Employee Absenteeism Prediction

Gursimran Singh

May 2019

Introduction

Employee absenteeism is a significant problem for most organizations. In 2017 the U.S. Department of Labour (DOL) estimated that almost 3 per cent of an employer's workforce was absent on any given day. The high price of absenteeism affects organizations even more when lost productivity, morale and temporary labour costs are considered. The cost of absence is often misunderstood, seen as not easily measured or dismissed as a negligible amount because the costs are largely included in payroll expenses. In addition, employers often fail to carefully track absenteeism.

Many firms still underestimate the magnitude of the problem of employee absenteeism. Instead, they consider such absences to be part of "the cost of doing business." From that perspective, they fail to quantify the full impact of employee absenteeism. In addition, they also fail to appreciate the value of solutions that can reduce the costs and lost time that result from employees' being off work.

But with the emergence of Machine Learning and Data Sciences technologies, most of the corporate giants started maintaining data of its employees and using that data to predict the absenteeism rate. Besides this companies are also able to predict the resigning rate of their employees from the number of their working hours or on other parameters of the collected data.

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 many losses every month can we project in 2011 if same trend of absenteeism continues?

Our task is to a build a machine learning model which will predict the absenteeism hours of the employees depending on various other factors. Besides this we have to assists the top business leaders in making decision by exploring data and answering their required questions. Sample from the whole dataset is shown below:-

Id	AbsentReason	AbsentMonth	WeekDay	Seasons	Expenses	ResidentDistance
11	26	7	3	1	289	36
36	0	7	3	1	118	13
3	23	7	4	1	179	51
7	7	7	5	1	279	5
11	23	7	5	1	289	36

Table 1.1: Sample Data

Service time	Age	AverageWorkLoad	HitTarget	DisciplineFailure	Education	Son
13	33	2,39,554	97	0	1	2
18	50	2,39,554	97	1	1	1
18	38	2,39,554	97	0	1	0
14	39	2,39,554	97	0	1	2
13	33	2,39,554	97	0	1	2

Table 1.2 : Sample Data

SocialDrinker	Social smoker	Pet	Weight	Height	Body mass index	AbsentTime
1	0	1	90	172	30	4
1	0	0	98	178	31	0
1	0	0	89	170	31	2
1	1	0	68	168	24	4
1	0	1	90	172	30	2

Table 1.3: Sample Data

Descriptions of the attributes are given below:-

Attribute	Description
Id	Individual identification
AbsentReason	Reason for absence 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).
AbsentMonth	Month of absence
WeekDay	Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
Seasons	Seasons (summer (1), autumn (2), winter (3), spring (4))
Expenses	Transportation expense
ResidentDistance	Distance from Residence to Work (kilometres)
ServiceTIme	Service time
Age	Age
WorkLoad	Work load Average/day
HitTarget	Hit target
DiscplineFailure	Disciplinary failure (yes=1; no=0)
Education	Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
Son	Son (number of children)
SocialDrinker	Social drinker (yes=1; no=0)
SocialSmoker	Social smoker (yes=1; no=0)
Pet	Pet (number of pet)
Weight	Weight
Height	Height
BodyMassIndex	Body mass index
AbsentTime	Absenteeism time in hours (target)

Table 1.3: Attribute Descriptions

Methodology

Any predictive modelling requires to look at the data before start modelling. 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 (EDA).

Exploratory data analysis (EDA) is a very important step which takes place after feature engineering and acquiring data and it should be done before any modelling. This is because it is very important for a data scientist to be able to understand the nature of the data without making assumptions.

The purpose of EDA is to use summary statistics and visualizations to better understand data, and find clues about the tendencies of the data, its quality and to formulate assumptions and the hypothesis of our analysis. EDA is not about making fancy visualizations or even aesthetically pleasing ones, the goal is to try and answer questions with data. A goal should be to be able to create a figure which someone can look at in a couple of seconds and understand what is going on. If not, the visualization is too complicated (or fancy) and something similar should be used.

EDA is also very iterative since we first make assumptions based on our first exploratory visualizations, and then build some models. We then make visualizations of the model results and tune our models.

Remember the quality of our inputs decide the quality of our output. So, once we have got our business hypothesis ready, it makes sense to spend lot of time and efforts here. Estimating, data exploration, cleaning and preparation can take up to 70% of our total project time.

Variable Identification

Variable identification is the first step in the exploratory data analysis. Identification of the variables in the dataset is totally dependent on the business requirements and need of the client. The main task here is to identify the Target variable on which future decision has to be made. Besides these identifying the independent variables are equally important because end result of the target variable are totally dependent on the independent/predictor variables. Brief introduction of predictor and target variables are given below:-

Predictor Variable

Predictor variables are those variables or attributes in the dataset on which the result of the target variable is totally dependent. These variables are those on which decisions are made by the clients to get the maximum profit from the business.

Target Variable

It is the variable or attribute in the whole dataset on which client is mostly interested. Based on the business requirements category of the target variable is identified by the data sciences experts.

Once the target variable and predictor variables are identified, our next task to identify the data types and categories of the variable. From analysing the dataset of bike renting company, the detailed description of all the variables are given below in the diagram on the different parameters.

In the further stages of exploratory data analysis process, we have to dive deep into the understanding of the each variable present in the dataset. From Business point of view each and every variable is crucial and even a minute mistake here can cause a loss of millions to your client. So to get the detailed summary of all the variables in the dataset on statistical parameters will help to better understanding of data to a data sciences expert.

Predictor Variables	Target Variable
Id	AbsentTime
AbsentReason	
AbsentMonth	
WeekDay	
Seasons	
Expenses	
ResidentDistance	
ServiceTIme	
Age	
WorkLoad	
HitTarget	
DiscplineFailure	

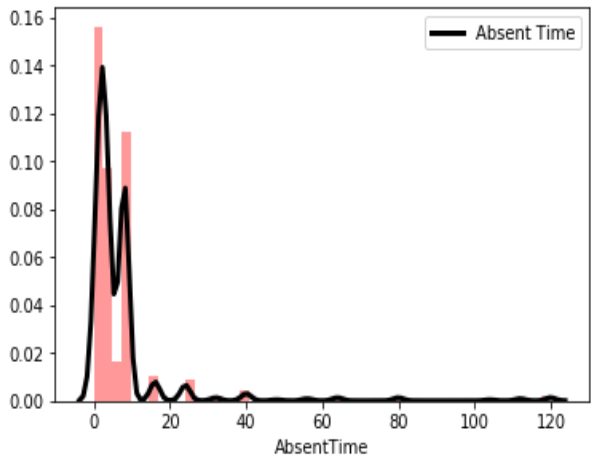
Education
Son
SocialDrinker
SocialSmoker
Pet
Weight
Height
BodyMassIndex

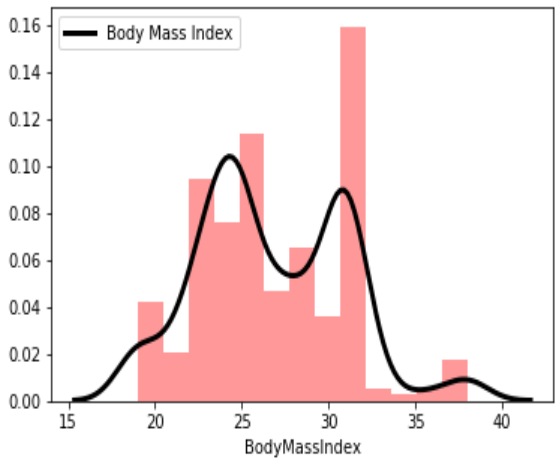
NOTE: For my better understanding I have changed the names of the dataset variables.

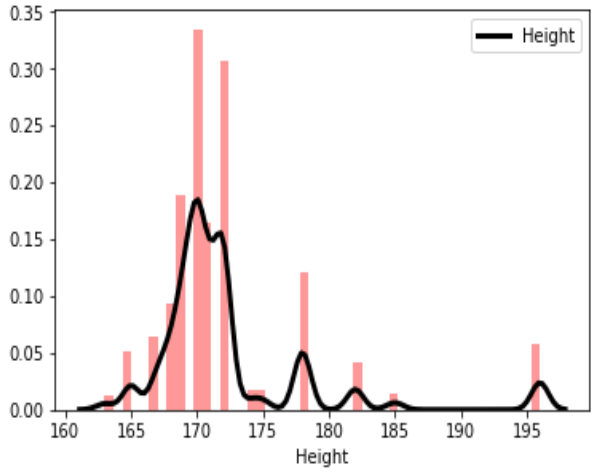
Univariate Analysis

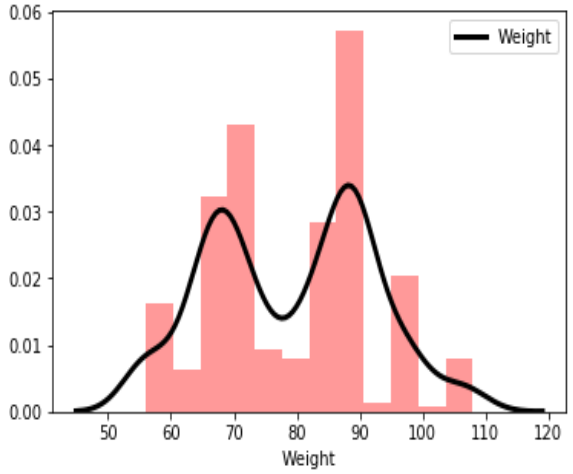
Univariate analysis is the simplest form of analysing data. “Uni” means “one”, so in other words data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and its major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.

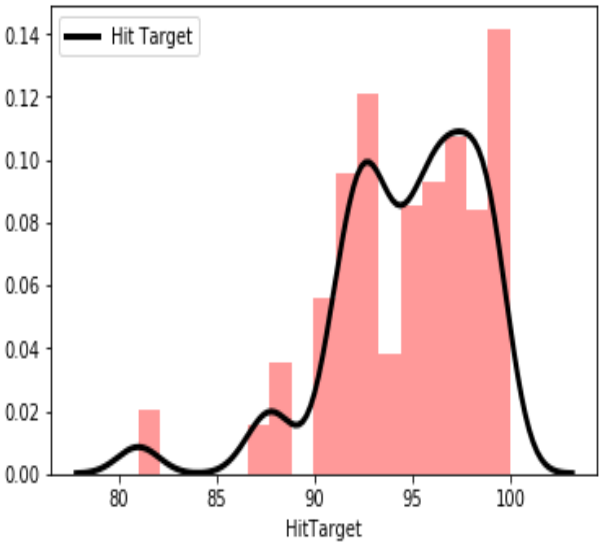
Method to perform univariate analysis will depend on whether the variable type is categorical or continuous. Plots of all continuous variables are shown below:-

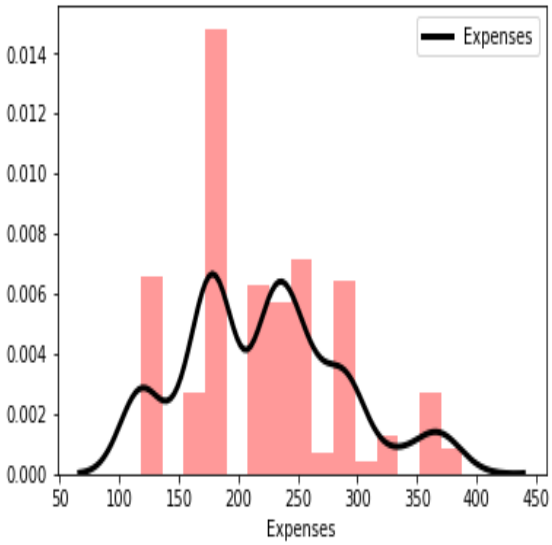


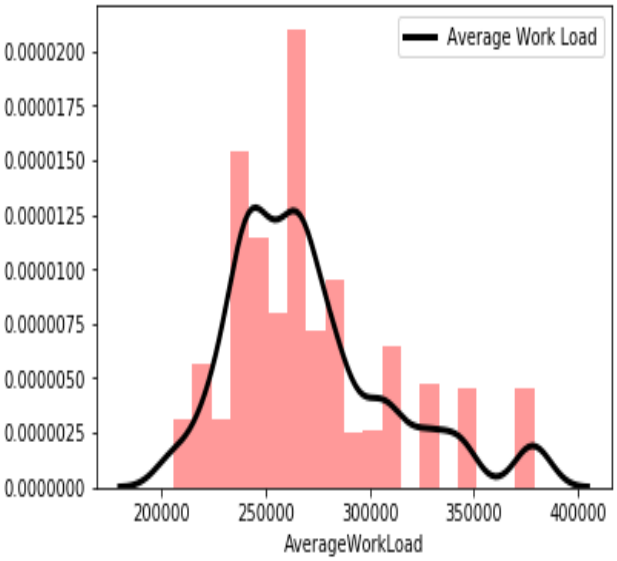


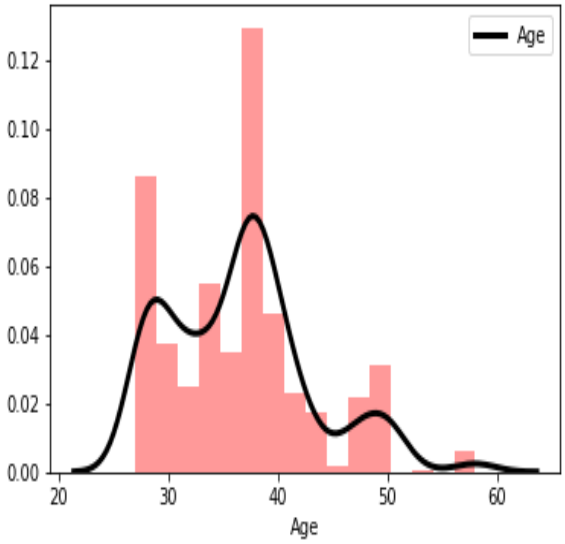


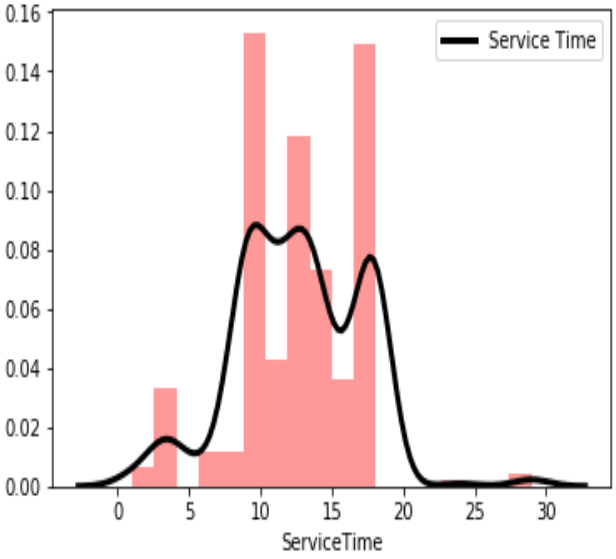


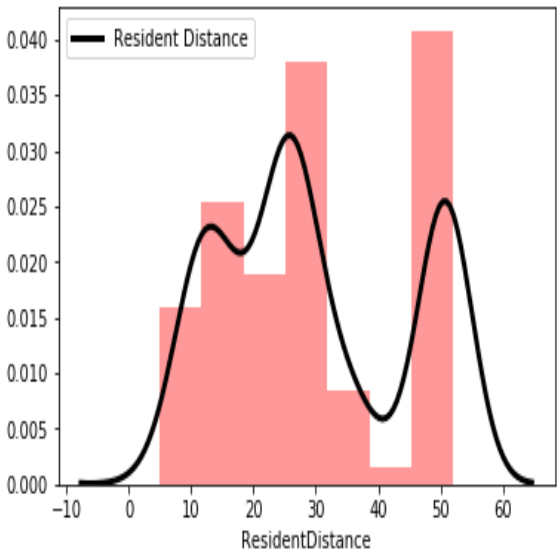

Hit Target
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
80
85
90
95
100
HitTarget

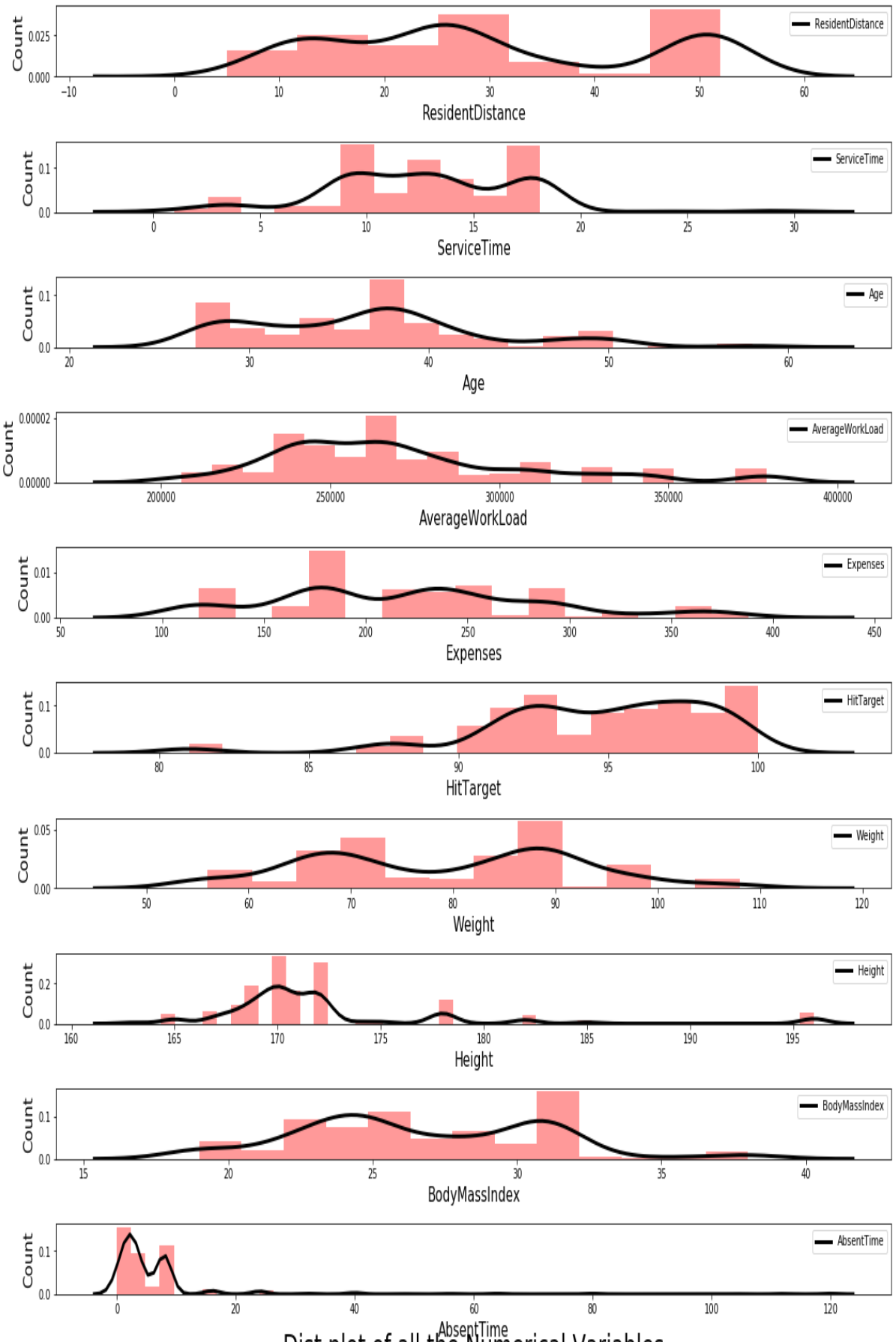

Expenses
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
50
100
150
200
250
300
350
400
450
Expenses


Average Work Load
0.0000200
0.0000175
0.0000150
0.0000125
0.0000100
0.0000075
0.0000050
0.0000025
0.0000000
200000
250000
300000
350000
400000
AverageWorkLoad


Age
0.12
0.10
0.08
0.06
0.04
0.02
0.00
20
30
40
50
60
Age


Service Time
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30
ServiceTime

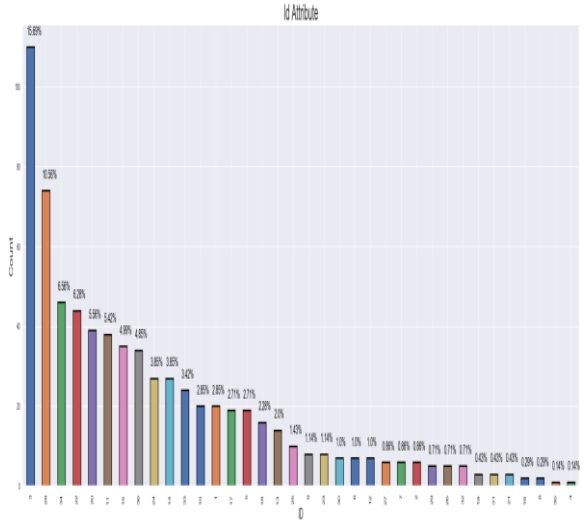

Resident Distance
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−10
0
10
20
30
40
50
60
ResidentDistance

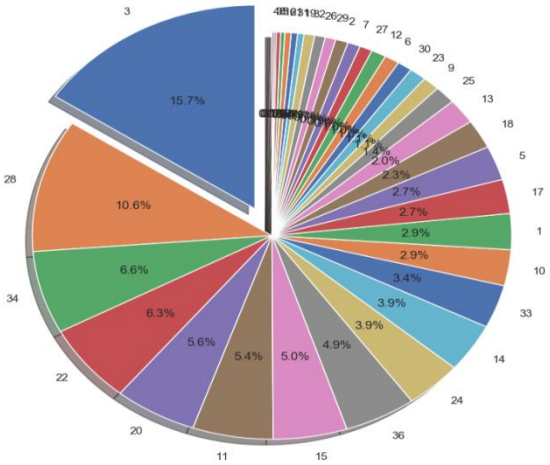


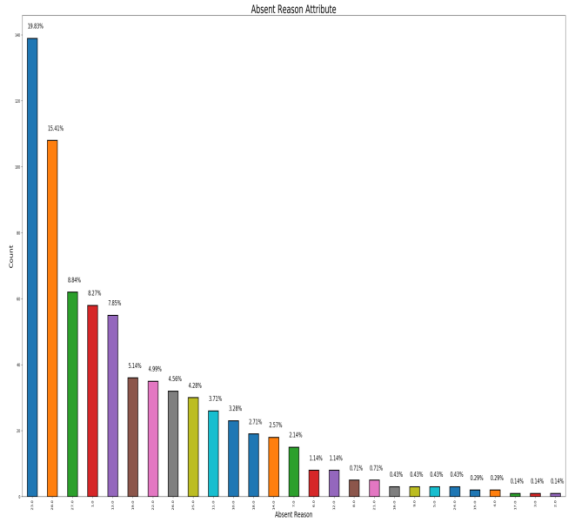
Dist plot of all the Numerical Variables

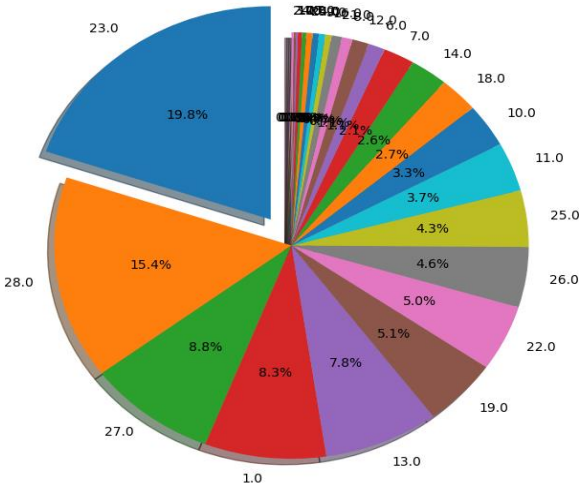
Categorical Variables

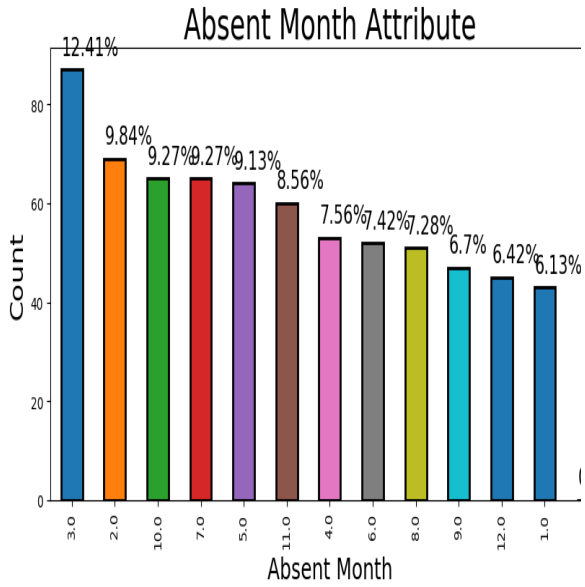
Categorical data are easier to interpret as compared to numerical variables. The best ways to analyze categorical variables are through bar graphs and pie charts. Bar graphs are mostly preferred to visualize the frequency of each category that falls into that variable, whereas pie charts are used to visualize the percentage of each category. Plots for analyzing categorical variables are shown below:-

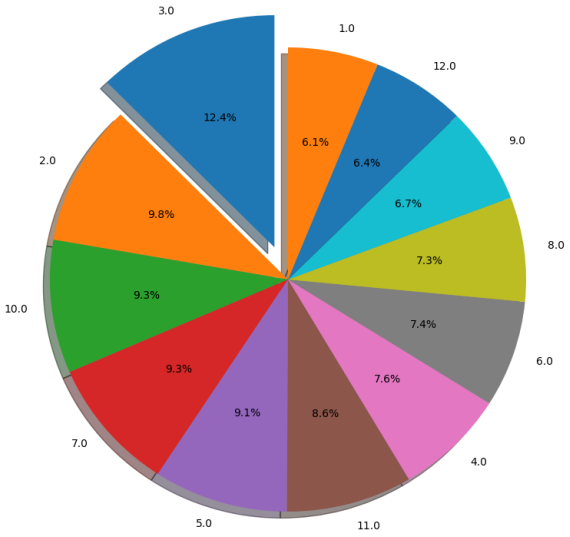


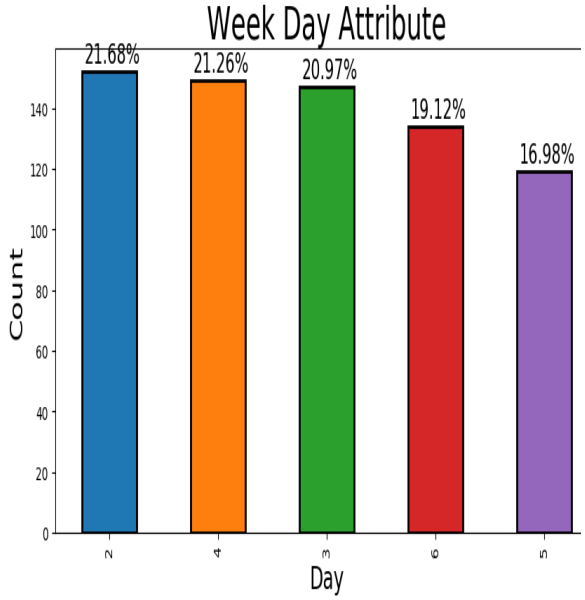


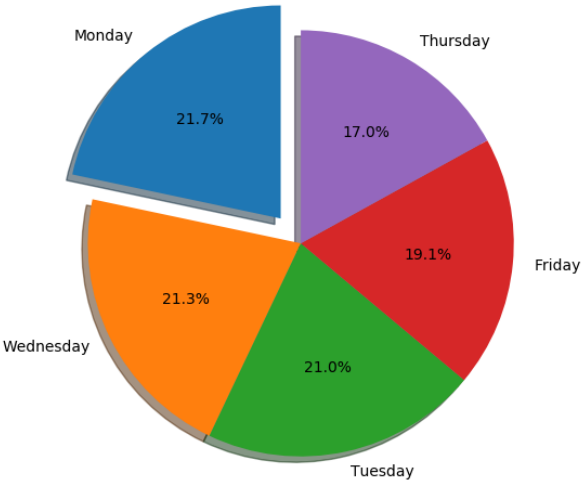


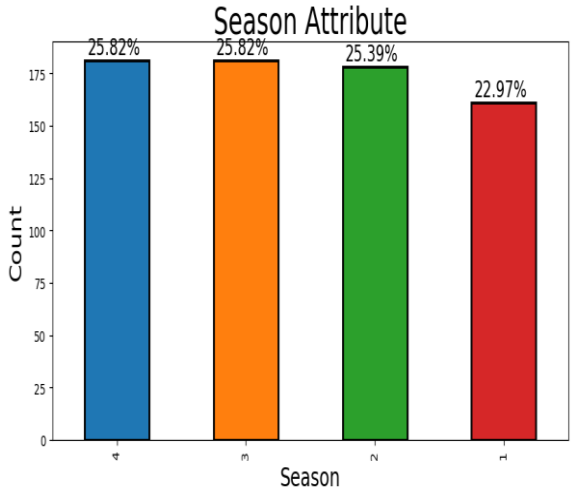


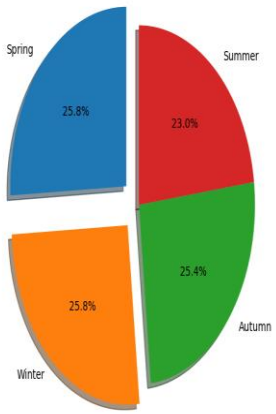

Absent Month Attribute
12.41%
9.84%
9.27%
9.27%
9.13%
8.56%
7.56%
7.42%
7.28%
6.7%
6.42%
6.13%
Count
Absent Month
3.0
2.0
10.0
7.0
5.0
11.0
4.0
6.0
8.0
9.0
12.0
1.0

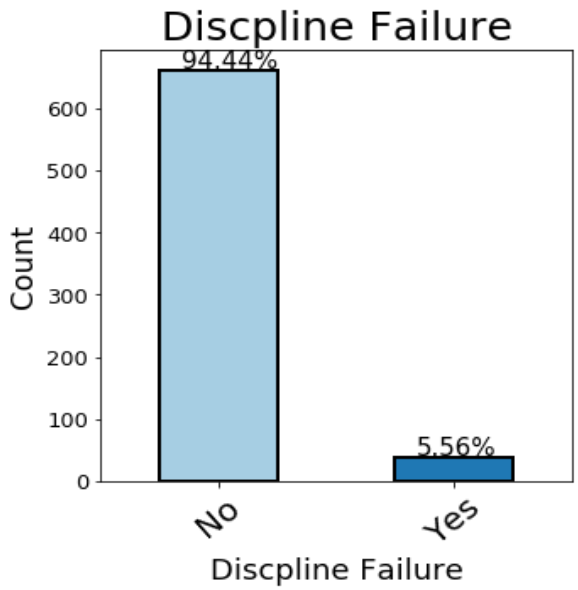

3.0
1.0
12.0
9.0
8.0
6.0
4.0
11.0
5.0
7.0
10.0
2.0
12.4%
6.1%
6.4%
6.7%
7.3%
7.4%
7.6%
8.6%
9.1%
9.3%
9.3%
9.8%

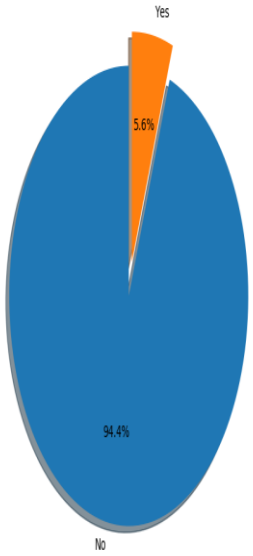

Week Day Attribute
21.68%
21.26%
20.97%
19.12%
16.98%
Count
Day
2
4
3
6
5

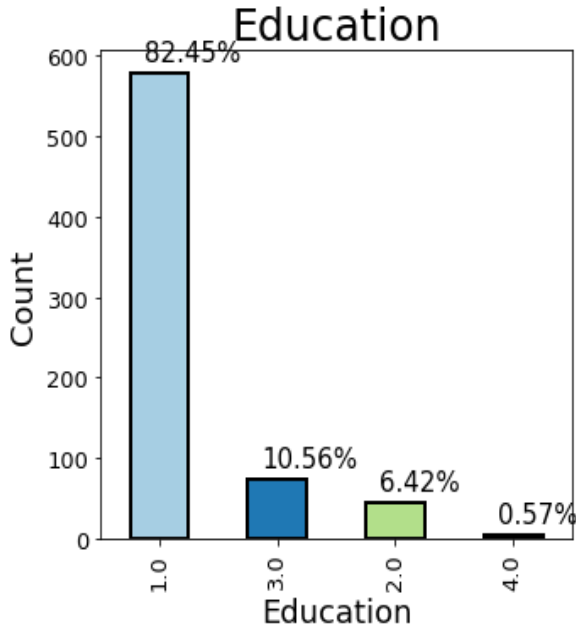

Monday
Thursday
Friday
Wednesday
Tuesday
21.7%
17.0%
19.1%
21.3%
21.0%

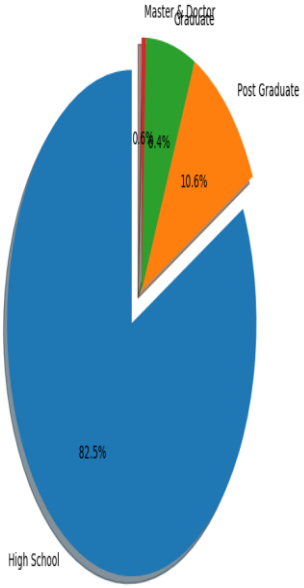

Season Attribute
25.82%
25.82%
25.39%
22.97%
Count
Season
4
3
2
1

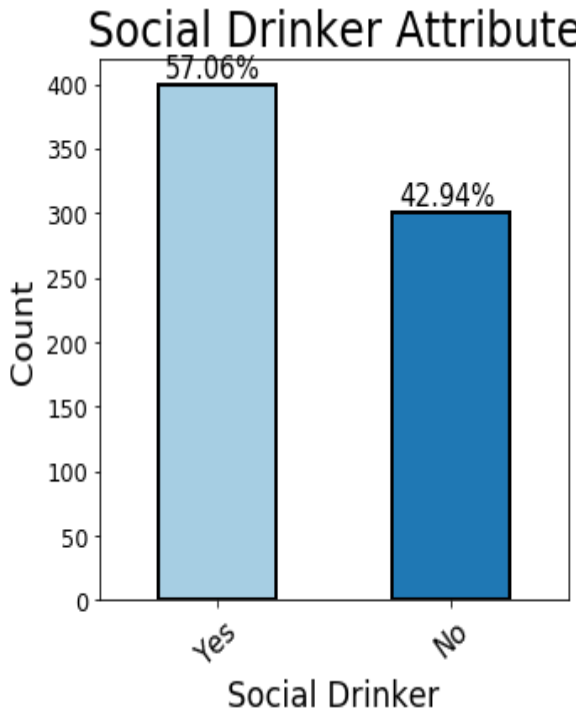

Spring
Summer
Autumn
Winter
25.8%
23.0%
25.4%
25.8%

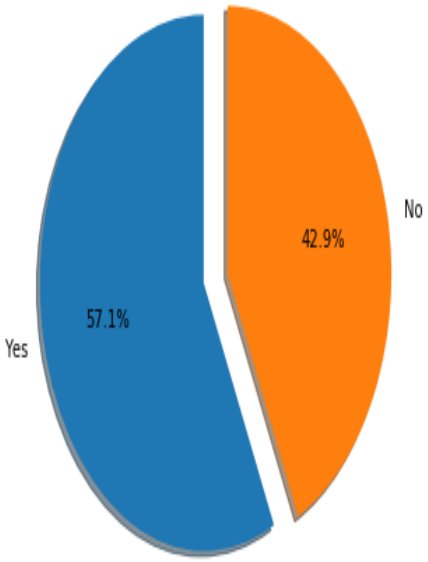

Discpline Failure
94.44%
5.56%
Count
No
Yes
Discpline Failure

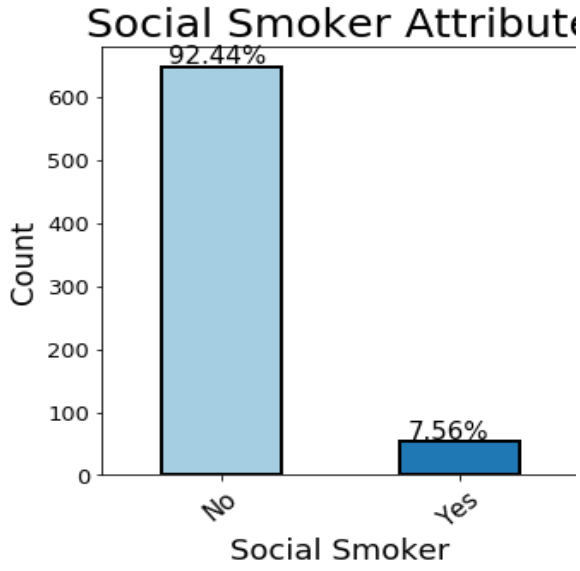

Yes
5.6%
94.4%
No

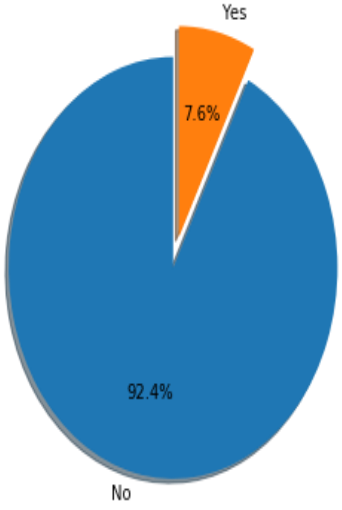

Education
82.45%
10.56%
6.42%
0.57%
Count
1.0
3.0
2.0
4.0
Education

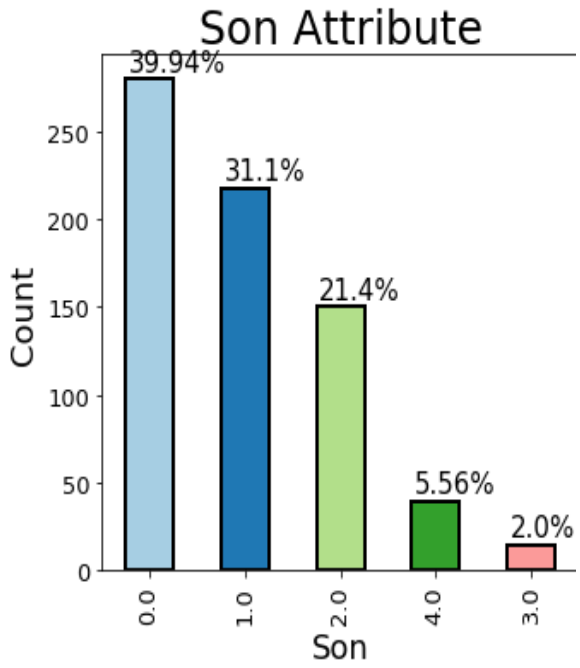

Master & Doctor
Graduate
Post Graduate
10.6%
82.5%
High School

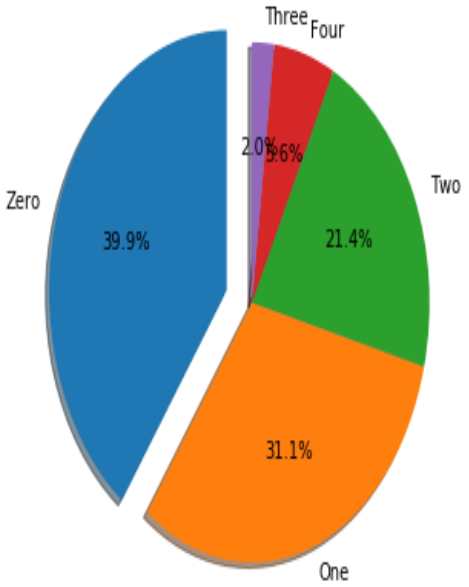

Social Drinker Attribute
57.06%
42.94%
Count
400
350
300
250
200
150
100
50
0
Yes
No
Social Drinker

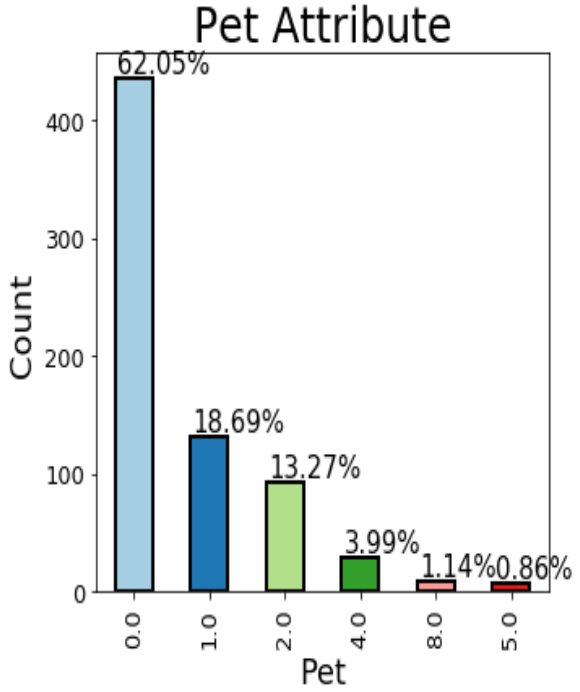

No
42.9%
57.1%
Yes

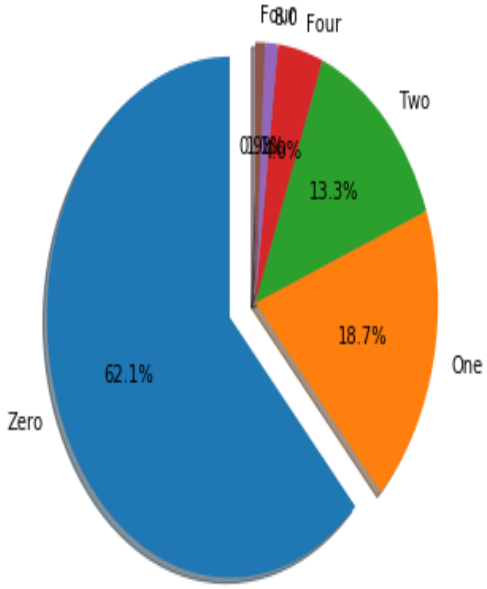

Social Smoker Attribute
92.44%
7.56%
Count
600
500
400
300
200
100
0
No
Yes
Social Smoker


Yes
7.6%
92.4%
No

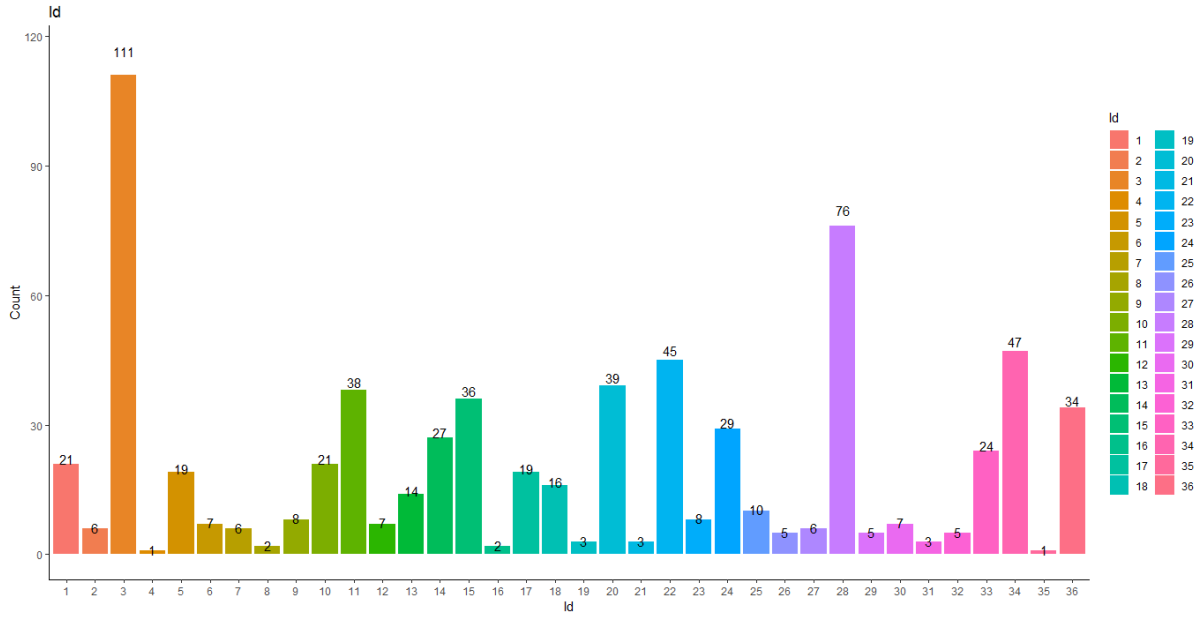

Son Attribute
39.94%
31.1%
21.4%
5.56%
2.0%
Count
Son
0.0
1.0
2.0
4.0
3.0

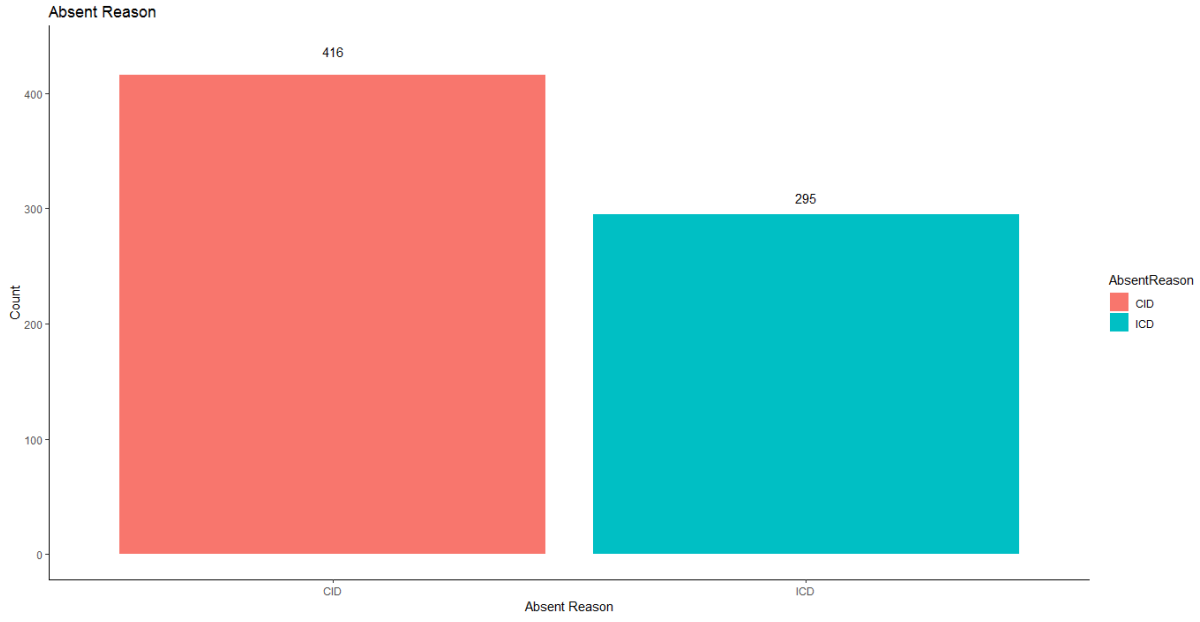

Three
Four
Two
Zero
One
39.9%
21.4%
31.1%

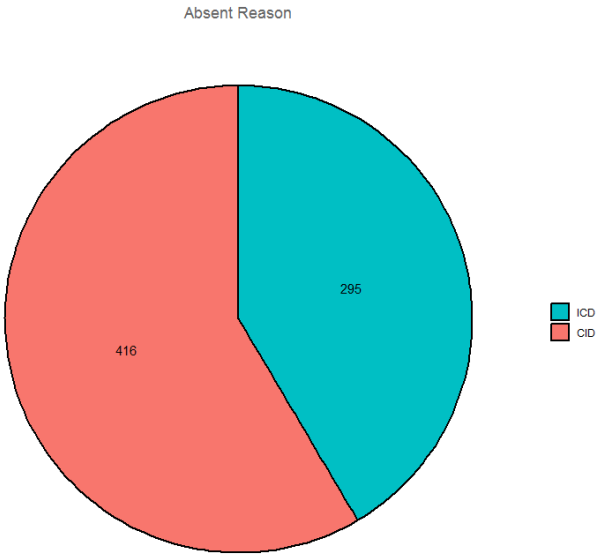

Pet Attribute
62.05%
18.69%
13.27%
3.99%
1.14%
0.86%
Count
Pet
0.0
1.0
2.0
4.0
8.0
5.0

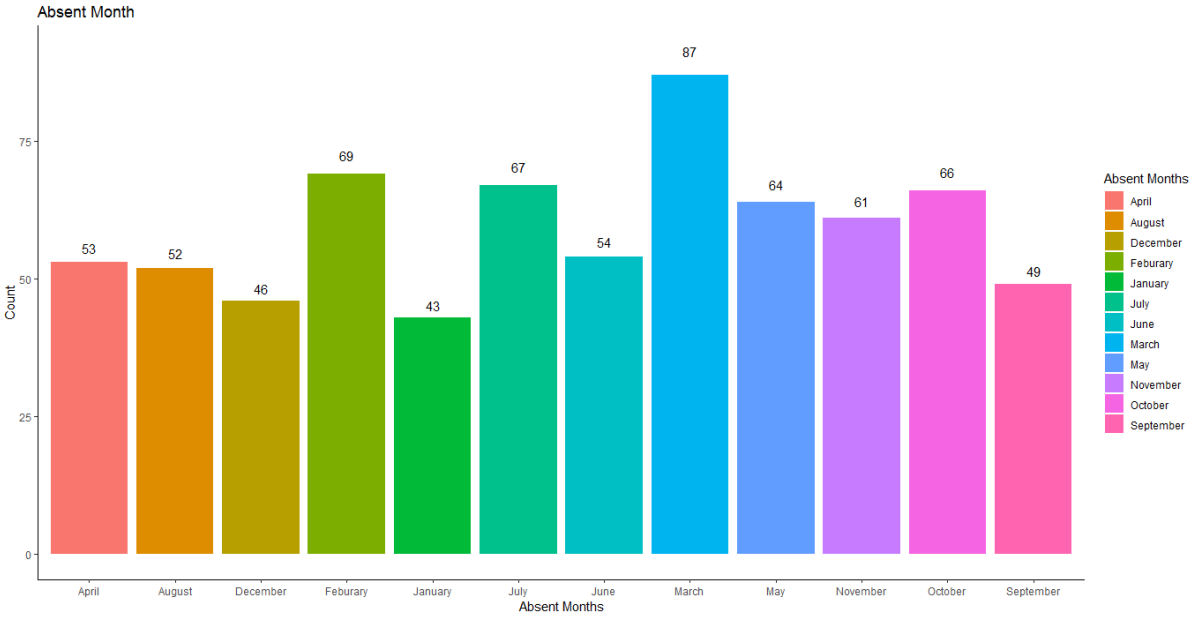

Four
Two
One
Zero
13.3%
18.7%
62.1%

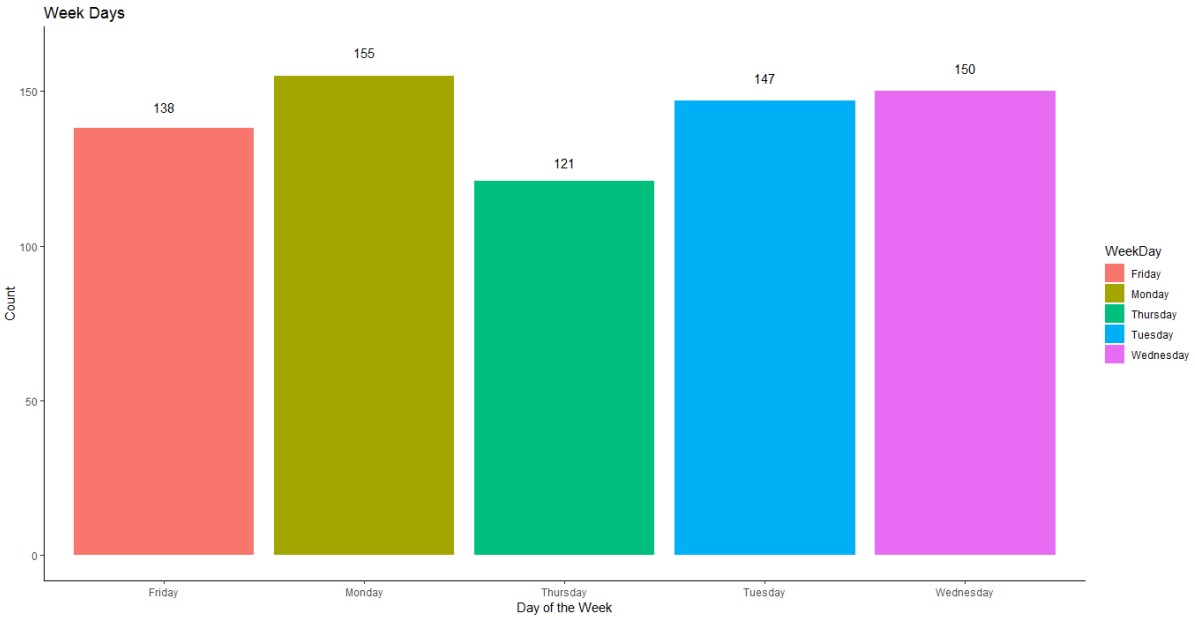
ggplot is very strong library written in R language and the graphs with better visualizations are shown below for categorical variables :-

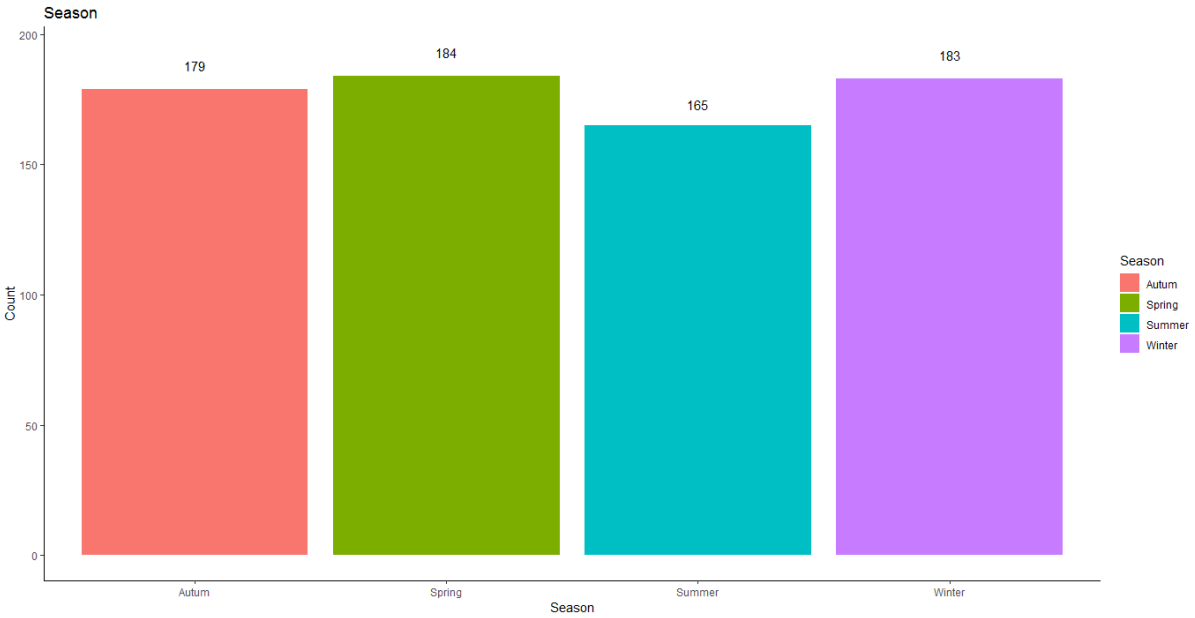


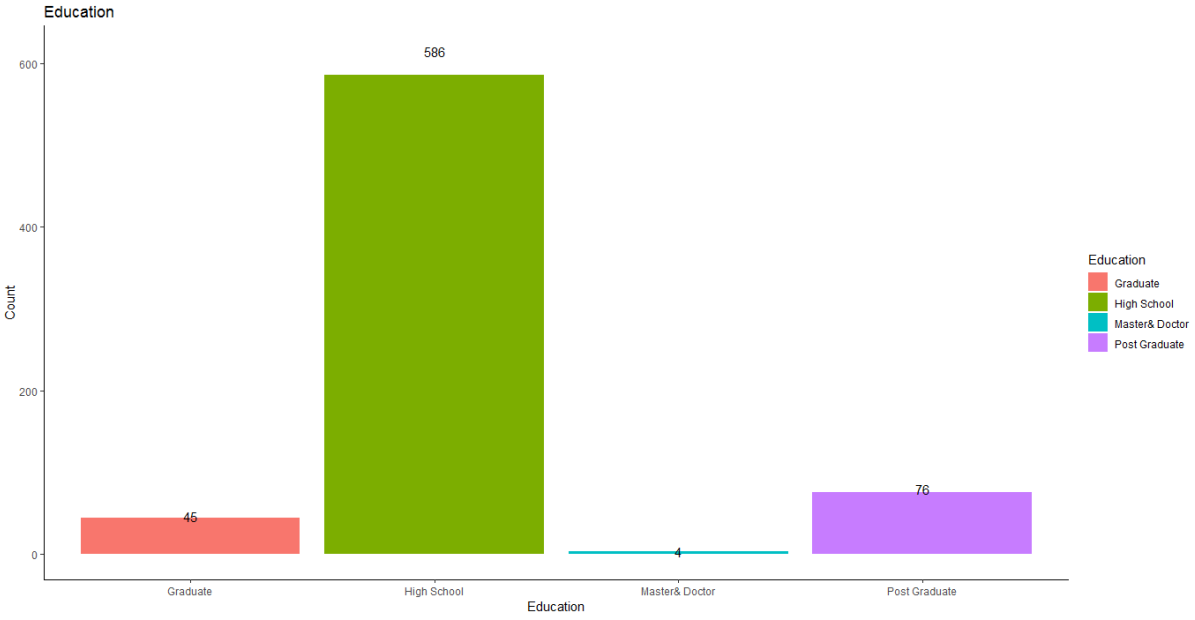


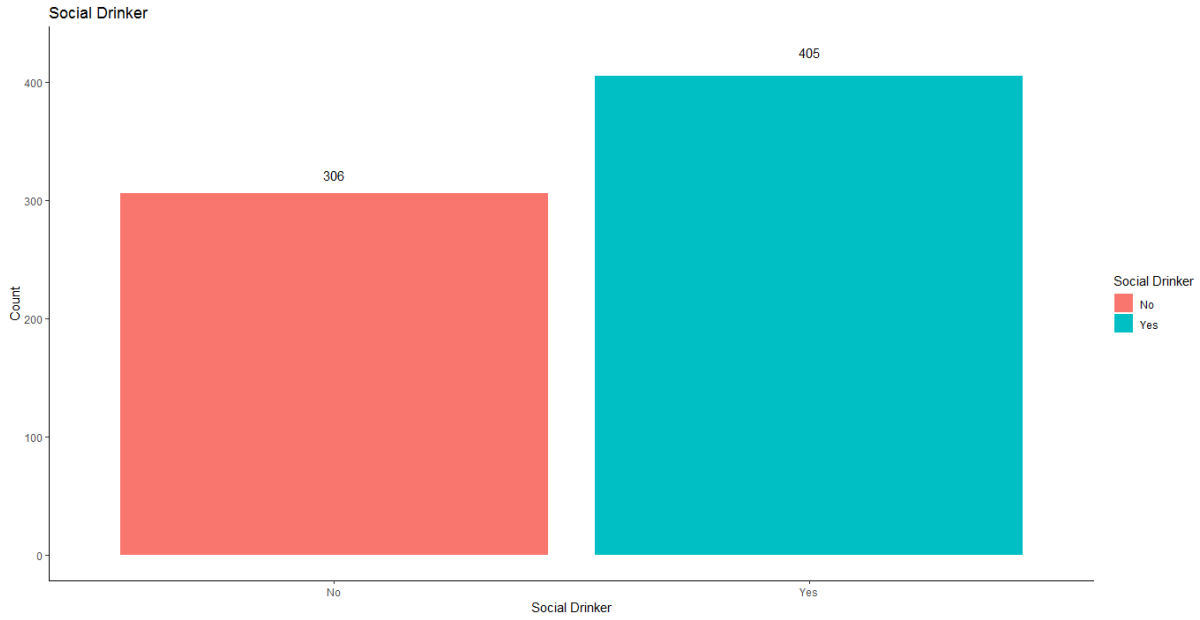

Absent Reason
295
416
ICD
CID

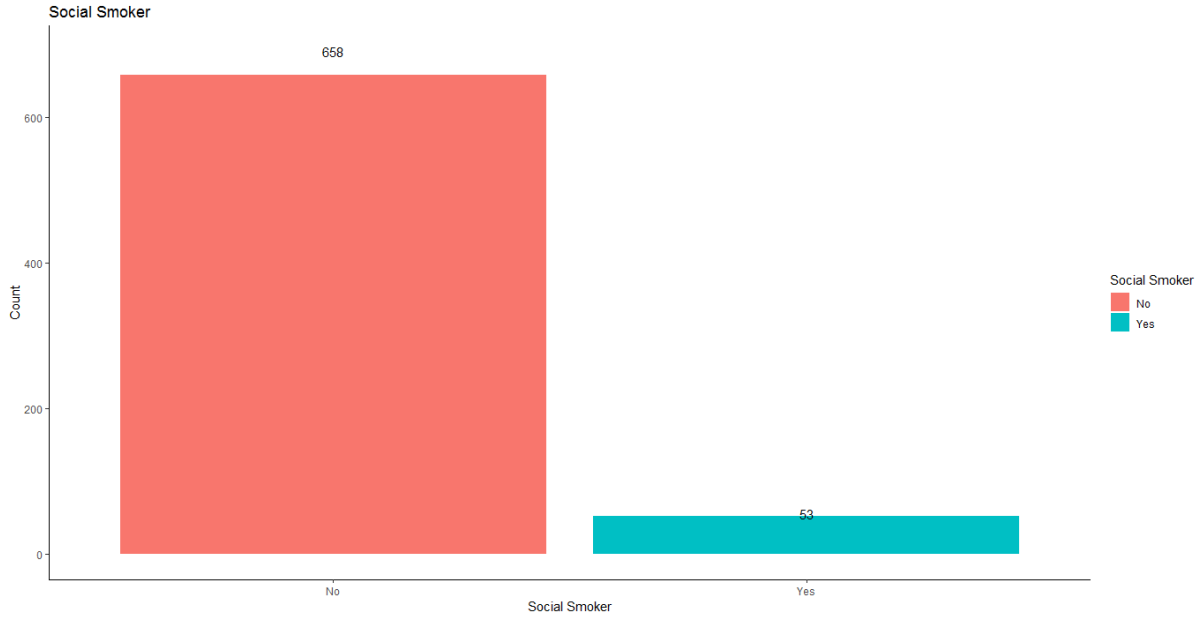

Absent Month
Count
75
50
25
0
53
52
46
69
43
67
54
87
64
61
66
49
April
August
December
Feburary
January
July
June
March
May
November
October
September
Absent Months
Absent Months
April
August
December
Feburary
January
July
June
March
May
November
October
September

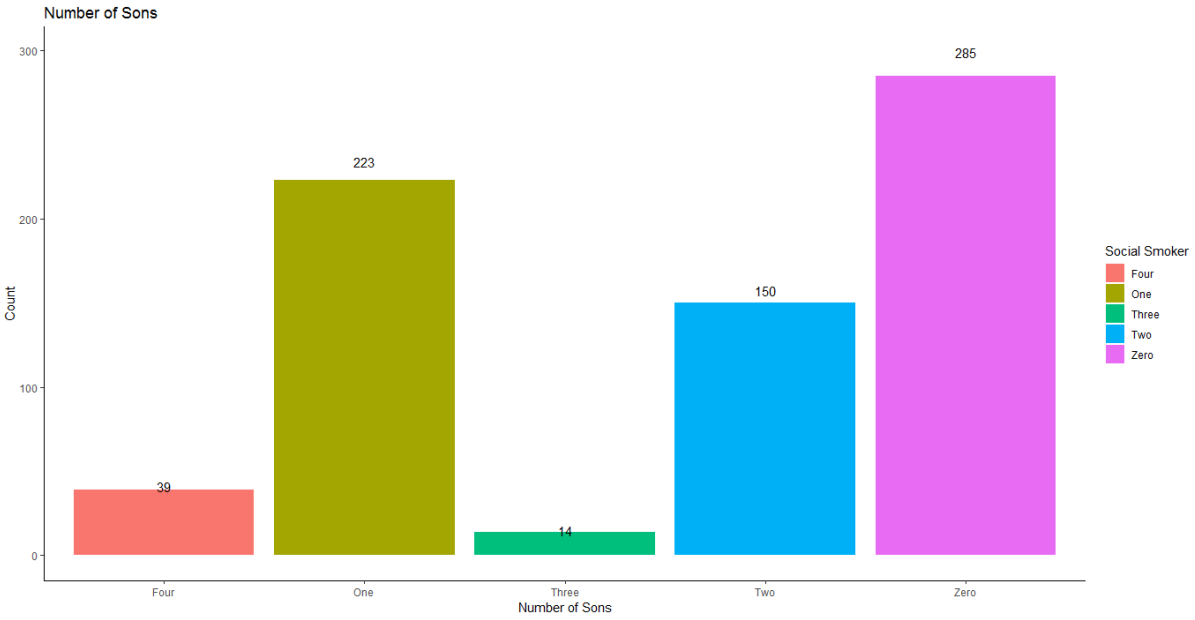

Week Days
155
147
150
138
121
150
100
50
0
Count
Friday
Monday
Thursday
Tuesday
Wednesday
Day of the Week
WeekDay
Friday
Monday
Thursday
Tuesday
Wednesday

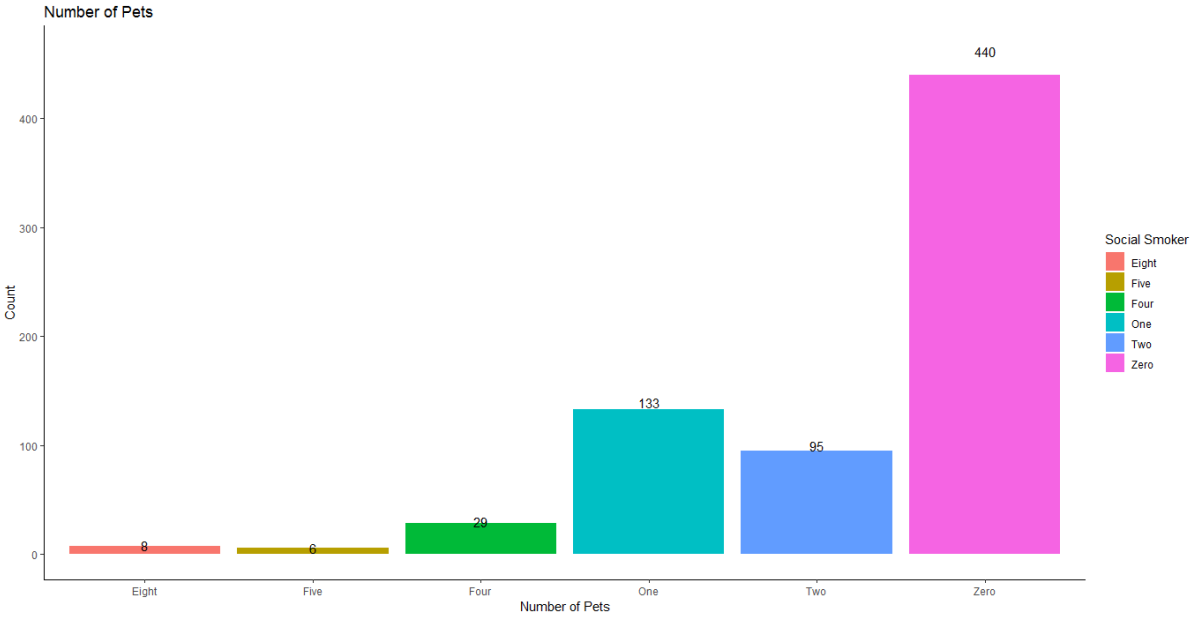

Season
200
150
100
50
0
Count
179
184
165
183
Autum
Spring
Summer
Winter
Season
Season
Autum
Spring
Summer
Winter


Education
600
400
200
0
Count
45
586
4
76
Graduate
High School
Master& Doctor
Post Graduate
Education
Education
Graduate
High School
Master& Doctor
Post Graduate


Social Drinker
405
306
400
300
200
100
0
Count
No
Yes
Social Drinker
Social Drinker
No
Yes


Social Smoker
658
600
400
200
0
Count
53
No
Yes
Social Smoker
Social Smoker
No
Yes


Number of Sons
300
200
100
0
Count
39
223
14
150
285
Four
One
Three
Two
Zero
Number of Sons
Social Smoker
Four
One
Three
Two
Zero

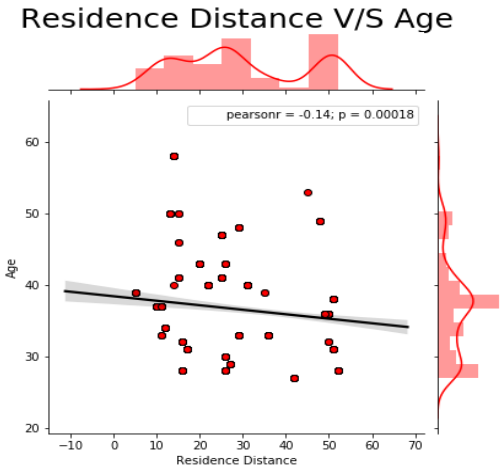

Number of Pets
400
300
200
100
0
Count
8
6
29
133
95
440
Eight
Five
Four
One
Two
Zero
Number of Pets
Social Smoker
Eight
Five
Four
One
Two
Zero

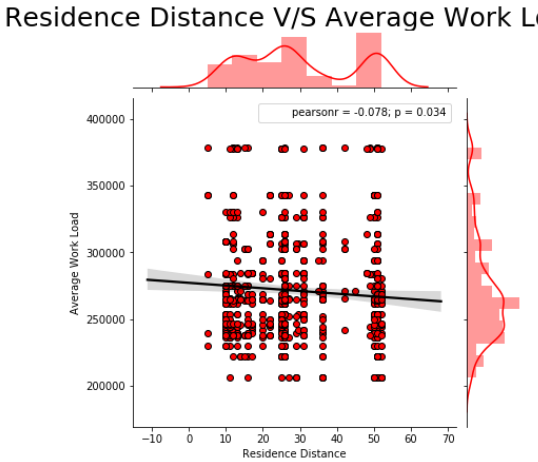
Bivariate Analysis

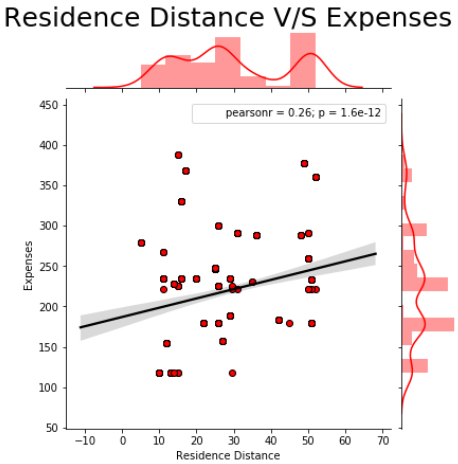
Bivariate analysis is the simultaneous analysis of two variables. It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

Continuous Variables

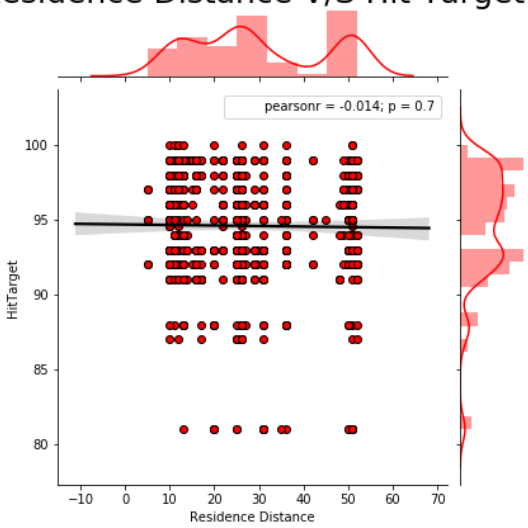
Scatter plot for all the continuous variables are shown below. Correlation value is shown in the legend and it describes the relationship between variables against which graphs are plotted.



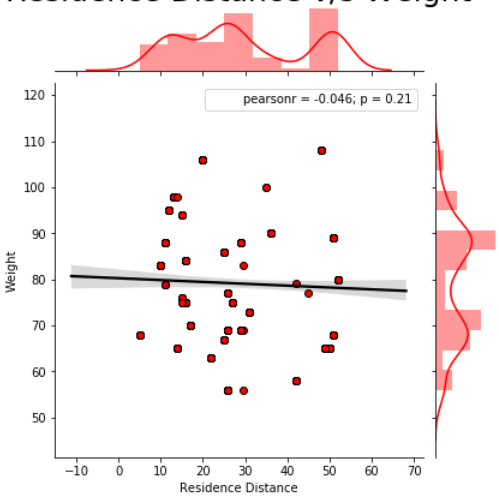




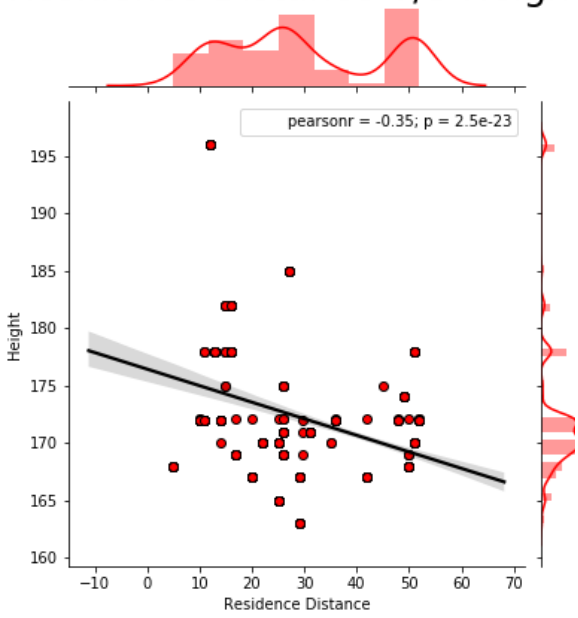
Residence Distance V/S Hit Target



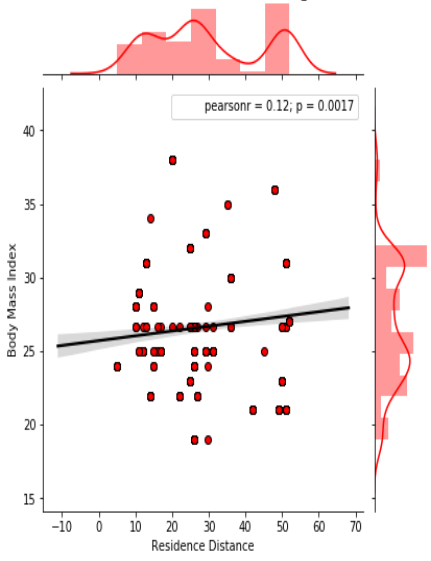
Residence Distance v/s Weight



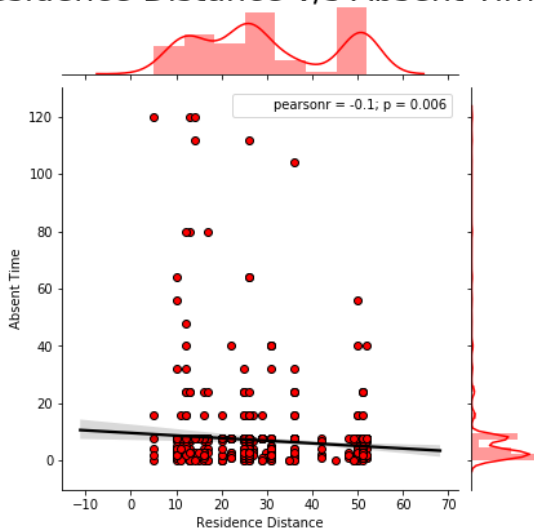
Residence Distance V/S Height



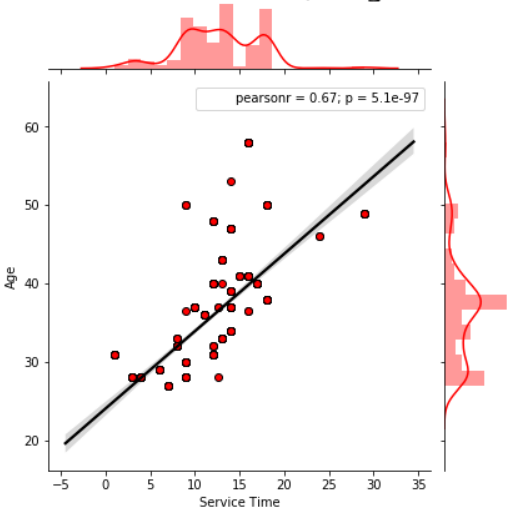
Residence Distance v/s Body Mass Index



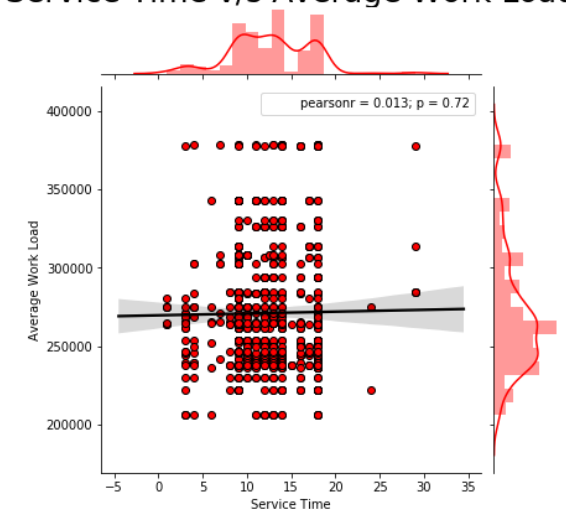
Residence Distance v/s Absent Time



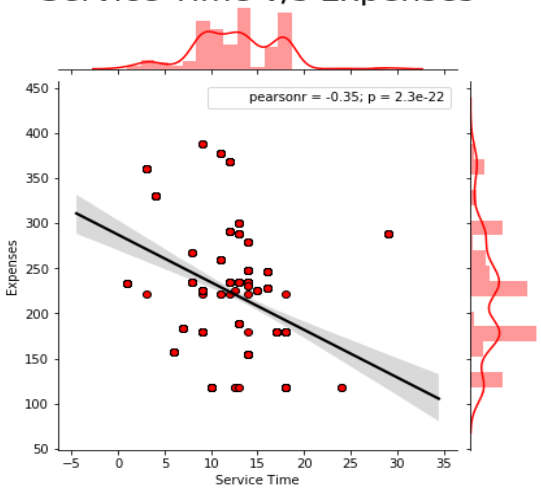
Service Time v/s Age



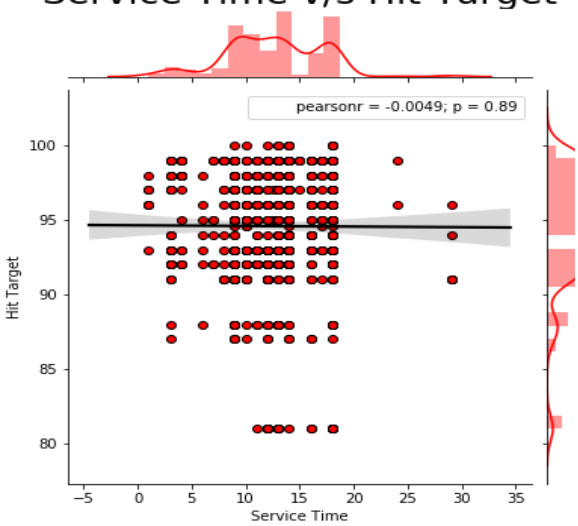
Service Time v/s Average Work Load



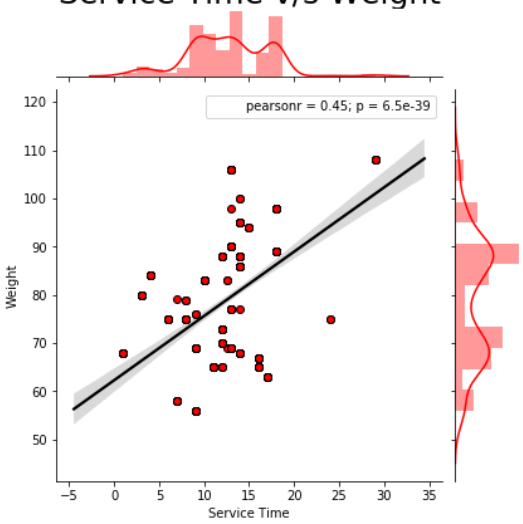
Service Time v/s Expenses



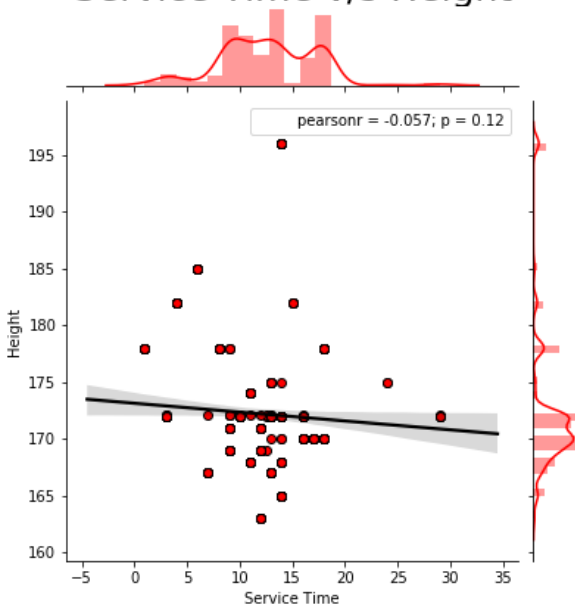
Service Time v/s Hit Target



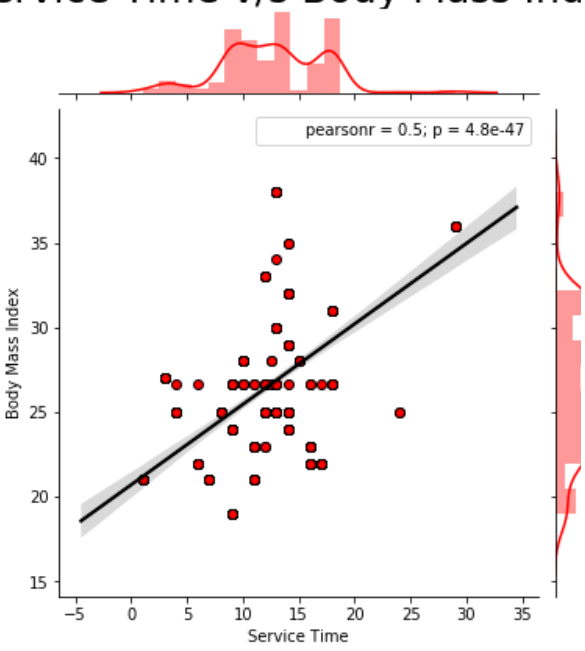
Service Time v/s Weight



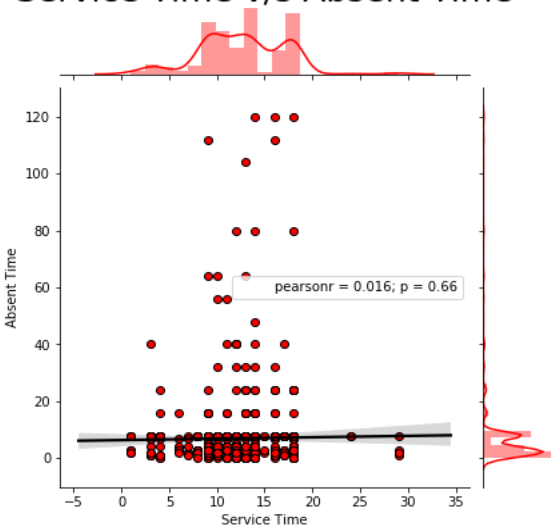
Service Time v/s Height



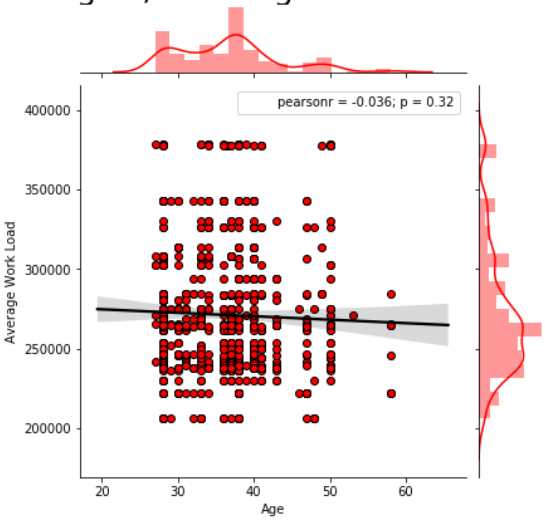
Service Time v/s Body Mass Ind



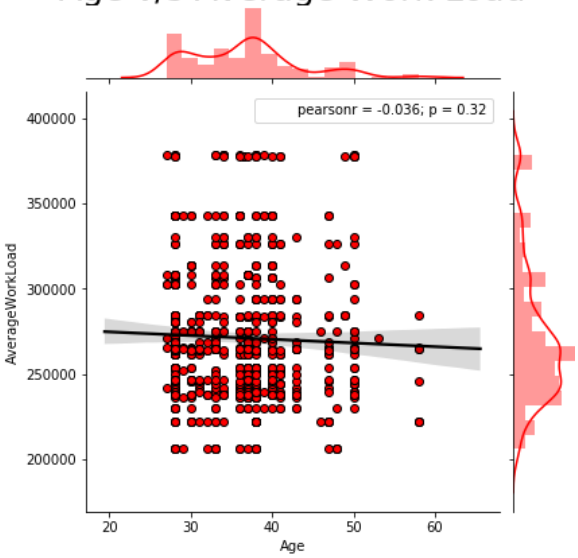
Service Time v/s Absent Time



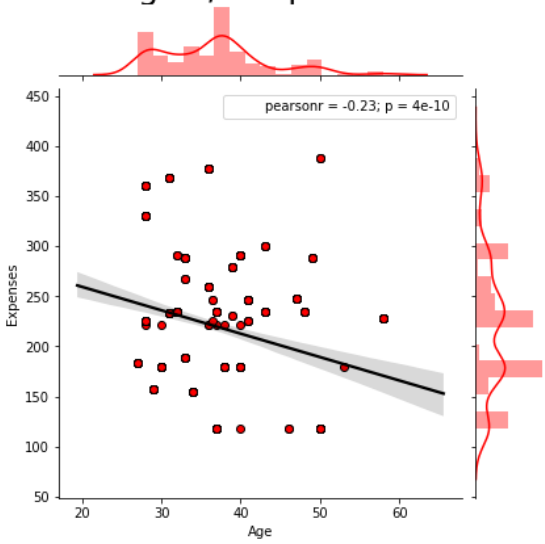
Age v/s Average Work Load



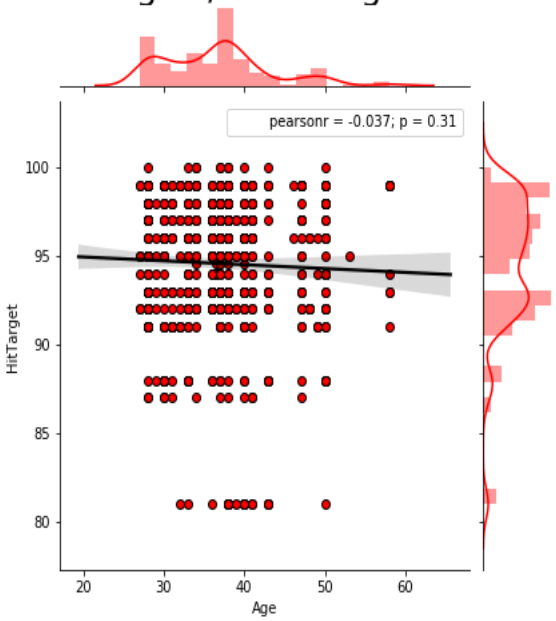
Age v/s Average Work Load



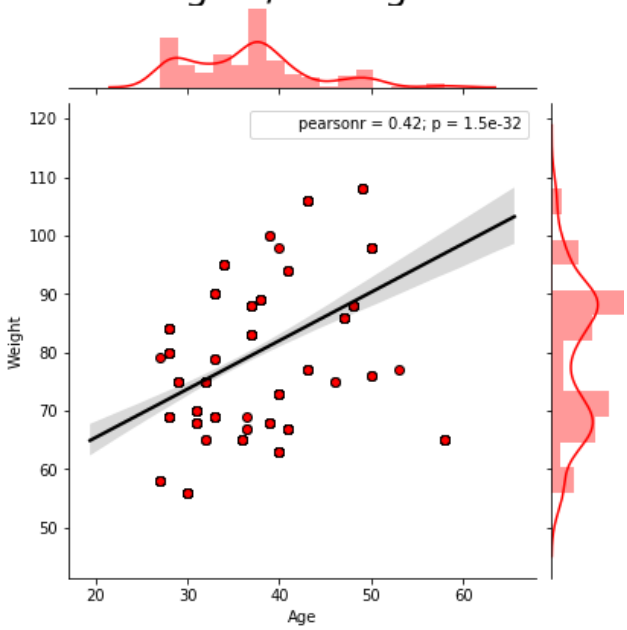
Age v/s Expenses



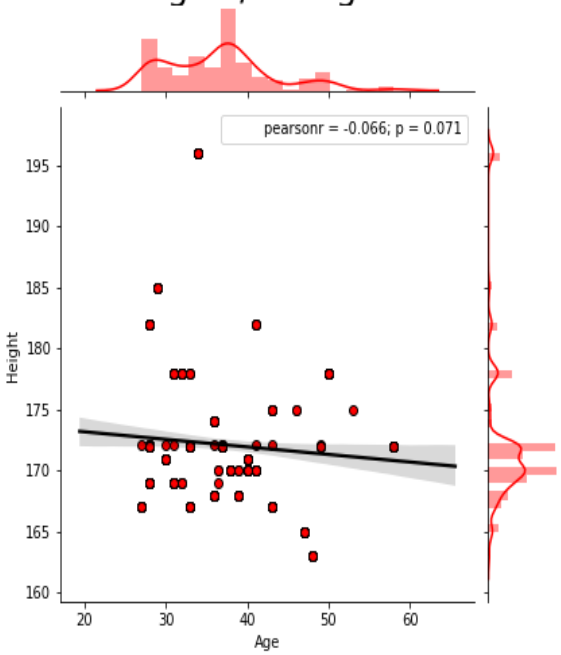
Age v/s Hit Target



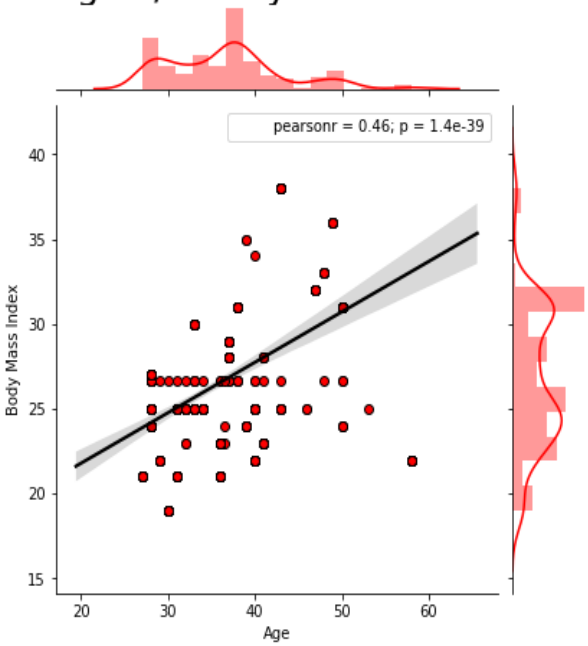
Age v/s Weight



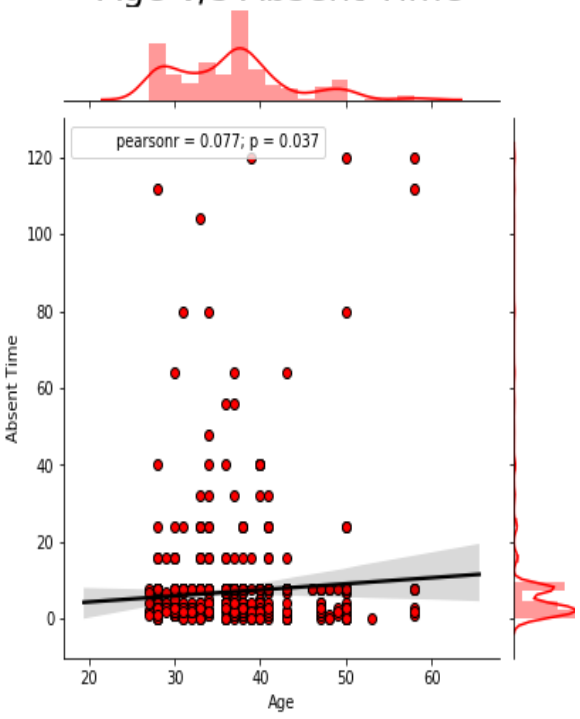
Age v/s Height



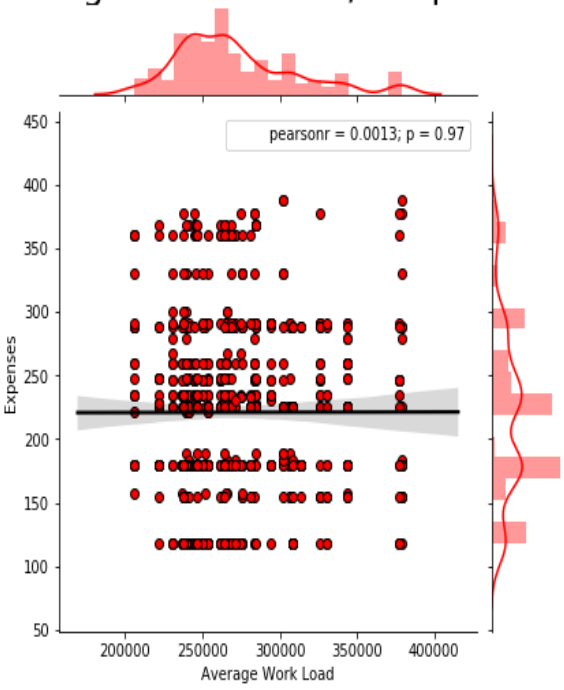
Age v/s Body Mass Index



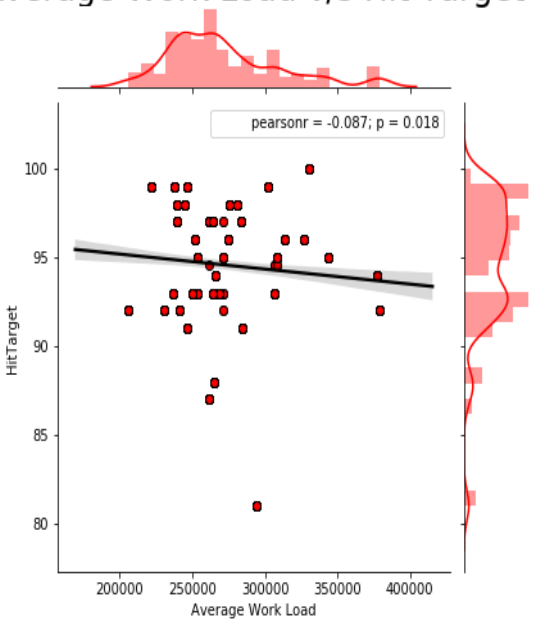
Age v/s Absent Time



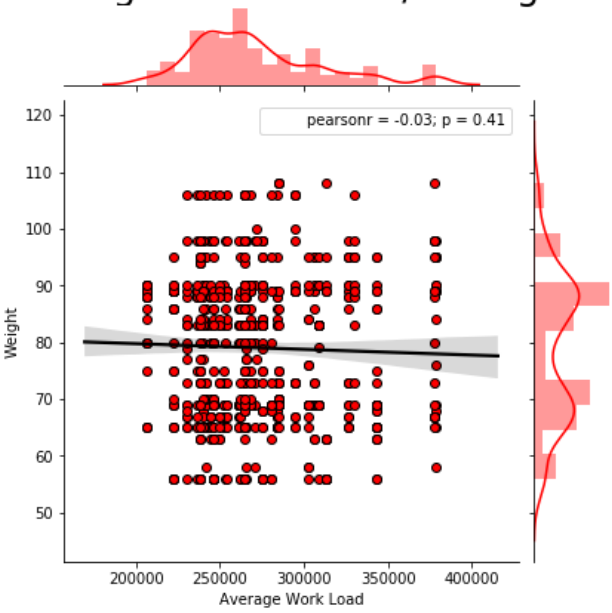
Average Work Load v/s Expenses



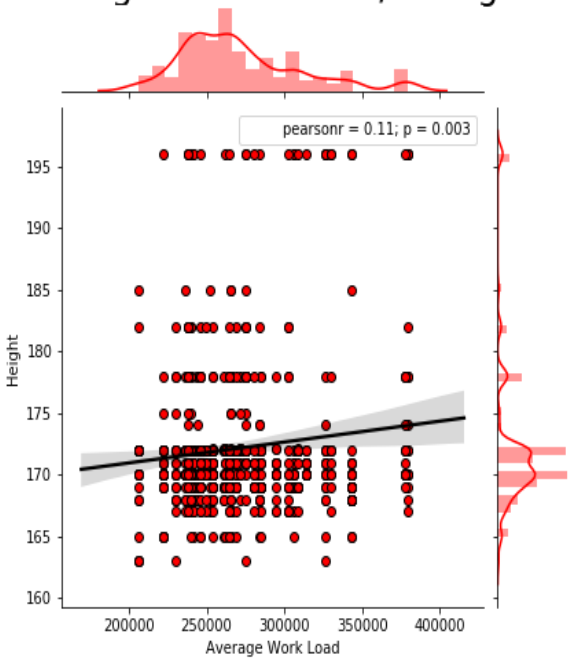
Average Work Load v/s Hit Target



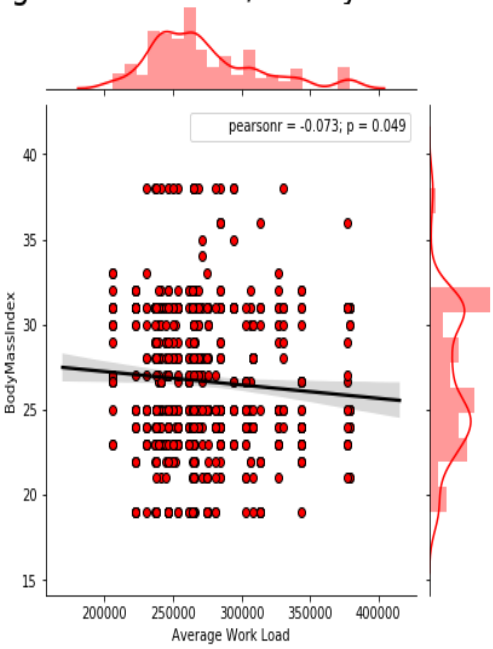
Average Work Load v/s Weight



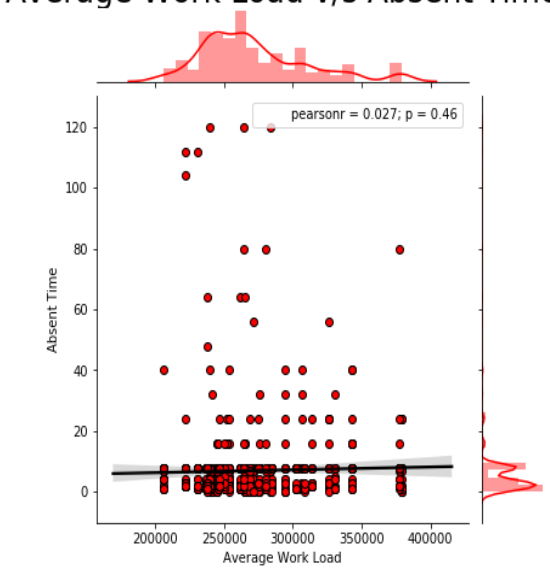
Average Work Load v/s Height



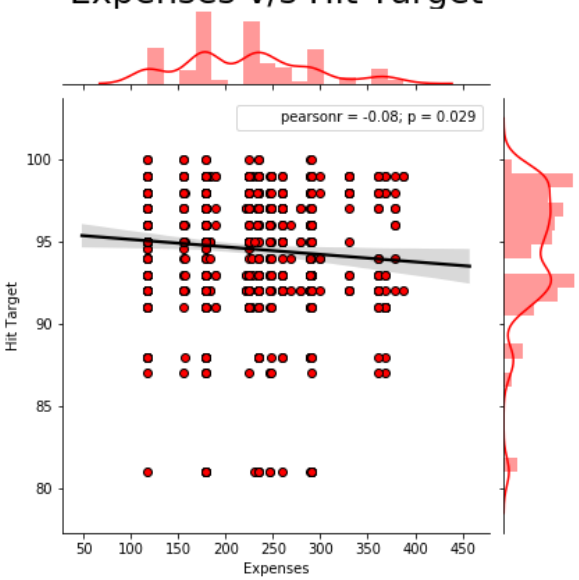
Average Work Load v/s Body Mass Index



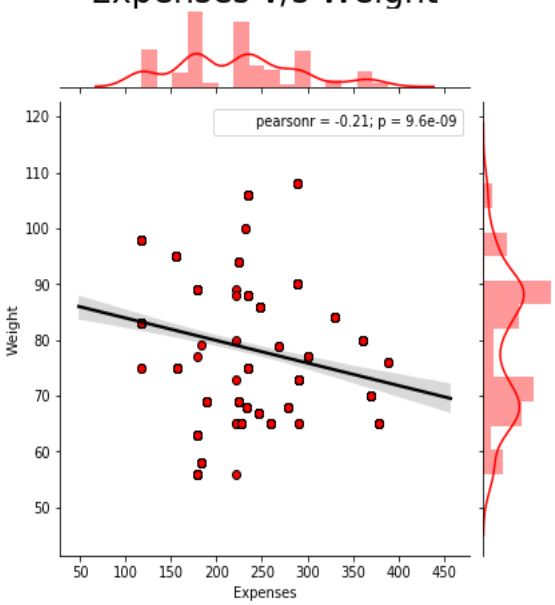
Average Work Load v/s Absent Time



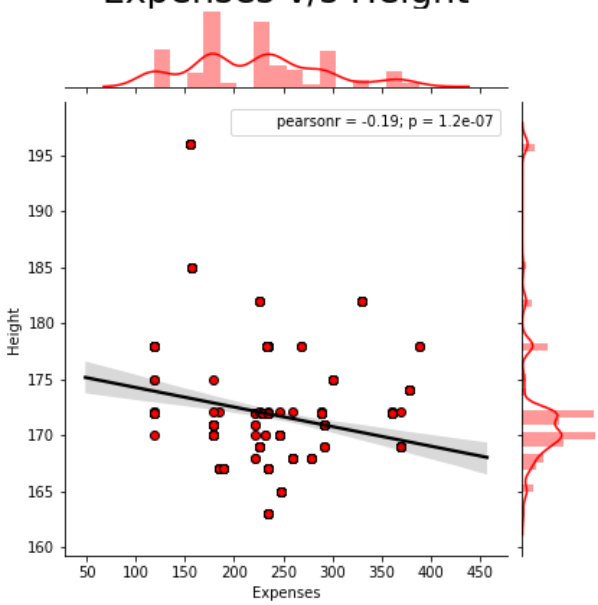
Expenses v/s Hit Target



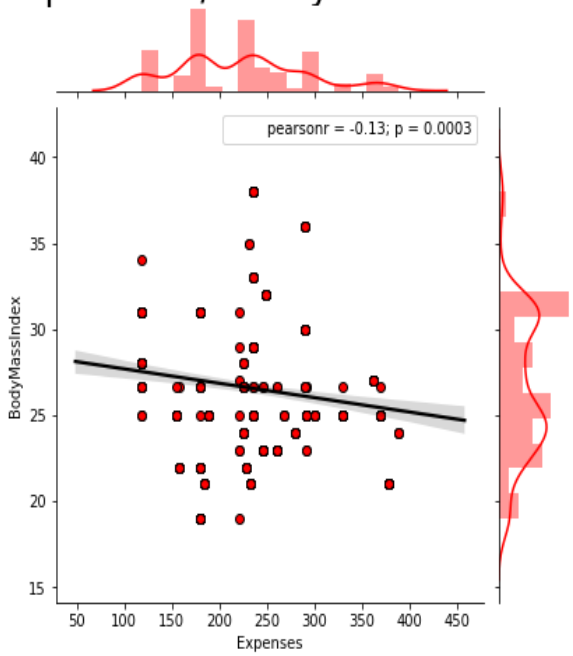
Expenses v/s Weight



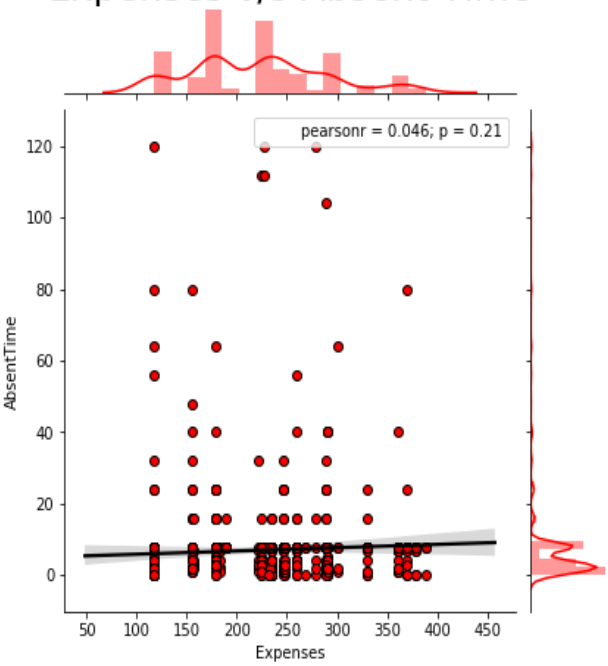
Expenses v/s Height



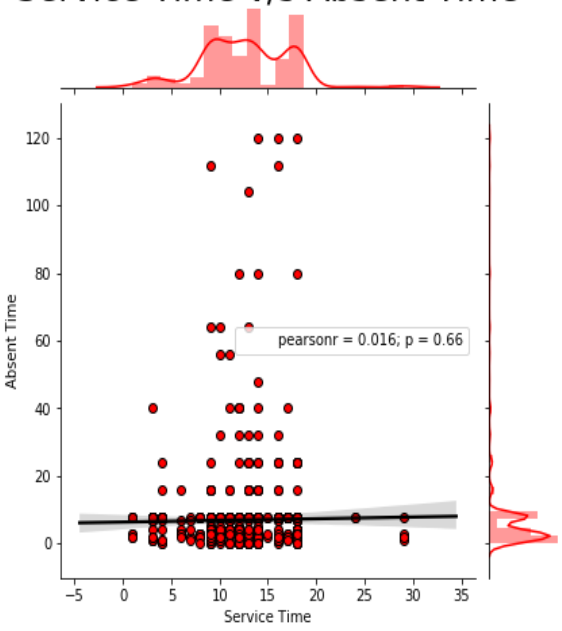
Expenses v/s Body Mass Index



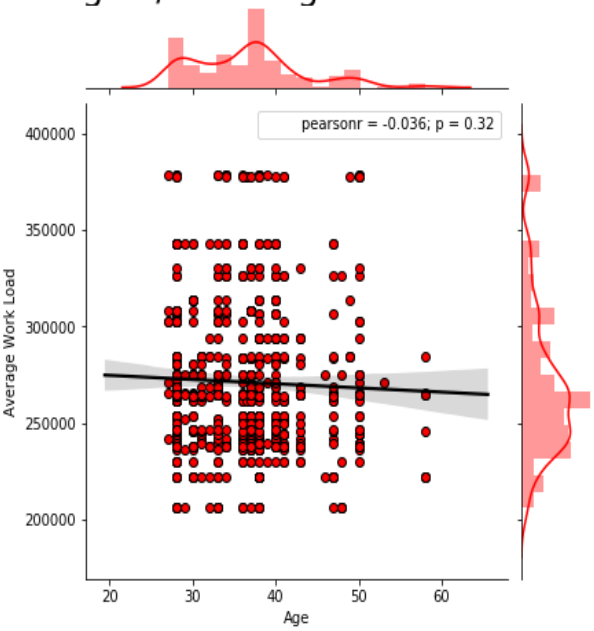
Expenses v/s Absent Time

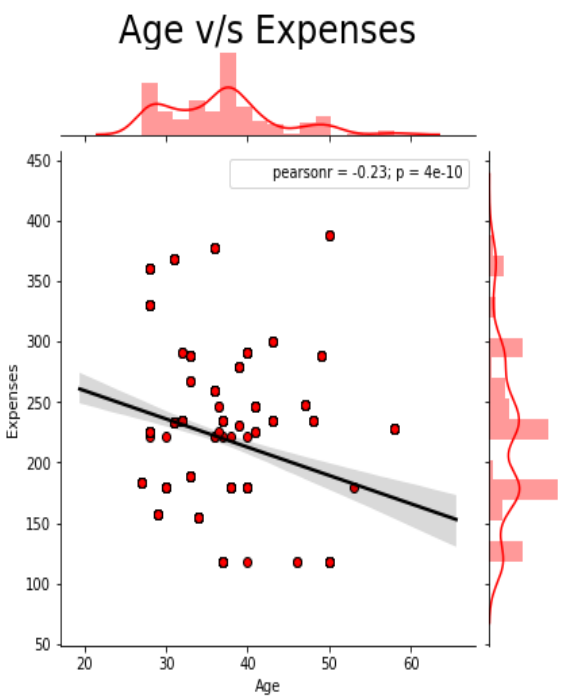


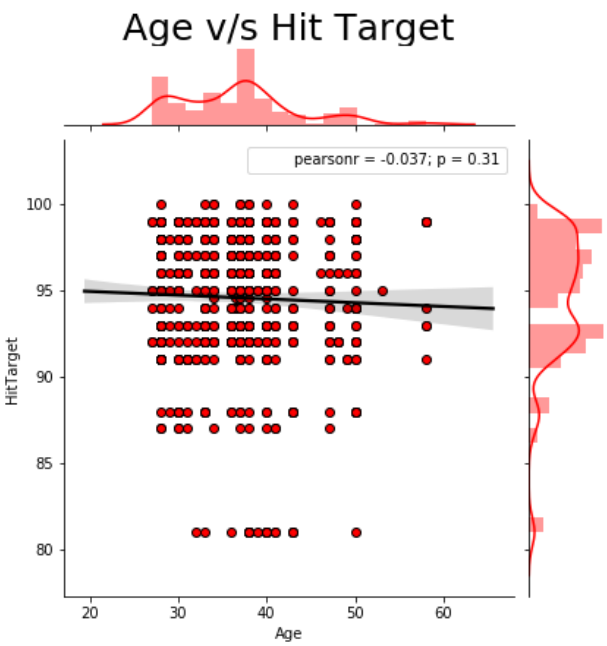
Service Time v/s Absent Time

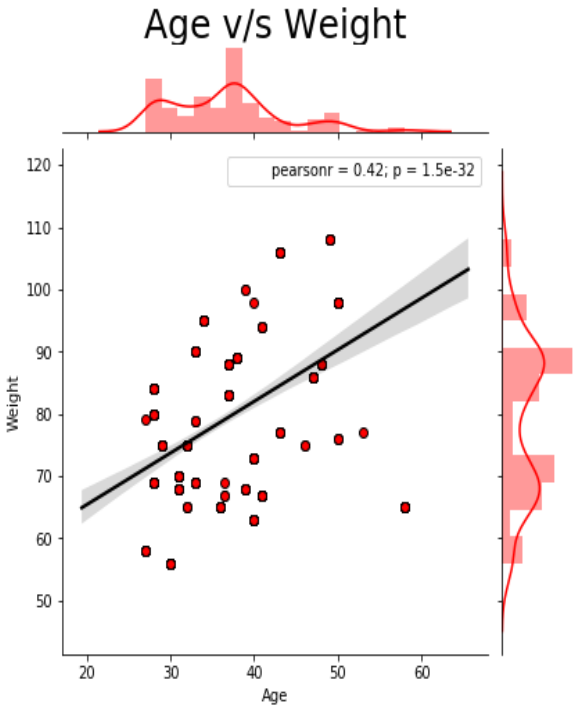


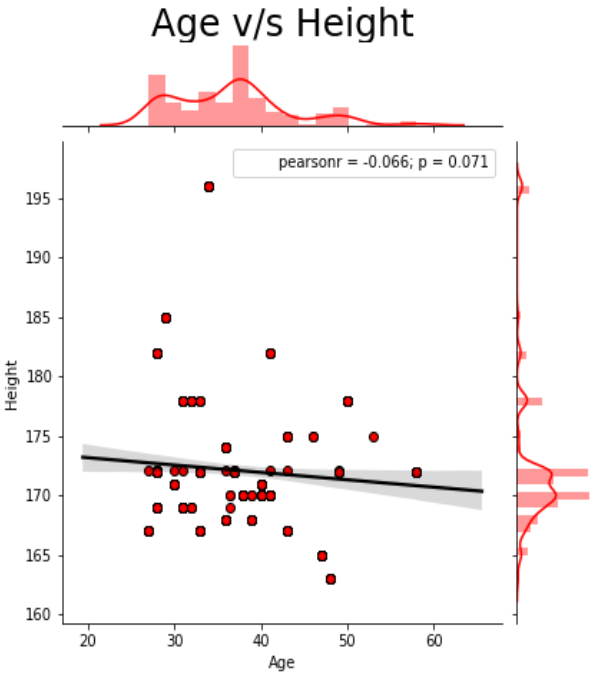
Age v/s Average Work Load



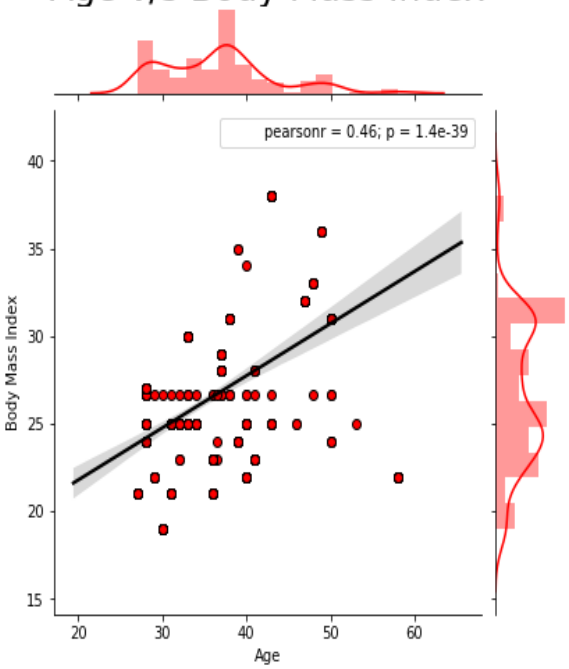

Age v/s Expenses
pearsonr = -0.23; p = 4e-10
Expenses
Age


Age v/s Hit Target
pearsonr = -0.037; p = 0.31
HitTarget
Age

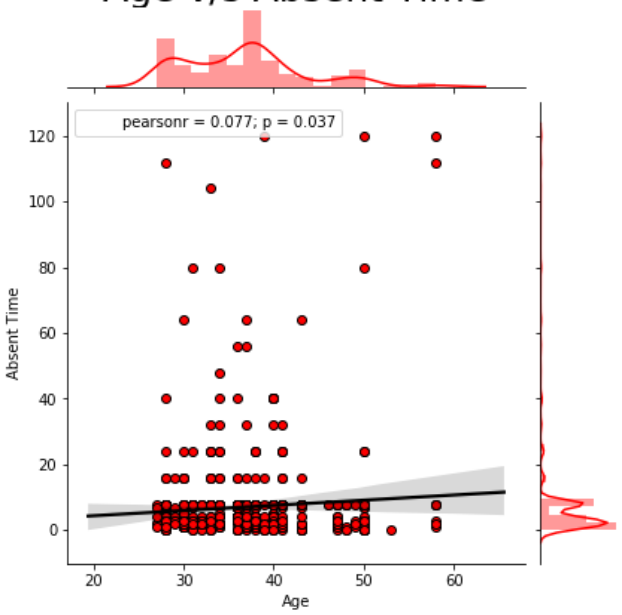

Age v/s Weight
pearsonr = 0.42; p = 1.5e-32
Weight
Age


Age v/s Height
pearsonr = -0.066; p = 0.071
Height
Age

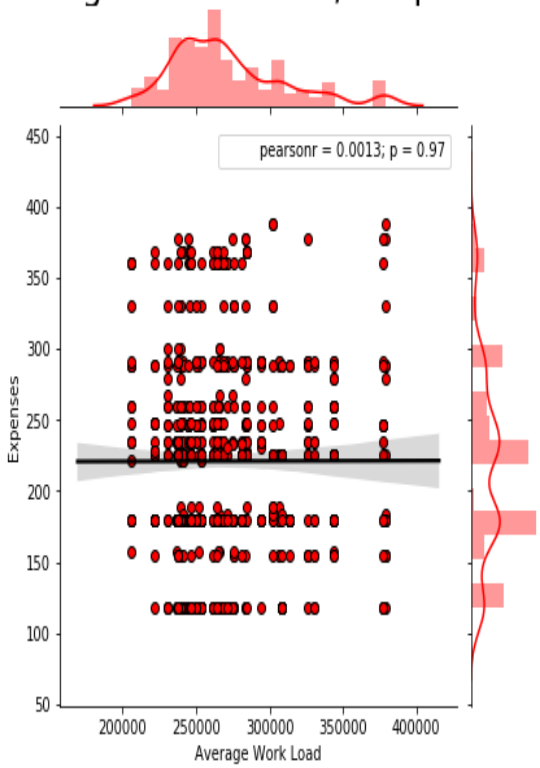
Age v/s Body Mass Index



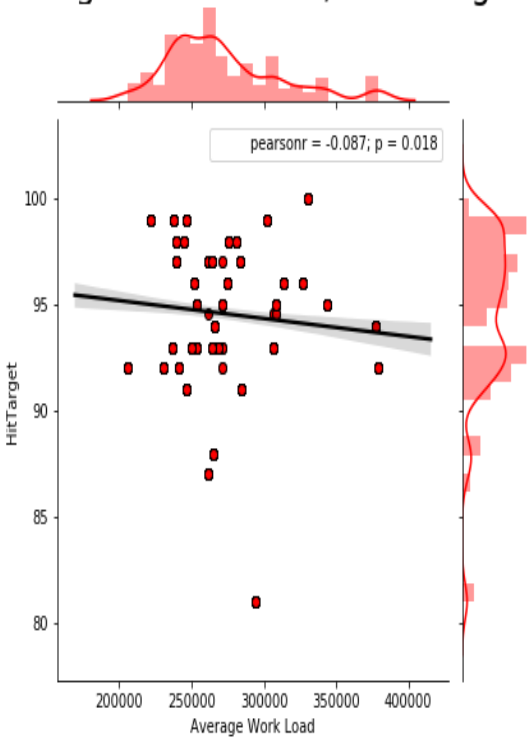
Age v/s Absent Time



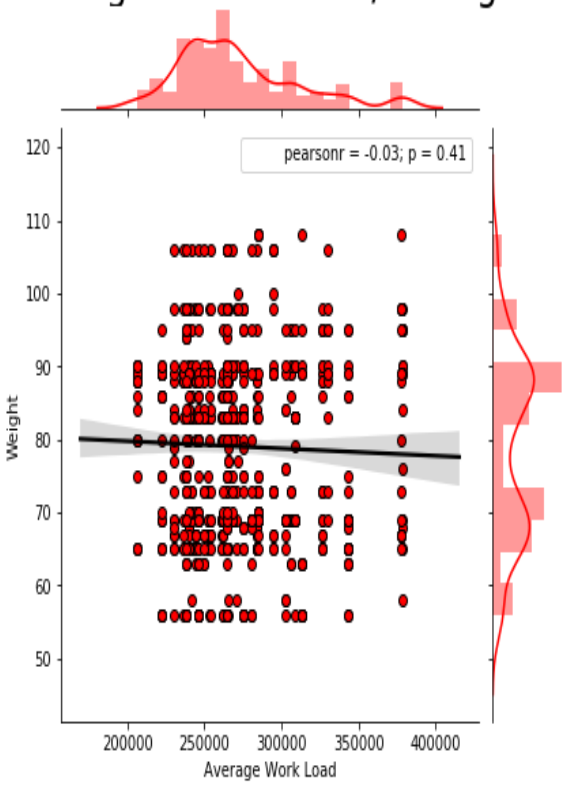
Average Work Load v/s Expenses



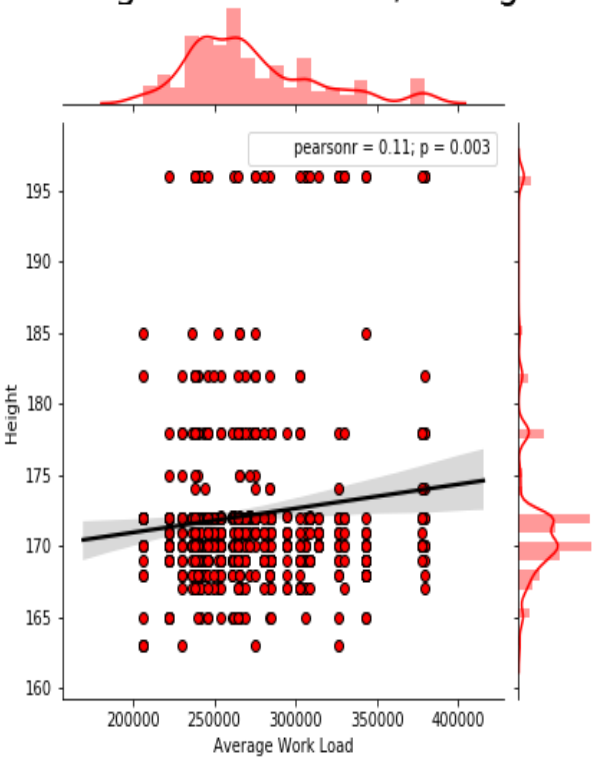
Average Work Load v/s Hit Target



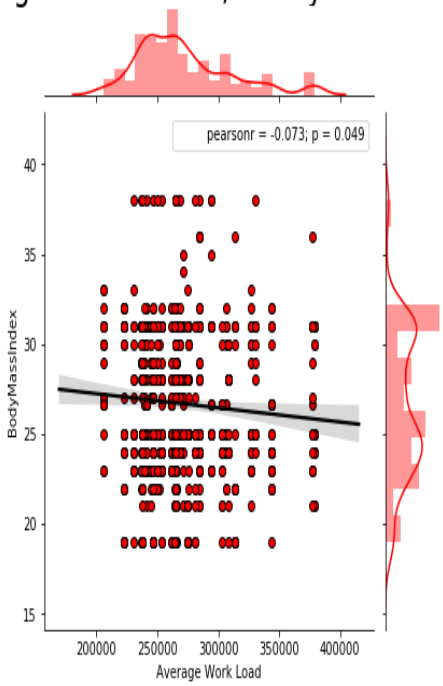
Average Work Load v/s Weight



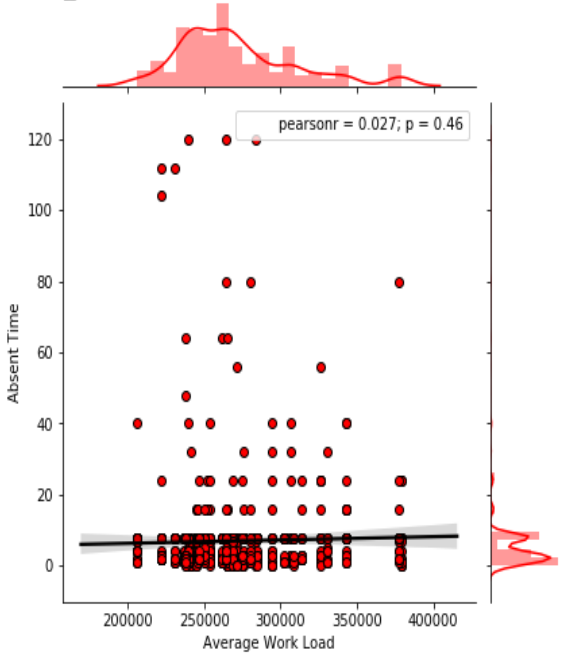
Average Work Load v/s Height



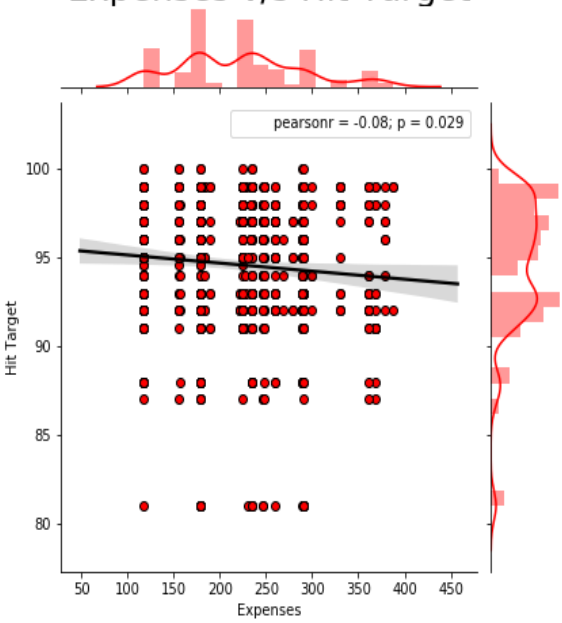
Average Work Load v/s Body Mass Index



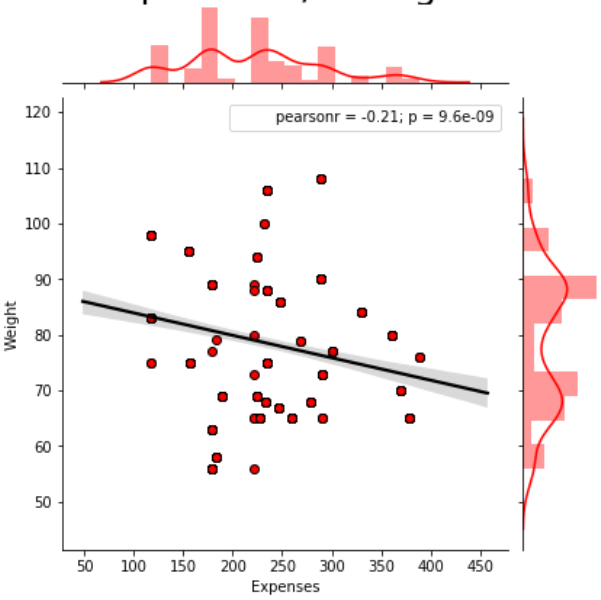
Average Work Load v/s Absent Time



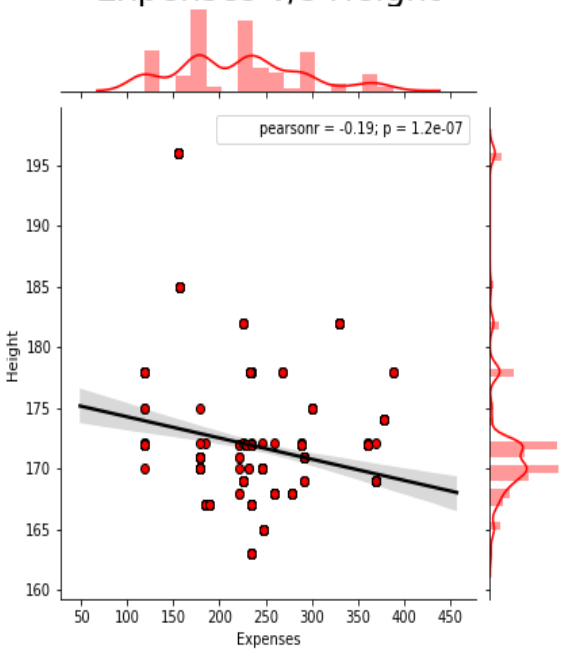
Expenses v/s Hit Target



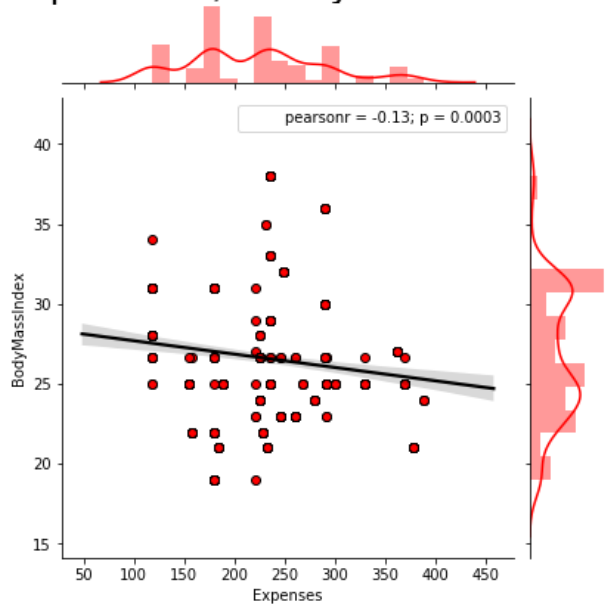
Expenses v/s Weight



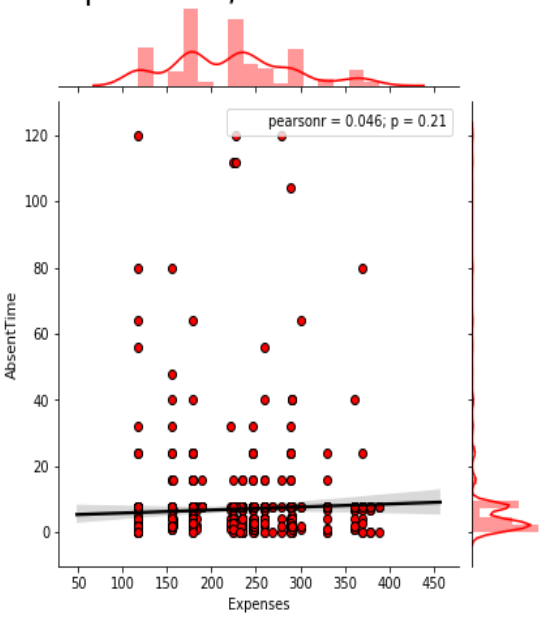
Expenses v/s Height



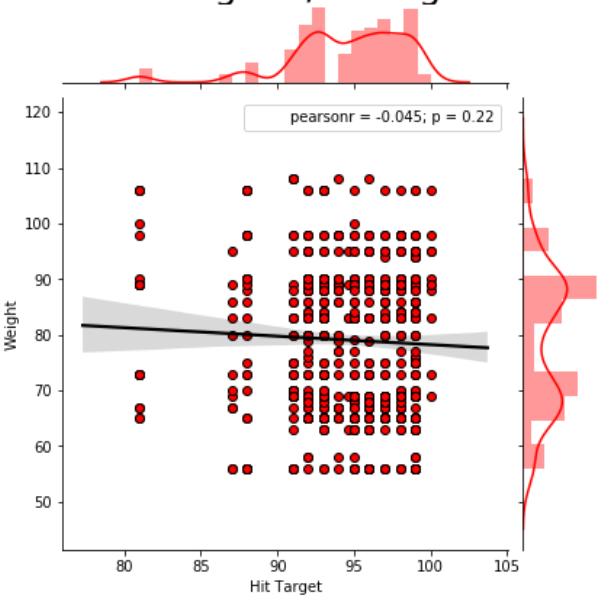
Expenses v/s Body Mass Index



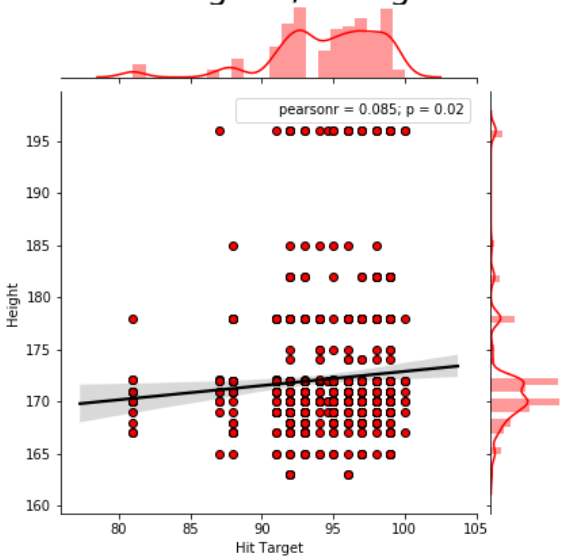
Expenses v/s Absent Time



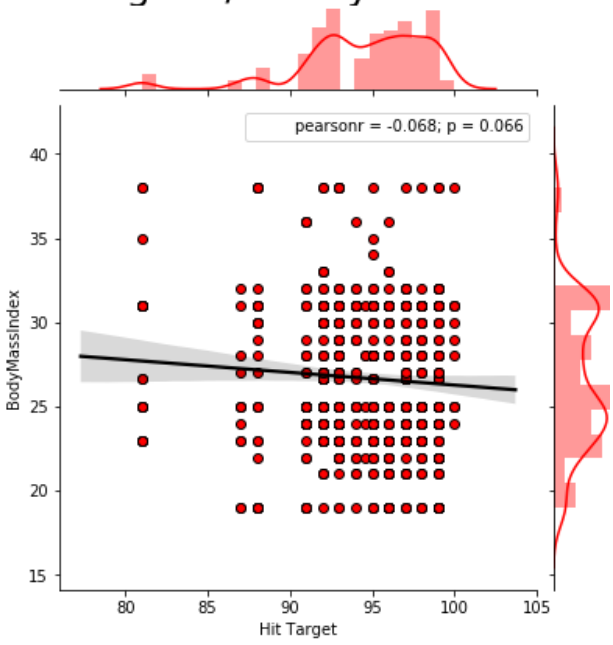
Hit Target v/s Weight



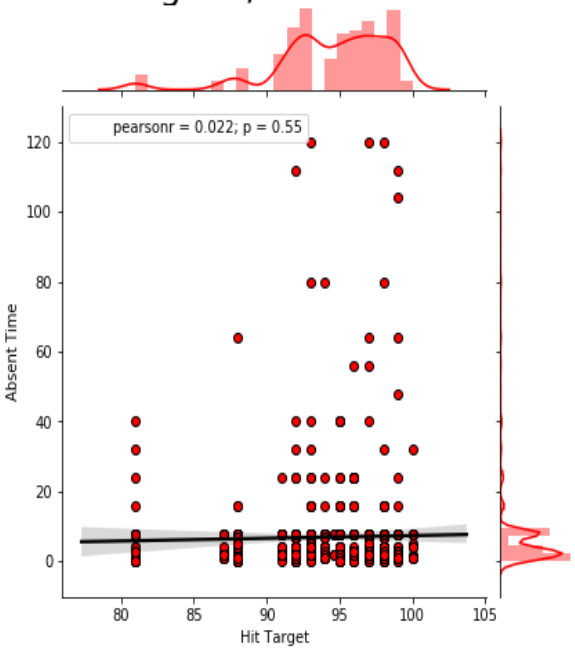
Hit Target v/s Height



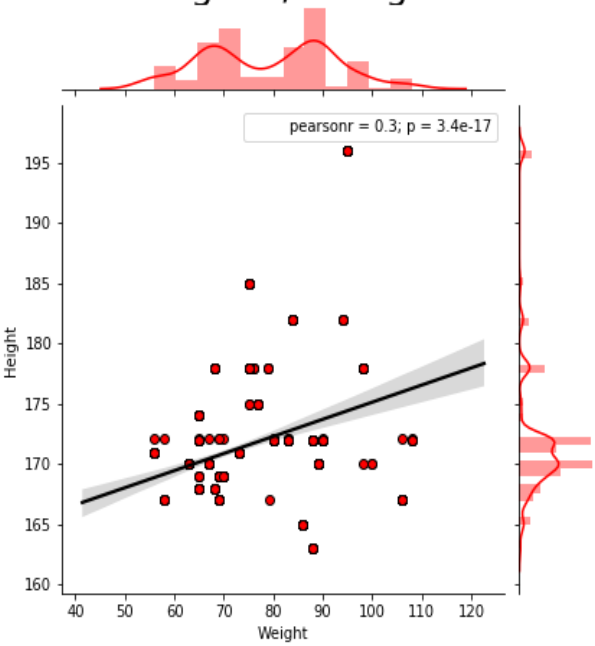
Hit Target v/s Body Mass Index



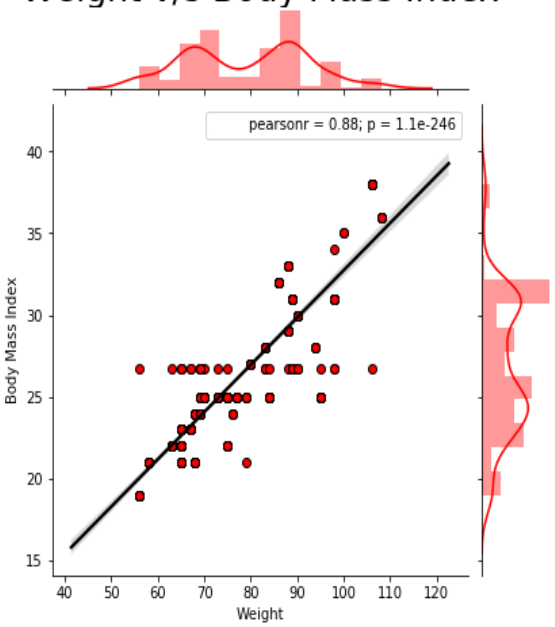
Hit Target v/s Absent Time



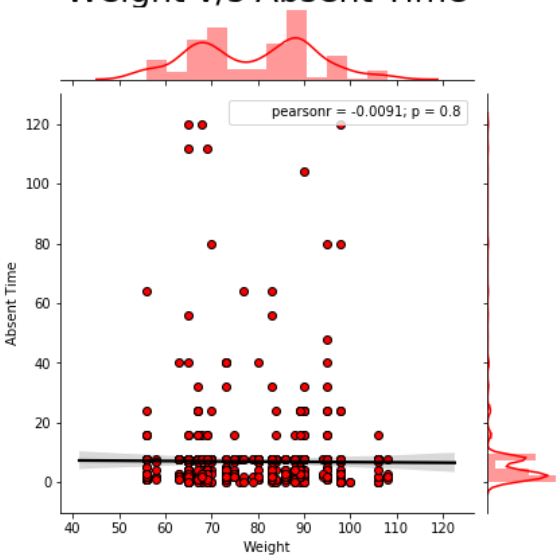
Weight v/s Height



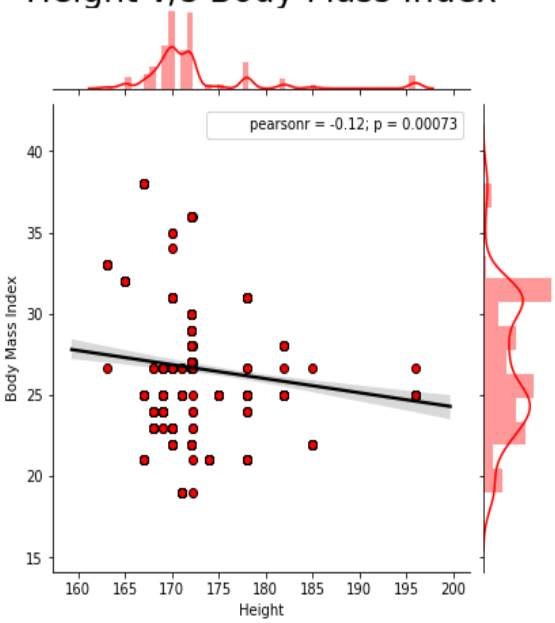
Weight v/s Body Mass Index



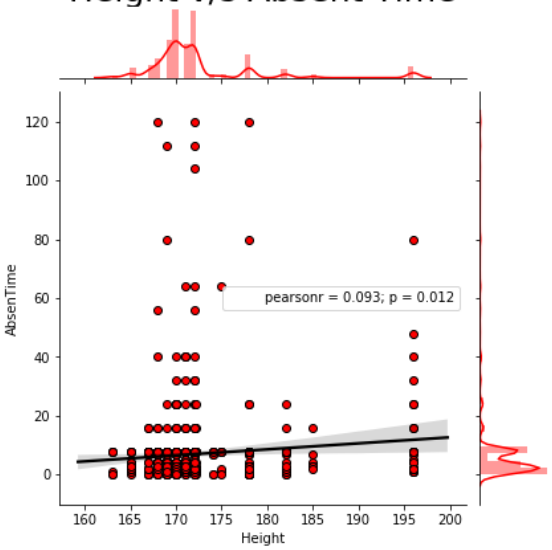
Weight v/s Absent Time



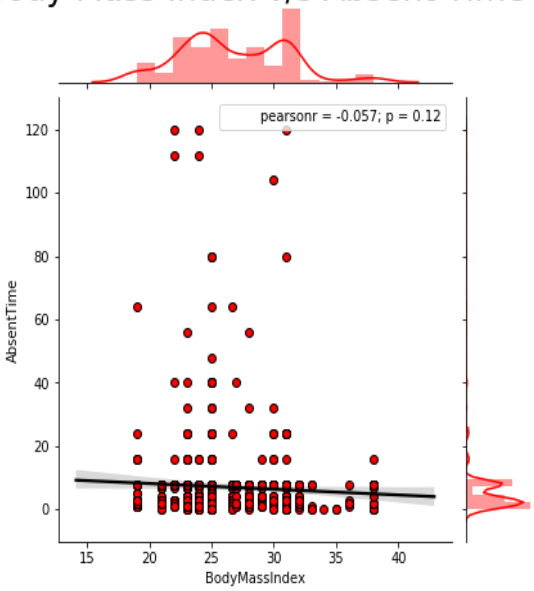
Height v/s Body Mass Index



Height v/s Absent Time

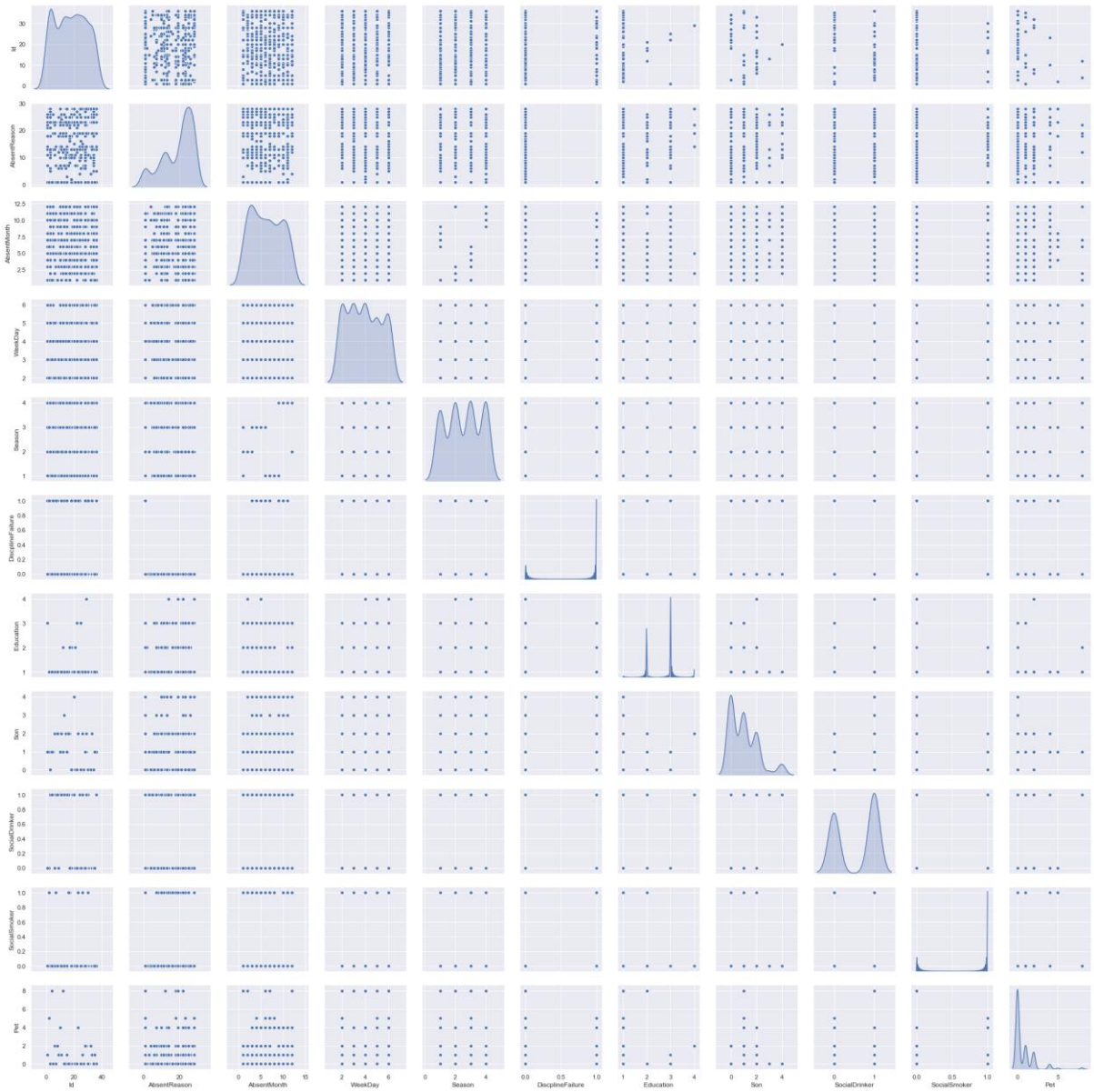


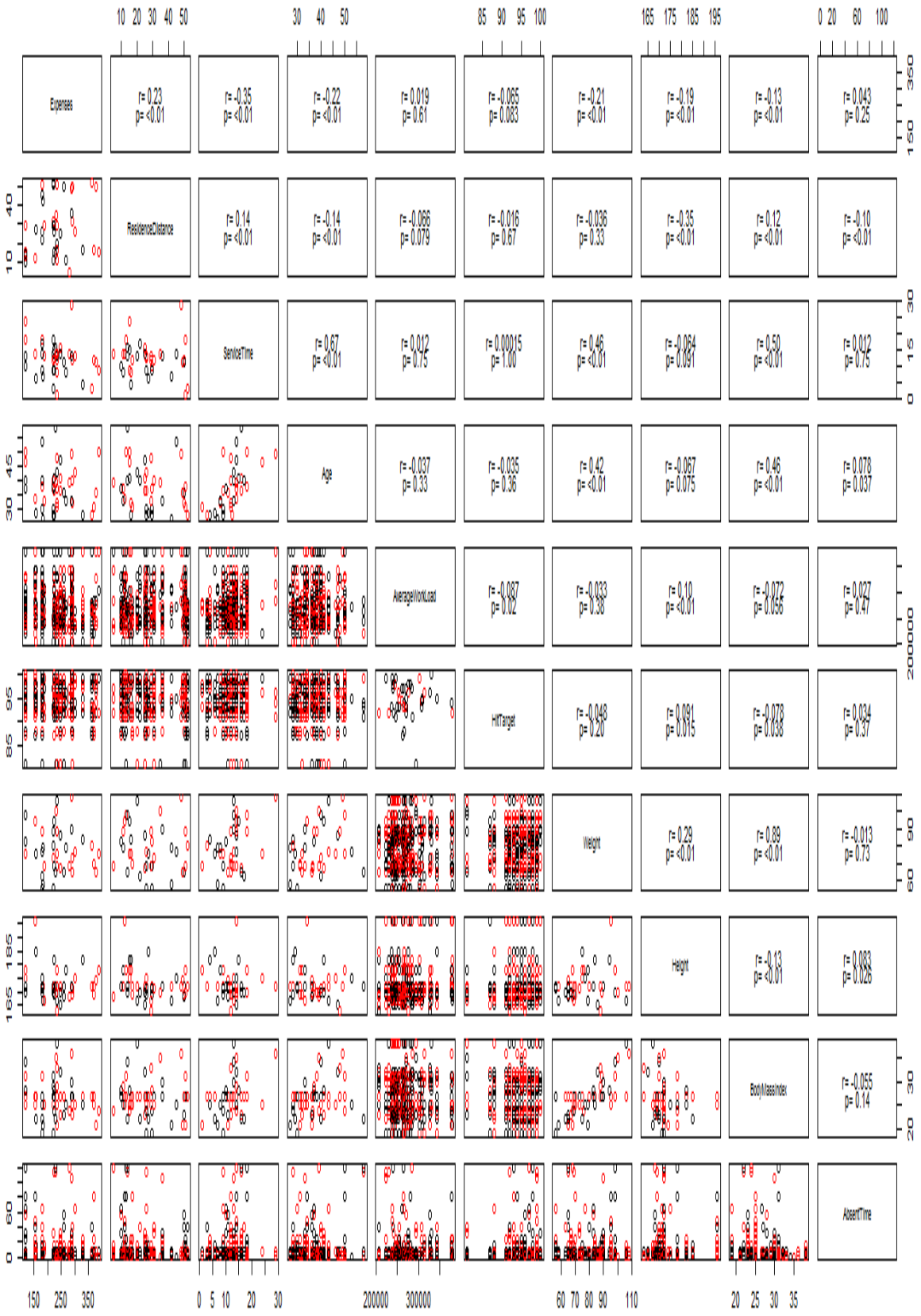
Body Mass Index v/s Absent Time



Pair Plot

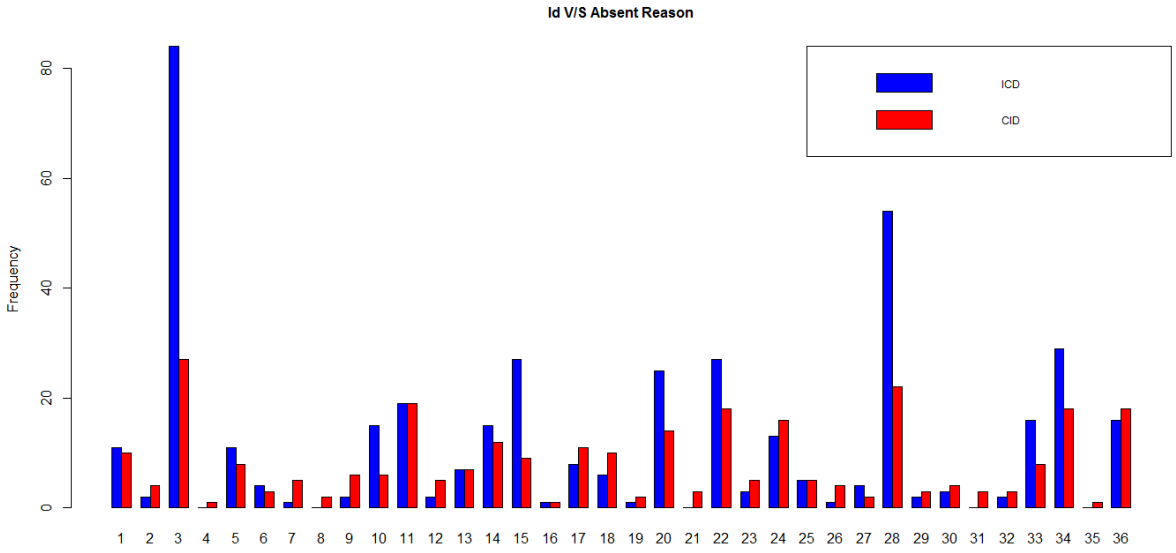
Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our dataset. Pair plots for all our numerical variables are shown below:-

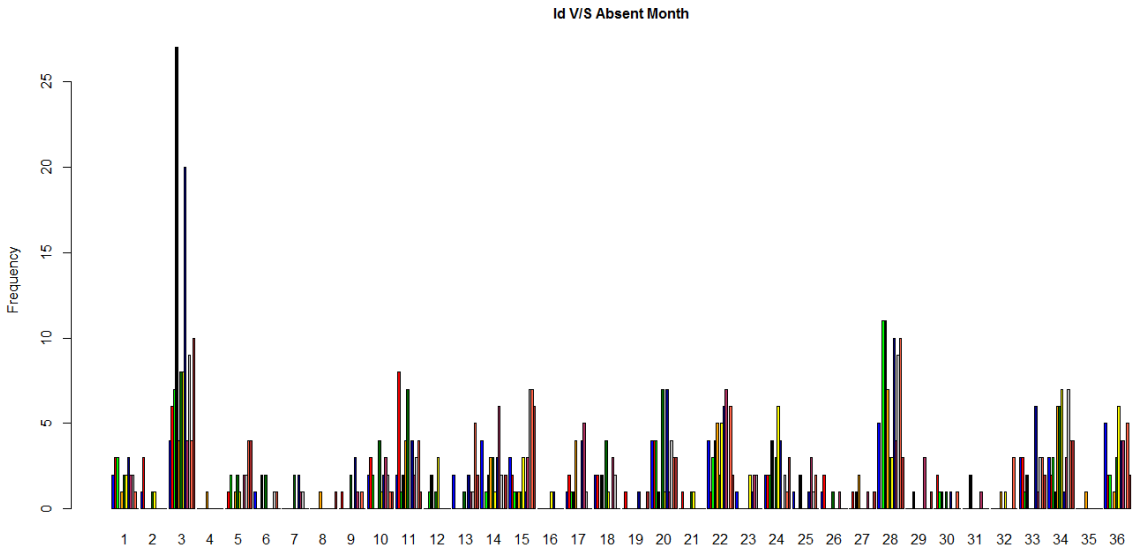


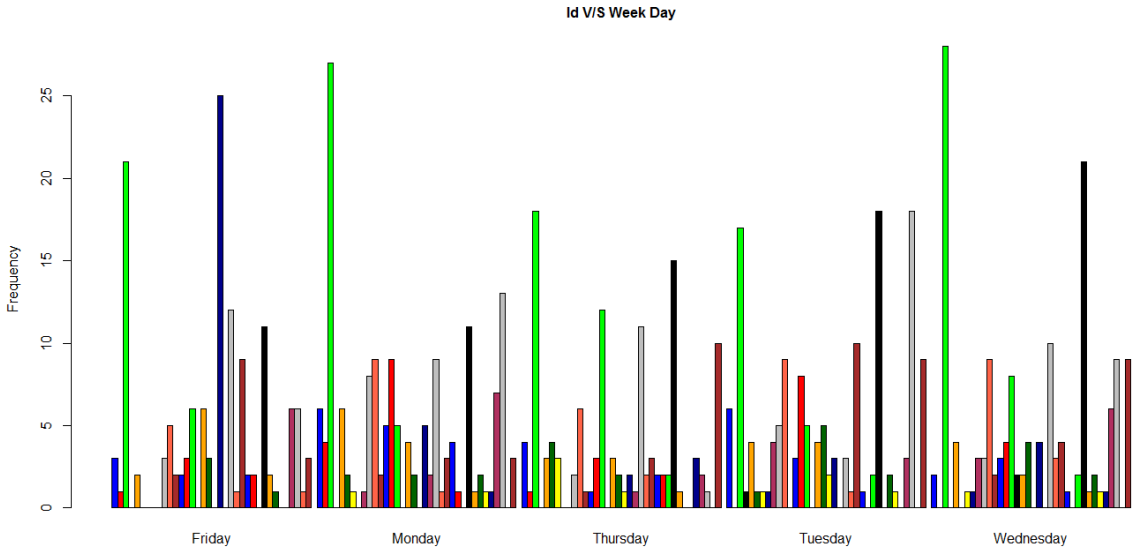

Expense
r= 0.23 p= <0.01
r= -0.35 p= <0.01
r= -0.22 p= <0.01
r= 0.019 p= 0.61
r= -0.065 p= 0.083
r= -0.21 p= <0.01
r= -0.19 p= <0.01
r= -0.13 p= <0.01
r= 0.043 p= 0.25
ResidenceDistance
r= 0.14 p= <0.01
r= -0.14 p= <0.01
r= -0.066 p= 0.079
r= -0.016 p= 0.67
r= -0.036 p= 0.33
r= -0.35 p= <0.01
r= 0.12 p= <0.01
r= -0.10 p= <0.01
ServiceTime
r= 0.67 p= <0.01
r= 0.012 p= 0.75
r= 0.00015 p= 1.00
r= 0.46 p= <0.01
r= -0.064 p= 0.091
r= 0.50 p= <0.01
r= 0.012 p= 0.75
Age
r= -0.037 p= 0.33
r= -0.035 p= 0.36
r= 0.42 p= <0.01
r= -0.067 p= 0.075
r= 0.46 p= <0.01
r= 0.078 p= 0.037
AverageWorkLoad
r= -0.007 p= 0.82
r= -0.033 p= 0.38
r= 0.10 p= <0.01
r= -0.072 p= 0.056
r= 0.027 p= 0.47
HitTarget
r= -0.048 p= 0.20
r= 0.091 p= 0.015
r= -0.078 p= 0.038
r= 0.034 p= 0.37
Weight
r= 0.29 p= <0.01
r= 0.89 p= <0.01
r= -0.013 p= 0.73
Height
r= -0.13 p= <0.01
r= 0.083 p= 0.026
BodyMassIndex
r= -0.055 p= 0.14
AbsentTime

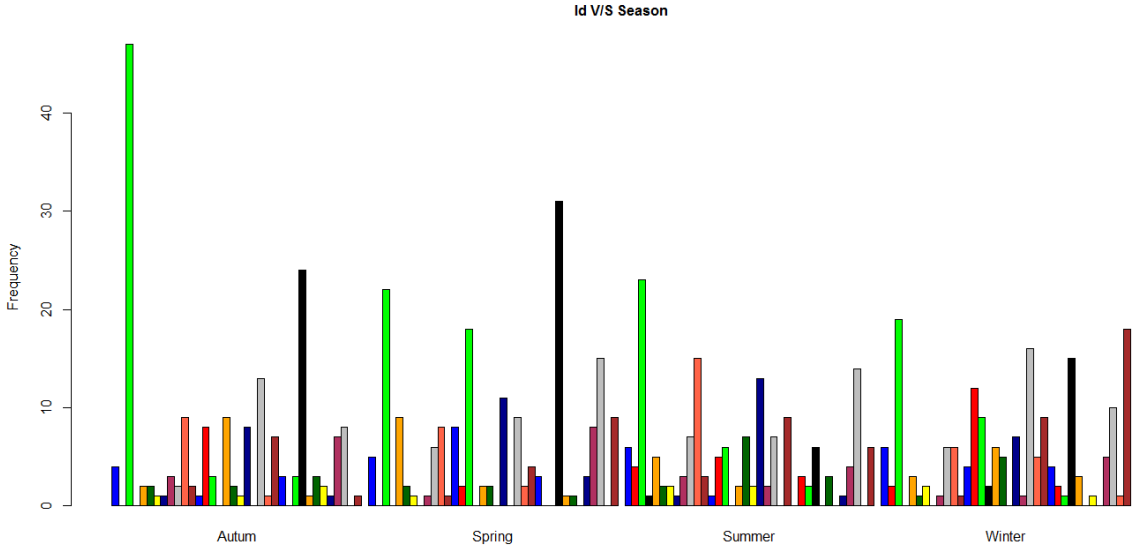
Bivariate Analysis of Categorical Variables

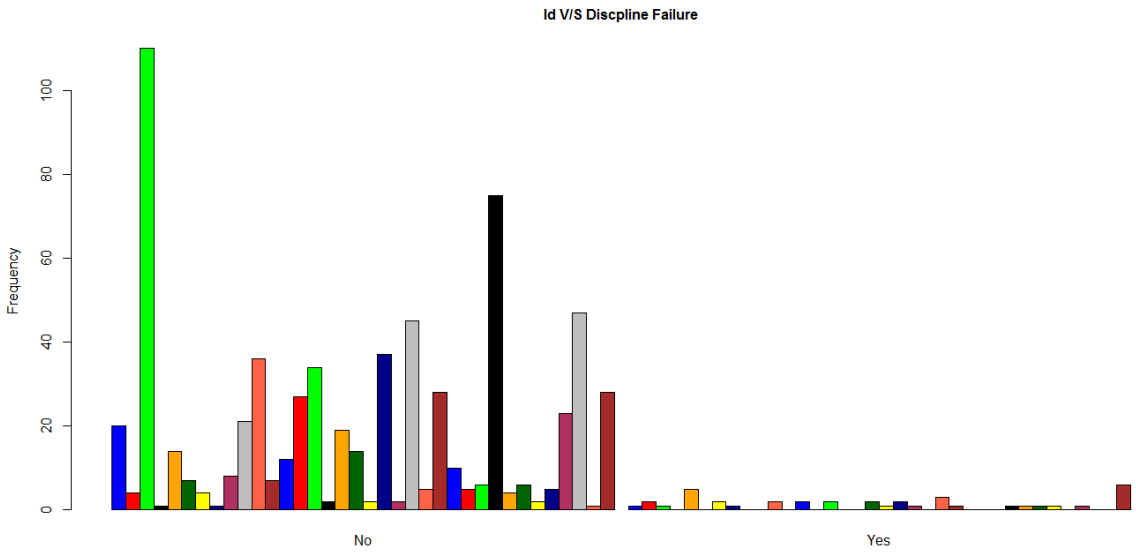
Bivariate analyses of categorical variables are shown below in the form of visualizations:-

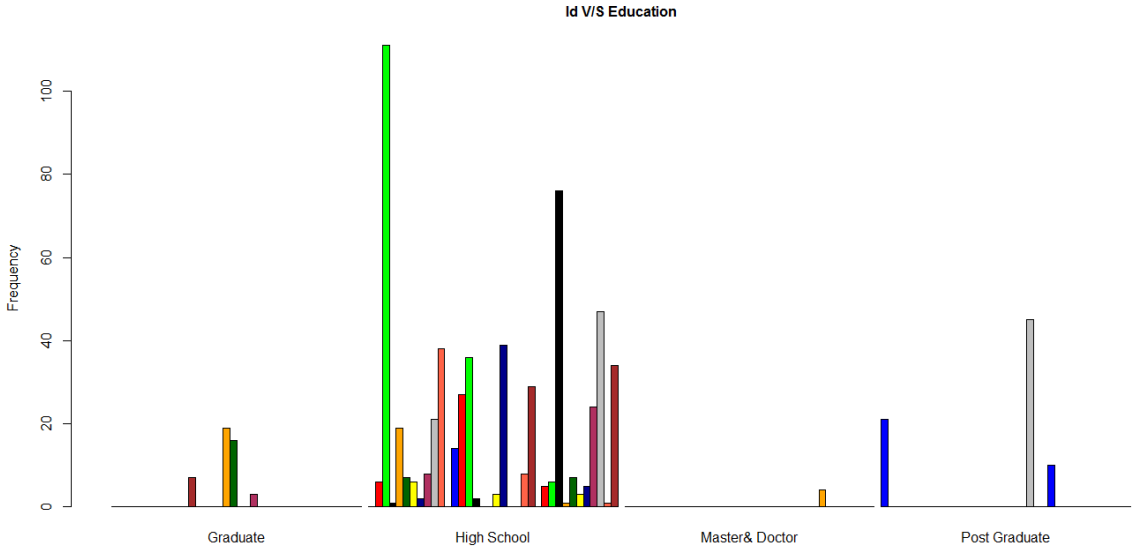


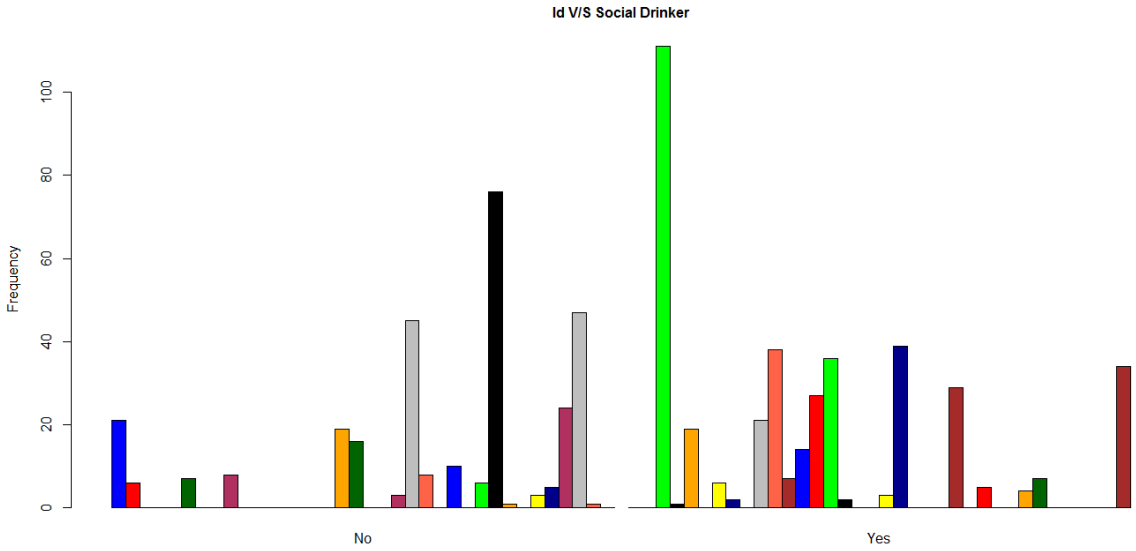


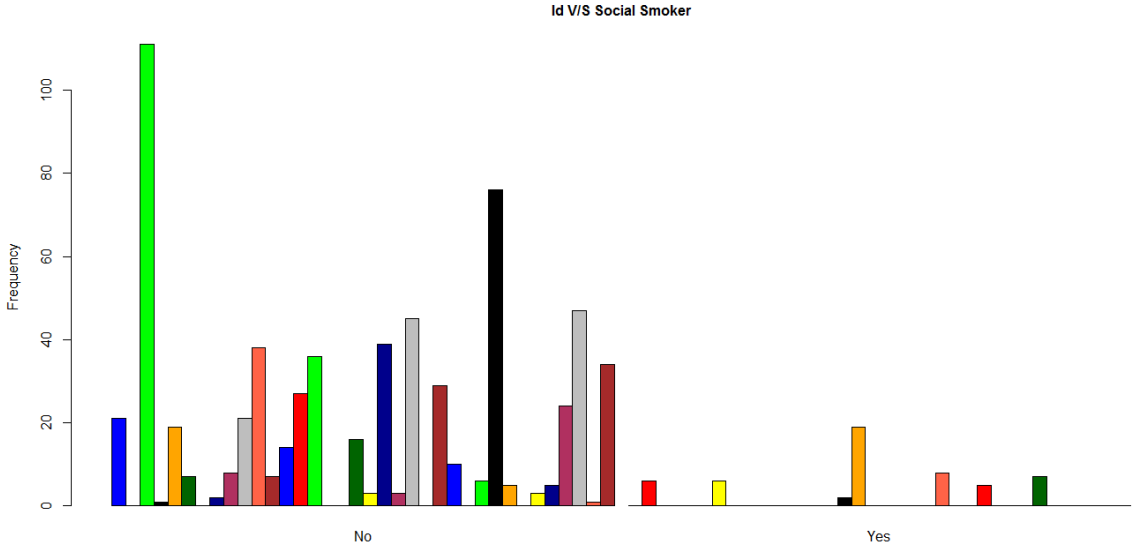

Id V/S Week Day
Frequency
0
5
10
15
20
25
Friday
Monday
Thursday
Tuesday
Wednesday

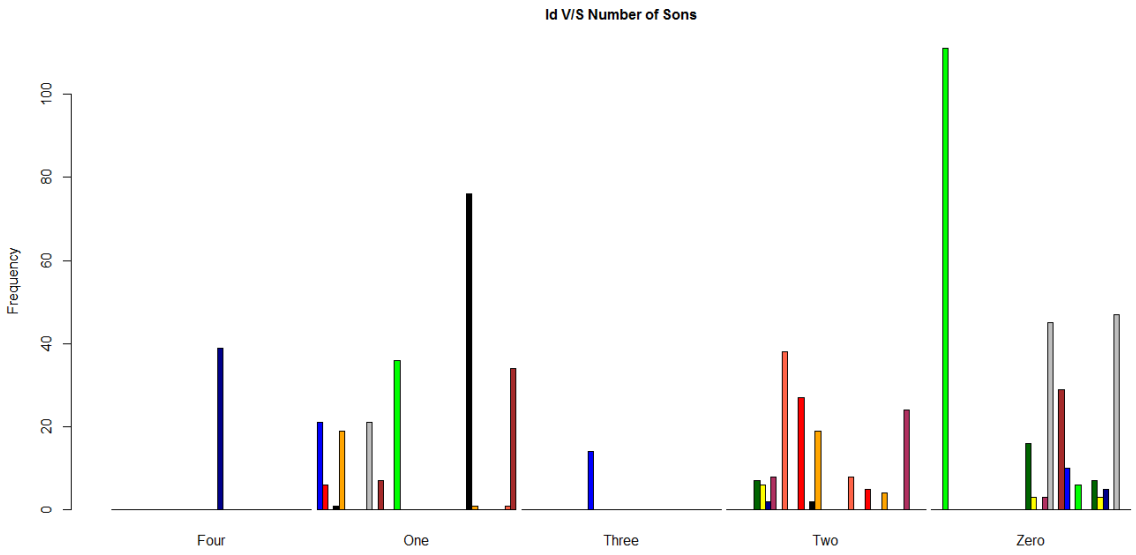

Id V/S Season
Frequency
0
10
20
30
40
Autum
Spring
Summer
Winter

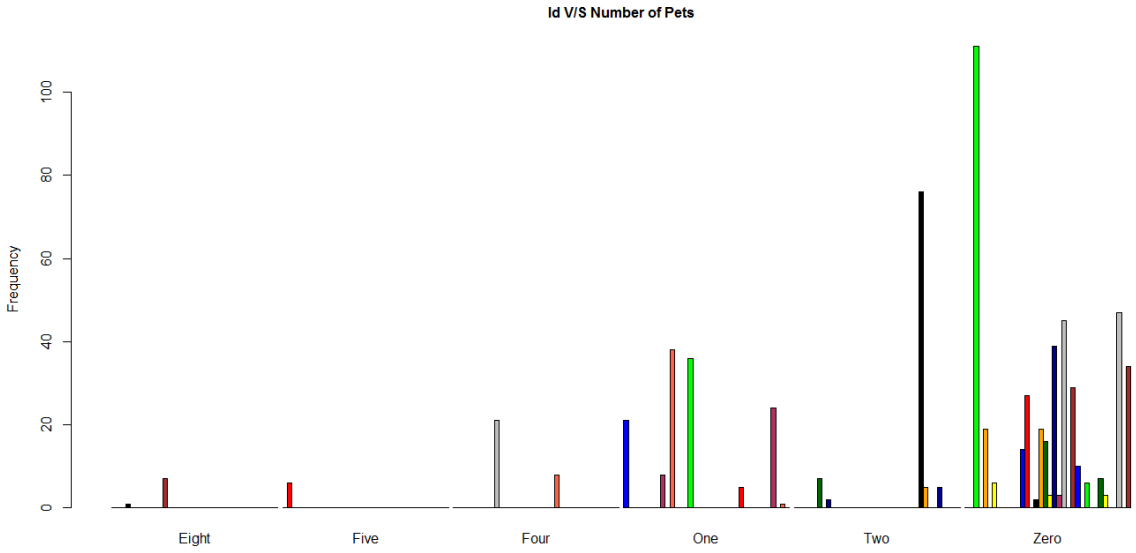

Id V/S Discpline Failure
Frequency
0
20
40
60
80
100
No
Yes

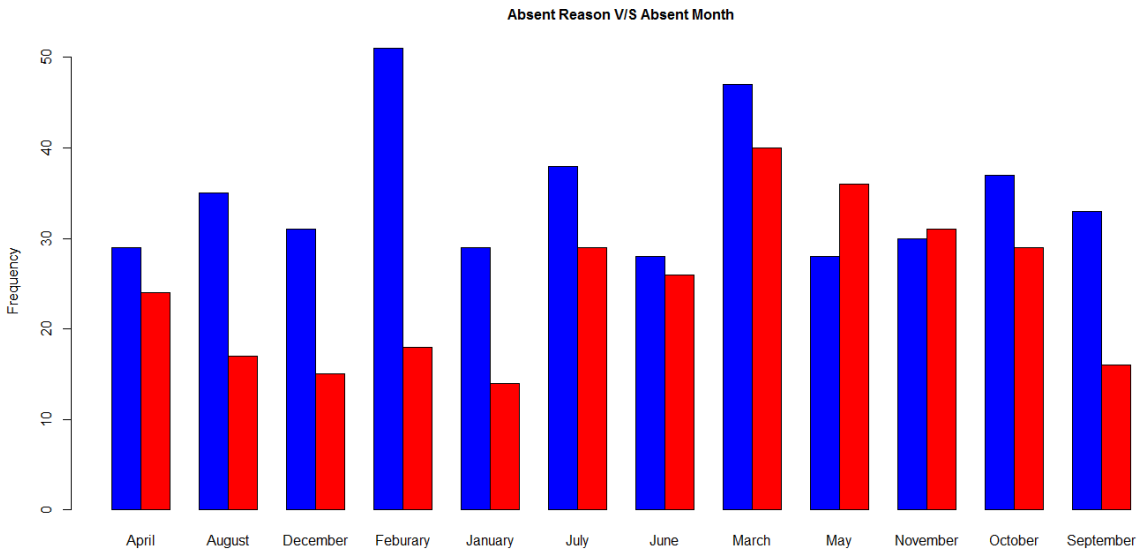

Id V/S Education
Frequency
0
20
40
60
80
100
Graduate
High School
Master& Doctor
Post Graduate

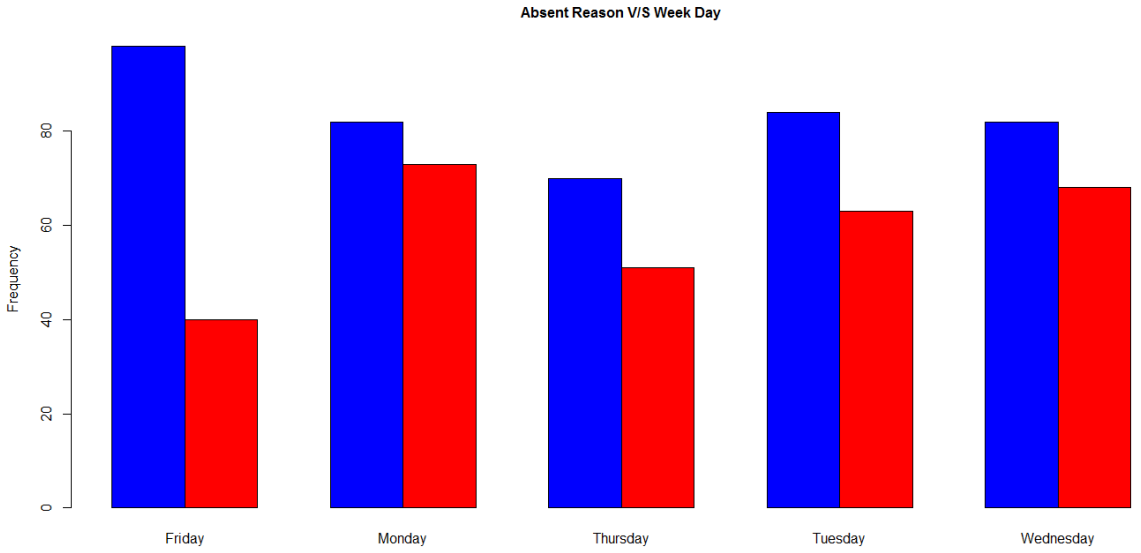

Id V/S Social Drinker
Frequency
0
20
40
60
80
100
No
Yes

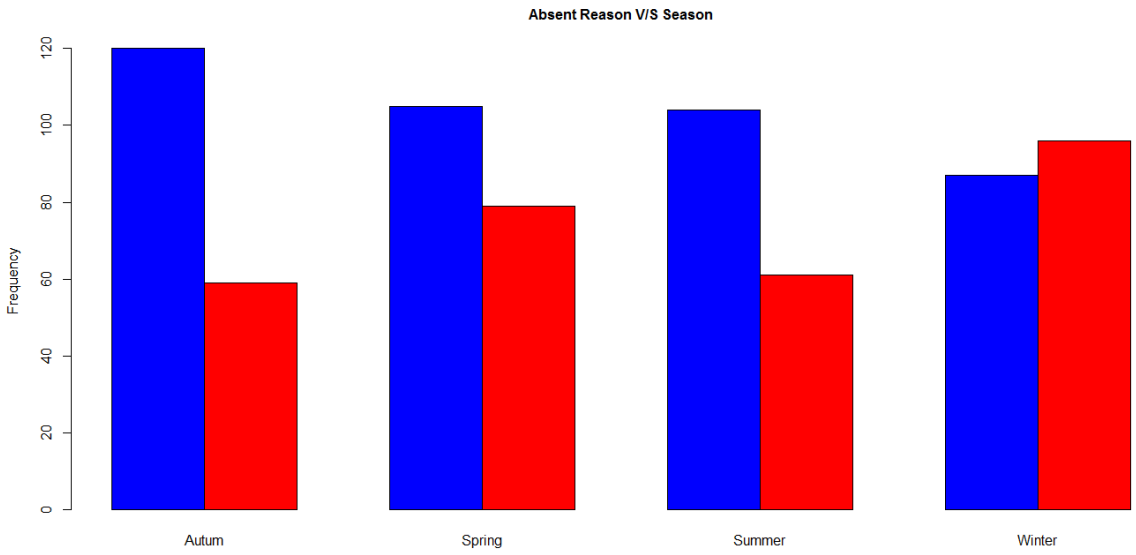

Id V/S Social Smoker
Frequency
0
20
40
60
80
100
No
Yes

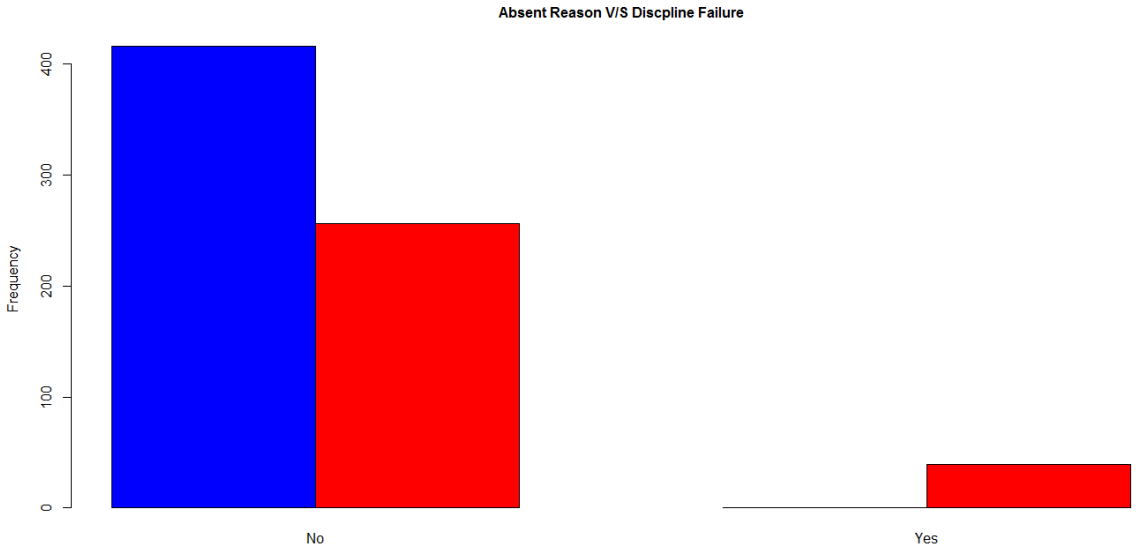

Id V/S Number of Sons
Frequency
0
20
40
60
80
100
Four
One
Three
Two
Zero

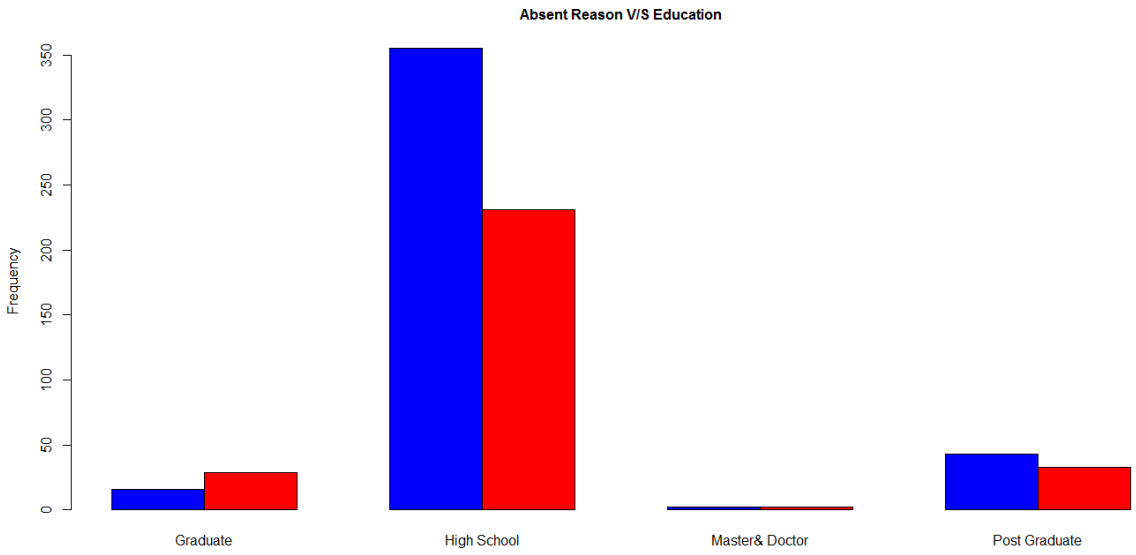

Id V/S Number of Pets
Frequency
0
20
40
60
80
100
Eight
Five
Four
One
Two
Zero

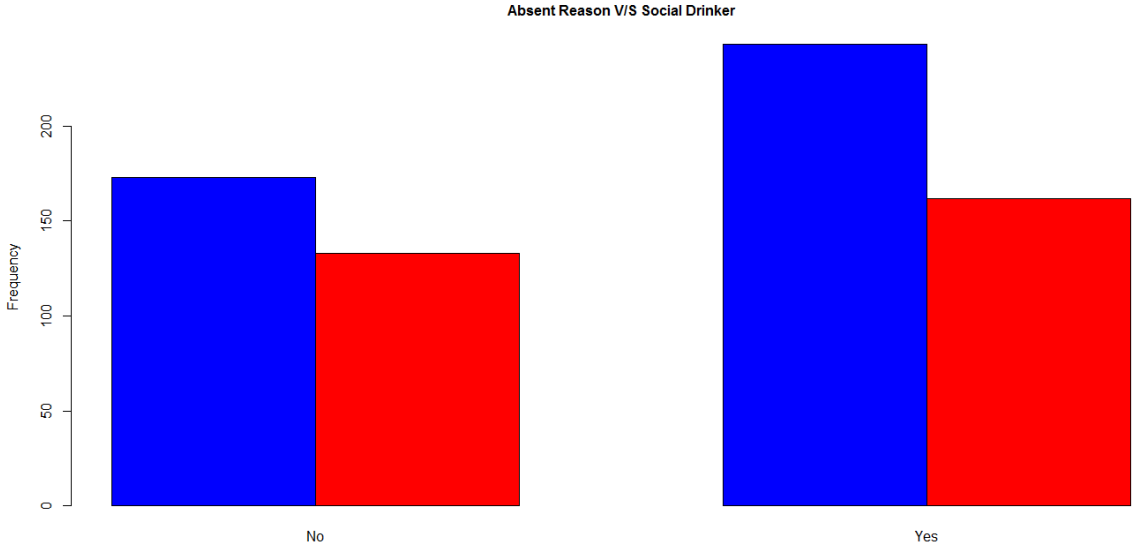

Absent Reason V/S Absent Month
Frequency
0
10
20
30
40
50
April
August
December
Feburary
January
July
June
March
May
November
October
September

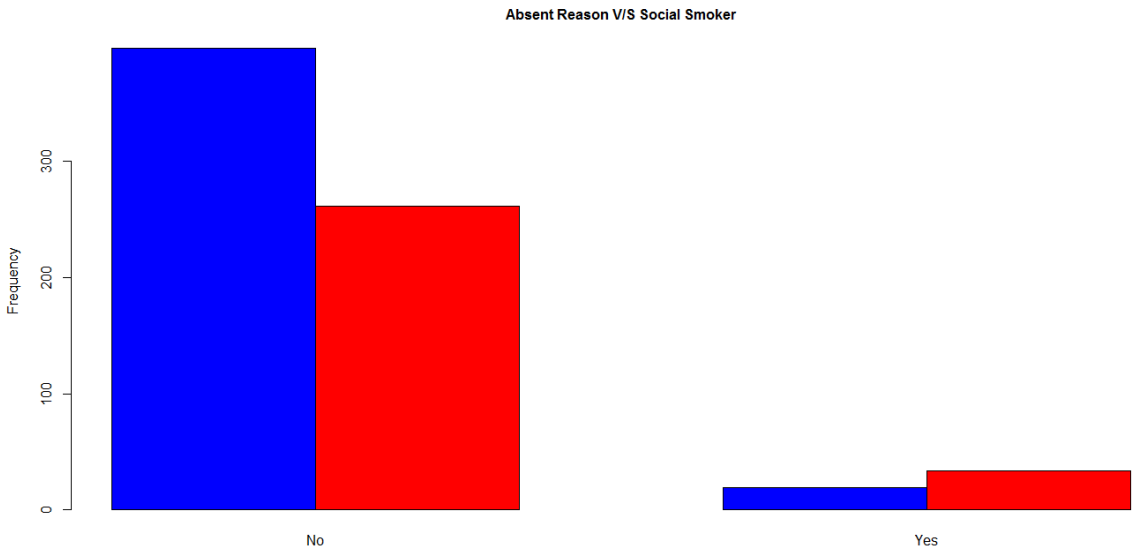

Absent Reason V/S Week Day
Frequency
0
20
40
60
80
Friday
Monday
Thursday
Tuesday
Wednesday

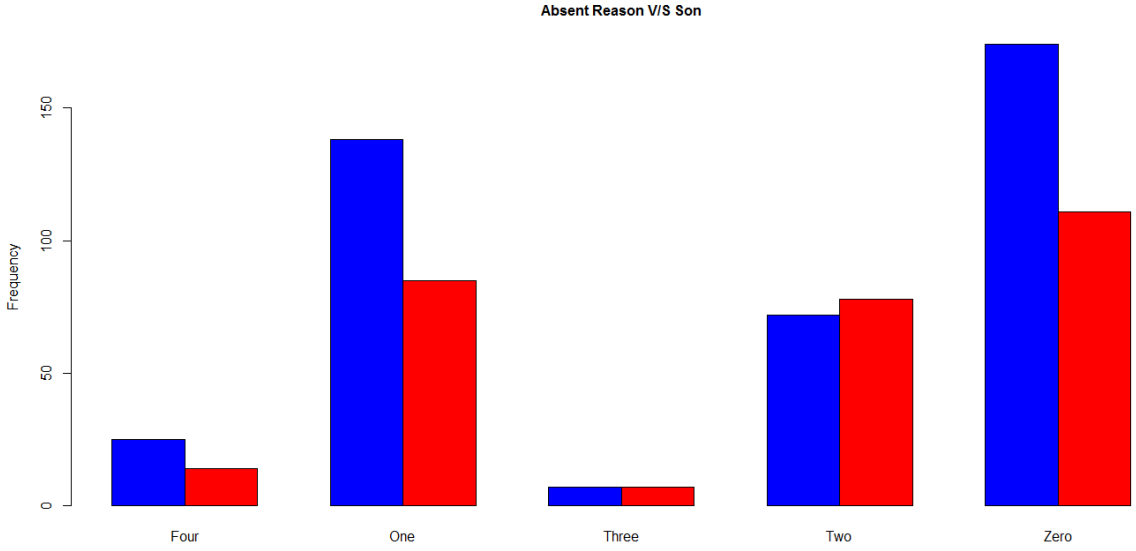

Absent Reason V/S Season
Frequency
0
20
40
60
80
100
120
Autum
Spring
Summer
Winter

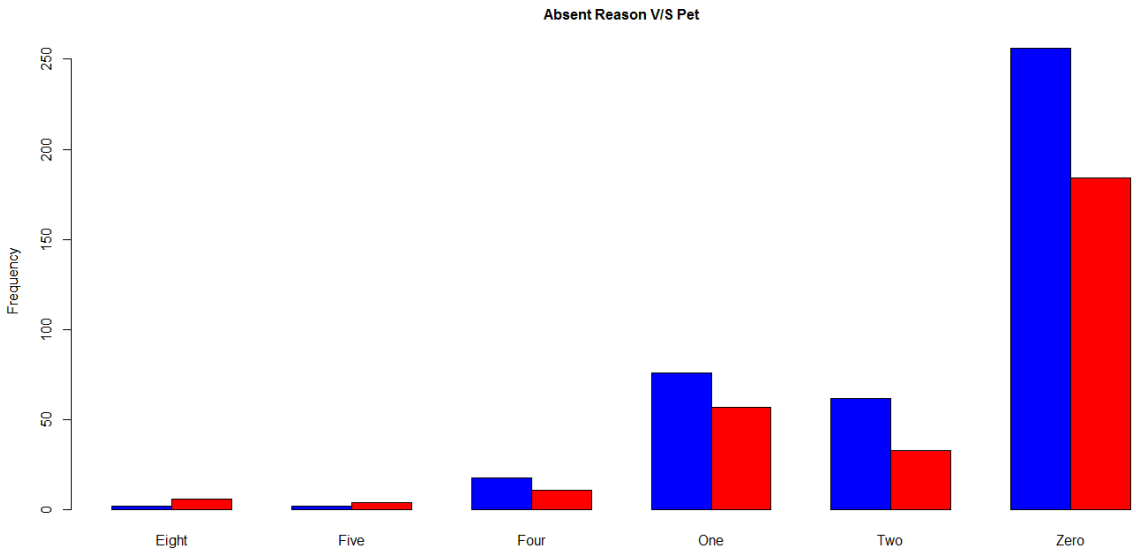

Absent Reason V/S Discpline Failure
Frequency
0
100
200
300
400
No
Yes

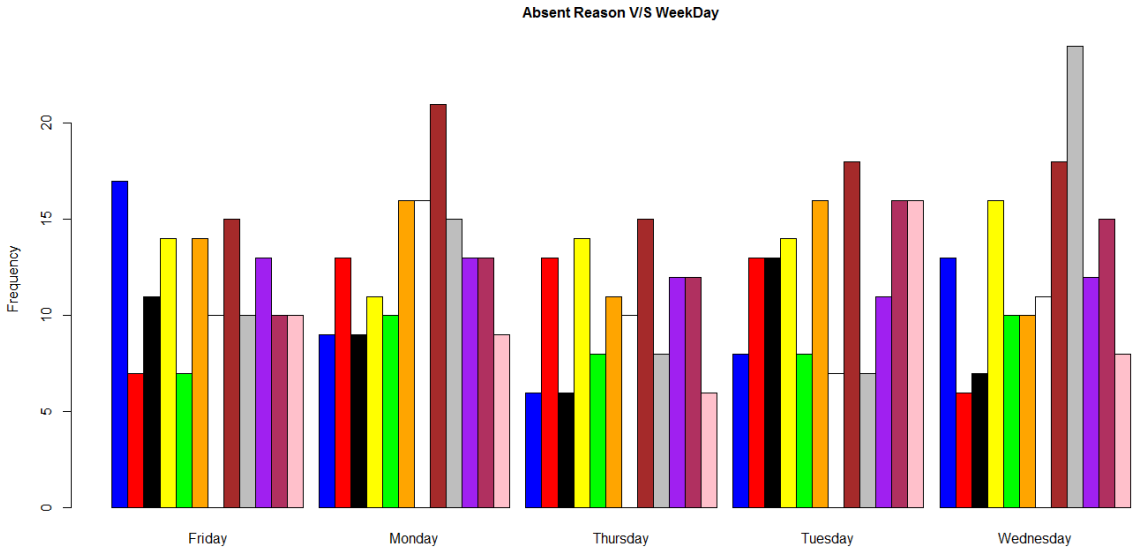

Absent Reason V/S Education
Frequency
0
50
100
150
200
250
300
350
Graduate
High School
Master& Doctor
Post Graduate

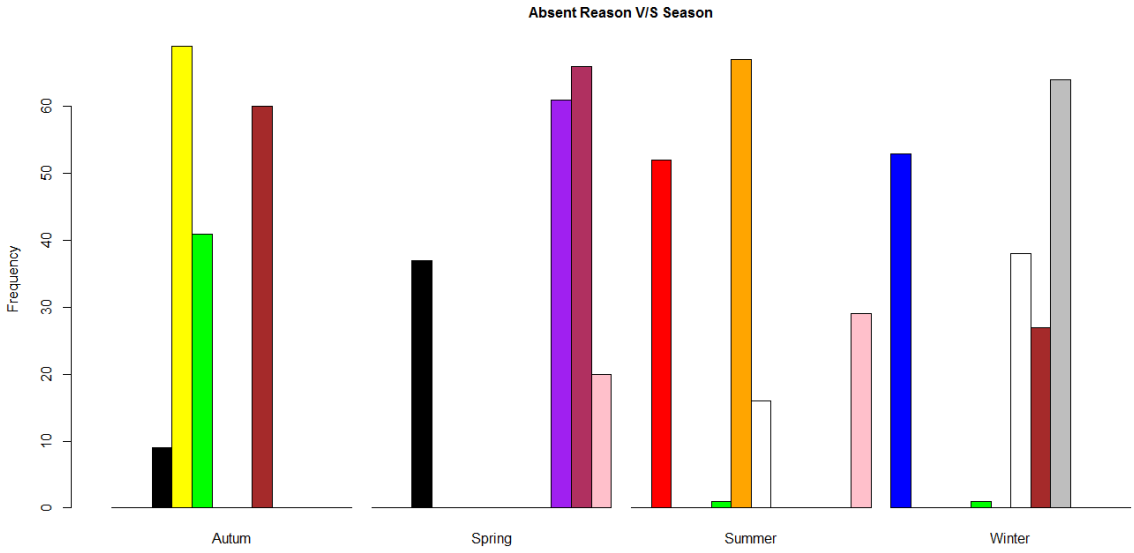

Absent Reason V/S Social Drinker
Frequency
0
50
100
150
200
No
Yes

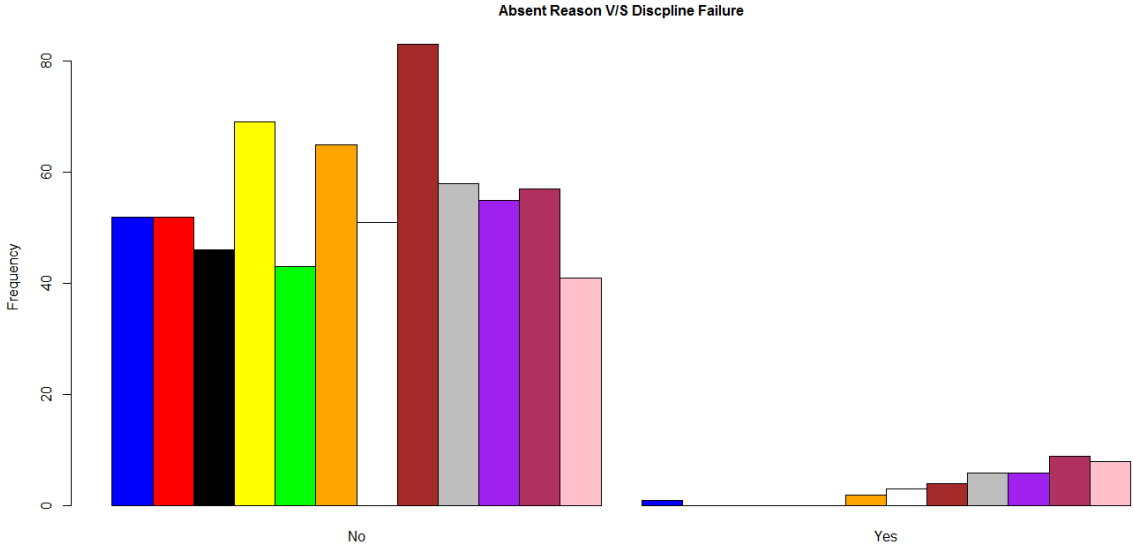

Absent Reason V/S Social Smoker
Frequency
0
100
200
300
No
Yes

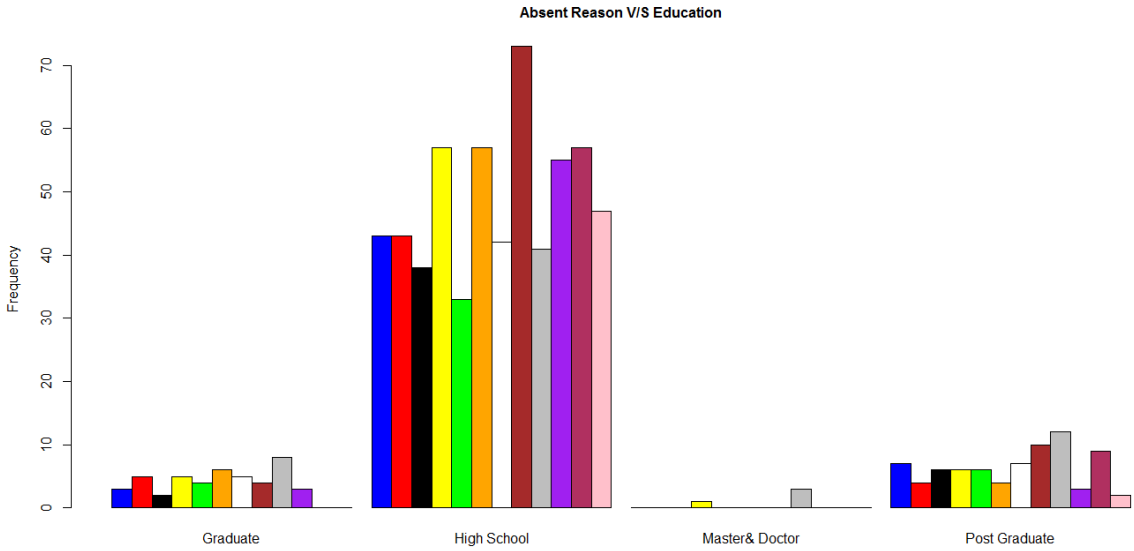

Absent Reason V/S Son
Frequency
0
50
100
150
Four
One
Three
Two
Zero

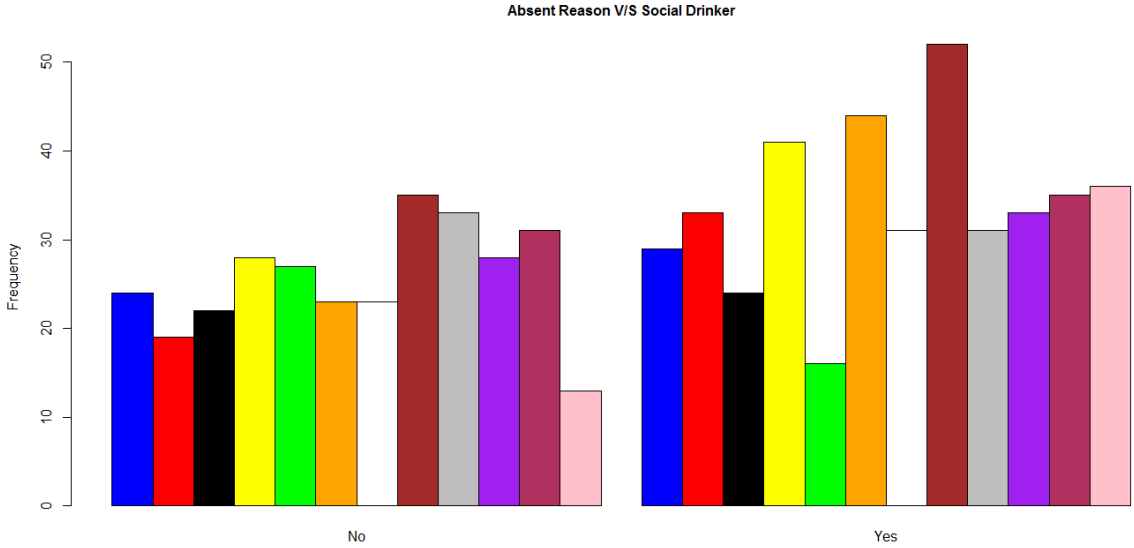

Absent Reason V/S Pet
Frequency
0
50
100
150
200
250
Eight
Five
Four
One
Two
Zero

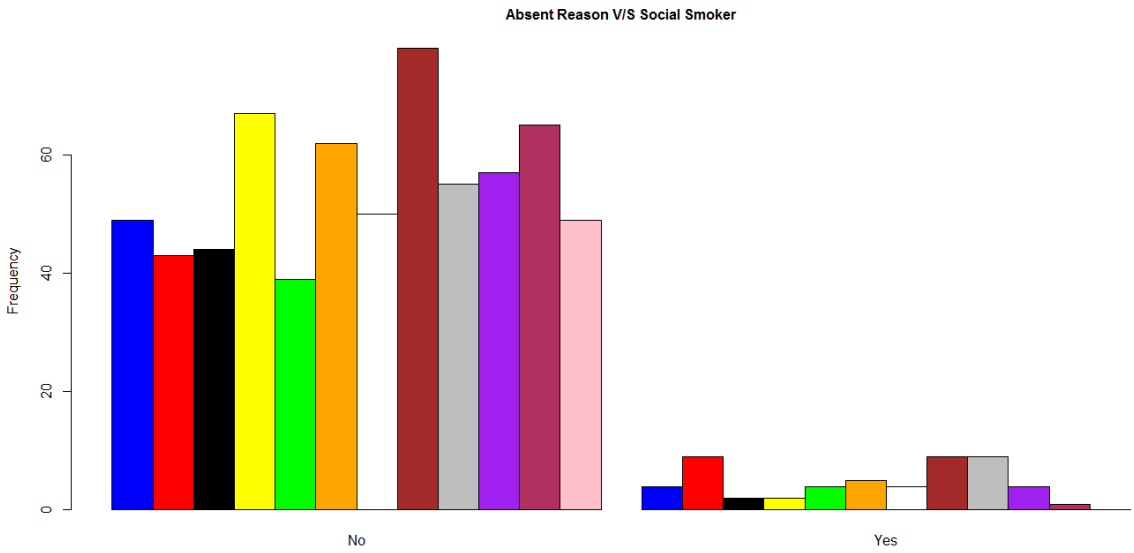

Absent Reason V/S WeekDay
Frequency
0
5
10
15
20
Friday
Monday
Thursday
Tuesday
Wednesday

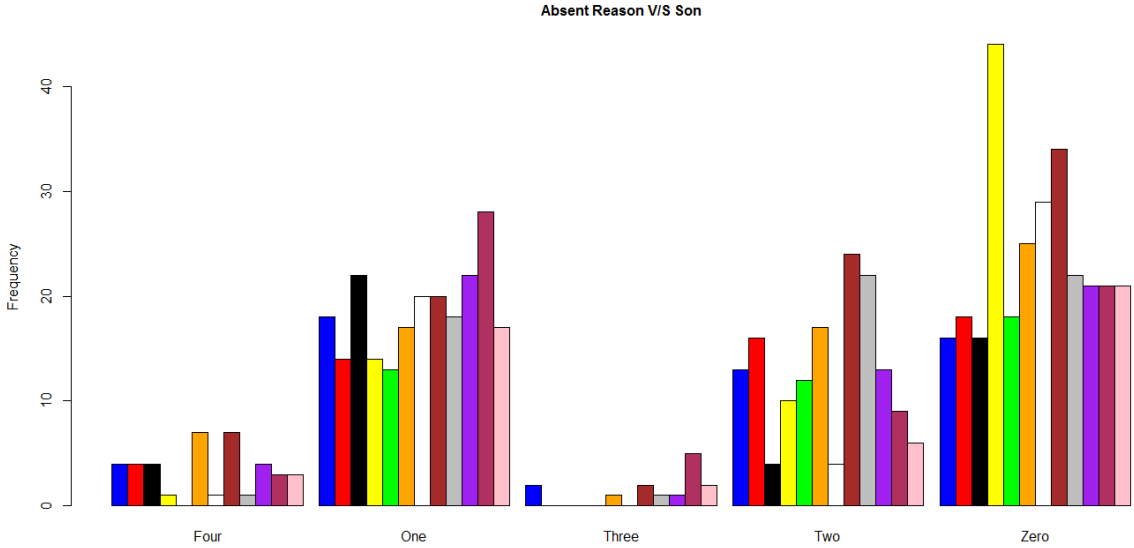

Absent Reason V/S Season
Frequency
0
10
20
30
40
50
60
Autum
Spring
Summer
Winter

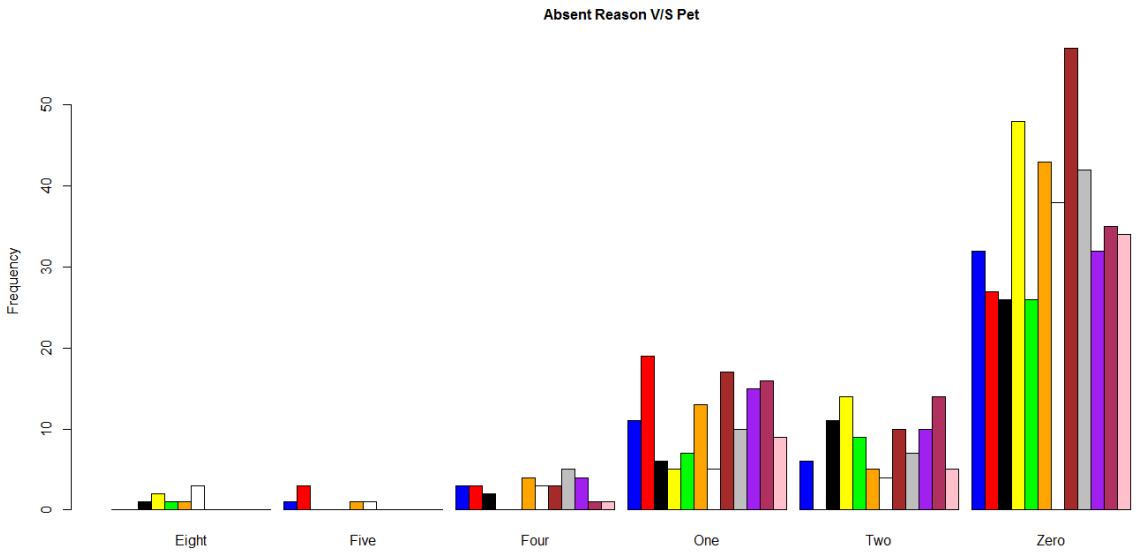

Absent Reason V/S Discpline Failure
Frequency
0
20
40
60
80
No
Yes

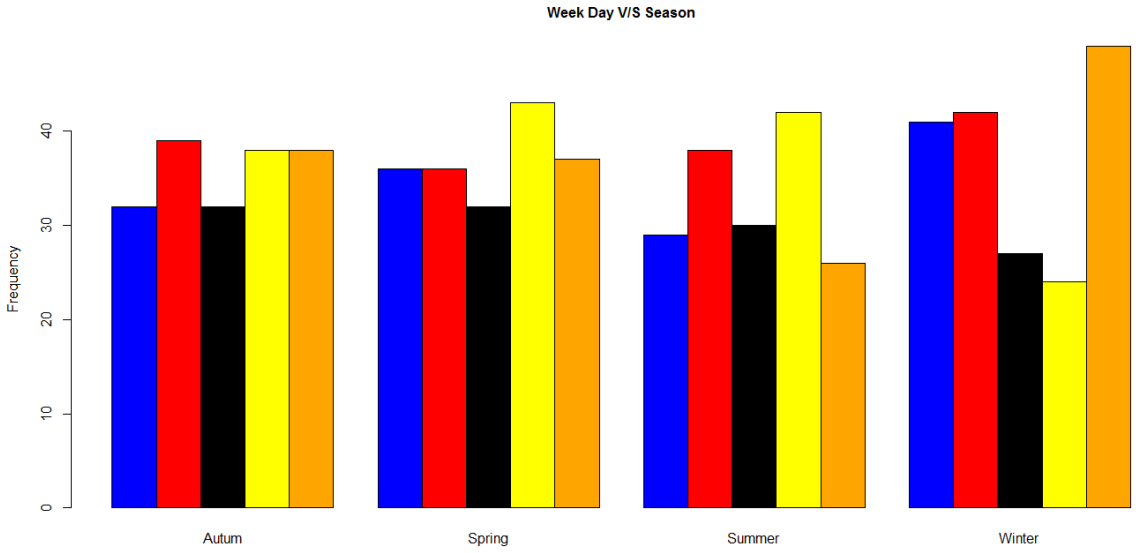

Absent Reason V/S Education
Frequency
0
10
20
30
40
50
60
70
Graduate
High School
Master& Doctor
Post Graduate

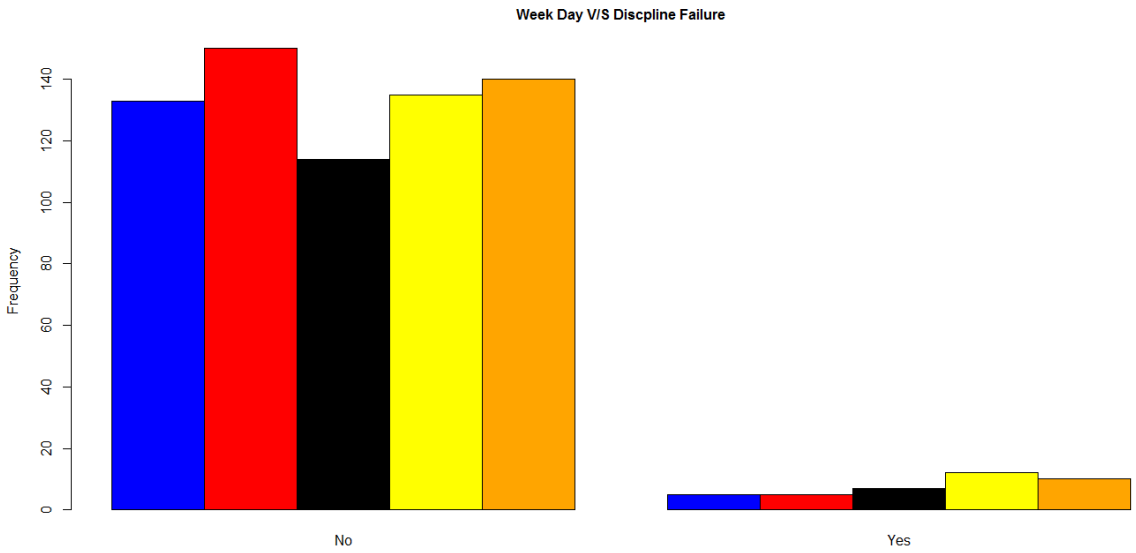

Absent Reason V/S Social Drinker
Frequency
0
10
20
30
40
50
No
Yes

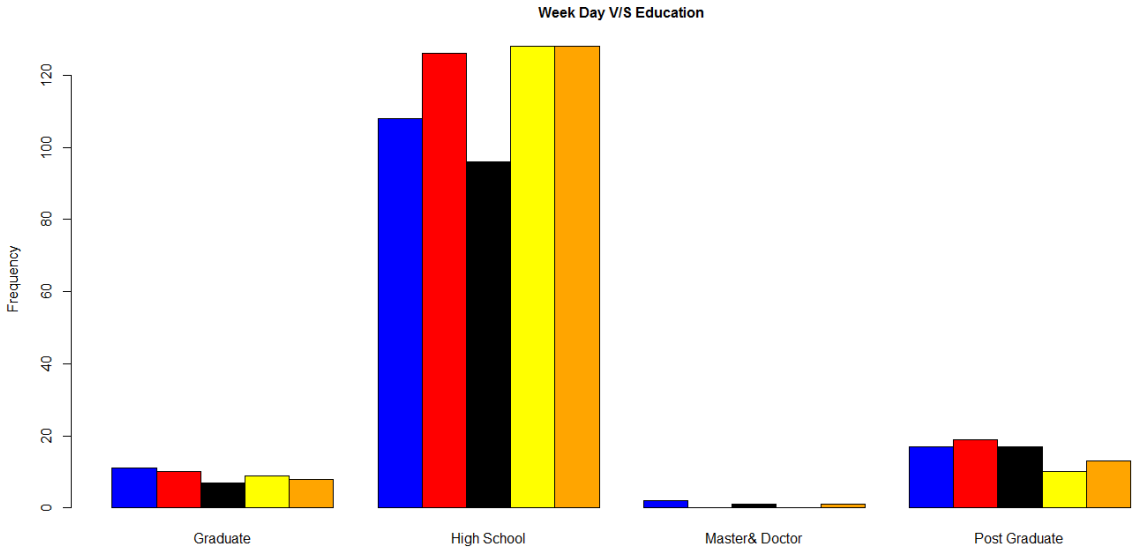

Absent Reason V/S Social Smoker
Frequency
0
20
40
60
No
Yes

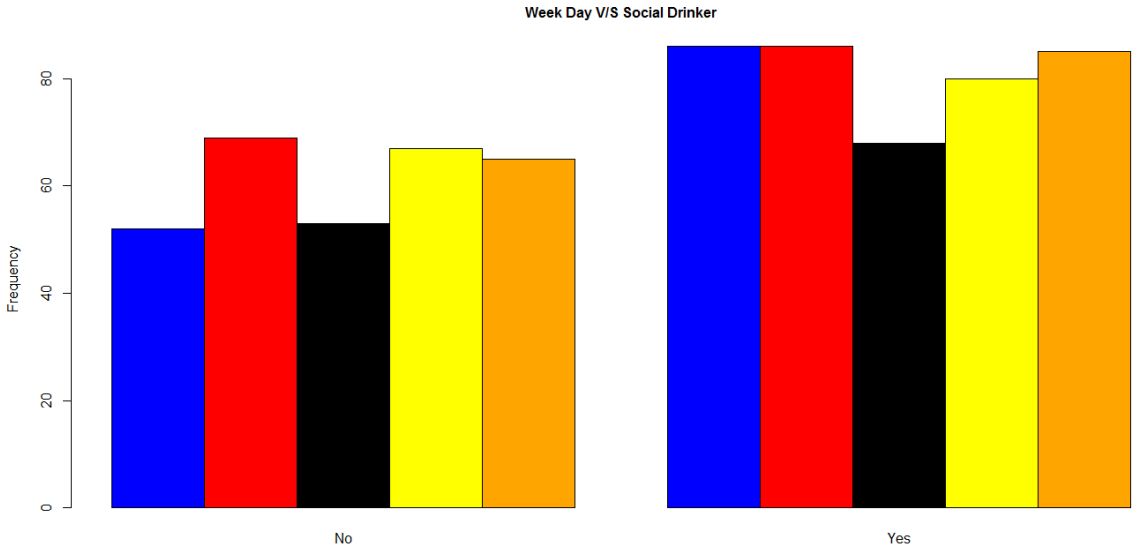

Absent Reason V/S Son
Frequency
0
10
20
30
40
Four
One
Three
Two
Zero

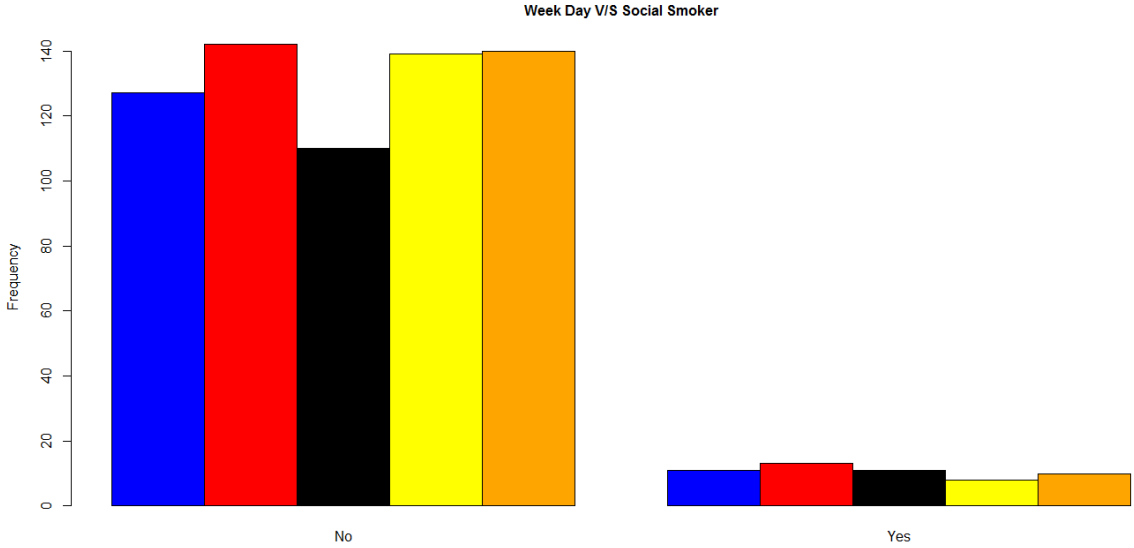

Absent Reason V/S Pet
Frequency
0
10
20
30
40
50
Eight
Five
Four
One
Two
Zero

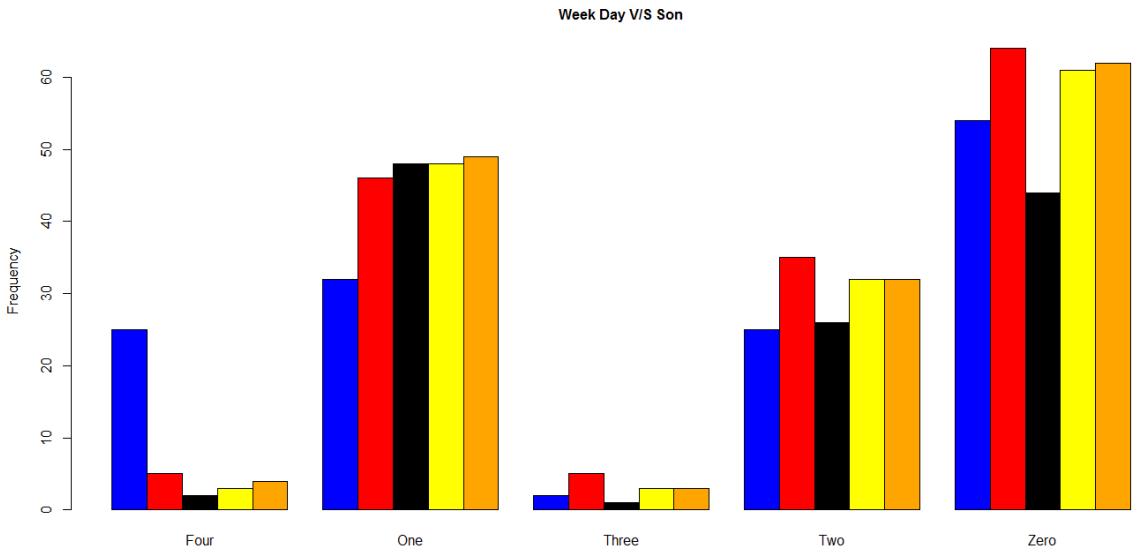

Week Day V/S Season
Frequency
0
10
20
30
40
Autum
Spring
Summer
Winter

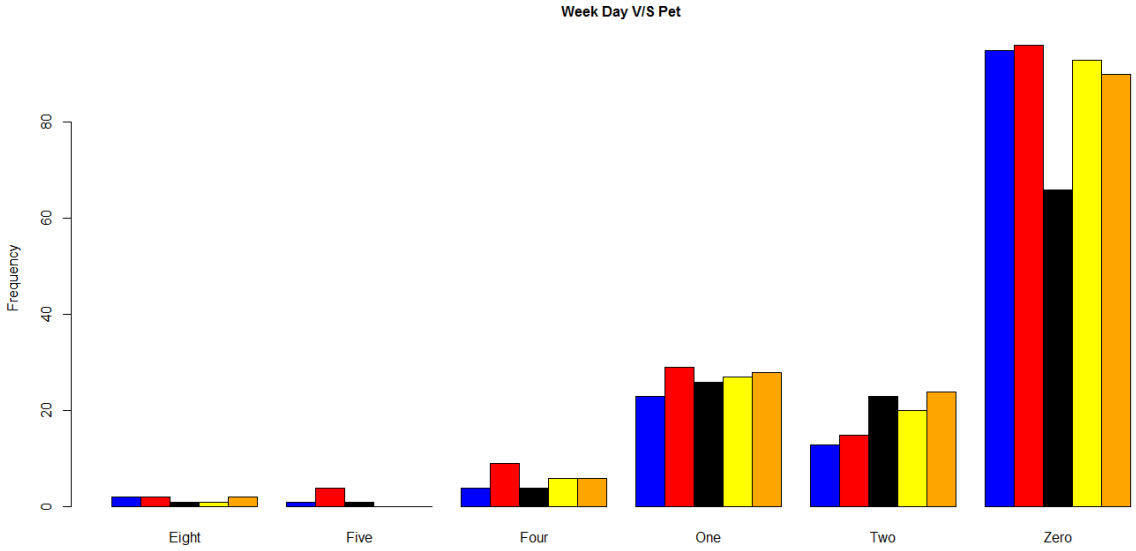

Week Day V/S Discpline Failure
Frequency
0
20
40
60
80
100
120
140
No
Yes

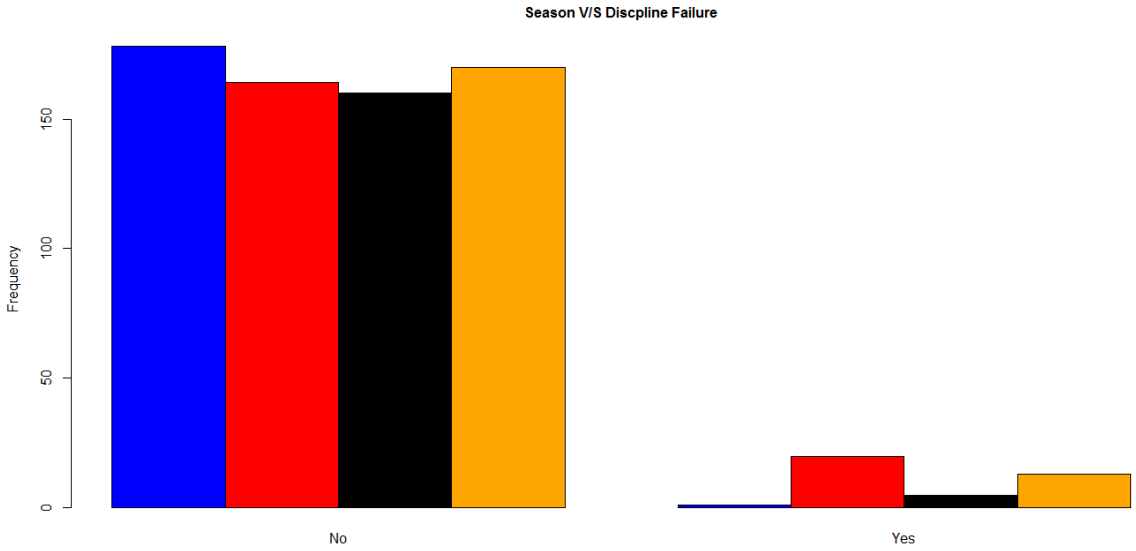

Week Day V/S Education
Frequency
0
20
40
60
80
100
120
Graduate
High School
Master& Doctor
Post Graduate

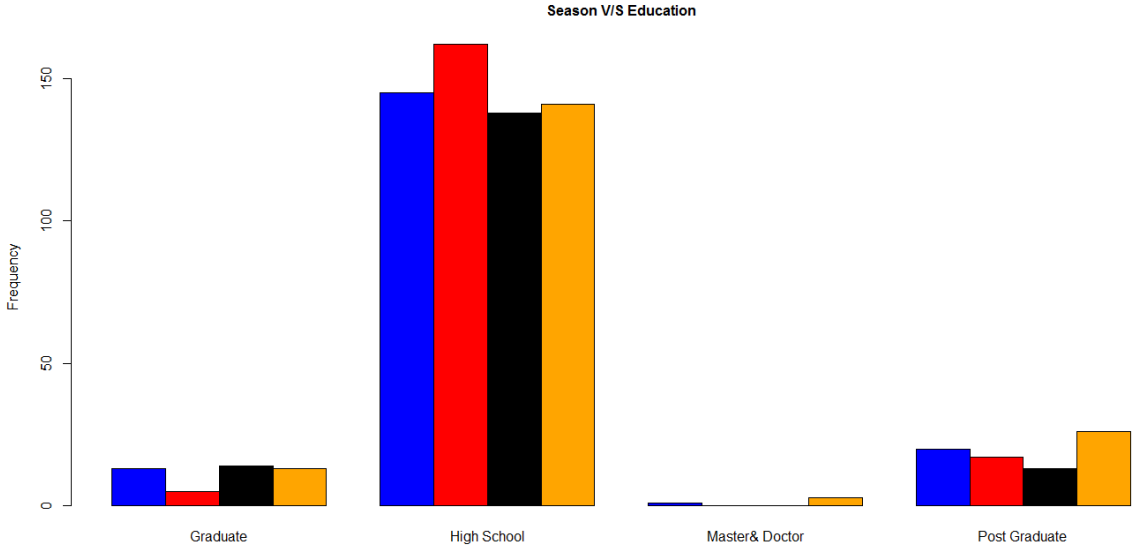

Week Day V/S Social Drinker
Frequency
0
20
40
60
80
No
Yes

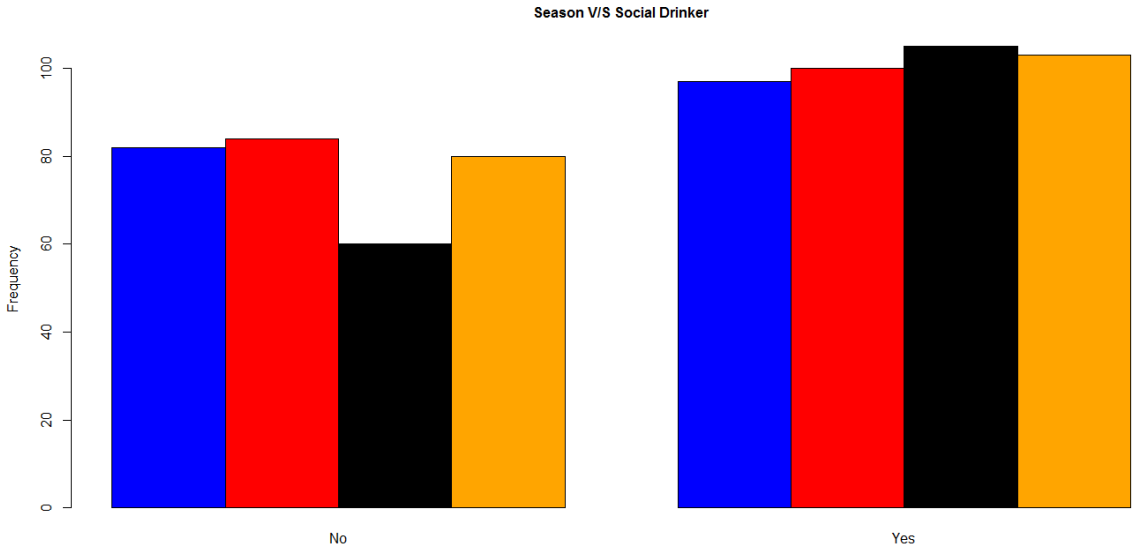

Week Day V/S Social Smoker
Frequency
0
20
40
60
80
100
120
140
No
Yes

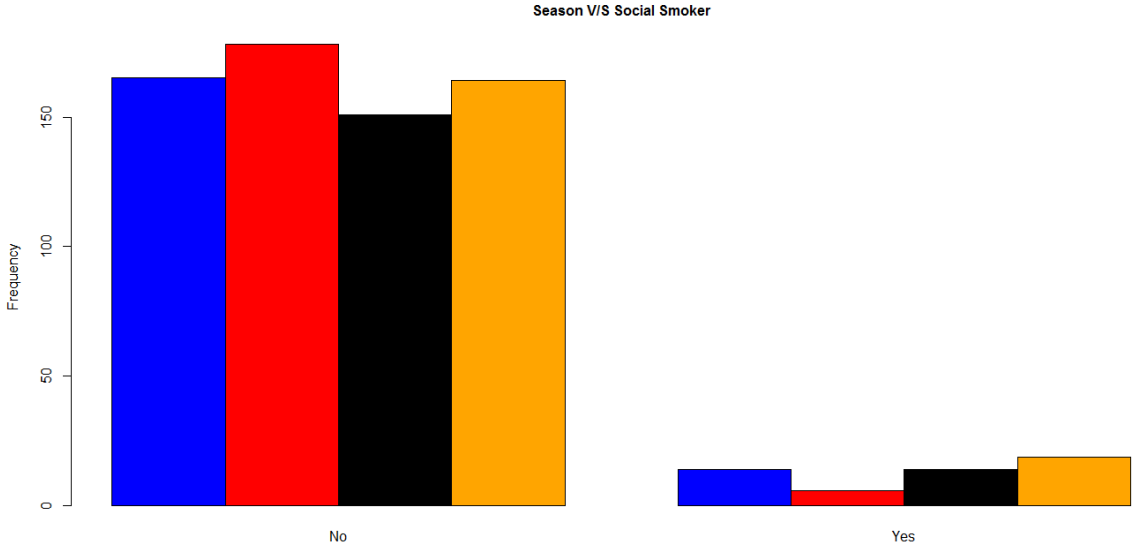

Week Day V/S Son
Frequency
0
10
20
30
40
50
60
Four
One
Three
Two
Zero

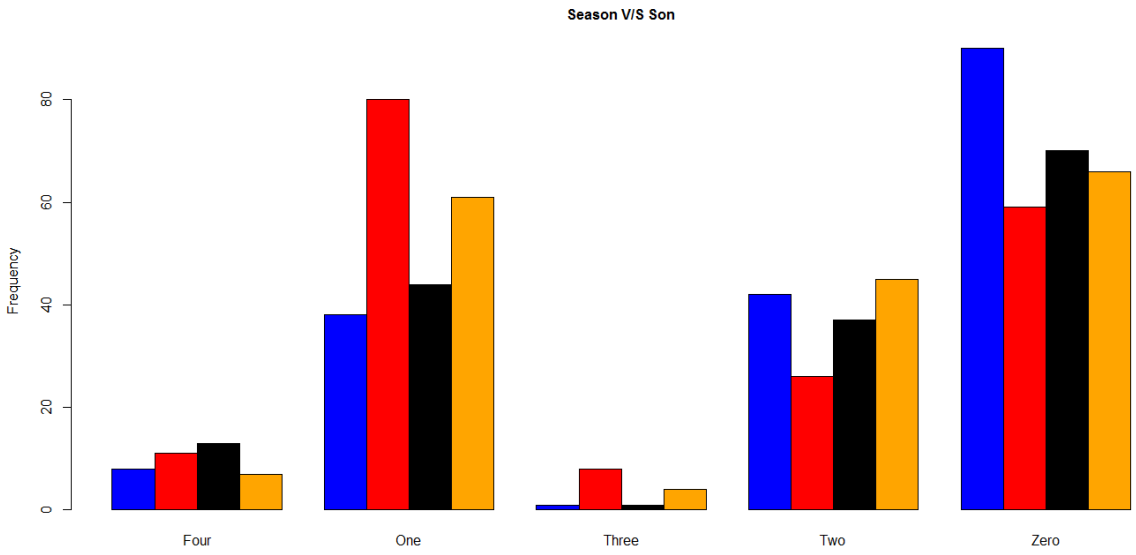

Week Day V/S Pet
Frequency
0
20
40
60
80
Eight
Five
Four
One
Two
Zero

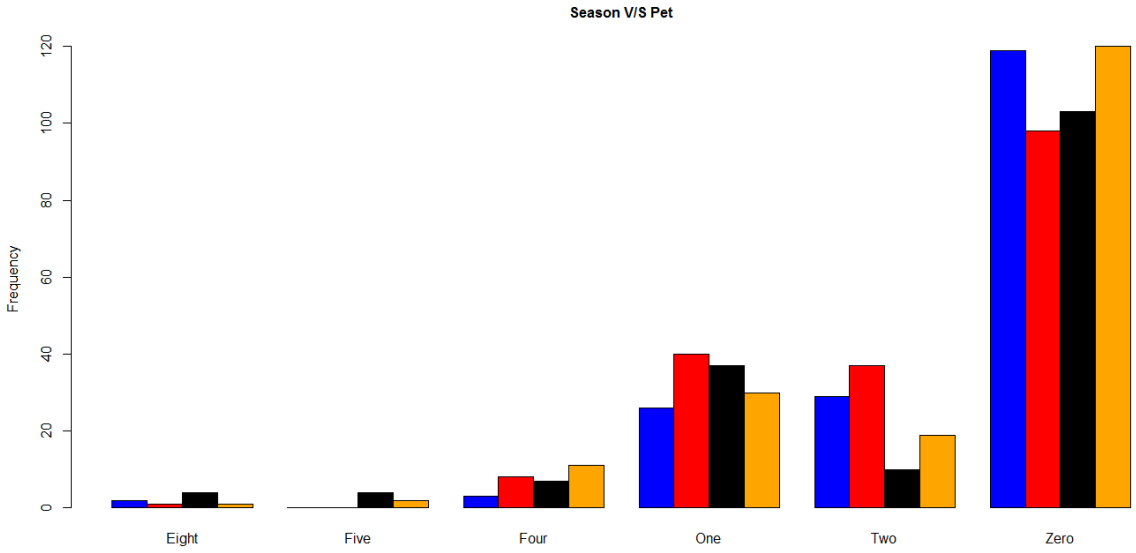

Season V/S Discpline Failure
Frequency
0
50
100
150
No
Yes

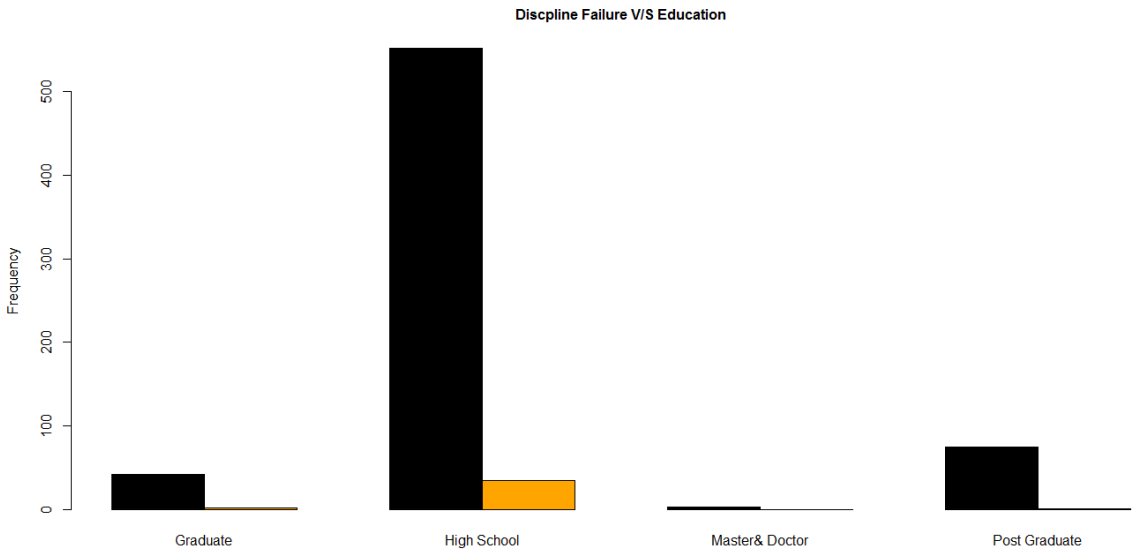

Season V/S Education
Frequency
0
50
100
150
Graduate
High School
Master& Doctor
Post Graduate

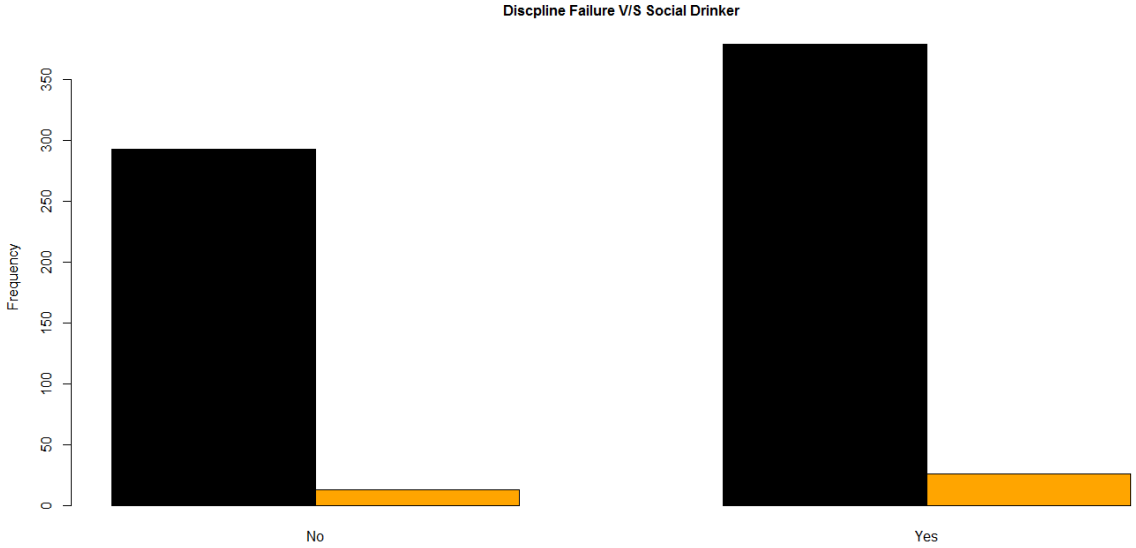

Season V/S Social Drinker
Frequency
0
20
40
60
80
100
No
Yes

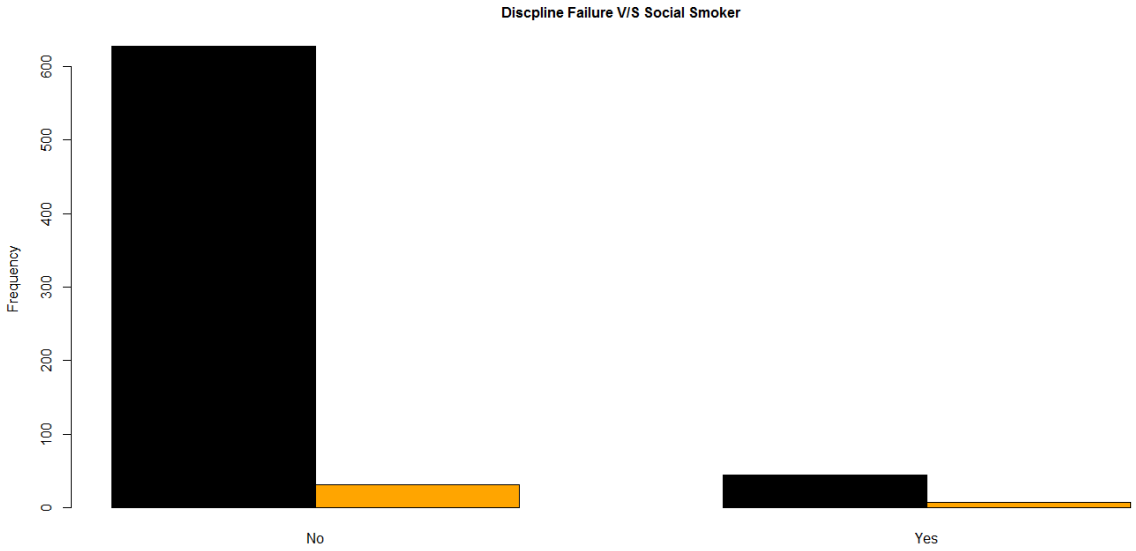

Season V/S Social Smoker
Frequency
0
50
100
150
No
Yes

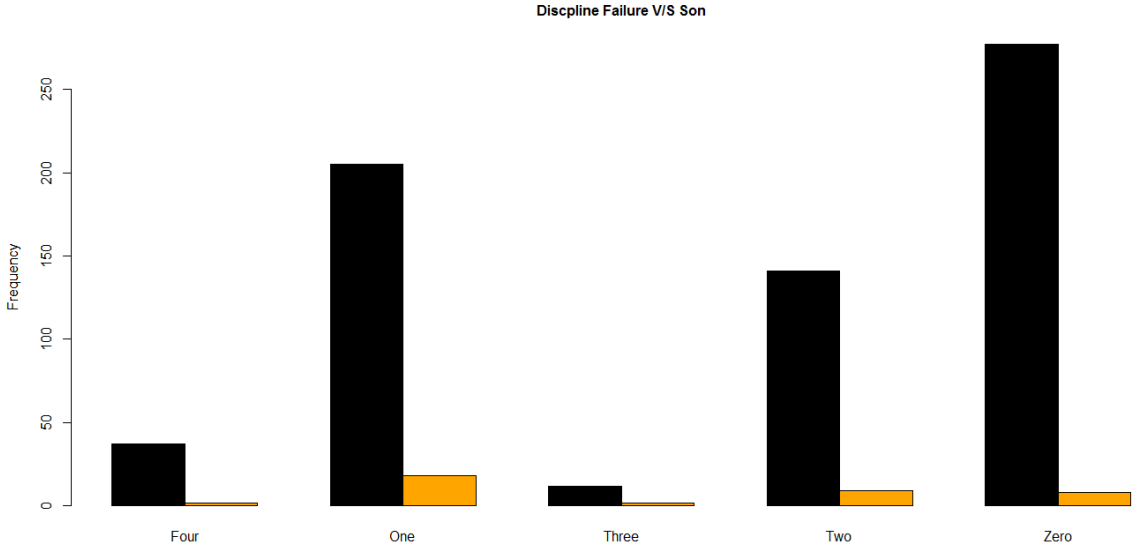

Season V/S Son
Frequency
0
20
40
60
80
Four
One
Three
Two
Zero

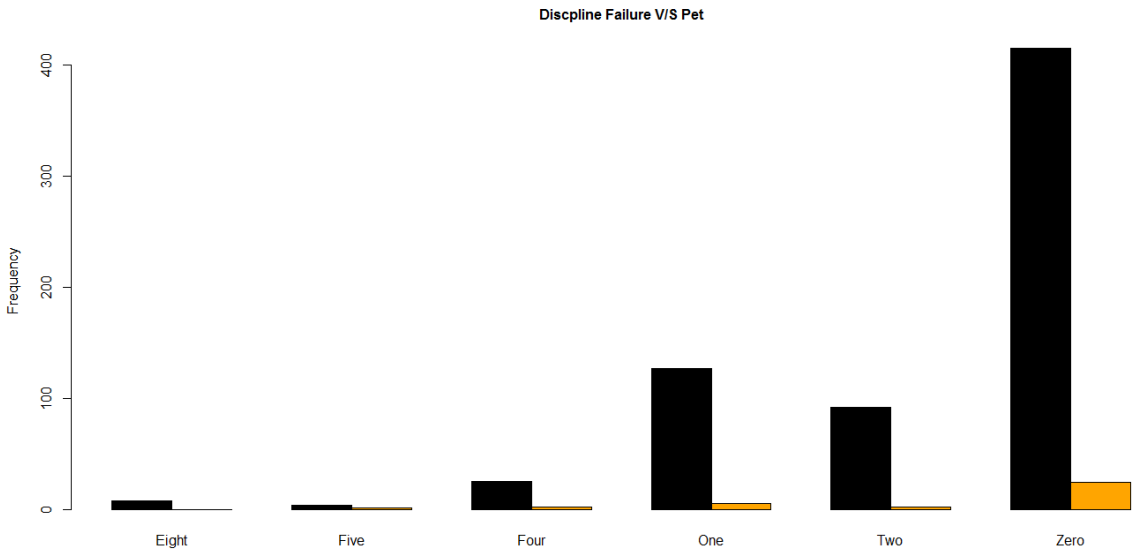

Season V/S Pet
Frequency
0
20
40
60
80
100
120
Eight
Five
Four
One
Two
Zero

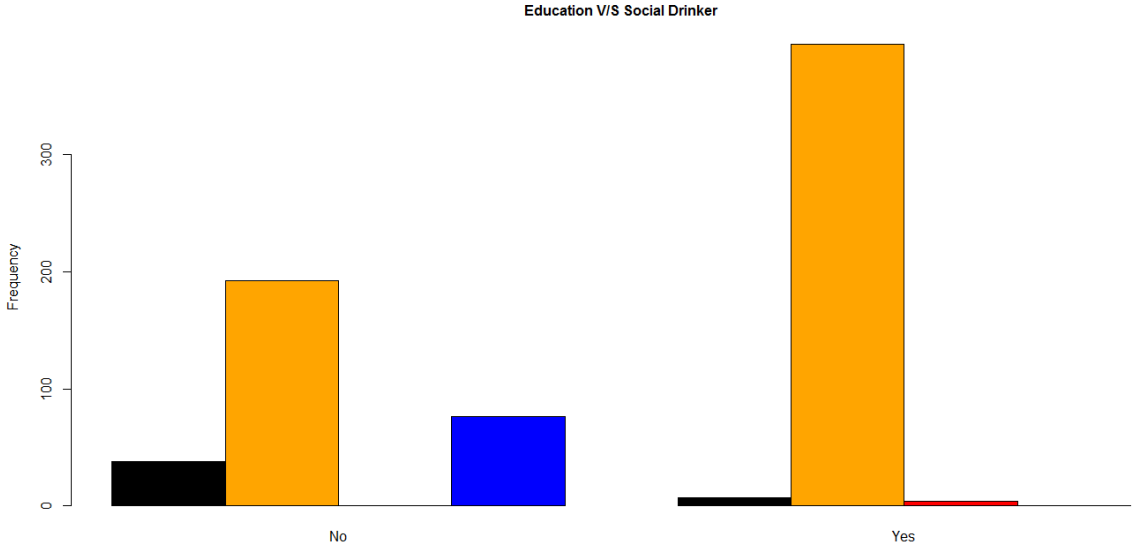

Discpline Failure V/S Education
Frequency
0
100
200
300
400
500
Graduate
High School
Master& Doctor
Post Graduate

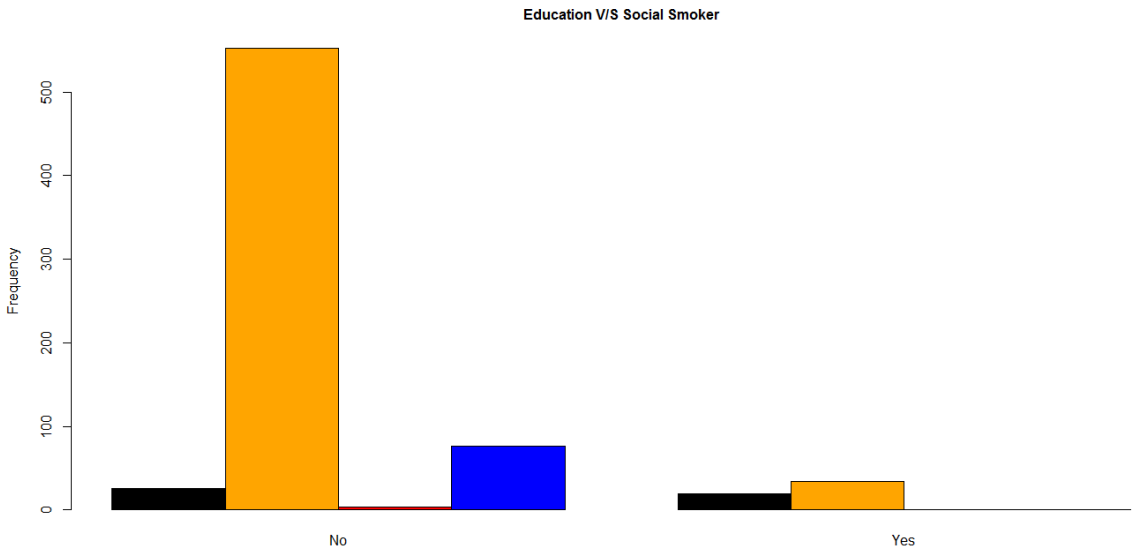

Discpline Failure V/S Social Drinker
Frequency
0
50
100
150
200
250
300
350
No
Yes

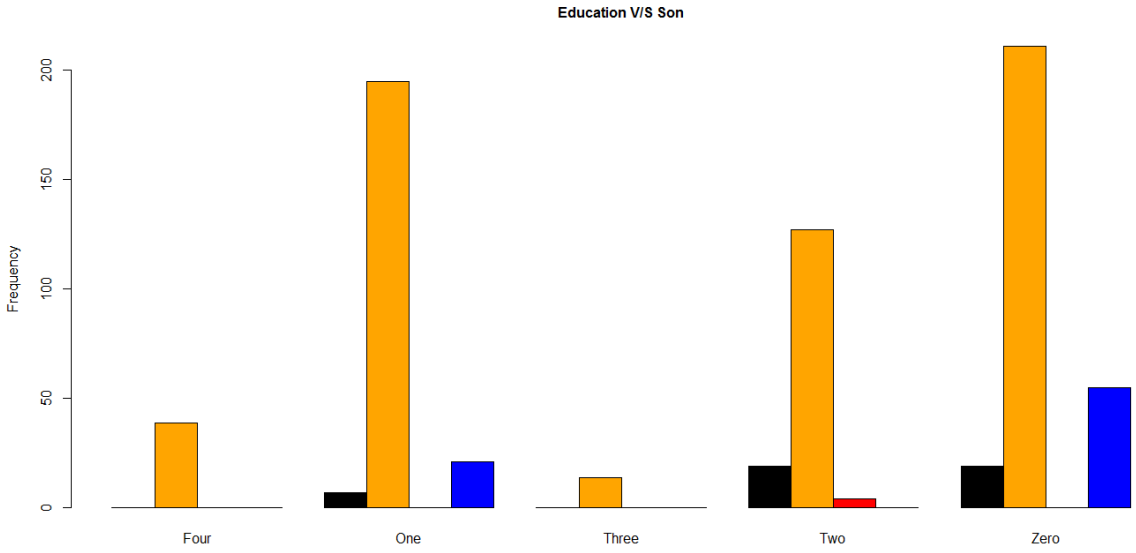

Discpline Failure V/S Social Smoker
Frequency
0
100
200
300
400
500
600
No
Yes

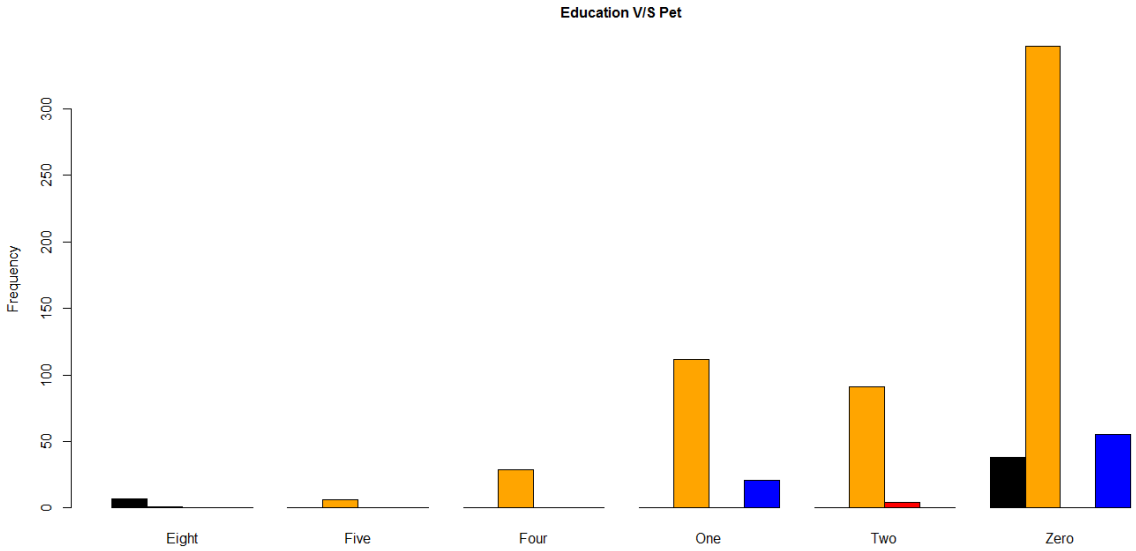

Discpline Failure V/S Son
Frequency
0
50
100
150
200
250
Four
One
Three
Two
Zero

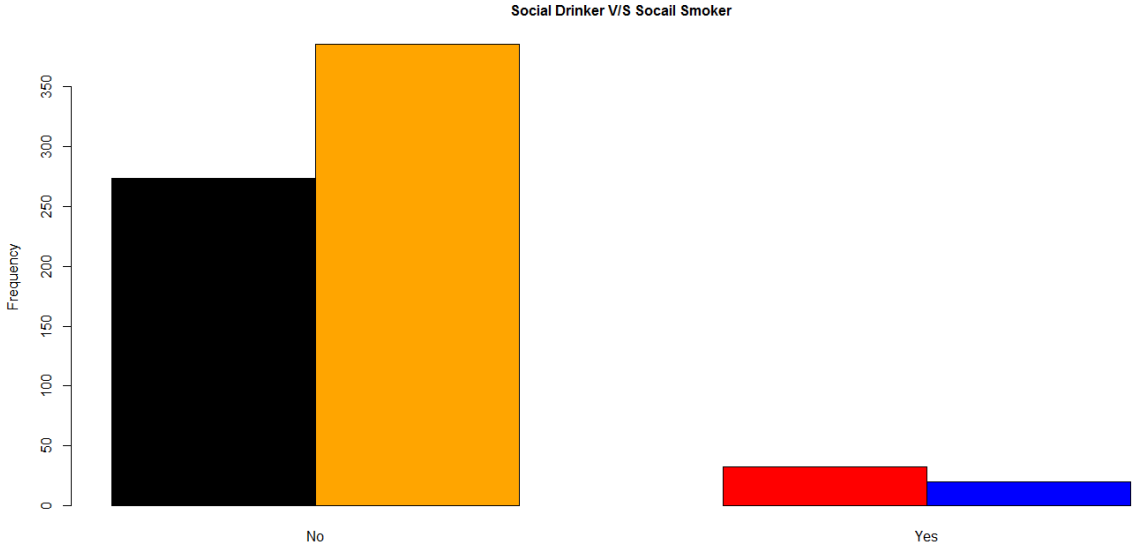

Discpline Failure V/S Pet
Frequency
0
100
200
300
400
Eight
Five
Four
One
Two
Zero

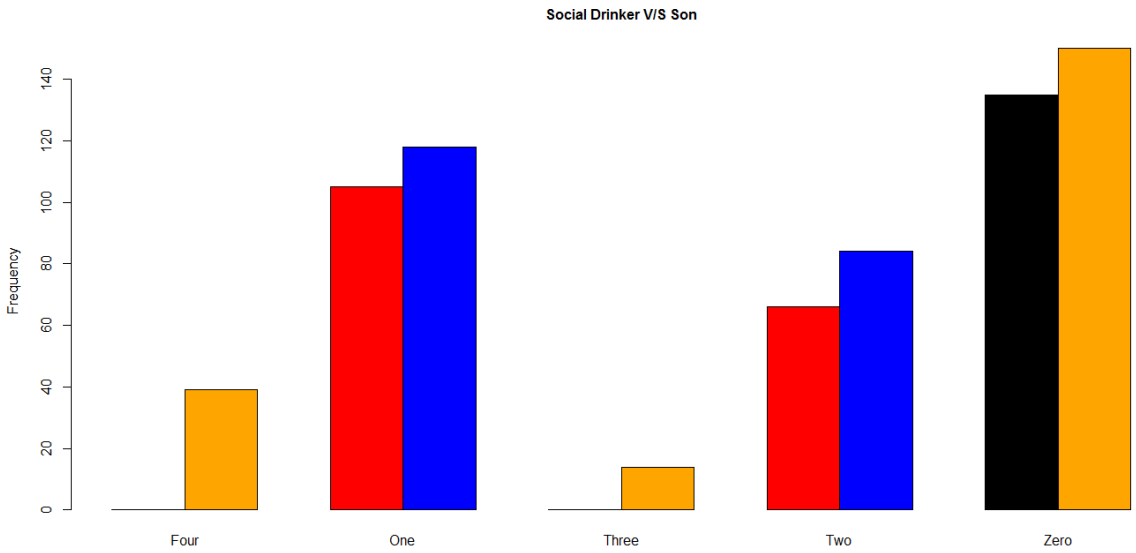

Education V/S Social Drinker
Frequency
0
100
200
300
No
Yes

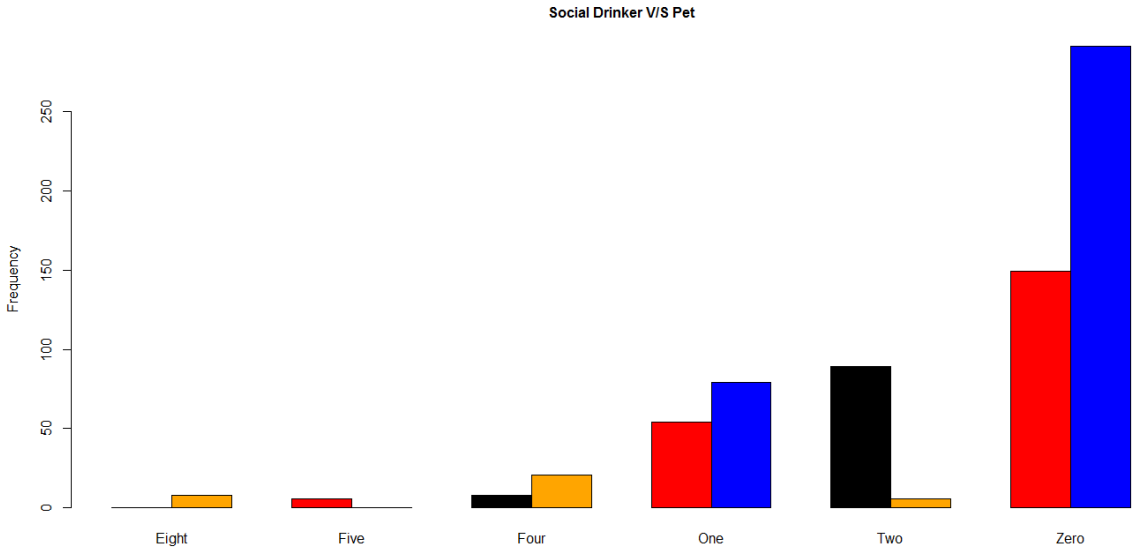

Education V/S Social Smoker
Frequency
0
100
200
300
400
500
No
Yes

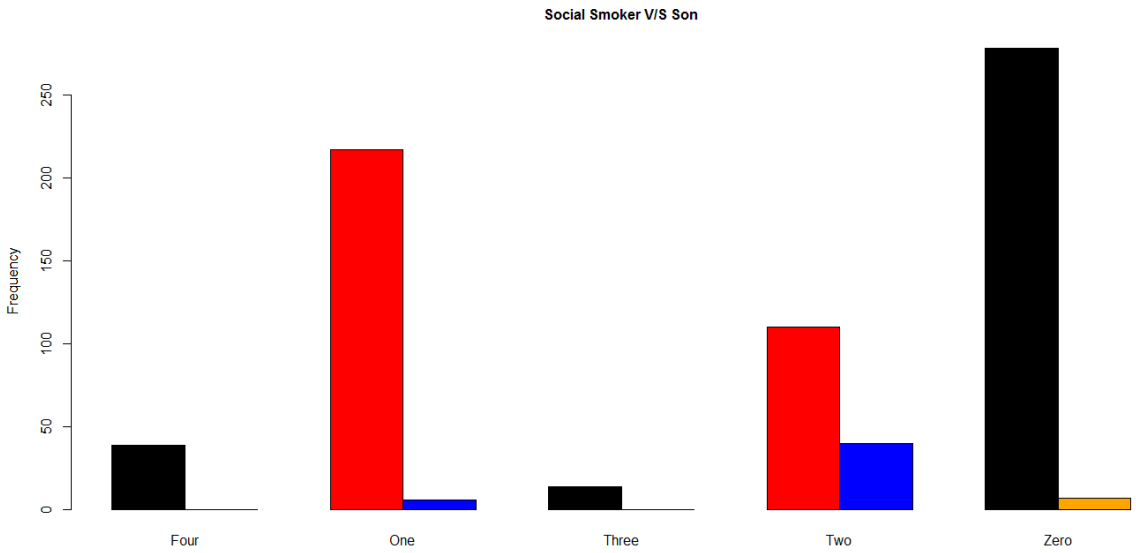

Education V/S Son
Frequency
0
50
100
150
200
Four
One
Three
Two
Zero

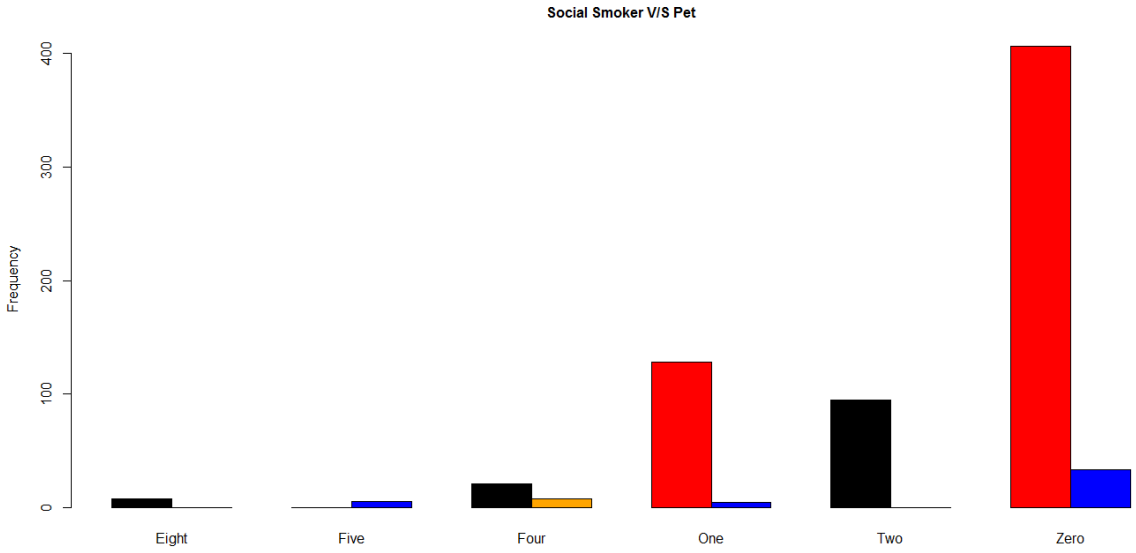

Education V/S Pet
Frequency
0
50
100
150
200
250
300
Eight
Five
Four
One
Two
Zero

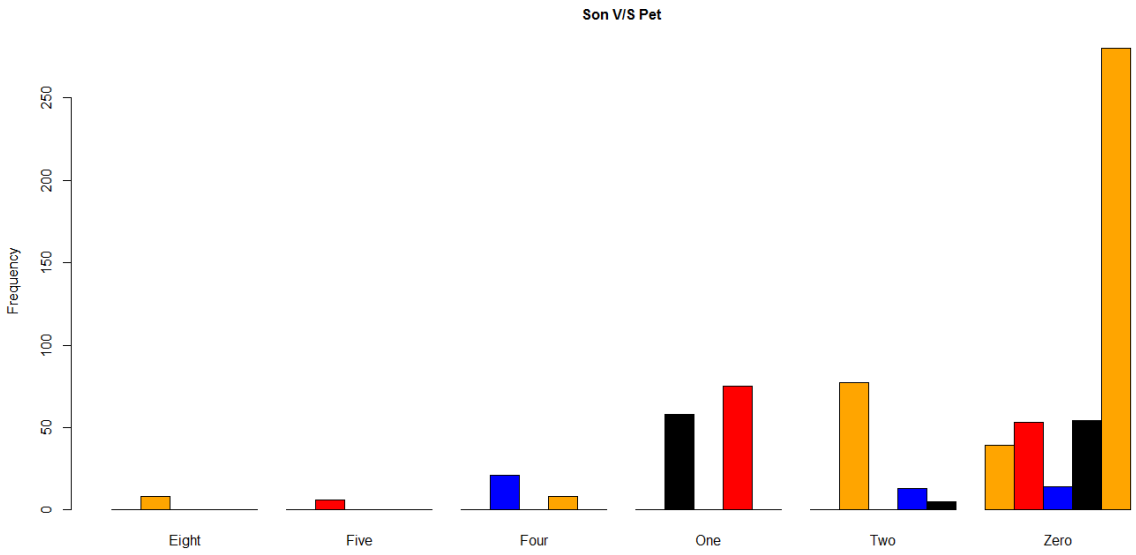

Social Drinker V/S Socail Smoker
Frequency
0
50
100
150
200
250
300
350
No
Yes


Social Drinker V/S Son
Frequency
0
20
40
60
80
100
120
140
Four
One
Three
Two
Zero


Social Drinker V/S Pet
Frequency
0
50
100
150
200
250
Eight
Five
Four
One
Two
Zero


Social Smoker V/S Son
Frequency
0
50
100
150
200
250
Four
One
Three
Two
Zero


Social Smoker V/S Pet
Frequency
0
100
200
300
400
Eight
Five
Four
One
Two
Zero


Son V/S Pet
Frequency
0
50
100
150
200
250
Eight
Five
Four
One
Two
Zero

Missing Values

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

Missing data can occur because of nonresponse: no information is provided for one or more items or for a whole unit ("subject"). Some items are more likely to generate a nonresponse than others

In our dataset, there is no missing value present. A table below will throw light on our dataset:-

```
In [69]: #GEtting sum of all missing values
         AbsentData.isnull().sum()

Out[69]: Id                    0
         AbsentReason          3
         AbsentMonth           1
         WeekDay               0
         Season                0
         Expenses              7
         ResidentDistance      3
         ServiceTime           3
         Age                   3
         AverageWorkLoad      10
         HitTarget             6
         DiscplineFailure      6
         Education            10
         Son                   6
         SocialDrinker         3
         SocialSmoker          4
         Pet                   2
         Weight                1
         Height               14
         BodyMassIndex        31
         AbsentTime           22
         dtype: int64
```

	Variables	MissingPercentage
0	BodyMassIndex	4.189189
1	AbsentTime	2.972973
2	Height	1.891892
3	AverageWorkLoad	1.351351
4	Education	1.351351
5	Expenses	0.945946
6	Son	0.810811
7	DiscplineFailure	0.810811
8	HitTarget	0.810811
9	SocialSmoker	0.540541
10	Age	0.405405
11	AbsentReason	0.405405
12	ServiceTime	0.405405
13	ResidentDistance	0.405405
14	SocialDrinker	0.405405
15	Pet	0.270270
16	Weight	0.135135
17	AbsentMonth	0.135135
18	Season	0.000000
19	WeekDay	0.000000
20	Id	0.000000

All the missing values must have to me imputed before training the models. There are different methods available to impute the missing values and it depends on the data scientist to choose the appropriate method. After imputing the missing values all NA in the dataset will be removed. All count of missing values must be zero after imputing as shown below:-

```
In [72]: #Checking Missing Values exsits after filling or not.
         AbsentData.isnull().any()
         #ALL Missing Values are filled.

Out[72]: Id                 False
         AbsentReason       False
         AbsentMonth        False
         WeekDay            False
         Season             False
         Expenses           False
         ResidentDistance   False
         ServiceTime        False
         Age                False
         AverageWorkLoad    False
         HitTarget          False
         DiscplineFailure   False
         Education          False
         Son                False
         SocialDrinker      False
         SocialSmoker       False
         Pet                False
         Weight             False
         Height             False
         BodyMassIndex      False
         AbsentTime         False
         dtype: bool
```

Outliers

An Outlier is a rare chance of occurrence within a given data set. In Statistics and Data Science, 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.

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

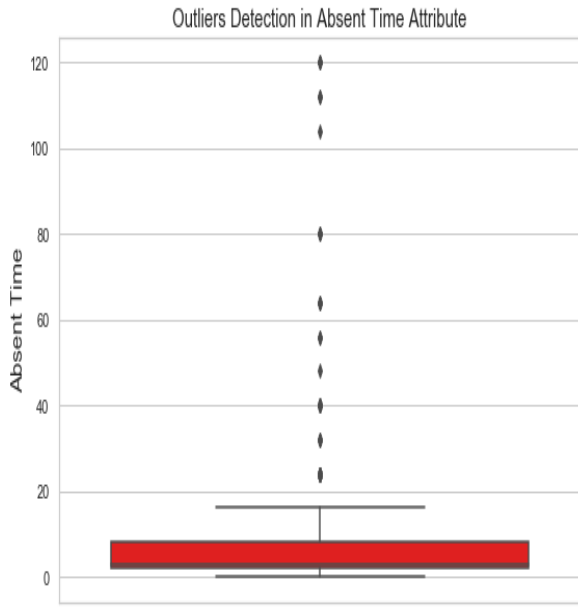
While outliers are attributed to a rare chance and may not necessarily be fully explainable, Outliers in data can distort predictions and affect the accuracy, if you don't detect and handle them.

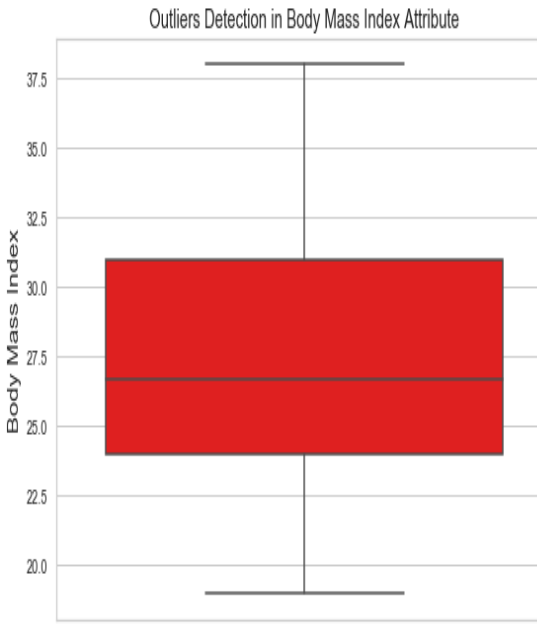
The contentious decision to consider or discard an outlier needs to be taken at the time of building the model. Outliers can drastically bias/change the fit estimates and predictions.

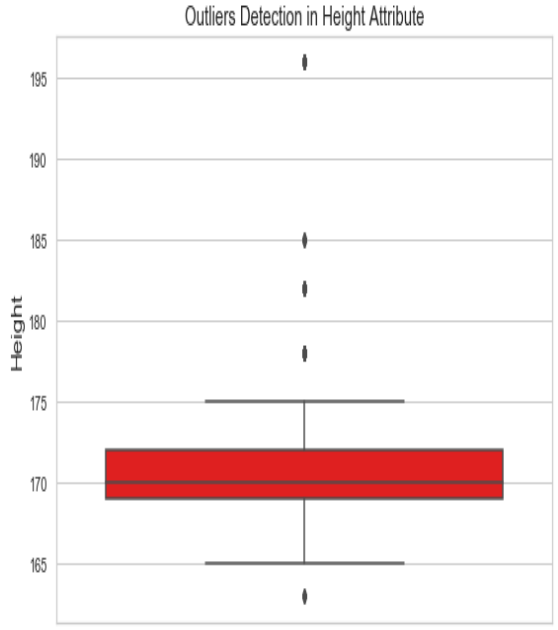
Detecting and Removing Outliers

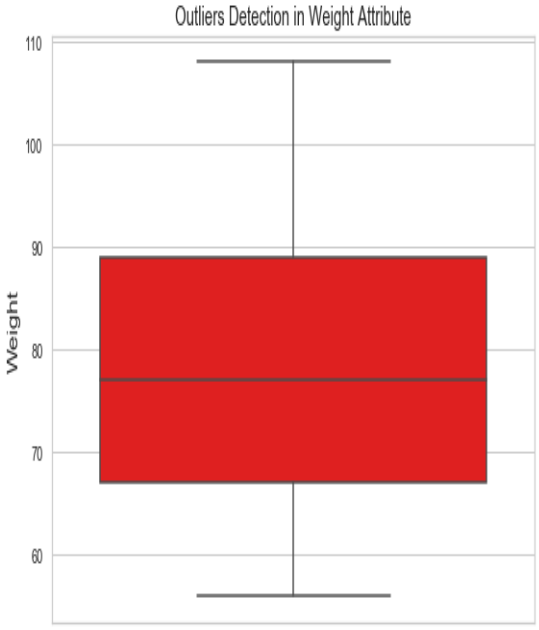
Mostly outliers are present in the continuous variables and box plot method is best and easy way to detect and remove outliers. Moreover, our dataset contains categorical variables that are already encoded so we will perform outlier detections only on continuous variables.

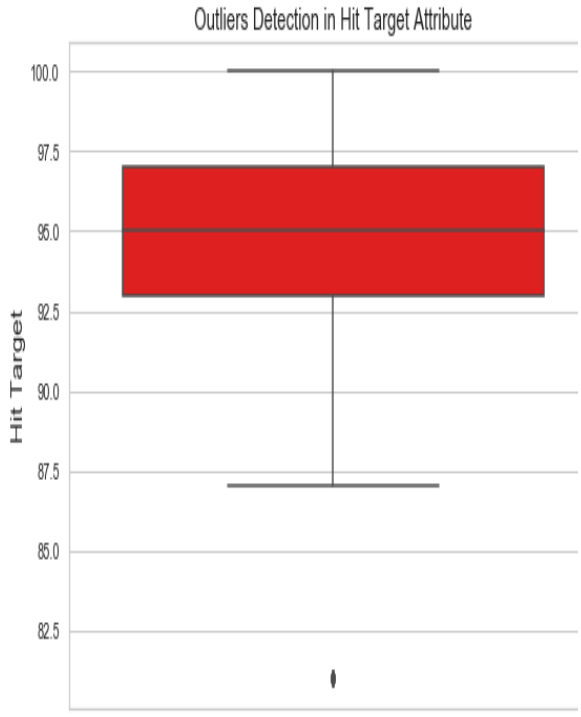
Box plot of all the variables are shown below:-

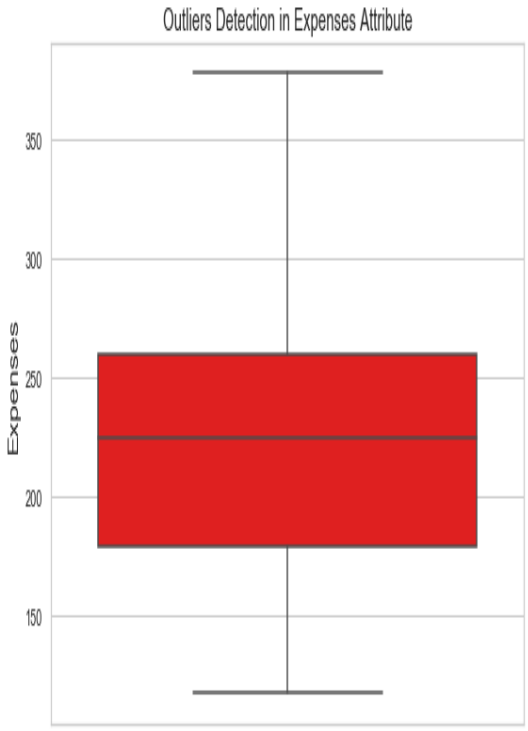


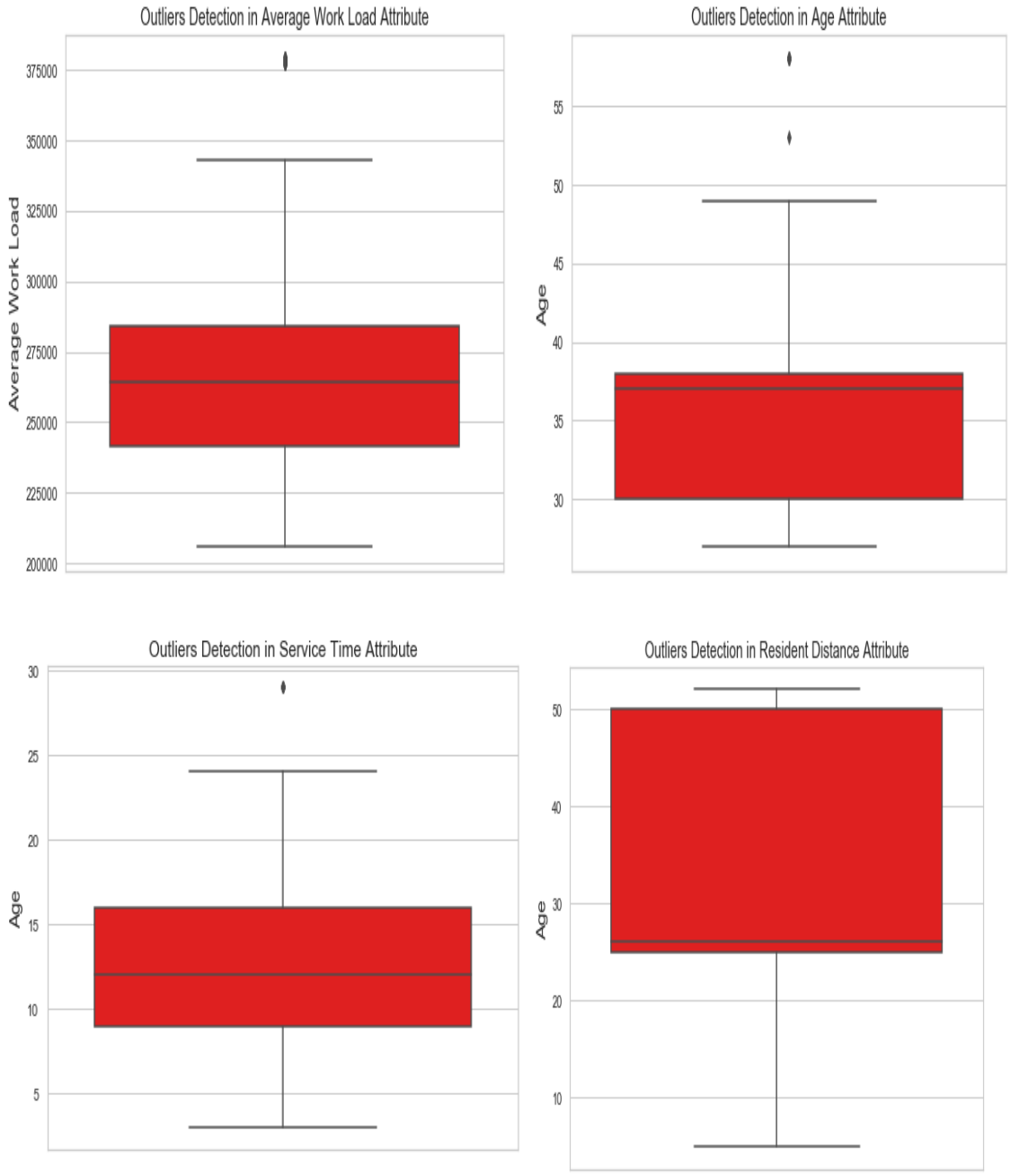



Outliers Detection in Height Attribute
Height
195
190
185
180
175
170
165


Outliers Detection in Weight Attribute
Weight
110
100
90
80
70
60


Outliers Detection in Hit Target Attribute
Hit Target
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5


Outliers Detection in Expenses Attribute
Expenses
350
300
250
200
150



Feature Selection

Machine learning works on a simple rule – if we put garbage in, we will only get garbage to come out.

This becomes even more important when the number of features is very large. We need not use every feature at our disposal for creating an algorithm. We can assist our algorithm by feeding in only those features that are really important. Feature subsets giving better results than complete set of feature for the same algorithm or – “Sometimes, less is better!”.

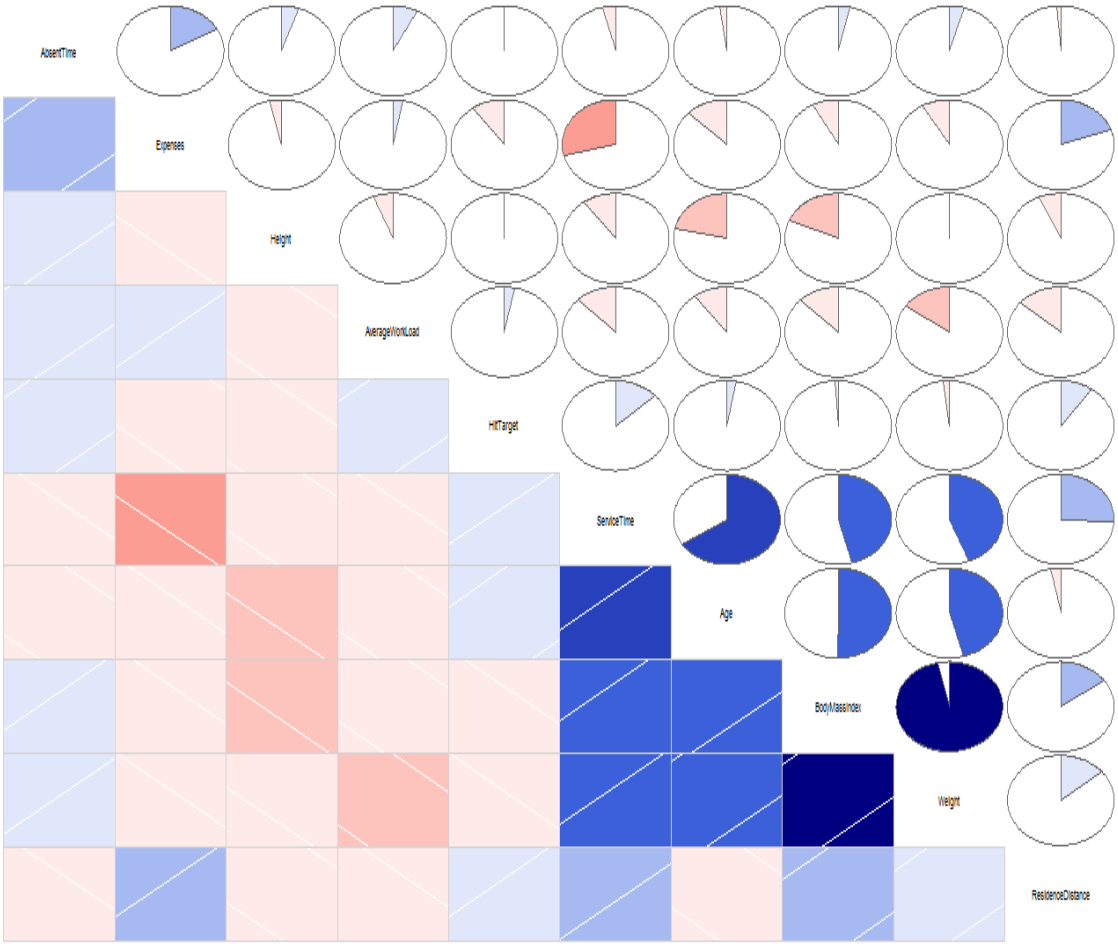
We should consider the selection of feature for model based on below criteria:-

1. The relationship between two independent variable should be less and
2. The relationship between Independent and target variables should be high.

Below figure shows a graphical display of a correlation matrix, called a correlogram. The cells of the matrix are coloured to show the correlation value.

	AbsentTime	BodyMassIndex	Height	Age	Weight	HitTarget	Expenses	AverageWorkLoad	ServiceTime	ResidentDistance
AbsentTime	1.0	-0.056	0.083	0.076	-0.013	0.035	0.042	0.028	0.012	-0.11
BodyMassIndex	-0.056	1.0	-0.13	0.46	0.89	-0.08	-0.13	-0.068	0.51	0.12
Height	0.083	-0.13	1.0	-0.068	0.29	0.091	-0.19	0.11	-0.064	-0.36
Age	0.076	0.46	-0.068	1.0	0.42	-0.035	-0.22	-0.037	0.67	-0.14
Weight	-0.013	0.89	0.29	0.42	1.0	-0.051	-0.21	-0.024	0.46	-0.038
HitTarget	0.035	-0.08	0.091	-0.035	-0.051	1.0	-0.067	-0.089	-7.9e-05	-0.02
Expenses	0.042	-0.13	-0.19	-0.22	-0.21	-0.067	1.0	0.015	-0.36	0.24
AverageWorkLoad	0.028	-0.068	0.11	-0.037	-0.024	-0.089	0.015	1.0	0.0051	-0.064
ServiceTime	0.012	0.51	-0.064	0.67	0.46	-7.9e-05	-0.36	0.0051	1.0	0.15
ResidentDistance	-0.11	0.12	-0.36	-0.14	-0.038	-0.02	0.24	-0.064	0.15	1.0

Correlation Plot



Feature Scaling

Feature scaling is done to reduce unwanted variation either within or between variables and to bring all of the variables into proportion with one another. I will use Normalization process to perform feature scaling. Formula for Normalization is given below:-

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Attribute after normalization are given below:-

Expenses	ResidentDistance	ServiceTime	Age	AverageWorkLoad	...
0.657692	0.659574	0.476190	0.30	0.312568	...
0.234615	0.978723	0.714286	0.55	0.312568	...
0.619231	0.000000	0.523810	0.60	0.312568	...
0.657692	0.659574	0.476190	0.30	0.312568	...
0.234615	0.978723	0.714286	0.55	0.312568	...

Modeling

Model Selection

In the case of this dataset we have to predict the count of total bike rented on basis of environmental and seasonal condition. The target variable here is a continuous variable and for a continuous variable we can use various Regression models. Trained model having less error rate and more accuracy will be our final model. Different machine learning methods which will be used to train our final model are mentioned below:-

1. Decision Tree Regression Model
2. Random Forest Model
3. Liner Regression Model

Final model with will be with the higher accuracy which we will able to decide at the end of the modelling process.

Decision Tree Regression Model

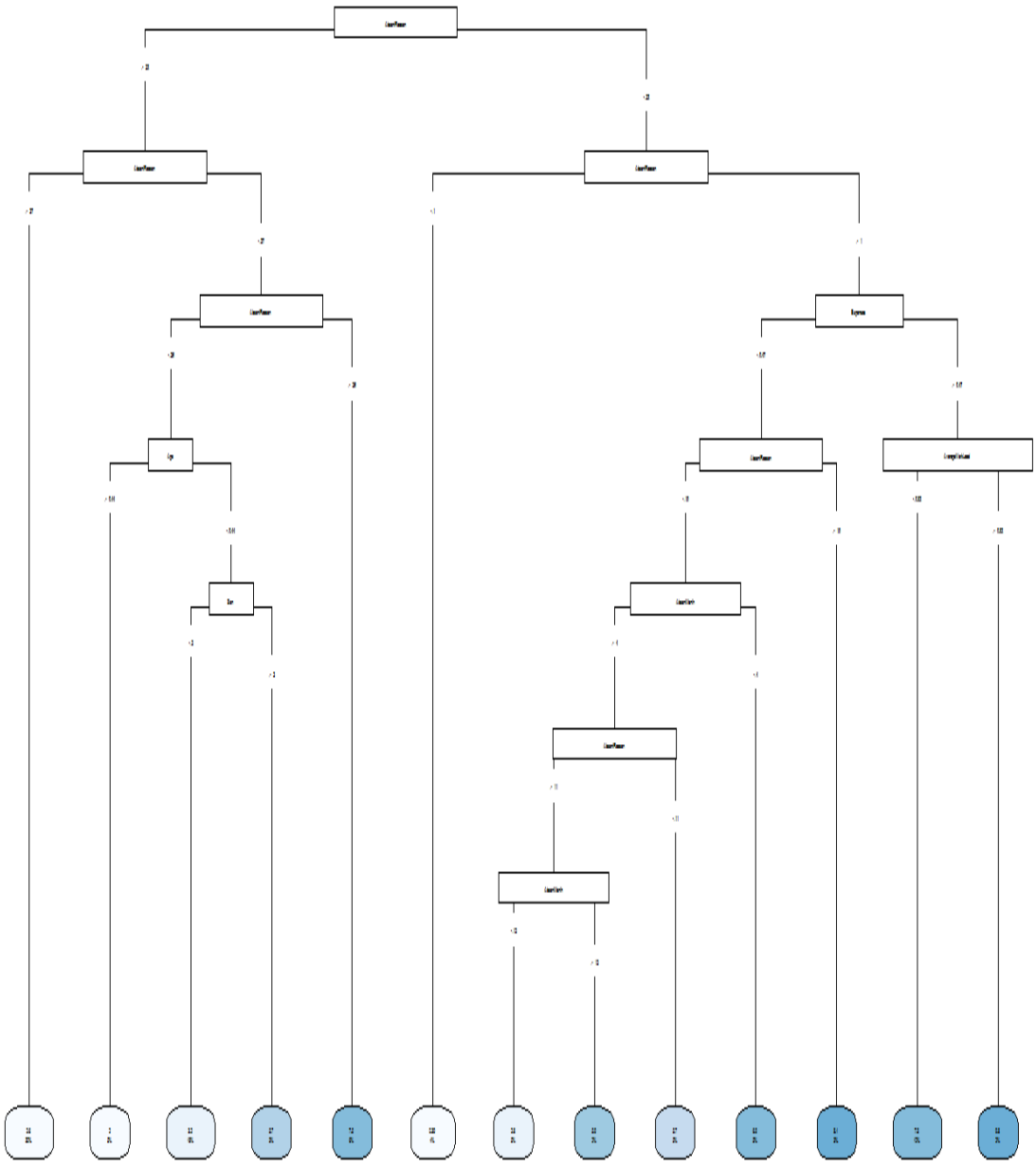
Decision tree builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an

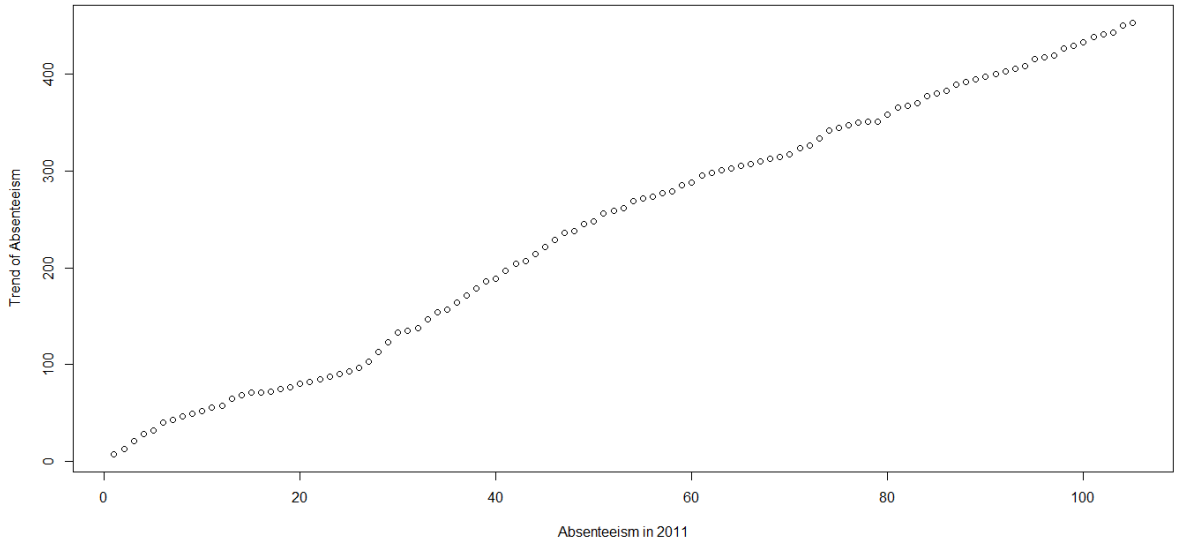
associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

```
#Training Decision tree
DecisionTree = DecisionTreeRegressor(max_depth=10,random_state=0)
DecisionTree.fit(X_Train,Y_Train)
```

```
DecisionTreeRegressor(criterion='mse', max_depth=10, 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=0, splitter='best')
```

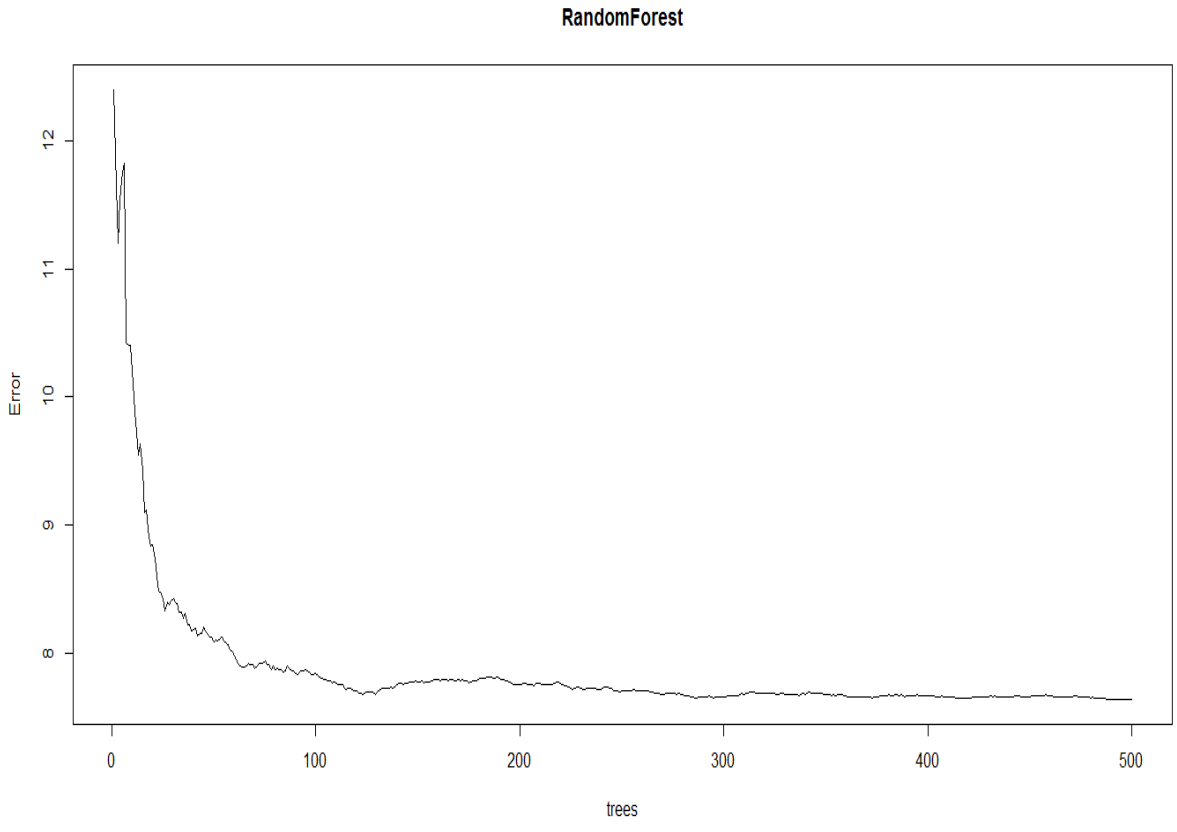
Graphical Visualization of a trained modal is shown below:



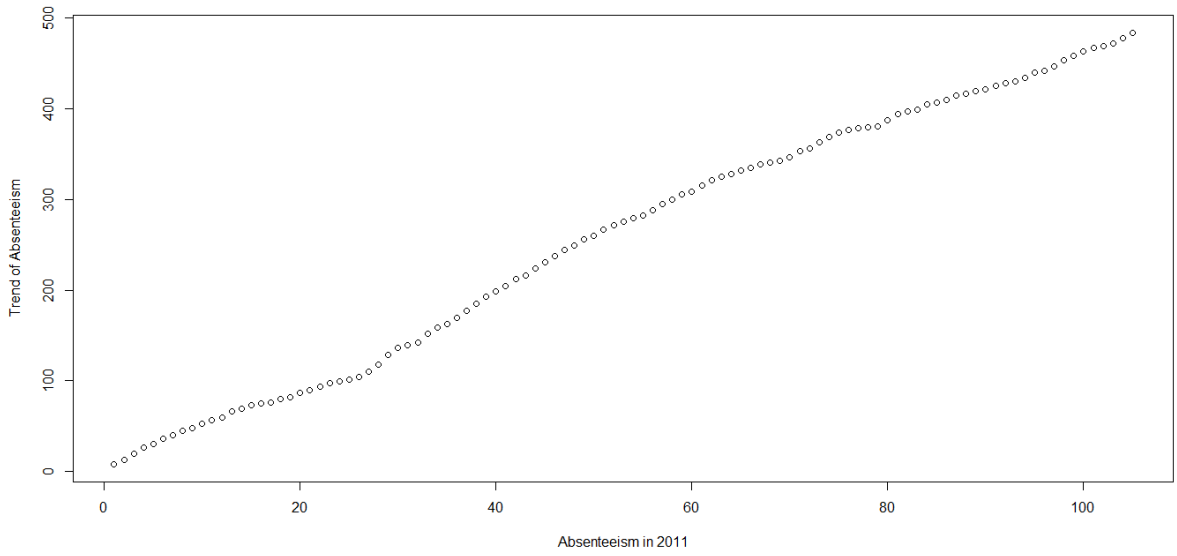


Random Forest Model

Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks.

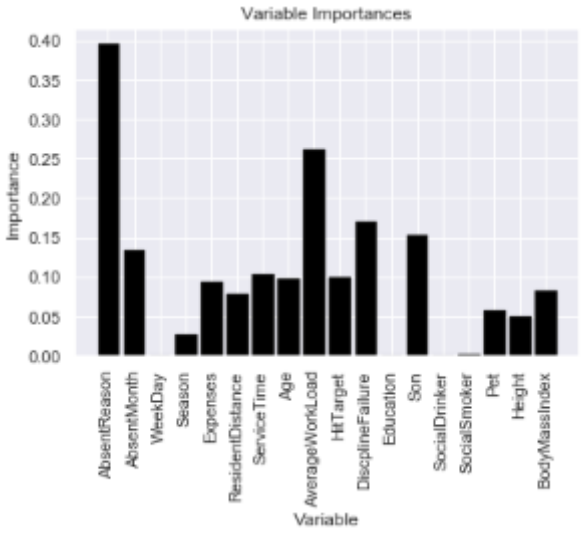


Visualization and Accuracy of the trained model is shown below:-

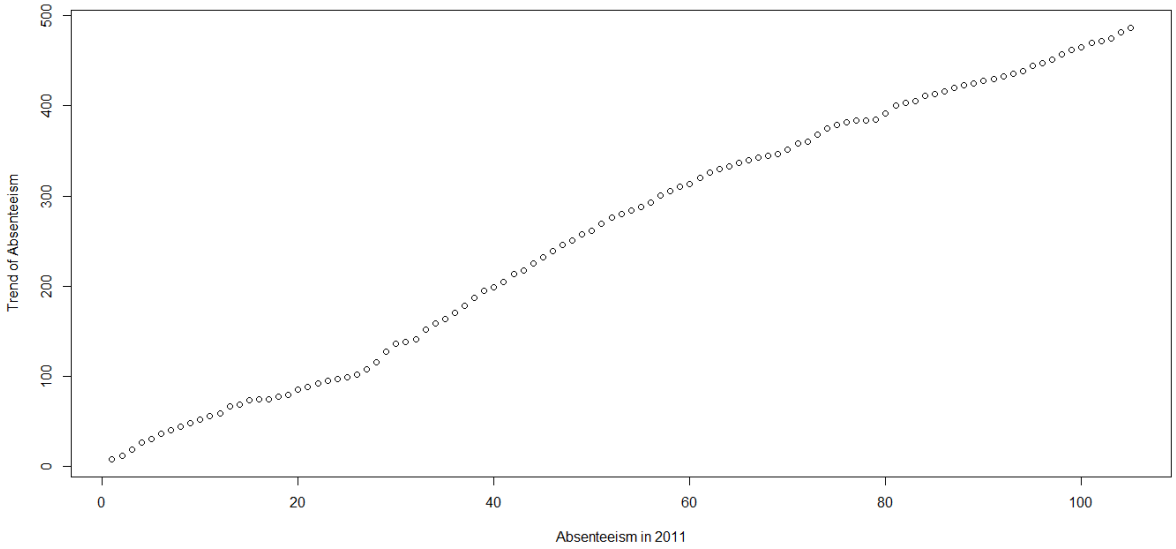


Over repetitions of iteration and training the model with all features were not increasing accuracy. So in order to increase the accuracy of the model we need to check importance of all features in the model then remove the features with least importance. Below graphs shows the importance of all the features present in the model.

	Importance
AbsentReason	0.289164
AverageWorkLoad	0.100921
DiscplineFailure	0.076935
HitTarget	0.071110
AbsentMonth	0.066839
WeekDay	0.065074
BodyMassIndex	0.057203
Expenses	0.050828
Age	0.040098
Height	0.036056
ResidentDistance	0.033223
Season	0.029051
ServiceTime	0.028654
Son	0.023746
Education	0.009844
Pet	0.009831
SocialDrinker	0.009393
SocialSmoker	0.002031

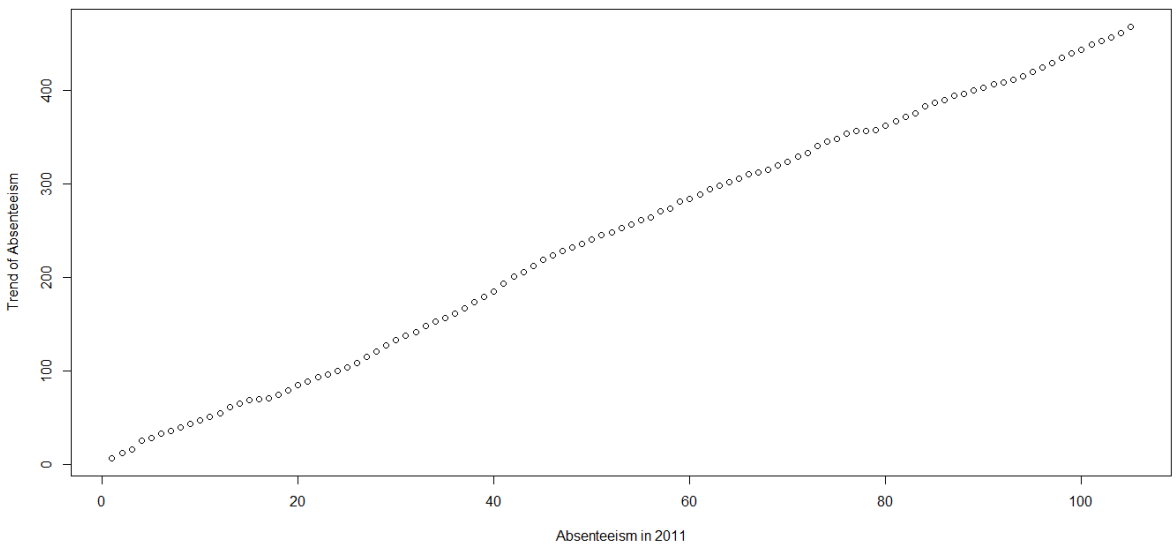


From the above graph, the next task is to remove the feature with less importance. So the graph of trained model after removing those features is shown below:-



Linear Regression

Multiple Linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regressions is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical. The trained model is shown below:-

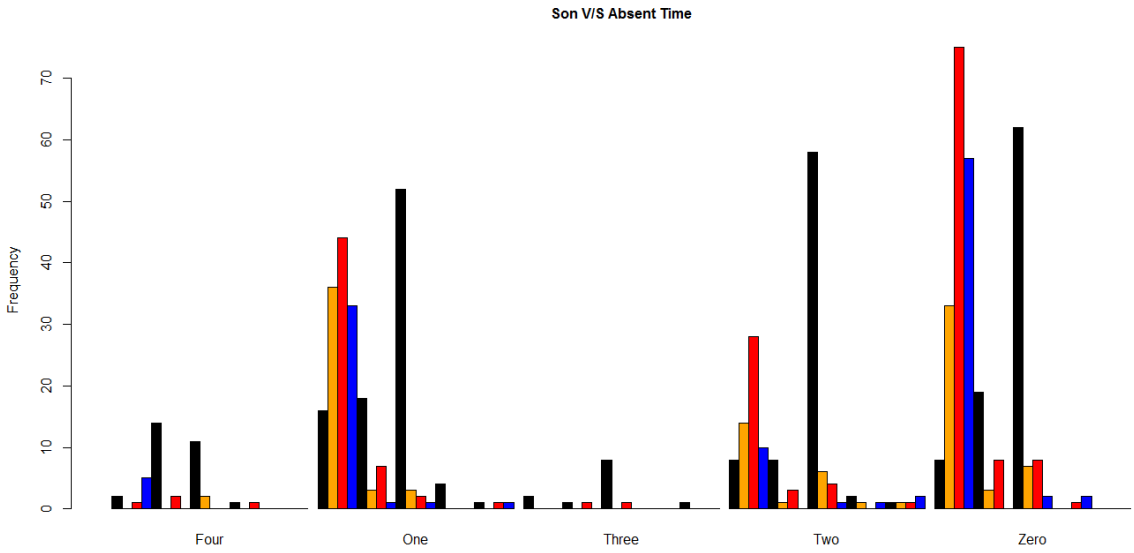


--------- VIFs of the remained variables --------

Variables	VIF
Id	7.887307
AbsentReason	1.460496
AbsentMonth	1.523459
WeekDay	1.111049
Season	1.519623
Expenses	3.806174
ServiceTime	5.380615
Age	3.837641
AverageWorkLoad	1.232699
DiscplineFailure	1.438735
Education	4.858774
Son	2.123952
SocialDrinker	8.340348
SocialSmoker	1.979498
Pet	3.546069
Height	1.662905
BodyMassIndex	4.054443

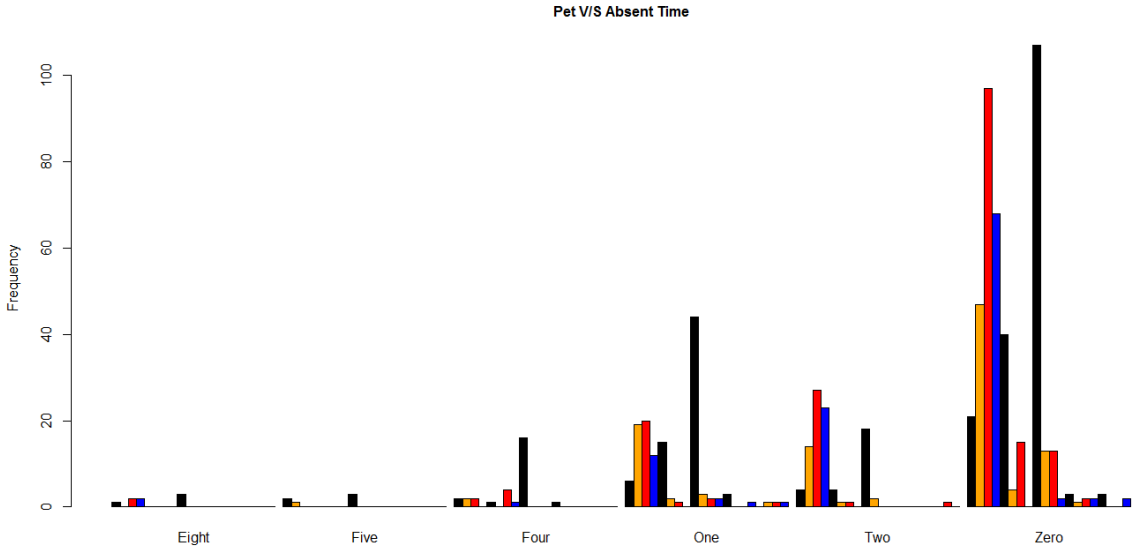
What changes company should bring to reduce the number of absenteeism?

- Absent Time v/s Number of Sons



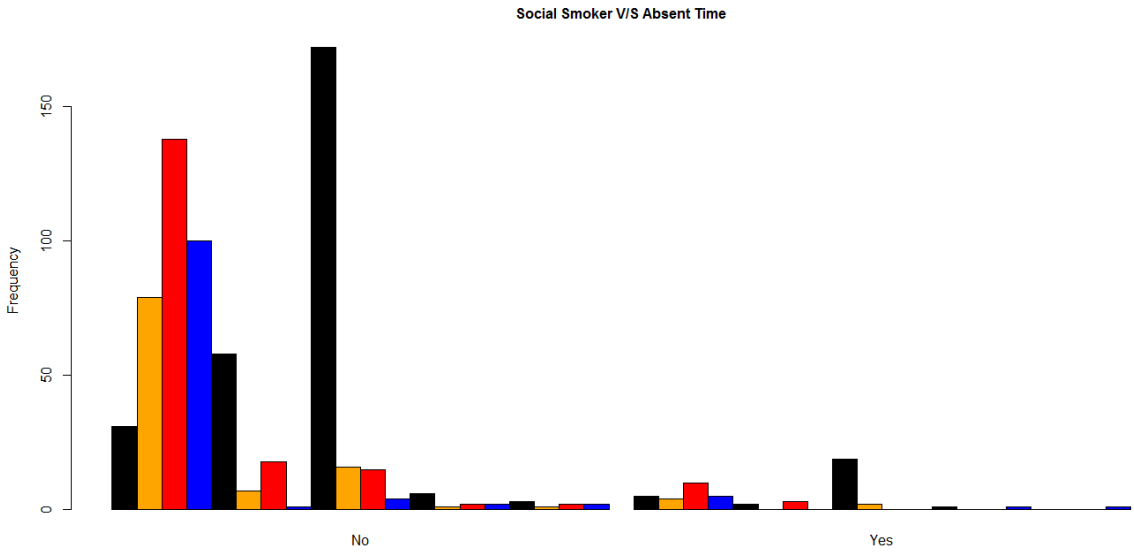
Employee with zero son are maximum time absent from their time. They may need some break from professional life so arranging some fun activities with the help of HR department can resolve this problem. Besides this employee with two and one child are also absent. It may be because their children are very small in age. Providing crèche facility can resolve this problem.

- Absent Time v./s Number of pets



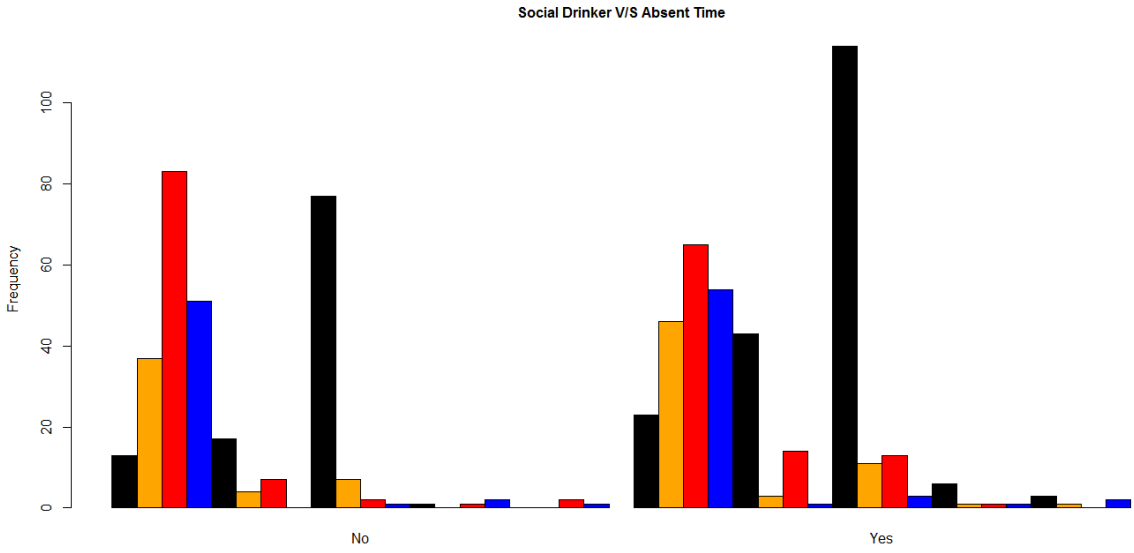
Employee with zero pets are most of the time absent.

- Social Smoker v/s Absent Time



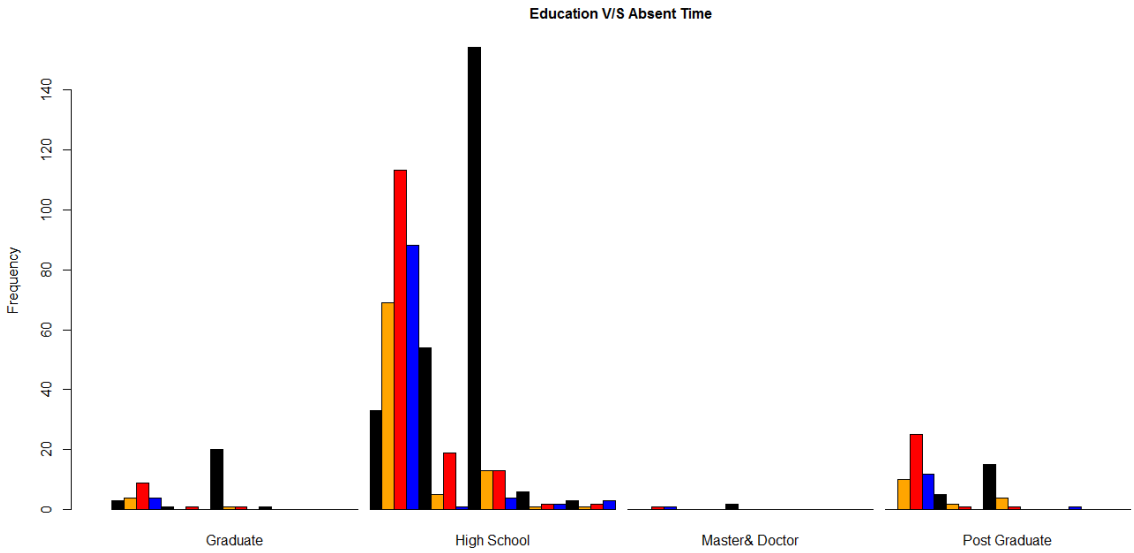
People who are not social smoker are mostly absent from their job.

- Social Drinker v/s Absent Time



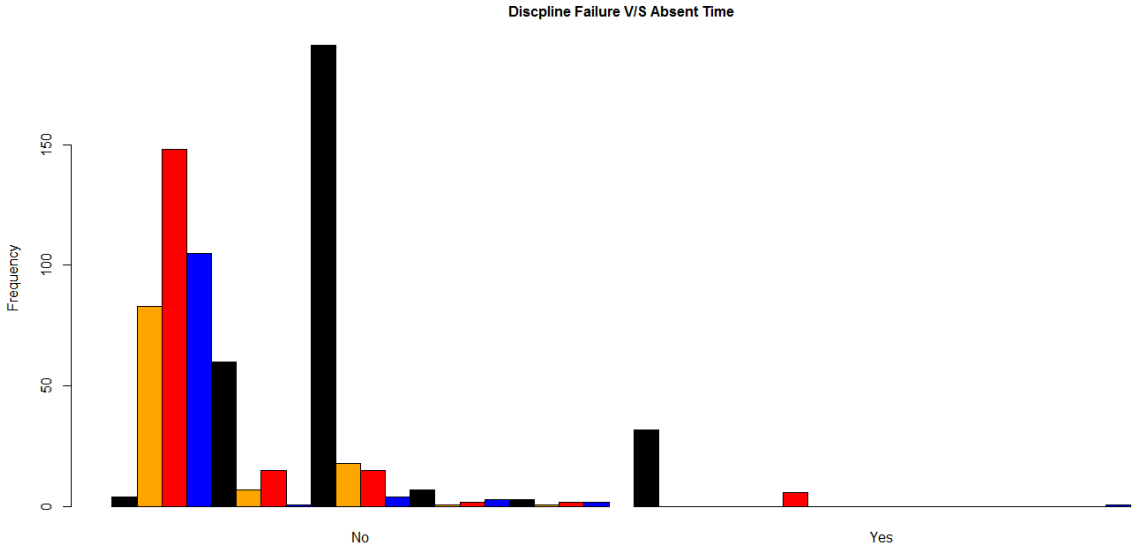
People who are not social smoker are mostly absent from their job.

- Absent Time v/s Education



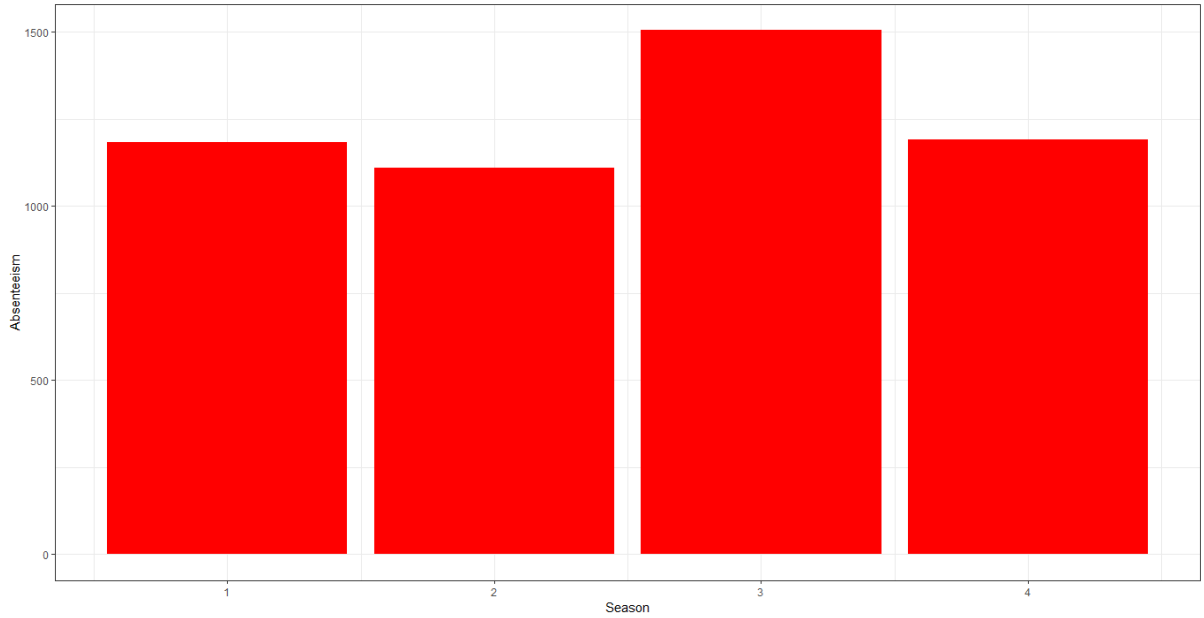
Employees who are less educated are mostly absent. Hiring more educated employees can resolve this issue.

- Absent Time v/s Discipline Failure



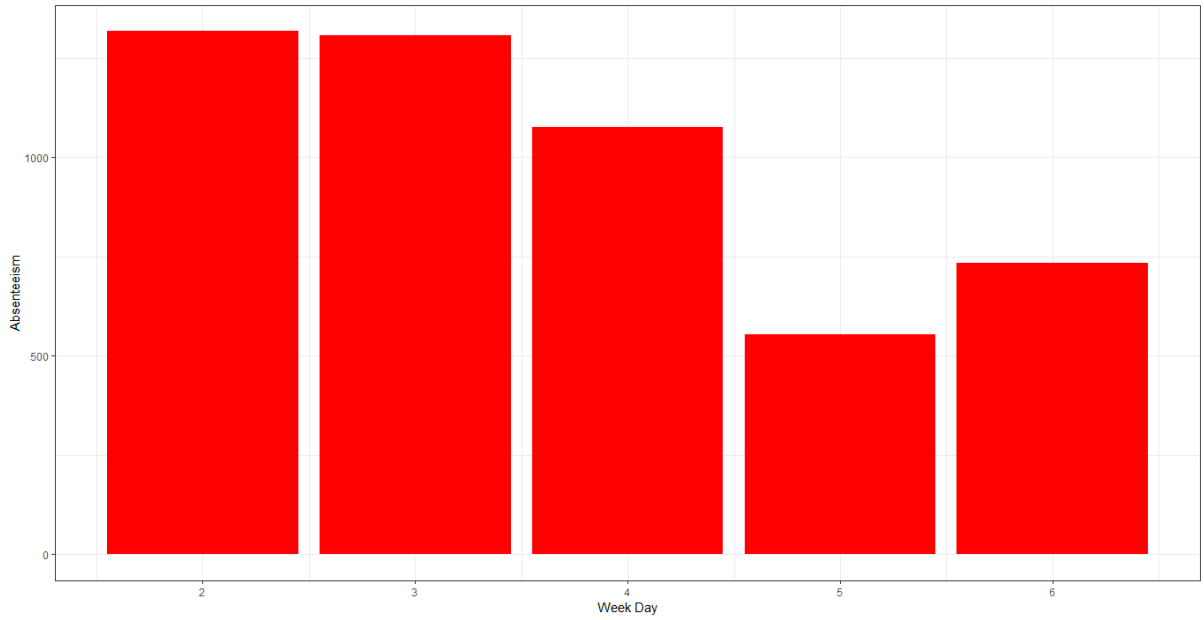
May be company is more strict about its policies. This is leading to more absent time. Giving some relaxation to their employee's can reduce employee absent time.

- Absent Time v/s Season



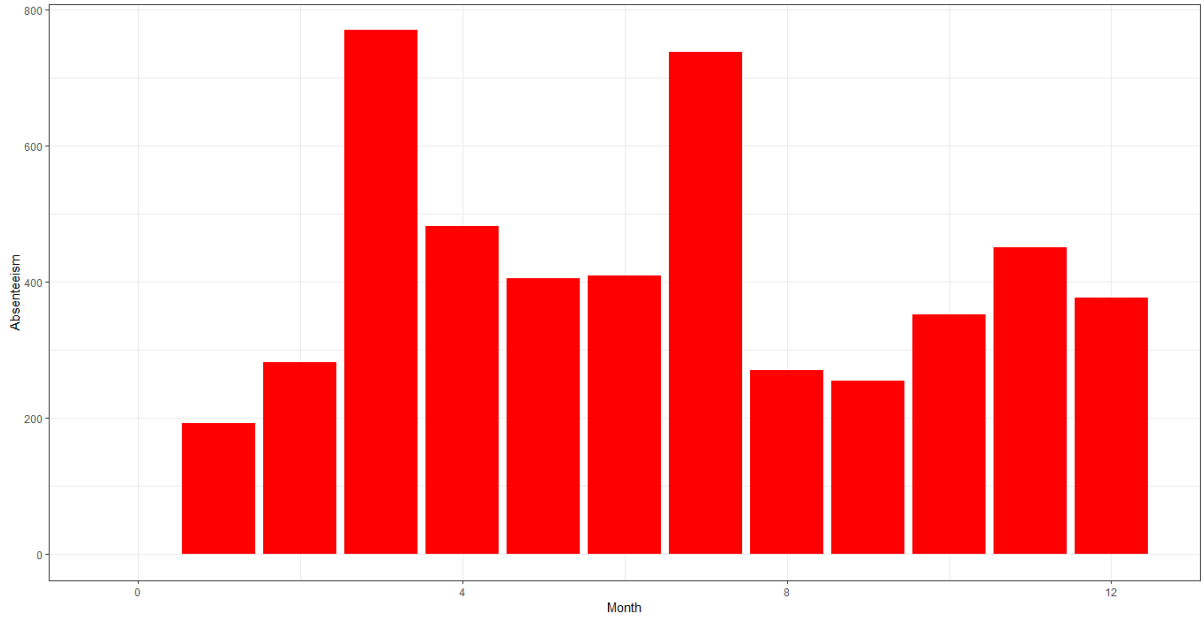
Most of the employees are absent in Winter Season. May be poor heating facility in office may leading to this. Providing blower can resolve this issue. Besides this providing them hot snacks with tea regular will pushes employee to join office daily.

- Absent Time v/s Week day



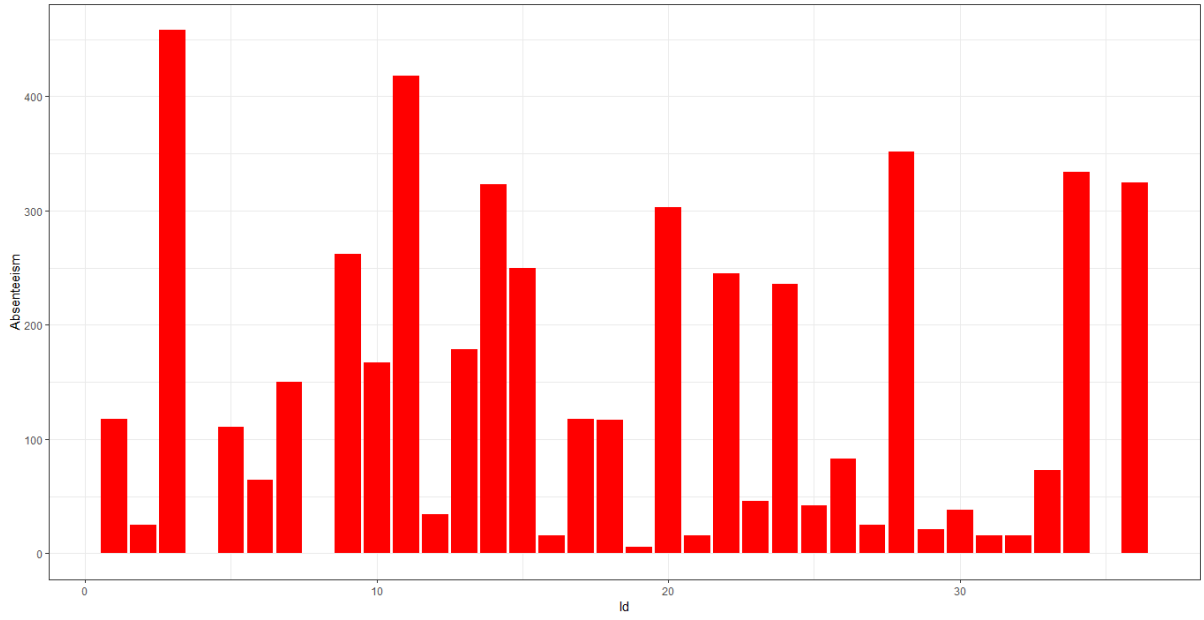
Employees are mostly absent on Monday and Tuesday. Creating meeting on these days will pushes employee to join office on these two days. Issuing warning can also help.

- Absent Time v/s Absent Month



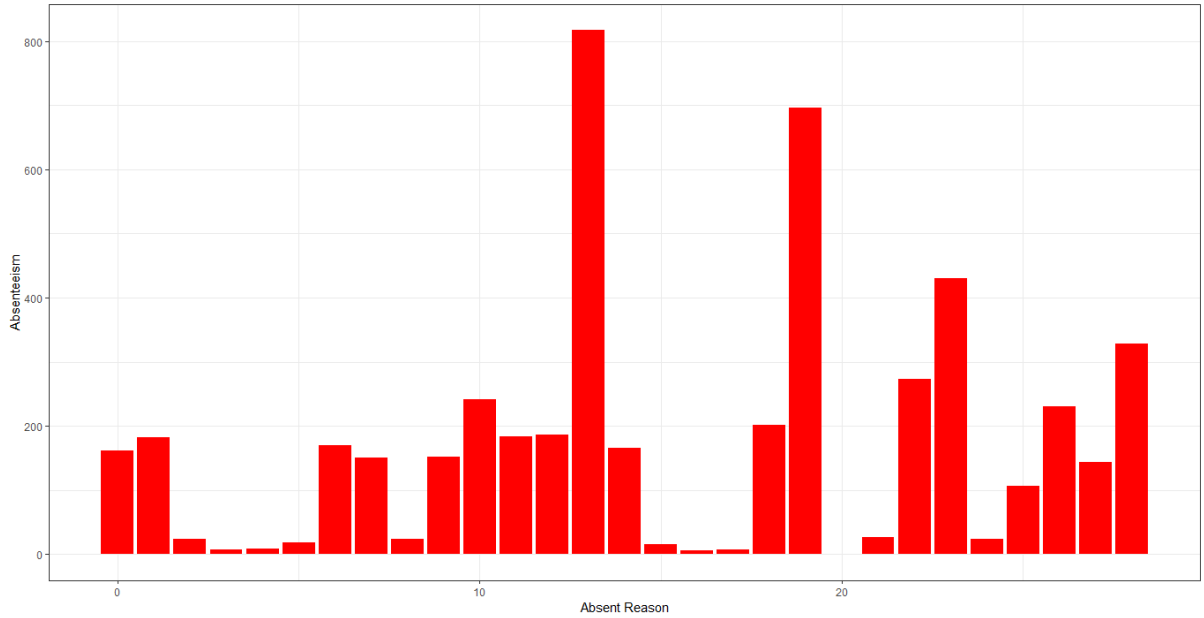
Employees are mostly absent in the month of March and July. Keeping Appraisal season around this month will resolve this issue.

- Absent Time v/s Id



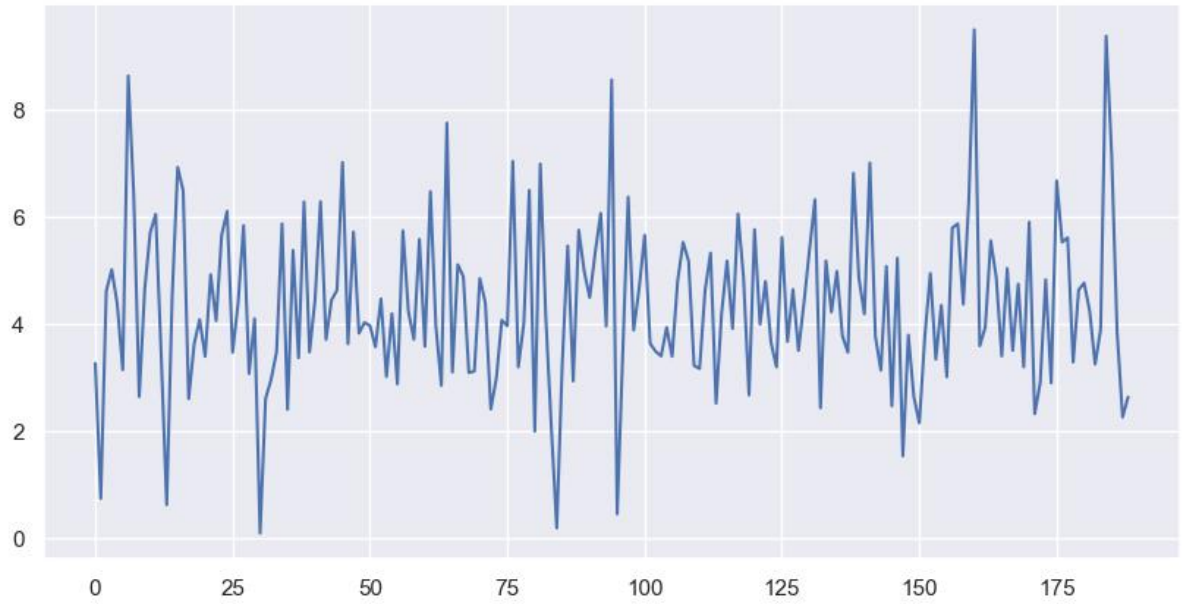
Maximum Absent Rate is with 3,11,20,22,28,34,36 a. giving them warning to be regular can help the firm. Company must interact with the employees of these Ids and motivate them to be regular. Arranging of some motivational talk can increase their moral and helps the firm to deal to reduce absenteeism time of its employee.

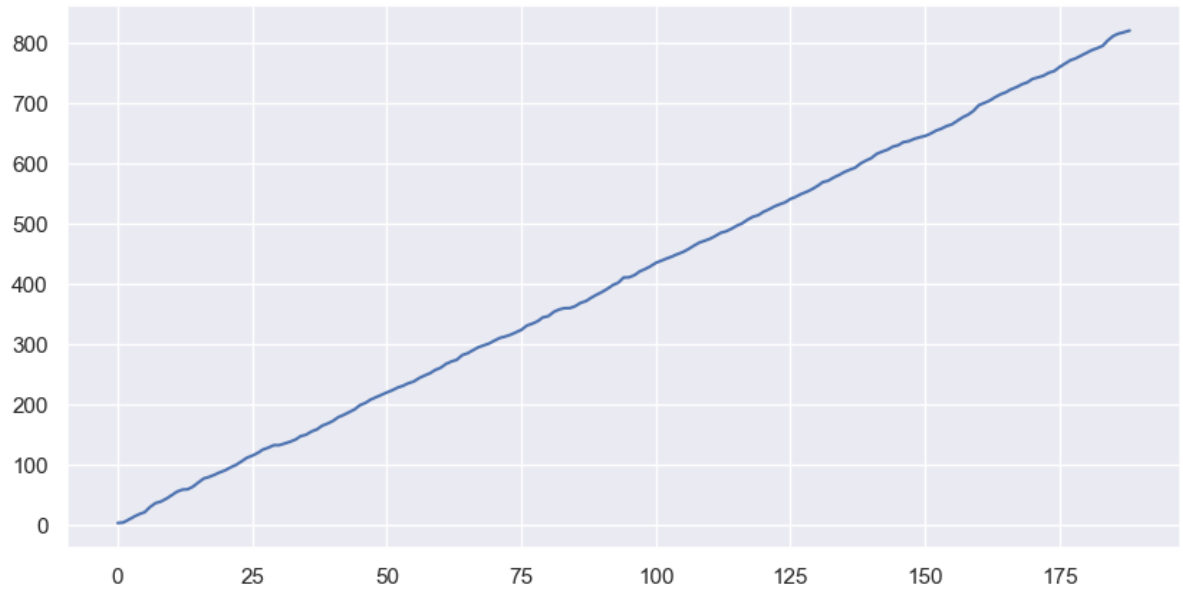
- Absent Time v/s Absent Reason



Employees having maximum absent time if they are suffering from disease number 23,26,27,28,13,10,1. Proper medical check up and treatment can help them to reduce employee absent time. Company can provide regular medical check up to their employees. Providing employee's with ESI facilities can also help to deal with this type of problem.

Graph of Prediction for future Business Decisions




800
700
600
500
400
300
200
100
0
0
25
50
75
100
125
150
175