Age | Work load Average/day | Hit target | Disciplinary failure |
| 36 | 13 | 33 | 239,554 | 97 | 0 |
| 13 | 18 | 50 | 239,554 | 97 | 1 |
| 51 | 18 | 38 | 239,554 | 97 | 0 |
| 5 | 14 | 39 | 239,554 | 97 | 0 |
| 36 | 13 | 33 | 239,554 | 97 | 0 |

Table 1.1: Absenteeism at work Sample Data (Columns: 1-6)

Table 1.2: Absenteeism at work Sample Data (Columns: 7-12)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Education | Son | Social drinker | Social smoker | Pet | Weight | Height | Body mass index |
| 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 |
| 1 | 1 | 1 | 0 | 0 | 98 | 178 | 31 |
| 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 |
| 1 | 2 | 1 | 1 | 0 | 68 | 168 | 24 |

Table 1.3: Absenteeism at work Sample Data (Columns: 13-20)

Table 1.4: Absenteeism at work Sample Data (Column: 21)

|  |
| --- |
| Absenteeism time in hours |
| 4 |
| 0 |
| 2 |
| 4 |

As we can see that there are 20 predictor variables using which we have to predict the absenteeism hours. They are:

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

As the dependent variable is continuous type so this is a regression type problem.

**1.3 Variables**

There are 21 variables in our data in which 20 are independent variables and 1 (Absenteeism time in hours) is dependent variable. Variable 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.

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 (KMs)

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**

## Methodology

## 

## 2.1 Pre Processing of the data

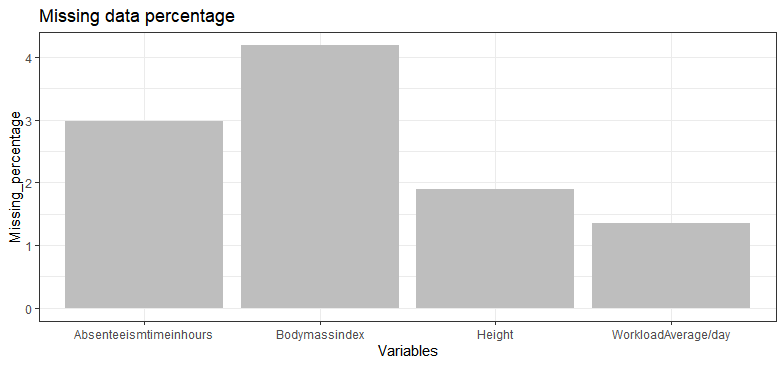
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. In this project we look at the continuous and categorical variables in the dataset. Also we will go through the missing value and outlier analysis.

## 2.2 Missing Value Analysis

Here we first count the null values present in any variable including the dependent variable. If the null values are more than 30 percent then it won’t be wise to impute or drop it. Otherwise we can either impute the null values or simply drop that record from the dataset. Here we have imputed the missing values using the mean and median method. Here we have not used KNN method of imputation due to system constraints in both R and Python languages.

Here we have first replaced a known value with null in the dataset and stored it in a variable. Next we have imputed the value using both the methods mean and median. After imputing the missing values using both methods each of the imputed value is compared with the actual value stored originally and the method which imputes the value closest to actual value is used for imputing missing values of that particular variable. In python this procedure has been performed using user defined function and loops, and in R it showing memory error so only user defined function has been used and no loops.

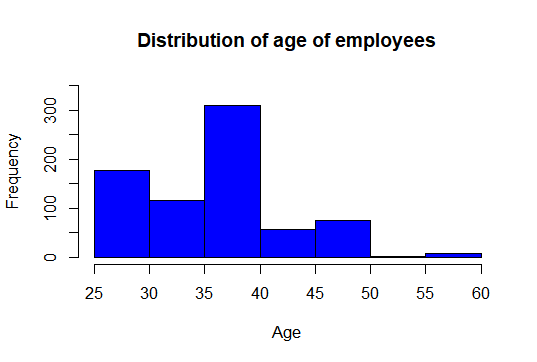
After performing the missing value analysis the data is cross checked so that no null values remain leftover.

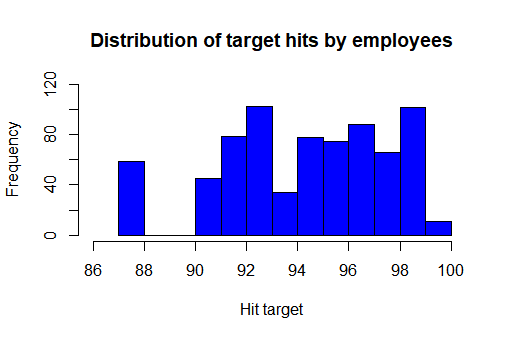
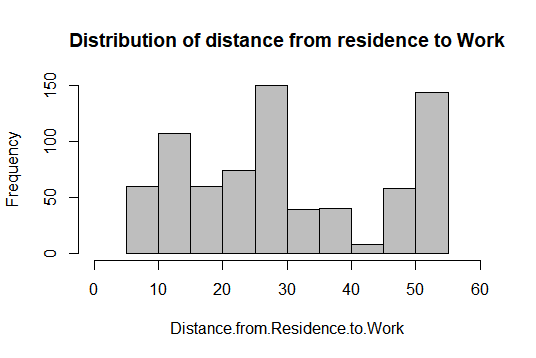


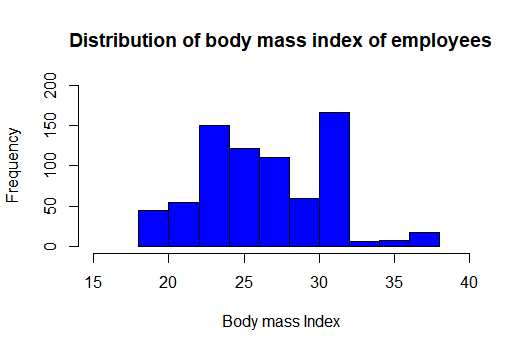
The figure above shows percentage of missing values present in some of the variables present in the dataset.

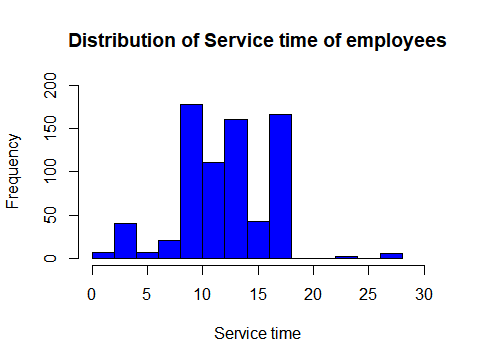
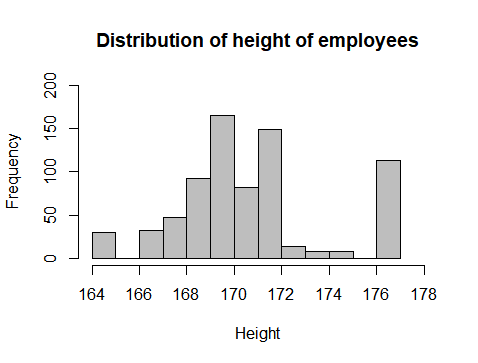
## 2.3 Exploratory Data Analysis

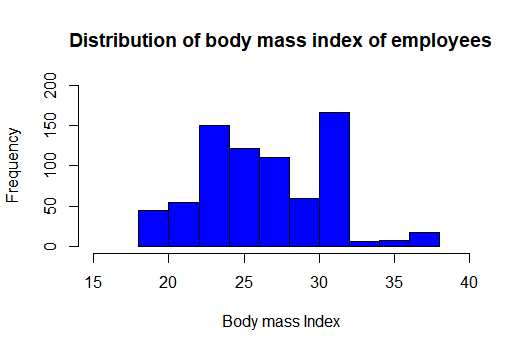
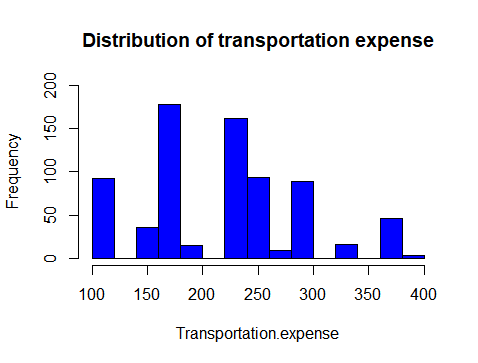
Here we have visualized some variables present in the dataset. For continuous variables we have used histograms to see their distribution in the dataset and for categorical variables we have used pie charts to see as to which category in a given categorical variable has major contribution in the dataset.

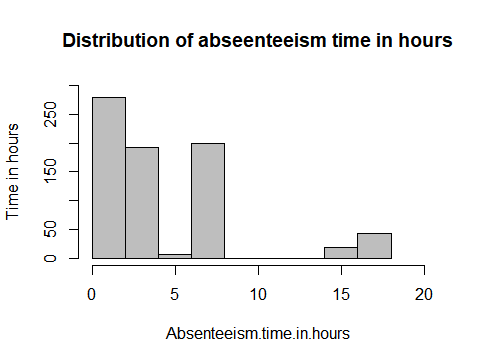


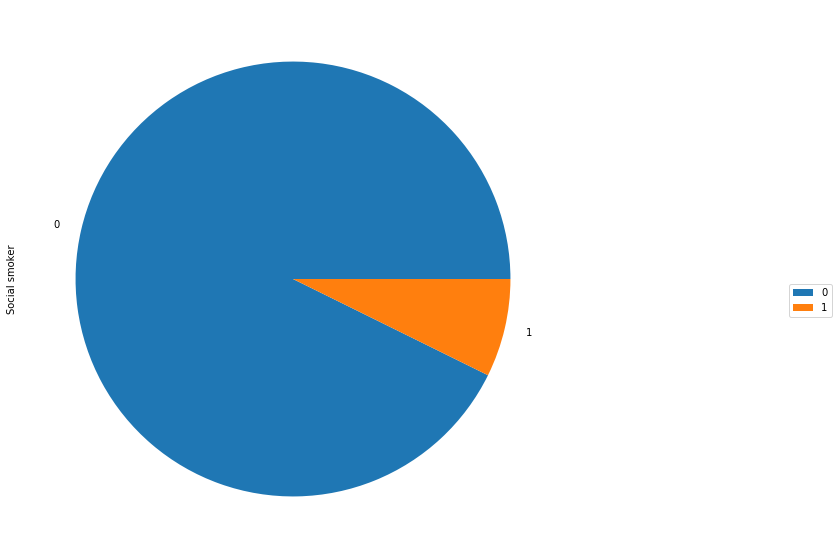




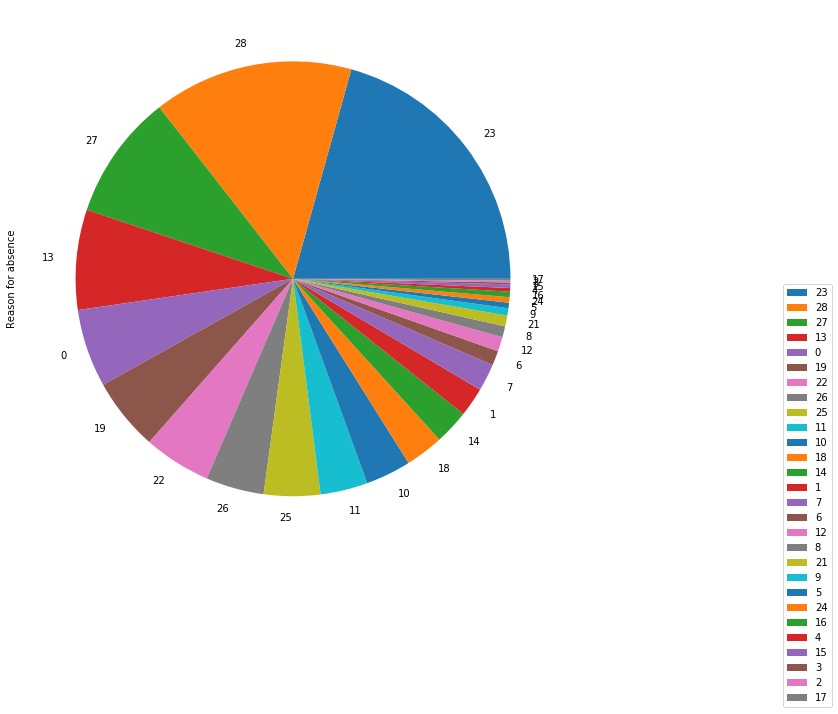
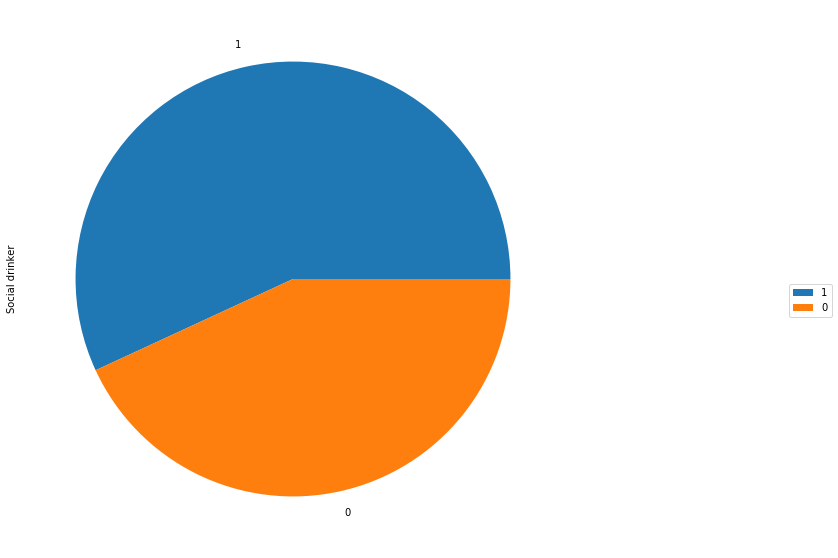






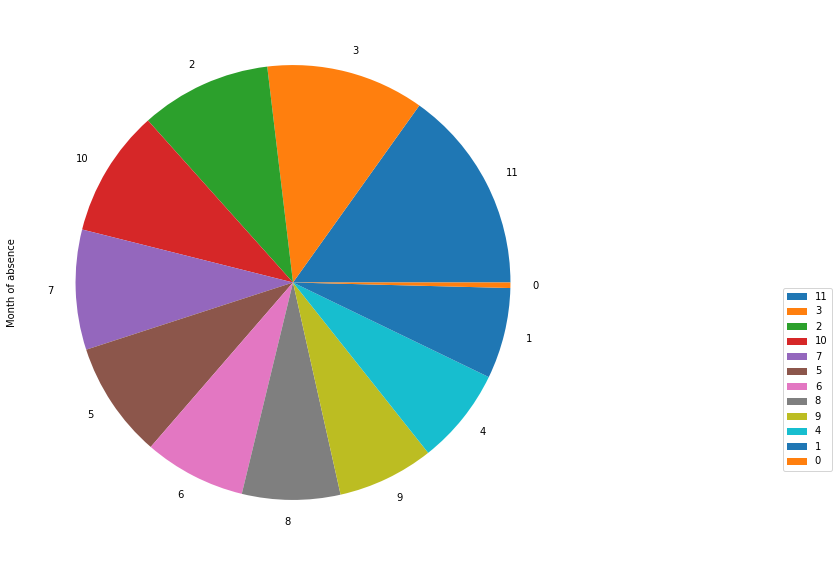


Pie chart-1

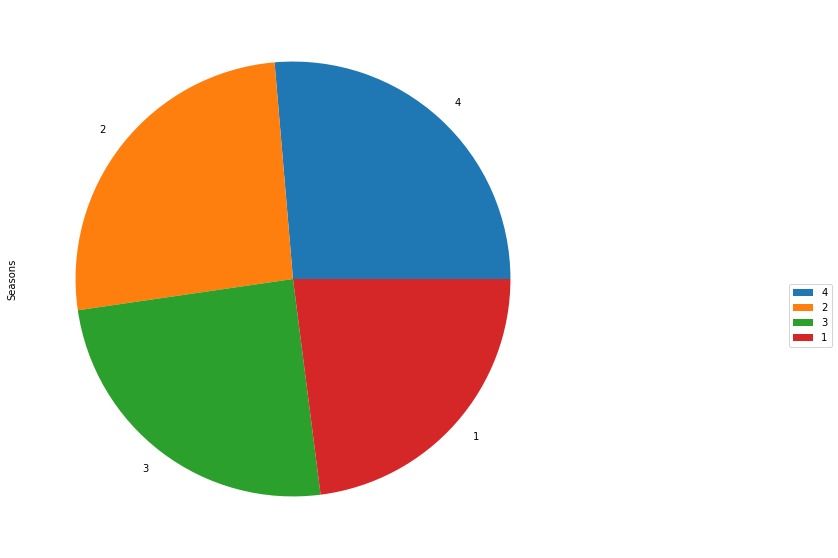


Pie chart-2

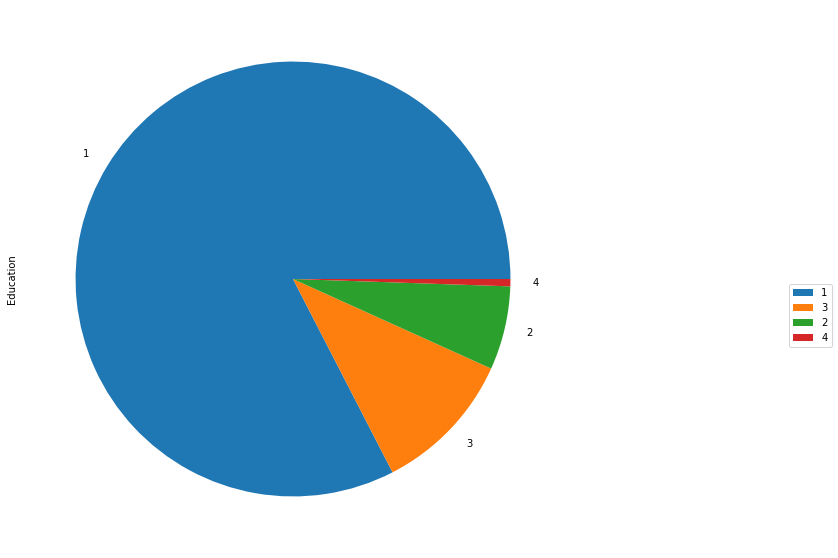
Pie chart-3



Pie chart-4



Pie chart-5



Pie chart-7

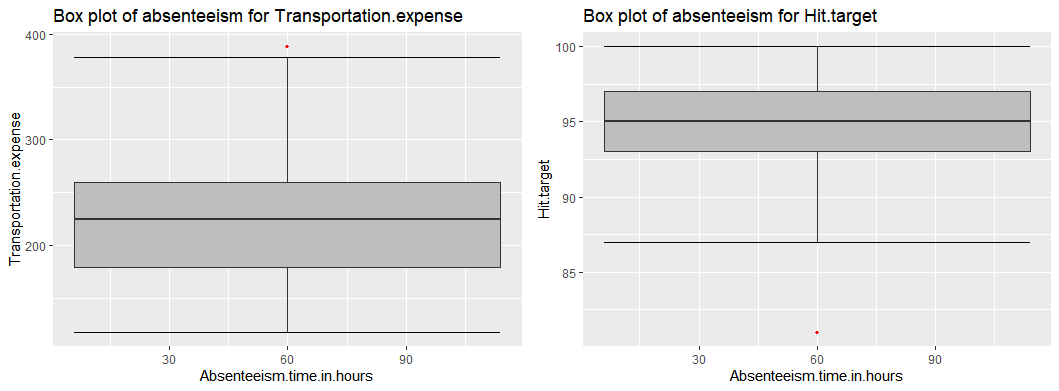
Pie chart-6

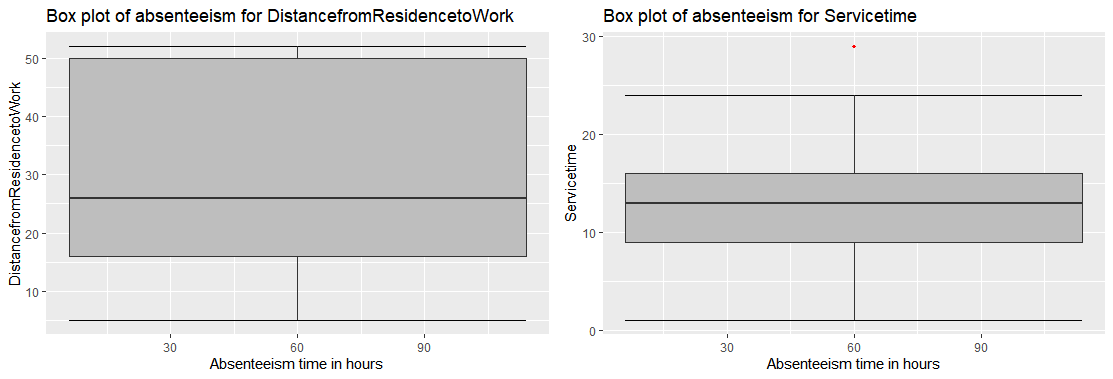
From the graphs above it is clear that

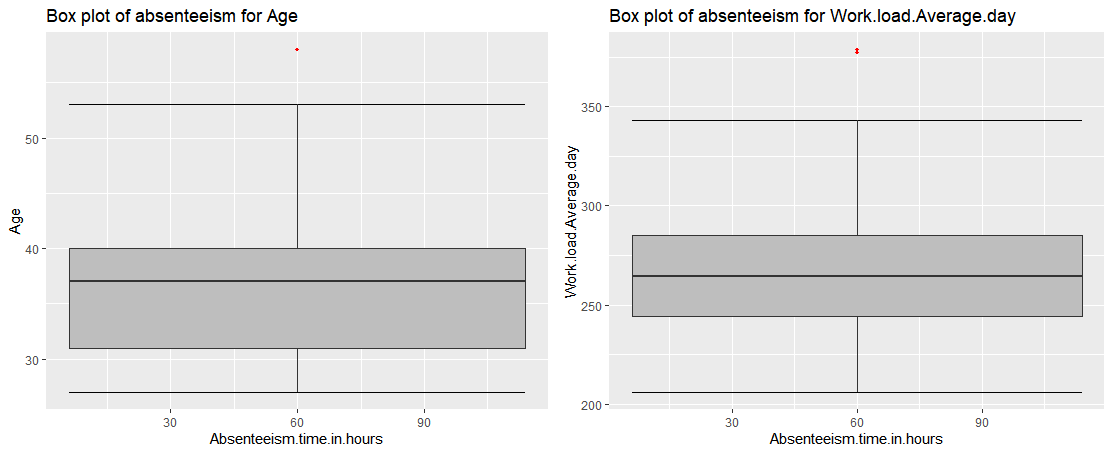
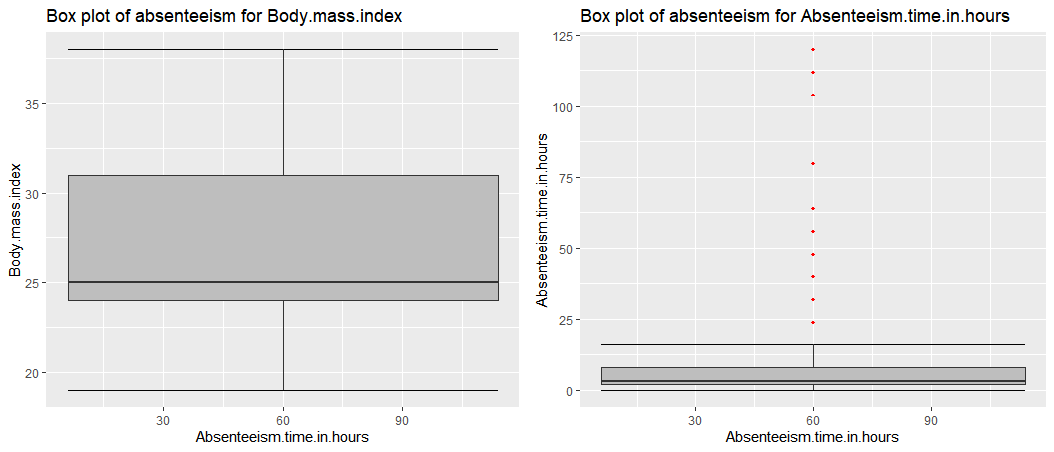
* The figure for distribution of absenteeism time in hours of the employees. It clearly shows that majority of the employees have absenteeism hours within 10 approximately. There are some outliers present which duration of more than 20 and few of the employees have absenteeism hours up to 120. it clearly indicates the presence of outliers in this variable.
* In the histogram distribution of transportation expense is plotted. It shows that majority of employees have transportation expense between 150 and 300. Also a bar between 350 and 400 clearly depicts presence of outliers in it.
* The figure showing distribution of distance from residence to work of employees. There are three peak bars between around 10,26 and 50km approximately, which shows that majority of employees have live around these.
* The histogram showing distribution of service time of employees i.e. for how long they have been working in the company. Majority of the employees are those who have served for more than 5 years up to 20 years. Also there are some old employees who have served for more than 25 years.
* The histogram for transportation expense shows that it has a wide range from 100 to 400 with peaks at approximately 170 and 230.
* The body mass index varies from 16 to 38 approximately for all the employees with majority of them having body mass index more than 25 which is not good. Body mass index for a healthy human being should be between 20 to 25. It may be one of the reason for absence of employees.
* The histogram for absenteeism shows that majority of employees have absenteeism hours less than 5 hours though some of them have absenteeism hours up to 20 also.
* From the pie chart-1 It is clear that majority of employees do not smoke.
* From pie chart-2 it is clear that majority of employees give reason 23 for absence which is for medical consultation.
* Pie chart-3 depicts that majority employees have the habit of drinking.
* Pie chart -4 shows that most of the employees make absent in month 11 i.e. (December as index here starts from 0).
* Pie chart-5 shows that season 4 has majority of absenteeism counts (not sum of absenteeism hours).
* Pie chart-6 shows that majority of the employees are not graduate and only have completed high school education which is education level 1.
* Pie chart-7 shows that distribution of employees in each day of the week is approximately constant.

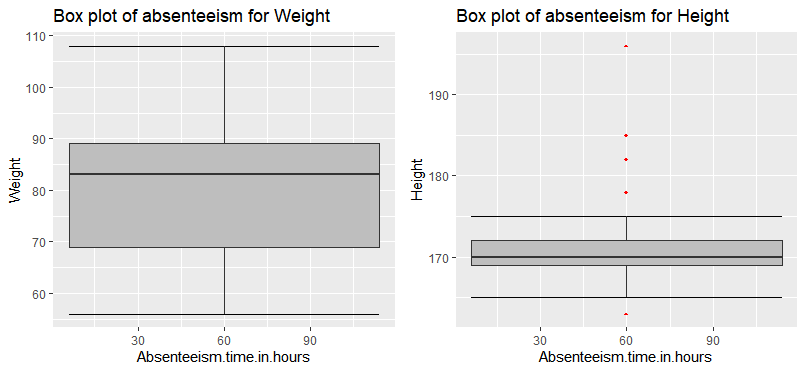
## 2.4 Outlier Analysis

From the histograms obtained for continuous variables it has been found that many of them have skewed distribution. To get a clear confirmation for outliers in a variable box plot analysis has been done for each continuous variable. Here the variable who have data points outside the whiskers are considered to have outliers. Below are shown box plot for each continuous variables.









From the above box plots it is clear that outliers are present in Transportation expense, Service time, Age, Hit target and Height.

So we need to remove these outliers. Here flooring and capping method has been used to remove the outliers. In this method we compare each value of a variable with two values

Min = q25-1.5\*iqr

Max = q75+1.5\*iqr

Where q25 = 25th percentile of that variable

q75= 75th percentile of that variable

iqr = inter-quartile range of that variable.

If a value exceeds Max then it is replaced with Max similarly when a value is less than Min it is replaced with min. Thus outliers are removed from every variable.

**Chapter 3**

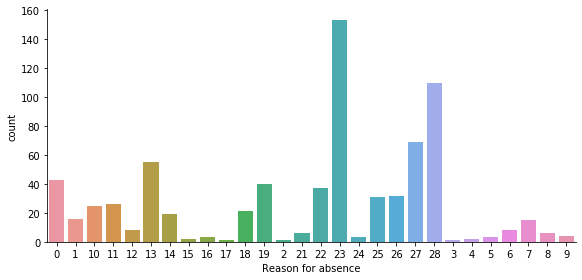
**Assessing the Data**

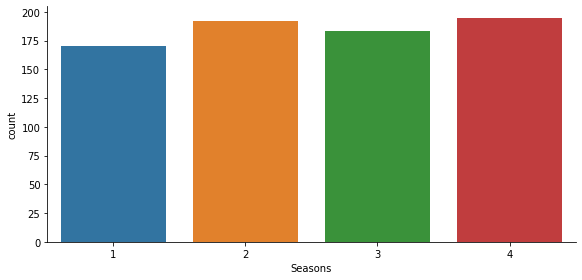
## 3.1 Exploratory Data Analysis

Here we have used multivariate plots to analyze how and how much each variable affects the absenteeism hours for employees in both R and Python.

Here we have first used count plot for each categorical variable to figure as to which category in a categorical variable has most number of counts of absenteeism hours.

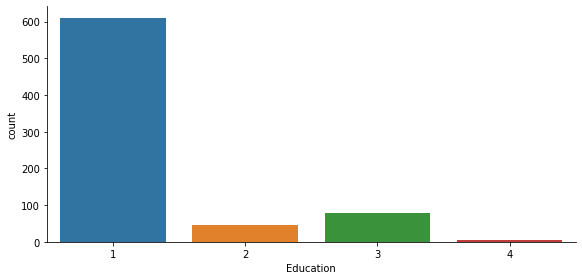
The first count plot is for reason for absence it clearly shows that that most of the employees who are absent give reason 23 followed by reason 28 and 27. Reason 23 is for medical consultation, reason 28 is for dental consultation and 27 for physiotherapy.



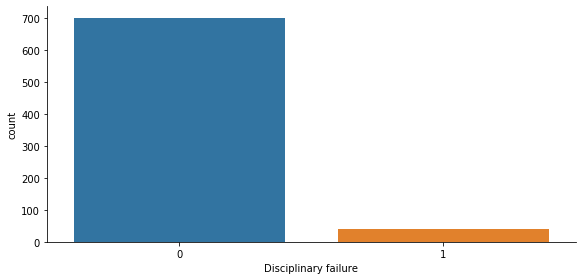
The figure above is count plot for seasons for no. of absenteeism counts of employees in a given season.

The bar plot clearly shows that absenteeism counts in every season is between 150 and 200. Among them season 4 i.e. spring followed by autumn and summer.

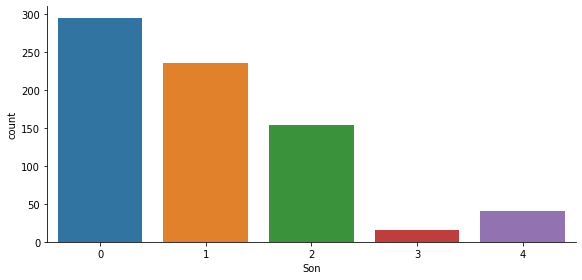
The figure below is count plot for education, which indicates the count of absenteeism for each level of education of employees.

The bar plot clearly indicates that employees having education level 1 (high school) have most number of absenteeism counts in the dataset.

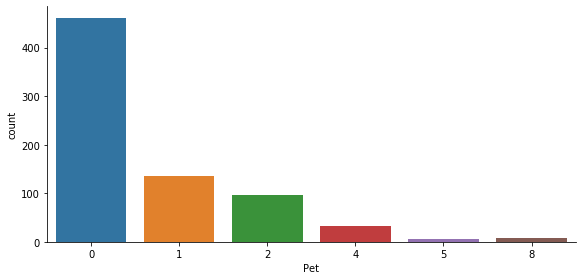
The figure below is count plot for disciplinary failure of employees.



It clearly indicates that of the employees making absent most of them are disciplined. It also shows that this parameter might not be much useful for predicting absenteeism.

The figure below is for count of absent made by employees based on the number of son they have.

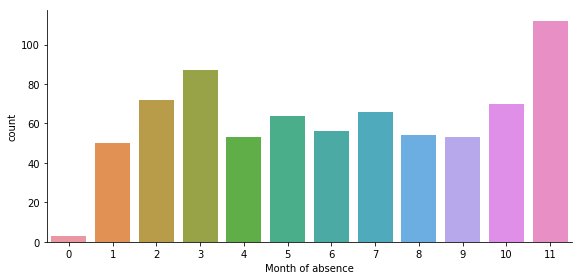
It clearly indicates that employees who do not have any son make the most absents. There could be reason that majority of employees are not married but no marital status has been provided in the data so we cannot conclude here.



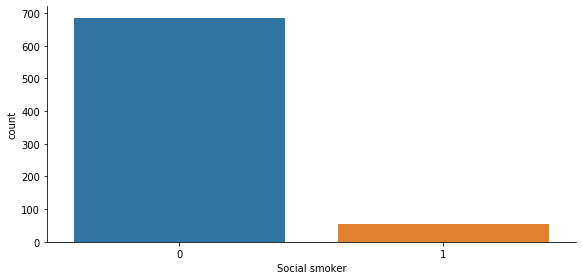
The figure below is for count of absent made by employee depending upon the no of pets they have. It clearly shows that employees who have no pets make the most number of absents and those who have more than 2 pets make the least no of absents. This may be a very important aspect when hiring a new employee the company can hire employees who have pets.

The figure below is for count of absents in each month. It shows that employees make the most absents in the month 11, followed by 3 and 2.

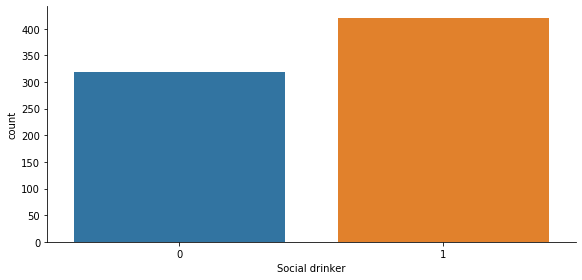
Month 11- December, 3- April, 2- March.

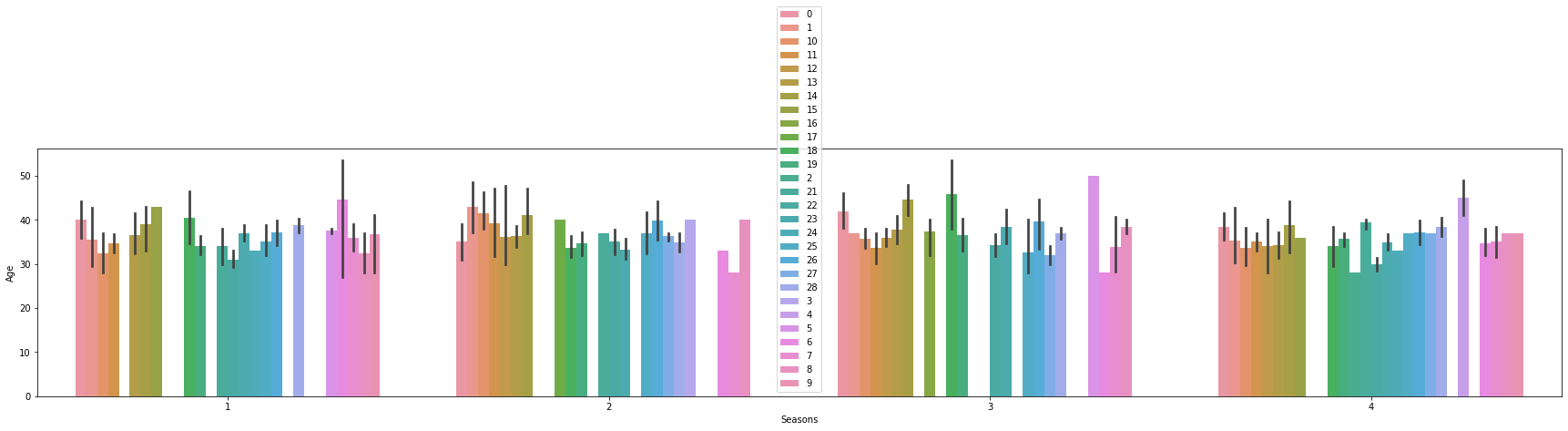


The figure shows the count of absents of employees based on whether they smoke or not. It clearly shows that employees who smoke make the most absents. This information can be used by the company while hiring new employees.



The figure below shows the count of absent of employees based on whether they drink or not. It quite evident from the figure below that employees who drink have higher probability of making absents compared to those who do not drink.

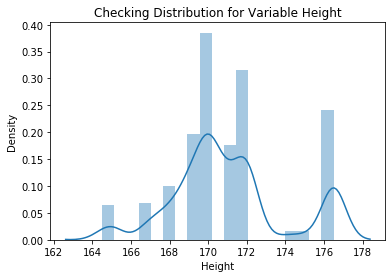


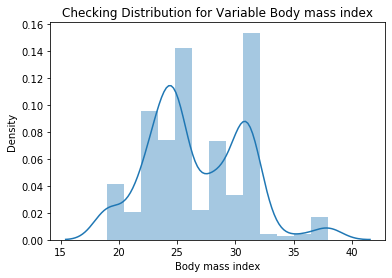
The figure below shows a barplot for age versus season with colors defining the reason of absence. It is very clear that employees in the age group 30-40 make the most reason for absence in every season. Previously from count plot we know that most used reason for absence is 23 which is medical consultation and here it is quite visible that employees giving this reason for absence are aged 30.

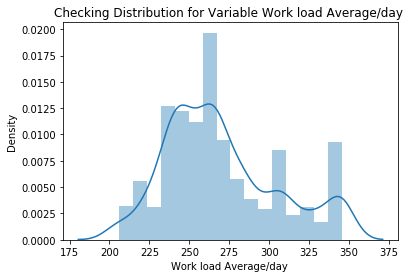
## 3.2 Feature Scaling

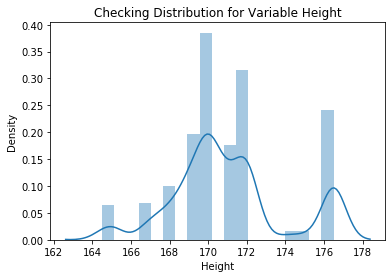
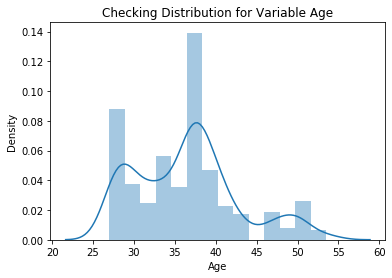
In the figures below histograms of individual continuous variables have been plotted so as to observe the range of values in that variable. It is very clear that each variable has different range of values due to this it may be possible that variable having data in higher range may appear to affecting the dependent variable more than the having lower range or it may be vice-versa also. Thus to overcome this issue we have normalized every continuous variable except the dependent variable.

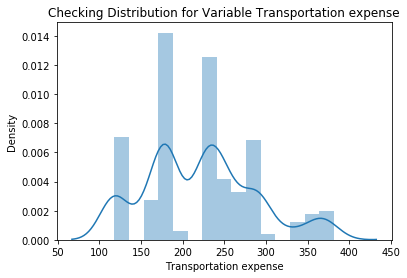
After normalization all the independent continuous variables are scaled in range from 0 to 1.



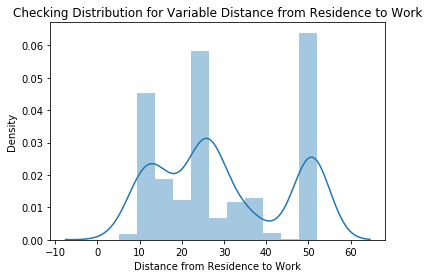








.



## Chapter 4

## Modeling the data

## 4.1 Building the Model

Here we have drawn correlation plot in both R and Python among all continuous variables to find any correlation among different continuous variables. A threshold of 0.8 is taken for assuming correlation in the given two variables.

Then we have calculated p-values for all categorical variables (except ID) using anova in both R and Python. The non-significant categorical variables having p-values greater than 0.05 have been dropped in both R and Python.

## 4.2 R

In R we have plotted the correlation plot first for the continuous variables then removed those variables which have correlation coefficient greater than 0.8 with any of the other variables. In the same way we have removed those categorical variables who have got p-values greater than 0.05.

It has been found that Weight has correlation coefficient greater than 0.8 i.e. 0.88 with Body mass Index. Seasons, Education and Social smoker have got p-values greater than 0.05. Thus in the final list for independent variables all these four variables have been dropped. Since regression is numerical analysis so we need numeric type data rather than factor type data. So we have added dummies for each categorical variable. After adding dummies one dummy variable from category has been dropped because the dummies of a categorical variable are linearly dependent and add up to one. So one dummy variable is dropped from each category to avoid any multi collinearity.

After this we split the entire dataset into two parts training and testing set in 80:20 ratio i.e. 80 percent is training dataset and 20 percent is testing set using random sampling as this is a regression problem.

## 4.2.1 Decision Tree

It is a predictive model based on a branching series of Boolean tests. It can be used for both regression and classification. There are different types of decision trees that can be used in machine learning algorithms.

It uses a model to predict a variable. It is like a flowchart structure. Each leaf/node represents an attribute or a class label. Decision tree is a rule. Each branch connects nodes with AND and multiple branches are connected by OR. It first selects a node and depending upon different categories present in that variable it will split. It will continue splitting until it covers the entire dataset. The table shows the Error metric RMSE of the Decision tree and r-squared.

|  |  |
| --- | --- |
| Error Metric | DT |
| RMSE(TRAIN) | 3.389 |
| RMSE(Test) | 3.054 |
| R2 | 0.4054 |

R-squared is basically calculation of goodness of fit of the model i.e. how much the developed model fits the dataset.

## 4.2.2 Random Forest

Random forest is an ensemble that consists of many decision trees. It builds different decision trees using different observations from the same dataset to improve accuracy and reduce weak learners hence it is called an ensemble technique.

It combines Bruimann’s bagging idea and the random selection of features. It feeds the error of one decision tree to another so as to improve accuracy. It randomly selects features to build a tree. It can be used for both regression and classification purpose.

For prediction a new sample is pushed down the tree. It is assigned the label of training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble and the average of all trees is reported as random forest prediction. In case of regression it uses mean of all trees to predict an output. The table below shows the error metric RMSE of the Random Forest and r-squared.

|  |  |
| --- | --- |
| Error Metric | RF |
| RMSE(TRAIN) | 1.9377 |
| RMSE(Test) | 2.897 |
| R2 | 0.464 |

## 4.2.3 Linear Regression

Linear regression model is basically a statistical model unlike decision tree and random forest which are machine learning models. In machine learning the model stores patterns of every variable (in the form of rules) whereas in statistical models there are coefficients of variables which are used to calculate the test data. It can also be used for imputation.

Here we first used variance inflation factor to detect multi collinearity but none of the variables were correlated. After developing the model we found that very few of the variables appear to be significant, so we kept them and removed all other variables and again performed linear regression. It was found that the RMSE of the model increased and r-squared decreased. So we decided to stick to the previous set of variables. The table below shows error metrics and r-squared of both linear regression models with original dataset and adjusted dataset.

|  |  |  |
| --- | --- | --- |
| Error Metric | LR | Adj LR |
| RMSE(TRAIN) | 3.315 | 4.185 |
| RMSE(Test) | 3.061 | 3.88 |
| R2 | 0.406 | 0.1013 |

## 4.2.4 XGBoost Regression

Gradient boosting is a [machine learning](https://en.wikipedia.org/wiki/Machine_learning) technique for [regression](https://en.wikipedia.org/wiki/Regression_(machine_learning)) and [classification](https://en.wikipedia.org/wiki/Classification_(machine_learning)) problems, which produces a prediction model in the form of an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning) of weak prediction models, typically [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning). It builds the model in a stage-wise fashion like other [boosting](https://en.wikipedia.org/wiki/Boosting_(machine_learning)) methods do, and it generalizes them by allowing optimization of an arbitrary [differentiable](https://en.wikipedia.org/wiki/Differentiable_function) [loss function](https://en.wikipedia.org/wiki/Loss_function). Like other boosting methods, gradient boosting combines weak "learners" into a single strong learner in an iterative fashion. It is easiest to explain in the least-squares [regression](https://en.wikipedia.org/wiki/Regression_analysis) setting, where the goal is to "teach" a model {\displaystyle F}to predict values{\displaystyle {\hat {y}}=F(x)} by minimizing the [mean squared error](https://en.wikipedia.org/wiki/Mean_squared_error). The table below shows the error metric RMSE of the XG boost and r-squared.

|  |  |
| --- | --- |
| Error Metric | XGBoost |
| RMSE(TRAIN) | 2.816 |
| RMSE(Test) | 3.699 |
| R2 | 0.422 |

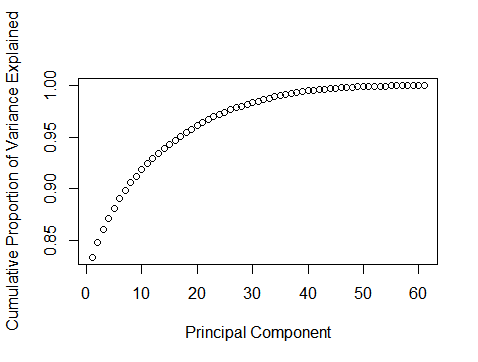
## 4.2.5 Support Vector Machines

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), support-vector machines (SVMs, also support-vector networks[[1]](https://en.wikipedia.org/wiki/Support-vector_machine#cite_note-CorinnaCortes-1)) are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis). Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-[probabilistic](https://en.wikipedia.org/wiki/Probabilistic_classification) [binary](https://en.wikipedia.org/wiki/Binary_classifier) [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. In simple regression we try to minimize the error rate. While in SVR we try to fit the error within a certain threshold.

Support Vector Regression (SVR) works on similar principles as Support Vector Machine (SVM) classification. One can say that SVR is the adapted form of SVM when the dependent variable is numerical rather than categorical. A major benefit of using SVR is that it is a non-parametric technique. Unlike SLR, whose results depend on Gauss-Markov assumptions, the output model from SVR does not depend on distributions of the underlying dependent and independent variables. Instead the SVR technique depends on kernel functions. Another advantage of SVR is that it permits for construction of a non-linear model without changing the explanatory variables, helping in better interpretation of the resultant model. The basic idea behind SVR is not to care about the prediction as long as the error (ϵi) is less than certain value. This is known as the principle of maximal margin. This idea of maximal margin allows viewing SVR as a convex optimization problem. The regression can also be penalized using a cost parameter, which becomes handy to avoid over-fit. SVR is a useful technique provides the user with high flexibility in terms of distribution of underlying variables, relationship between independent and dependent variables and the control on the penalty term. The table below shows error metric RMSE and r-squared on training and testing set.

|  |  |
| --- | --- |
| Error Metric | SVR |
| RMSE(TRAIN) | 4.189 |
| RMSE(Test) | 4.74 |
| R2 | -0.01824 |

After obtaining the above result we have used principal component analysis. Here we have first plotted variance graph of every variable.



From the graph above it is clear that out of the given variables around 25 of them explain more than 95% of variance in the dataset. So we reduce the number of independent variables to 25 and again proceed towards building the model and obtain the RMSE and r-squared for each of them.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Error Metric | DT | RF | LR | SVR | XGBoost |
| RMSE(TRAIN) | 3.613 | 3.337 | 3.319 | 4.302 | 3.714 |
| RMSE(Test) | 4.066 | 3.872 | 3.854 | 4.932 | 4.08 |
| R2 | 0.252 | 0.322 | 0.395 | -0.099 | 0.2479 |

From the table above it is clear that of the given algorithms linear regression has least RMSE and maximum r-squared of about 0.395. Generally we prefer r-squared of about 0.9 but here the data is also limited to 740 records, it is quite possible that if we had a larger dataset the accuracy of the model would have been better and r-squared would have been higher.

## 4.3 Python

Here we have plotted the correlation plot first for the continuous variables then removed those variables which have correlation coefficient greater than 0.8 with any of the other variables. In the same way we have removed those categorical variables who have got p-values greater than 0.05.It has been found that Weight has correlation coefficient greater than 0.8 i.e. 0.88 with Body mass Index. Month of absence has got p-value greater than 0.05. Thus in the final list for independent variables all these two variables have been dropped. Since regression is numerical analysis so we need numeric type data rather than factor type data. So we have added dummies for each categorical variable. After adding dummies one dummy variable from category has been dropped because the dummies of a categorical variable are linearly dependent and add up to one. So one dummy variable is dropped from each category to avoid any multi collinearity.

After this we split the entire dataset into two parts training and testing set in 80:20 ratio i.e. 80 percent is training dataset and 20 percent is testing set using random sampling as this is a regression problem.

Thereafter different algorithms viz Decision tree, Random forest, Linear regression, XGBoost and support vector machine, have been used to create machine learning model and evaluate it’ s error metrics and r-squared. R-squared is basically calculation of goodness of fit of the model i.e. how much the developed model fits the dataset.

We have first developed model using the above mentioned algorithms and calculated RMSE (root mean squared error) for each of the model on testing and training set. Also r-squared has been calculated for each of them.

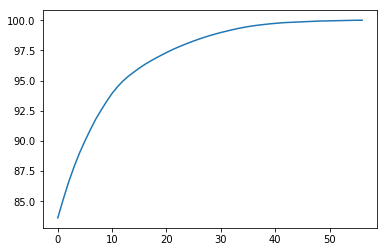
In linear regression model, after developing the model we found that very few of the variables appear to be significant, so we kept them and removed all other variables and again performed linear regression. It was found that the RMSE of the model further decreased along with r-squared. So we decided to stick to the previous training and testing set.

Correlation Plot

The table below shows the RMSE for each of the machine learning models along with R-squared value.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Error Metric | DT | RF | LR | SVR | XGBoost |
| RMSE(TRAIN) | 153.77 | 31.99 | 133.933 | 13.49 | 6.895 |
| RMSE(Test) | 151.305 | 159.31 | 147.801 | 12.71 | 13.058 |
| R2 | 0.057 | 0.007 | 0.0796 | -0.0063 | -0.061 |

After obtaining the above result we have used principal component analysis. Here we have first plotted variance graph of every variable.



From the graph above it is clear that out of the given variables around 20 of them explain more than 95% of variance in the dataset. So we reduce the number of independent variables to 20 and again proceed towards building the model and obtain the RMSE and r-squared for each of them.

From the table below it is clear that of the given algorithms linear regression has least RMSE for testing set and maximum r-squared of about 0.1127. Generally we prefer r-squared of about 0.9 but here the data is also limited to 740 records, it is quite possible that if we had a larger dataset the accuracy of the model would have been better and r-squared would have been higher.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Error Metric | DT | RF | LR | SVR | XGBoost |
| RMSE(TRAIN) | 11.29 | 4.94 | 11.67 | 12.21 | 4.69 |
| RMSE(Test) | 17.18 | 15.81 | 15.11 | 15.89 | 16.91 |
| R Squared | -0.146 | 0.028 | 0.1127 | 0.0186 | -0.111 |

## Chapter 5

## Conclusion

## 5.1 Model Evaluation

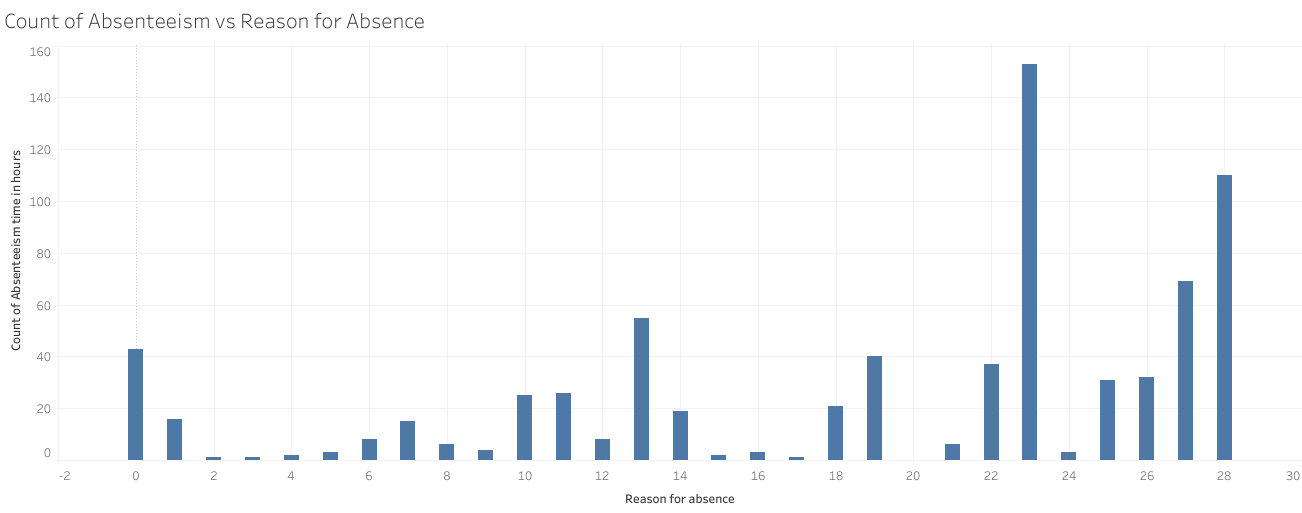
From comparison of error metrics of above machine learning models in R and Python it can be concluded that in R and Python both the linear regression model is the best as it has got low RMSE and high r-squared among all the given models. Generally we prefer r-squared of about 0.9 but here the data is also limited to 740 records, it is quite possible that if we had a larger dataset the accuracy of the model would have been better.

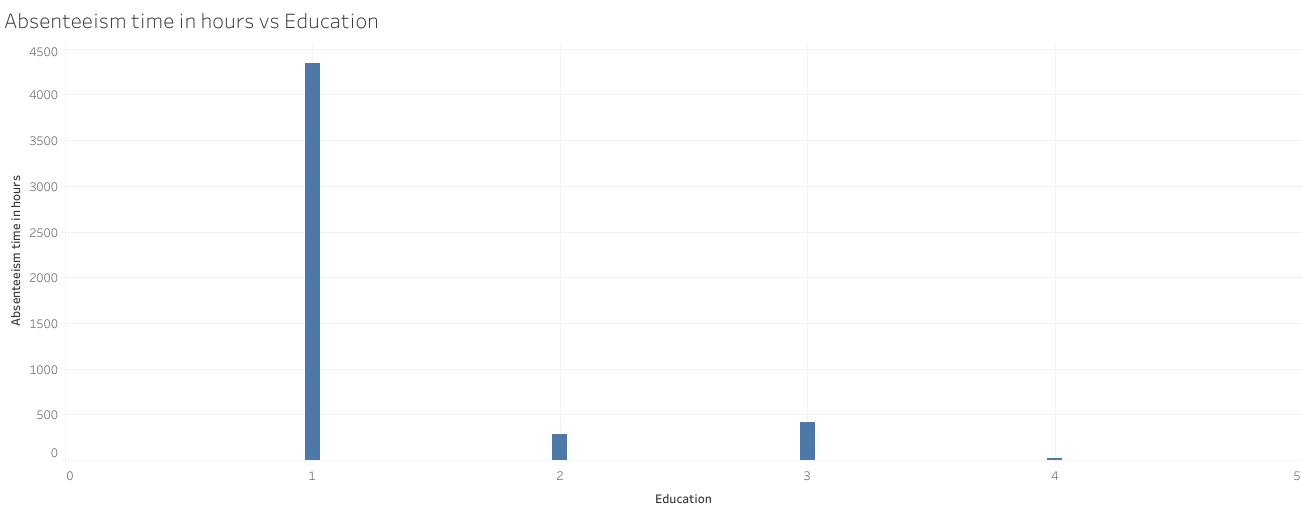
Lower value of RMSE and higher value of R-squared reveal that the model fits the data very well.

## 5.2 Solutions of Problem Statement

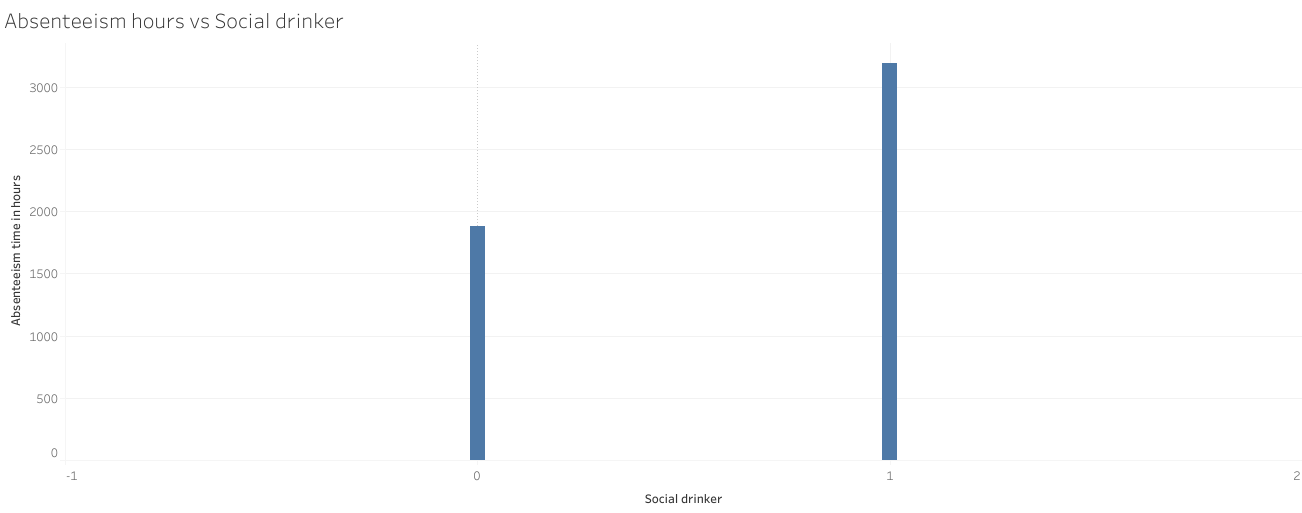
1. What changes company should bring to reduce the number of absenteeism?

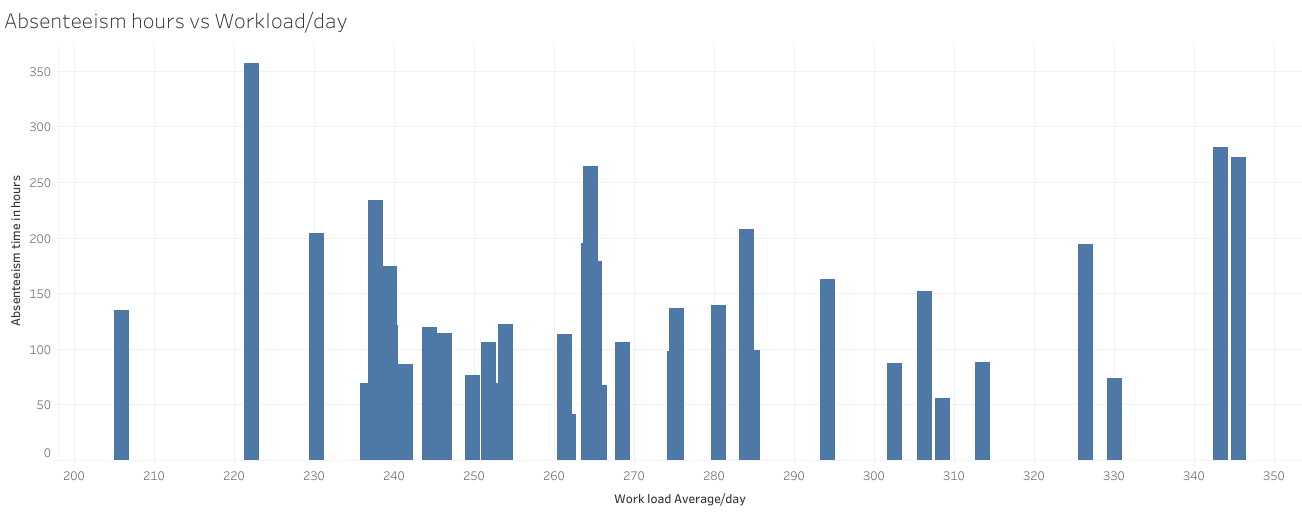
Solution:

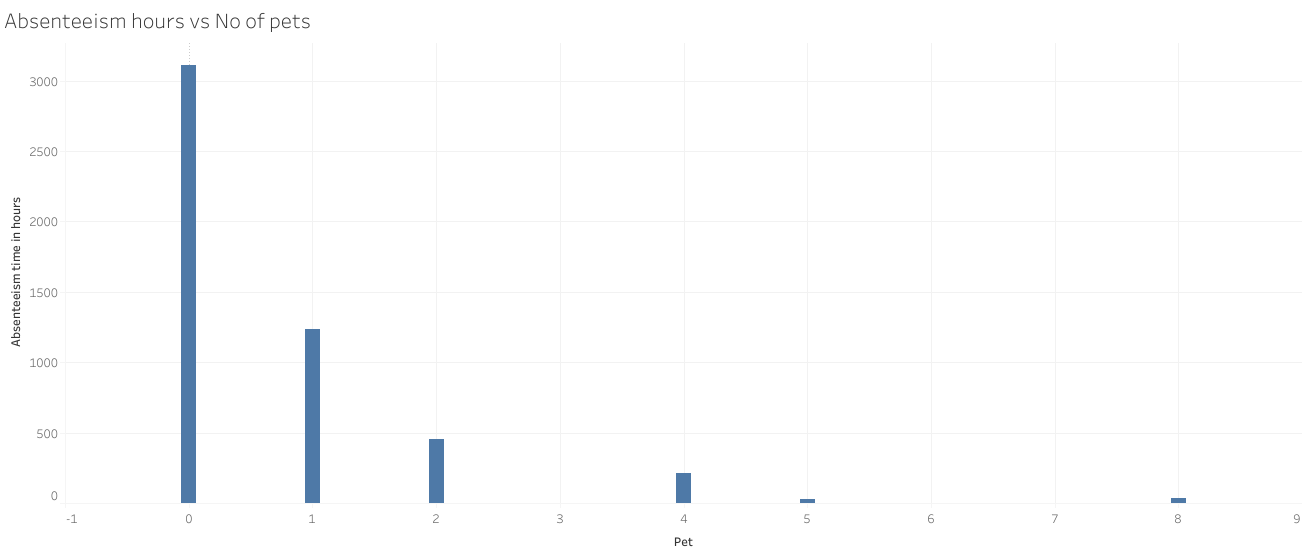
1. From the graph it is clear that the major reason given by employees for absence is reason 23 which is for medical consultation. This issue can be overcome if the XYZ company provides free medical consultation for its employees at regular intervals.
2. The graph shows that majority of the employees belong to category 1 i.e. they have completed high school education who make the most absent. The company can improve upon this by hiring employees who are at least graduate.

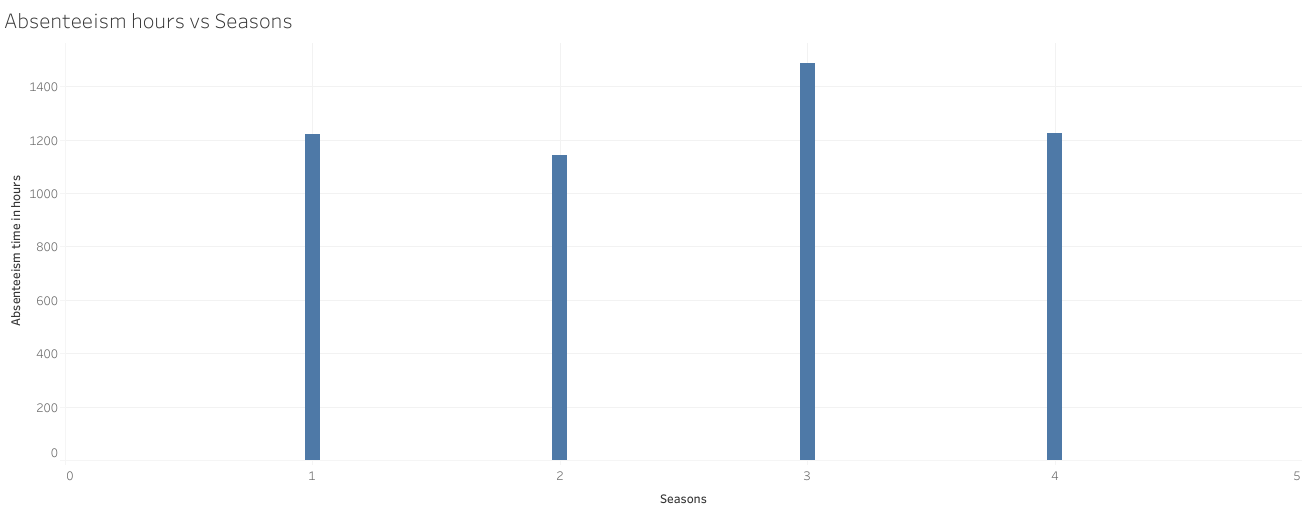


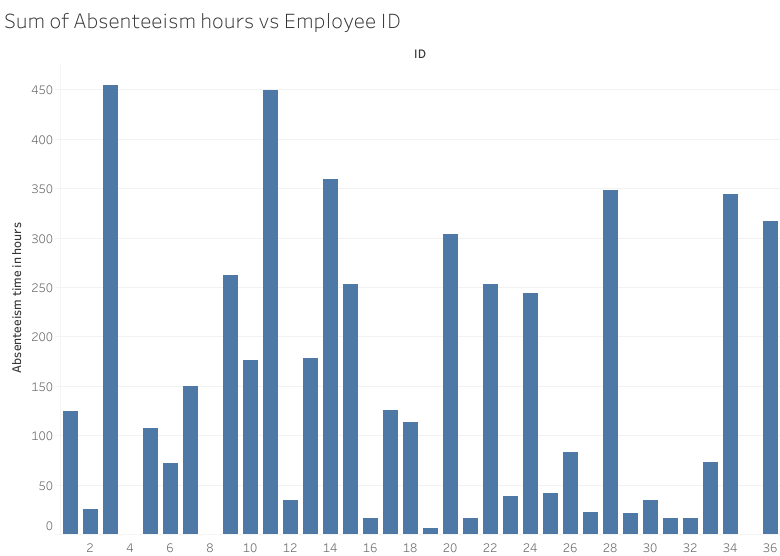
1. From the graph it is clear that majority of the employees making absent are social drinkers. Company can pay heed to this information by hiring employees who do not drink. It will help them reduce absenteeism drastically.



1. The graph below shows absenteeism hours versus workload on employees per day. It is quite evident from the graph that there is a peak in absenteeism hours of employees who have low workload around 220 as compared to other employees. Also there are high bars for employees who have very high workload. Thus company should ensure that workload is equally distributed among the employees irrespective of their education level.
2. The graph below displays how the sum of absenteeism hours of employees varies on no of pets an employee has. From the graph it is clear that employees who have got no pets make the most absent. This factor may be quite useful for the company when hiring a new employee and the company can also teach the importance of pets to it’s employees it might help in reducing the absenteeism hours.

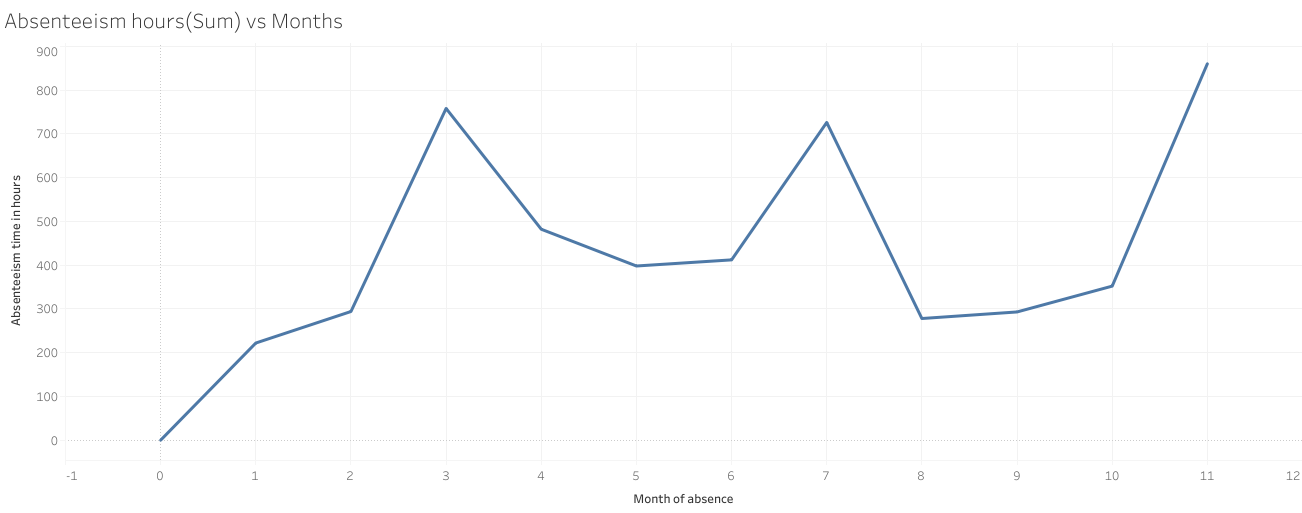


1. The graph below shows that that the most absenteeism hours are in the season 3 which is winter. This may be due to health issue may be the company can schedule free annual health checkup of employees depending upon their attendance it will work two way reduce absenteeism hours and also provide medical consultation to its employees which is major reason for absence.
2. From the graph of employee ID and absenteeism hours it is clear that the employees having ID 3,11,28,33 make the most absent. The company should talk to each of them and figure out if there is something or else warn them for a strict action against them in future.



1. How much losses should the company predict in every month of 2011 if the same trend of absenteeism continues?

Solution:- Assuming the losses here to be in terms of absenteeism hours because no such data regarding revenue loss of the company has been provided.



The graph above shows absenteeism hours of employees versus months. It is quite evident from the graph that in month three i.e. April (as month index start from 0 so 3 is the 4th index), month 7 i.e. August and month 11 i.e. December the absenteeism hours are extremely high as compared to other months in the year. To reduce it the company may organize different functions to engage the employees during this period.