Employee Absenteeism

-Project Report

Divakar Sunkara

15/02/2020

Contents

1 Introduction

1.1 Problem Statement

Human capital plays an important role in courier companies for work like collection ,transportation and delivery.However,absenteeism poses serious threat to the profitability of the company.Our problem at hand is to assist XYZ courier company in :

i) Formulating policies for reducing the number of changes.

ii)To project monthly loss in 2011,if the same trend continues.

1.2 Data

Dataset Characteristics: Time Series Multivariate

Number of Attributes: 21

Missing Values : Yes

Attribute 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 (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)

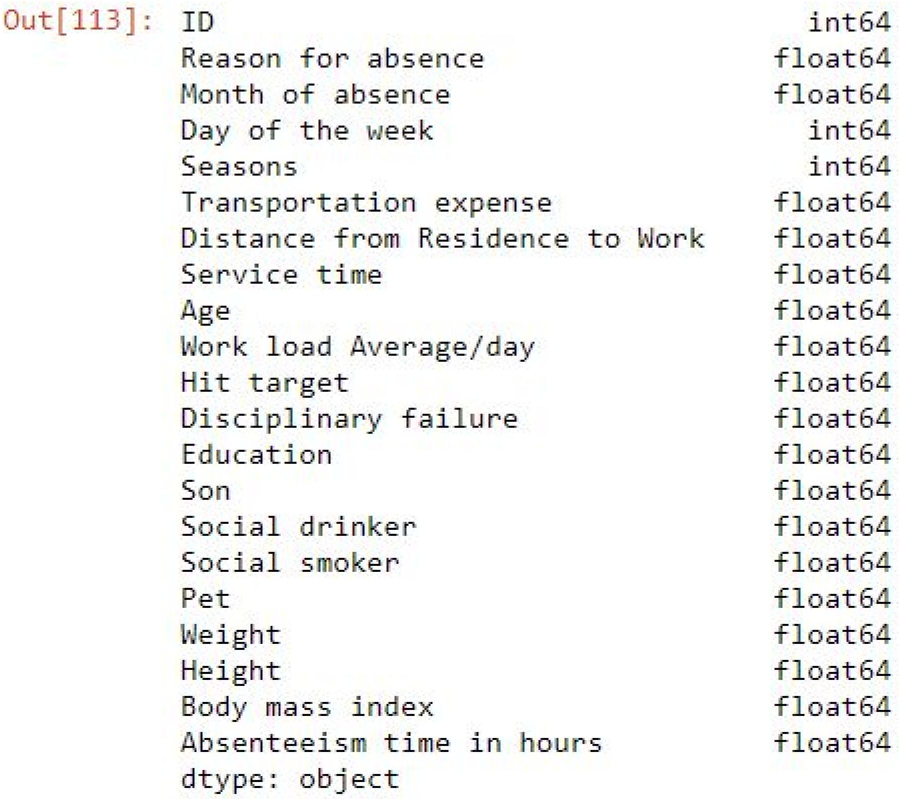
18. Weight

19. Height

20. Body mass index

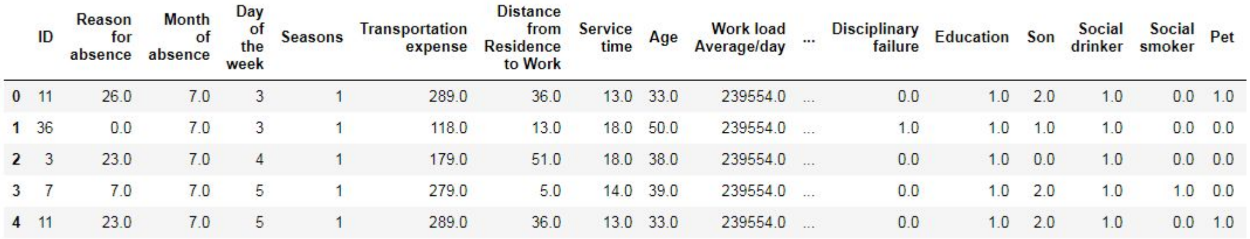
21. Absenteeism time in hours (target)

The size of the given data is **(740, 21)** with 21 columns with different data types as shown below.

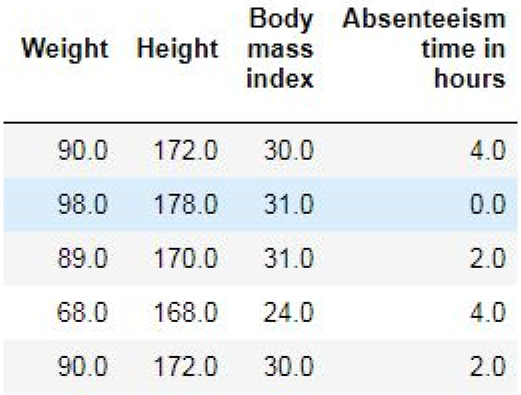


```
Out[113]: ID                                  int64
          Reason for absence                float64
          Month of absence                  float64
          Day of the week                     int64
          Seasons                             int64
          Transportation expense            float64
          Distance from Residence to Work   float64
          Service time                      float64
          Age                               float64
          Work load Average/day             float64
          Hit target                        float64
          Disciplinary failure              float64
          Education                         float64
          Son                               float64
          Social drinker                    float64
          Social smoker                     float64
          Pet                               float64
          Weight                            float64
          Height                            float64
          Body mass index                   float64
          Absenteeism time in hours         float64
          dtype: object
```

Sample data:



	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet
0	11	26.0	7.0	3	1	289.0	36.0	13.0	33.0	239554.0	...	0.0	1.0	2.0	1.0	0.0	1.0
1	36	0.0	7.0	3	1	118.0	13.0	18.0	50.0	239554.0	...	1.0	1.0	1.0	1.0	0.0	0.0
2	3	23.0	7.0	4	1	179.0	51.0	18.0	38.0	239554.0	...	0.0	1.0	0.0	1.0	0.0	0.0
3	7	7.0	7.0	5	1	279.0	5.0	14.0	39.0	239554.0	...	0.0	1.0	2.0	1.0	1.0	0.0
4	11	23.0	7.0	5	1	289.0	36.0	13.0	33.0	239554.0	...	0.0	1.0	2.0	1.0	0.0	1.0



Weight	Height	Body mass index	Absenteeism time in hours
90.0	172.0	30.0	4.0
98.0	178.0	31.0	0.0
89.0	170.0	31.0	2.0
68.0	168.0	24.0	4.0
90.0	172.0	30.0	2.0

We need to find/predict the **Absenteeism time in hours** from the given data. So our target variable is **Absenteeism time in hours** which is a continuous variable. So it is a regression problem.

Target variable : **Absenteeism time in hours**

Based on the type of problem, We need to decide the type of models and metrics that we are going to apply on the data.

Below are the models that we need to apply on the data to train the models.

1. KNN

2. Linear regression

3. Decision tree

4. Ensemble methods.

Below are the metrics that we need to apply for evaluating model performance.

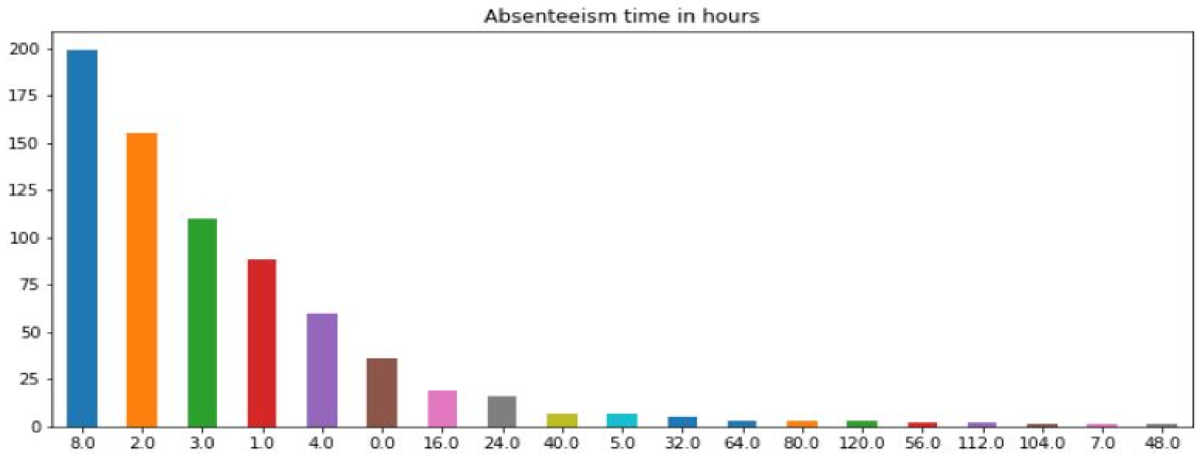
1. RSME (Root mean square error)

2. MSE (Mean square error)

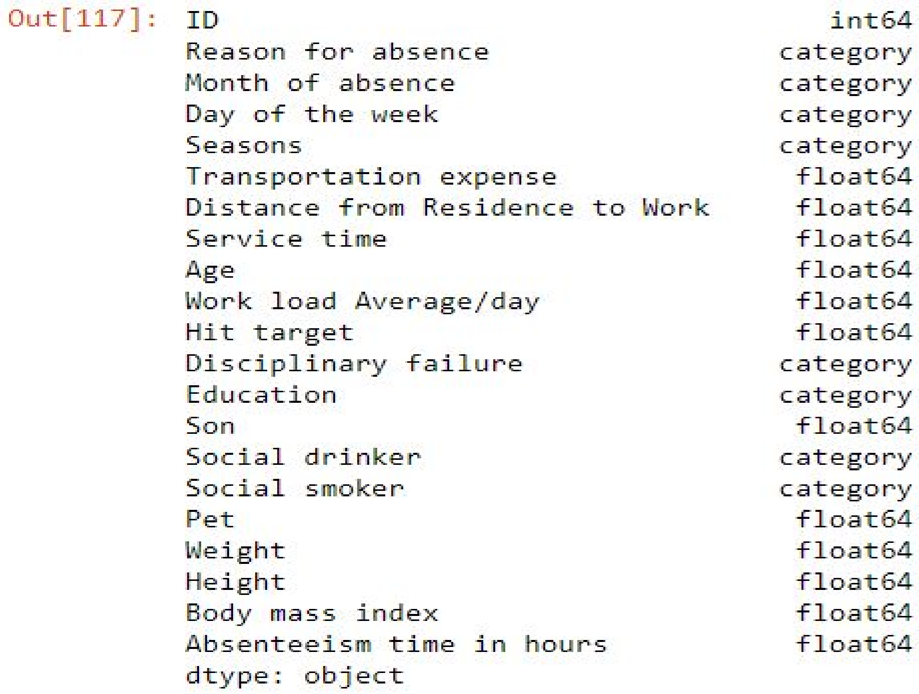
3. R square

4. Adjusted R square.

Distribution of Absenteeism in hours as shown below for the given data.



Given data is not in appropriate format (Appropriate data type), So we need to convert given data into required formats (float to category etc…)



```
Out[117]: ID                                  int64
          Reason for absence               category
          Month of absence                 category
          Day of the week                  category
          Seasons                          category
          Transportation expense            float64
          Distance from Residence to Work   float64
          Service time                      float64
          Age                               float64
          Work load Average/day             float64
          Hit target                        float64
          Disciplinary failure             category
          Education                        category
          Son                               float64
          Social drinker                   category
          Social smoker                    category
          Pet                               float64
          Weight                            float64
          Height                            float64
          Body mass index                   float64
          Absenteeism time in hours         float64
          dtype: object
```

2 Data Processing

2.1 Missing Value Analysis

We can categorise our feature set into numerical and categorical feature set consisting of following features in the data:

Numerical features set :

"ID","Transportation expense","Distance from Residence to Work","Service time","Age","Work load Average day","Hit target","Son","Pet","Height","Weight","Body mass index","Absenteeism time in hours"

Categorical features set :

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

Before proceeding with any analysis ,we must get a feel of the data set at hand .We will first evaluate missing value in the data.

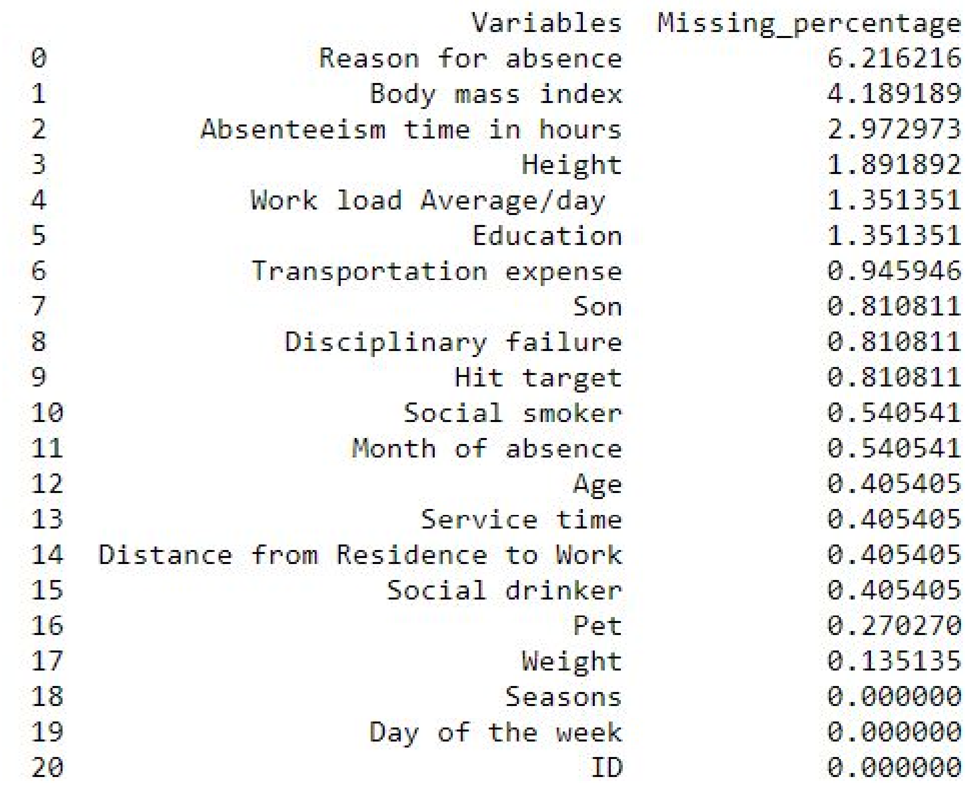
Many times, data set has missing value due to various reasons may be error in collection or error in reporting the data .Let us find out how many data points are missing in our data set.

Also in the data set it was observed that some predictors like

"Reason for absence","Month of absence","Day of the week","Seasons","Education","ID","Age","Weight","Height"Body mass index" had ‘0’ value in the observation .

Logically ‘0’ values for these predictors are not acceptable and can be treated as missing values. We replace these values with NA in the data and then do missing value analysis.

Below is a summary of the missing value in our data set .



```
                          Variables  Missing_percentage
0                Reason for absence            6.216216
1                   Body mass index            4.189189
2       Absenteeism time in hours              2.972973
3                            Height            1.891892
4            Work load Average/day             1.351351
5                         Education            1.351351
6            Transportation expense            0.945946
7                               Son            0.810811
8              Disciplinary failure            0.810811
9                        Hit target            0.810811
10                    Social smoker            0.540541
11                 Month of absence            0.540541
12                              Age            0.405405
13                     Service time            0.405405
14  Distance from Residence to Work            0.405405
15                   Social drinker            0.405405
16                              Pet            0.270270
17                           Weight            0.135135
18                          Seasons            0.000000
19                  Day of the week            0.000000
20                               ID            0.000000
```

However we see that no column has more than 30 % of the missing data. Thus we keep all the features set for our further analysis.

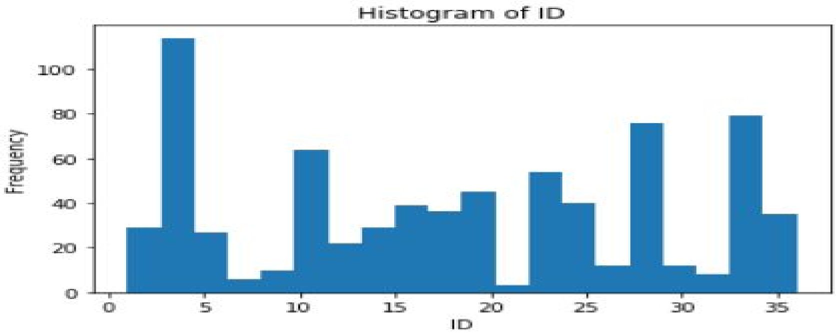
Out of the many methods ,we test following 3 methods i.e, mean mode method ,median mode method and knn method We see that median mode suits best in our dataset.Not only this ,we have also observed many features has ‘0’ as input which makes no sense .Thus we replace them as “NA” data point.

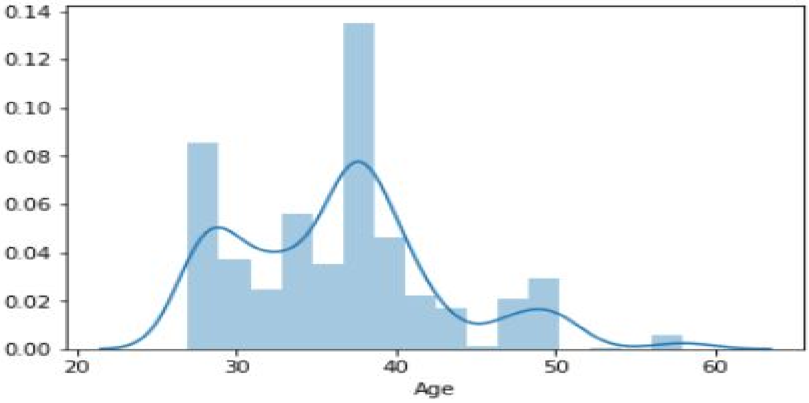
2.2 Data Visualisation

After imputation let us visualise our dataset ,to get a pictorial representation:

Data visualisation is a must for better understanding the data. There are many plots used to explain the data and to find the anomalies in the data. This will help to preprocess the data for model building and inferential statistics to assume the hypothesis.

Numerical type of data:





From the multiple histogram charts, pair plots etc.. on the numerical type of columns, we have identified few interesting details like below.

ID :The frequency distribution of the “ID” feature set shows a non uniform.However,around ID 3 is observed to be occuring most frequently as compared to others.

Transportation expense : Most of the transportation expense is around 175 and the distribution is non uniform.

Distance from resident to work :Residential distance is found to occur mostll for around 25 units or more than 50 units.

Service time : The observation set has employees with service time around 7-17 hours.

Age : The sample set in our study is mainly in the age group of 25 – 40 as depicted in the age frequency plot.

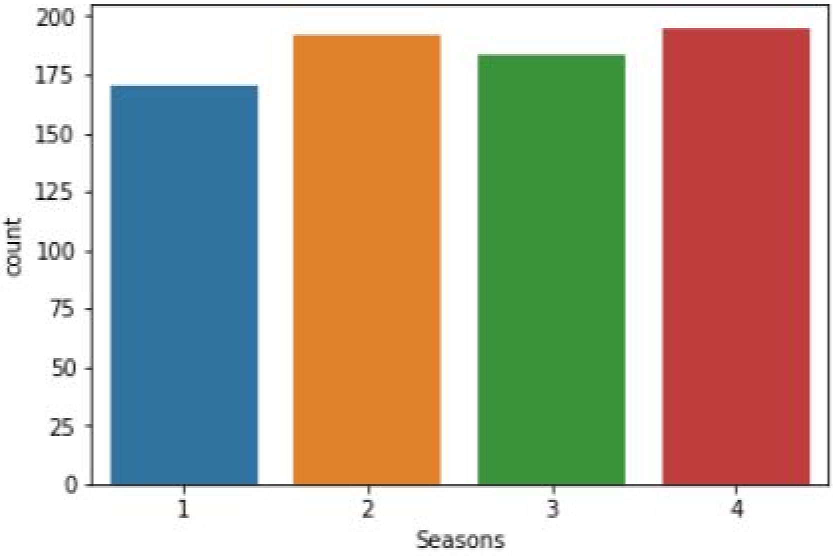
Hit target : Hit target is around 92 to 99 in our data set.On plotting ,it against our target variable ,we see there is some higher amount of absenteeism hours around 92 – 93 and around 98-99 units.Some concentration of absenteeism is also seen on the lower end of hit target

Work load average day :From the frequency plot we see that work load average is spread across the entire range with peak around 260000 – 270000 units

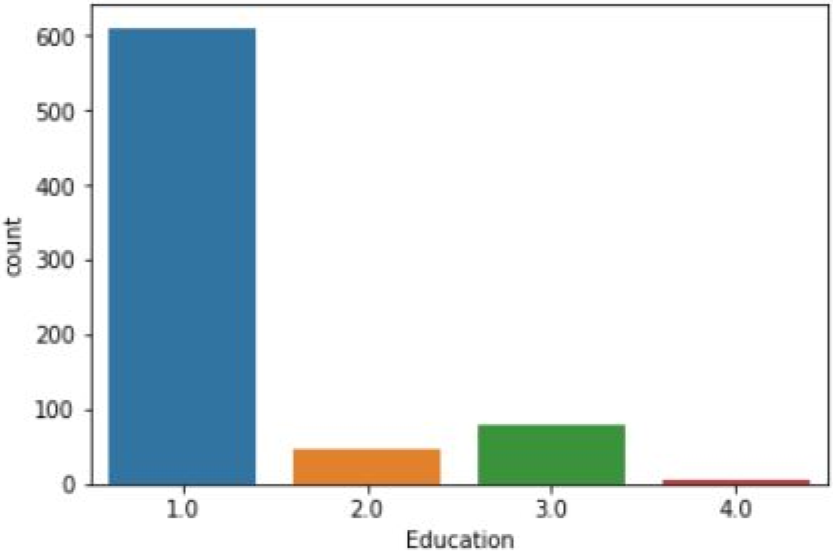
Son frequency : Most of the employees in the data set have less than or equal to 2 sons.Higher absenteeism hours are seen for employees with 1 and 2 sons.

Pet : Most of the employees don’t have any pets .However,some employees have 1 or 2 pet.We can see some data points around 4 and 8 which can be considered as outliers.Absenteeism hours is high around 0 and 1 pet.

Categorical feature set



Season:The frequency of season in our data set is almost uniform and so is the median and range of absenteeism hours.

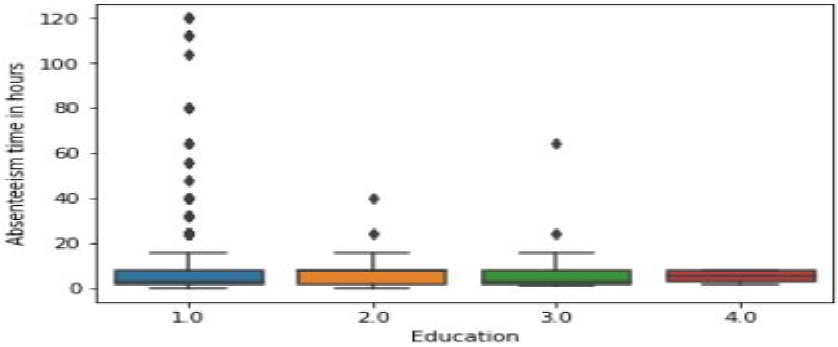


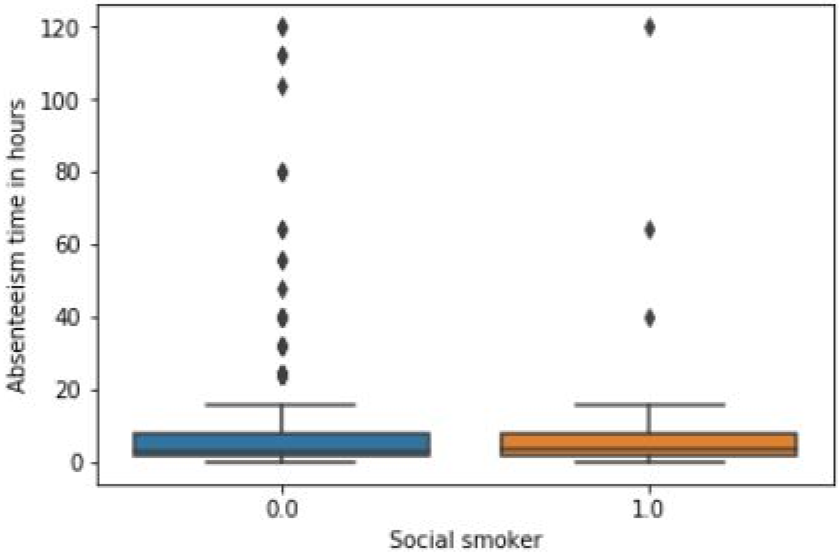
Education : Sample sets of the employees in XYZ company are mostly found to be high school educated.The range and median of absenteeism hours grouped by the education level is mostly uniform.Also high school educated employees show more absenteeism hours.

2.3 Outlier Analysis

Before proceeding further with the analysis , we would like to do outlier analysis using boxplot method, which means that any data point that is less than 1.5*IQR(InterQuartile range) times the 25th percentile and more than 1.5*IQR the 75th percentile ,is to be treated as an outlier .We replace these items with NaN in the dataset and then impute it with the median values.

Below is the Boxplots of the numerical features with outliers.



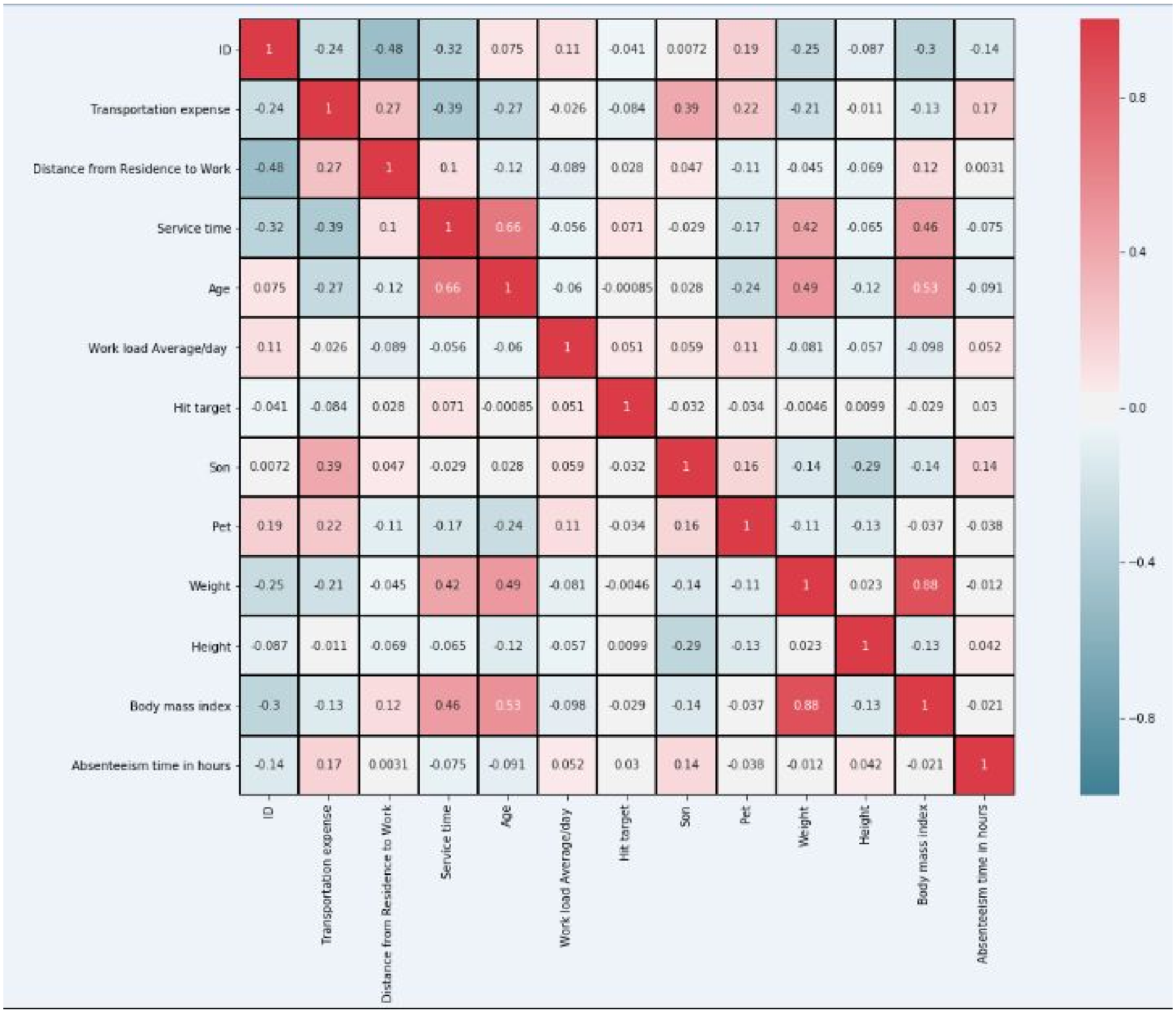


2.4 Feature Selection

One of the key tasks in any data science operation is to choose the right set of predictors. This is because ,although more features implies more knowledge of our dataset but high dimension in the data set can also lead to higher variance which might fail to generalise on the test data leading to higher test MSE(Mean Square Error). This is also known as the curse of dimensionality. Apart from this , higher dimensional data in our model can also be computationally expensive. Thus we need to perform feature selection before supplying predictors to our model.

So based on the correlation analysis, We need to remove the highly correlated data from the entire data. It is called feature selection.

We plot **correlation plot** of our numerical data set :



	ID	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Son	Pet	Weight	Height	Body mass index	Absenteeism time in hours
ID	1	-0.24	-0.48	-0.32	0.075	0.11	-0.041	0.0072	0.19	-0.25	-0.087	-0.3	-0.14
Transportation expense	-0.24	1	0.27	-0.39	-0.27	-0.026	-0.084	0.39	0.22	-0.21	-0.011	-0.13	0.17
Distance from Residence to Work	-0.48	0.27	1	0.1	-0.12	-0.089	0.028	0.047	-0.11	-0.045	-0.069	0.12	0.0031
Service time	-0.32	-0.39	0.1	1	0.66	-0.056	0.071	-0.029	-0.17	0.42	-0.065	0.46	-0.075
Age	0.075	-0.27	-0.12	0.66	1	-0.06	-0.00085	0.028	-0.24	0.49	-0.12	0.53	-0.091
Work load Average/day	0.11	-0.026	-0.089	-0.056	-0.06	1	0.051	0.059	0.11	-0.081	-0.057	-0.098	0.052
Hit target	-0.041	-0.084	0.028	0.071	-0.00085	0.051	1	-0.032	-0.034	-0.0046	0.0099	-0.029	0.03
Son	0.0072	0.39	0.047	-0.029	0.028	0.059	-0.032	1	0.16	-0.14	-0.29	-0.14	0.14
Pet	0.19	0.22	-0.11	-0.17	-0.24	0.11	-0.034	0.16	1	-0.11	-0.13	-0.037	-0.038
Weight	-0.25	-0.21	-0.045	0.42	0.49	-0.081	-0.0046	-0.14	-0.11	1	0.023	0.88	-0.012
Height	-0.087	-0.011	-0.069	-0.065	-0.12	-0.057	0.0099	-0.29	-0.13	0.023	1	-0.13	0.042
Body mass index	-0.3	-0.13	0.12	0.46	0.53	-0.098	-0.029	-0.14	-0.037	0.88	-0.13	1	-0.021
Absenteeism time in hours	-0.14	0.17	0.0031	-0.075	-0.091	0.052	0.03	0.14	-0.038	-0.012	0.042	-0.021	1

Multicollinearity can be checked through VIF(**Variance Inflation Factor**) values. We obtain the following VIF values for our data set.

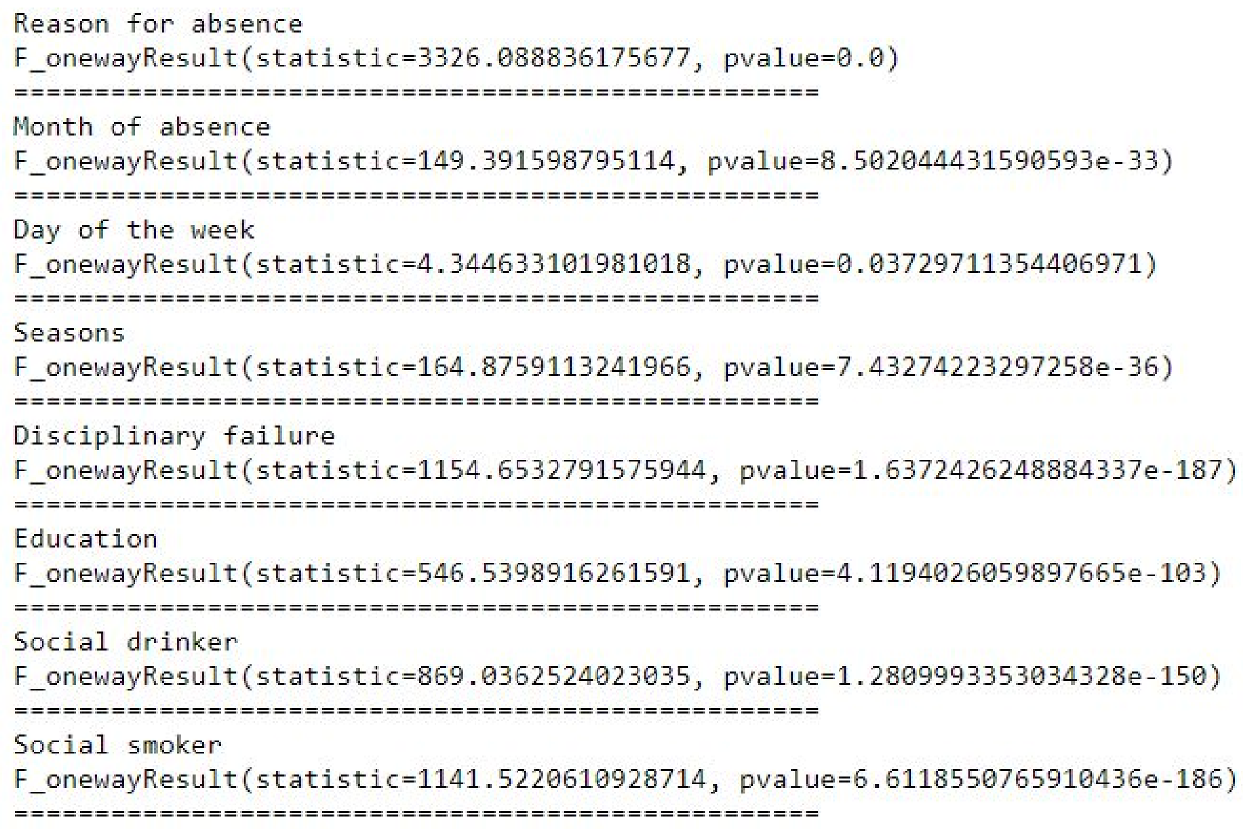
```
ID                                        2.555528
Transportation expense                    2.199472
Distance from Residence to Work           1.593952
Service time                              3.443374
Age                                       3.501619
Work load Average/day                     1.042375
Hit target                                1.024523
Son                                       1.532156
Pet                                       1.458590
Weight                                    6.039502
Height                                    1.328574
Body mass index                           7.227522
Absenteeism time in hours                 1.082891
const                                 14722.123211
dtype: float64
```

Any feature set having more than 0.80 correlation will be removed .

Also, the feature set with VIF > 5 ,will be removed. Thus, we remove one of the features out of weight and Body Mass Index .We chose to remove one of the variables that is Weight

Feature selection on categorical data set :

As our target variable is continuous data ,we select anova test for performing feature selection on categorical data set .



```
Reason for absence
F_onewayResult(statistic=3326.088836175677, pvalue=0.0)
==================================================
Month of absence
F_onewayResult(statistic=149.391598795114, pvalue=8.502044431590593e-33)
==================================================
Day of the week
F_onewayResult(statistic=4.344633101981018, pvalue=0.03729711354406971)
==================================================
Seasons
F_onewayResult(statistic=164.8759113241966, pvalue=7.43274223297258e-36)
==================================================
Disciplinary failure
F_onewayResult(statistic=1154.6532791575944, pvalue=1.6372426248884337e-187)
==================================================
Education
F_onewayResult(statistic=546.5398916261591, pvalue=4.1194026059897665e-103)
==================================================
Social drinker
F_onewayResult(statistic=869.0362524023035, pvalue=1.2809993353034328e-150)
==================================================
Social smoker
F_onewayResult(statistic=1141.5220610928714, pvalue=6.6118550765910436e-186)
==================================================
```

Taking 95% as our confidence interval,we would select only those features whose p value is less than 0.05 i.e "Reason.for.absence","Month.of.absence","Disciplinary.failure"

Thus ,we reduce our overall dimension of 21 predictors to 15 predictors.

3 Modeling

We choose Random Sampling to divide our dataset into test and train set based on Reason of absence feature which is continuous variable.

Our problem statement wants us to predict the fare_amount. This is a Regression problem. So,

We are going to build regression models on training data and predict it on test data. In this

project I have built models using 5 Regression Algorithms:

I. Linear Regression

II. LinearRegression with statsmodels - OLS

III. Lasso Regression, Ridge Regression

IV. Decision Tree

V. Random Forest

VI. Xgboost Regression

We will evaluate performance on validation dataset which was generated using Sampling. We

will deal with specific error metrics like – Regression metrics for our Models: -

- r square

- Adjusted r square

- MAPE (Mean Absolute Percentage Error)

- MSE(Mean square Error)

- RMSE(Root Mean Square Error)

3.1 Model Performance

Here, we will evaluate the performance of different Regression models based on different Error Metrics-

LinearRegression

```
===================Building the model... ===========================

Model LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False) ran
successfully..

======================== Score's =====================================

r square :  0.1619149462136612

Adjusted r square : 0.11090107337449273

MSE : 7.563276053739892
RMSE : 2.7501410970602747
```

LinearRegression with statsmodels

```
                    OLS Regression Results
==============================================================================
Dep. Variable:                   y   R-squared:                     0.711
Model:                         OLS   Adj. R-squared:                0.703
Method:              Least Squares   F-statistic:                   84.70
Date:             Thu, 13 Feb 2020   Prob (F-statistic):        5.93e-120
Time:                     09:54:10   Log-Likelihood:              -1237.9
No. Observations:              495   AIC:                           2504.
Df Residuals:                  481   BIC:                           2563.
Df Model:                       14
Covariance Type:         nonrobust
==============================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
ID                       -0.0540      0.020     -2.713      0.007      -0.093      -0.015
Reason for absence       -0.1626      0.020     -7.963      0.000      -0.203      -0.122
Month of absence         -0.0218      0.045     -0.483      0.630      -0.111       0.067
Transportation expense    0.0060      0.003      1.986      0.048    6.34e-05       0.012
Distance from Residence to Work  -0.0251      0.011     -2.311      0.021      -0.046      -0.004
```

```
Service time                -0.0614     0.061     -1.007     0.314     -0.181      0.058
Age                       -0.0239     0.043     -0.562     0.574     -0.108      0.060
Work load Average/day          6.835e-06   4.31e-06      1.586      0.113   -1.63e-06    1.53e-05
Hit target                  0.0115     0.047      0.243      0.808     -0.081      0.104
Disciplinary failure            -4.0336     0.638     -6.323      0.000     -5.287      -2.780
Son                        0.4247      0.141      3.012      0.003      0.148      0.702
Pet                       -0.5629     0.228     -2.471      0.014     -1.011      -0.115
Height                      0.0362      0.029      1.236      0.217     -0.021      0.094
Body mass index                 0.0275      0.044      0.628      0.530     -0.058      0.113
==============================================================================
Omnibus:                 147.752   Durbin-Watson:                1.849
Prob(Omnibus):              0.000   Jarque-Bera (JB):            423.674
Skew:                   1.433   Prob(JB):                  1.00e-92
Kurtosis:                 6.511   Cond. No.                   1.28e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.28e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
RMSE value : 2.749753681855755
```

Ridge

```
==================Building the model... ==========================
Model Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None,
   normalize=False, random_state=None, solver='auto', tol=0.001) ran successfully..



======================= Score's ===================================
r square :  0.1629013254104088
Adjusted r square : 0.1119474930440858
MSE : 7.554374501176742
RMSE : 2.748522239527405
```

Lasso

```
==================Building the model... ==========================
Model Lasso(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=1000,
   normalize=False, positive=False, precompute=False, random_state=None,
   selection='cyclic', tol=0.0001, warm_start=False) ran successfully..
```

```
======================= Score's ===================================
r square :  0.10909471500822399
Adjusted r square : 0.05486569766089855
MSE : 8.039950811300876
RMSE : 2.8354807019799795
```

KNeighborsRegressor

```
===================Building the model... ===========================
Model KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
          metric_params=None, n_jobs=1, n_neighbors=5, p=2,
          weights='uniform') ran successfully..


======================= Score's ===================================
r square :  -0.19719804908306182
Adjusted r square : -0.270070973809857
MSE : 10.80408163265306
RMSE : 3.286956287000644
```

DecisionTreeRegressor

```
===================Building the model... ===========================
Model DecisionTreeRegressor(criterion='mse', max_depth=None, max_features=None,
           max_leaf_nodes=None, min_impurity_decrease=0.0,
           min_impurity_split=None, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           presort=False, random_state=None, splitter='best') ran successfully..


======================= Score's ===================================
r square :  -0.668078651256208
Adjusted r square : -0.7696138735065858
MSE : 15.053530977319122
RMSE : 3.8798880109249443
```

RandomForestRegressor

```
===================Building the model... ===========================
Model RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=2,
           max_features=10, max_leaf_nodes=None, min_impurity_decrease=0.0,
           min_impurity_split=None, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           n_estimators=10, n_jobs=1, oob_score=False, random_state=0,
           verbose=0, warm_start=False) ran successfully..


======================== Score's ===================================
r square :  0.21175865984268571
Adjusted r square : 0.16377875218093607
MSE : 7.1134627990866495
RMSE : 2.667107571712594
```

XGBRegressor

```
===================Building the model... ============================
[09:59:16] WARNING:
C:/Jenkins/workspace/xgboost-win64_release_0.90/src/objective/regression_obj.cu:152: reg:linear
is now deprecated in favor of reg:squarederror.
Model XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
       colsample_bynode=1, colsample_bytree=1, gamma=0,
       importance_type='gain', learning_rate=0.1, max_delta_step=0,
       max_depth=3, min_child_weight=1, missing=None, n_estimators=100,
       n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
       reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
       silent=None, subsample=1, verbosity=1) ran successfully..


======================== Score's ===================================
r square :  0.2517689470391494
Adjusted r square : 0.20622444816327157
MSE : 6.752391037110787
RMSE : 2.5985363259171086
```

3.2 Best model

Out of all the models, XGBoost gave the best results with an RMSE value of 2.59. So we can use XGboost for this problem.

3.3 Deployment

We can store the model in many different ways like json, pickle, csv etc in python.. But here I use pickle way of sterilizing the model.

Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script.
When our model is trained , we need to dump or store the model. To do so we use a dump method in pickle to store the model data. When we want to predict the on test data we load the pickle object which we stored for the best model and will use a predict method to predict the test data..

4.Conclusions

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

Companies should bring a good medical plan for their employees, reduce the workload on each employee by recruiting the people and motivate the people towards work by providing rewards or incentives on a timely basis.

Summary from the visualization:

Absenteeism time in hours:
More no.of persons (200) are absent for 8 hours.
Less no.of persons (~10) are absent for 48 hours.

Highest no.of hours a few people are absent from work is 120 hours.

Numerical data summary:
Most of the people are in the age of 30-40 years with workload average of 250000/day and 95-100% successful delivery rate

Most people are in the height range of 170 CM, weight ranges of 60-70 & 80-90KG's with BMI ranges 20 - 30

Categorical data summary:
Most people said the reason was medical consultation (23) for their absence.
We can see more absences in the march month.
Almost everyday, 140-160 are absent from work. This is same for every season
Most of the people who work are from high school education and there are very less no of persons from master graduate & doctor's
There are more drinkers than smokers.

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

Now that we have analysed our data set and selected or predictive model.We can see that the most important predictor in our absenteeism prediction is Reason for absence (fig no. 2.8). From the bar plot and boxplot chart (fig no.2.3) of Reason for absence we see maximum frequency of absent reason is reason no. 23(medical consultation) and 28 (dental consultation).Thus , it is evident that maximum absenteeism is due to health related reasons.

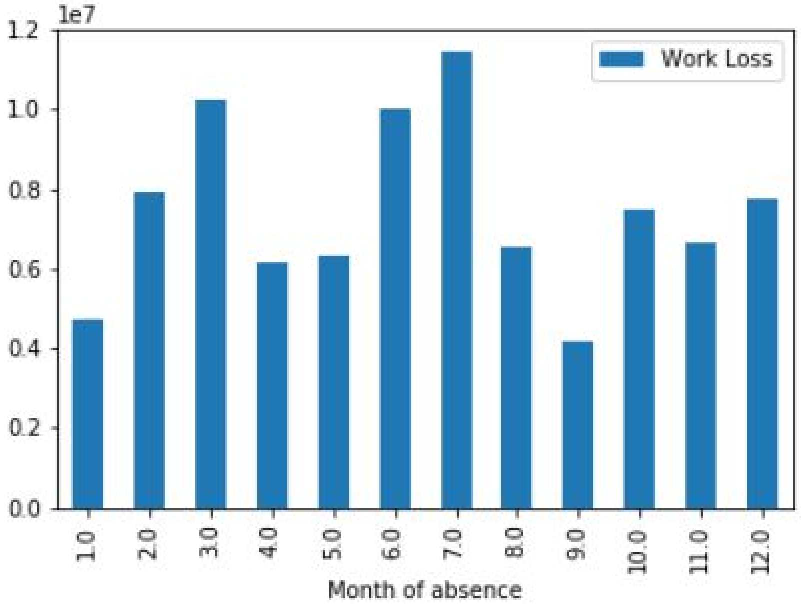
On summarising the Count, Sum of Absenteeism hours and Mean of absenteeism hours Reason wise, we see that medical consultation(Reason no. 23) and dental consultation(Reason no. 28) is a common cause of absenteeism .Hence the company can arrange for free regular medical consultation and dental consultation in coalition with some hospitals and other promotion camps in its office.

Our second problem at hand is to predict monthly work loss for the company is the same trend continues.We calculate work loss as :

Work Loss = (Work.load.Average.day/Service.time)*Absenteeism.time.in.hours

Work loss for 2011 ,considering the same trend in the absenteeism pattern is :

Month of absence	Work Loss
1	4,730,330.00
2	7,938,029.00
3	10,237,680.00
4	6,140,413.00
5	6,341,453.00
6	10,033,180.00
7	11,431,100.00
8	6,520,187.00
9	4,159,296.00
10	7,494,952.00
11	6,674,416.00
12	7,742,550.00



Final Summary

Work loss for 2011 ,considering the same trend in the absenteeism pattern is shown above Below are the absent no.of hours for all the employees.

We can see that July has more no.of absent hours.

References

Scikit-Learn,
Edwisor,
Stackoverflow etc..