**Employee Absenteeism Report**

**Prajwal R**

**01-12-2018**

**Contents**

**1. Introduction**

1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 Exploratory Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

**2. Methodology**

2.1 Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

2.1.1 Missing Value Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .6

2.1.2 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .7

2.1.3 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

2.1.4 Feature Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .8

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2.1 Decision Tree. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2.2 Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10

2.2.3 Linear Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.2.4 KNN. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

**3. Conclusion**

3.1 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

3.2 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

3.3 Answers. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**Appendix**

Extra Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

**References**

**Chapter 1**

**Introduction**

* 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 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**

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)

**1.3 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. Missing valuesare present in this dataset.

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)

From EDA we came to know that apart from Target variable, there are 10 continuous variables and 10 categorical variables.

**Continuous variables:**

ID,Transportation expense,Distance from Residence to Work,Service time,Age,Work load Average/day,Hit target, Weight,Height,Body mass index

**Categorical variables:**

Reason for absence,Month of absence,Day of the week,Seasons,Disciplinary failure, Education,Son,Social drinker, Social smoker,Pet

**Target variable:**

Absenteeism time in hours

**Chapter 2**

**Methodology**

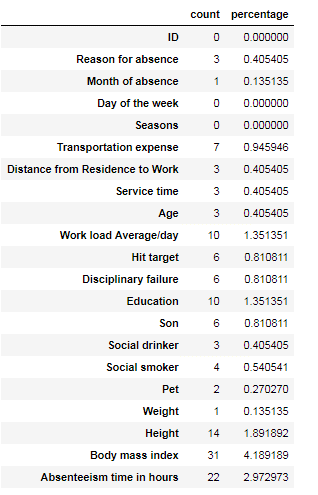
Before feeding the data to the model we need to clean the data and convert it to a proper format. It is the most crucial part of data science project. Few pre-processing techniques are appliedon the data set to bring it to proper shape.

**2.1 Pre-Processing**

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

**2.2.1 Missing Value Analysis**

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. If acolumns has more than 30% of missing values, either we ignore the entire column or we ignore those observations. In the given data the maximum percentage of missing value is 4.189% for **body mass index** column. So we will compute missing value for all the columns.

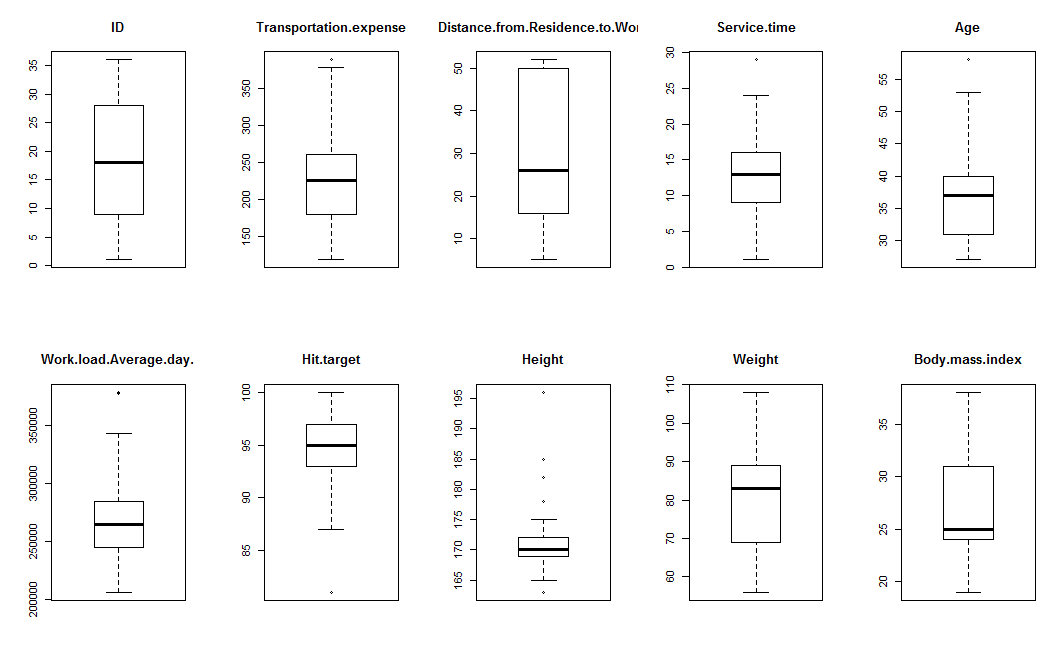


**In this project we have used KNN imputation method to impute missing value**.

**2.1.2 Outlier Analysis**

One of the pre-processing step after imputing the missing values is checking for the presence of outliers. In this case we use a classic approach of removing outliers. We visualize the outliers using boxplots.

Boxplots are plotted for each of the continuous variables and based on the insights and proper understanding about the data, outliers are removed from the variables.

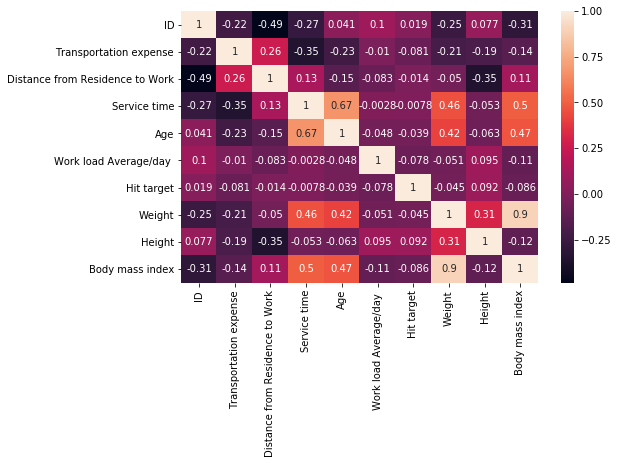


From the boxplot almost all the variables except **“Distance from residence to work”, “Weight” and “Body mass index”** consists of outliers.

Here we have converted the outliers from the variable Work load Average/day to NA. Then missing values are imputed using**KNN** imputation method.

**2.1.3 Feature Selection**

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem. Selecting subset of relevant columns for the model construction is known as Feature Selection. We cannot use all the features because some features may be carrying the same information or irrelevant information which can increase overhead. To reduce overhead we adopt feature selection technique to extract meaningful features out of data. This in turn helps us to avoid the problem of multi collinearity. In this project we have selected **Correlation** Analysis for numerical variable and **Kruskal Wallis test** for categorical variables since target variable is continuous and also values in target variable are not uniformly distributed.



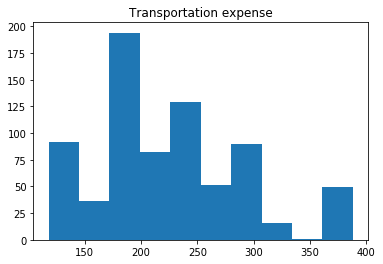
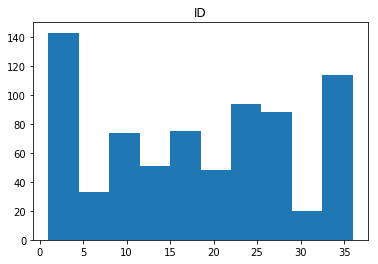
From correlation analysis we have found that Weight and Body mass index has high correlation of 0.9. So we have excluded the Weight column for modeling.

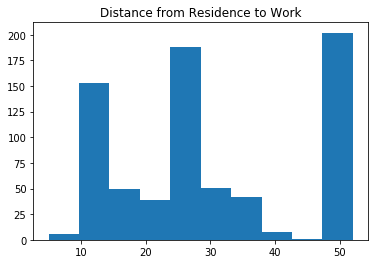
**2.2.4 Feature Scaling**

Feature scaling is a method used to bring the data to a suitable range of values. Normalization and Standardization are the two major techniques used to scale the data. 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.

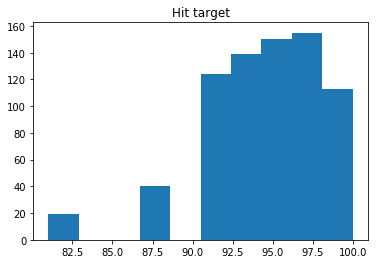
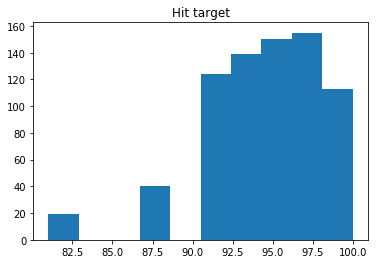
We have plotted the histogram to check the distribution of data across all the continuous variables.

Since most of the variables are not uniformly distributed, data is fed to the model as is.











After pre-processing the data set contains 19 independent variables and 1 dependent variables without any missing values and unwanted outliers. Will freeze this dataset and apply different regression models.

**2.2 Modeling**

After a thorough preprocessing we will be using some regression models on our processed data to predict the target variable. Data set is divided into train and test. Models are trained using train dataset and evaluated using test dataset.

**2.2.1 Decision Tree**

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 utility. Each branch connects 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. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users.

Decision tree model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Mean absolute error** is calculated.

|  |  |  |
| --- | --- | --- |
| Decision Tree | R | PYTHON |
| MAE | 4.27 | 4.75 |

**2.2.2 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.

Random Forest model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Mean absolute error** is calculated.

|  |  |  |
| --- | --- | --- |
| Random Forest | R | PYTHON |
| MAE | 4.45 | 4.70 |

**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.

Linear regression model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Mean absolute error** is calculated.

|  |  |  |
| --- | --- | --- |
| Linear Regression | R | PYTHON |
| MAE | 6.319 | 5.97 |

**2.2.4 KNN**

KNN is simple algorithm that stores all available cases and classifies new cases based on a similarity measure. KNN measures the distance between given data point and all other data points. Then based on the specified k value, it selects the closest neighbors to that data point. In case of regression, it will take the average of neighbor values and assigns that value.

KNN model is applied on the train dataset and the values are predicted for the test data using predict function. Predicted values and actual values of test data are compared and **Mean absolute error** and is calculated.

|  |  |  |
| --- | --- | --- |
| KNN | R | PYTHON |
| MAE | 6.34 | 5.79 |

**Chapter 3**

**Conclusion**

In this chapter we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

**3.1 Model Evaluation**

In the previous chapter we have calculated the Mean Absolute Error (MAE) and Mean for different models. Mean Absolute Error (MAE) is the average of absolute errors or deviations. Lower values of MAE, higher the accuracy of model.

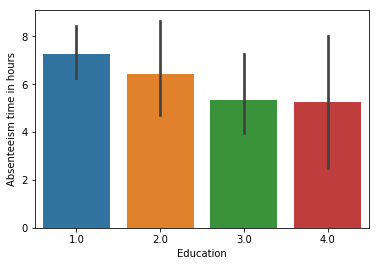
**3.2 Model Selection**

After observing the MAE values of all the models, we can conclude that Random forest has lower values of MAE. We can freeze Random forest model for this data set.

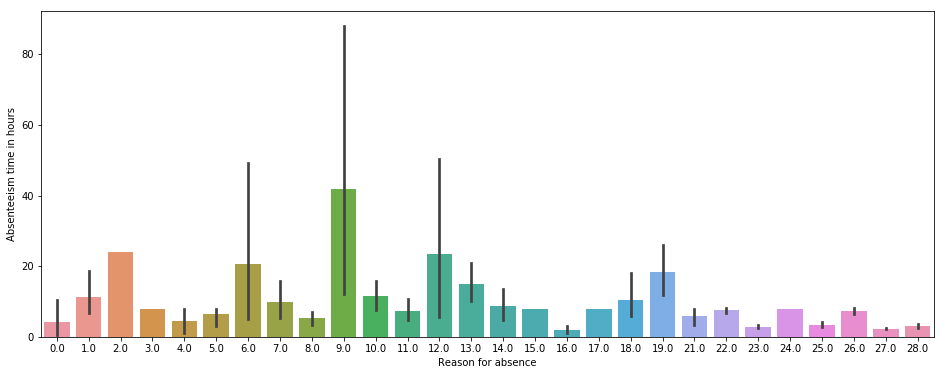
**3.2 Answers**

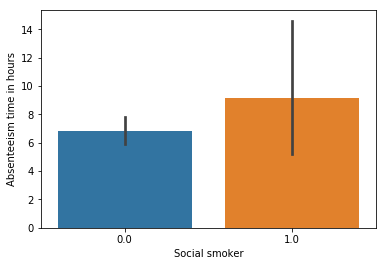
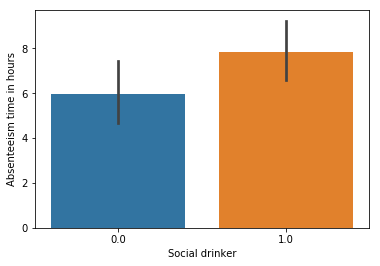
The Changes which company should bring to reduce the number of absenteeism are

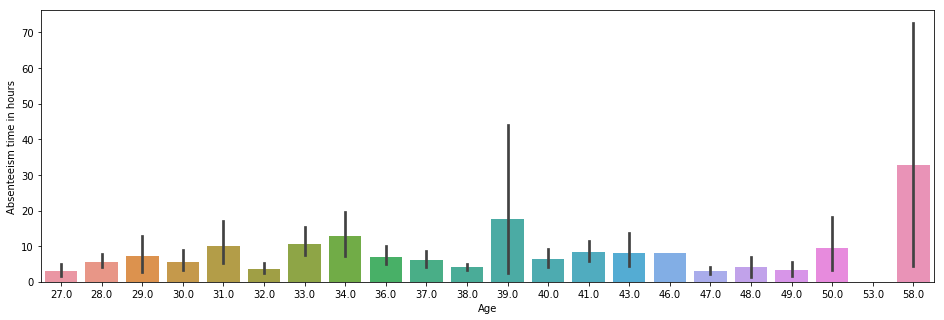
1. It is observed that employees with low education have more number of absenteeism time in hours.



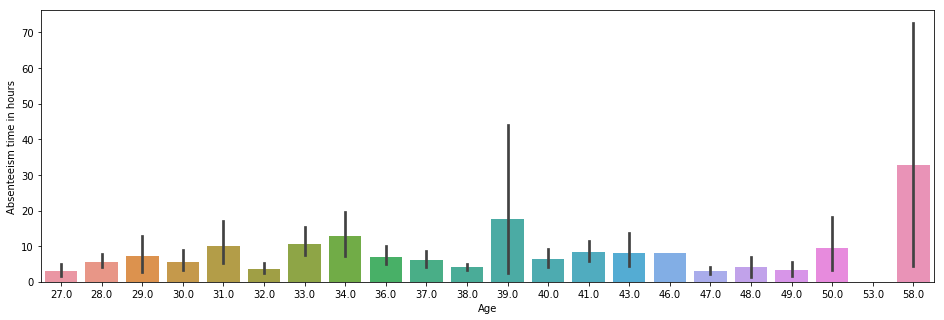
1. Employees are often giving a medical reason of Diseases of the circulatory system to take leaves.



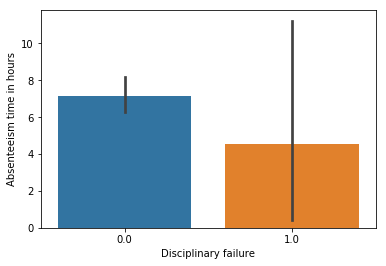
1. Employees who are social smokers and social drinkers have more absenteeism time.
2. Aged employees have more absenteeism time.



1. Employees who stays close to office have more absenteeism time.



1. Employees having disciplinary failure gave more absenteeism time.

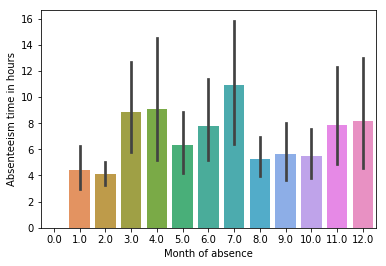


**Appendix**

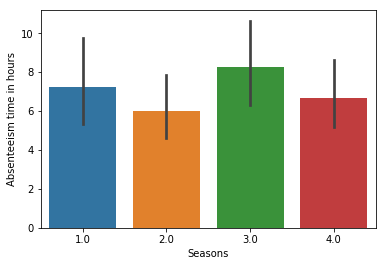
**Extra Figures**

Relationship of our target variable (Absentee time in hour) with other variables.

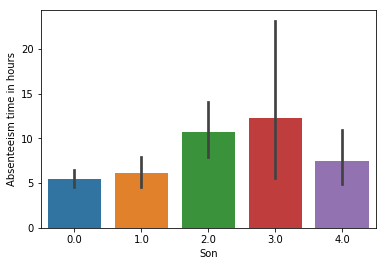
1. Graph of 'Absenteeism time in hours' vs 'Month of absence'



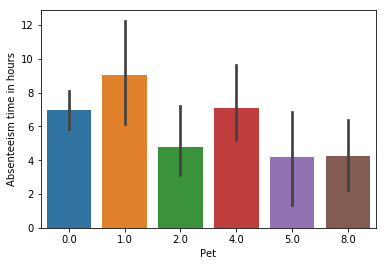
1. Graph of 'Absenteeism time in hours' vs 'Seasons'



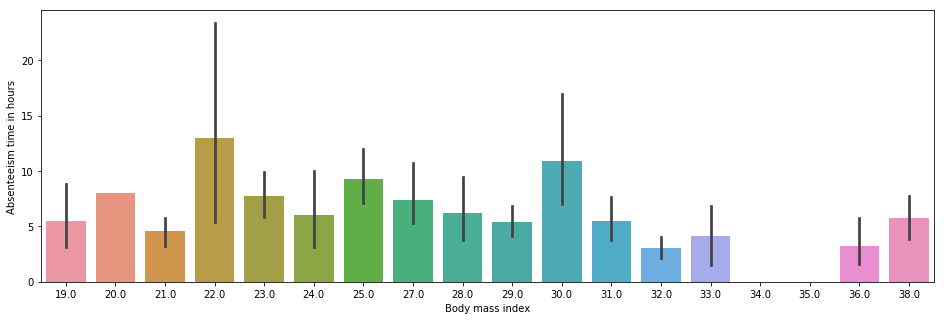
1. Graph of 'Absenteeism time in hours' vs 'Son'



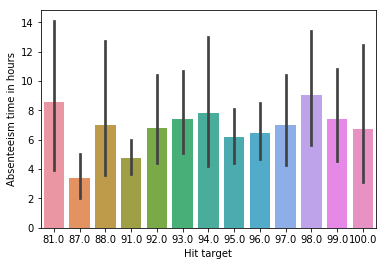
1. Graph of 'Absenteeism time in hours' vs 'Pet'



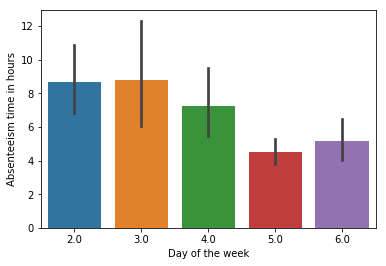
1. Graph of 'Absenteeism time in hours' vs 'Body mass index'



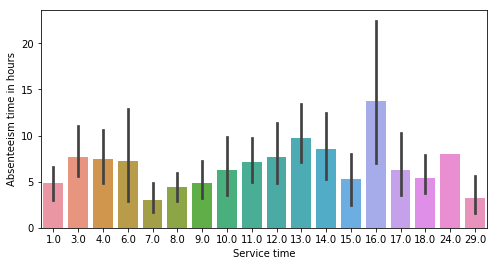
1. Graph of 'Absenteeism time in hours' vs 'Hit target'



1. Graph of 'Absenteeism time in hours' vs 'Day of the week'



1. Graph of 'Absenteeism time in hours' vs 'Service time'



**References**

1. For Data Cleaning and Model Development -

<https://edwisor.com/career-data-scientist>

1. For Kruskal wallis test -

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html>