Name - Neeraj Appa.i
F-129

Q11] Explain the property and f distribution using R software

→ The fisher's F distribution: The F distribution is a distribution of all possible value of F statistic. It is a continuous random variable. Consider two random variable Y and Z such that nY ~ Chi square n and mZ ~ chi square m where n and m are degrees of freedom. The random variable X is defined by $X = \frac{Y/n}{Z/m}$

Properties of F distribution

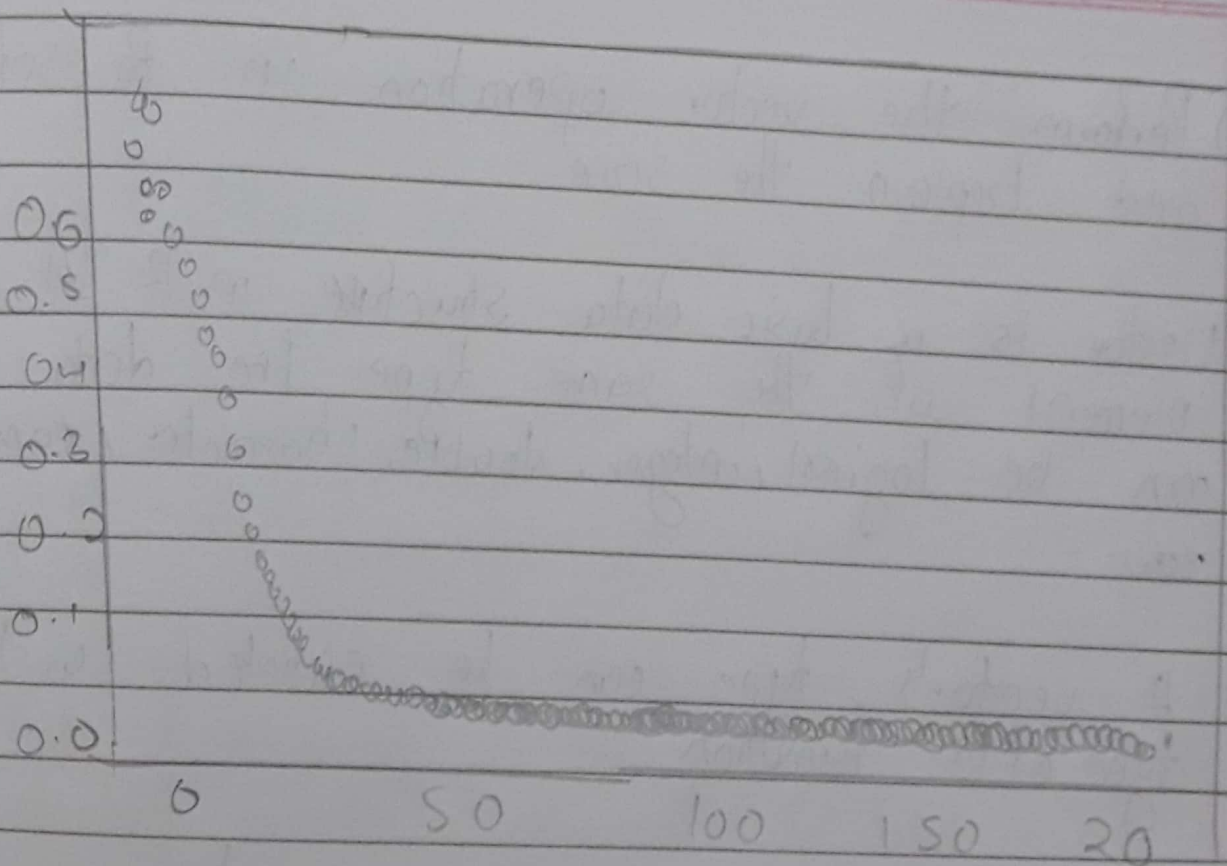
1) The F distribution is positively skewed and with increase in degrees of freedom n and m its skewness decreases
2) The value of F distribution is always positive or zero since the variance are the square of the deviation and hence cannot be negative
3) It is used to calculate the value of mean and variance
4) The shape of F distribution depends on its parameter and n and m degrees of freedom

Code

```
> x_df <- seq(0, 20, by = 0.1)
> y_df <- df(x_df, df1 = 3, df2 = 5)
> plot(y_df)
```

F-12a

Page No.
Date: ALMRAO



Name - Neeraj Appari
Roll no - 120

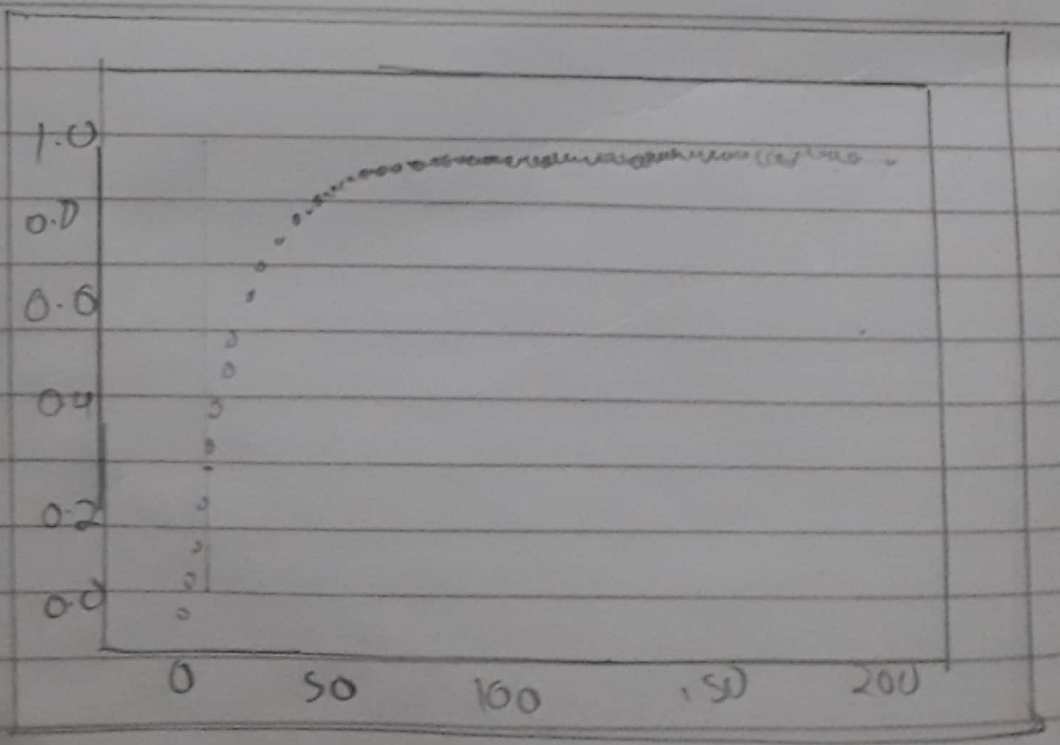
A.L.N.RAO

Q42) Generate the plot for f distribution also mention the application of f distribution

→ A probability density function that is used especially in analysis of variance and is a function of the ratio of two independant random variables each of which has a chi square distribution and is divided by its number of degress of freedom

Code

```
x-pf <- seq(0,20,by=0.1)
y-pf <-pf(x-pf, df1=3, df2=5)
plot(y-pf)
```



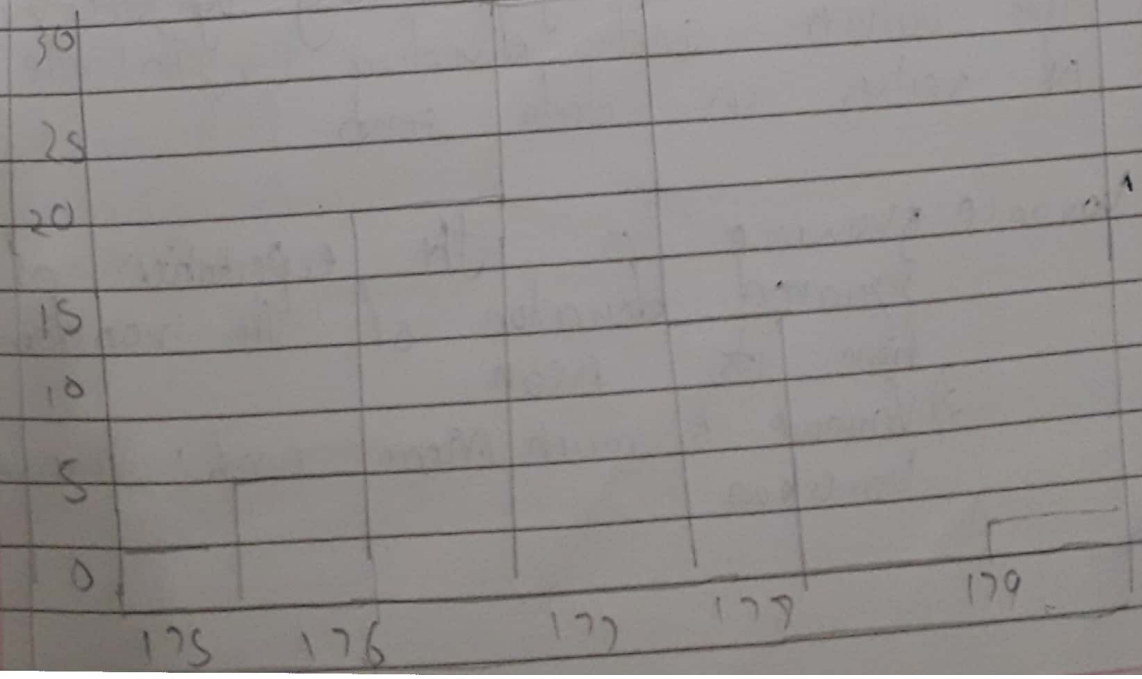
Neeraj Appani
F-129

Page No.
Date ACNIRAO

(Q3) Calculate the f software in ANOVA using R Software and explain process

Code

```
> deep.Data: read.csv(file.choose())
>summary (deep.data)
>hist (deep.Data $yield)
>deep.oneway <- aov (yield ~ fertilizer, data= deep.Data
>summary (deep.oneway)
>deep.twoway <- aov (yield ~ fertilizer + density,
                      data = deep.Data)
> summary (deep.twoway)
>deep.twowayinteraction <- aov (yield ~ fertilizer * density,
                                 data = deep.Data)
> summary (deep.twowayinteraction)
> deep.twowayblocking <- aov (yield ~ fertilizer + density +
      block, data = deep.Data)
>summary (deep.twoway)
```



Name - [illegible]
F-129

Date: [illegible]

Q4) Generate the RUN chart in R software to understand the RUN in data set)

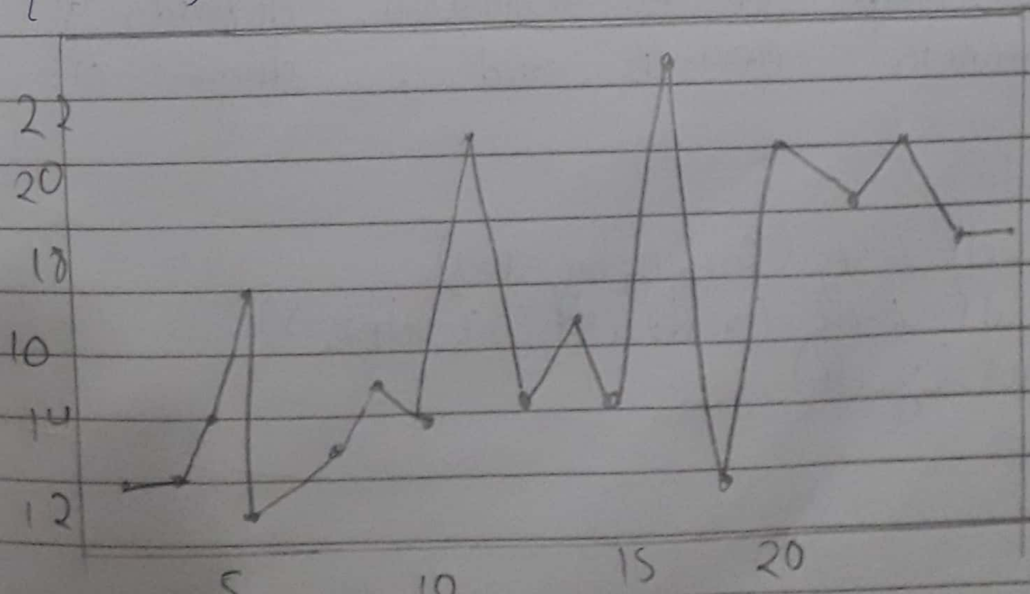
→ 1) A run chart is a line graph of data plotted over time. By collecting and charting data over time, you can find find trends or pattern in process. Because they do not use

2) control limits, run charts cannot tell you if a process is stable. However they can show you how the process is running.

3) The run chart is can be valuable tool at the beginning of the project, as it reveals important information about a process before you have collected enough data.

Code:

```
library(qicharts)
set.seed(9)
y <- rpois(24,16
qic(y)
y[13:24] <- rpois(12,24)
qic(y)
```



Name - Neeraj Appari
I-120

Page No.
Date 21/11/20

Q4) For a data set generate the steam and leaf plot and explain the same

1) A stem and leaf plot is a special table where each data value is split into a 'stem' the first digit and a 'leaf' usually the last digits of digits.
2) For example "32" is split into '3' (stem) and '2' leaf. The stem values are listed down and the leaf values are listed next to them
3) This way the stem groups the data and each leaf indicates a score within that group
4) Unlike histogram stem and leaf display retain the original ~~temperature~~ data atleast two significant data

Code:

```
grades <- c(87, 53, 74, 82, 76, 93, 68, 47, 88, 85, 71, 79,
            90, 65, 93, 86, 87, 63, 24)
> stem(grades)
```

The decimal point is 1 digit(s) to the right of

```
4 | 7
5 | 3
6 | 358
7 | 1469
8 | 1025677
9 | 1335
```

Name - Neeraj Appori
f-129

Date A.L.N.RAO

stem (grouds (scale = 2)

The decimal point is 1 digit (s) to right of the |

4 | 7
5 | 3
5 |
6 | 3
6 | 58
7 | 44
7 | 69
8 | 02

Scanned with CamScanner

Q16) Explain the post Hoc analysis in R software

1) Post-hoc means to analyse the results of your experimental data
2) Without all data in an experiment post hoc analysis is not possible.
3) Post-hoc analysis is performed to analyse the difference between data of groups
4) Post hoc is the tests attempt to control the experiment wise error rate.

Code

```
drug testing <- data.frame(Test = c(5.40, 5.50, 5.55,
                                    5.85, 5.70, 5.75,
                                    5.20, 5.60, 5.50
                                    5.55, 5.50, 5.40
                                    5.90, 5.85, 5.70
                                    5.40, 5.40, 5.35)
                                    5.45, 5.50, 5.35)
                                    5.25, 5.20, 5.10
                                    5.65, 5.55, 5.45)
                                    5.05, 5.00, 4.95)
                                    5.50, 5.50, 5.40
                                    5.45, 5.55, 5.35)
                                    6.30, 6.30 6.35))
sc1 = factor(rep(c("Drug1", "Drug2", "Drug3"),
with, factor(rep(1:20, rep(3,20)))
with(drug testing, boxplot(Test ~ sc1))
friedman.test with post.hoc(Test ~ sc1)
```

Output

Drug 2 - Drug 1 - 0.79300209

Drug 3 - Drug 1 - 0.01602425

Drug 3 - Drug 2 - 0.08080336

F-129

Page No.
Date FL-N R17U

Q47) Explain pairwise mean difference test in R software?

1) What is Pairwise mean difference test.
2) Pairwise mean difference test is also called as the paired t test
3) The paired t-test is used to compare the means between two related group of samples
4) In this case, you have two values for samples
5) The purpose of test is to determine whether there is statistical evidence that the mean difference between paired observations on a particular outcome is significantly different from zero. The Paired T-Test is a parametric test

Code

```
before <-c(200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2,
           185.5, 205.2, 193.7)
after <-c(392.4, 393.2, 345.2, 345.1, 393, 434, 427.9,
          422, 383.9, 392.3, 352.2)
res <-t.test(before, after, paired = TRUE)
res
```

Output

```
data: before and after
t = -20.883, df = 9, p-value = 6.2e-09
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
-215.5581 -173.4219
Mean of differences
-194.49
```

Scanned with CamScanner

Name - Neeraj Appar
F-129

(Q12)
Explain Tukey's method and write a program in R software to understand the output.

→ Tukey's range test also known as Tukey's test Tukey method
- The Tukey method applies simultaneously to the set of all pairwise comparisons
- The confidence coefficient for the set when all the sample size are equal exactly $1-\alpha$ for unequal sample size
  E.g max(x) - min(x) / RMS / n

Code

```
library(multcomp)
set.seed(1) treatment <- rep(c("A", "B", "C", "D", "E"),
                             each = 20)
value = c(sample(2.5, 20, replace = T), sample(6:10, 20,
  replace T), sample(1:7, 20, replace T),
model = lm(data$value ~ data$treatment)
anova = aov(model)
Tukey <- TukeyHSD(Anova, data$treatment,
                  conf.level = 0.95)
plot(Tukey, las = 1, col = brown)
```

Name - Aleeraj Appoo
F729

Page No.
Date

ALMRAO

Output:

B-A
C-A
D-A
E-A
C-B
D-B
E-D
D-C
E-C
E-D

-5 0 5 10

Neeraj Appa...
F-129

Page No.
Date ACHIRAO

Aug) Write a code to understand the critical region in R Software of Normal distribution

:) The A function that represents the distribution of many random variables as a symmetrical bell-shaped graph is known as normal distribution
Critical region
The set of all the values of the test statistic that would cause us to reject the null hypothesis is called Critical region or rejection region. It indicates that if the value of the test statistic lies in the region. If the test statistic fall into the region of non-rejection area do not reject the null hypothesis
~~If the test statistic~~

Ex
> xbar = 9.0
> mu0 = 10
> sig = 1.1
> N = 30
> z = (xbar - mu0)/(sig/sqrt(n))
> z
[1] -0.4979296

> alpha = 0.5
> z.alpha = qnorm(1-alpha)
> z.alpha

Name - Neeraj Appari
F-129

A.I.N RAO

Q.30) Display the types of Error in Hypothesis testing using R software

Type 1 error
It occurs when rejecting null hypothesis when True

```
type1.tester<-function(mu0, sigma,n, alpha, iteration=1000000){
pvals<-rep(NA, iteration)
for (i in 1:iteration){
temporary.sample<-rnorm(n=n, mean=mu0, sd=sigma)
temporary.mean<-mean(temporary.sample)
temporary.sd<-sd(temporary.sample)
pvals[i]<-1-pt((temporary.mean-mu0)/temporary.sd
 /sqrt(n)), df=n-1)
}
return (mean(pvals<alpha))
}
type1.tester(mu0=5, sigma=0.25, n=100, alpha=0.03)
output- 0.00472
```

Type 2 error
It occurs when failing to reject null hypothesis when it is actually false
~~The probability~~

```
typeII.tester<-function(mu0, TRUEmu, sigma, n, alpha,
    iteration=10000){
pvals<-rep(NA, iteration){
temporary.sample<-rnorm(n=n, mean=TRUEmu
                       sd=sigma)
```

Name - Neeraj Appam
F-12a

A.L.N.RAO

(Q51) Plot the properties of discrete random variable and explain the same.

→ A discrete random variable has a countable number of possible values. The probability of each value of a discrete random variable is between 0 and 1, and the sum of all probabilities is equal to 1. A continuous random variable takes on all the values in some interval of numbers

Code

```
library (discreteRV)
x <- RV (outcomes = 1:6, probs = 1/6)
X probs(x)
plot (x)

Sum (probs (X)*outcomes (x))
```

The easy way is to use the Expectation function in the discrete RV package:

E(X)

Similary, you can calculate variance, by calculating the weighted average of the squared deviations $\Sigma p(x)(x-E[x])^2$ $\Sigma p(x)(x-E(x))^2$

Fridays Appavi

Sum (prob(x))* (outcomes(x) - 1(x))^2)

Or you can use the variance function V, and the standard deviation function, SD:

V(x) & SD(x)

More dice:

What if you roll dice and add the results? You could treat this as the sum of two independant and identically distributed (iid) random variables. In the discrete RV package, the SoPID function sums of iid random variables:

SoPID(X, n=2)

Two dice

<- SoPID(X, n=2)

What if you roll 20 dice and add them up?

Twenty_dice <- SoPID(X, n=20) plot(twenty_dice)

Neeraj Appavi
E129

ALHRAO

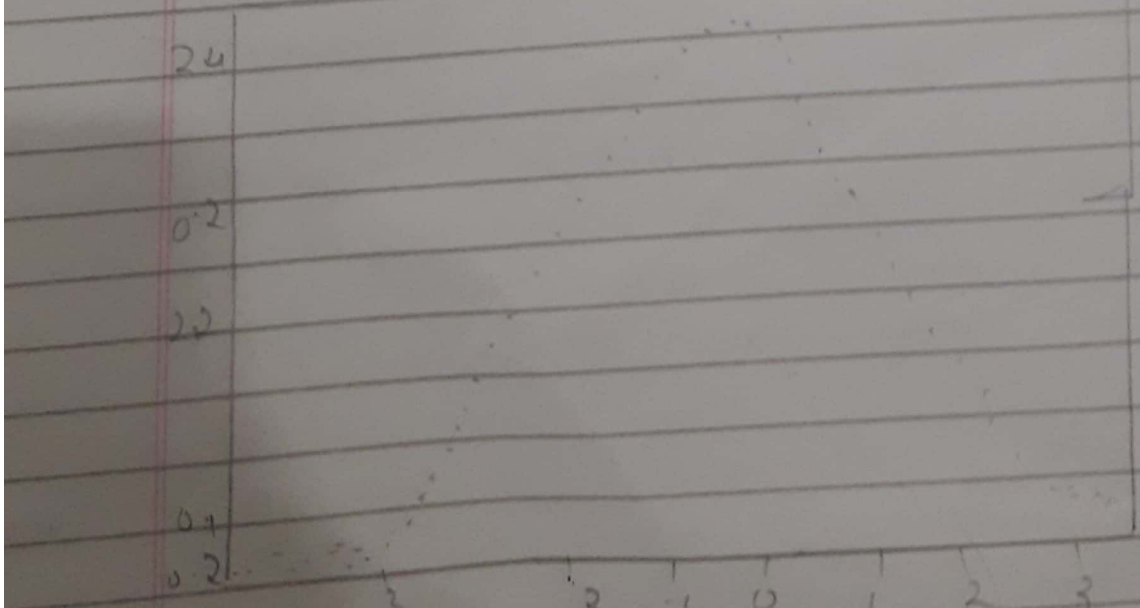
(Q52) Plot the properties of continous variable and explain the same.

-) The probability density function $f(x)$ of a continuous random variable x with a supporting box B

a) $f(x)$ is positive everywhere in the supporting box B, that is $f(x) > 0$ for all in x in B
b) The area under the curve $f(x)$ in box supports is 1 ie $\int_B f(x)\,dx = 1$
c) If $f(x)$ is the PDF of x then the probability that x belongs to A, where A is some interval within the range, is given by the integral of $f(x)$

Code

```
x <- seq(from = -3, to = +3, length.out = 100)
plot(x, dnorm(x))
```



Name- Neeraj Appari
F-129

Page No.
Date 21.12.20

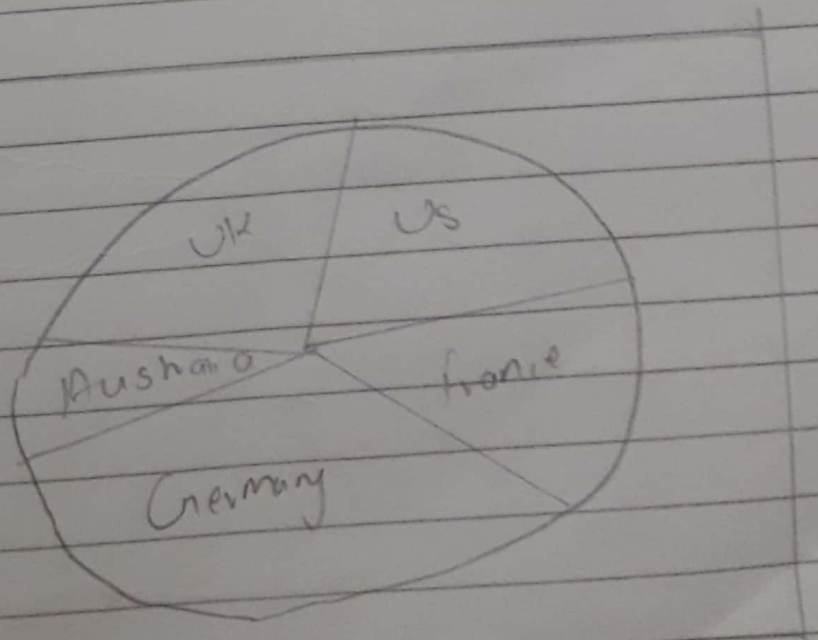
Q53) Plot a pie digram using R software

1) A pie chart is a representation of values as slices of a circle with different colors
2) In R pie chart is created using the pie() function which takes positive numbers as a vector input
3) Pie charts are not recommended in the R documentation, and their features are somewhat limited

Code:

```
Slices <- c(10,12,4,16,8)
lbls <- c("US","UK","Australia","Germany","France")
Pie(slices, labels=lbls, main="Pie Chart of Countries")
```

Output



Name - Neeraj Appari
R.124

Page No.
Date D.L.N.RAO

Q5(b) Explain and plot the bar diagram using R software.

-> 1) A bar graph is a visual tool that uses bars to compare data among categories.
2) A bar graph may run horizontally or vertically. ~~The important thing~~
3) Bar graph consists of two axes. On a vertical bar graph, the horizontal axis (x-axis) and on vertical y axis)

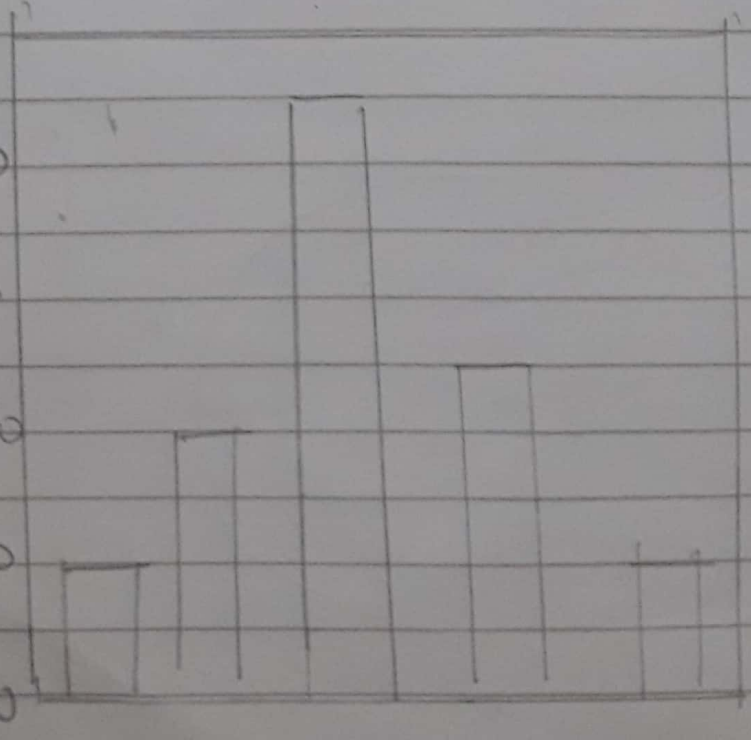
Code

```
H <- c(7,12,28,3,41)

png(file = "bar-chart.png")

barplot(H)

dev.off()
```



Ass) Calculate the mean and variance in R software for a data set and explain the application of both in day to day life

Any numeric variable will have mean, median and mode

Mean

It is calculated by taking by the sum of the values and dividing with the number of values in data sets

Variance (variance is the expectation of the squared deviation of the random variable from its mean

2) Variance is much more useful measure of variation

Meeraj Appar 472a

Mean

Suppose there are frequencies of bus coming in a bus station
i.e = 13, 16, 20, 17, 18, 17, 15, 16)

> x c-c (13, 16, 20, 17, 18, 17, 15, 16)
> x
> mean c

16.5

Vaurce

The company produce random sample of 45 such bulbs which is selected out of which 25% are there to be defective. Find vaurce and SD

> xc - Binom (size = 45, prob = 0.21)
> E (x)
[1] 11.2
> var (x)
Var (x) = 8.432
> sd (x)
Sd (x) = 2.904738

Name - Neeraj Appari
F.12a

Q 56) Perform the vector operation in R software and Explain the same.

Vector is a basic data structure in R. It contains element of the same type. The data types can be logical, integer, double, character, complex or raw.

A A vector's type can be checked with the typeof() function

1) Creating a vector using : operator

```
> x <- 1:7; x
[1] 1 2 3 4 5 6 7
```

2) Creating a vector using seq() function

```
> seq (1,2, by = 0.2)
[1] 1.0 1.2 1.4 1.6 1.8 2.0
```

Using integer vector as index

Vector index in R starts from 1, unlike most programming language where it start from 0

```
> x
[1] 0 2 4 6 8 0
> x[3]
[1] 4
```

Scanned with CamScanner

2) calculate the relationship
3. The coefficient

"A" Name- Neeraj Appavi
I-120

Page No.
Date 21.11.20

3) Using logical vector as index

```
x[c(TRUE, FALSE, FALSE, TRUE)]
[1] -3  3

x[x<0]
[1] -3 -1
```

4) Using character vector as index
This type of indexing is useful when dealing with named vectors.

```
x<-c("first"=3, "second"=0, "third"=9)
names(x)
[1] "first"  "second" "third"

x["second"]
second
     0
```

5) To modify a vector in R

We can modify a vector using assignment operator

```
x<-x
x
[1] -3 -2 -1  0  1  2
x[2]<-0; x
[1] -3  0 -1  0  1  2
```

Name - Neeraj Appavi
1--120

Page No.
Date 12.1.2022

To Delete a vector

We can delete a vector by simply assigning a NULL to it

```
> x
 [1] -3 -2 -1  0  1  2
> x <- NULL
> x
NULL
```

Name - Neeraj Appu
Roll no - F128

Q52] Plot a line of regression in R software and explain the same.

→ 1) A linear regression is a statistical model that analyzes the relationship between a response variable (X) and one or more variables (Y) and their interactions
2) It is used to predict the value of an outcome variable X based on one or more input predictor
3) It is the line that best fits the data

formula $Y = \beta_1 + \beta_2 X + \varepsilon$

Code

```
> x -> weight <- c(70,45,50,102,59,57,63,49,50,60)
> height <- c(164,144,133,175,148,64,173,155,143,175)

> plot(weight, height)
> lm(height ~ weight)
```

Call:
lm(formula = height ~ weight)

Coefficients:

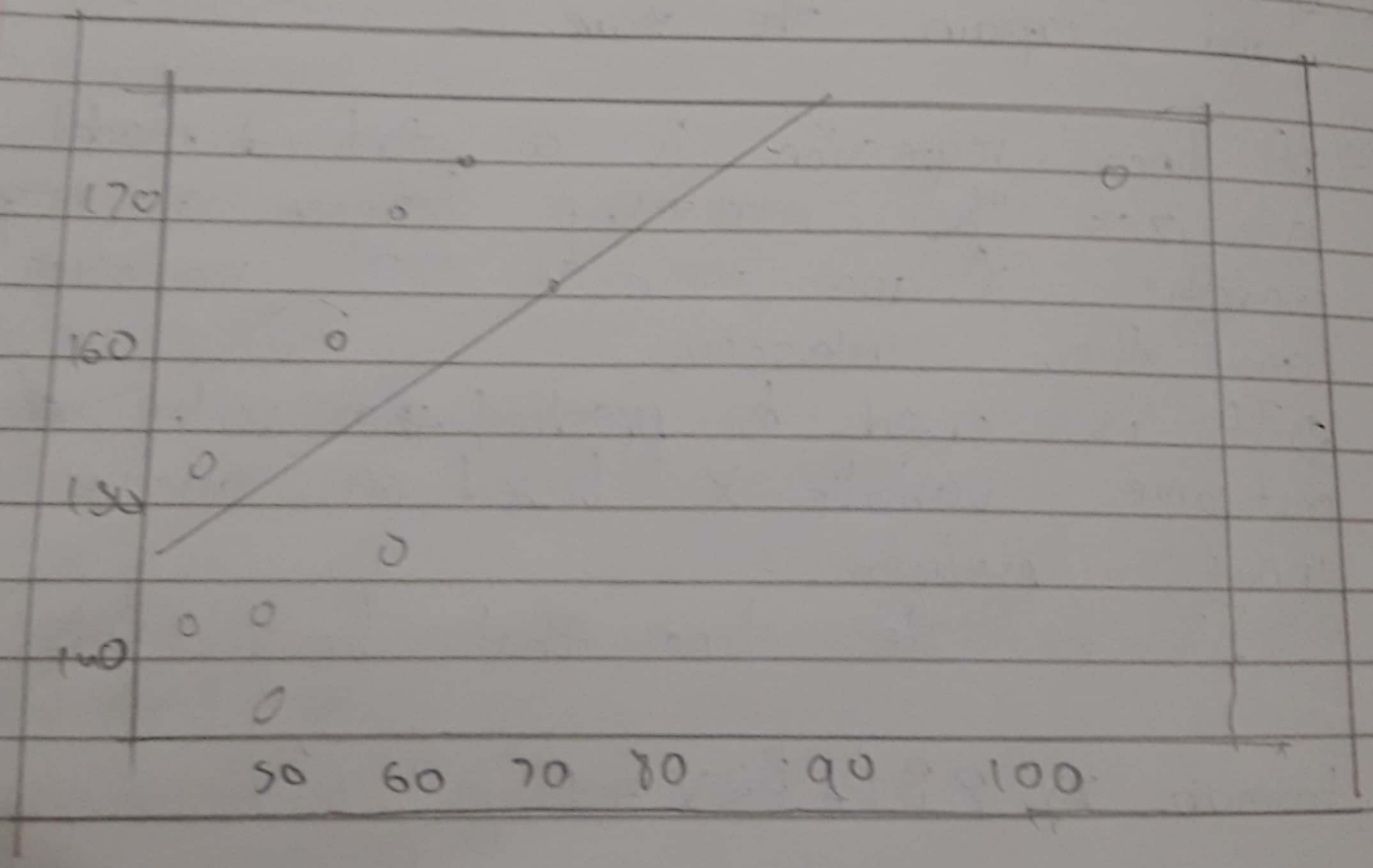
(Intercept)	weight
117.7507	0.640

```
> abline(117.7507, 0.640)
```

F-129

A.L.NIRAO

Output:



Neeraj Appaw
#12a

ALNIRAO

(OSS) Prepare the data set and plot the bayes theorem fa stahstical Analysis and explain Bayes theorem

→ 1) It is mathematical formula for determing conditional pwbability

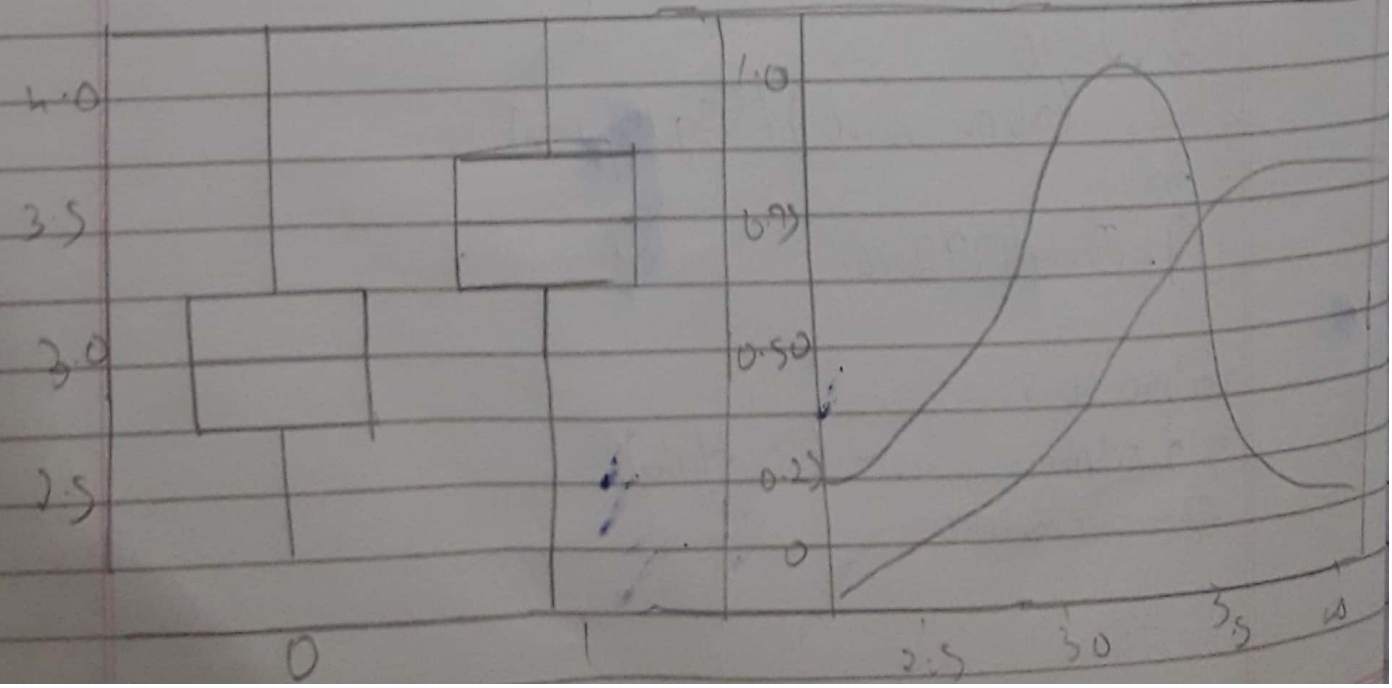
2) This theorem provids a way to revise existing predictions a theories given new or additional evidence

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

→ pairs.panels(data[-1])

- data %>% %>%
  ggplot(aes(x=admit, y=gpa, fill=admit)) +
  geom_boxplot + ggtitle("Box plot")

2) data %>% %>%
  ggplot(aes(x=gpa, fill=admit)) + geom_density(alpha=0.2, colour='black') + ggtitle("Density")



Name - Neeraj Appat
Roll no: 129

Name - Neeraj App
F 129

ALNIRAO

Q8) Plot the Decision Tree in R software and explain the tree

```
install.packages("rpart")
install.packages("rpart.plot")
path <- "C:\\Users\\App.csv"
fit <- read.csv(path)
tail(fit)
```

> tail fit

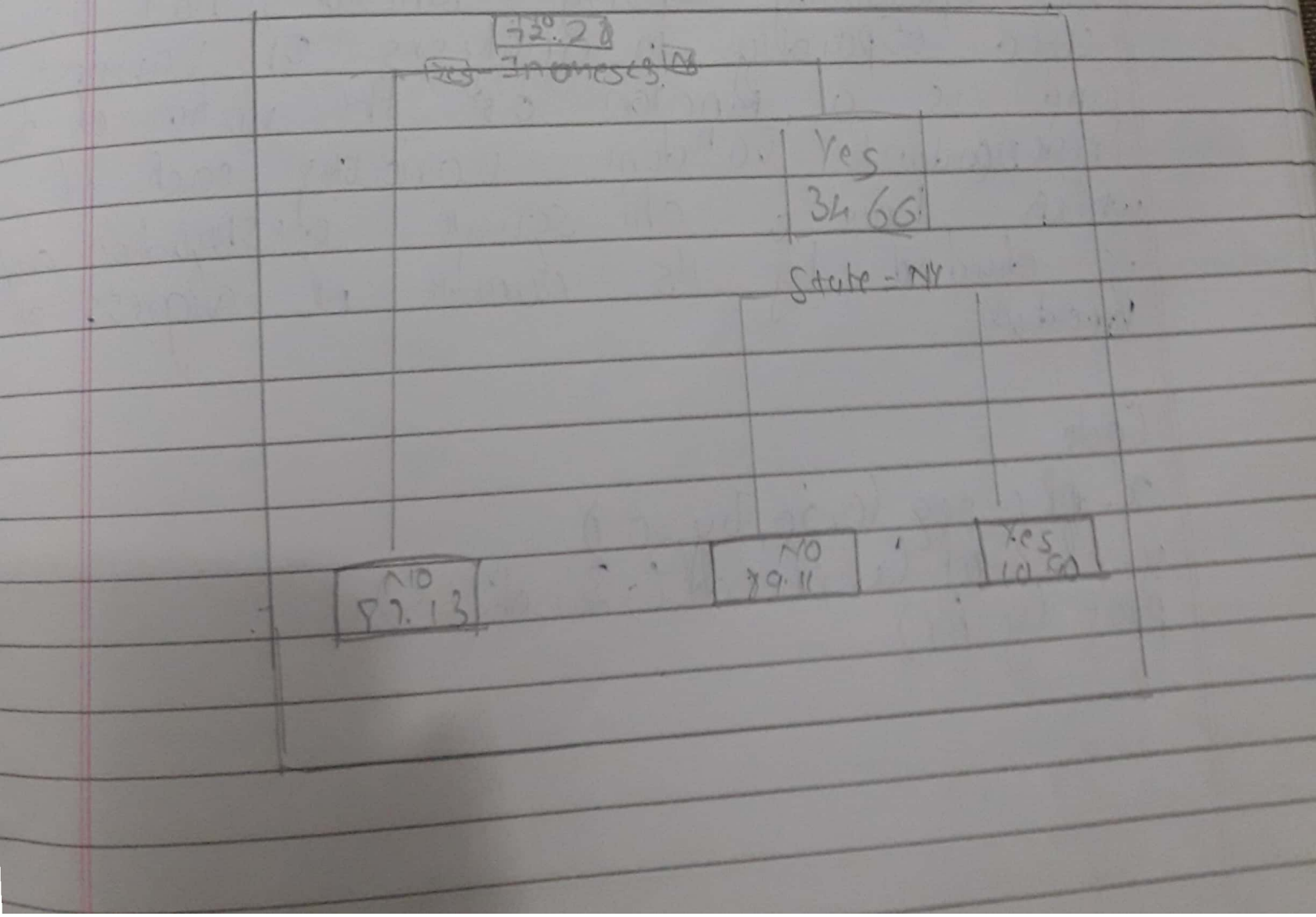
	Income	GymVisits	State	Hours	PayOrNot
95	50	2	MP	6.2	No
96	80	3	CA	6.2	No
97	75	3	WP	7.4	Yes
98	65	4	SD	6.0	No
99	90	3	TY	7.6	Yes

```
fitness <- read.table("C:\\Users\\App.csv", sep=",", head=T)
library("rpart")
tree <- rpart(PayORNot ~ Income + GymVisits + State
  data = fitness)
tree
n=100
node), split, n, loss, yval, yprob
  (dindes) terminal node
1) node 100 28 No
2) Income < 95 71 9
3) Income >= 95 29 10
c) State: NY . 9 1 No
```

Roll no - 129

Page No.
Date A.L.N.R.A.G

```
library ("rpart.plot")
rpart.plot (tree, extra=4)
```



Name - Neeraj Appan
F-12a

ALAI RAJ

Q66. Karl Pearson's Coefficient of Corelation

→ The Karl Pearson's coefficient of corelation is a measure of shenght of a linear association between two variables and used wherein the numerical expression is used to calculate the degreee and direction of the relationship between linear related variables. The coefficient of corrlations in denoted by "r"

formula

$$r = \Sigma(x-\bar{x})(y-\bar{y}) / \sqrt{\Sigma(x-\bar{x})^2 - \sqrt{(y-\bar{y})^2}}$$

where

$\bar{x}$ = mean of x variable

$\bar{y}$ = mean of y variable

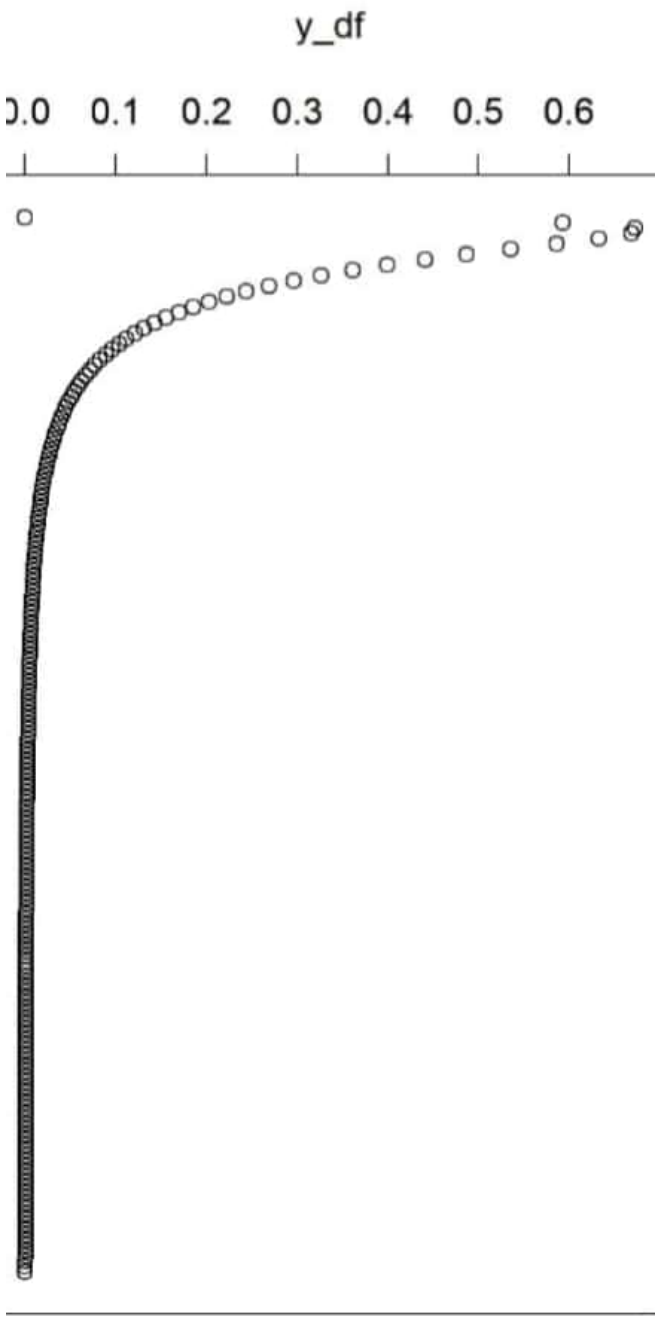
Code

```
cor.test(trees$Girth, trees$Height)
```

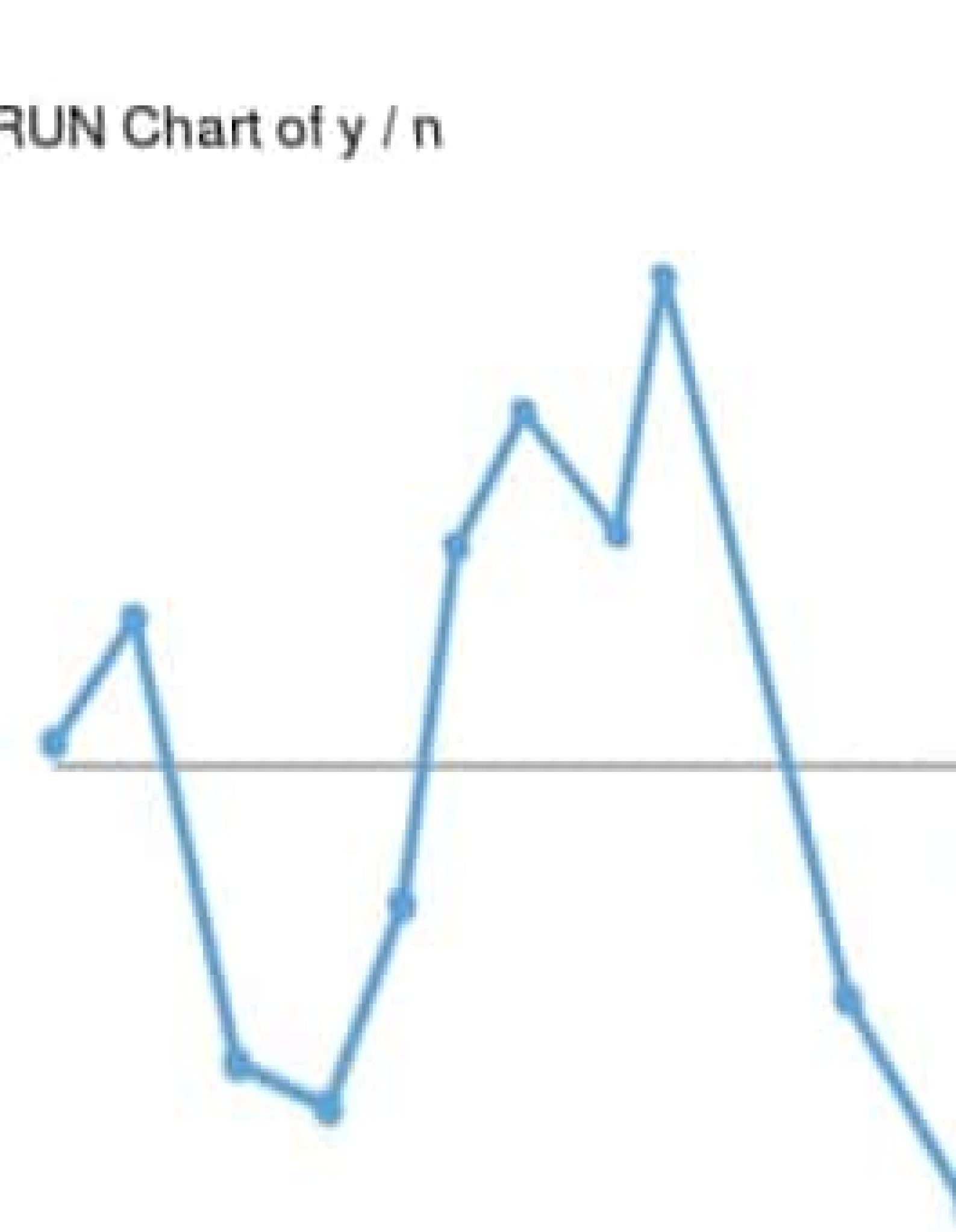
→ cor.test(trees$Girth, trees$Height)

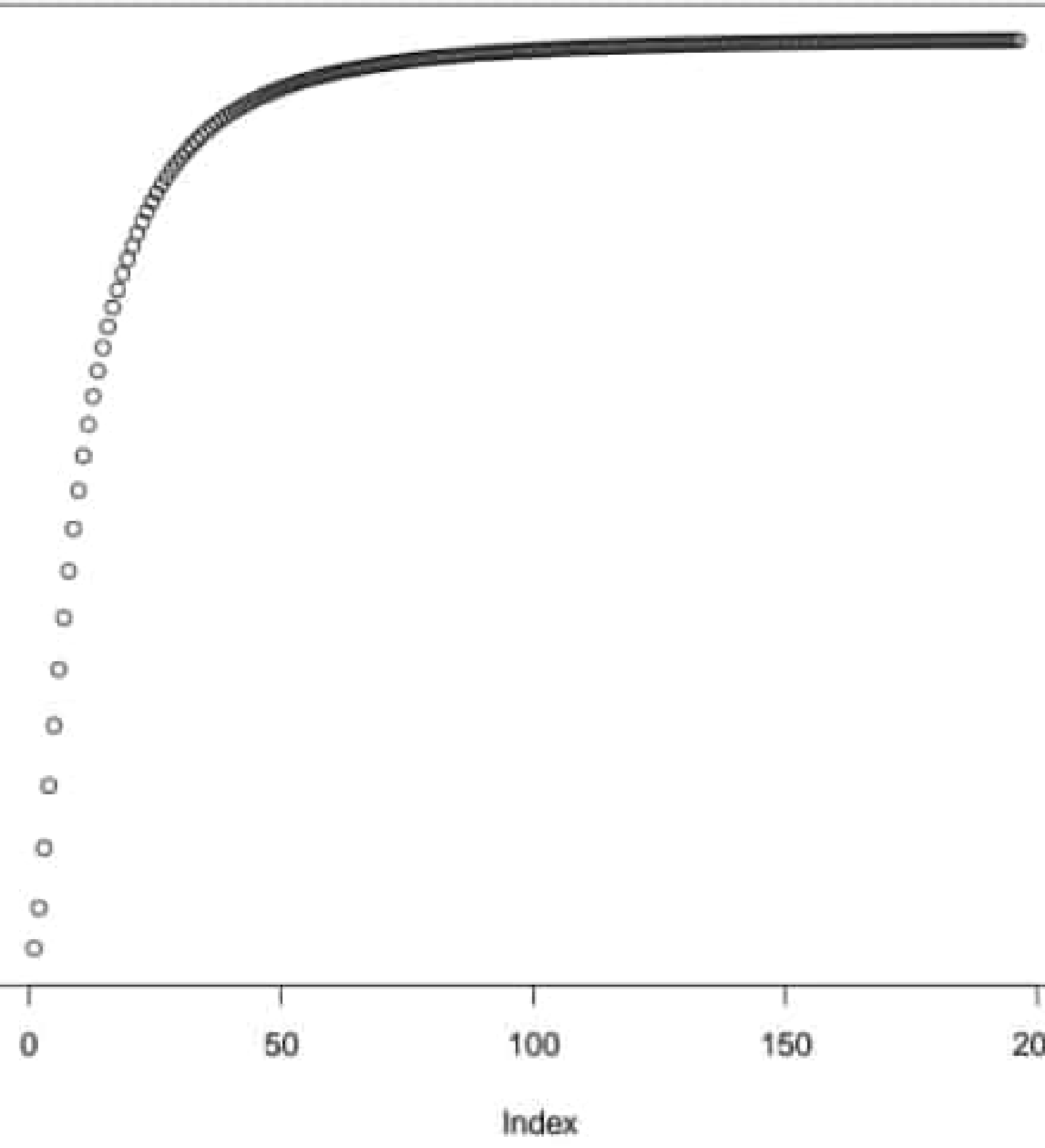
Person's product - moment correlation

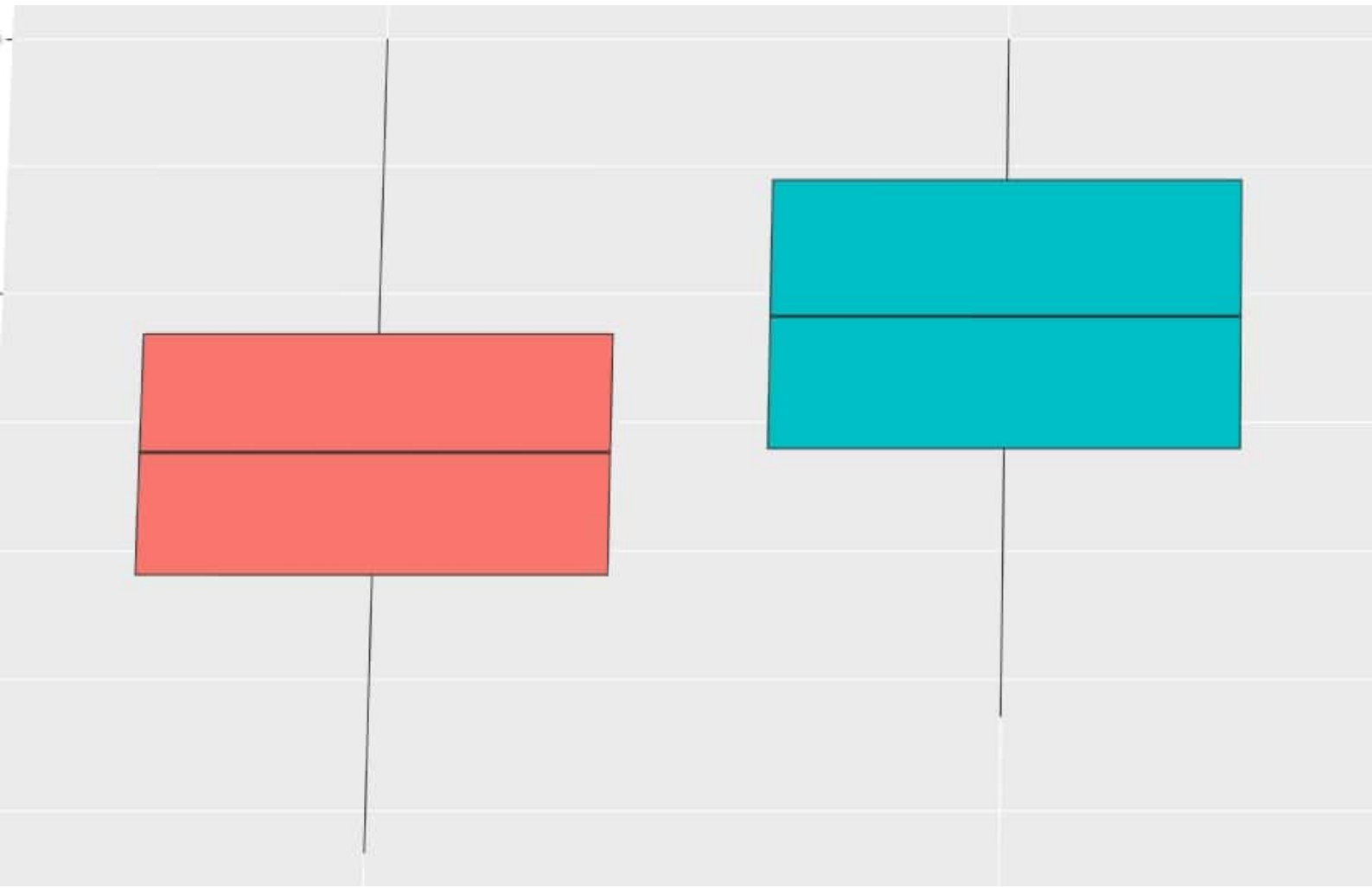
data trees$Girth and trees$Heigh
t = 3.8272, df = 29, p-value = 0.00275
95 percent con
alternative hypothesis: true correlation is not equal to
95 percent confidence interval:
0.2021327 0.7378538
Sample estimates
cor
0.6931801

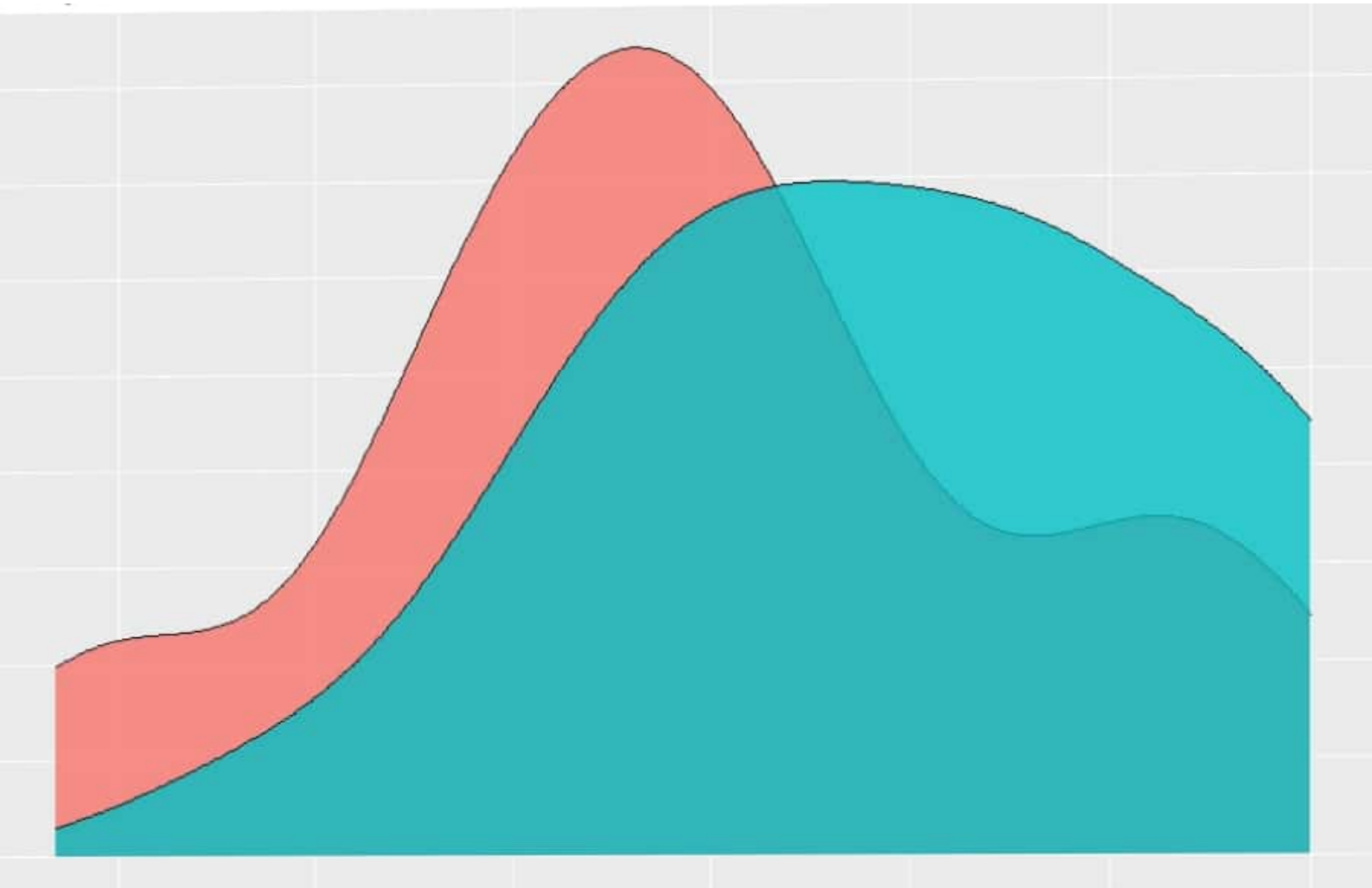


RUN Chart of y / n









Pie Chart of Countries