Chapter 1

1. Introduction:-

1.1. Problem Statement:-

XYZ is a courier company. The company is passing through genuine issue of Absenteeism. The company has shared its 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 Hypothesis Generation:-

1) Miscellaneous: - It is commonly expected that low pay, poor benefits and high workloads will be the major causes. However, in numerous employee surveys absenteeism generally has been identified as a symptom of low job satisfaction, sub-standard working conditions and consistent negative and unfair treatment received by first-line supervisors.

2) Management behavior's issue:- If they find their supervisor or job unpleasant - really unpleasant then they look for legitimate excuses to stay home and find them with things such as upset stomachs or splitting headaches.

3) Age: - Absenteeism is generally high workers below 30 years of age and those above 40. ? The younger employees are not regular and punctual?. Presumably because of the employment of a large no. of new comers among the younger age groups, while the older people are not able to withstand the strenuous nature of the work

The younger employees are often restless. They want to spend time with their friends and have fun, rather than being tied down with work responsibility. This lack of ownership often leads them to take unauthorized time off. With age, people gain experience and maturity, which makes them focused and responsible. Their approach is rather professional and they prefer to stick to their chairs to get the work done. If ever they are found absent, then it could be due to sickness.

4) Means of Transport: The Commission also stated that the transport facilities also play very important to contribute the absenteeism of the worker in the industries. It has been pointed out that, the rate of absenteeism is higher in those factories where transport facilities are not easily available as compared to those where such facilities are easily available or provided by the factory itself.

5) Drinking and amusement: Drinking and amusements are also responsible for absenteeism. Since drinking and amusements in the late hours of night make it difficult for the workers to reach in time on their duties. They like to become absent rather than late since they know that.

6) Workload Avg Per day: -The pressure at work sometimes takes a toll on the employees. This results in increased levels of stress. The employees then resort to excuses that can help them stay away from work.

7) Infrastructure: - If there's not good infrastructure in company in terms of food counters in cafeteria or games in company then employees tend to go out for lunch, snacks or dinner or to play games.

7) Job Satisfaction: If employees do not find their job challenging, dissatisfaction creeps in. That leads to more absenteeism in the workplace.

1.4 Data:-

The dataset shows 12 months of data . The data has 740 observations and 21 features."Absenteeism.time.in.hours" is our target variable. We are required to suggest following based upon the various features available in dataset-

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?

Employees's Absenteeism sample data:-

```
> glimpse(data)
Observations: 740
Variables: 21
$ ID                              <int> 11, 36, 3, 7, 11, 3, 10, 20, 14, 1, 2
0, 20, 20, 3, 3, 24, 3, 3, 6,...
$ Reason.for.absence              <dbl> 26, 0, 23, 7, 23, 23, 22, 23, 19, 22,
1, 1, 11, 11, 23, 14, 23, 21...
$ Month.of.absence                <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7
, 7, 7, 7, 7, 7, 7, 8, 8, 8, ...
$ Day.of.the.week                 <int> 3, 3, 4, 5, 5, 6, 6, 6, 2, 2, 2, 3, 4
, 4, 4, 6, 6, 2, 5, 4, 4, 2, ...
$ Seasons                         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ Transportation.expense          <dbl> 289.0000, 118.0000, 179.0000, 279.000
0, 289.0000, 179.0000, 221.03...
$ Distance.from.Residence.to.Work <dbl> 36, 13, 51, 5, 36, 51, 52, 50, 12, 11
, 50, 50, 50, 51, 51, 25, 51,...
$ Service.time                    <dbl> 13, 18, 18, 14, 13, 18, 3, 11, 14, 14
, 11, 11, 11, 18, 18, 16, 18,...
$ Age                             <dbl> 33, 50, 38, 39, 33, 38, 28, 36, 34, 3
7, 36, 36, 36, 38, 38, 41, 38...
$ Work.load.Average.day           <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8
, 8, 8, 8, 8, 8, 8, 2, 2, 2, ...
$ Hit.target                      <dbl> 97, 97, 97, 97, 97, 97, 97, 97, 97, 9
7, 97, 97, 97, 97, 97, 97, 97...
$ Disciplinary.failure            <chr> "0", "1", "0", "0", "0", "0", "0", "0
", "0", "0", "0", "0", "0", "...
$ Education                       <chr> "1", "1", "1", "1", "1", "1", "1", "1
", "1", "3", "1", "1", "1", "...
$ Son                             <chr> "2", "1", "0", "2", "2", "0", "1", "4
", "2", "1", "4", "4", "4", "...
```

```
$ Social.drinker                  <chr> "1", "1", "1", "1", "1", "1", "1", "1
", "1", "0", "1", "1", "1", "...
$ Social.smoker                   <chr> "0", "0", "0", "1", "0", "0", "0", "0
", "0", "0", "0", "0", "0", "...
$ Pet                             <chr> "1", "0", "0", "0", "1", "0", "4", "0
", "0", "1", "0", "0", "0", "...
$ Weight                          <dbl> 90, 98, 89, 68, 90, 89, 80, 65, 95, 8
8, 65, 65, 65, 89, 89, 67, 89...
$ Height                          <dbl> 172, 178, 170, 168, 172, 170, 172, 16
8, 196, 172, 168, 168, 168, 1...
$ Body.mass.index                 <dbl> 30.00000, 31.00000, 31.00000, 24.0000
0, 30.00000, 31.00000, 27.000...
$ Absenteeism.time.in.hours       <dbl> 4.000000, 0.000000, 2.000000, 4.00000
0, 2.000000, 6.977716,
```

Predictor Variables:-

As we can see in the table below we have the following 21 variables/features, using which we have to provide solution

"ID"	"Reason.for.absence"	"Month.of.absence"	"Day.of.the.week"
"Seasons"	"Transportation.expense"	"Distance.from.Residence.to.Work"	"Service.time"
"Work.load.Average.day	"Age"	"Hit.target"	"Disciplinary.failure"
"Education"	"Social.smoker"	"Son"	"Social.drinker"
"Pet"	"Weight"	"Height"	"Body.mass.index"

Chapter 2

Methodology

2.1 Data Pre Processing or Data Exploration:-

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis. To start this process we will first try and look at all the histograms of the variables. Most analysis like regression, require the data to be normally distributed. We can visualize that in a glance by looking at the distributions of the

variable.

We should check for Missing values and Outliers in data, if there’re any.

1. Pet- We see that most of the people have no pet or 1 pet.

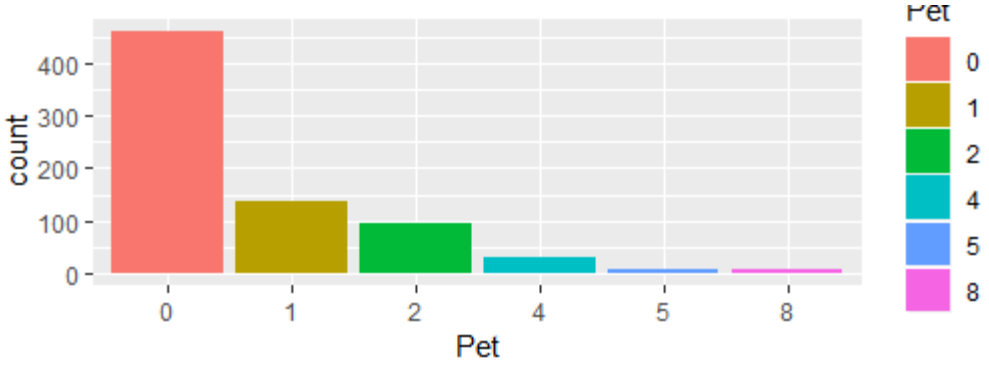
2. Son- We see that most of the people have no son or 1 son. But It’ll be safe to say that most of majority of employees are having 0 to 2 sons.

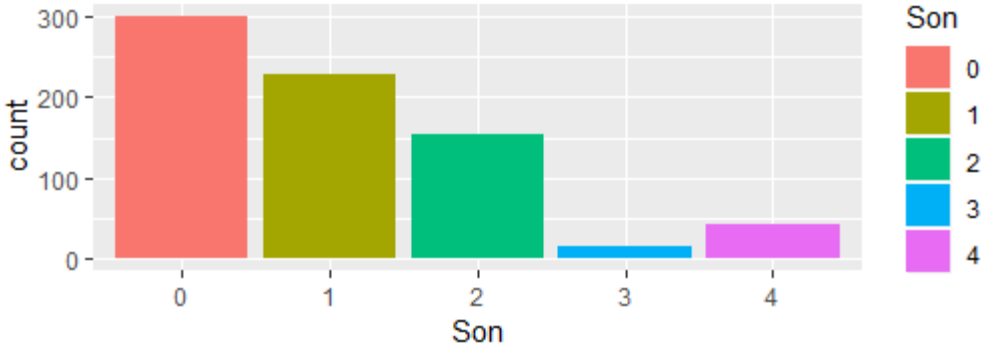
3. Social smoker V/S Social Drinker :- we see that people are not social smoker but they’re social drinker which adds in count of social drinker who are social smoker as well.

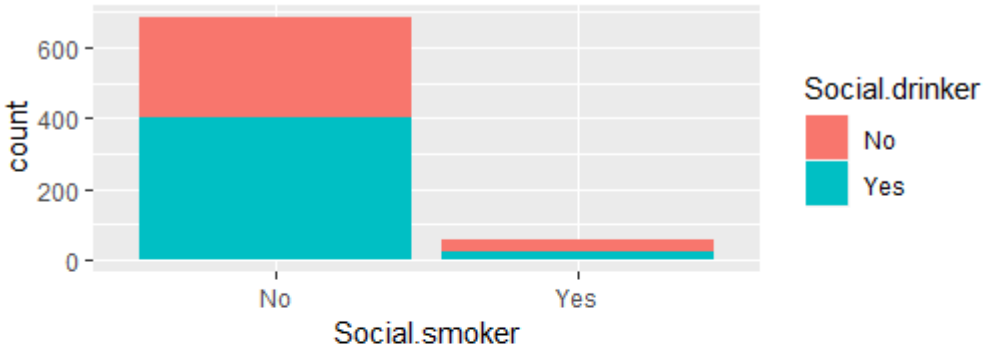
Reason for absence: - Here the various types of reasons for absence attribute is analyzed -

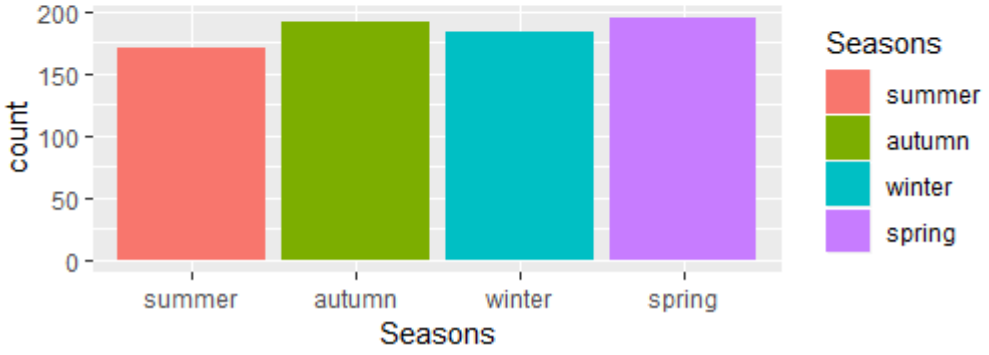
The top four of them cover 50% of the reasons for absence

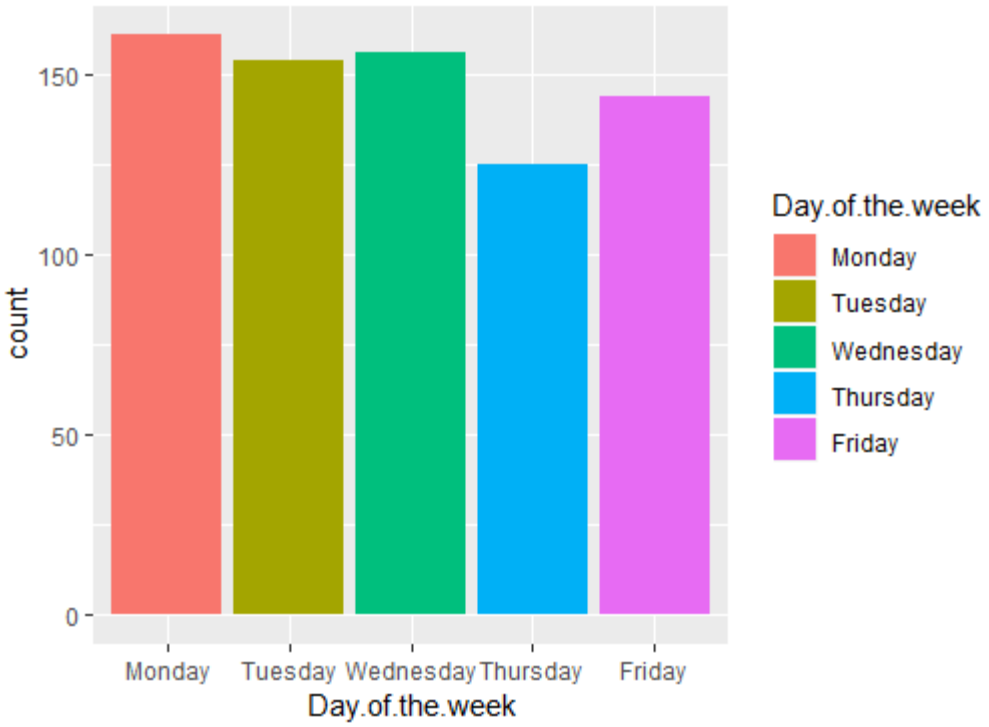
1) Medical consultation
2) Dental consultation
3) Physiotherapy
4) Disease of genitourinary system
5) The unjustified absence amounts to 4.7% of total.

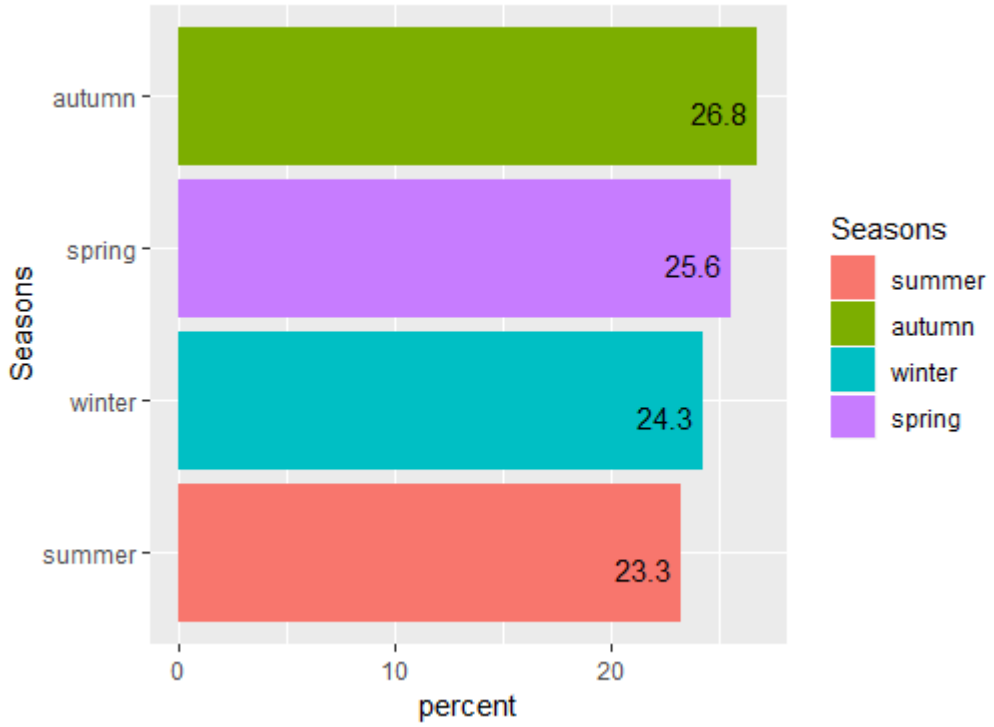

count
400
300
200
100
0
0
1
2
4
5
8
Pet
Pet
0
1
2
4
5
8

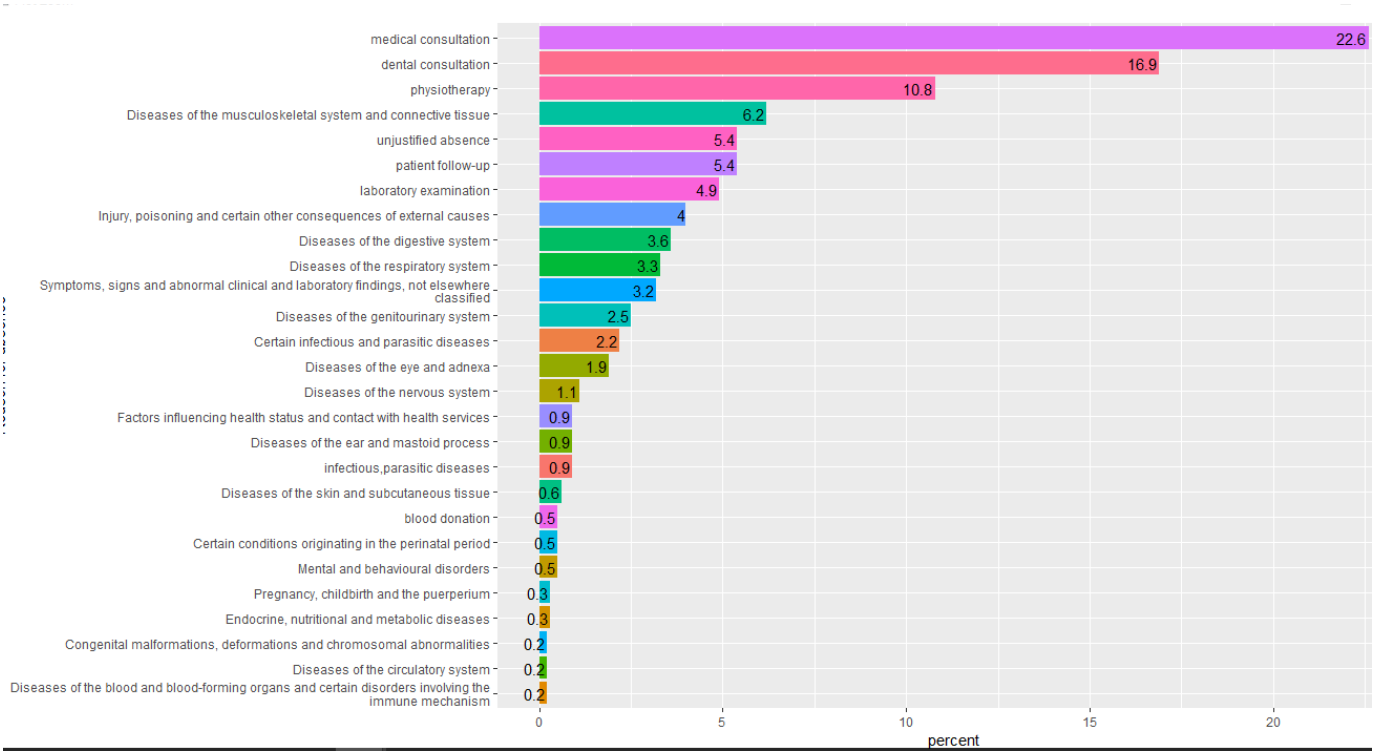

count
300
200
100
0
0
1
2
3
4
Son
Son
0
1
2
3
4


count
600
400
200
0
No
Yes
Social.smoker
Social.drinker
No
Yes


count
200
150
100
50
0
summer
autumn
winter
spring
Seasons
Seasons
summer
autumn
winter
spring


150
100
50
0
count
Monday
Tuesday
Wednesday
Thursday
Friday
Day.of.the.week
Day.of.the.week
Monday
Tuesday
Wednesday
Thursday
Friday

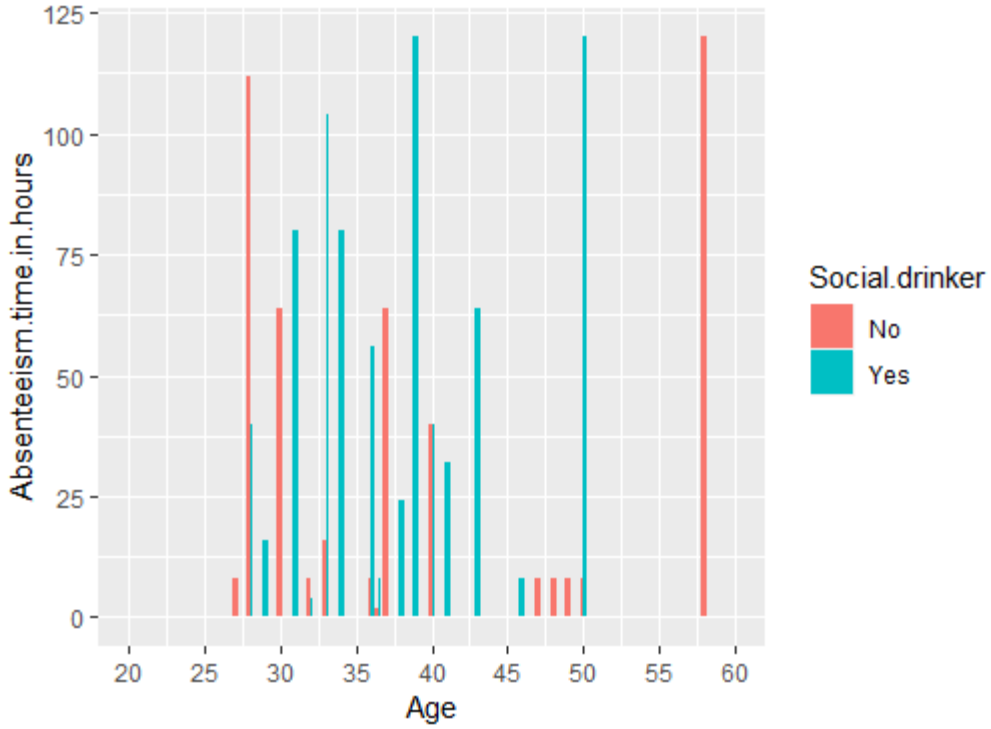

autumn
spring
winter
summer
Seasons
26.8
25.6
24.3
23.3
0
10
20
percent
Seasons
summer
autumn
winter
spring



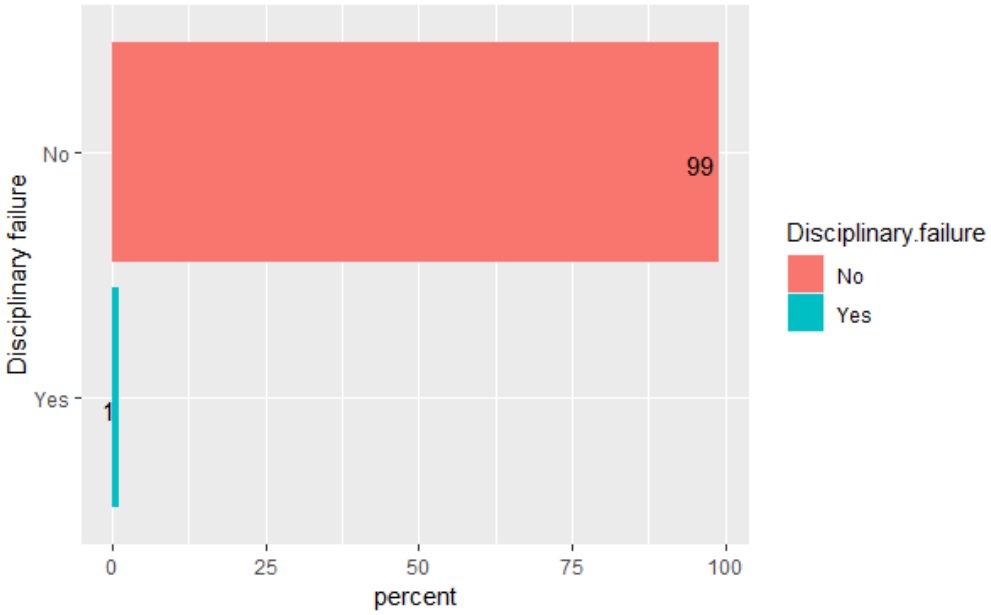
We see that people with Age close to 40 or 50 are social drinker and contributing more in our target variable Absenteeism.time.in.hours. Similar inference can be drawn for people with Age group of 30 to 35. They too are contributing quite high in our target variable Absenteeism.time.in.hours.

Close to the half of employees drink alcohol (320/420),so the attempted analysis can be

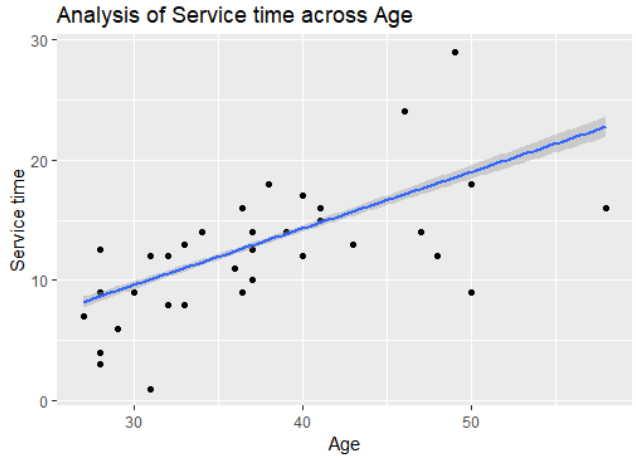
taken into consideration that it can be an element that influence the target variable.



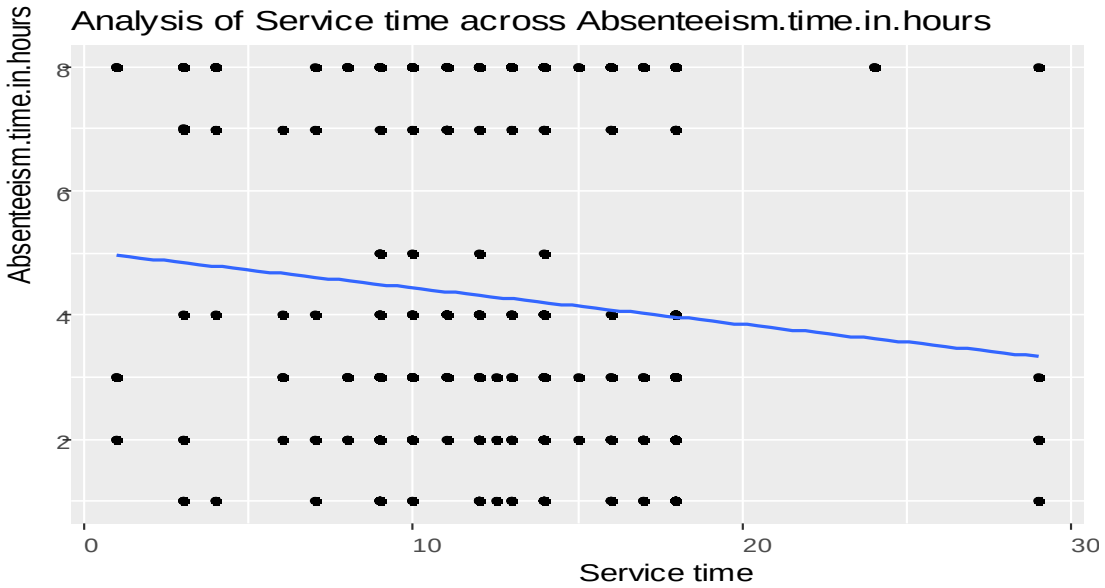
The attribute disciplinary failure is taken into consideration and it was found it had no obvious part on target variable



Here trend of service time across age is taken. And they have positive correlation

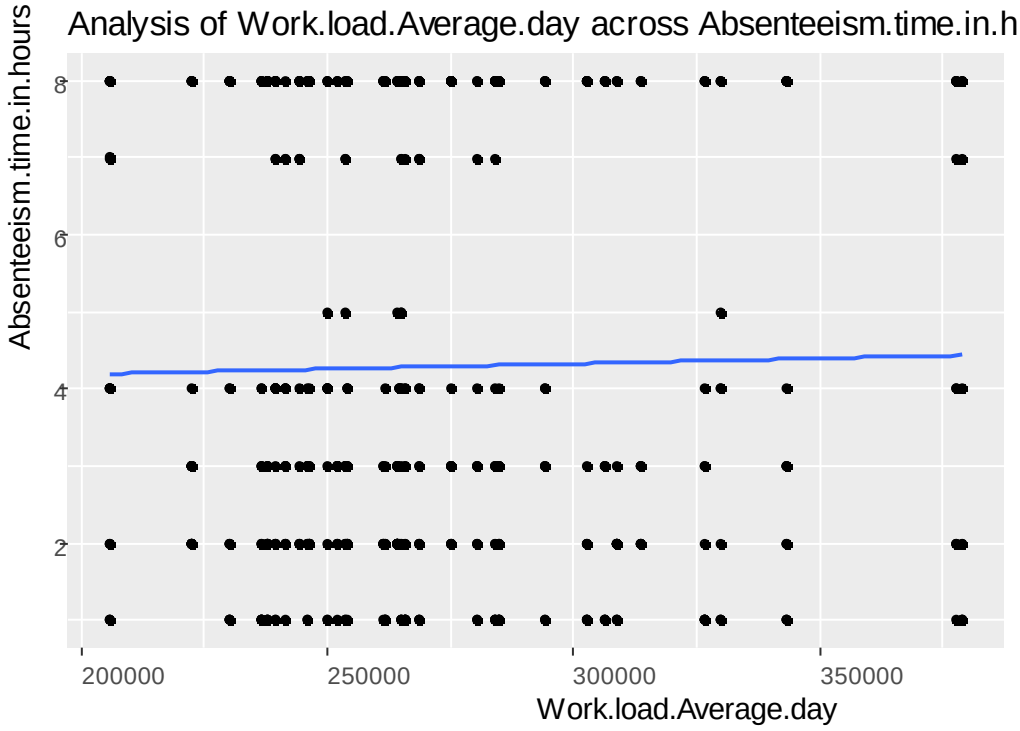


Here trend of age across **Absenteeism.time.in.hours** is taken. And they have Negative correlation. Here our 3rd Hypothesis is justified that People become more responsible as Age increases or Service time increases

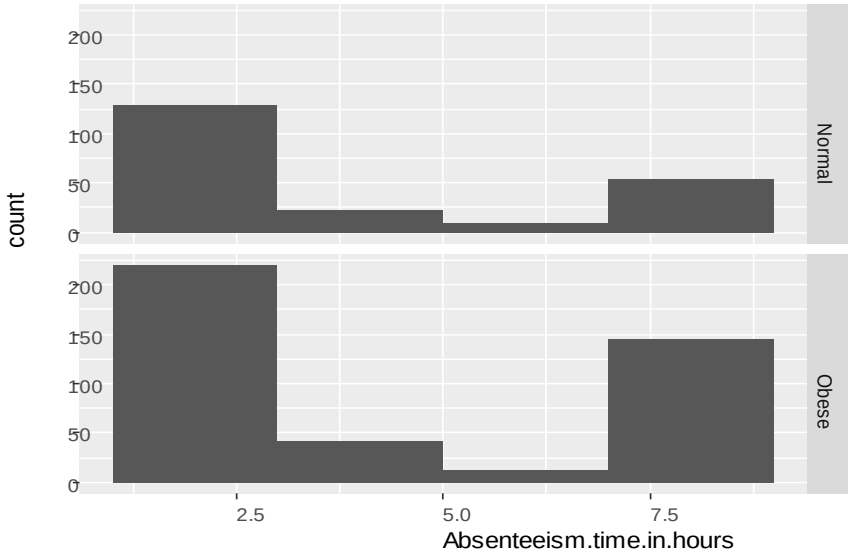


Here trend of age across Absenteeism.time.in.hours is taken. And they have Negative correlation. Here our 6th Hypothesis is justified that .The pressure at work sometimes

takes a toll on the employees. This results in increased levels of stress. The employees then resort to excuses that can help them stay away from work.



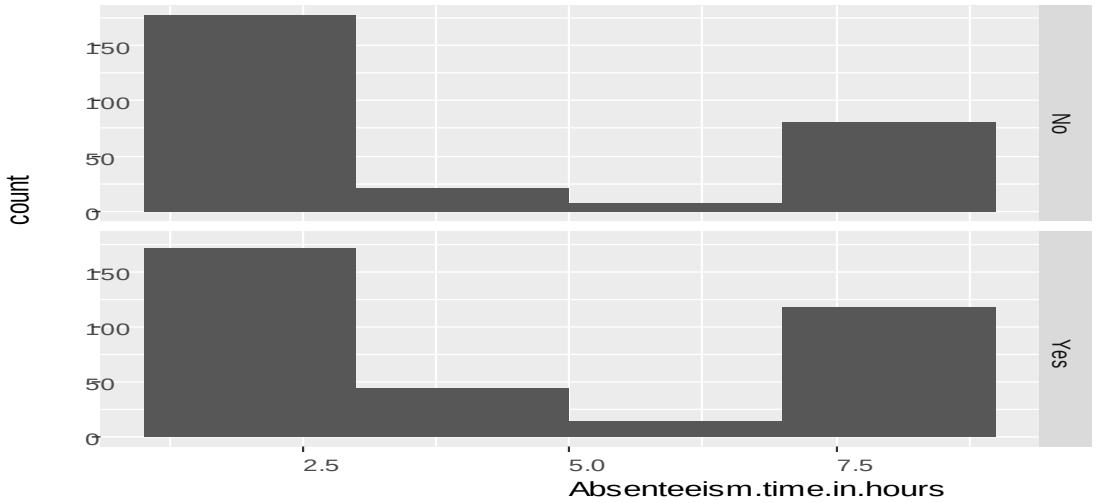
##We see that Obese people tend to take long hours of absence or full day leaves comparatively. Reason might be they've developed serious health issues which results them to consult doctor or take therapies so they need to take leaves or long time offs.



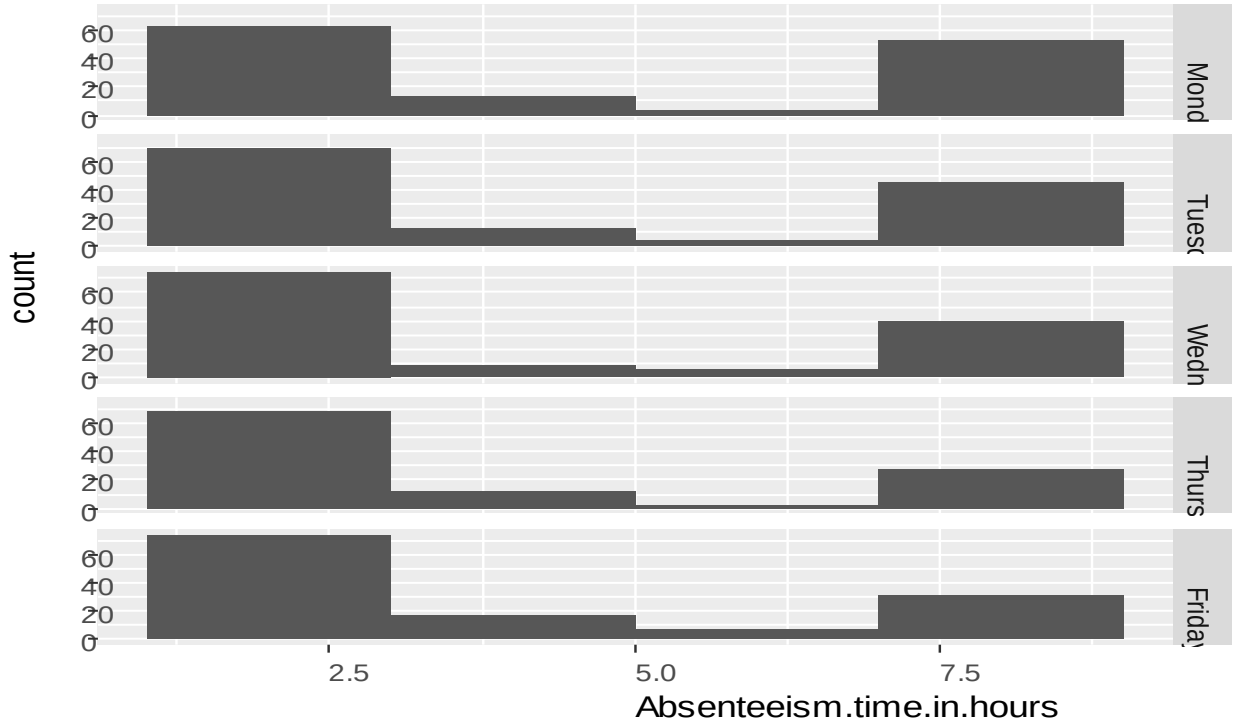
###############Social Drinker analysis

We see that there're longer hours of absence or full day leaves on Mondays comparatively our 5th Hypothesis is justified, that drinking and amusements in the late hours of night make it difficult for the workers to reach in time on their duties. They like to become absent rather than late since they know that.

#Analysis Day.of.the.week V/S Absenteeism.time.in.hours

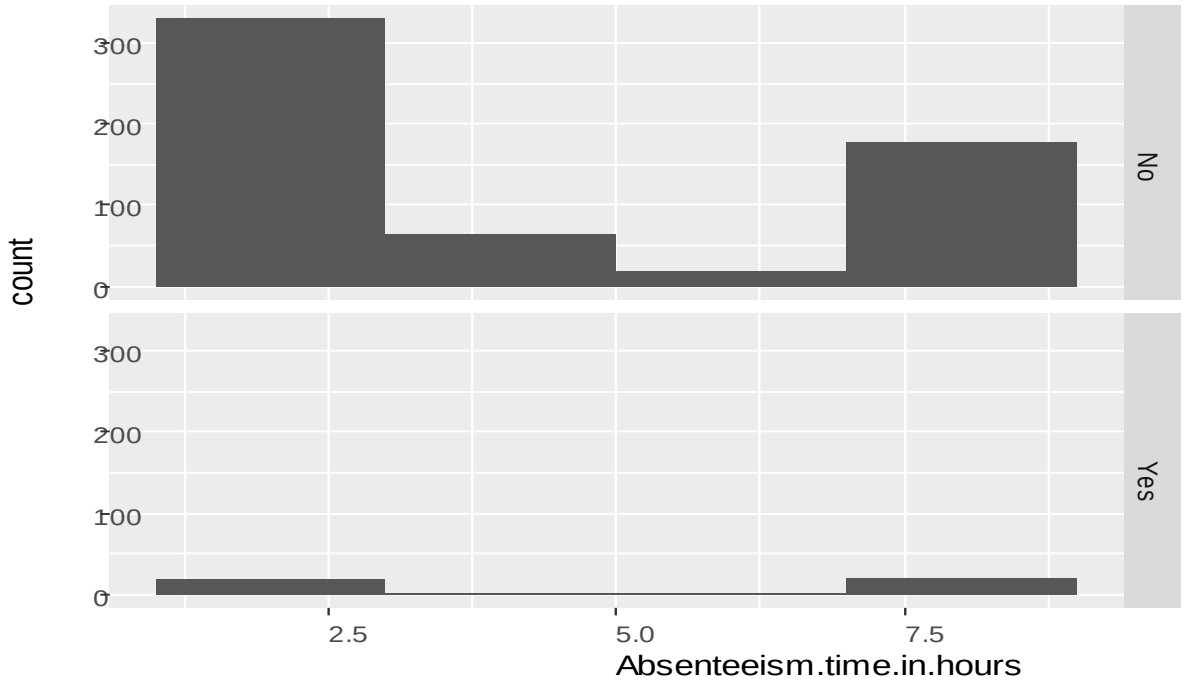


#Analysis Social.drinker V/S Absenteeism.time.in.hours



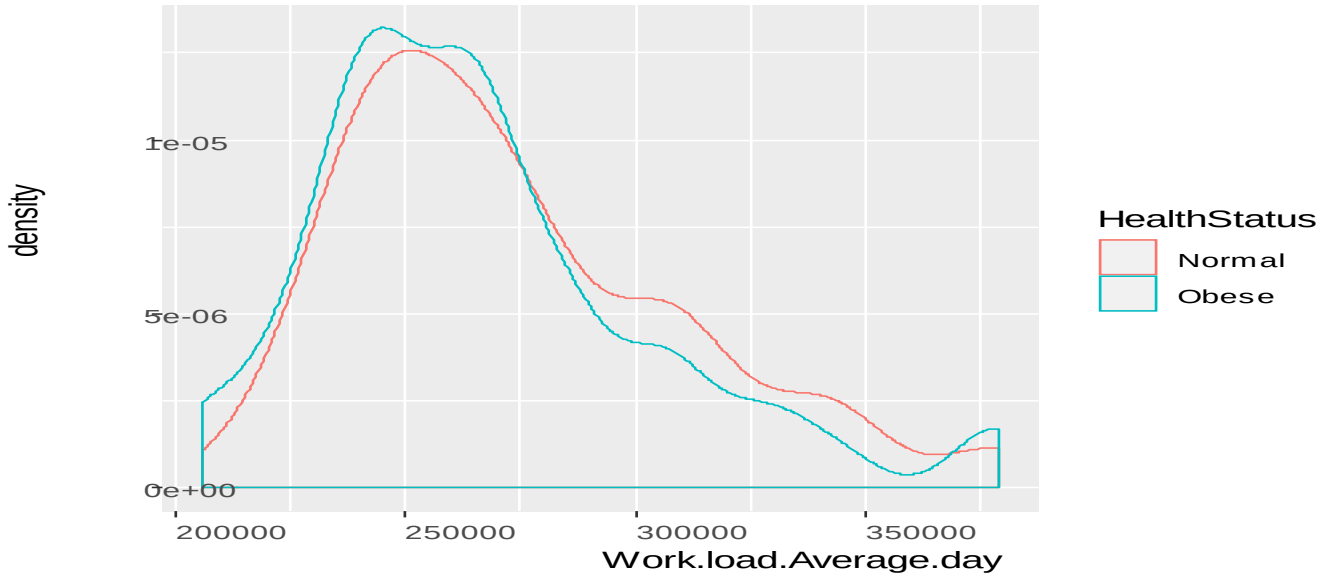
#Analysis Social.smoker V/S Absenteeism.time.in.hours

This kinda justifies our 7th Hypothesis .We see that Non-Smokers are being absent for 1 to 3 hours more than smokers. Do q quick survey for it, whether our cafeteria is having enough food counters and do we've enough games available in company itself for employees



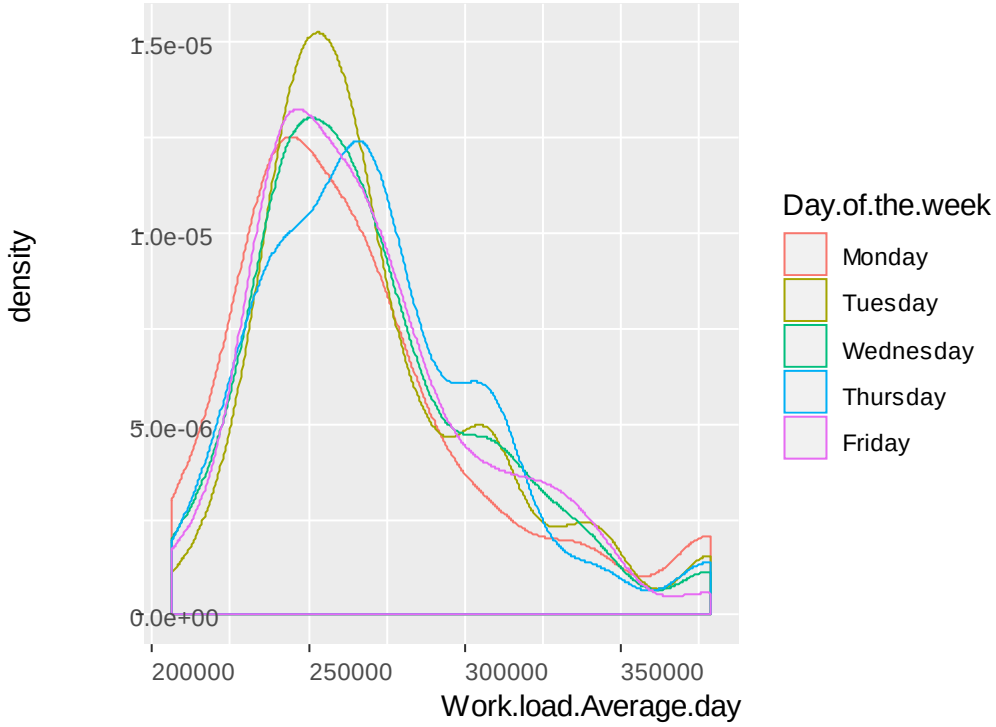
#Analysis HealthStatus V/S Work.load.Average.day

We see that **obese** people are having more Work load.Average per day. Or it's vice versa since few people have been so engrossed in work that work load is taking toll on their health and They've become couch potato and resulted them being Obese.

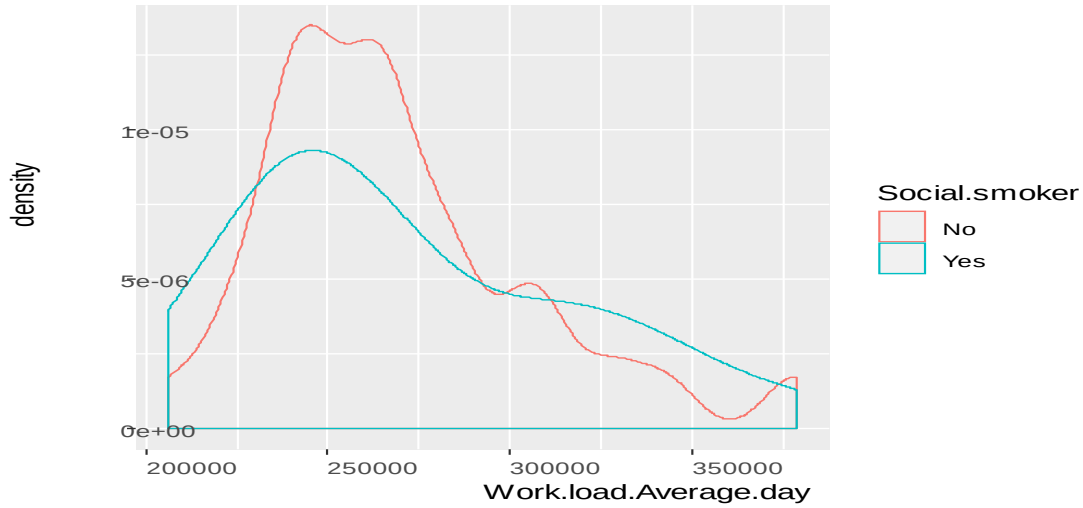


#Analysis Day.of.the.week V/S Work.load.Average.day:-

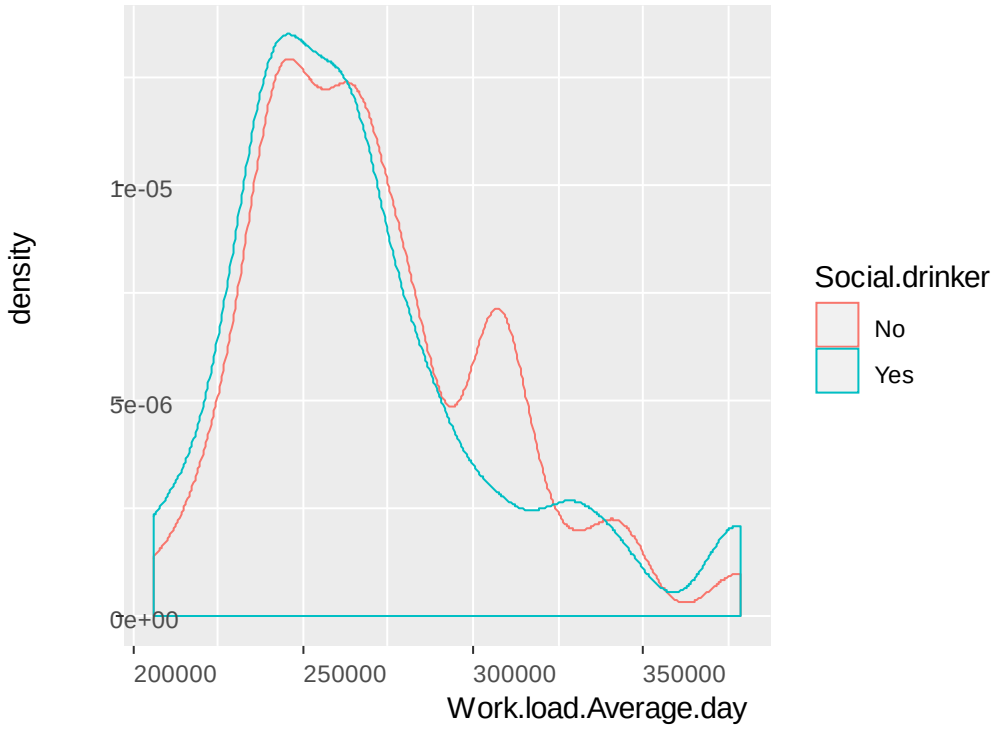
We see that Work load average per day is quite higher on Tuesday comparatively so we can draw conclusion that since a decent number of people took off on Monday or been absent on Monday and this resulted in Work load average per day being high on Tuesday.



#Analysis Social.smoker V/S Work.load.Average.day

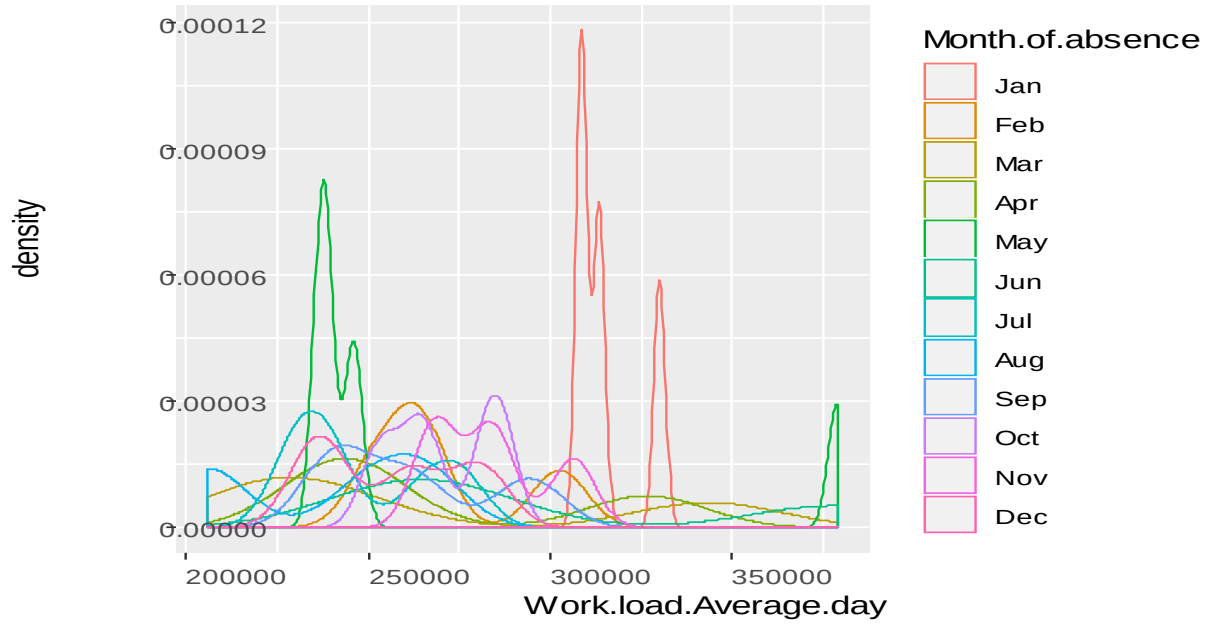


#Analysis Social.drinker V/S Work.load.Average.day



#Analysis Month.of.absence V/S Work.load.Average.day

We see that Work.load.Average.day is relatively higher in Jan & Feb, since most of It companies have shutdown time in December and this might result in high Work.load.Average.day in Jan & Feb



Chapter 3

Predict losses without building any Model/Algorithm-

1) Workloss:- We can a new variable called Workloss in terms of Work.load.Average.day & Absenteeism.time.in.hours. So equation would be like below-

Workloss = (Work.load.Average.day/8)* Absenteeism.time.in.hours

Current workloss is 9,05,94,032 units in a year ie. 7,549,503 units of workloss per month.

2) Transport Expense:-

Assuming Company pays to Transport Vendor on monthly basis; If an employee is absent for a day then still company has to pay his monthly amount to the vendor; So considering this also a kind of loss to the company

Loss of company in transportation expenses this year is 47,788 rs. ie.3982.333 rs. per month

Predict losses with algorithm-

RandomForest:-

1) Working Hour loss ie Absenteeism.time.in.hours:-

2648.963 working Hours will be lost if same trend continues.

2) Work loss:-

Again create a new variable using below equation and calculate work loss.

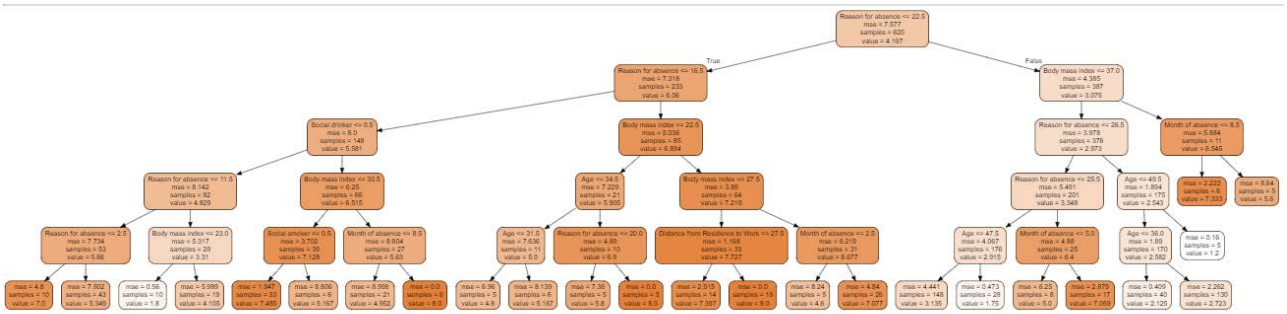
Workloss = (Work.load.Average.day/8)* Absenteeism.time.in.hours

Hence we can say that 89,379,007 units in a year ie. 7,448,251 units of work loss per month will occur if same trend continues.

3) Transport Expense:-

Transport loss will be same assuming there's no hike or inflation in transport Expense for the next year also.

We can create Decision tree rules like below.



Chapter 4

Strategies to Curb Absenteeism:-

1) Get a Clinic or small Hospital in company campus so primary medical ads can be done at campus itself. We see a major proportion in absenteeism is contributed by employees seeking Medical Consultation ie. 22.6%.
2) We see that People are taking time off of 1to 3 Hours from work is quite high in frequency. Don't micro manage employees but don't let employees take longer coffee breaks. Build strong policies so employees don't take longer breaks other than lunch breaks.
3) Take a survey if our Cafeteria is having good food options available at their counter. Check if we're having good games in our campus because we see that Non-smokers are taking quite frequent **work off** of 1 to 3 hours. Is the good food or not good company infrastructure is responsible for it.
4) We see that obese people are having high workload comparatively or it's vice versa, Workload has been quite high that few employees are taking toll on their health and they're becoming obese. So set up regular medical checkup for employees whose BMI is higher than 25, It could be weekly as well.

Other various strategies to curb absenteeism are:

1) High collaborative culture.

2) Be aware of problems that may affect employee attendance or performance

3) Develop open communication between managers, supervisors and employees.

4) Employees are encouraged to voice their concerns so their perceptions of the work place are clear and can be dealt with.

5) Cooperation with union representatives can be very helpful in attendance management and should be encouraged

6) Regularly scheduled department meetings are an excellent way not only to hear employee perceptions and concerns but also to communicate organizational goals

7) An employee's relationship with their supervisor can greatly influence their feelings about their work, their coworkers and thus their attendance at work

8) More openness and transparency on the part of management.

9) Encourage risk taking and experimentation among members.

10) Make each employee aware that they are a valued member of the "team", that they play an important role in your organization and that their attendance is critical

11) Hold regular meetings, keep your staff informed and involved

12) Know your employees; without prying show an interest in their personal lives

13) Familiarize with community programs which you can recommend to an employee if he/she has a need for assistance (i.e. marital or financial counseling)

14) Awareness, commitment and involvement by all levels of staff

15) Match the attendance records during a period of "high" workload to a period of "normal" workload

16) Bonus for unused sick leave.

17) Official warnings.

18) Develop a comprehensive and collaborative continuous improvement program throughout the department.

19) Counsel individual employees. Discuss with all employees problems of unjustifiable time off.

20) Bonus work (e.g. Saturday) should not be available to any worker who has been absent during the week.

21) Introduce an incentive scheme to reward those who don't have an absent day. This is measured quarterly and annually.

22) Greater attention by supervisors and more accountability of operations management and other management has improved sick leave.

23) Front line management will be held accountable for attendance management performance.

24) A management structure be engineered to identify and execute objectives relating to absence prevention, disability management and attendance control.

25) Effective training and development program.

26) Effective career planning and development program.

27) Each worksite should develop and maintain an attendance management policy.

28) Employers should track attendance and assign costs based on reliable data.

29) The attendance management team should be given the ability to develop complex case management strategies and aggressively intervene where appropriate.

30) Employees should be encouraged to give as much notice as possible for anticipated absences. Absent employees should be requested to keep contact with their employer.

31) The employer should be informed of any changes in the employee's health status.

32) Employees should be called if they are not keeping contact with the employer. The purpose is to show concern and desire for the employee to regain a healthy status and return to work.

33) In addition to individual counseling make use of family counseling methods.