**Predicting Losses Due to Absenteeism in 2011**

**Preeti Chauhan**

**28 November 2018**

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?

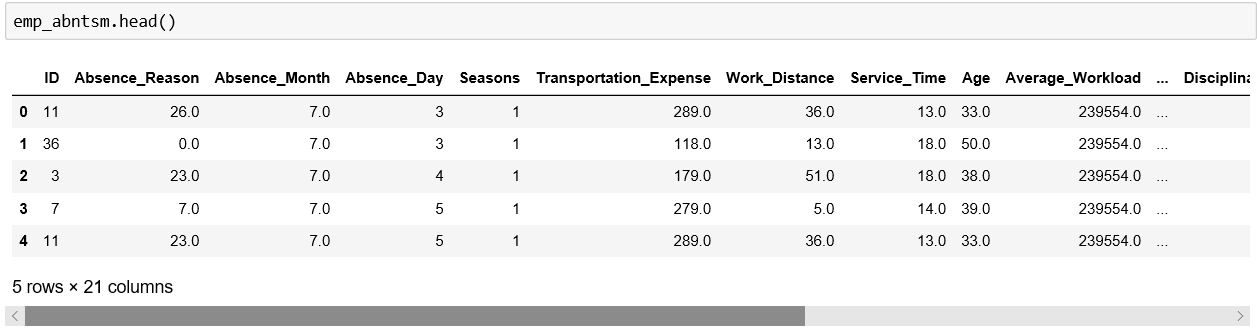
2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

* 1. **Data**

Data provided with the problem is **Absenteeism\_at\_work\_Project.xlsx**.

Let’s have a look at sample of dataset:

Note: I have renamed the columns(removed spaces and shortened) for ease of operations



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column Name** | **Description** | **Expected Values** | **Value Meaning** |  |
| ID | Record index | Unique numeric identifier |  |  |
| Reason for absence | Date | Any valid Date |  |  |
| Month of absence | Month of Absence | 1  2  3  .  .  . | 1: January,  2: Feb,  3: March,  .  .  . |  |
| Day of the week | Day of absence instance | 2  .  .  6 | Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6) |  |
| Seasons | Seasons | 1  2  3  4 | 1:springer, 2:summer, 3:fall, 4:winter |  |
| Transportation expense | Expense of Transportation |  |  |  |
| Distance from Residence to Work |  |  |  |  |
| Service time | Service duration in the company |  |  |  |
| Age | Age of employee |  |  |  |
| Work load Average/day |  |  |  |  |
| Hit target |  |  |  |  |
| Disciplinary failure |  |  |  |  |
| Education |  |  |  |  |
| Son |  |  |  |  |
| Social drinker |  |  |  |  |
| Social smoker |  |  |  |  |
| Pet |  |  |  |  |
| Weight |  |  |  |  |
| Height |  |  |  |  |
| Body mass index |  |  |  |  |
| Absenteeism time in hours (target) |  |  |  |  |

Here Target variable(dependent variable), which needs to be predicted is: **‘**Absenteeism time in hours’

And others are the predictors(independent variables) on basis of which target needs to be predicted are.

Let’s explore further to use this data to build a predictive model.

1. PROJECT APPROACH

To execute this project, I’ll go along with tried and tested CRISP- DM process

* 1. **CRISP-DM process**

CRISP – DM process

2

Data understanding

1

Business Understanding

Data preparation

Data

3

Deployment

6

Modelling

4

Evaluation

5

Above process, shows phases of data science projects, each phase having its importance.

* 1. **CRISP - DM In a nutshell**
* **Get a decent understanding of business domain.** Its very important to relate to problem before you figure out the solution . e.g. , if you want to build a model for predicting stock prices, you should have a basic understanding of stock market.
* **Get a precise Business Objective.** Sometimes you get directly from client, else you have to derive and get it verified.
* **Get the Data to analyze** and find stats, patterns, existing trends. This will help you toward a better judgement.
* **Categorize the problem statement.** We need to put the problem in a problem category(all the problem categories are described in answer to Question 3). \this is very important as specific machine learning algorithm is used for specific category.

**Eg.** Unsupervised learning problem will use any of the unsupervised learning algorithm, which may not be suitable for a prediction problem.

* **Know the prerequisites of the selected data models** and see if the data given to you is meeting that requirement. eg. If for a banking problem, decision tree machine learning model is selected to address a problem statement, then we should know that Decision tree does not allow missing values as input.
* **You have to prepare the data as per selected model.** eg. If you have selected decision tree then, you know that it doesn’t take missing values, then you need to deal with missing value before feeding it to the model.
* **Build the Model(s)** using the machine learning algo(s) you have selected.
* **Compare and Evaluate the models**: Compare various models on their performance matrices and choose the best one fitting your requirements.
* Deploy the selected Model to prod.

1. DETAILED PROJECT IMPLEMENTATION

Below are the different phases of the implementation of project

* 1. **Define the Project ROADMAP**

I have sketched the plan to implement the project in different phases, using CRISP-DM process. Some steps in CDM process have been left out which are out of the scope of this project report, like deployment.

The whole project is divided in 7 phases (and further subphases). Below are the phases defined.

* Define and categorize problem statement
* Gather the data
* Prepare data for consumption
* Perform Exploratory Data Analysis
* Modelling
* Evaluate and compare Model performances and choose the best model
* Hypertune the selected model
* Produce sample output with tuned model
  1. **Implement the Project ROADMAP**

As per the above roadmap, let’s start the project, exploring each phase.

* + 1. **Categorize Problem**

The problem statement is “to predict the monthly losses due to employee absenteeism”

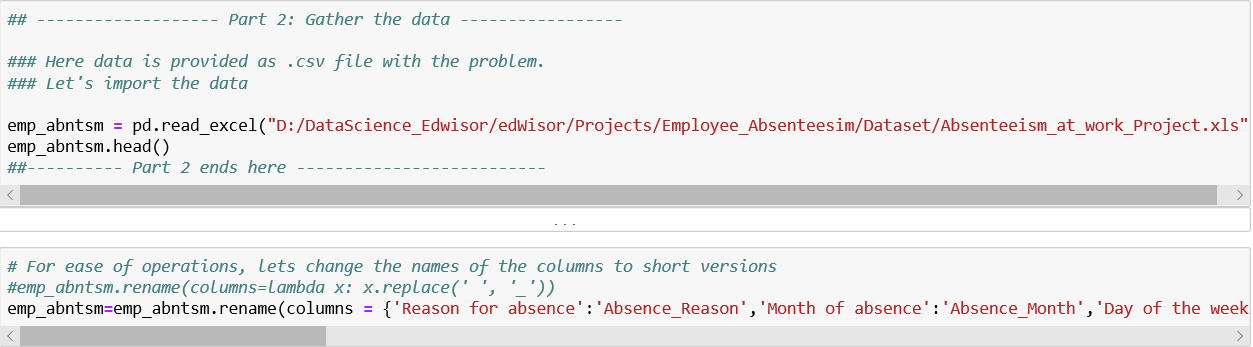
It is evident from the problem statement above that based on the predictors values (input, both numerical and categorical), the output dependent value(numerical) needs to be predicted.

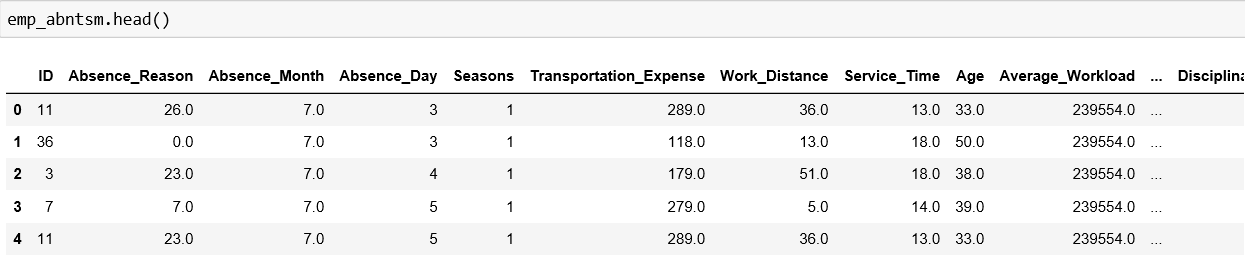
So, clearly this problem is of category – **Supervised Machine Learning Regression Problem.**

* + 1. **Gather the Data**

The data is given to us on platter. Simply import the data and have a glance.

Note: Data is well described in section 1.2 . Use this section to understand the data better.



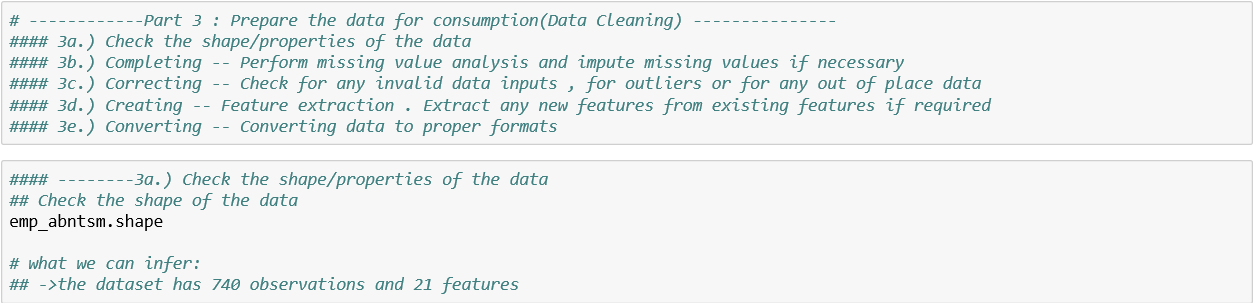


* + 1. **Prepare Data**

Next step is to prepare the data for consumption. In this case, the compiled data set is given to us in 1 file. We do not need to join different sources to prepare data for the analysis.

What we need to do here is more of data cleaning activity and make it ready for EDA and modelling. I performed following steps to achieve this.

* + - 1. **Check the shape/properties of the data**
    - There are 16 features and 731 observations in the dataset
    - Int64(11) ,float64(4) and object(1) datatypes are used in this dataset.
    - None of the columns in dataset has nulls

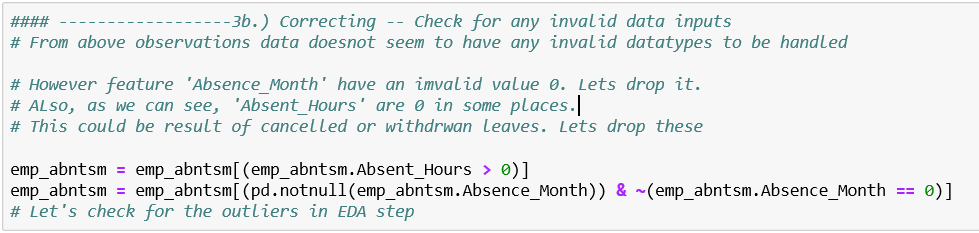




* + - 1. **Correcting**

Perform Check for any invalid data inputs , for outliers or for any out of place data.

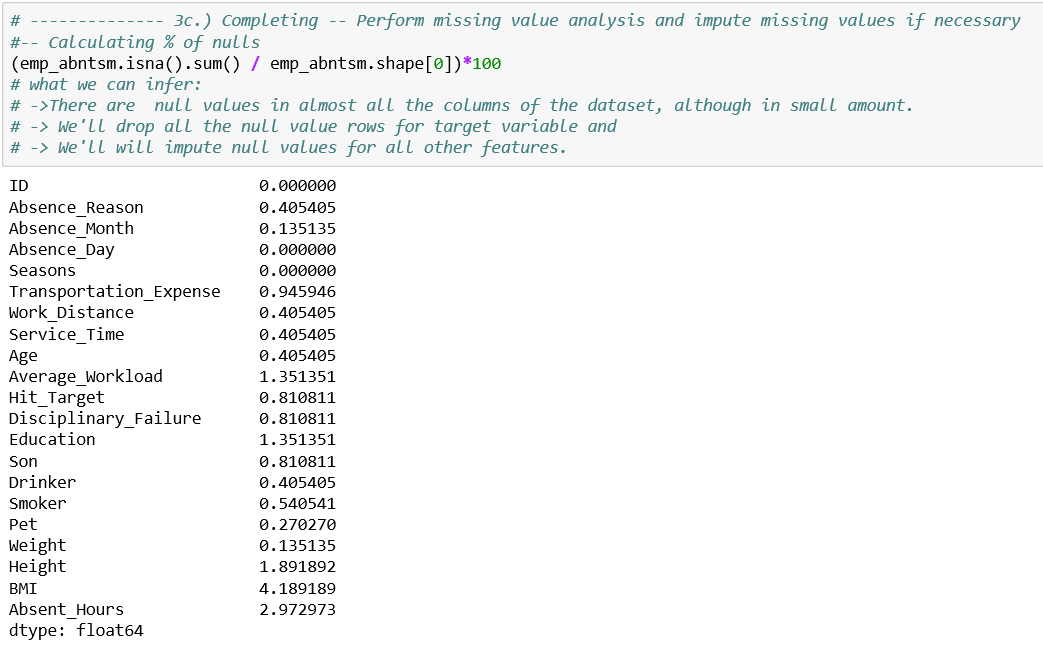
* + - There are few observations where ‘Absent\_Hours’ is 0.
    - This could be the result of cancelled or withdrawn leaves. Dropping such observations
    - Drop observations where month is invalid.

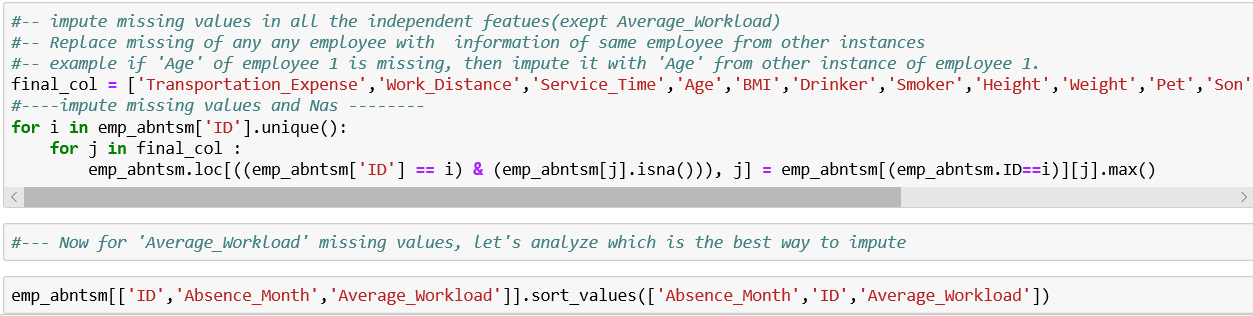
****

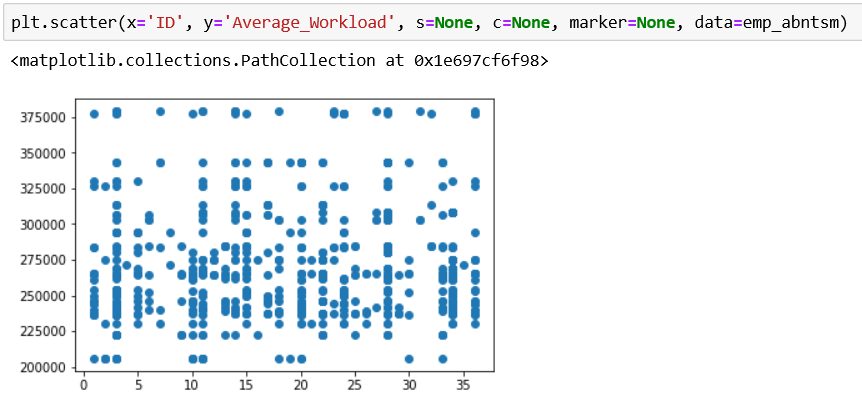
* + - 1. **Completing**

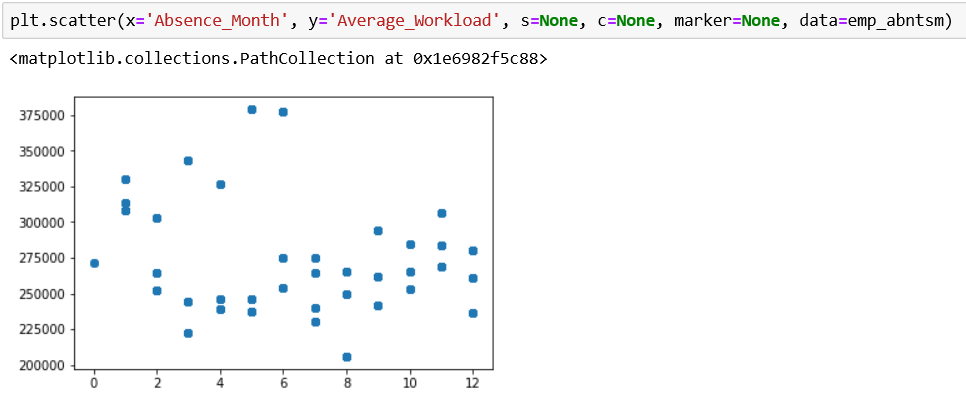
Perform missing value analysis and impute missing values if necessary

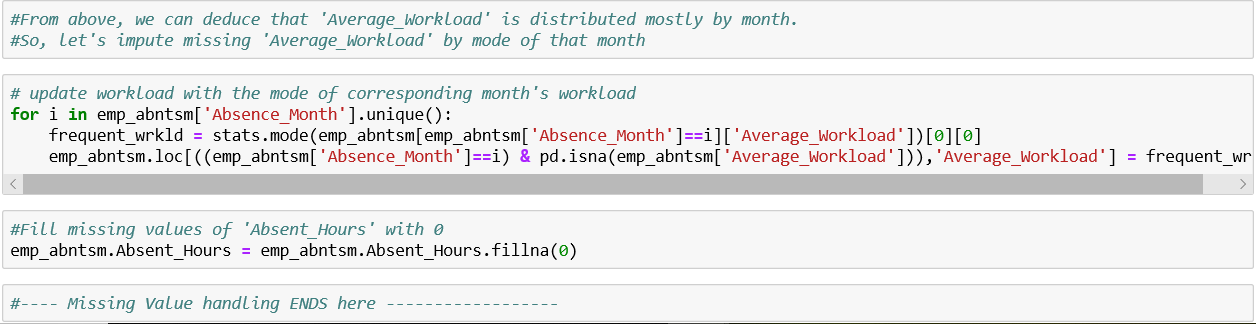
* + - There are many null values in the dataset
    - These needs to be dropped or imputed
    - Lets’ impute missing values of all the features(except ‘Average\_Workload’ and target feature ‘Absent\_Hours’ with the any value of that feature for same employee from other instance
    - E.g. if ‘Age’ is missing for employee 1, then impute it with ‘Age’ of employee 1 which is there in another observation
    - ‘Average\_Workload’ as can be seen is mostly distributed over the month. So impute missing value of ‘Average\_Workload’ with the mode(most frequent) of it for that month.
    - Finally, impute missing value of ‘Absent\_Hours’ with 0.











* + - 1. **Creating**

**Feature extraction**. Extract any new features from existing features if required.

In this case, feature extraction is not required, however the dataset needs to be aggregated over month, since we need to predict absences by month.

* + - 1. **Converting**

Converting data to proper formats.

Columns like ‘Season’, ‘Education’ are imported as numeric columns. However, these are categorical in nature. So, these needs to be converted to ‘categorical’ datatypes.

It is important to convert the categorical values to category as ‘numeric’ and ‘categories’ have different features. E.g.

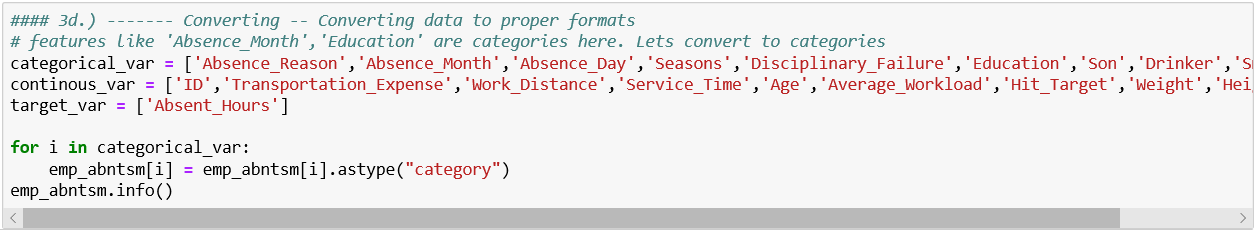
X\_category = [1,2,3,4]

X\_numeric = [1,2,3,4]

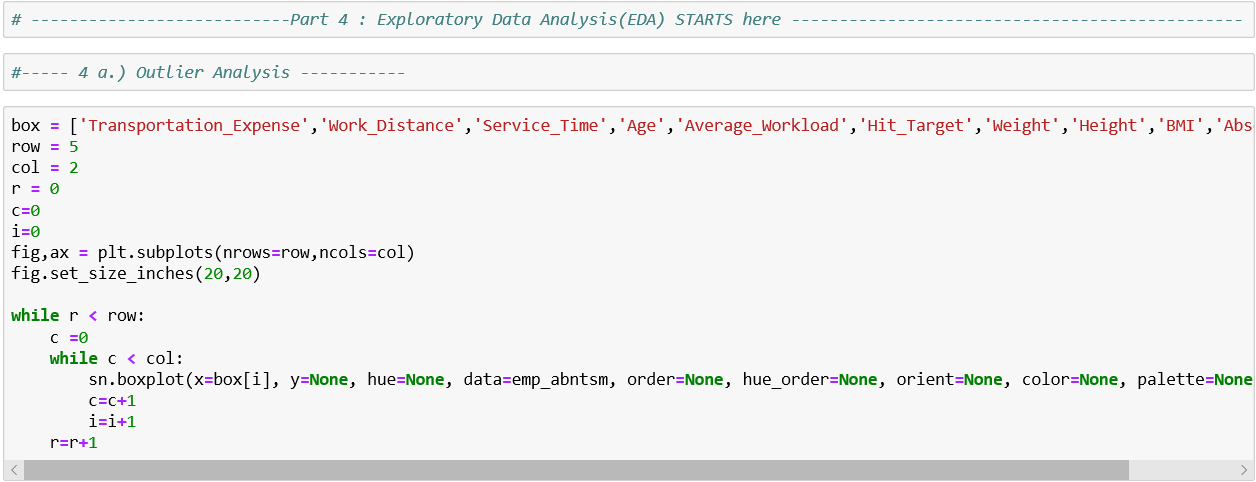
Here X\_category and X\_numeric may seem same but they are not, X\_numeric has order attached to values like 1<2 etc.

But X\_category does not, 1, 2,34 does not have any order, they are simply categories.

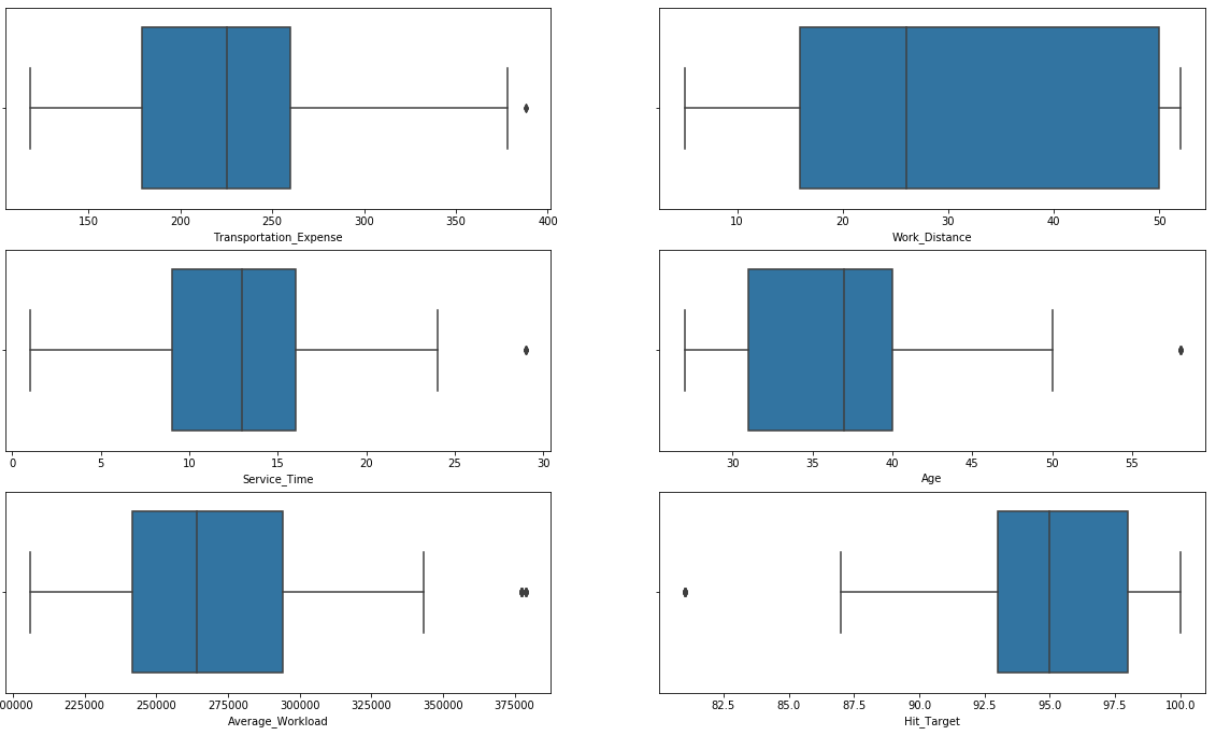
Feeding wrong datatypes to model may affect the results.

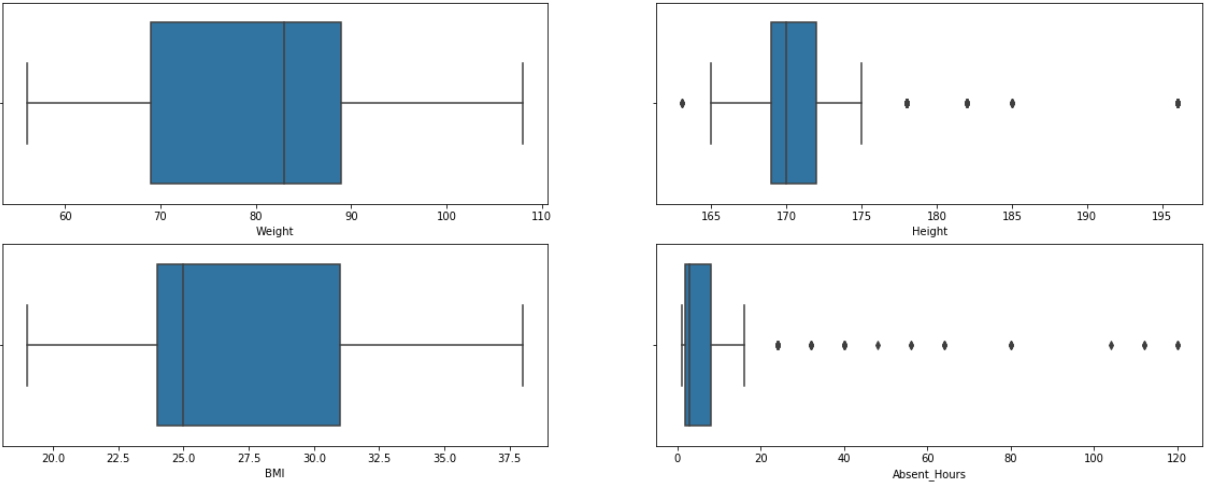


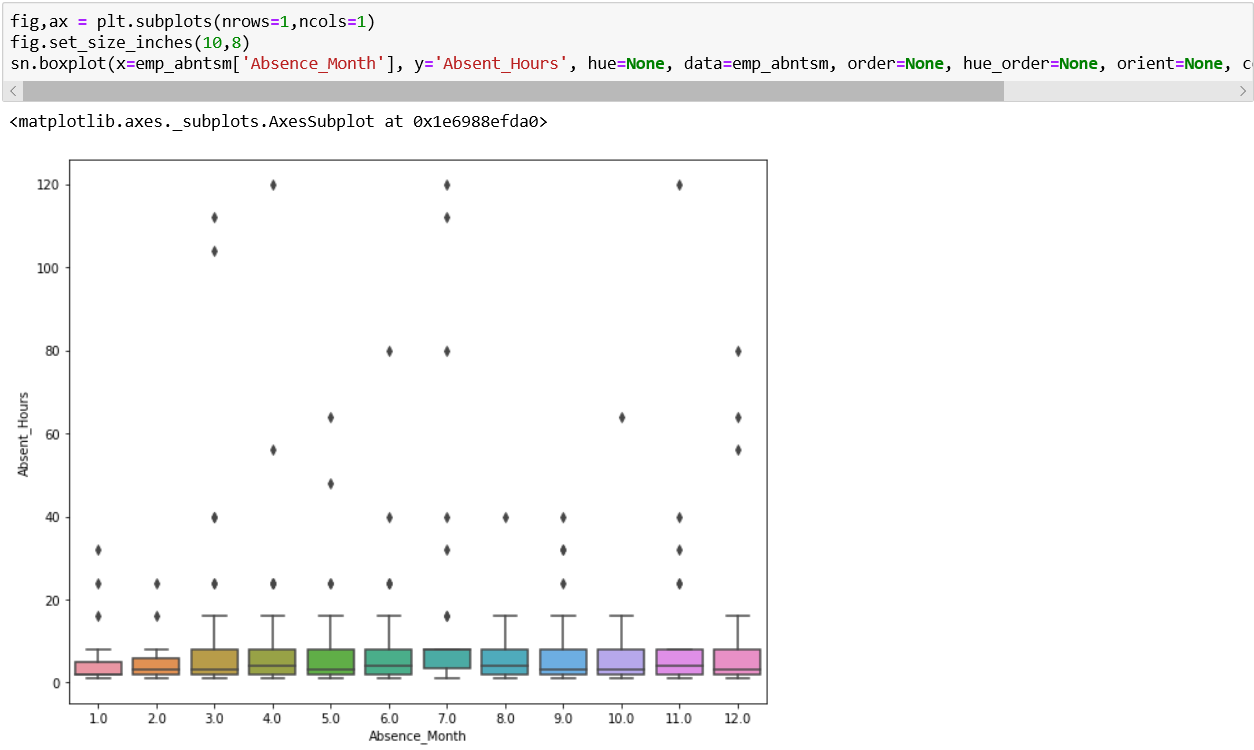
* + 1. **Perform EDA**
       1. **Outlier Analysis using Boxplots**

****

**Boxplot :**

****

****

****

* This can be seen that the target feature ‘Absence\_Month’ has so many outliers
* 'Service\_Time','Age','Average\_Workload' also have outliers .
* These outliers need to be handled, either dropped or imputed with some other value.
* We’ll handle outliers after aggregating the data. Let’s keep it as such for now.
* By removing outliers, we’ll make the model more generalized.
  + - 1. **Analysis of Numerical Features**

1. **Correlation Analysis**

We’ll analyse the

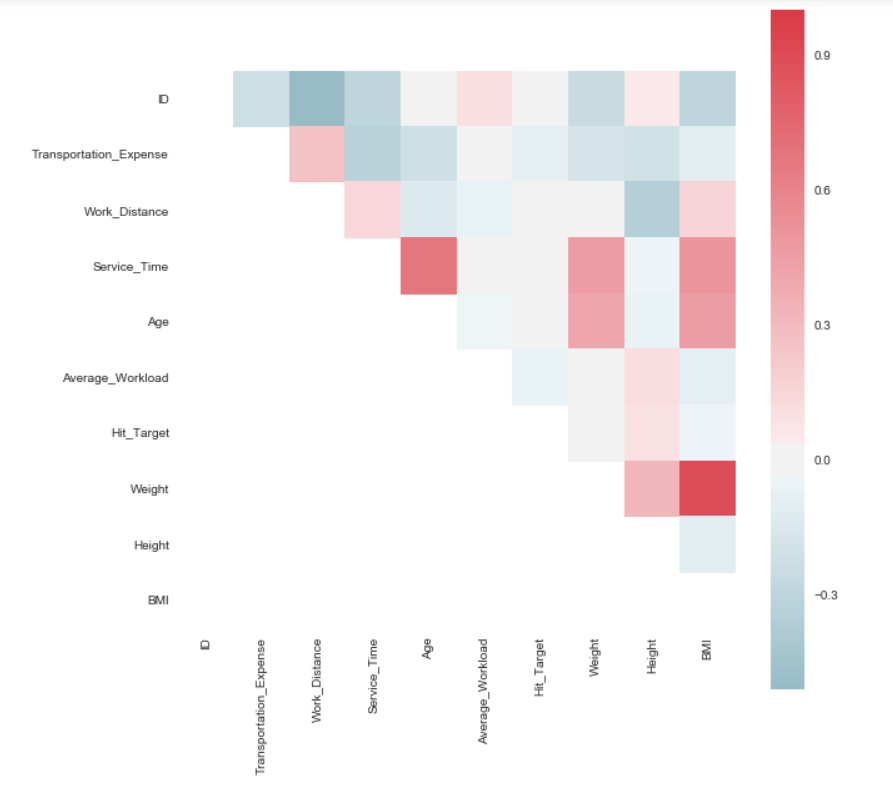
* + - relationship of all numeric independent variables with target variable(‘Absent\_Hours’)
    - relationship of all numeric variable among themselves (to detect multi collinearity)

1. **Barplot for Correlation of features with the response variable**



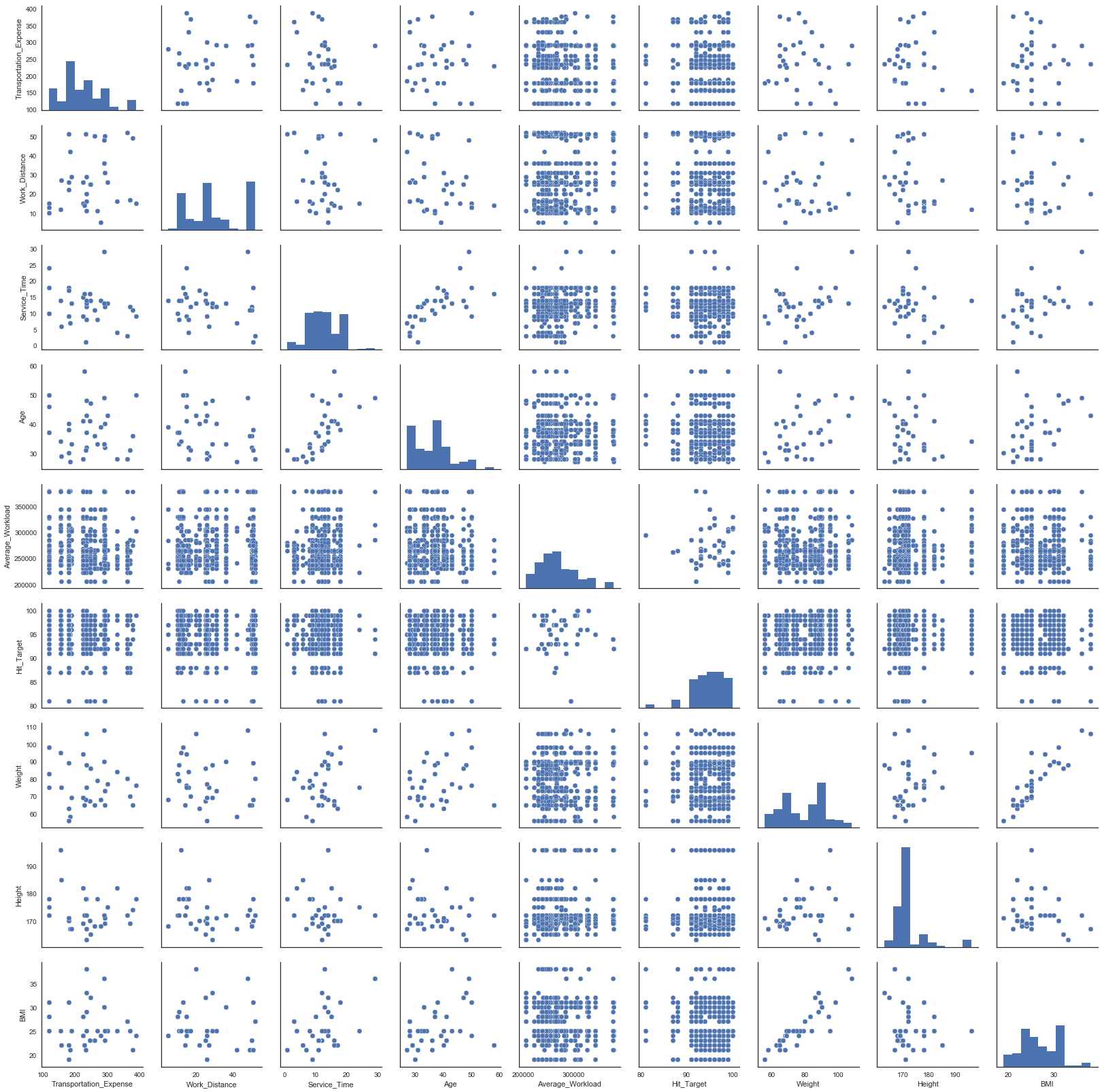
* + - This is evident that ‘Transportation\_Expense’,’Work\_Distance’, ‘Age’, ‘Height’ has good correlation with target feature ‘Absent\_Hours’.
    - Most of the other variables have low correlation with target feature.

1. **Heatmap for Correlation of features with each other**



* + - It is evident that ‘BMI and ‘Weight’ are highly correlated and that **multicollinearity** exists.
    - To remove multicollinearity, drop one either one of ‘BMI’ or ‘Weight’ needs to be dropped.

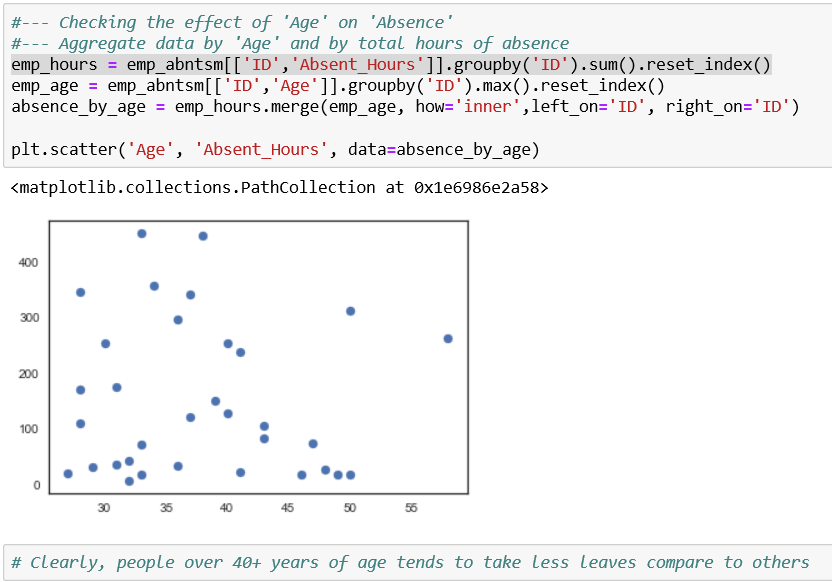
1. **Explore using pairtplots**
   * + See more clear relationships among independent variables using pairplots
     + Most of these simply reiterates, which we have seen in correlation heatmap matrix

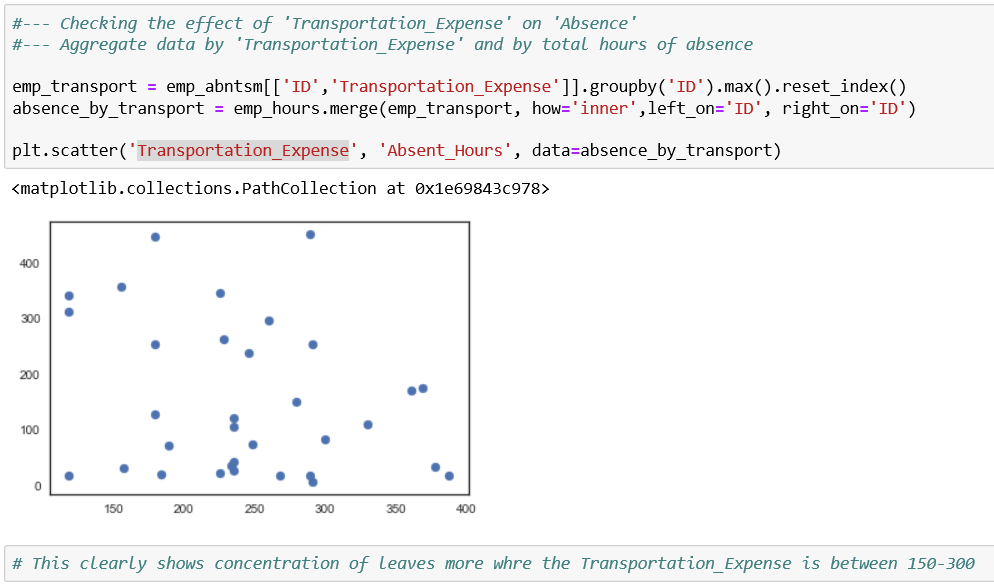


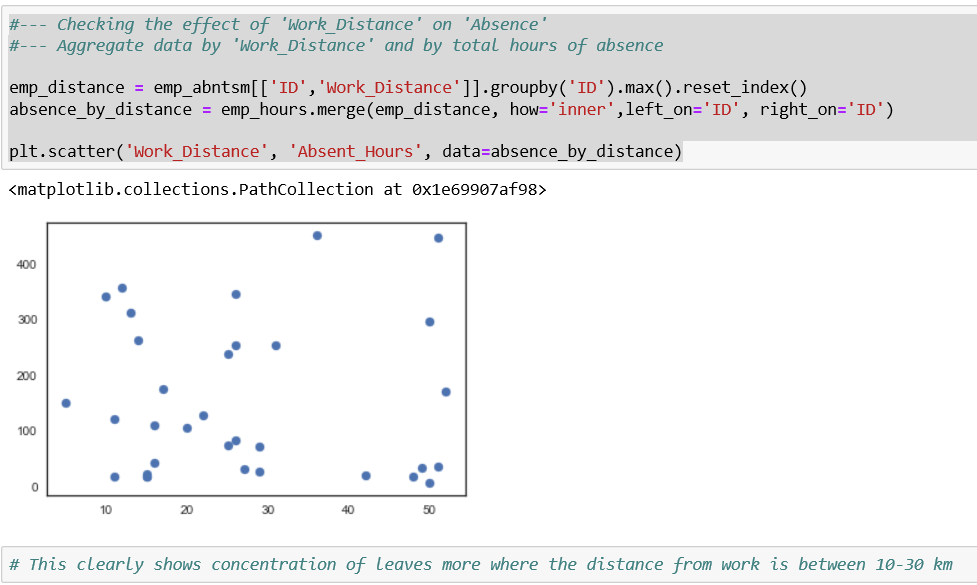
1. **Explore through Jointplots**

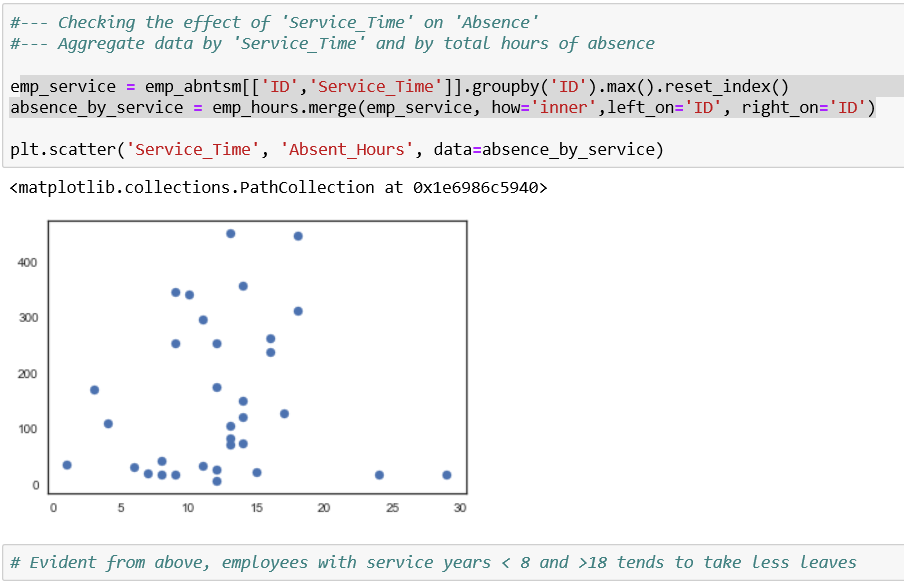


1. **Explore through pointplots**







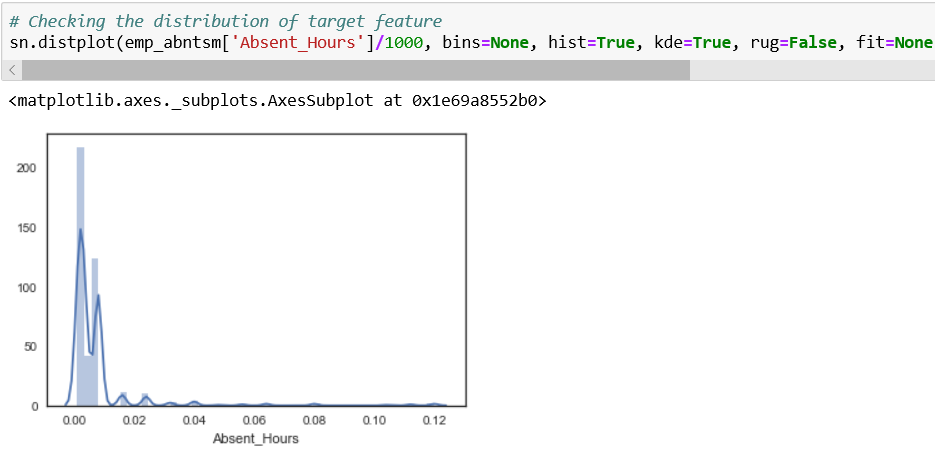


1. **Distribution of target variable through Distplot**

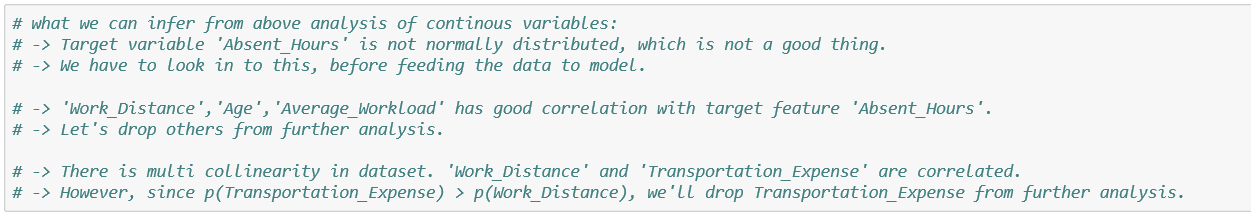
For many models, it is important that the target variable follows normal distribution.

If that is not the case, many a times, we apply some techniques (like taking log) to convert the distribution to normal.

Let explore the distribution of our target variable:



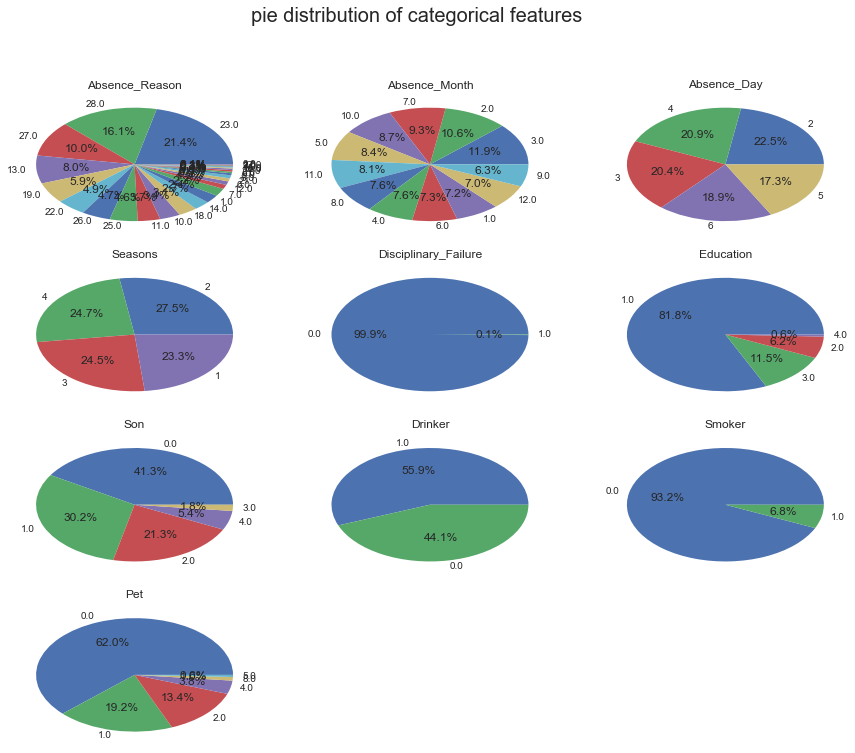
* + - Clearly, target variable does not have normal distribution, so we need to apply some techniques here. Let’s see this after aggregation of data.

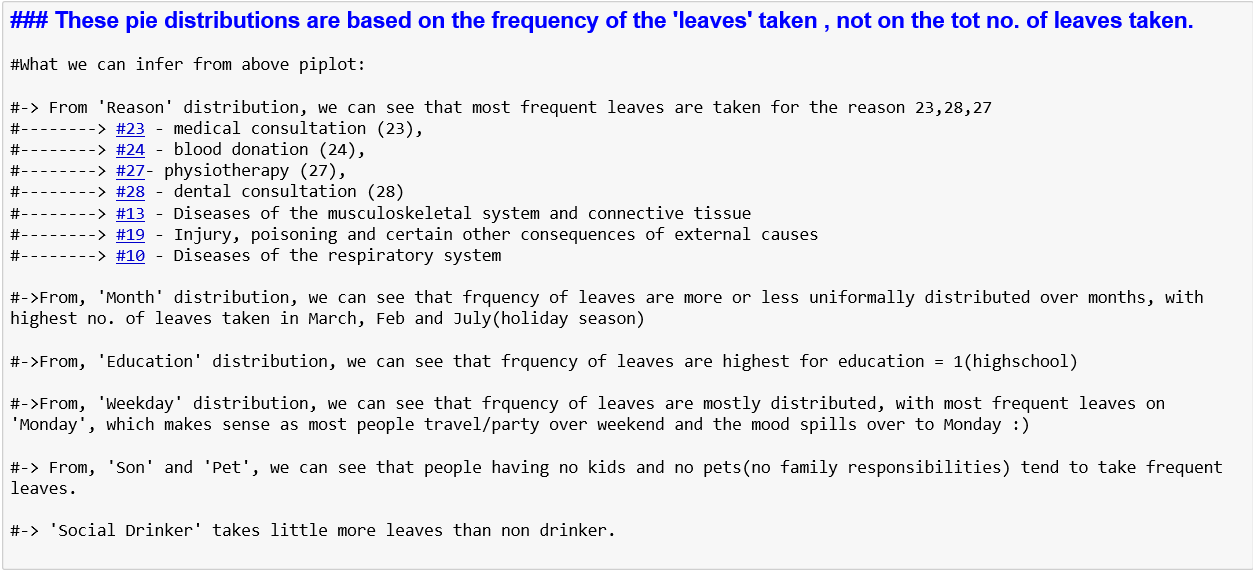
****

* + - 1. **Exploratory Analysis of Categorical Features**

Let’s explore categorical features now through various graphs.

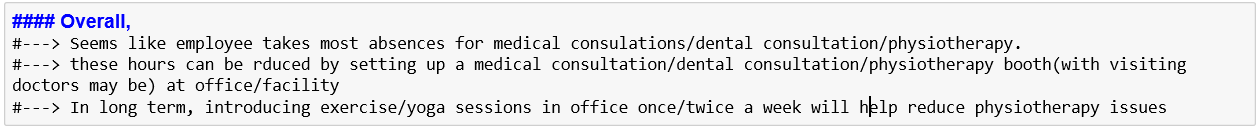
1. **Pie Distribution of Categorical Features**



****

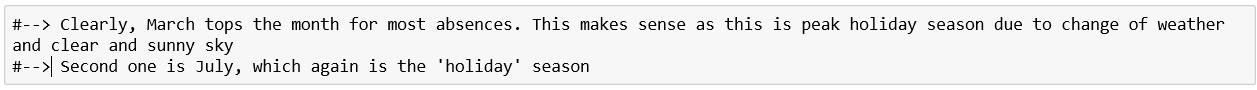
1. **Barplot for Absence Vs. Categorical features**

****

****

****

****

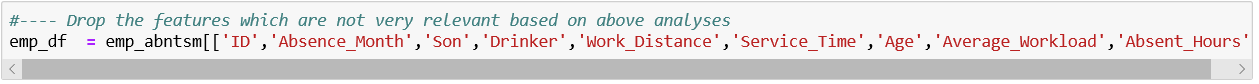
****

Let’s see how these categorical variables individually effects the count of rented bikes

* Does 'yr' affect count of rented bikes
* YES. the count has an upward trend wrt year
* Does 'season' affect count of rented bikes
  + YES, it seems ppl rent more bikes during season 3 and 2, i.e. highest in fall and summer and less in winter and springs. This makes sense as weather is good to ride during summer and fall.
* Does 'month' affect count of rented bikes
  + YES.ppl are likely to rent bikes more btwn the months May- October and lowest in month of Jan,Feb and Dec(in that order). This again makes sense, as this trend is in sync with favourable weather conditions
* Does 'holiday' affect count of rented bikes
  + YES. ppl rent more bikes on non-holiday than holiday. It makes sense as bikers who commute to work/school will be less on holiday.
* Does 'weekday' affect count of rented bikes
  + To some extent Yes. ppl seems to rent lesser bikes on Sat/ Sun. ie. over the weekend. Again makes sense as school and offices are closed on weekend.Monday also has lesser count of rented bikes. It may be possible the ppl visit to other places/cities over weekend and travel back in car on Monday, istead of renting bikes.
* Does 'season' affect count of rented bikes
  + Most definately YES. noone rented bike on extreme weather(season=4). ppl rent maximum bikes during a clear day (weathersit=1)
* Does 'date' affect count of rented bikes
  + Well there is no set trends. It seems to be random. Let explore bit more of it over the 12 months using pairplot
  + From the pointplot over 2 years, it looks like count distribution by ‘date’ has little similarity on trends for both years. So, let’s leave the ‘date’.
    1. **Preparing Data for Modelling**

1. **Drop the Features**

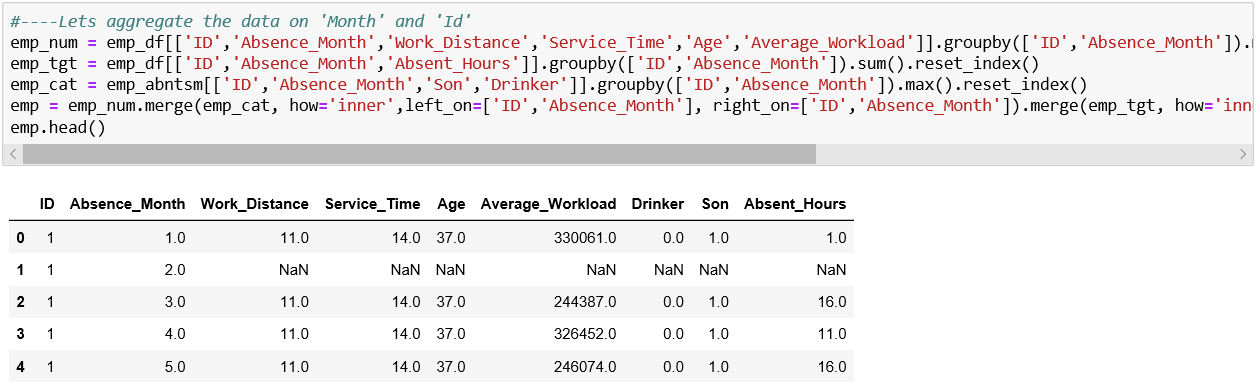
Drop the features which are not relevant for modelling as per the above analysis.



1. **Aggregate the Dataset**

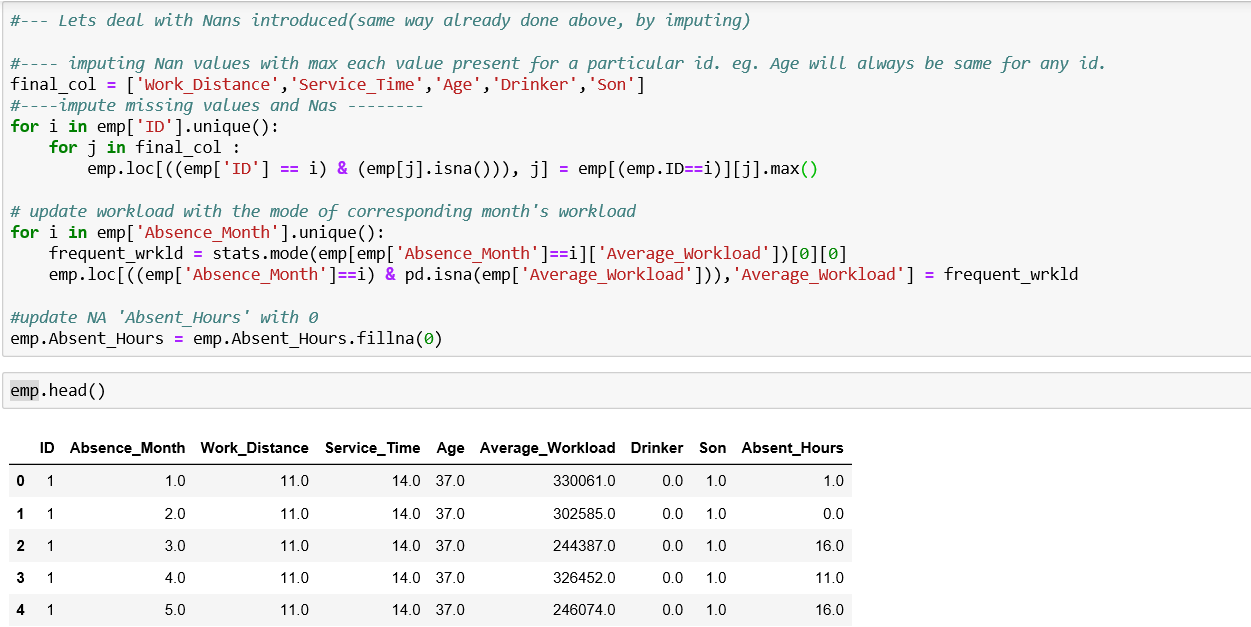
We need to predict absences(losses) per month. However, dataset we have all instances of leaves taken in all the months.

Let’s aggregate dataset for each employee, for each month before feeding to model

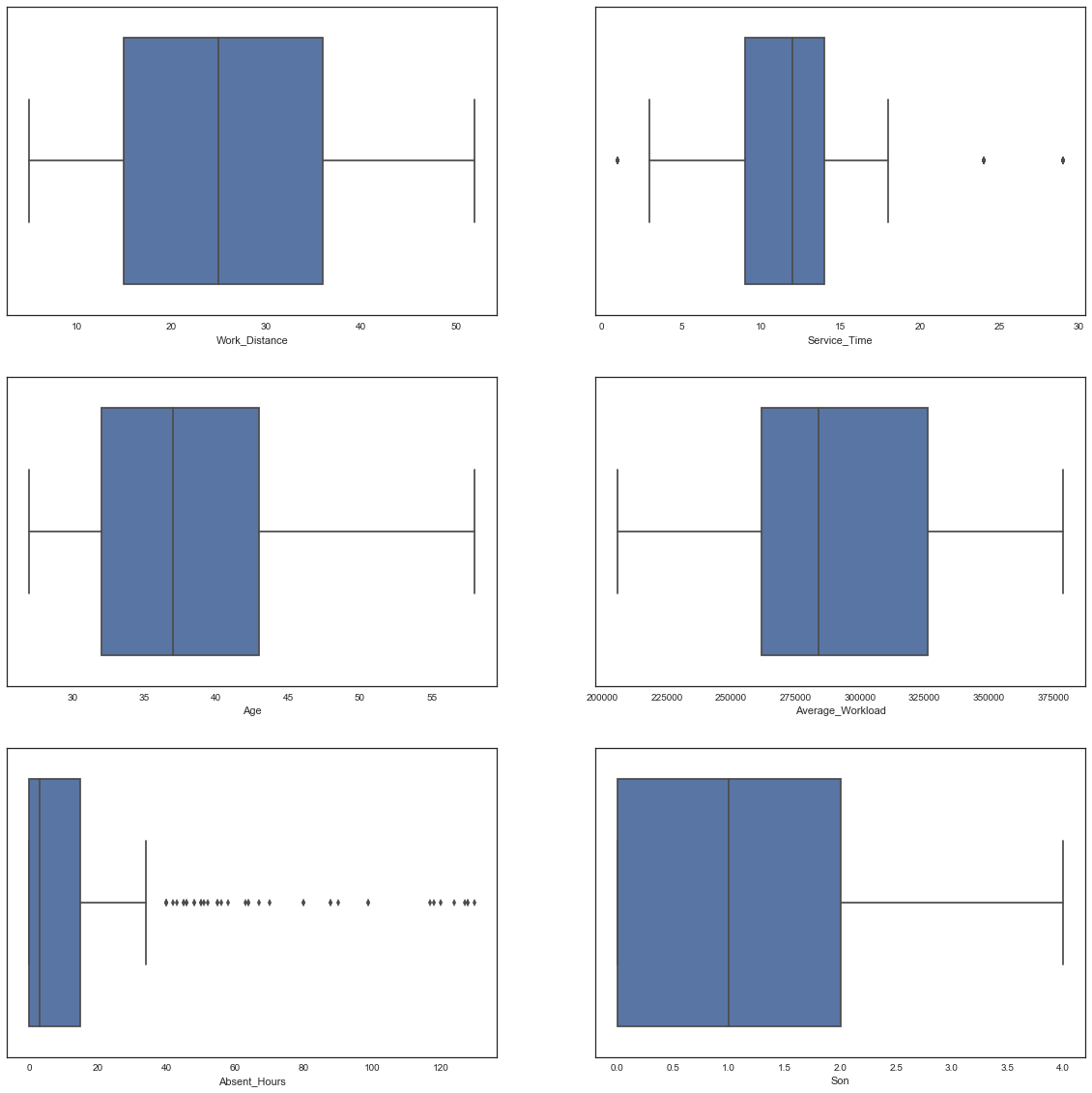


1. **Check Nans in Aggregated Dataset**

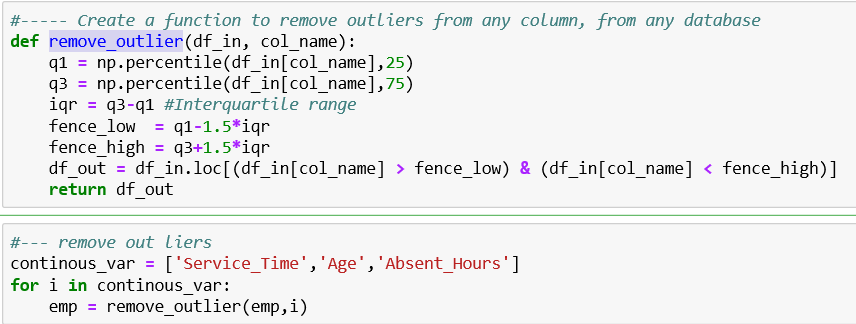
Few Nan values are introduced due to aggregation in the dataset. Let’s Drop/Impute these values accordingly

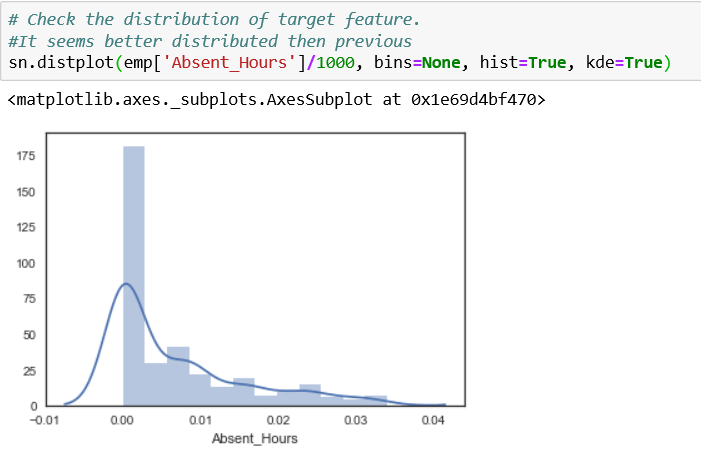


1. **Check for outliers in Aggregated Dataset**

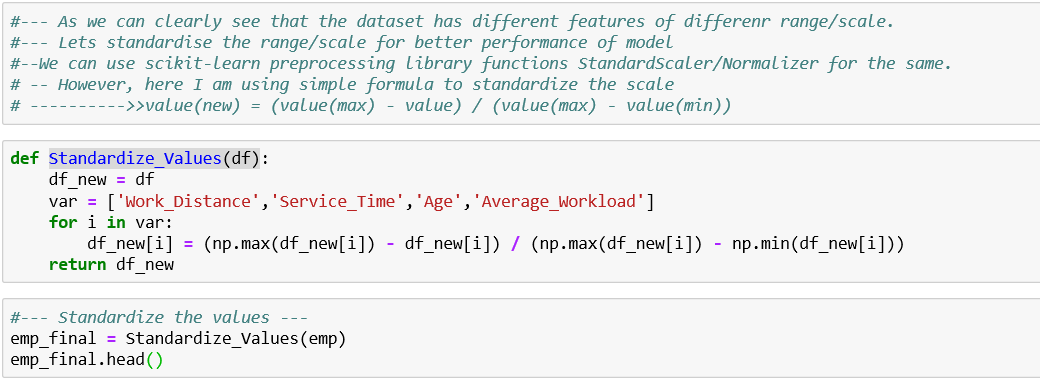


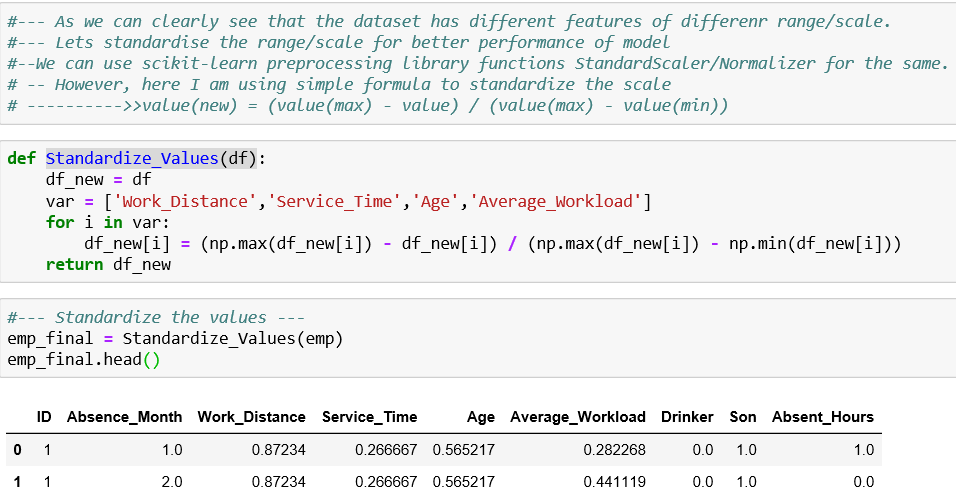
* + - We can see ‘Service\_Time’, ‘Absent\_Hours’ still has some outliers.
    - Lets remove these outliers.
    - Use the custom function ‘remove\_outlier’ for the same





1. **Standardize/Normalize dataset**

****



* + 1. **Modelling**
       1. **Choosing the ML Models**

Now, the exploratory analysis is done, we need to decide on the machine learning algos we’ll use to build the predictive models.

I am going to use 3 ML algo to build 3 different models and later compare them to decide on the best model. Below are the 3 ML algo I have chosen to build model on and compare:

* + - Linear Regression Model
    - Random Forest Model
    - Gradient Boosting Model

on of subset data is used to train their decision trees. As a result, we get an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree classifier.

Now, lets get on to the models we are going to build.

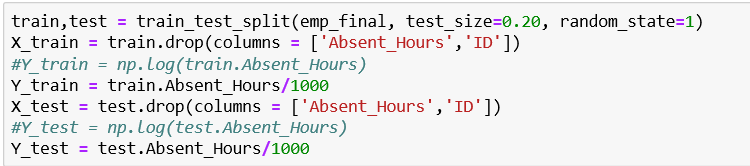
* + - 1. **Choosing the Performance Measures for the Models**

We are working on regression problem, so I think the best performance matrix could be

* + - R-SQAURED
    - RMSE
    - MSE
    - MAE
      1. **Building the ML predictive Models**

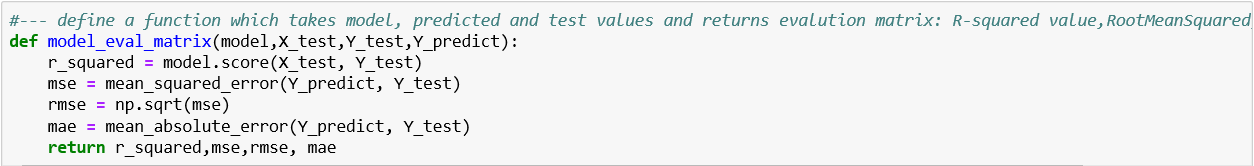
Let’s start building models. Follow the following steps:

1. **Divide Datasets for Cross Validation**



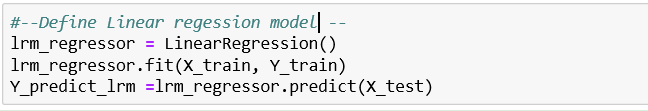
80% of data is used for training while 20% is reserved for testing(cross validating)

1. **Create function to Measure Performance**

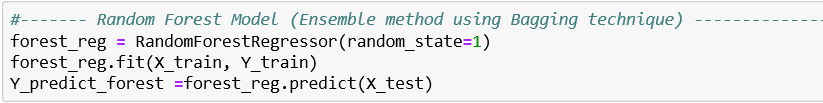


I have created a function, which takes the model, X\_test, Y\_test,Y\_predict values of any model and returns r\_squared,mse,rmse and mae. This is a generic function which would return the performance parameters of any model.

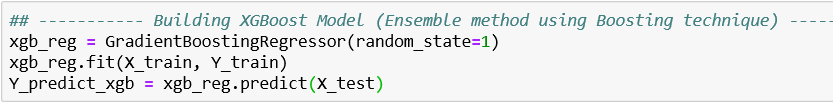
1. **Build, Train & Predict with Linear Regression Model**



1. **Build, Train & Predict with Random Forest Model**



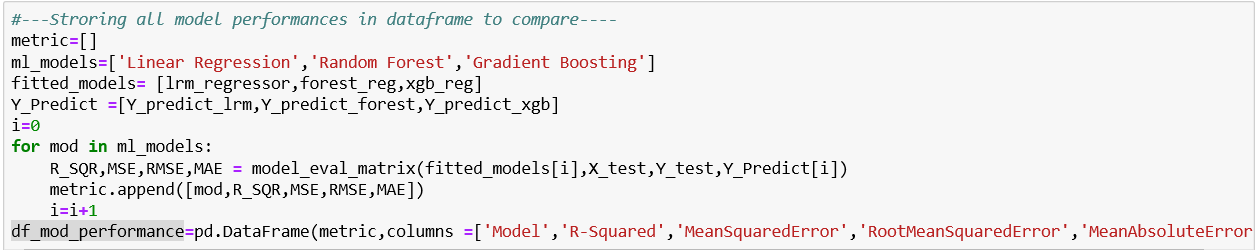
1. **Build, Train & Predict with GradientBoost Model**



Now, all the 3 models are built , lets get the performance scores for the models and store in a dataframe

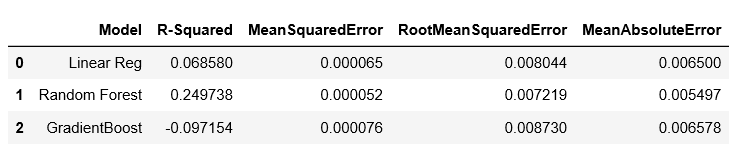
* + - 1. **Building the Performance Matrix dataframe**

Now when all the models are built, let’s call the function to measure performance for each model and store in a dataframe.



* + 1. **Model Evaluation and Comparision**

Results are as below:

**Random Forest model has highest R-squared(goodness of fit) and lowest MeanAbsoluteErorr.**

**Clearly this Random Forest model is our model of choice**

Let’s visualize performance of various models:

1. **Visualize Performance Results with Pointplots**
   * 1. **Hypertuning the Selected Model**

Now, Gradient Boosting is the final model, parameter hypertuning can be performed on the model to find the best parameters which will give the maximum performance.

Functions like **GRIDSearchCV** from **GridSearch** library of python can be used for this.

However, I tried here simple approach of ‘hit and trial’, where I changed parameter few times and found a set which gave me maximum performance.

**Final verdict: we are going to use Random Forest as our final model with the following parameters**:

n\_estimators=2000,

criterion='mse',

max\_depth=10,

min\_samples\_split=5,

min\_samples\_leaf=1,

min\_weight\_fraction\_leaf=0.0,

max\_features='auto',

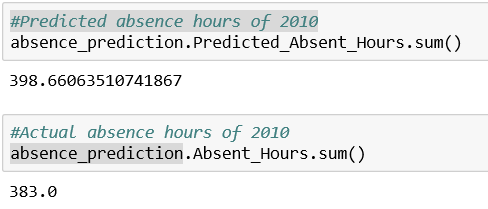
max\_leaf\_nodes=20,

min\_impurity\_decrease=0.00,

min\_impurity\_split=None,

* + 1. **Predicting on test Dataset**

Using Forest Model, we have predicted on test dataset

****

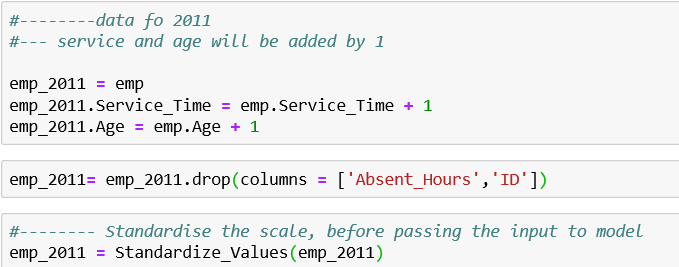
**That’s fair prediction.**

****

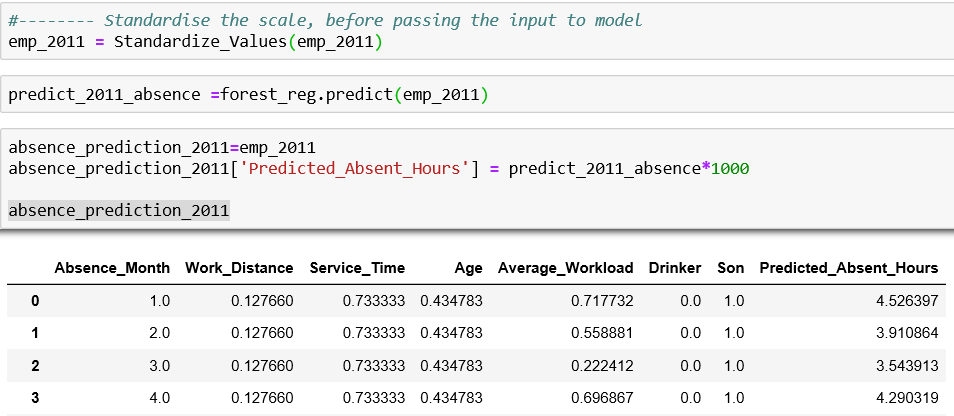
1. Predicting 2011 losses

Now we’ll use the model built above to predict the losses due to absences for 2011

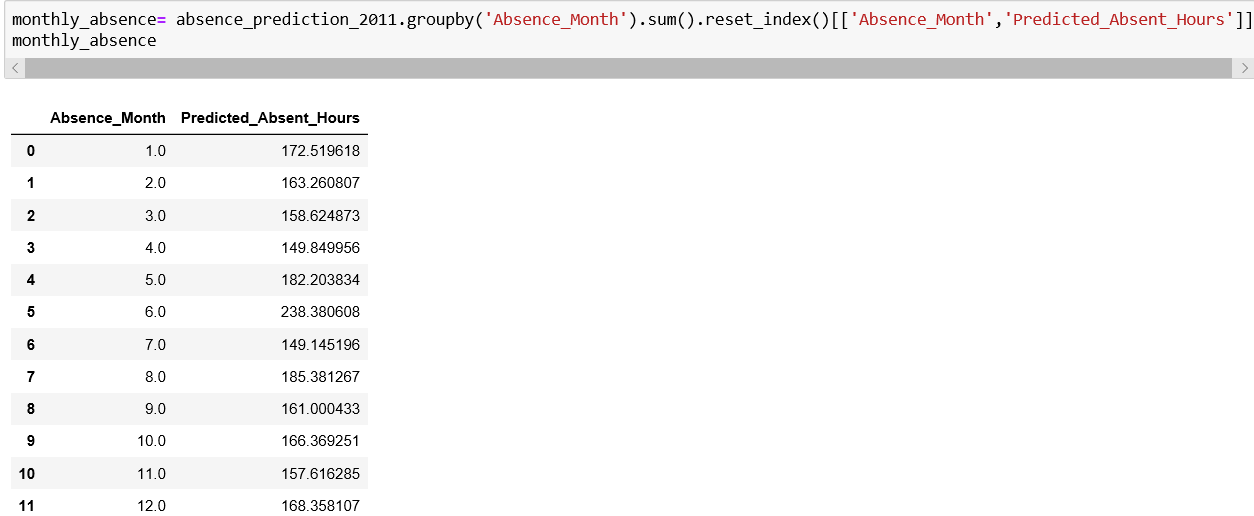
1. **Let’s prepare data for 2011.** 
   * We’ll assume that company retained all the employees and all trend remains same
   * Will derive 2011 input dataset from 2010 dataset(given in the problem)
   * Add +1 to ‘Service\_Time’ in 2010 dataset to obtain ‘Service\_Time’ in 2011
   * Add +1 to ‘Age’ in 2010 dataset to obtain ‘Age’ in 2011
   * All other feature remains same
   * We’ll consider only independent variables and input them to model to predict the target variable ‘Absent\_Hours’



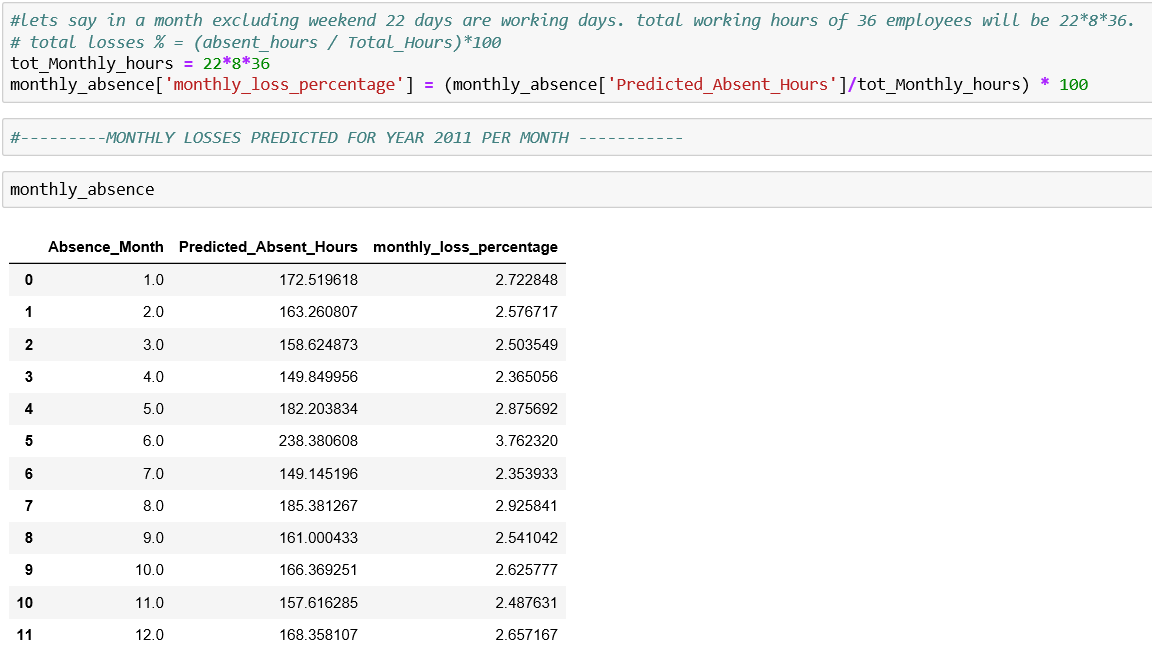
1. **Predict using model**

****

**After Aggregating monthwise:**

****

1. **Calculations to get monthly losses**



1. LAST WORDS: ANSWERING THE QUESTIONS

So now we are done with Analysis and Modelling, let’s answer the 2 questions asked in the problem.

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

As per the exploratory analysis(done in above steps), following are the top reasons for the frequent leaves:

* + medical consultation (23),
  + dental consultation (28)
  + physiotherapy (27),

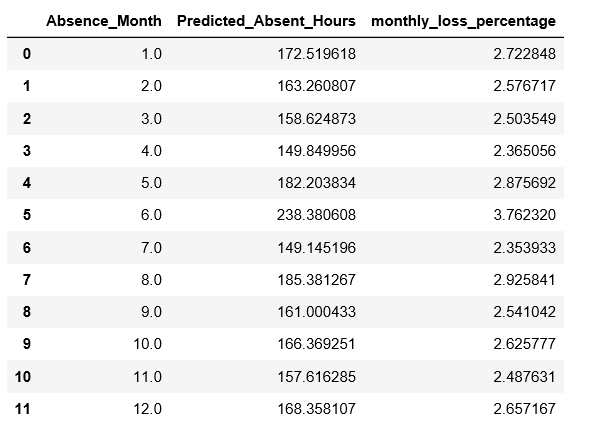
It seems that most employees take leaves for visiting doctor for medical(general consultation)/dental consultation/ physiotherapy

* These hours can be reduced by setting up a medical consultation/dental consultation/physiotherapy booth(with visiting doctors may be) at office/facility, where employees can consult the doctors without taking hours of absence
* Many Physiotherapy cases also indicates that employees might be ‘over worked’. This needs to be taken care of.
* Also, employees with longest service time with the company tends to take less leaves. So, company should try to retain the old employees
* Other factors affecting absence are ‘Travel\_expense’, ‘Distance from office’. Company could provide subsidised travel to employees to rule out absences due to high travel expenses and long distance from work.
* Number of kids/pets area another factor which affect absences to some extent. It is observed that people with no kids/pets tends to take frequent leaves and people with more than 2 kids tend to take less leaves.
* Also, social drinker tends to take more leaves than non drinkers.
* Also, above factors needs to be considered while screening the employees.

1. **How much losses every month can we project in 2011 if same trend of absenteeism continues?**

This has been already answered in section 5. Please check the table below.

Considering that there are 36 employees working for 22 days in a month and 8 hours per day.



1. PYTHON CODE

Python code is attached here: 

Python code is also submitted as separate file employee\_absenteeism.ipynb

1. RCODE

R code is attached here:

R code is also submitted as separate file employee\_absenteeism.R