***Project Report***

***Employee Absenteeism***

**“Akshay Verma”**

**======11th February 2019========**

**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?

**Dataset Details:**

Dataset Characteristics: Timeseries Multivariant

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

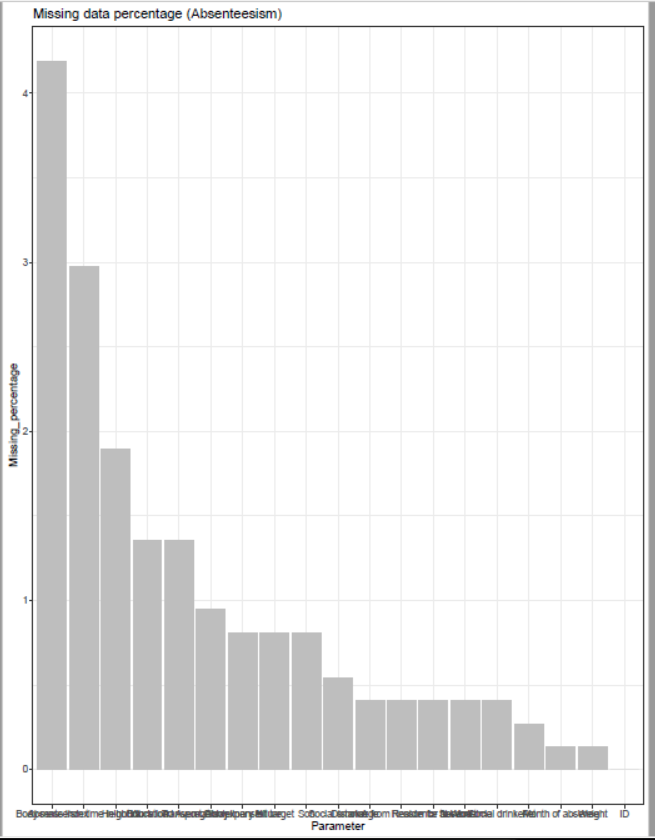
21. Absenteeism time in hours (target)

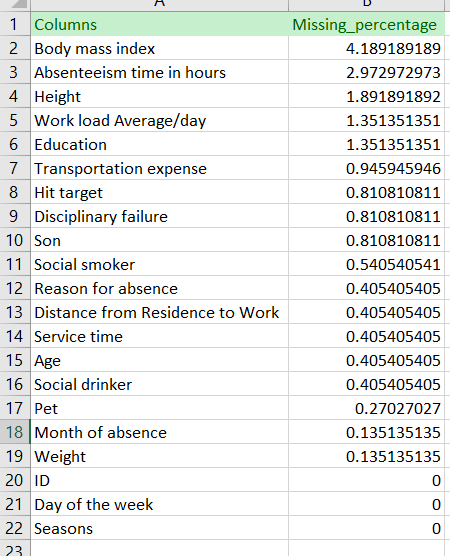
Now we will explain the pre – processing steps one by one and will apply the codes in R and Python..

1. **Missing Value Analysis**: Missing values are the values which are blank in dataset. Missing values occurred because of many reasons such as skipping few of questions while filling surveys and not answering few questions.. If the dataset is huge the chances of missing value there is high and our first step after understanding the data is to find the missing values and impute accordingly in order to meet the accuracy..

Below is the bar graph of missing values that shows how many missing values are there in each variable of dataset..

I have also pasted the and extracted missing value data by percentage..

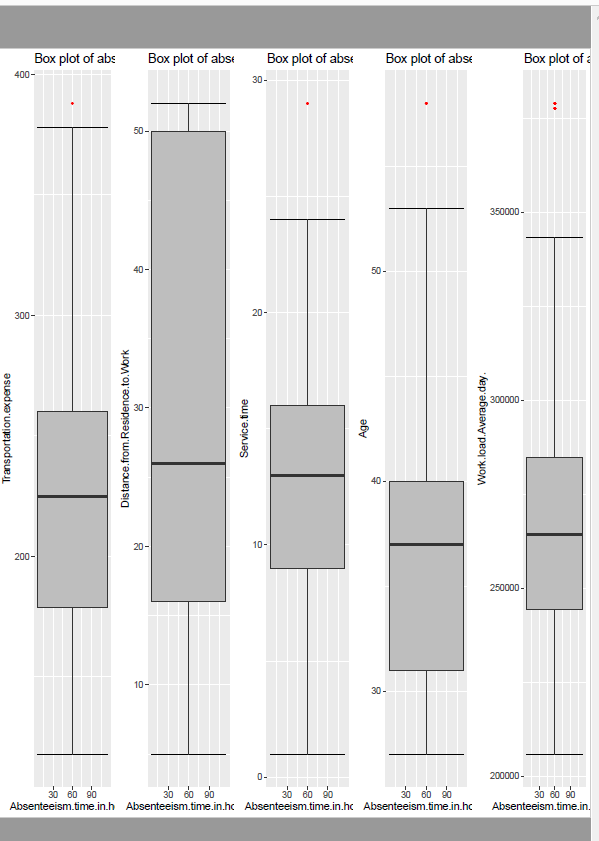


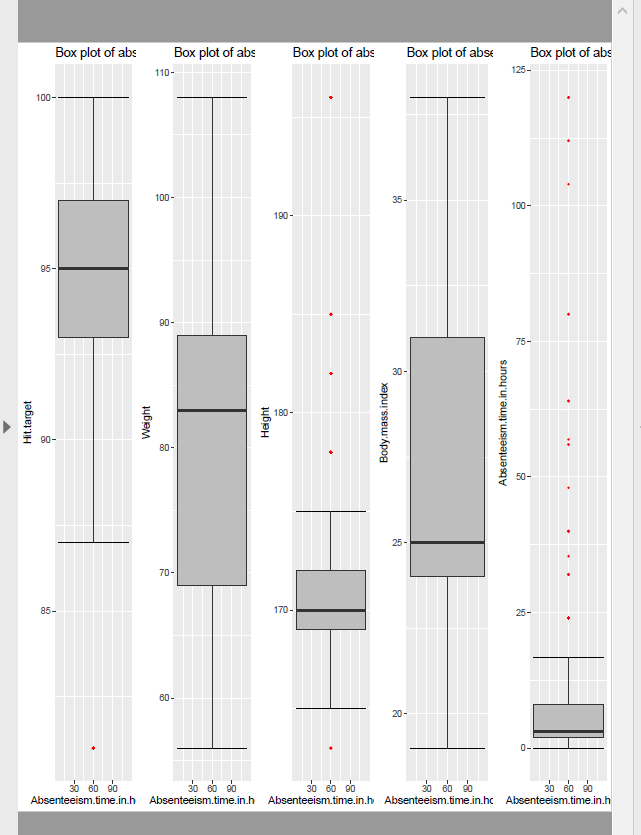


And then we can fill the missing value using knn imputation , median and mean method. Here in this project I have used KNN imputation in R and median method in python for comparing accuracies and RMSE..

1. **Outlier Analysis**: In statistics, an **outlier** is an observation point that is distant from other observations. An **outlier** may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An **outlier** can cause serious problems in statistical **analysis.**

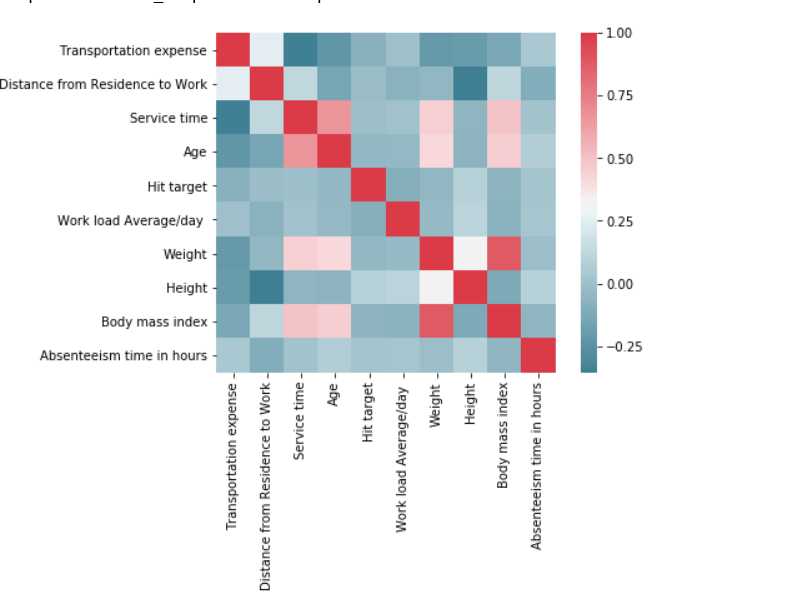
So , below are the outlier plots of each numerical variable using R..





So from our project we have 10 continuous variables and all the boxplots for outlier analysis have been plotted above.. And then we have removed outlier and replaced them with NA and we have used KNN imputation and median method for R and python respectively...

1. **Feature selection** : It becomes imperative to analyze ad check each variable and its importance to dataset as many times it is possible that a particular variable is not explaining or not contributing that much to a data set and to find that variable or variables we will use feature selection as a part of data pre processing technique.. For continuous variable we have used correlation plot and for categorical we have used ANOVA test.. Below is the plot heat map which we have extracted from python..

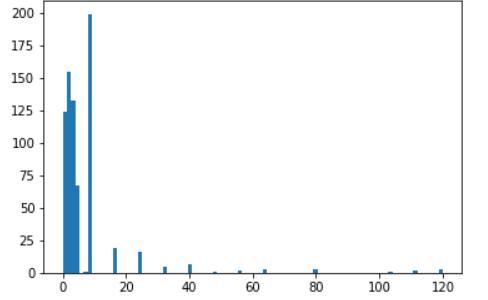


Weight and Body mass index are positively correlated to each other..

1. **Feature scaling** : In order make data in same range and to avoid biasing of output we will use feature scaling as a part of data pre processing technique.. Feature scaling is of 2 type

* Normalization
* Standardization

Here in this project we have use Normalization as a part of feature scaling because our data is not normally distributed..



1. **Sampling techniques**: Now our data is ready to work, but before that we will need to divide the data into train and test.. As our target variable is continuous so we will use **Simple Random Sampling** to sample out data into train and test..

Now after dividing data into train and test we will move towards selecting the model..

**Model Development**

As our target variable is continuous , so it is basically a Regression problem.. So here we have used Decision tree for Regression, Linear Regression and Random Forest to develop our model.. Now we will talk about all the models one by one..

* **Decision Tree for Regression** : After doing the codes in R and python, we have found below Root mean square value. Here Root mean square(RMSE) plays a important role as it is a time series data and for time series data we will consider RMSE as our parameter to find accuracy.

|  |  |
| --- | --- |
| R | Python |
| RMSE = 0.14 | RMSE = 0.12 |

* **Linear Regression Model** : Below is the RMSE calculated in R and python

|  |  |
| --- | --- |
| R | Python |
| RMSE= 0.15 | RMSE = 0.12 |

* **Random forest for Regression** : Below is the RMSE calculated in R and python..

|  |  |
| --- | --- |
| R | Python |
| RMSE = 0.13 | RMSE =0.11 |

So the values of RMSE are pretty close enough. But here we will choose **Random Forest as our model for further prediction..**

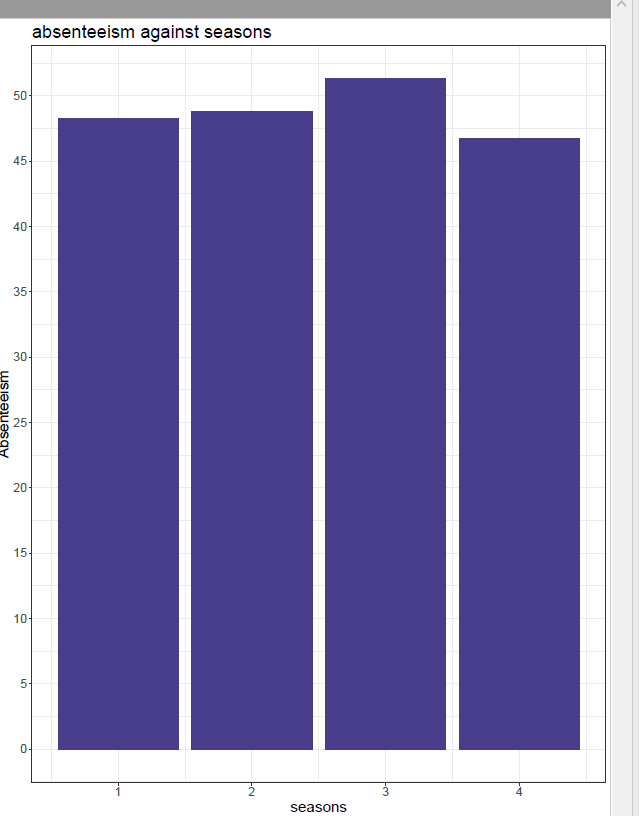
Now we will use some visualization graphs in order to answer the questions which are asked in project.

Answer 1) **Changes Company should bring in order to reduce Absenteeism**:

The graph tells you all about this.. we will plot the bar graphs keeping target variable on Y – Axis and all other categorical variable one by one on X – axis..

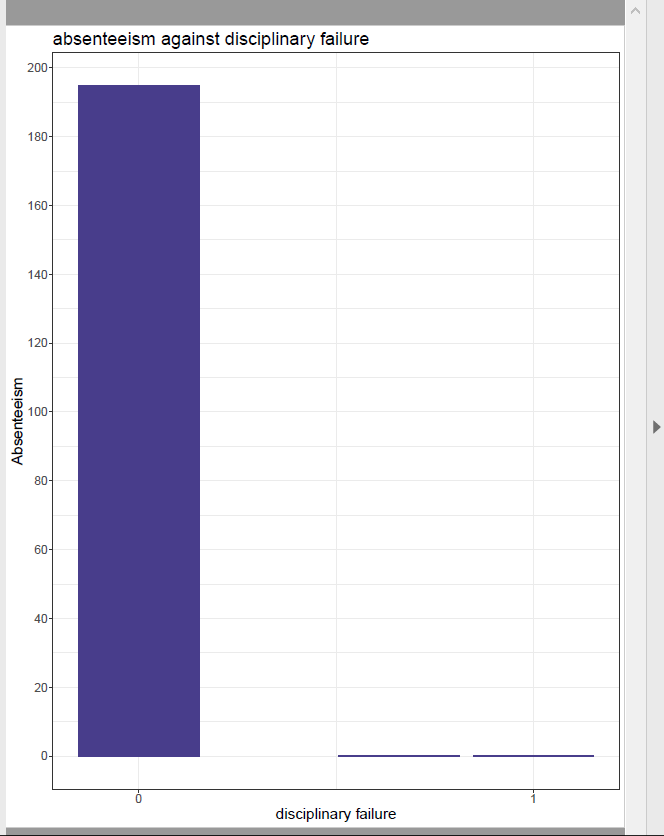
***Absenteeism V/s Seasons:***

From the graph it is confirmed that season is not playing any role in reducing or increasing absenteeism..



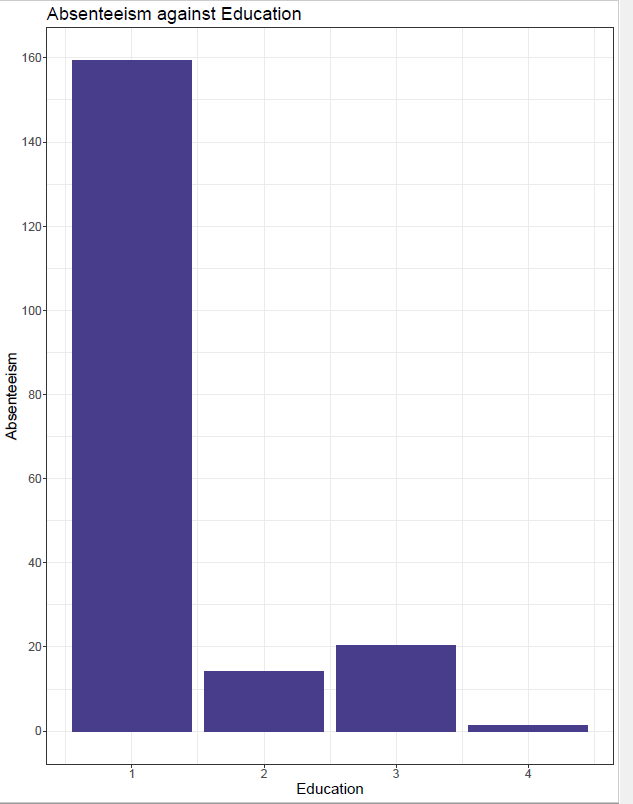
***Absenteeism V/s Disciplinary Failure :***

It is clear that Disciplinary failure of 0 has a complete rise in absenteeism..

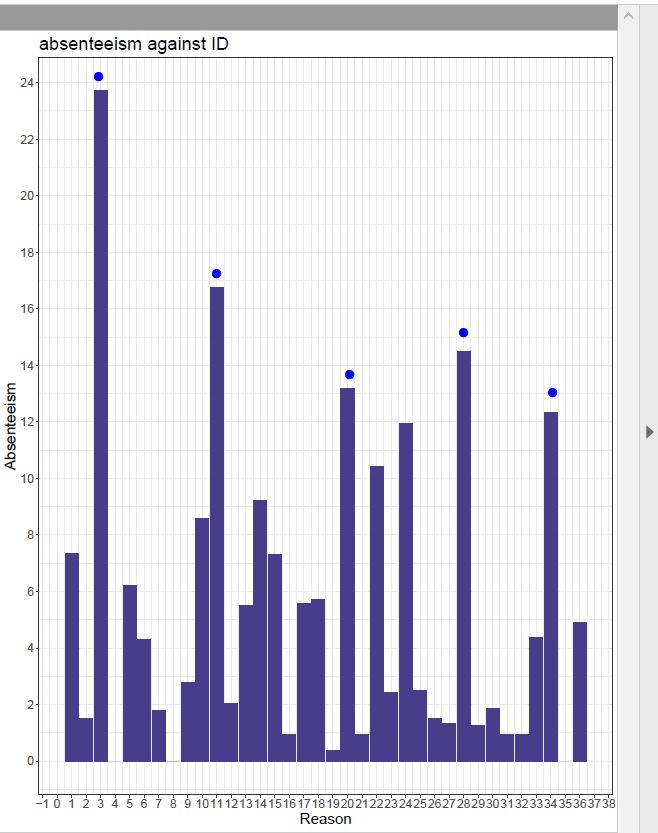


Company should adhere to there guidelines and follow the protocols which are there to follow so that employee also follow the same..

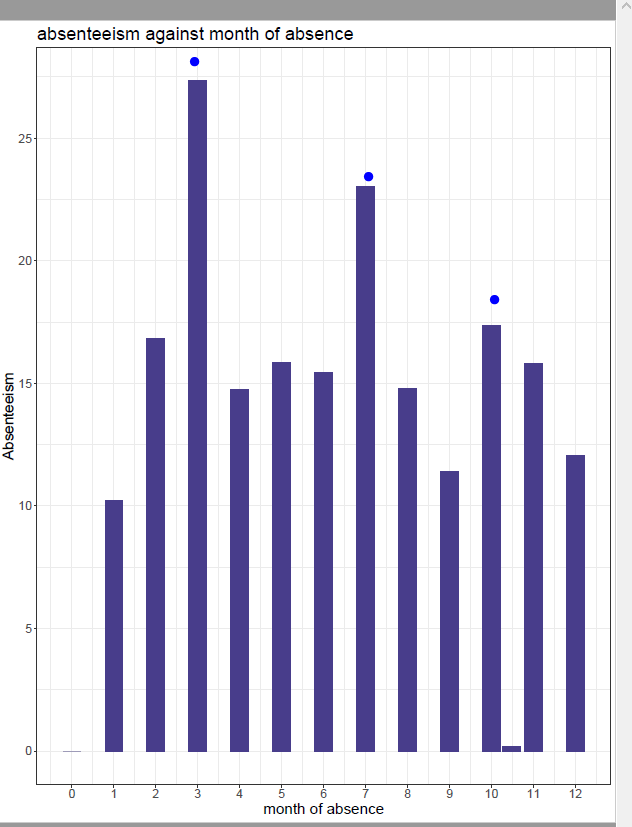
***Absenteeism V/s Education :*** The Graph shows that the absenteeism rate of those employees are high who comes under Education 1 category. So company should hire people who are graduates , post graduates and master and doctor..



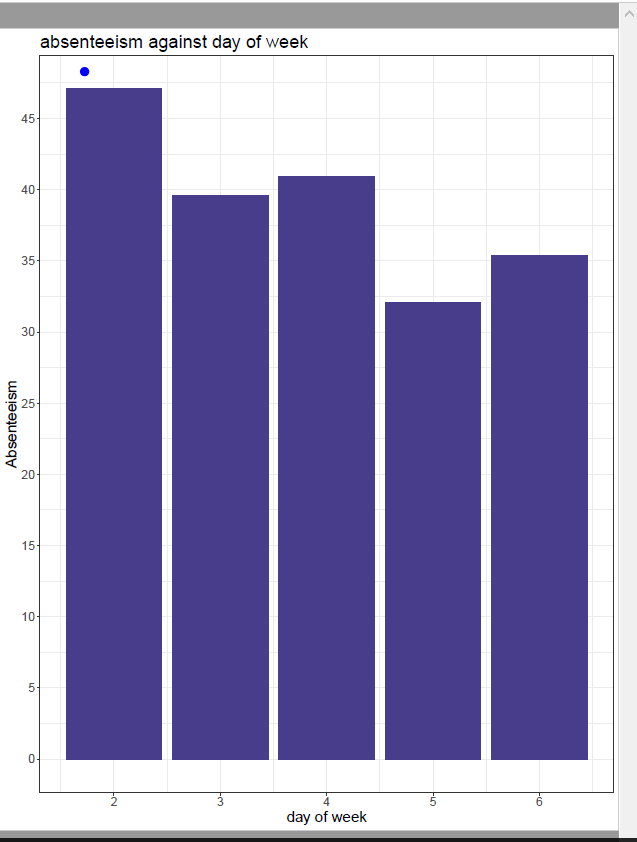
***Absenteeism V/s ID*** : It has been seen that ID (3,11,20,28,34) have higher absenteeism rate compared to others.



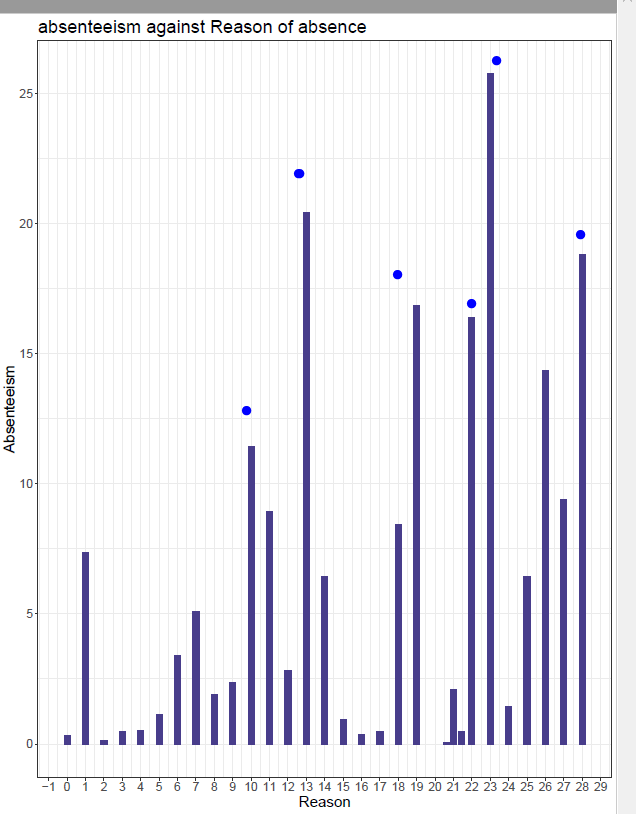
***Absenteeism V/S Month***: Month 3,7 and 10 have high absenteeism rate. Company can improve there audit during this period..



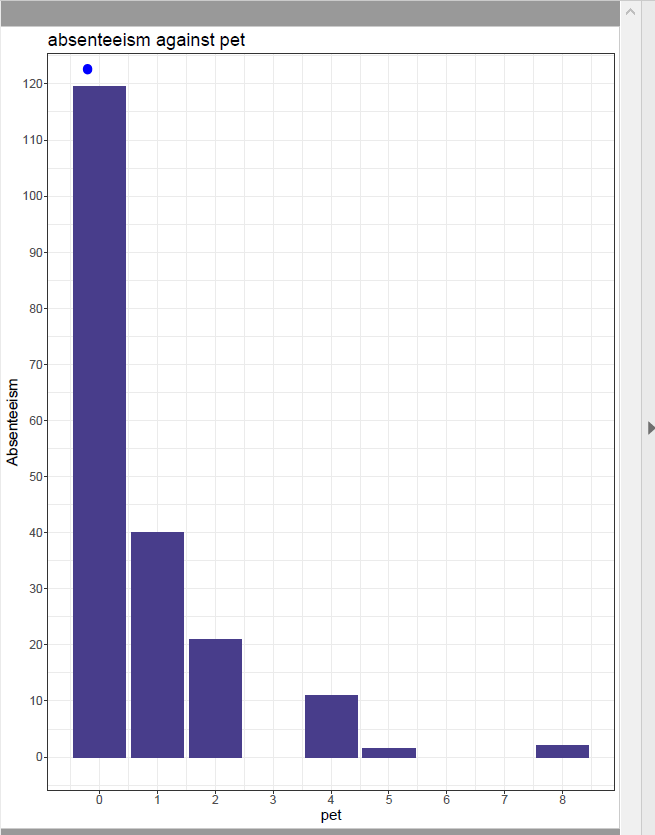
***Absenteeism V/s day of week :*** It seems Monday is the problem.. So after weekend if the employ is having off then there is a chance that absent hours will increase at the start of next week i.e Monday..



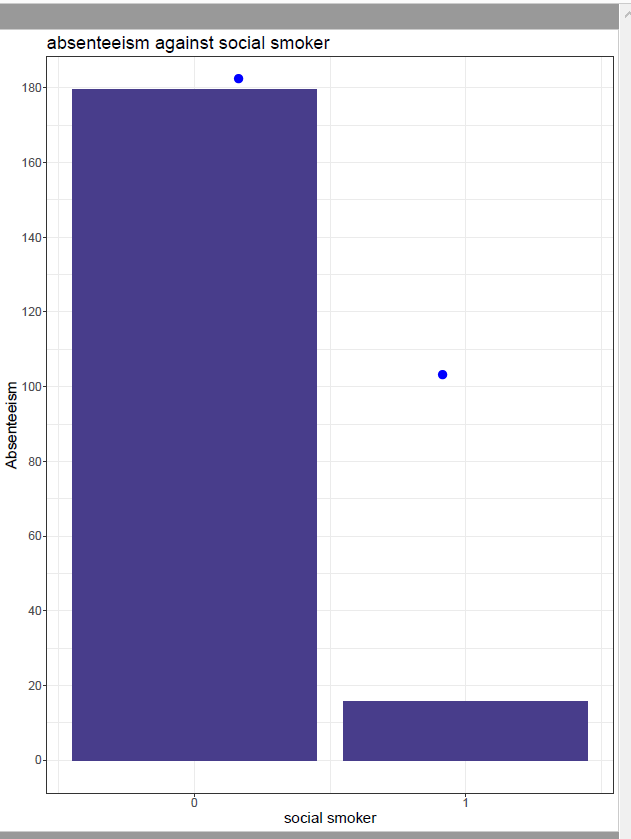
***Absenteeism V/s Reason*** : 13,19,22,23(highest),28 categories have the highest rate of absenteeism. As these are all medical related issues. Company can bring an action play to introduce free medical check ups inside the company campus..

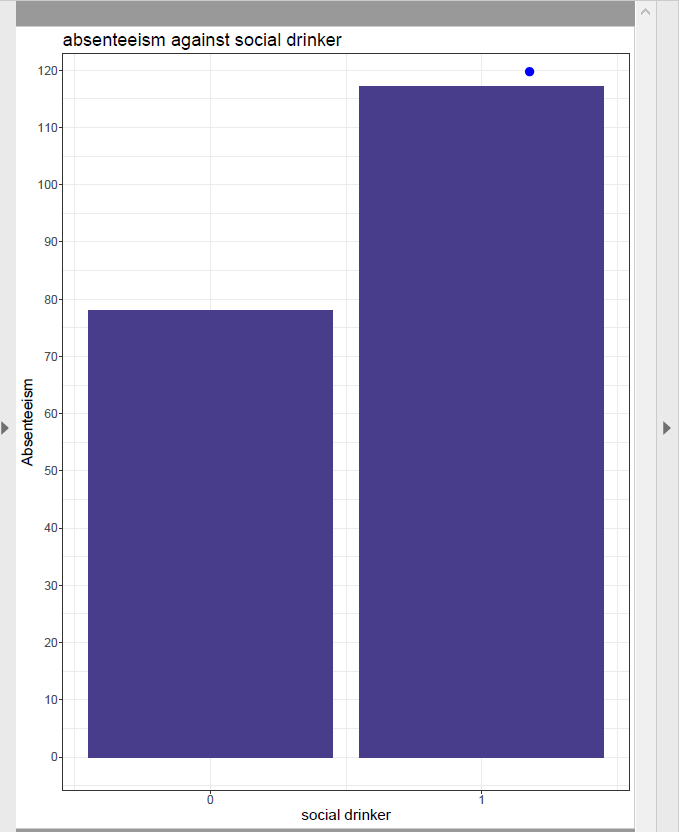


***Absenteeism V/s Pets*** : No pets refers to high absent rate. Company can d nothing about this. This is just a statistical data.



***Absenteeism V/s Social drinker and Social Smoker*** : It has been observed that Social drinker has high absent rate . Company can release some programs to guide them the ill effects of drinking and make an audit regarding this in order to reduce these things.. But contrasting to this social smokers have opposite features.





So by curbing above issues, company can reduce the absenteeism rate and it could be a big respite for them in coming years in terms of profit as well as Human capital plays an important role to this courier company..

And if the same trend continues then the company will have to suffer a serious loss in terms of money as well as it is a courier company and human capital will play a key role in making this company big..

Thanks..

Note : The code written in R and Python will be separately attached..