
IT TAKES
21 DAY
TO MAKE
OR BREAK
A HABIT

Gokhale Education Society's
H.P.T. Arts and R.Y.K. Science College
Prin. T. A. Kulkarni, Vidya Nagar, Nashik- 422005
E-Mail : prinhptryknsk@rediffmail.com ☎ : 0253-2572153
"Higher Education for All"

▶ Permanently Affiliated to Savitribai Phule Pune University (ID No. : PU/NS/AS/001(1924) ▶ NAAC Re-Accredited : 'A' Grade ▶ ISO 9001:2015 Certified

DEPARTMENT OF STATISTICS

T.Y.B.Sc. 2022-2023

Project Title: - "Analysis of proven technique 21 days challenge"

Ahire Vaibhav (1)

Deochake Nupur (7)

Joshi Devashri (13)

Kurade Neha (19)

Patro Saikrishna (25)

Shukla Shruti (31)

.....

Mrs. V. S. Joshi

HEAD OF DEPARTMENT

.....

Mrs. V. P. Aher

PROJECT GUIDE

➢ ACKNOWLEDGEMENTS

We would like to express my heartfelt thanks to all those who have supported me throughout this statistics project. First and foremost, I extend my gratitude to my instructor who has provided me with valuable guidance, feedback, and motivation throughout the project. I am grateful for their insights, expertise, and support that have helped me to better understand the concepts and principles of statistics.

We wish to thank our HOD of statistics department **Mrs. V.S.Joshi** ma'am and our project guide **Ms. V.P.Aher** ma'am for the valuable support and guidance. Thanks to the teaching faculty of our college for their support in completing the work successfully

I would also like to thank our classmates for their helpful feedback and constructive criticism during our group discussions. Their insights and ideas have been invaluable in shaping our understanding of the project's goals and objectives.

Thank you all for your support, guidance, and encouragement throughout this project.

INDEX

INTRODUCTION

In today's world, where people are struggling to achieve their goals and make changes in their lives, the concept of building habits has become increasingly important. We can create positive changes by following a structured approach and making small, incremental changes in our behaviour.
In this statistics project, we aim to analyze the impact of following the "21 Days Challenge" as recommended by James Clear in his book, "Atomic Habits." The challenge involves making small changes in our daily routine for 21 days, which can help us build positive habits and break negative ones.

21 day rule is a very natural process to understand science of brain if you will follow some habit for 21days continue then it will become part of life and even if you forget to do that but you will do that ,that how human nature is …a very famous saying perhaps u herd ..That...

NATURE AND SIGNATURE NEVER CHANGES.

Through this project, we hope to provide valuable insights into the effectiveness of the 21 Days Challenge and its potential to help people make positive changes in their lives.

METHODOLOGY

1) Diagrammatic representation

a) Pie Chart

Pie charts were invented by William Playfair in 1801.

A pie chart (or a circle chart) is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice is proportional to the quantity it represents.

It presents the relationship of different parts of the data. One would easily see the biggest or smallest share of the total data, by simply looking at the pie chart.

While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented.

The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sector of a pie chart forms a specific portion of the total(percentage). The sum of all the data is equal to 360°.

Pie charts are very widely used in the business world and the mass media.

b)Multiple Bar Diagram

Multiple bar diagram is identical to a regular bar graph with the exception that there are two or more bars in each category, one for each subdivision. This diagram is created using the same method as a straightforward bar chart with the exception that we use various tones, hues and/or dots to distinguish between various phenomena. If the sum of various phenomena is meaningless, we often draw multiple bar charts.

Advantages of Multiple Bar Diagram

Multiple bar diagram advantages are as follows:

It can be applied to contrast two or more data points from a specific data set.

No need to create two different diagrams.

A single diagram contains information about two or more variables side by side.

It looks visually appealing and is very simple to understand.

c) Spike and Rod Plot

Spike plot produces a frequency plot for a variable in which the frequencies are depicted as vertical lines from zero. The frequency may be a count, a fraction, or the square root of the count (Tukey's rootogram, circa 1965).

2) Testing

a) Chi Square Test Of Independence

Suppose that the given data are classified into r levels of attributes A denoted by A1,A2,... Aj..... Ar and s levels of attribute B represented by B1,B2,... Bj,...Bs

N= Σ Σ Oij = Total observed frequency,

(Ai)= Σ Oij= Total of observed frequencies in the ith row, i = 1,2... r

(Bj)= Σ Oij = Total of observed frequencies in the jth column, j=1,2... s

A \ B	B_1	B_2.....Bj.....Bs	total
A_1	O_{11} (e_{11})	O_{12}.....O_{1j}.....O_{1s} (e_{12}) (e_{1j}) (e_{1s})	A_1
A_2	o_{21} (e_{21})	O_{22}.....O_{2j}.....O_{2s} (e_{22}) (e_{2j}) (e_{2s})	A_2
:	:	:	:
Ai	Oi_1 (ei_1)	Oi_2.....Oij.....Ois (ei_2) (eij) (eis)	Ai
:	:	:	:
Ar	Or_1 (er_1)	Or_2.....Orj.....Ors (er_2) (erj) (ers)	Ar
total	B_1	B_2 Bj Bs	N

We wish to test Ho: Two attributes are independent against H: The attributes are not independent.

e ij = (A i)(B j)/N; i=1,2...,r j=1,2...s

= [(Total of observed frequency in the i^{th} row) x (Total of observed frequencies in the ith column)]/ Grand total of all observed frequencies.

where eij: expected frequency of $(i, j)^{th}$ cell.

Hence, using the above formula we can find all expected frequencies.

Criteria:

We reject H0 at % level of significance if X^2 (r-1)(s-1) >= X^2(r-1)(s-1); and accept H0 otherwise.

If H0, is true, the statistic

$X^2 = \sum\sum \frac{(O_{ij} - e_{ij})^2}{e_{ij}} = \sum\sum \left[\frac{(O_{ij}^2)}{e_{ij}} \right] - N$

b) Proportion Test

Two sample Z test of proportions is the test to determine whether the two populations differ significantly on specific characteristics. In other words, compare the proportion of two different populations that have some single characteristic. It calculates the range of values that is likely to include the difference between the population proportions.

Let,
$\mathbf{P_1}$=proportion of specific items in first population.
$\mathbf{P_2}$=proportion of specific items in second population.
$\mathbf{n_1}$=size of sample drawn from first sample.
$\mathbf{n_2}$=size of sample drawn from second sample.
$\mathbf{X_1}$-Number of items of specific type in first sample.
$\mathbf{X_2}$-Number of items of specific type in second sample.
$\mathbf{P_1}=X_1/n_1$=proportion of specific items in first sample.
$\mathbf{P_2}=X_2/n_2$= proportion of specific items in second sample.

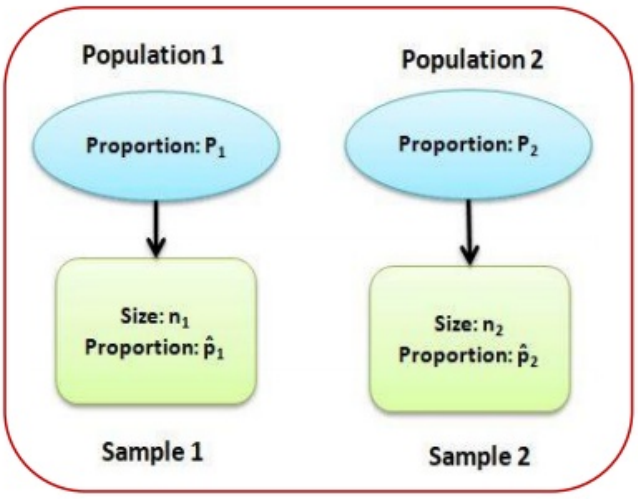
The hypothesis for such problems will be:
Null Hypothesis, H_0: $P_1=P_2$
V/s Alternative Hypothesis,
$H_1: P_1 \neq P_2$
$H_1: P_1 < P_2$
$H_1: P_1 > P_2$

R commands for null hypothesis H0: P1=P2
(a) Consider the alternative hypothesis: H1:P1=P2
prop.test (x,n,conf.level=c)

(b) Consider the alternative hypothesis H1:P1>P2.

prop.test (x,n,conf.level=c alternative="greater")
(c) Consider the alternative hypothesis H1:P1<P2.
prop.test(x,n,conf.level=c alternative="less")

CRITERIA: Here level of significance α% is less than p-value then we may accept H_0.



3) Data Analytics Algorithm

a) K-Nearest Neighbor(K-NN)

Evelyn Fix and Joseph Hodges are credited with the initial ideas. K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

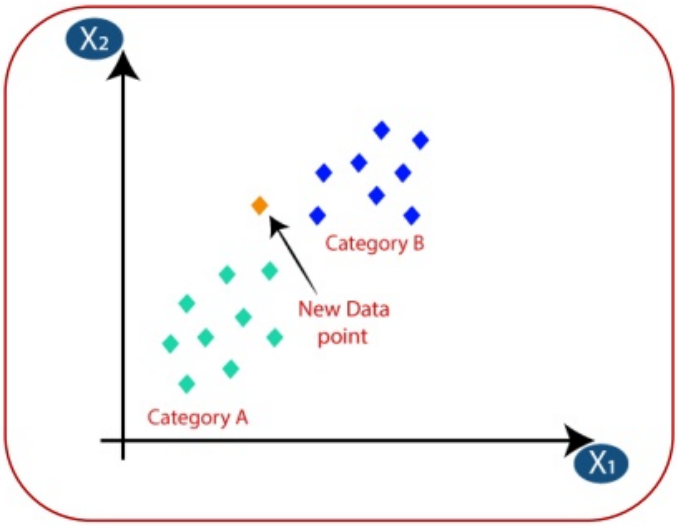
K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K in KNN is a parameter that refers to the number of the nearest neighbors to include in the majority voting

How do we choose K?

Sqrt(n), where n is a total number of data points

Here, we use distance formula i.e. Euclidean distance to find the distance between any two points

$$d=\sqrt{((x2-x1)^2+(y2-y1)^2)}$$



b) Logistic Regression Model

The logistic regression as a general statistical model was originally developed and popularized primarily by Joseph Berkson.

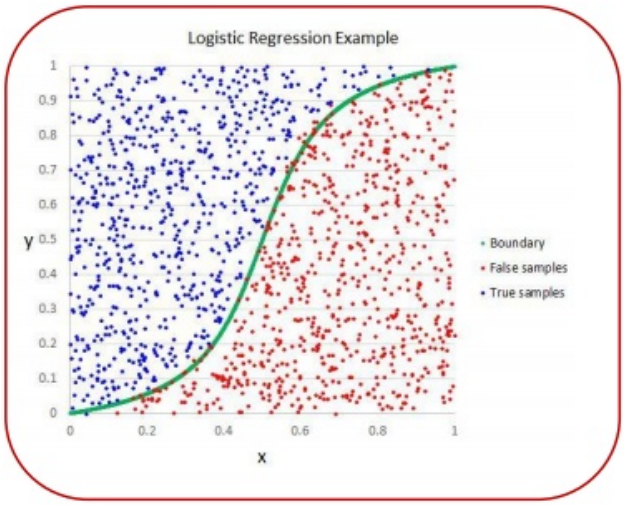
Logistic Regression is one of the basic and popular algorithms to solve a classification problem. It is named 'Logistic Regression' because its underlying technique is quite the same as Linear Regression. The term "Logistic" is taken from the Logit function that is used in this method of classification.

Logistic regression is used when your Y variable can take only two values, and if the data is linearly separable, it is more efficient to classify it into two separate classes.

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The result is the impact of each variable on the odds ratio of the observed event of interest.

The logistics regression model can be written as:

Y= e^(b0+b1X)/1+e^(b0+b1X)



c) Naïve Bayes

Thomas Bayes is the guy who founded Bayes theorem which Naive Bayes Classifier is based on.

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

It is mainly used in text classification that includes a high-dimensional training dataset.

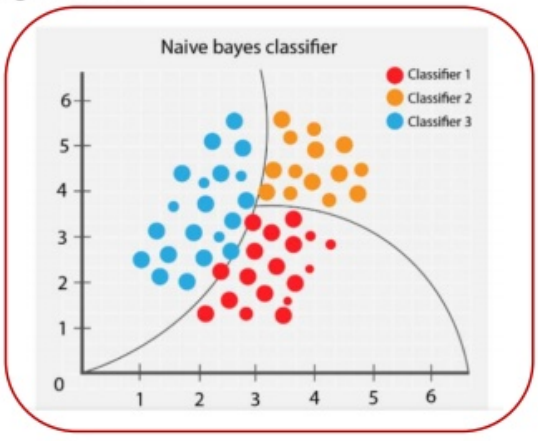
Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

The formula for Bayes' theorem is given as:

Naïve Bayes Classifier Algorithm

$P(A|B) = P(B|A) * P(A) / P(B)$



d) Decision Tree

A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm.

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

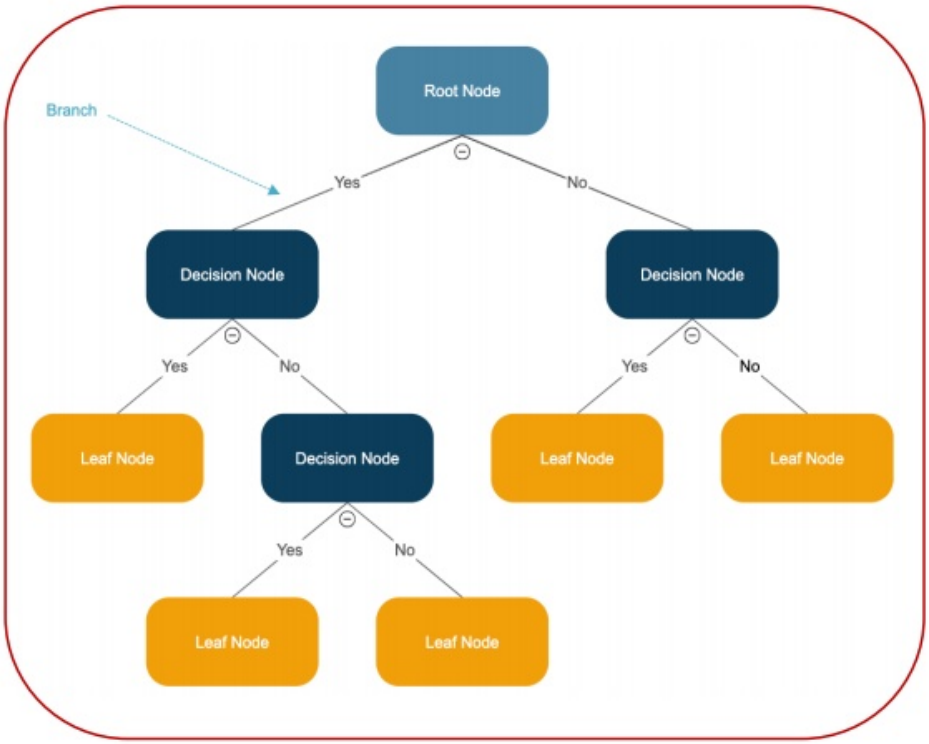
A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The benefits of having a decision tree are as follows −

a)It does not require any domain knowledge.

b)It is easy to comprehend.

c)The learning and classification steps of a decision tree are simple and fast.



SOFTWARE

- Statistical software packages used:
  - MS-Excel
  - MS-Word
  - R software
  - R studio

DATA

Week 1

Age				
Under 18	18-24	25-34	35-44	Above 45
0	1	2	3	4

Gender	Male	Female
	0	1

Do you follow any routine?	
Yes	No
1	0

Have you ever followed 21 days challenge?	
Yes	No
1	0

Was it effective?				
Agree	Strongly Agree	Neutral	Disagree	Strongly disagree
0	1	2	3	4

Do you struggle to be consistent?				
Never	Often	Rarely	Sometimes	Every time
0	1	2	3	4

Do you think consistency is the key for everything?			
Agree	Strongly Agree	Neutral	Strongly disagree
0	1	2	3

Do you believe this 21 days challenge will bring change in your life?				
Agree	Strongly Agree	Neutral	Strongly disagree	Blank
0	1	2	3	4

Week 2

Have you followed your new routine?		
No	Ye	NA
0	1	2

Are you reading any self-help book? Or have you ever read any self-help books?		
No	Ye	NA

0	1	2

Do you feel enthusiastic to continue your routine?		
No	Ye	NA
0	1	2

Week 3

Did you follow your routine today?		
No	Ye	NA
0	1	2

Today we are readily available with all these factors, but how easy(on the scale of 1-5) it is to implement them in our day to day lives?					
1	2	3	4	5	NA

Does the motivation gained from these motivational factors stay for longer periods?					
Agree	Strongly Agree	Neutral	Disagree	Strongly disagree	NA
1	2	3	4	5	NA

Do you think people around us has a role to play too help us being consistent?			
No	Yes	Maybe	NA
0	1	2	NA

Week 4

Do you think diet and consistency in life are interrelated?			
No	Yes	Maybe	NA
0	1	2	NA

Did you follow your routine today?		
No	Ye	NA
0	1	2

Again, do you think sleep cycle and consistency in life are interrelated too?			
No	Yes	Maybe	NA
0	1	2	NA

Week 5

Did you complete your 21 days challenge successfully?			
No	Yes	Maybe	NA

0	1	2	NA

How many days did you skip?					
0	1	2	3	4	5

Are you going to continue the routine even from tomorrow?		
No	Yes	Maybe
0	1	2

Do you think physical and mental exercise help us being consistent?		
No	Yes	Maybe
0	1	2

What happened when you missed any day?					
Feel Bad	Frustration	Impatience	Longing	Nothing	NA
0	1	2	3	4	NA

After the completion of 21 days challenge do you think you can continue that routine every day?		
No	Yes	Maybe
0	1	2

Did you achieve your 21 days goal?		
No	Yes	Maybe
0	1	2

Did you find this survey useful ??	
No	Yes
0	1

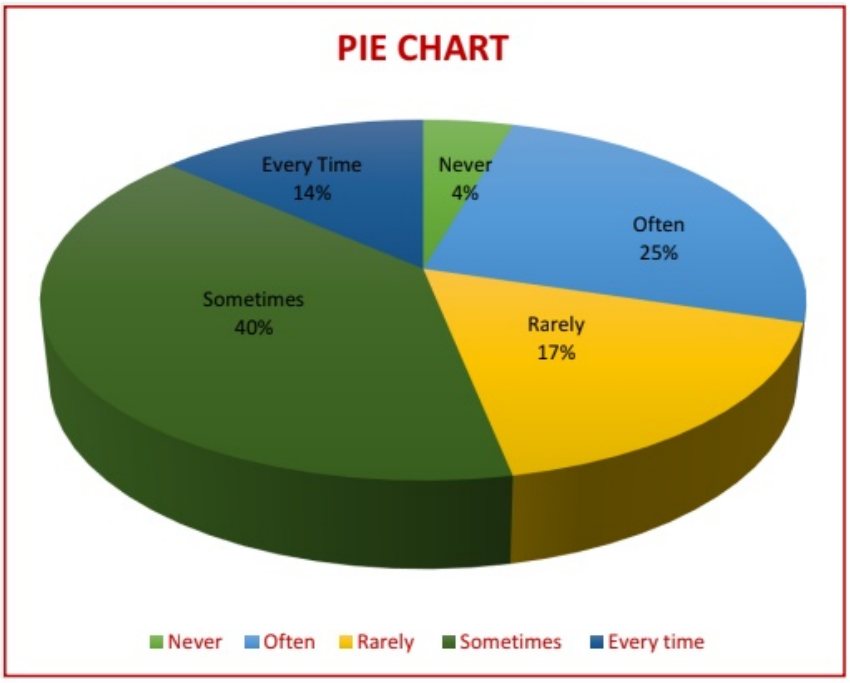
DATA ANALYSIS

1) Diagrammatic Representation

Pie Chart

Q.1) Do you struggle to be consistent?

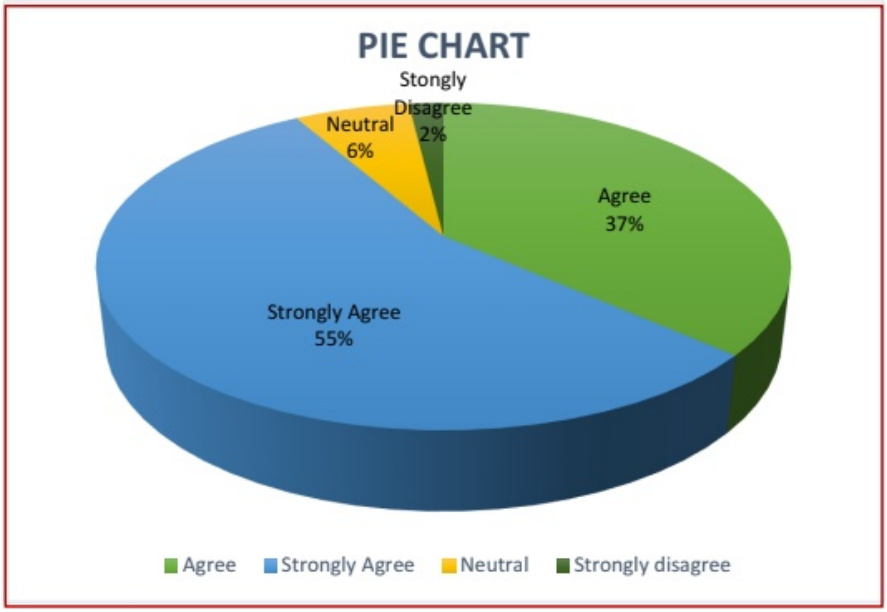
Never	Often	Rarely	Sometimes	Every time
5	28	19	44	15



Conclusion: For this data we conclude that most of the persons struggle sometimes to being consistent.

Q.2) Do you think consistency is the key for everything?

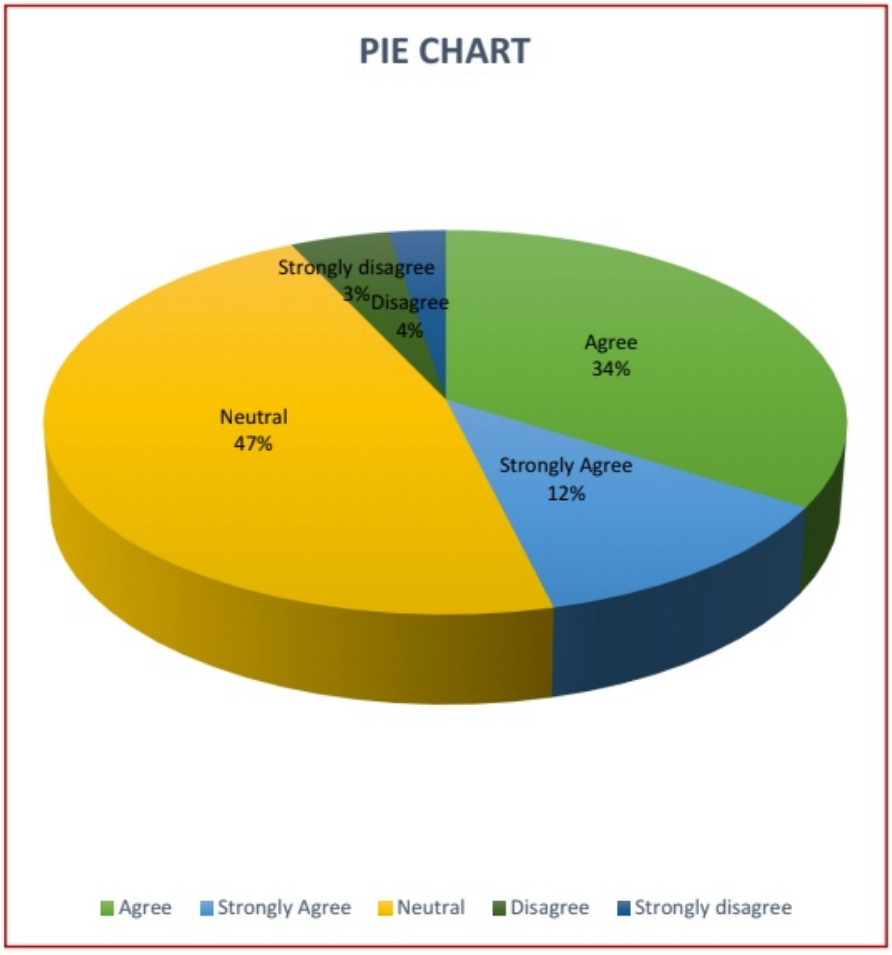
Feel Bad	Frustration	Impatience	Longing	Nothing	Others
57	16	11	19	40	13



Conclusion: For our data, we conclude that most of the peoples are strongly agree to think that consistency is the key for everything.

Q.3) Was your 21 days challenge effective?

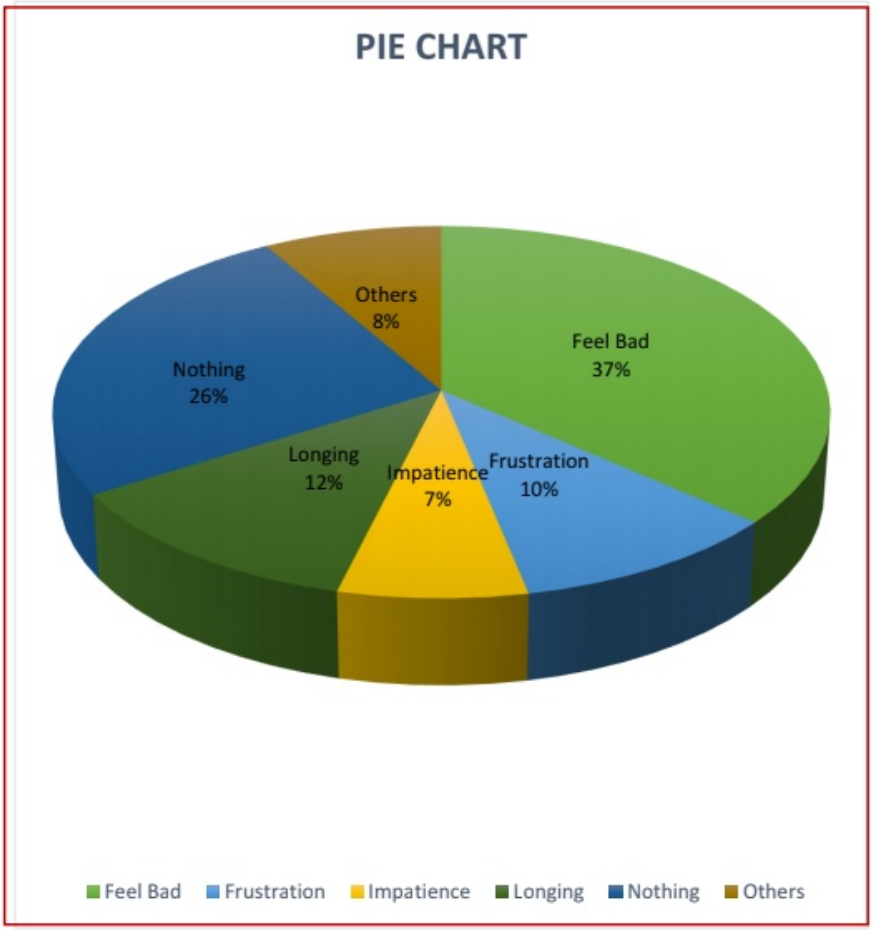
Agree	Strongly Agree	Neutral	Disagree	Strongly disagree
53	19	73	7	4



Conclusion: According to our data, most of the people challenges were not as much as they expected.

Q.4) What happen when you missed any day of your 21 days challenge?

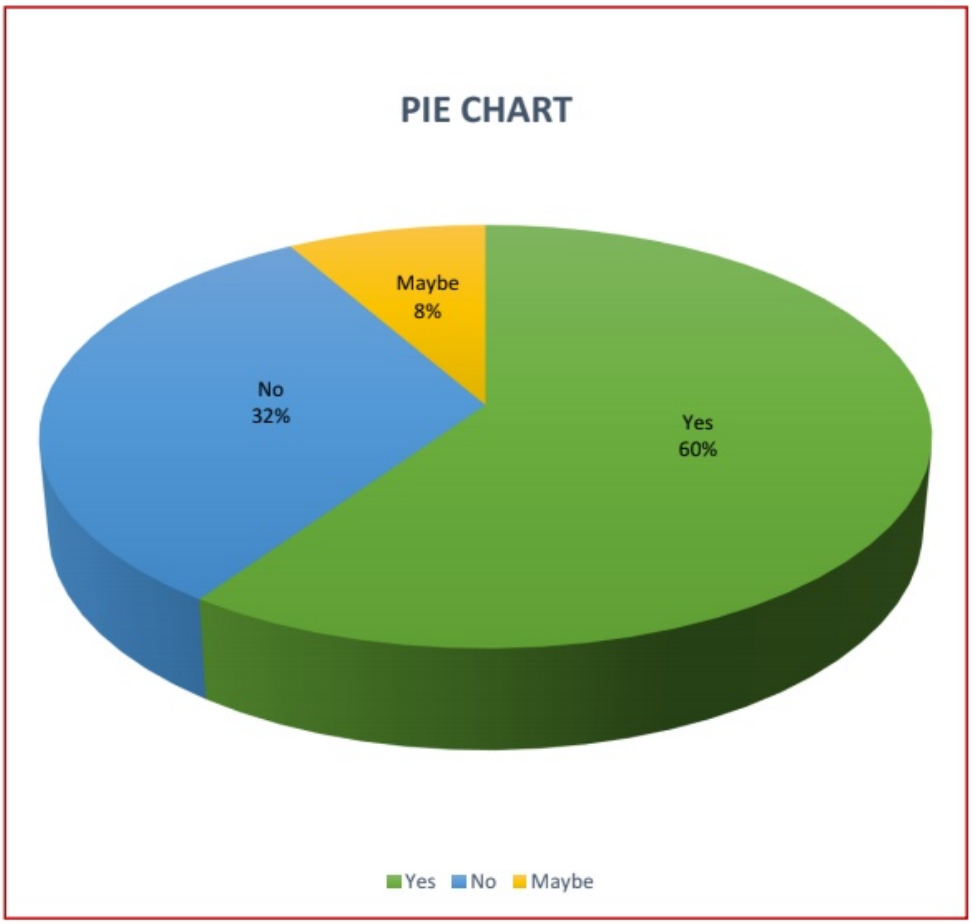
Feel Bad	Frustration	Impatience	Longing	Nothing	Others
57	16	11	19	40	13



Conclusion: According to our data it may conclude that the most of the people feels bad when they missed any data in their 21 days challenge.

Q.5) Do you think you have improved in these 21 days?

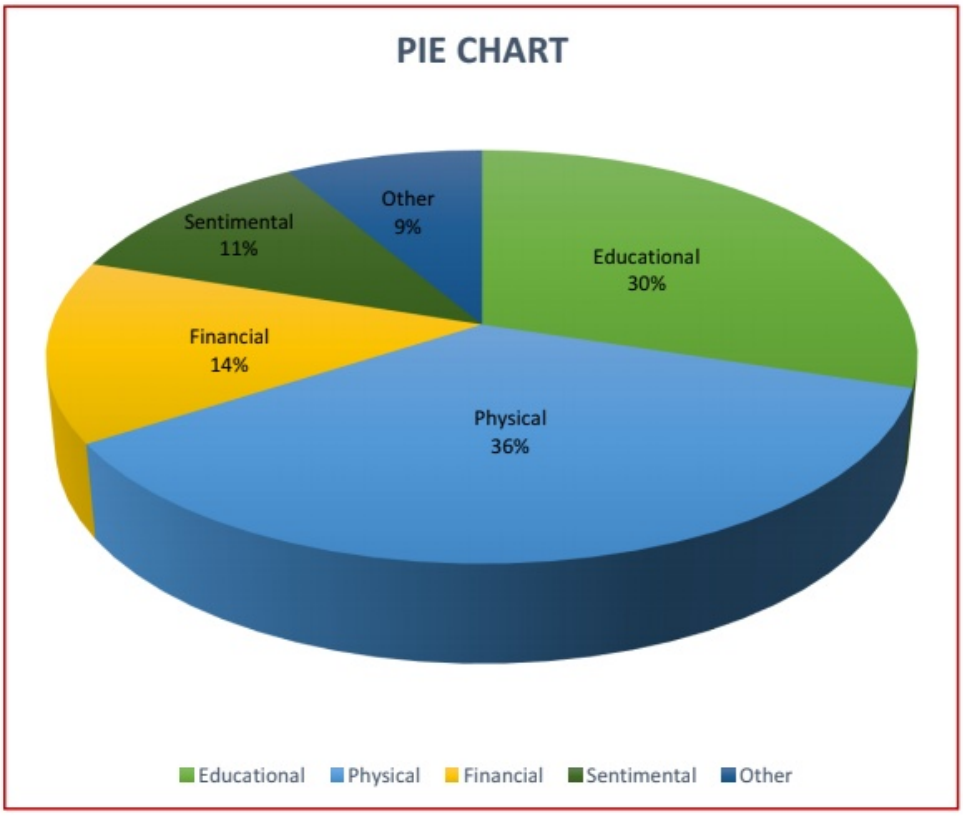
Yes	No	Maybe
93	50	13



Conclusion: From above Pie-Diagram We may conclude that most of the people have seen improvement during 21 days.

Q.6) In which area of life you want to work on?

Educational	Physical	Financial	Sentimental	Other
63	76	30	24	18



Conclusion: From above Pie-Diagram we may conclude that most of the people want to work on physical area of their life.

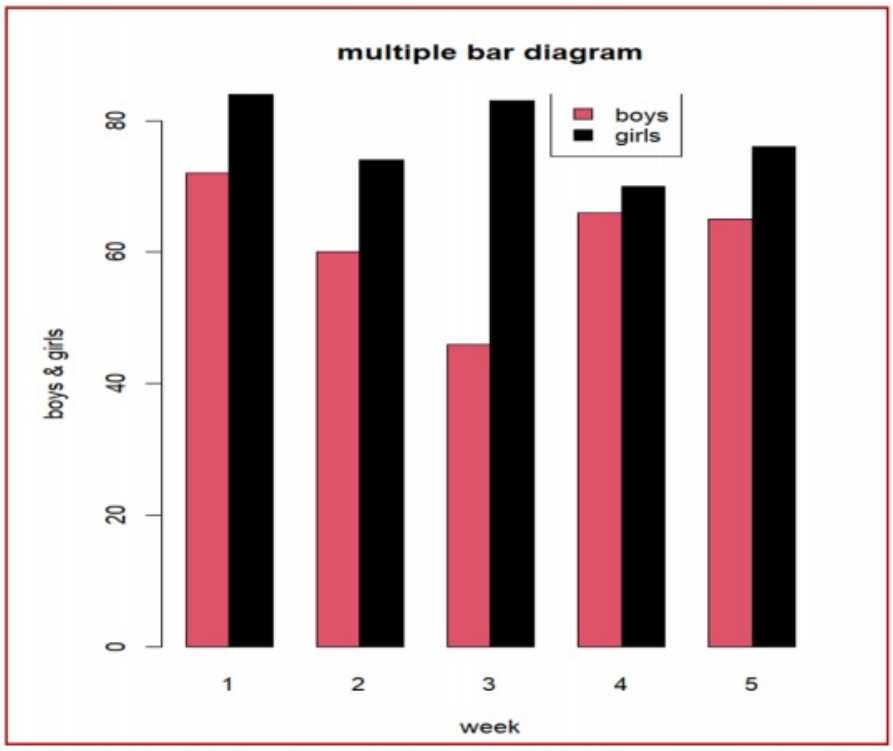
Multiple Bar Diagram

Q.7) Who were responded more in 1 to 5 weeks? Boys or girls?

CODE:

```
week=1:5
boys=c(72,60,46,66,65)
girls=c(84,74,83,70,76)
data=data.frame(boys,girls)
data
d1=as.matrix(data)
d1
 barplot(t(d1),beside=T,xlab="week",ylab="boys & girls",main="multiple bar diagram",col=2:1,names.arg=week)
legend(locator(1),legend=c("boys","girls"),fill=2:1)
```

OUTPUT:



Conclusion: In our data we can observe with the help of multiple bar diagram Girl's responses in every week is more than Boys responses.

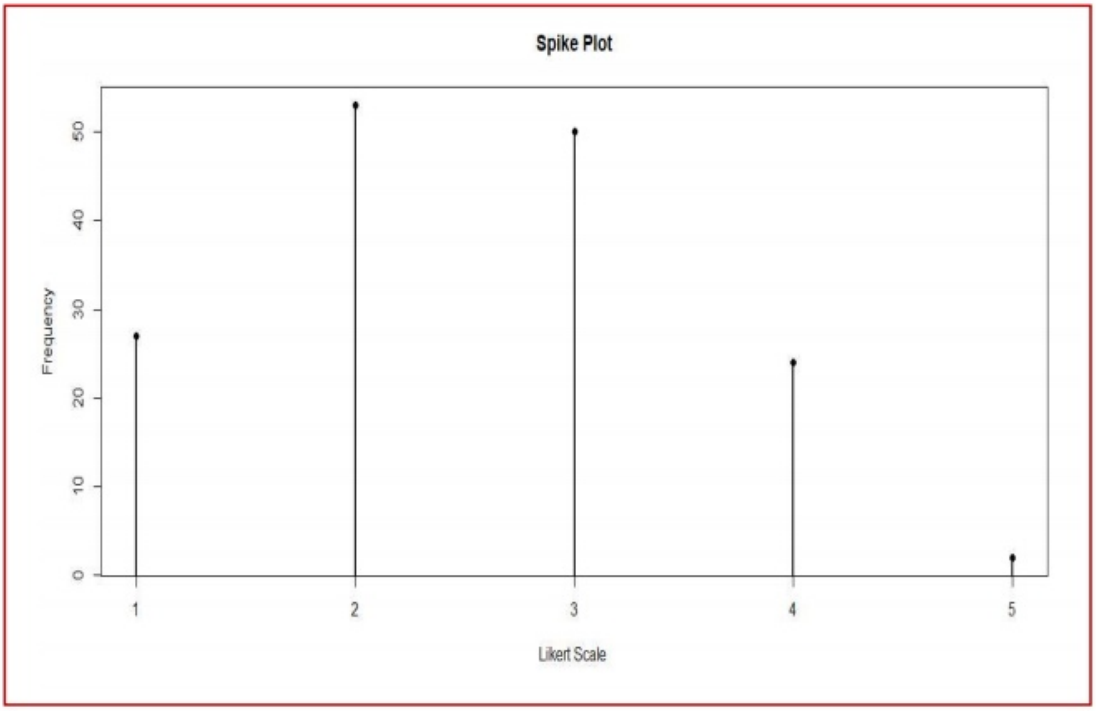
Spike And Rod Plot

Q.8) Does the motivation gained from self-help book, articles, motivational videos and podcasts stay for longer period ?

Agree	Strongly Agree	Neutral	Disagree	Strongly disagree
53	27	50	24	2

CODE:

```
x=1:5

f=c(27,53,50,24,2)

plot(x,f,"h",xlab="Likert Scale",ylab="Frequency",lwd=2,main="Spike Plot")

points(x,f,pch=16)
```



Conclusion: From above spike plot we can conclude that most of the people are agree with the fact that motivation gained from self-help book, articles, motivational videos and podcasts stay for longer period.

2) Testing

Chi-square test

a) For testing Balance diet and consistency

H_0: Balanced diet and consistency are independent of each other.
H_1: Balanced diet and consistency are not independent of each other.

Whether consume balanced diet or not	Consistent	Inconsistent
Yes	64	18
No	47	8

Code

```
> x=c(64,47,18,8)
> y=matrix(x,nrow=2,ncol=2)
> y
     [,1] [,2]
[1,]   64   18
[2,]   47    8
> T=chisq.test(y,correct=F)
> T
```

Pearson's Chi-squared test

data: y
X-squared = 1.1742, df = 1, p-value = 0.2785

Criteria: Reject **H_0** if p-value is less than α , otherwise accept **H_0**.
Decision: Here p-value > 0.05 therefore, we may accept Ho at 5% l.o.s.

Conclusion: Balanced diet and consistency are independent of each other. This means a person's consistency is not assured even if he/she consumes balanced diet regularly.

b) Hours of sleeping and Consistency

H_0: Hours of sleeping and consistency are independent of each other.
H_1: Hours of sleeping and consistency are not independent of each other.

No. of hours of sleeping	Consistent	Inconsistent
5	11	2
6	31	5
7	38	10
8	16	7
More than 8	15	2

Code

```
> x=c(11,2,31,5,38,10,16,7,15,2)
> u=matrix(x,nrow=5,ncol=2,byrow=TRUE)
> u
     [,1] [,2]
[1,]   11    2
[2,]   31    5
[3,]   38   10
[4,]   16    7
[5,]   15    2
> T=chisq.test(u,correct=TRUE)
> T
```

Pearson's Chi-squared test

data: u
X-squared = 3.3616, df = 4, p-value = 0.4992
Criteria: Reject $\mathbf{H_0}$ if p-value is less than α , otherwise accept $\mathbf{H_0}$.
Decision: Here p-value(=0.4992) > 0.05 therefore, we may accept Ho at 5% l.o.s.
Conclusion: Hours of sleeping and consistency are independent of each other. That means a person being consistent does not depend on how many ours he/she sleeps.

c) Reading habit and Consistency

H_0: Reading self-help books and consistency are independent of each other.
H_1: Reading self-help books and consistency are not independent of each other.

Whether has reading habit	Consistent	Inconsistent
Yes	49	18
No	62	6

Code

```
> x=c(49,62,18,6)
> y=matrix(x,nrow=2,ncol=2)
> y
     [,1] [,2]
[1,]   49   18
[2,]   62    6
> T=chisq.test(y,correct=F)
> T
```

Pearson's Chi-squared test

data: y
X-squared = 7.5155, df = 1, p-value = 0.006117

Criteria: Reject **H_0** if p-value is less than α , otherwise accept **H_0**.
Decision: Here p-value < 0.05 therefore, we may reject Ho at 5% l.o.s.

Conclusion: Reading habit and consistency are not independent of each other. This means people who read self help books are more likely to be consistent.

d) Exercise and Consistency

H_0: Exercise and consistency are independent of each other.
H_1: Exercise and consistency are not independent of each other.

Whether exercise daily	Consistent	Inconsistent
Yes	80	22
No	28	11

Code

```
> x=c(80,28,22,11)
> y=matrix(x,nrow=2,ncol=2)
> y
     [,1] [,2]
[1,]   80   22
[2,]   28   11
> T=chisq.test(y,correct=F)
> T

        Pearson's Chi-squared test

data:  y
X-squared = 0.69315, df = 1, p-value = 0.4051
```

Criteria: Reject **H_0** if p-value is less than α , otherwise accept **H_0**.
Decision: Here p-value > 0.05 therefore, we may accept Ho at 5% l.o.s.

Conclusion: Exercising daily and consistency are independent of each other. This means a person who exercise daily may or may not be consistent.

e) Struggle to be consistent and consistency

Ho: Consistency and different levels of struggle to be consistent are independent
H1: Consistency and different levels of struggle to be consistent are not independent

How often ↓	Consistent	Inconsistent
Often	28	10
Rarely	19	5
Sometimes	45	21
Never	4	2
Everytime	15	8

Code on R

```
> x=c(10,28,5,19,21,45,2,4,8,15)
> y=matrix(x,nrow=5,ncol=2,byrow=TRUE)
> y
     [,1] [,2]
[1,]   10   28
[2,]    5   19
[3,]   21   45
[4,]    2    4
[5,]    8   15
> T=chisq.test(y,correct=TRUE)
> T
```

Pearson's Chi-squared test

data: y
X-squared = 1.5768, df = 4, p-value = 0.813

Criteria: Reject H_0 if p-value is less than α , otherwise accept H_0.
Decision: Here p-value > 0.05 therefore, we may accept Ho at 5% l.o.s.

Conclusion: Consistency is independent of struggle to be consistent. This means even if a person struggles to be consistent often, rarely, sometimes, everytime, or never (in general), he may or may not be consistent throughout the routine.

f) Effect of people on us and Consistency

Ho : Effect of people on us and consistency are independent.
H1 : Effect of people on us and consistency are not independent.

	Consistent	Inconsistent
Yes	71	25
No	11	4
Maybe	29	10

Code

```
> x=c(71,11,29,25,4,10)
> y=matrix(x,nrow=3,ncol=2)
> y
     [,1] [,2]
[1,]   71   25
[2,]   11    4
[3,]   29   10
> T=chisq.test(y,correct=TRUE)
> T
```

Pearson's Chi-squared test

data: y
X-squared = 0.0061637, df = 2, p-value = 0.9969

Criteria: Reject H_0 if p-value is less than α , otherwise accept H_0.
Decision: Here p-value > 0.05 therefore, we may accept Ho at 5% l.o.s.

Conclusion: The effect of people around us and consistency are independent of each other. That means People around us has no role to play in consistence.

Proportion Test

Q) Does gender affects a person's consistency?

Gender	Were the YES	People consistent NO	Total
Male	48	12	60
Female	63	34	97
Total	111	46	157

Let,
Y=no. of males who were consistent.
X=no. of females who were consistent.
P1=proportion of males who were consistent.
P2=proportion of females who were consistent.
N1=total no. of males who participated.
N2=total no. of females who participated.
Now,
Here, we want to find is there sufficient evidence at alpha=0.05, that the two populations –males & females-differ significantly with respect to their consistency.
So, we want to test,
H0=P1 is equal to P2
v/s H1=P1 is not equal to P2
using R software we get,

CODE

```
y=48;y
[1] 48
> n1=60;n1
[1] 60
> x=63;x
[1] 63
```

```
> n2=97;n2
[1] 97
> prop.test(x=c(48,63),n=c(60,97))

        2-sample test for equality of proportions with continuity correction

data:  c(48, 63) out of c(60, 97)
X-squared = 3.3601, df = 1, p-value = 0.06679
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.001751509  0.302782436
sample estimates:
   prop 1    prop 2
0.8000000 0.6494845
```

Now, here
Criteria : Reject H0, when p-value is less than or equal to specified value of level of significance ,otherwise accept H0.

Decision : Here, p-value=0.06679
& alpha=0.05
As , p-value > alpha
i.e 0.06679 > 0.05
we may accept H0 at alpha% l.o.s

Conclusion: There is no significant difference between proportion of consistency of male and females. i.e. P1 is equal to P2.
In short, gender does not affect consistency of a person.

3) Data Analytics Algorithm

K-Nearest Neighbour(K-NN)

Q) Did the people who skip the routine and struggled to be consistent achieve their goal?

x2	8
y2	4
k	10

x1	x2	y	pred
6	5	**1**	2.236068
8	1	**2**	9.848858
6	5	**2**	6.403124
7	0	**1**	7
6	0	**1**	6
6	1	**1**	6.082763
6	2	**0**	6.324555
7	1	**0**	7.071068
5	0	**2**	5
7	5	**2**	8.602325
6	2	**1**	6.324555
7	1	**2**	7.071068
7	2	**1**	7.28011
8	0	**1**	8
6	2	**1**	6.324555
7	2	**2**	7.28011
7	5	**1**	8.602325
7	2	**1**	7.28011
6	0	**1**	6
8	3	**1**	8.544004
9	2	**1**	9.219544
7	2	**1**	7.28011
6	4	**1**	7.211103
7	0	**1**	7
6	3	**2**	6.708204
7	3	**2**	7.615773
6	5	**1**	7.81025
5	1	**1**	5.09902
9	2	**2**	9.219544
6	8	**2**	10
7	4	**2**	8.062258

6	5	**1**	7.81025
7	4	**1**	8.062258
8	0	**1**	8
8	3	**2**	8.544004
6	2	**1**	6.324555
8	3	**1**	8.544004
7	4	**1**	8.062258
8	0	**1**	8
6	3	**1**	6.708204
6	0	**1**	6
7	6	**2**	9.219544
7	5	**1**	8.602325
6	2	**1**	6.324555
7	0	**1**	7
7	3	**2**	7.615773
5	2	**1**	5.385165
8	0	**1**	8
6	5	**1**	7.81025
8	3	**1**	8.544004
7	1	**1**	7.071068
8	5	**1**	9.433981
8	0	**1**	8
7	2	**2**	7.28011
6	1	**1**	6.082763
8	0	**1**	8
7	13	**0**	14.76482
6	6	**2**	8.485281
8	3	**1**	8.544004
7	0	**1**	7
8	10	**2**	12.80625
7	1	**1**	7.071068
6	3	**1**	6.708204
8	0	**1**	8
7	10	**2**	12.20656
9	3	**1**	9.486833
7	2	**1**	7.28011
7	5	**1**	8.602325
7	0	**0**	7
6	2	**2**	6.324555
6	0	**1**	6
8	1	**1**	8.062258
7	5	**1**	8.602325
7	2	**0**	7.28011
6	3	**1**	6.708204
8	5	**0**	9.433981
8	1	**2**	8.062258

7	5	**0**	8.602325
6	1	**2**	6.082763
6	0	**2**	6
5	0	**1**	5
4	0	**1**	4
7	5	**2**	8.602325
7	5	**0**	8.602325
6	0	**1**	6
5	1	**0**	5.09902
5	0	**0**	5
8	2	**1**	8.246211
6	0	**1**	6
9	2	**1**	9.219544
7	1	**0**	7.071068
7	3	**0**	7.615773
5	4	**1**	6.403124
7	2	**1**	7.28011
6	1	**1**	6.082763
8	3	**0**	8.544004
6	6	**0**	8.485281
8	0	**1**	8
8	4	**1**	8.944272
9	3	**2**	9.486833
5	5	**0**	7.071068
8	1	**1**	8.062258
5	2	**1**	5.385165
8	1	**0**	8.062258
7	3	**1**	7.615773
8	1	**1**	8.062258
7	1	**0**	7.071068
7	2	**0**	7.28011
6	1	**1**	6.082763
5	3	**1**	5.830952

Conclusion: for our data, a person who maybe struggle to be consistent and skips near about 3 days can achieve their 21 days goal.

Naïve Bayes

```
#Att 1 : Surrounding and consistency are interrelated
#Att2:Diet and consistency are interrelated
#Att3:sleep cycle and consistency are interrelated
#Att4:physical or mental exercise and consistency are interrelated
r=read.csv("C:\\Users\\ABHISHEK\\Desktop\\Devashri Naive bayes.csv")
r
d=data.frame(r)
sam=sample(2,nrow(d),prob = c(.8,.2),replace = TRUE)
sam
data_train=d[sam==1, ]
data_test=d[sam==2, ]
model=naiveBayes(y~.,data_train)
model
pred=predict(model,newdata=data_test)
pred
cm=table(pred,data_test[,5])
cm
acc=(sum(diag(cm))/sum(cm))*100
acc
output;
r=read.csv("C:\\Users\\ABHISHEK\\Desktop\\Devashri Naive bayes.csv")
> r
   x1 x2 x3 x4 y
1   1  1  3  1 2
2   1  1  1  3 3
3   1  1  2  0 1
```

4	3	3	1	1	1
5	3	3	1	1	2
6	2	1	1	1	2
7	2	1	2	1	2
8	1	1	1	1	1
9	1	1	0	1	1
10	1	1	1	3	3
11	2	2	1	1	1
12	1	1	1	1	0
13	3	3	3	1	0
14	2	2	2	1	0
15	1	1	1	1	2
19	1	1	0	1	2
20	2	1	1	1	1
21	1	1	1	1	1
22	1	2	1	1	1
23	2	0	1	3	3
24	1	1	1	1	2
25	2	0	1	0	1
26	3	3	1	3	3
27	1	1	0	1	1
28	2	2	0	1	1
29	0	1	1	0	1
30	2	2	1	0	1
31	1	1	1	1	1
32	1	1	2	1	1
33	2	2	2	1	2

34 3 3 1 0 1

35 0 1 3 1 1

36 0 0 1 1 1

37 2 2 1 1 1

38 3 3 0 0 2

39 1 1 1 0 1

40 0 0 1 0 2

41 2 1 1 1 2

42 3 3 2 1 1

43 1 1 1 0 1

44 0 2 2 1 1

45 1 1 1 0 1

46 0 1 1 1 2

47 1 1 1 1 2

48 2 2 1 1 2

49 1 1 1 0 1

50 3 3 2 1 2

51 0 0 1 1 1

52 2 1 1 1 1

53 0 1 1 3 3

54 2 2 1 1 1

55 1 2 1 1 2

56 3 3 2 0 1

57 0 0 1 1 1

58 2 1 2 1 1

59 1 2 3 0 2

60 2 1 1 1 1

61 1 1 2 1 1

62 3 3 1 0 2

63 0 0 2 1 1

64 2 2 0 1 1

65 1 1 1 1 1

66 1 1 2 1 2

67 1 1 2 0 1

68 3 3 1 1 2

69 1 1 0 1 1

70 1 1 1 1 1

71 1 1 2 0 2

72 0 2 1 0 1

73 0 2 0 1 1

74 1 1 1 2 1

75 3 3 1 1 2

76 0 0 1 0 1

77 1 0 2 1 1

78 2 1 2 0 1

79 1 1 0 3 3

80 0 2 1 1 1

81 2 0 1 1 2

82 1 1 1 1 1

83 3 3 0 0 1

84 0 1 1 3 3

85 2 0 1 1 1

86 2 2 2 1 1

87 1 0 1 1 0

88 1 1 2 1 2

89 1 2 2 1 1

90 1 1 1 1 1

91 3 3 2 1 1

92 0 0 1 1 2

93 2 1 1 0 1

94 2 2 1 0 1

95 0 0 1 1 1

96 1 1 2 1 2

97 1 1 1 1 1

98 2 1 2 0 1

99 0 0 1 1 1

100 3 3 3 0 2

101 0 2 1 1 0

102 1 1 0 1 2

103 1 2 1 1 2

104 0 0 3 0 0

105 2 2 0 0 1

106 1 1 1 0 1

107 0 0 2 1 1

108 2 1 1 1 0

109 3 3 1 3 3

110 1 1 1 1 1

111 2 0 1 1 0

112 1 1 2 0 2

113 0 0 1 2 0

114 0 2 1 1 2

115 3 3 1 0 2

116 0 0 2 3 3

117 2 2 1 1 2

118 2 2 0 1 1

119 0 0 1 1 1

120 2 0 0 0 2

121 1 1 1 1 0

122 0 2 2 1 1

123 1 1 1 0 0

124 2 0 1 1 0

125 3 3 2 1 1

126 2 0 1 1 1

127 1 1 0 1 1

128 1 1 1 1 1

129 0 2 1 1 1

130 1 1 1 1 0

131 3 3 2 0 1

132 0 1 2 0 0

133 0 0 1 0 1

134 2 2 1 1 1

135 0 0 2 0 1

136 2 1 1 3 3

137 2 2 1 0 0

138 1 1 1 1 0

139 0 2 1 1 1

140 1 1 1 1 2

141 0 0 2 0 1

142 1 1 1 3 3

143 0 0 1 1 2

144 2 2 1 1 0

145 0 1 2 1 1

146 0 1 1 1 1

147 2 0 2 3 3

148 2 2 2 0 0

149 1 1 1 1 1

150 1 1 2 1 1

151 1 0 2 1 0

```
152  2  1  2  1 0
153  3  3  1  1 1
154  2  1  1  3 3
155  0  1  2  1 1
156  1  2  1  0 1
> d=data.frame(r)
> sam=sample(2,nrow(d),prob = c(.8,.2),replace = TRUE)
> sam
  [1] 1 1 1 1 1 2 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 2 1 2 1 1 2 1 1 1 1 1 1
1 1 2 1 1 1 1 1 1 1 1 2 1 1 1
 [56] 1 1 1 1 1 2 1 1 1 1 1 1 2 1 2 2 1 1 2 1 1 2 1 2 1 2 1 1 1 1 2 2 1 1 2 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 2 1 2 2
[111] 1 2 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1
> data_train=d[sam==1, ]
> data_test=d[sam==2, ]
> model=naiveBayes(y~.,data_train)
> model
Naive Bayes Classifier for Discrete Predictors


Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)


A-priori probabilities:
Y
      0       1       2       3
0.1484375 0.5234375 0.2421875 0.0859375
Conditional probabilities:
```

x1

Y	[,1]	[,2]
0	1.315789	0.8852264
1	1.179104	0.9988686
2	1.483871	1.0286226
3	1.272727	1.0090500

x2

Y	[,1]	[,2]
0	1.1052632	0.8752610
1	1.2835821	0.9179353
2	1.5161290	1.0286226
3	0.9090909	0.8312094

x3

Y	[,1]	[,2]
0	1.473684	0.6966923
1	1.194030	0.6568706
2	1.322581	0.8712863
3	1.181818	0.4045199

x4

Y	[,1]	[,2]
0	0.7894737	0.5353034
1	0.6716418	0.4731602
2	0.7741935	0.4250237

```
  3 3.0000000 0.0000000
> pred=predict(model,newdata=data_test)
> pred
 [1] 3 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 3
Levels: 0 1 2 3
> cm=table(pred,data_test[,5])
> cm


pred  0  1  2  3
   0  0  0  0  0
   1  3 15  3  0
   2  0  2  1  0
   3  0  0  0  3
> acc=(sum(diag(cm))/sum(cm))*100
> acc
[1] 70.37037
```

Conclusion: From Naïve Bayes it can be conclude that the accuracy of completing 21 days challenge is 70.37037

Logistic Regression

Logistic regression using R-software

The response variable Y is whether the person completed his/her 21 days challenge. There are 2 regressors. The regressor x1 is how many hours he/she sleeps daily. Second regressor x2 is how many days he/she failed to follow their routine. Fit univariate and multivariate logistic model. Provide an interpretation of parameter b1 in this model.

y=c(1,1,1,1,1,1,0,1,1,1,1,0,1,1,1,0,1,1,1,1,1,1,1,1,0,1,0,1,0,1,0,1,1,1,0,1,1,0,1,1,
1,0,1,1,1,1,1,1,1,0,1,1,1,1,1,1,0,0,1,0,0,1,1,0,0,1,1,1,1,1,1,1,0,0,0,0,1,0,1,0,1,0,0
,1,0,1,1,0,1,1,0,0,1,1,0,0,1,1,1,0,0,1,1,0,0,1,1,0,1,1)

x1=c(6,8,6,7,6,6,6,7,5,7,6,7,7,8,6,7,7,7,6,8,9,7,6,7,6,7,6,5,9,6,7,6,7,8,8,6,8,7,8,6
,6,7,7,6,7,7,5,8,6,8,7,8,8,7,6,8,7,6,8,7,8,7,6,8,7,9,7,7,7,6,6,8,7,7,6,8,8,7,6,6,5,4,
7,7,6,5,5,8,6,9,7,7,5,7,6,8,6,8,8,9,5,8,5,8,7,8,7,7,6,5)

x2=c(5,1,5,0,0,1,2,1,0,5,2,1,2,0,2,2,5,2,0,3,2,2,4,0,3,3,5,1,2,8,4,5,4,0,3,2,3,4,0,3
,0,6,5,2,0,3,2,0,5,3,1,5,0,2,3,0,13,6,3,0,10,1,3,0,10,3,2,5,0,2,0,1,5,2,3,5,1,5,1,0,
0,0,5,5,0,1,0,2,0,2,1,3,4,2,1,3,6,0,4,3,5,1,2,1,3,1,1,2,1,3)

#if we fit univariate logistic regression model with x1 as single variable

Code to fit logistic model:

```
logreg1=glm(y~x1,family=binomial)

summary(logreg1)
```

Output:

```
Call:
glm(formula = y ~ x1, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0720     1.3688   1.514    0.130
x1           -0.1969     0.1961  -1.004    0.315

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 139.09  on 109  degrees of freedom
Residual deviance: 138.07  on 108  degrees of freedom
AIC: 142.07
```

Number of Fisher Scoring iterations: 4

#if we fit univariate logistic regression model with x2 as single variable

Code to fit logistic model:

```
logreg1=glm(y~x2,family=binomial)

summary(logreg1)
```

Output:

```
Call:

glm(formula = y ~ x2, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.38073    0.33477   4.124 3.72e-05 ***
x2          -0.24762    0.09501  -2.606  0.00916 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 139.09  on 109  degrees of freedom
Residual deviance: 131.34  on 108  degrees of freedom
AIC: 135.34

Number of Fisher Scoring iterations: 4
```

if we fit multiple regression model :

Code to fit logistic model:

```
logreg1=glm(y~(x1+x2),family=binomial)

summary(logreg1)
```

Output:

```
Call:
glm(formula = y ~ (x1 + x2), family = binomial)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.8239	1.4811	1.907	0.05657	.
x1	-0.2089	0.2065	-1.012	0.31164	
x2	-0.2505	0.0964	-2.598	0.00937	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 139.09 on 109 degrees of freedom
Residual deviance: 130.30 on 107 degrees of freedom
AIC: 136.3

Number of Fisher Scoring iterations: 4

Conclusion:

Comparison of Two model's accuracy

Decision Tree

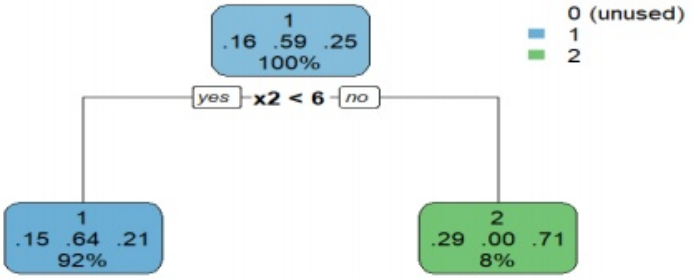
x1= How many hour do you sleep daily?
x2= How many days did you skip?
Y= Did you achieve your 21 Days challenge?
Calculate the accuracy for the following data

```
y=c(1,2,2,1,1,1,0,0,2,2,1,2,1,1,1,2,1,1,1,1,1,1,1,1,2,2,1,1,2,2,2,1,1,1,2,1,1,1,1,1,
1,2,1,1,1,2,1,1,1,1,1,1,1,2,1,1,0,2,1,1,2,1,1,1,2,1,1,1,0,2,1,1,1,0,1,0,2,0,2,2,1,1,2
,0,1,0,0,1,1,1,0,0,1,1,1,0,0,1,1,2,0,1,1,0,1,1,0,0,2,1)

x1=c(6,8,6,7,6,6,6,7,5,7,6,7,7,8,6,7,7,7,6,8,9,7,6,7,6,7,6,5,9,6,7,6,7,8,8,6,8,7,8,6
,6,7,7,6,7,7,5,8,6,8,7,8,8,7,6,8,7,6,8,7,8,7,6,8,7,9,7,7,7,6,6,8,7,7,6,8,8,7,6,6,5,4,
7,7,6,5,5,8,6,9,7,7,5,7,6,8,6,8,8,9,5,8,5,8,7,8,7,7,6,5)

x2=c(5,1,5,0,0,1,2,1,0,5,2,1,2,0,2,2,5,2,0,3,2,2,4,0,3,3,5,1,2,8,4,5,4,0,3,2,3,4,0,3
,0,6,5,2,0,3,2,0,5,3,1,5,0,2,3,0,13,6,3,0,10,1,3,0,10,3,2,5,0,2,0,1,5,2,3,5,1,5,1,0,
0,0,5,5,0,1,0,2,0,2,1,3,4,2,1,3,6,0,4,3,5,1,2,1,3,1,1,2,1,3)

d=data.frame(x1,x2,y)

sam=sample(2,nrow(d),prob=c(0.8,0.2),replace = TRUE)

sam

data_train=d[sam==1, ]

data_test=d[sam==2, ]

model_dt= rpart(y~.,data_train,method='class')

model_dt

rpart.plot(model_dt)

pred_dt = predict(model_dt,data_test,type="class")

pred_dt

cm=table(pred_dt,data_test[,3]);cm

accuracy=(sum(diag(cm))/sum(cm))*100

accuracy
```

OUTPUT:

```
y=c(1,2,2,1,1,1,0,0,2,2,1,2,1,1,1,2,1,1,1,1,1,1,1,1,2,2,1,1,2,2,2,1,1,1,2,1,1,1,1,1,
1,2,1,1,1,2,1,1,1,1,1,1,1,2,1,1,0,2,1,1,2,1,1,1,2,1,1,1,0,2,1,1,1,0,1,0,2,0,2,2,1,1,2
,0,1,0,0,1,1,1,0,0,1,1,1,0,0,1,1,2,0,1,1,0,1,1,0,0,2,1)
> x1=c(6,8,6,7,6,6,6,7,5,7,6,7,7,8,6,7,7,7,6,8,9,7,6,7,6,7,6,5,9,6,7,6,7,8,8,6,8,7,8
,6,6,7,7,6,7,7,5,8,6,8,7,8,8,7,6,8,7,6,8,7,8,7,6,8,7,9,7,7,7,6,6,8,7,7,6,8,8,7,6,6,5,
4,7,7,6,5,5,8,6,9,7,7,5,7,6,8,6,8,8,9,5,8,5,8,7,8,7,7,6,5)
> x2=c(5,1,5,0,0,1,2,1,0,5,2,1,2,0,2,2,5,2,0,3,2,2,4,0,3,3,5,1,2,8,4,5,4,0,3,2,3,4,0
,3,0,6,5,2,0,3,2,0,5,3,1,5,0,2,3,0,13,6,3,0,10,1,3,0,10,3,2,5,0,2,0,1,5,2,3,5,1,5,1,
0,0,0,5,5,0,1,0,2,0,2,1,3,4,2,1,3,6,0,4,3,5,1,2,1,3,1,1,2,1,3)
> d=data.frame(x1,x2,y)
> sam=sample(2,nrow(d),prob=c(0.8,0.2),replace = TRUE)
> sam
  [1] 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1
2 1 2 1 1 1 1 1 2 1 2 1 1 1 2
 [56] 2 1 1 2 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 2 2 2 1 1 2 1 2 1
1 1 2 1 1 1 2 1 1 1 1 2 1 1 1
> data_train=d[sam==1, ]
> data_test=d[sam==2, ]
> model_dt= rpart(y~.,data_train,method='class')
> model_dt
n= 88

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 88 36 1 (0.1590909 0.5909091 0.2500000)
  2) x2< 5.5 81 29 1 (0.1481481 0.6419753 0.2098765) *
  3) x2>=5.5 7  2 2 (0.2857143 0.0000000 0.7142857) *
> rpart.plot(model_dt)
> pred_dt = predict(model_dt,data_test,type="class")
> pred_dt
  2  18  29  41  43  49  51  55  56  59  67  68  83  86  87  88  89  92  94  98 102 1
07
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
Levels: 0 1 2
> cm=table(pred_dt,data_test[,3]);cm

pred_dt  0  1  2
      0  0  0  0
      1  4 15  3
      2  0  0  0
```

```
> accuracy=(sum(diag(cm))/sum(cm))*100
> accuracy
[1] 68.18182
```



Conclusion: For our data it can be conclude that the chance of achieving goal is 68.18182. i.e. Accuracy is 68.18182.

Naïve Bayes

Y = Did you achieve your 21 days goal ?

X1 = How many hours do you sleep daily ?

X2 =How many days did you skip ?

R commands :

```
y=c(1,2,2,1,1,1,0,0,2,2,1,2,1,1,1,2,1,1,1,1,1,1,1,1,2,2,1,1,2,2,2,1,1,1,2,1,1,1,1,1,
1,2,1,1,1,2,1,1,1,1,1,1,1,2,1,1,0,2,1,1,2,1,1,1,2,1,1,1,0,2,1,1,1,0,1,0,2,0,2,2,1,1,2
,0,1,0,0,1,1,1,0,0,1,1,1,0,0,1,1,2,0,1,1,0,1,1,0,0,2,1)
```

```
x1=c(6,8,6,7,6,6,6,7,5,7,6,7,7,8,6,7,7,7,6,8,9,7,6,7,6,7,6,5,9,6,7,6,7,8,8,6,8,7,8,6
,6,7,7,6,7,7,5,8,6,8,7,8,8,7,6,8,7,6,8,7,8,7,6,8,7,9,7,7,7,6,6,8,7,7,6,8,8,7,6,6,5,4,
7,7,6,5,5,8,6,9,7,7,5,7,6,8,6,8,8,9,5,8,5,8,7,8,7,7,6,5)

x2=c(5,1,5,0,0,1,2,1,0,5,2,1,2,0,2,2,5,2,0,3,2,2,4,0,3,3,5,1,2,8,4,5,4,0,3,2,3,4,0,3
,0,6,5,2,0,3,2,0,5,3,1,5,0,2,3,0,13,6,3,0,10,1,3,0,10,3,2,5,0,2,0,1,5,2,3,5,1,5,1,0,
0,0,5,5,0,1,0,2,0,2,1,3,4,2,1,3,6,0,4,3,5,1,2,1,3,1,1,2,1,3)

data=data.frame(x1,x2,y)

sam=sample(2,nrow(data),prob=c(.6,.4),replace=TRUE)

sam

library(dplyr)

length(sam)

data_train=data[sam==1,]

View(data_train)

data_test=data[sam==2,]

View(data_test)

model_nb=naiveBayes(y~.,data=data_train)

pred_nb=predict(model_nb,newdata=data_test)

pred_nb

confusion_matrix=table(pred_nb,data_test[,3])

confusion_matrix

accuracy=(sum(diag(confusion_matrix))/sum(confusion_matrix))*100

accuracy
```

output:

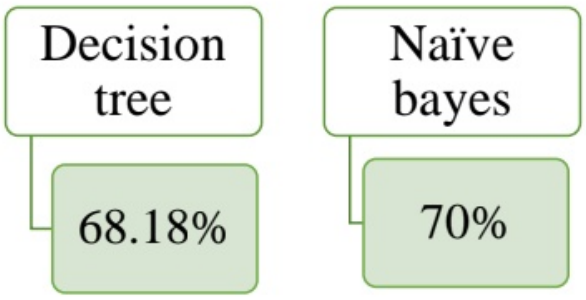
```
y=c(1,2,2,1,1,1,0,0,2,2,1,2,1,1,1,2,1,1,1,1,1,1,1,1,2,2,1,1,2,2,2,1,1,1,2,1,1,1,1,1,
1,2,1,1,1,2,1,1,1,1,1,1,1,2,1,1,0,2,1,1,2,1,1,1,2,1,1,1,0,2,1,1,1,0,1,0,2,0,2,2,1,1,2
,0,1,0,0,1,1,1,0,0,1,1,1,0,0,1,1,2,0,1,1,0,1,1,0,0,2,1)

>
x1=c(6,8,6,7,6,6,6,7,5,7,6,7,7,8,6,7,7,7,6,8,9,7,6,7,6,7,6,5,9,6,7,6,7,8,8,6,8,7,8,6
```

```
,6,7,7,6,7,7,5,8,6,8,7,8,8,7,6,8,7,6,8,7,8,7,6,8,7,9,7,7,7,6,6,8,7,7,6,8,8,7,6,6,5,4,
7,7,6,5,5,8,6,9,7,7,5,7,6,8,6,8,8,9,5,8,5,8,7,8,7,7,6,5)

>
x2=c(5,1,5,0,0,1,2,1,0,5,2,1,2,0,2,2,5,2,0,3,2,2,4,0,3,3,5,1,2,8,4,5,4,0,3,2,3,4,0,3
,0,6,5,2,0,3,2,0,5,3,1,5,0,2,3,0,13,6,3,0,10,1,3,0,10,3,2,5,0,2,0,1,5,2,3,5,1,5,1,0,
0,0,5,5,0,1,0,2,0,2,1,3,4,2,1,3,6,0,4,3,5,1,2,1,3,1,1,2,1,3)

> data=data.frame(x1,x2,y)

> sam=sample(2,nrow(data),prob=c(.6,.4),replace=TRUE)

> sam

  [1] 2 2 1 1 1 2 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 1 2 2 1 2 1 2 2 1 1 1 1 1 2 2 2 1 2 1
2 1 1

 [44] 2 2 1 1 2 1 1 1 1 1 1 2 1 2 1 1 2 2 1 1 1 1 2 2 1 1 1 2 1 2 2 1 2 1 2 2 1 1 1 1
1 2 1

 [87] 1 1 2 1 1 2 1 2 2 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1

> library(dplyr)

> length(sam)

[1] 110

> data_train=data[sam==1,]

> View(data_train)

> data_test=data[sam==2,]

> View(data_test)

> model_nb=naiveBayes(y~.,data=data_train)

> pred_nb=predict(model_nb,newdata=data_test)

> pred_nb

 [1] 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Levels: 0 1 2

> confusion_matrix=table(pred_nb,data_test[,3])

> confusion_matrix
```

```
pred_nb  0  1  2
      0  0  0  0
      1  4 27  6
      2  1  1  1
> accuracy=(sum(diag(confusion_matrix))/sum(confusion_matrix))*100
> accuracy
[1] 70
>
```

Conclusion: According to our data, we conclude that the accuracy for above fitted model of Naïve Bayes is 70%.

Comparison between two models accuracy:



Conclusion: As the accuracy of Naive bayes is more than Decision tree, hence we may conclude that the best fitted model for the given variable is Naive Bayes.

CONCLUSION

- According to the data, we may conclude that before following 21 days challenge most of the people around 47% there challenges were not effective as much as they expected. But after following the 21 days challenge around 60% people have seen improvement.
- By both the models of classifier (Naïve Bayes and Decision Tree) it can be observed that there are 70% and 68% chances respectively of the peoples for achieving there 21 days goal with the attributes of daily sleeping hours and skipping days in challenge.
- We may conclude that surrounding, Diet, sleep cycle, physical or mental health, reading habits have some impact to achieve a goal. In short these attributes contribute more likely to form habit.
- We can conclude that As consistency does not depend upon any of the major attributes hence a person can become consistent only with his will power.

LIMITATIONS

- Unexpectedly our data is mostly dichotomous/binary in nature, so we can not fit models easily.
- IF we collect data from different cities or from different areas then accuracy can be increases.
- Since we collected the data from different persons so, according to there responses data is heterogenous so it may be unbiased in nature.
- As all the participants haven't filled the weekly Google forms hence there was loss of data.
- As the data collected is mostly from the students ie people of the age 18-24, hence the data can be biased.
- The data has been collected through Google forms, this means only people with smart phones were been able to access the Google forms and gave us the data. Hence the data is biased again.