Employee Absenteeism

By C Sai Amogh

Contents:-

1. INTRODUCTION

1.1 Business Problem:- 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?

1.2 Metric for the problem

Mean square error:- measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated.

Root mean square error:- Square root of mean square error.

1.3 Data

Dataset Characteristics: Timeseries Multivariant
Number of Attributes: 21

1.4 Sampling techniques

1.4.1 Random sampling:- A simple random sample is a subset of a statistical population in which each member of the subset has an equal probability of being chosen here Training and testing is done by random sampling.

1.5 :- Mongo DB

Dataset has been stored in Mongo DB

2 METHODOLOGY

2.1 Data cleaning:-

2.1.1 Train and Test Dimensions

Dataset has 740 rows and 21 variables

2.1.2 Missing values

Dataset has missing values, following are the percentage of missing values in the variables

Body mass index	4.189189
Absenteeism time in hours	2.972973
Height	1.891892
Education	1.351351
Work load Average/day	1.351351
Transportation expense	0.945946
Disciplinary failure	0.810811
Hit target	0.810811
Son	0.810811
Social smoker	0.540541
Social drinker	0.405405
Age	0.405405
Service time	0.405405
Distance from Residence to wk	0.405405
Reason for absence	0.405405
Pet	0.270270
Weight	0.135135
Month of absence	0.135135
Seasons	0.000000
Day of the week	0.000000
ID	

2.1.3 Filling the Missing values

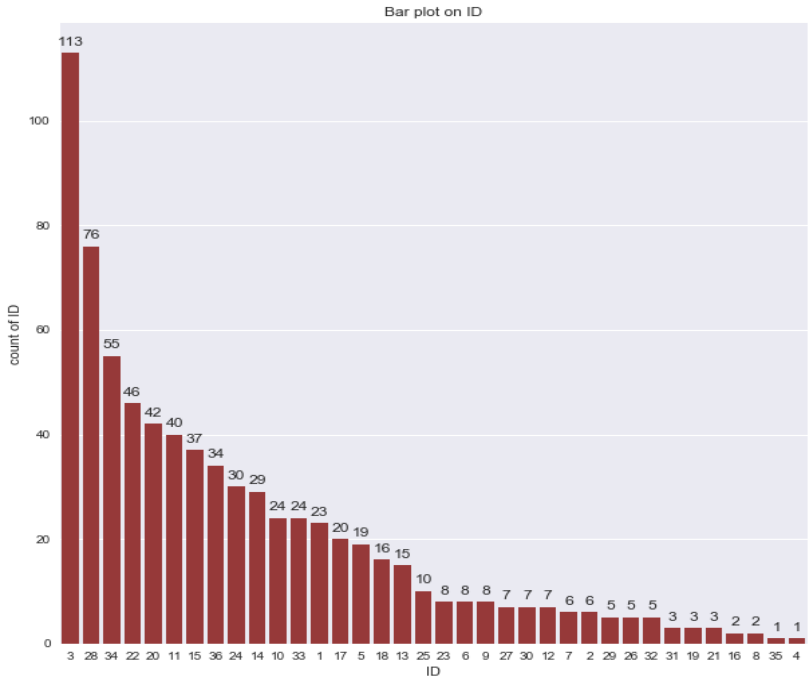
- In month of absence variable we find that there are 3 months with 0, I replaced 0 with frequent month by season(1,2,3).
- Filled Absenteeism time in hours variable by taking the mean of Absenteeism time in hours based on reason for absence.
- Work load avg/ day is filled by taking the mean with the help ID
- Disciplinary failure was filled with 0 (mode)
- Reason for absence was filled by mode
- Rest of the variables are filled by mode based on ID

2.2 EDA and Statistical analysis:-

2.2.1 Statistical analysis:-

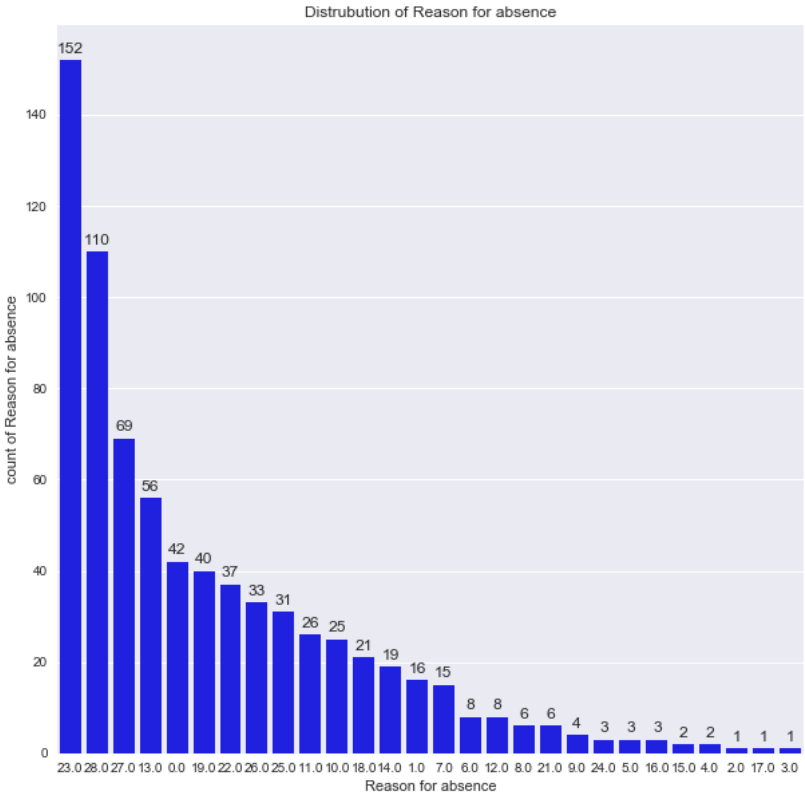
- Following are the findings after doing descriptive statistics:
- Absenteeism time in hour :- minimum is 0 which means that are people who are not absent and max in 120 hours, median 3, 75 percentile of people are absent for 8 hours.
- Body mass index:- max – 38 there are people with over weight
- Hit target:- mean - 94.4, max – 100
- Min :- 19, mean in 26.6
- Height:- max height is 196, min is 163, mean and median – 172
- Age – 27 to 58, most of the people in the dataset are around 26-27

2.2.2 Bar plots:-

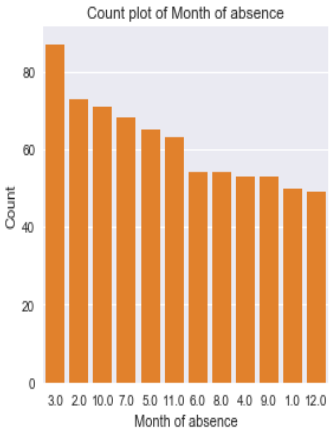


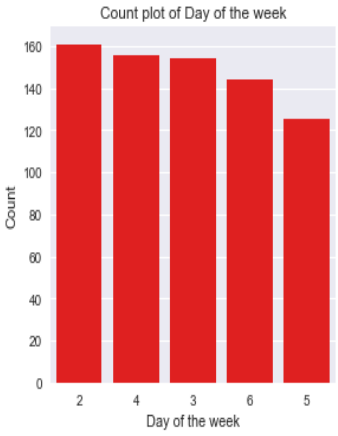
From the above plot **:-** Id how is a absent for the highest time ?

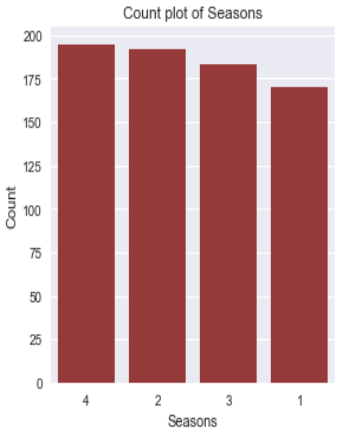
3(113 times) followed by 28(78 times), 46(55 times)



Reason for absence:- 23 toped which is medical consultation followed by dental consultation (28) and physiotherapy (27) and lest is 3 (Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism) and 7 (I Diseases of the eye and adnexa)



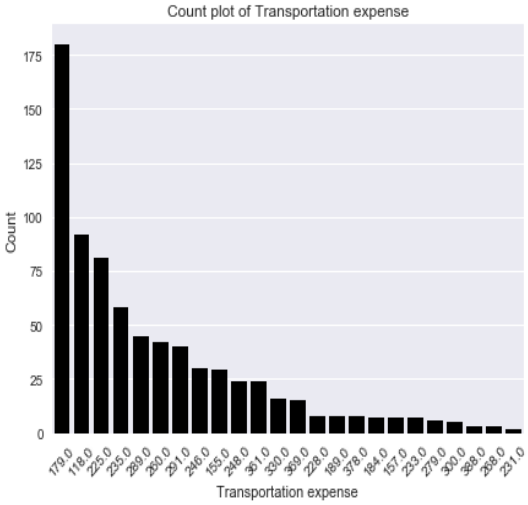


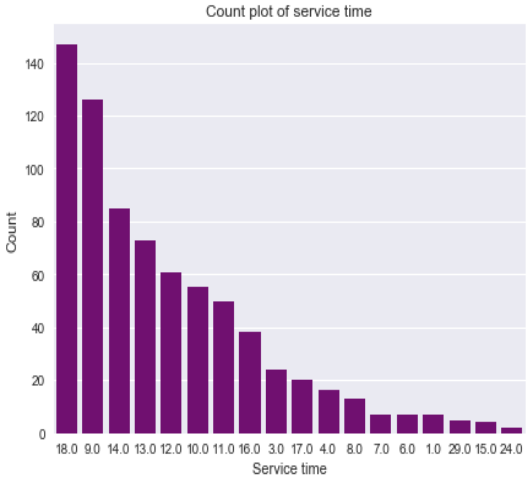


From the above plots we can see the following
Highest absentees found

- 3rd month,

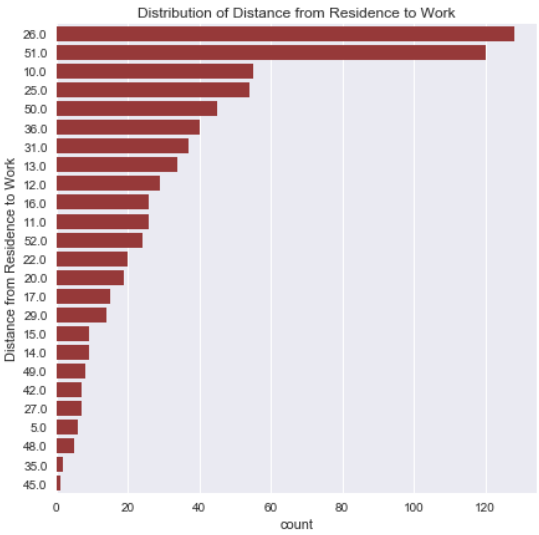
- 2nd day of the week followed 4th day of the week ,
- 4th season

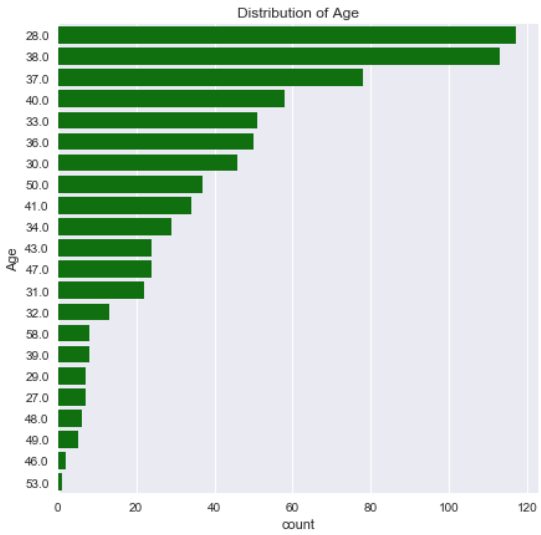


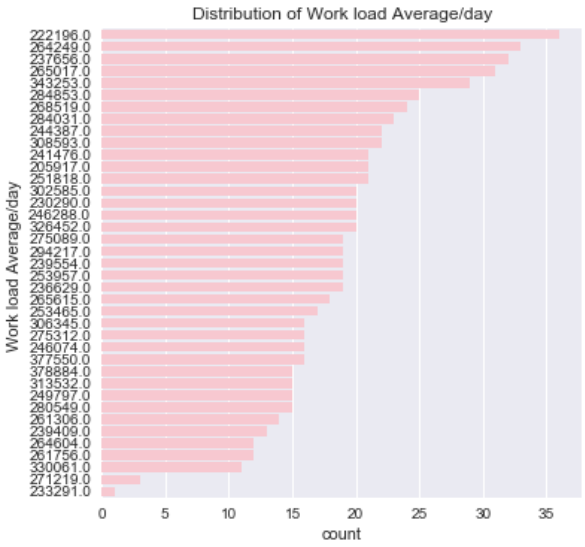


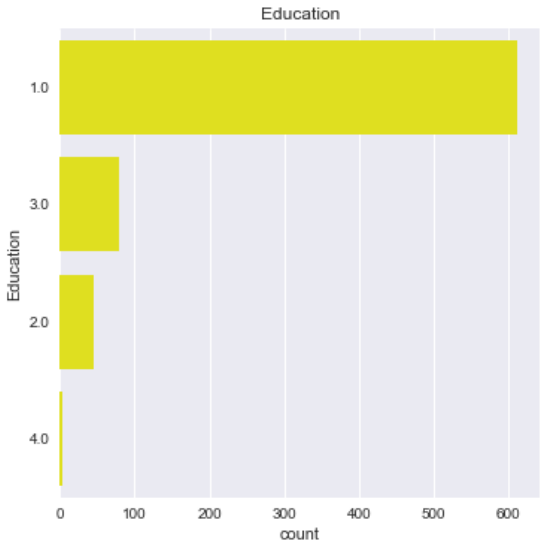
From the above plots we can find the following:-

- most of the people are working for 18 hours followed by 9 hours
- most of the people transportation cost is around 179 – 118



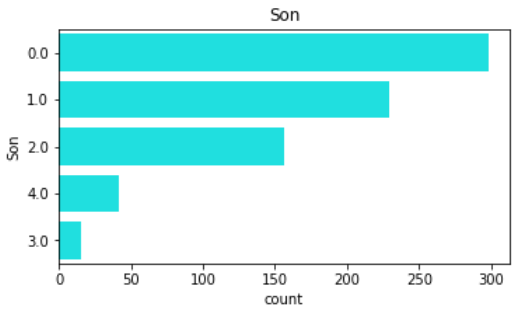


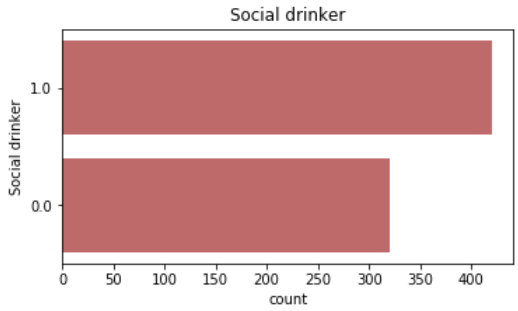


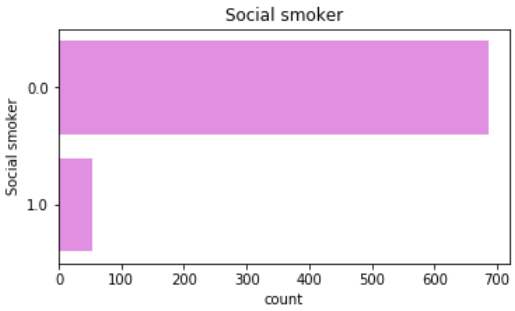


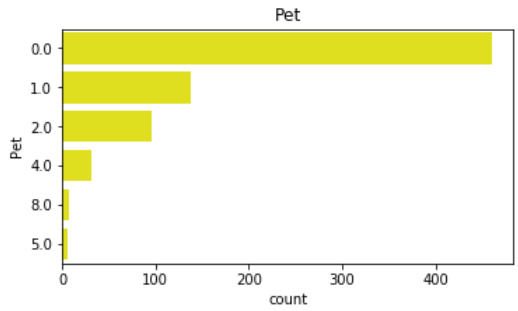
From the above plots we get the following:-

- There are more people with age around 28 and 38 and education with high school followed by masters
- Distance of travel is around 26 to 51
- work load per day is around 222196 followed by 264249

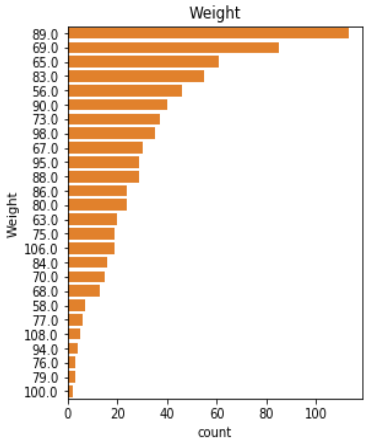


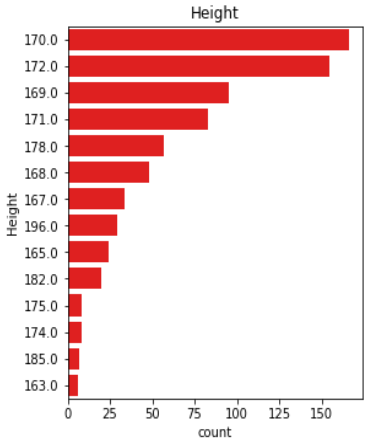


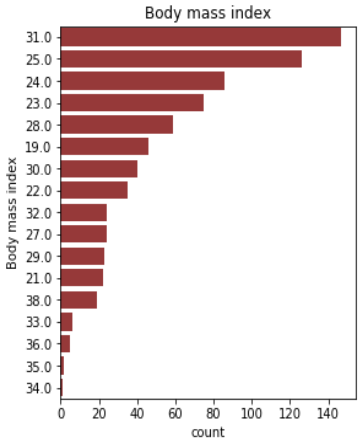




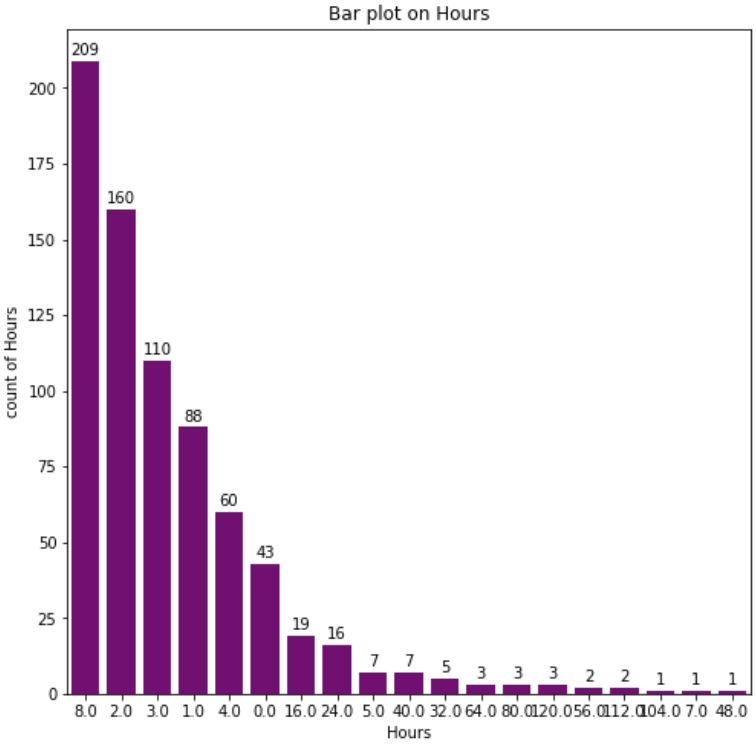
- Most of the people are social drinkers and most are not nonsmokers further most don't have pets
- Around 225 have 1 kid





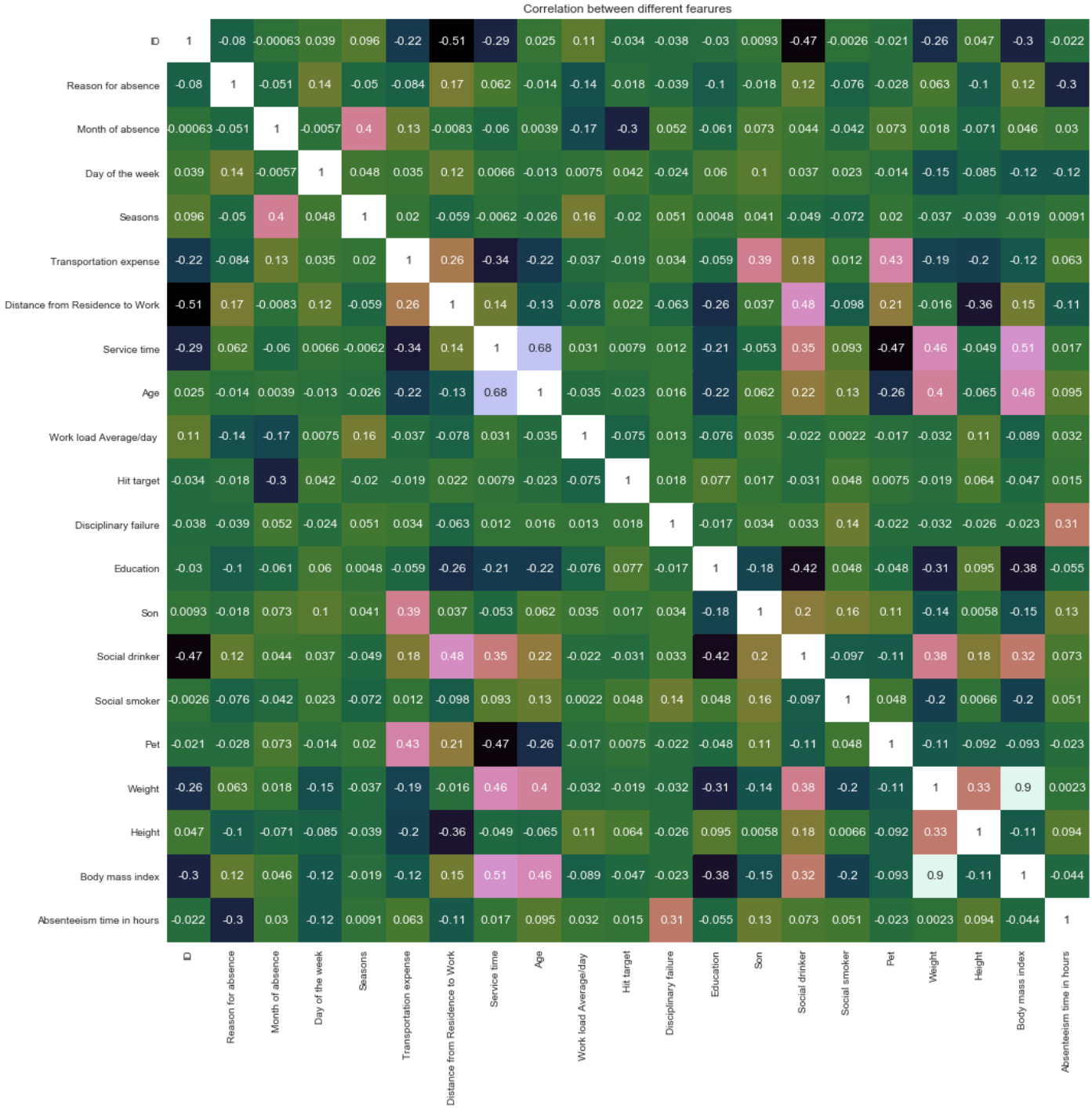


- highest frequent weight is around 89 followed by 69 and least is 100
- Body mass index few are obese as the index is 38,36 and most have around 25 to 31,
- height is 169-172 for the most and few have 196



- people are most absent for 8 hours followed by 2 and 120 is the max hours of absent
- one person was absent for 48,7,104 hours

2.2.3 Correlation between variables

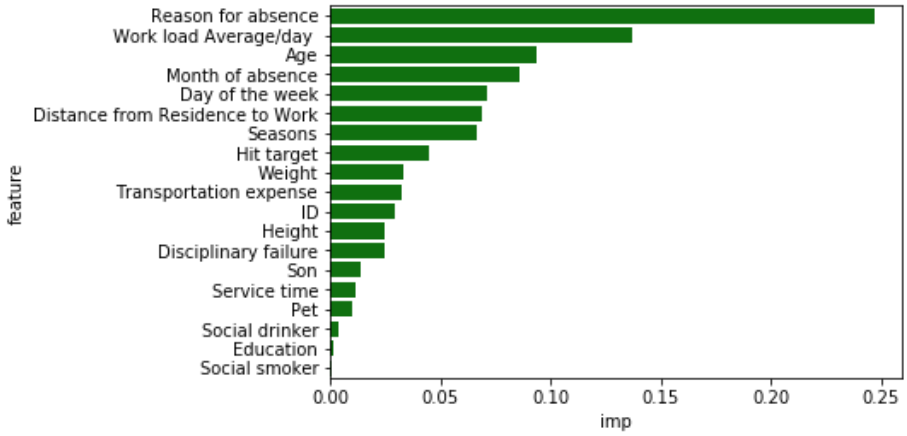


Following are from the correlation plot

- weight and body mass index are correlated and this is logical
- There is a 33 % correlation b/w weight and height
- there is a 38 % correlation b/w social drinkers and may be social drinkers have put on weight
- there is correlation of about 0.31,0.13 between absentees and delivery failures ,sons

2.3 Feature importance, Hypothesis Testing and aggregations

2.3.1 Feature importance by random forest.



2.3.2 Aggregations (group by):-

1. Group by reasons and median by distance form residence and work

Reason 27, has highest median of 51

2. Group by reasons and mean of service time

Reason 3 – mean of service time is 22 hours

3.Group by reasons and mean of social drinker

Reason 3, 24, 15 have social drinker 1

4.Group by reasons and mean of work load average/day

Reason 17 has the highest work load per day

5.Group by reasons and service time

Reason 4 has the highest service time

6. Group by reason for absence and Age

Reason 4 has the highest age

7. group by month of absence , son and mean on hours

We can see that month 7 with 3 sons have highest absent hours

8 group by day of the week, social drinker and mean absent hours

2nd day of the week social drinker 1 has mean of absent hours of 11

9 group by reasons and body mass index of mean > 30

Reason 27 has the effect of body mass index

2.3.3 Ztest after group by analysis

1) Ztest on distance form residence to work and absenteeism time in hours on reason 27(physiotherapy)

Null:- There is no effect of residence to work and absenteeism time in hours on reason 27 (means are same)

Alternate:- There is an effect of residence to work and absenteeism time in hours on reason 27

(17.47147422452035, 2.3630900461602933e-68)

As the p score is less we reject the null and accept the alternate hypothesis

2 Null:- There is no effect of Body mass index and Absenteeism time in hours on reason 27(physiotherapy) (means are same)
Alternate:-There is an effect of Body mass index and Absenteeism time in hours on reason 27(physiotherapy)
(42.801864116966485, 0.0)

As the p score is less we reject the null and accept the alternate hypothesis

3 Null:- There is no effect of Work load Average/day and Absenteeism time in hours on reason 13(tissue) (means are same)
Alternate:-There is an effect of Work load Average/day and Absenteeism time in hours on reason 13(tissue)
(44.93932872423466, 0.0)

As the p score is less we reject the null and accept the alternate hypothesis

4 Null:- There is no effect of Son and Month of absence (means are same)
Alternate:-There is an effect Son and Month of absence

(-40.35683995856594, 0.0)

As the p score is less we reject the null and accept the alternate hypothesis

5 Null:- There is no effect of Social drinker and Day of the week
Alternate:-There is an effect of Social drinker and Day of the week

(-60.47717574687599, 0.0)

As the p score is less we reject the null and accept the alternate hypothesis

2.3.4 suggestions

1) Reason for absent

23 (blood donation) so most of the works are going for blood donation
solution: - even though blood donation is a good thing, company has to bring a policy that blood donation should be done on weekends which are off and if there is an emergency for blood to be given a small statement from victims family or recent has to be shown. on an average about 3 ho urs of work is lost

28 dental consultations
A circular should be given on dental consultation as dental condition is not a serious disease cons ultation should be done on weekends further exceptional cases with poofs should be given permi ssion. almost 3hr is lost on average per worker

27 physiotherapy

Distance from the home, higher body mass index so next Time Company should hire workers wh o are less than 35 to 40 km distance from the office, morning or evening yoga or physical exercis e programs should be run

13 Diseases of the musculoskeletal system and connective tissue

Solution: on average heavy work on per day is the higher s for this cause; company has to take pr ecautions that work has to be distributed properly.

lets look at the most mean absent cause

These are the diseases where rest is needed

9 Diseases of the circulatory system

2 Neoplasms
12 Diseases of the skin and subcutaneous tissue
6 Diseases of the nervous system

unjustified absence
(25), unjustified absence :- unjustified absence should not be entertained

2) Sons and seasons

- 6 and 7th month is a season effect .. company can reduce work load and hit target in 7th month
- further people with kids are taking leave of absent and reason for absence should be cross verified properly every 7^{th} month of year

3) day of the week

- If workers are taking holidays for unwanted reasons on Monday, they have to work on Saturday company should introduce this policy.
- Further few of social drinker may be absent on Monday and cross checked with reasons.

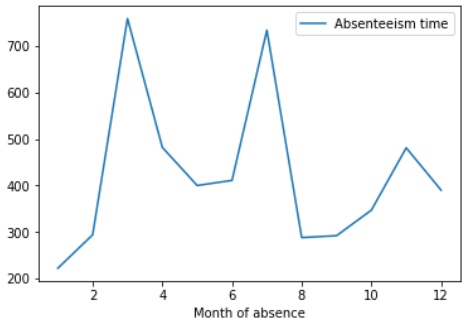
4) ID:- Few people are getting absent frequently they should be cautioned for the absent days so that they don't repeat the same.

3 Modeling

2.4.1 Month vs. absentees using tree based models

Group by Month and count by absent hours

Line plot



I applied following algorithms on the above data points

Decision trees, Random forest , GBM

Did not divide into train and test dude to 12 data points only All the three where perfect hours on month vs. absent fit with 0 RMSE

2.4.2 Applying machine leaning on Features provided.

Training and testing datasets was done by random sampling

Kernal SVM:- support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used
for classification and regression analysis

Kernal RBF which is most frequent used kernel for non linear problems

Hyper turning for hyper parameters was done by grid search , cv of 5

Following are the results for c = 1/alpha, gamma

{'C': 13, 'gamma': 0.05}

Random forest:- Random forests or random decision forests are an ensemble learning method fo r classification, regression and other tasks.

Hyper tuning of parameters is done by grid search the following are the results :-
{'n_estimators': 350}

{'max_depth': 3}
{'min_samples_split': 20}
{'min_samples_leaf': 10}
{'max_leaf_nodes': None}
{'max_features': 'auto'}

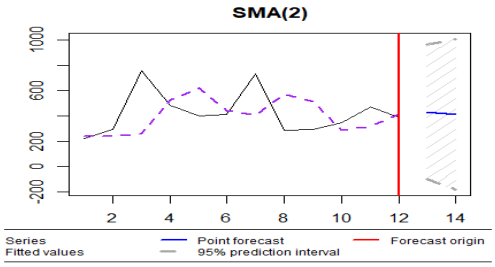
Gradient boosting methods : is a machine learning technique for regression and classification p roblems, which produces a prediction model in the form of an ensemble of weak prediction mode ls, typically decision trees. It builds the model in a stage-wise fashion like other boosting method s do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Hyper turning was done with grid search :-

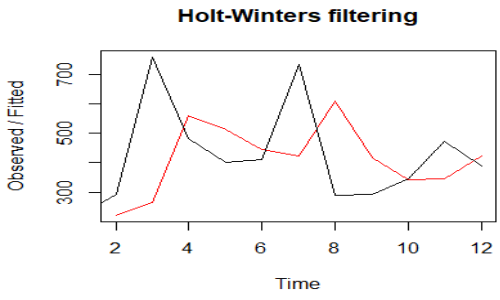
{'n_estimators': 10}
{'max_depth': 3}
{'min_samples_split': 2}
{'min_samples_leaf': 30}
{'max_leaf_nodes': None}
{'max_features': 'auto'}

2.4.3 Time series analysis:-

Simple moving averages:-
Order or window is 2

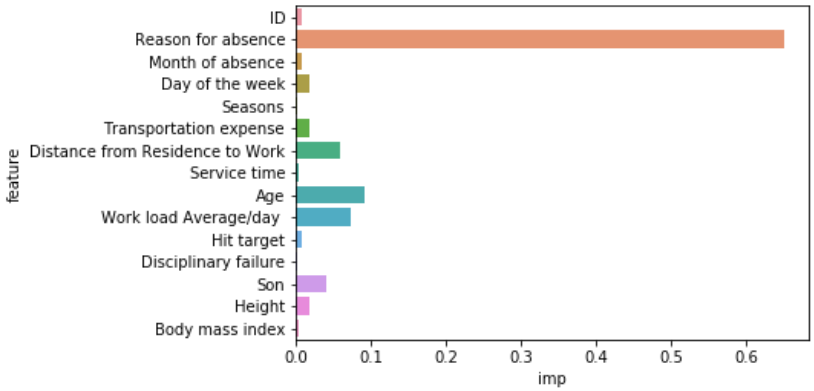


SES:-
Alpha = 0.6

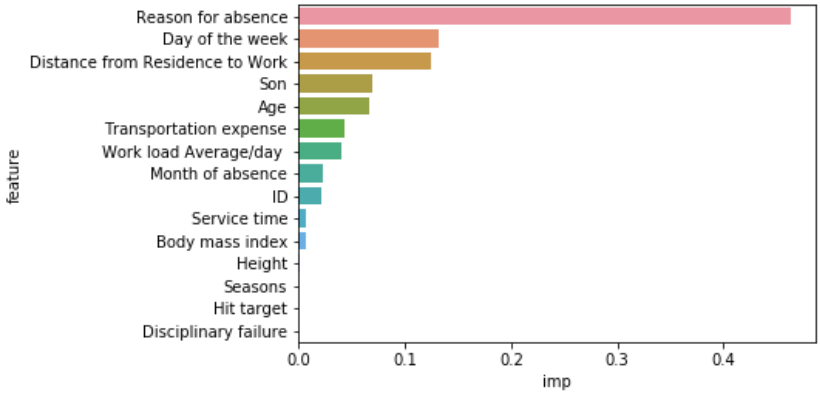


2.4.4 Feature importance from machine leaning for business purpose:-

Random forest:-



GBM:-



3 Results and conclusions

3.1 Results:-

RMSE for months vs. absentees is 0 on the original dataset is 0,

Machine leaning on features:-

SVM:-

Mean Square error:- 282.08

RMSE :- 16.7

Random forest:-

MSE:- 160.3

RMSE :- 12.6

GBM:-

MSE:- 95

RMSR:- 9.7

Time series:-

SMA:-

RMSE :- 162

SES:-

RMSE:- 134

Conclusions:- I was able to successfully suggest the changes to be made by the company and was able to predict the absentees in house using machine leaning with features where GBM performed well and performed month vs. hours by machine leaning and time series, in time series SES performed well.