```
#NOTE:
1. THIS NOTEBOOK CONSISTS OF EXERCISES OF LAB1
2. BOX PLOTS,QQPLOT, ECDF PLOTS ARE PLOTTED TO KNOW THE DISTRIBUTION OF DATA
3. RUG PLOT IS ALSO PLOTTED
4. WE NEED TO UNDERSTAND THE TYPE BY USING CLASS
5. ELIMINATE THE NA AND ?
6. WHILE PLOTTING THE HISTOGRAM WITH SEQUENCE, WE NEED TO FIRST UNDERSTAND THE RANGE OF THE DATA
```

```
install.packages('xlsx')
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘rJava’, ‘xlsxjars’


Warning message in install.packages("xlsx"):
“installation of package ‘rJava’ had non-zero exit status”
Warning message in install.packages("xlsx"):
“installation of package ‘xlsxjars’ had non-zero exit status”
Warning message in install.packages("xlsx"):
“installation of package ‘xlsx’ had non-zero exit status”
```

```
install.packages("readxl")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
library(readxl)
```

```
##################READ_EXCEL TO READ EACH OF THE SHEETS#########
EPI_data<-read_excel("2010EPI_data.xls",
                     sheet = 2)


# For viewing the details of sheet 2. This can be done for other sheets as well

head(EPI_data)

```

A tibble: 6 × 12

Index	Objectives	Objective Codes	Objective Weight (% of EPI)	Policy Categories	Policy Category Codes	Policy Category Weight (% of EPI)	Indicators	Indicator Codes	Indicator Weight in EPI %	S
<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<dbl>	<chr>	<chr>	<dbl>	
EPI	Environmental Health	ENVHEALTH	50	Environmental burden of disease	EH	25.000000	Environmental burden of disease	DALY	25.000000	
NA	NA	NA	NA	Air pollution (effects on humans)	AIR_H	12.500000	Indoor air pollution*	INDOOR	6.250000	Develo Indi
NA	NA	NA	NA	NA	NA	NA	Outdoor air pollution*	PM10	6.250000	Develo Indi
NA	NA	NA	NA	Water (effects on humans)	WATER_H	12.500000	Access to water*	WATSUP	6.250000	Develo Indi
NA	NA	NA	NA	NA	NA	NA	Access to sanitation*	ACSAT	6.250000	Develo Indi
NA	Ecosystem Vitality	ECOSYSTEM	50	Air Pollution (effects on ecosystem)	AIR_E	4.166667	Sulfur dioxide emissions per populated land area	SO2	2.083333	EDGAR UNF RE

```
help("stem")
```

```
##########make the first row as the header of the dataset
EPI_data_csv<-read.csv("EPI_Data.csv")
data_2010EPI <- EPI_data_csv[-1, ]
```

```
head(data_2010EPI)
```

A data.frame: 6

	code	ISO3V10	Country	EPI_regions	GEO_subregion	GDPCAP07	Population07	Landarea	PopulationDensity	Landlock
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
2	4	AFG	Afghanistan	South Asia	South Asia	NA	NA	634924.74	NA	1
3	24	AGO	Angola	Sub-Saharan Africa	Southern Africa	4875.36	17554585.0	1251895.62	14.02	0
4	660	AIA	Anguilla	Latin America and Caribbean	Caribbean	NA	NA	82.83	NA	0
5	8	ALB	Albania	Eastern Europe and Central Asia	Central Europe	6811.38	3132458.0	28346.12	110.51	0
6	20	AND	Andorra	Europe	Western Europe	NA	82180.0	463.79	177.19	1
7	530	ANT	Netherlands Antilles	Latin America and Caribbean	Caribbean	NA	191328.8	818.07	233.88	0

```
summary(data_2010EPI$EPI)
##################DISPLAY THE MEAN, MEDIAN, Minimum, Lower Hinge (Q1),Median (Q2), Upper Hinge (Q3) and Maximum after
fivenum(data_2010EPI$EPI, na.rm = T)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  32.10   48.60   59.20   58.37   67.60   93.50      67
32.1 · 48.6 · 59.2 · 67.6 · 93.5
```

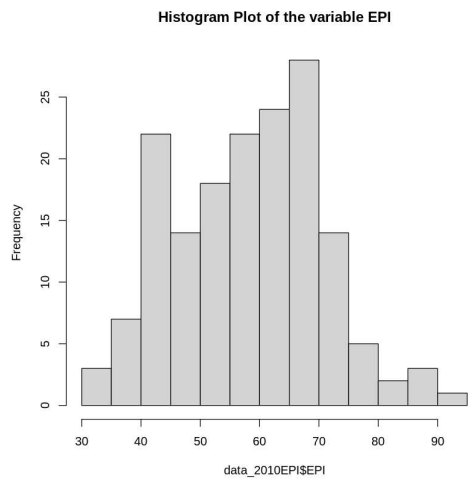
```
#########UNDERSTAND THE STEM PLOT###########
help(stem)
```

```
###########STEM PLOT WITH SCALE WITH SCALE CONTROLLING THE PLOT LENGTH###########
#NOTE#: The  number on the left is the stem and the number on right is the leaf
#####USEFUL FOR CHECKING THE PRECISION OF NUMBERS IN THE DATASET
stem(data_2010EPI$EPI,scale=2)
```

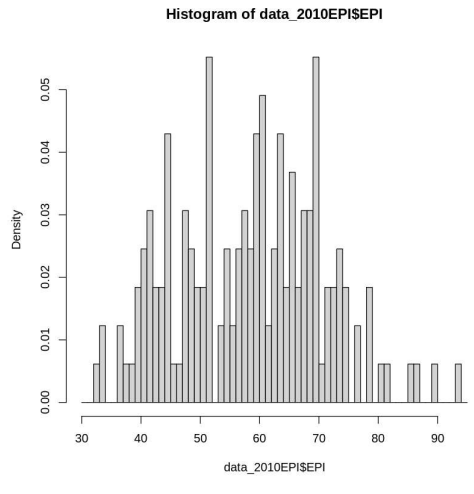
```
  The decimal point is at the |

  32 | 137
  34 |
  36 | 346
  38 | 4456
```

```
40 | 27803789
42 | 033819
44 | 033466679
46 | 01389
48 | 03390289
50 | 138112333446
52 |
54 | 00234639
56 | 13401339
58 | 01280112367
60 | 0445668902
62 | 024591455678
64 | 600467799
66 | 4401348
68 | 022347112334689
70 | 6446
72 | 5501234
74 | 257
76 | 38
78 | 112
80 | 61
82 |
84 |
86 | 04
88 | 1
90 |
92 | 5
```

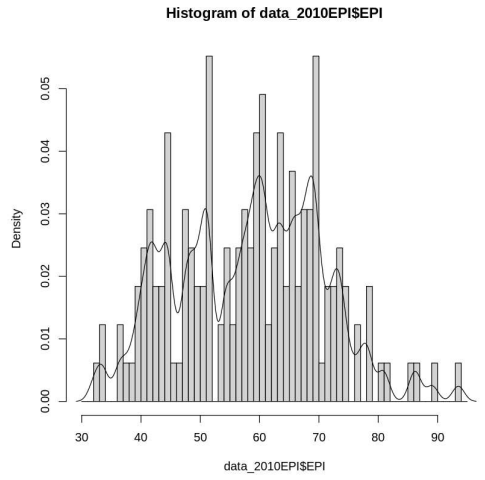
```
1 ########outcome: Values in the range 60-70 are in highest count for the EPI variable#
2 hist(data_2010EPI$EPI,main="Histogram Plot of the variable EPI")
```



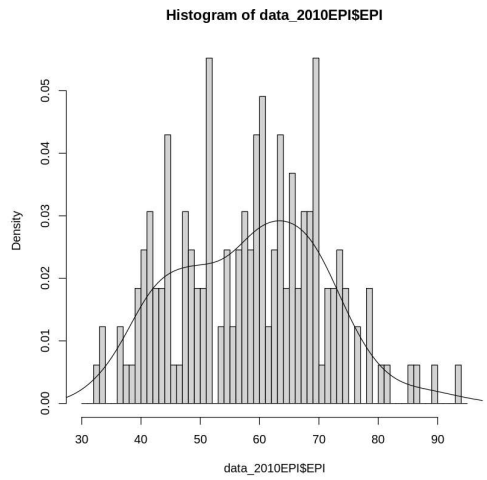

```
1 hist(data_2010EPI$EPI, seq(30., 95., 1.0), prob=TRUE)
```




```
1 hist(data_2010EPI$EPI, seq(30., 95., 1.0), prob = TRUE)
2
3 #######WE SET THE BANDWIDTH=1 HERE
4 lines(density(data_2010EPI$EPI,na.rm=TRUE,bw=1.)) #bw=“SJ”
```

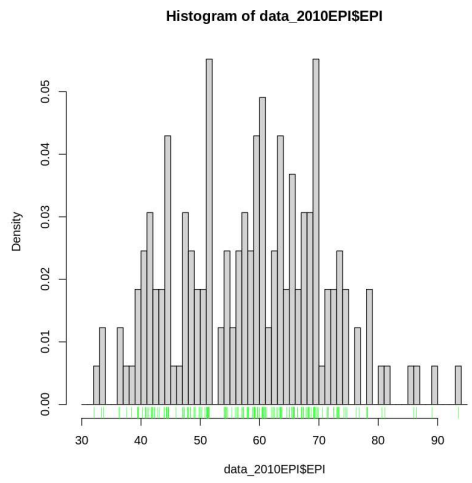



```
1 #Sheather-Jones (SJ) bandwidth selection method:automatically selects an appropriate bandwidt
2 hist(data_2010EPI$EPI, seq(30., 95., 1.0), prob = TRUE)
3 lines(density(data_2010EPI$EPI,na.rm=TRUE,bw="SJ"))
```

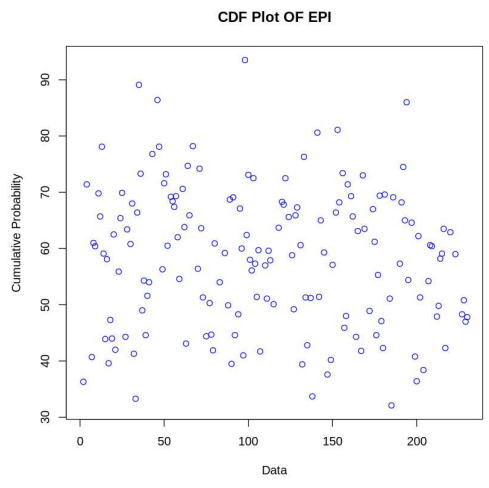


```
1 help(rug)
```

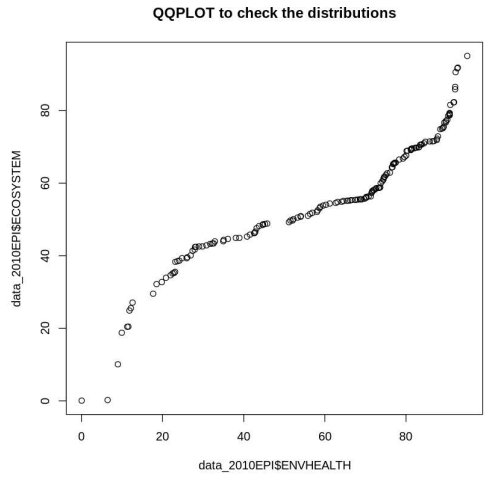
```
1 hist(data_2010EPI$EPI, seq(30., 95., 1.0), prob = TRUE)
2 rug(data_2010EPI$EPI,col="green")
```



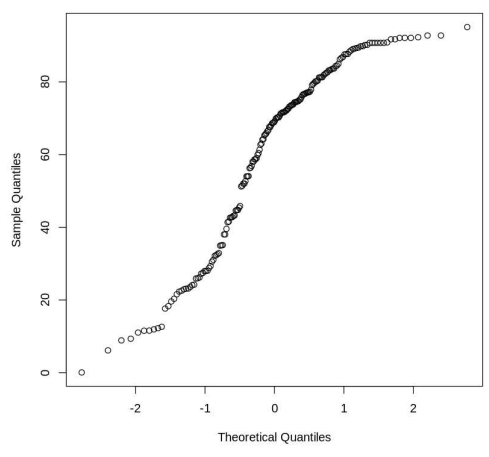

```
1 #CDF plot will show the cumulative distribution of the data, allowing to understand how the data values are
2 #spread across the entire range and what proportion of data falls below or equal to specific values.
3
4 plot(data_2010EPI$EPI, main = "CDF Plot OF EPI", xlab = "Data", ylab = "Cumulative Probability", col = "blue")
```




```
1 #If both sets of quantiles came from the same distribution, then the points on the plot should roughly form a straigh
2 qqplot(data_2010EPI$ENVHEALTH, data_2010EPI$ECOSYSTEM,main="QQPLOT to check the distributions")
```




```
1 qqnorm(data_2010EPI$ENVHEALTH, main = "Q-Q Plot of EPI DALY", xlab = "Theoretical Quantiles", ylab = "Sample Quantile
```




```
1 ###############EXERCISE 2 BEGINS:
2 Your exercise: do the same exploration and
3 fitting for another 2 variables in the EPI_data,
4 i.e. primary variables (DALY, WATER_H, …)
5 • Try fitting other distributions – i.e. as ecdf or qq
```

```
1 EPI_data_csv<-read.csv("EPI_Data.csv")
2 data_2010EPI <- EPI_data_csv[-1, ]
```

```
1 summary(data_2010EPI$DALY)
2 ##################DISPLAY THE MEAN, MEDIAN, Minimum, Lower Hinge (Q1),Median (Q2), Upper Hinge (Q3) and Maximum after the removal of NA values
3 fivenum(data_2010EPI$DALY, na.rm = T)
```

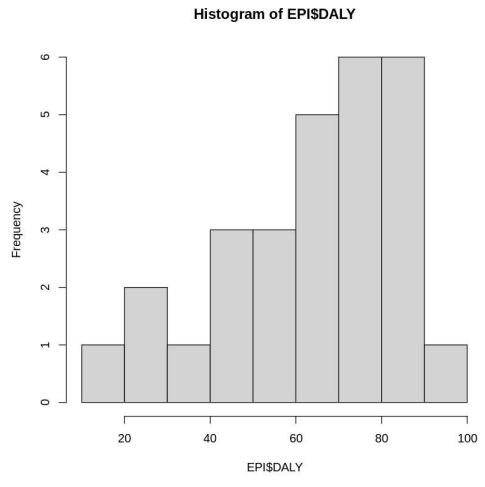
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   37.19   60.35   53.94   71.97   91.50      38
```

0 · 36.955 · 60.35 · 72.32 · 91.5

```
1 #############REMOVE THE NA BEFORE PROCESSING ANY FURTHER#######
2 # Create a data frame with a column containing NA values
3 df <- read.csv("EPI_Data.csv")
4
5 # Use na.omit() to remove rows with NA values
6 EPI <- na.omit(df)
7
```

```
8
9
```

```
1 hist(EPI$DALY)
```




```
1 ######SHORTER STEM PLOT#################
2 stem(EPI$DALY)
```

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

1 | 8
2 | 59
3 | 7
4 | 135
5 | 159
6 | 12268
7 | 024999
8 | 133577
9 | 2
```

```
1 ######Detailed STEM PLOT#################
2 stem(EPI$DALY,scal=2)
3 #######WE SEE THE DISTRIBUTION OF THE DATA CAN BE BETTER UNDERSTOOD HERE#######
```

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

1 | 8
2 |
2 | 59
3 |
3 | 7
4 | 13
4 | 5
5 | 1
5 | 59
6 | 122
6 | 68
7 | 024
7 | 999
8 | 133
8 | 577
9 | 2
```

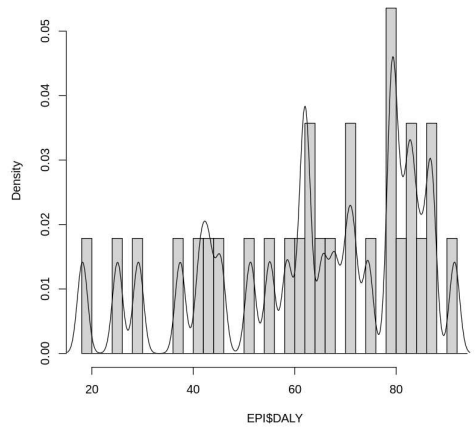
```
1 data_range <- range(EPI$DALY)
2 print(data_range)
3
4
```

```
[1] 18.16 91.50
```

```
1 print(class(EPI$DALY))
```

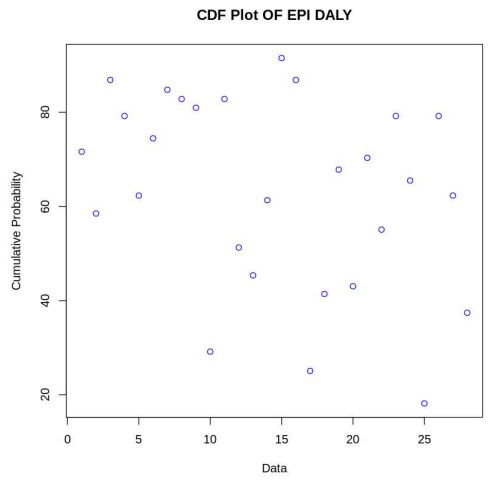
```
[1] "numeric"
```

```
1 hist(EPI$DALY, seq(18., 95., 2.0), prob = TRUE)
2 lines(density(EPI$DALY,na.rm=TRUE,bw=1.))
```

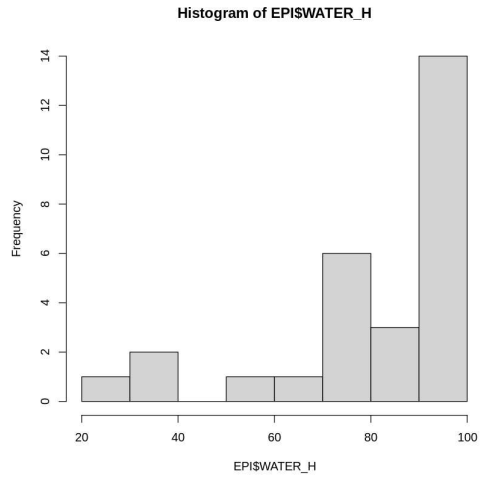



```
1 # Create the Q-Q plot against the standard normal distribution
2 qqnorm(EPI$DALY, main = "Q-Q Plot of EPI DALY", xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
3 ##########: A straight diagonal line indicates a good fit between the data and the theoretical distribution.
4 ####Deviations from a straight line suggest variation from the assumed distribution.
5 ##########This is near to straight diagonal
```

```
#CDF plot will show the cumulative distribution of the data, allowing to understand how the data values are
#spread across the entire range and what proportion of data falls below or equal to specific values.

plot(EPI$DALY, main = "CDF Plot OF EPI DALY", xlab = "Data", ylab = "Cumulative Probability", col = "blue")
```



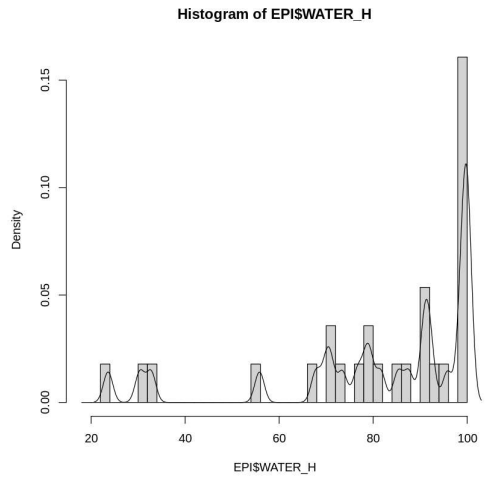

```
######WATER QUALITY WITH VALUE 100 ARE HIGHER#######
hist(EPI$WATER_H)
```




```
data_range <- range(EPI$WATER_H)
print(data_range)
```

```
[1]  23.61 100.00
```

```

hist(EPI$WATER_H, seq(18., 100., 2.0), prob = TRUE)
lines(density(EPI$WATER_H,na.rm=TRUE,bw=1.))
```



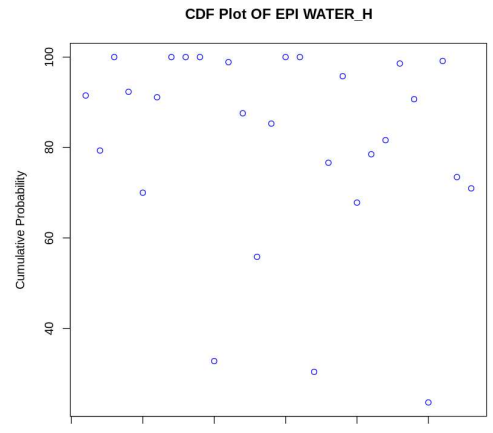

```
stem(EPI$WATER_H)
```

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

   2 | 4
   3 | 03
   4 |
   5 | 6
   6 | 8
   7 | 013799
   8 | 258
   9 | 11226999
  10 | 000000
```

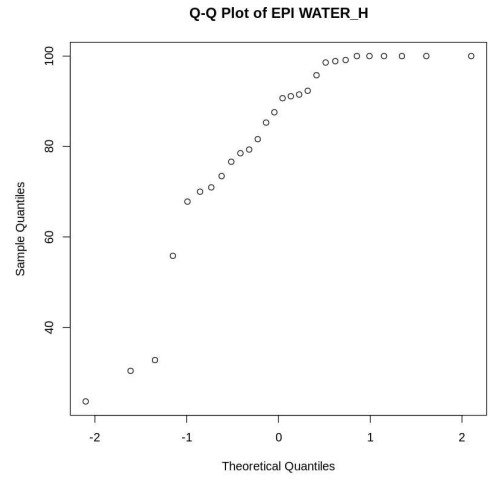
```
stem(EPI$WATER_H,scale=2)
```

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

   2 | 4
   2 |
   3 | 03
   3 |
   4 |
   4 |
   5 |
   5 | 6
   6 |
   6 | 8
   7 | 013
   7 | 799
   8 | 2
   8 | 58
   9 | 1122
   9 | 6999
  10 | 000000
```

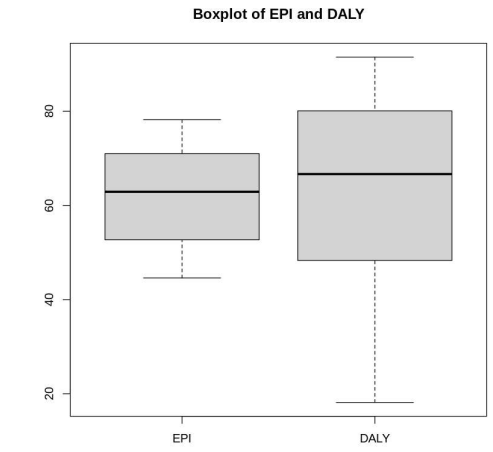
```

plot(EPI$WATER_H, main = "CDF Plot OF EPI WATER_H", xlab = "Data", ylab = "Cumulative Probability", col = "blue")
```



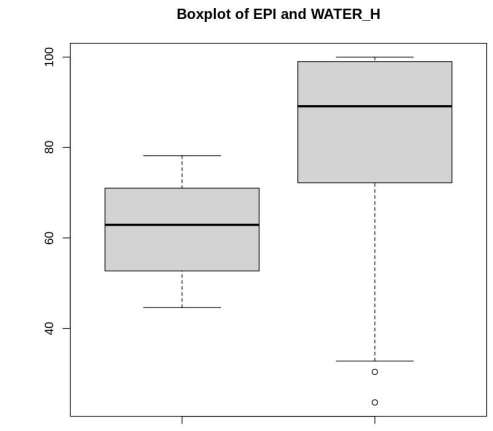

```
1 qqnorm(EPI$WATER_H, main = "Q-Q Plot of EPI WATER_H", xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
```



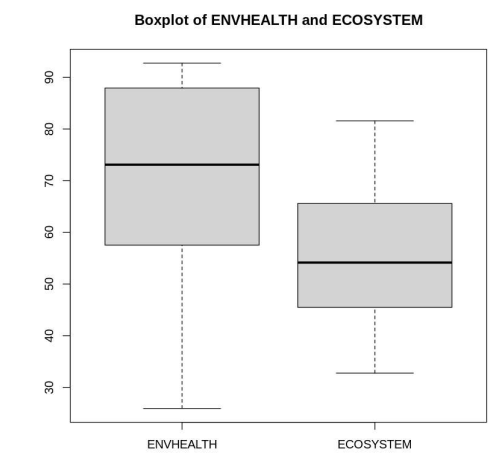

```
1 boxplot(EPI$EPI,EPI$DALY,main = "Boxplot of EPI and DALY",names = c("EPI", "DALY"))
```



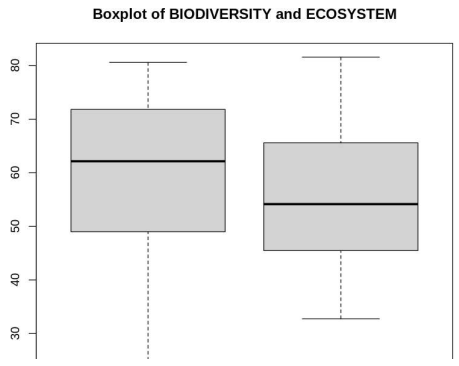

```
1 #########WE SEE SOME OUTLIERS HERE#########
2 boxplot(EPI$EPI,EPI$WATER_H,main = "Boxplot of EPI and WATER_H",names = c("EPI", "WATER_H"))
```



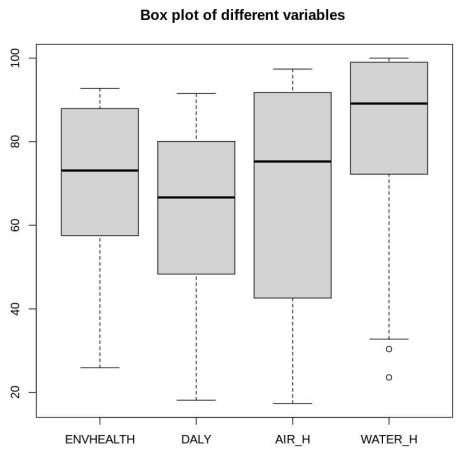

```
1
2  #############FOR THE BOXPLOT WE CAN USE THE MAIN TO DISPLAY THE TITLE AND NAMES TO DISPLAY THE NAMES OF THE PLOTS
3 boxplot(EPI$ENVHEALTH, EPI$ECOSYSTEM, main = "Boxplot of ENVHEALTH and ECOSYSTEM",names = c("ENVHEALTH", "ECOSYSTEM")
```




```
1
2  #############FOR THE BOXPLOT WE CAN USE THE MAIN TO DISPLAY THE TITLE AND NAMES TO DISPLAY THE NAMES OF THE PLOTS
3 boxplot(EPI$BIODIVERSITY, EPI$ECOSYSTEM, main = "Boxplot of BIODIVERSITY and ECOSYSTEM",names = c("AIR_EWATER_E", "EC
```




```
1 ENVHEALTH <- EPI$ENVHEALTH
2 DALY <- EPI$DALY
3 AIR_H<- EPI$AIR_H
4 WATER_H<- EPI$WATER_H
5
6 boxplot(ENVHEALTH,DALY,AIR_H,WATER_H, names = c("ENVHEALTH", "DALY", "AIR_H", "WATER_H"), main="Box plot of different
```




```
1 ################Exercise 2: filtering (populations)
```

```
1 ########WE CANNOT DIRECTLY ACCESS THE VARIABLE, HENCE ALWAYS USE THE DOLLAR SIGN######
2 EPILand<-EPI[!EPI$Landlock]
3
```

```
1 head(EPILand)
```

A data.frame: 6 × 160

	code	ISO3V10	Country	EPI_regions	GEO_subregion	GDPCAP07	Population07	Landarea	PopulationDensity	Landlock	⋯	OZONE_ttr	SO2_ttr	NOX_ttr	NMVOC_ttr	WATSTR_ttr	MPAEEZ_ttr	AGWAT_ttr	GHGCA
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	⋯	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	
9	32	ARG	Argentina	Latin America and Caribbean	South America	12501.62	39503466	2736296.0	14.44	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
29	76	BRA	Brazil	Latin America and Caribbean	South America	9145.54	190119995	8511043.6	22.34	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
35	124	CAN	Canada	North America	North America	36260.04	32976000	9458906.8	3.49	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
37	152	CHL	Chile	Latin America and Caribbean	South America	13087.43	16594596	721229.3	23.01	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
38	156	CHN	China	East Asia and the Pacific	Northeast Asia	5083.66	1318309724	9198093.5	143.32	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
48	192	CUB	Cuba	Latin America and Caribbean	Caribbean	9100.00	11257013	111198.9	101.23	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2

```
1 Eland <- EPILand[!is.na(EPILand)]
2 head(Eland)
3 #####FIND THE CLASS OF THIS VARIABLE########
4
5
```

' 32' · ' 76' · '124' · '152' · '156' · '192'

```
1 EPIWater<-EPI[!EPI$No_surface_water]
```

```
1 head(EPIWater)
```

A data.frame: 6 × 160

	code	ISO3V10	Country	EPI_regions	GEO_subregion	GDPCAP07	Population07	Landarea	PopulationDensity	Landlock	⋯	OZONE_ttr	SO2_ttr	NOX_ttr	NMVOC_ttr	WATSTR_ttr	MPAEEZ_ttr	AGWAT_ttr	GHGCA
	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	⋯	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	
9	32	ARG	Argentina	Latin America and Caribbean	South America	12501.62	39503466	2736296.0	14.44	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
29	76	BRA	Brazil	Latin America and Caribbean	South America	9145.54	190119995	8511043.6	22.34	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
35	124	CAN	Canada	North America	North America	36260.04	32976000	9458906.8	3.49	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
37	152	CHL	Chile	Latin America and Caribbean	South America	13087.43	16594596	721229.3	23.01	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
38	156	CHN	China	East Asia and the Pacific	Northeast Asia	5083.66	1318309724	9198093.5	143.32	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2
48	192	CUB	Cuba	Latin America and Caribbean	Caribbean	9100.00	11257013	111198.9	101.23	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.2

```
1 install.packages("dplyr")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
1 library(dplyr)
```

```
Attaching package: ‘dplyr’
```

```
The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

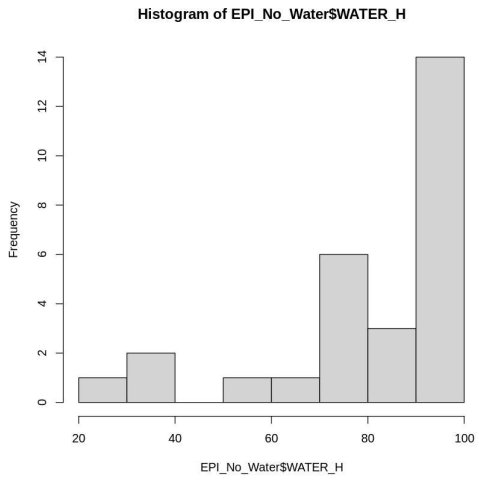
```
1
2 filterdf <- EPI%>% filter(No_surface_water==0)
3
4 # Print filtered data frame
5 filterdf
```

A data.frame: 28 × 160

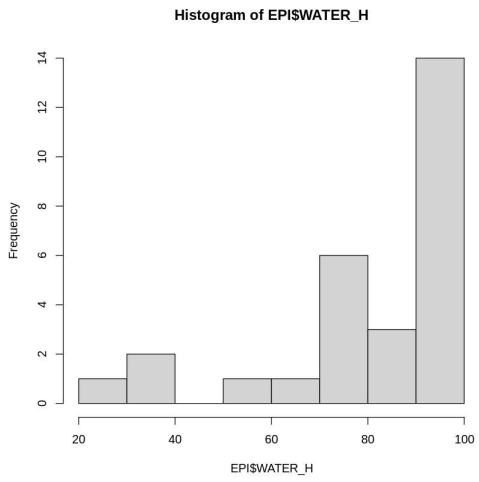
code	ISO3V10	Country	EPI_regions	GEO_subregion	GDPCAP07	Population07	Landarea	PopulationDensity	Landlock	⋯	OZONE_ttr	SO2_ttr	NOX_ttr	NMVOC_ttr	WATSTR_ttr	MPAEEZ_ttr	AGWAT_ttr	GHGCAP_
<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	⋯	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<d
32	ARG	Argentina	Latin America and Caribbean	South America	12501.62	39503466	2736296.00	14.44	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
76	BRA	Brazil	Latin America and Caribbean	South America	9145.54	190119995	8511043.60	22.34	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
124	CAN	Canada	North America	North America	36260.04	32976000	9458906.77	3.49	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
152	CHL	Chile	Latin America and Caribbean	South America	13087.43	16594596	721229.34	23.01	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
156	CHN	China	East Asia and the Pacific	Northeast Asia	5083.66	1318309724	9198093.51	143.32	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
192	CUB	Cuba	Latin America and Caribbean	Caribbean	9100.00	11257013	111198.91	101.23	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
724	ESP	Spain	Europe	Western Europe	28536.42	44878945	505283.96	88.82	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
250	FRA	France	Europe	Western Europe	31624.73	61707072	547106.71	112.79	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
826	GBR	United Kingdom	Europe	Western Europe	33716.77	61001341	247168.89	246.80	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
288	GHA	Ghana	Sub-Saharan Africa	Western Africa	1290.32	22870966	231729.96	98.70	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
300	GRC	Greece	Europe	Western Europe	26928.23	11193366	131890.98	84.87	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
320	GTM	Guatemala	Latin America and Caribbean	Meso America	4333.36	13348222	108523.47	123.00	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
360	IDN	Indonesia	East Asia and the Pacific	South East Asia	3504.21	225630065	1897811.61	118.89	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
364	IRN	Iran	Middle East and North Africa	South Asia	10345.55	71021039	1590351.27	44.66	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
376	ISR	Israel	Middle East and North Africa	Western Europe	24824.21	7180100	21878.11	328.19	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
380	ITA	Italy	Europe	Western Europe	28681.57	59374701	299286.24	198.39	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
404	KEN	Kenya	Sub-Saharan Africa	Eastern Africa	1456.45	37530726	579617.08	64.75	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
144	LKA	Sri Lanka	South Asia	South Asia	4007.59	20010000	65830.04	303.96	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
458	MYS	Malaysia	East Asia and the Pacific	South East Asia	12766.19	26549518	330798.79	80.26	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
586	PAK	Pakistan	South Asia	South Asia	2357.26	162481399	785319.81	206.90	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
591	PAN	Panama	Latin America and Caribbean	Meso America	10756.84	3340605	74515.22	44.83	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
608	PHL	Philippines	East Asia and the Pacific	South East Asia	3181.65	88718185	295386.88	300.35	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
620	PRT	Portugal	Europe	Western Europe	21168.74	10608335	91426.39	116.03	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
792	TUR	Turkey	Eastern Europe and Central Asia	Central Europe	11967.68	73003736	768693.62	94.97	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
834	TZA	Tanzania	Sub-Saharan Africa	Southern Africa	1118.13	41276209	891021.57	46.32	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
840	USA	United States of America	North America	North America	43102.28	301290000	9210753.38	32.71	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
704	VNM	Viet Nam	East Asia and the Pacific	South East Asia	2455.18	85154900	328819.59	258.97	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
710	ZAF	South Africa	Sub-Saharan Africa	Southern Africa	9224.22	47850700	1217643.11	39.30	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252

```
1 # Use na.omit() to remove rows with NA values
2 EPI_No_Water <- na.omit(filterdf)
```

```
1 hist(EPI_No_Water$WATER_H)
```



```
1 hist(EPI$WATER_H)
```




```
1 ######NOTES: SURFACE WATER DOES NOT AFFECT THE WATER HEALTH###########
```

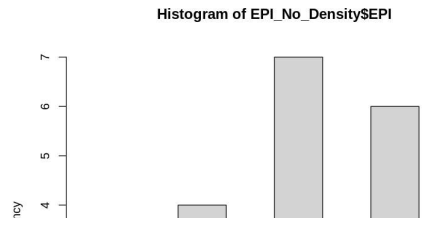
```
1 filterdf_density <- EPI%>% filter(High_Population_Density==0)
2
3 # Print filtered data frame
4 filterdf_density
```

A data.frame: 26 × 160

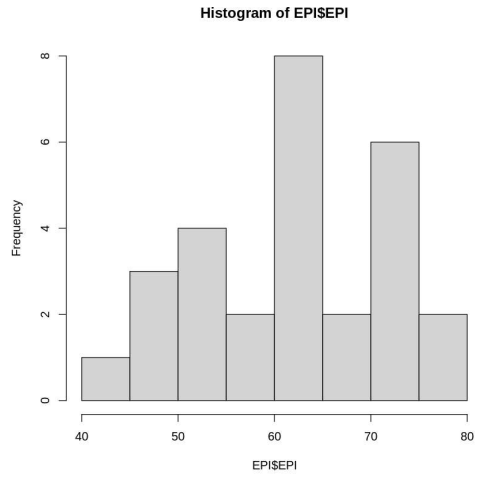
code	ISO3V10	Country	EPI_regions	GEO_subregion	GDPCAP07	Population07	Landarea	PopulationDensity	Landlock	⋯	OZONE_ttr	SO2_ttr	NOX_ttr	NMVOC_ttr	WATSTR_ttr	MPAEEZ_ttr	AGWAT_ttr	GHGCAP_
<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	⋯	<int>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<d
32	ARG	Argentina	Latin America and Caribbean	South America	12501.62	39503466	2736296.00	14.44	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
76	BRA	Brazil	Latin America and Caribbean	South America	9145.54	190119995	8511043.60	22.34	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
124	CAN	Canada	North America	North America	36260.04	32976000	9458906.77	3.49	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
152	CHL	Chile	Latin America and Caribbean	South America	13087.43	16594596	721229.34	23.01	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
156	CHN	China	East Asia and the Pacific	Northeast Asia	5083.66	1318309724	9198093.51	143.32	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
192	CUB	Cuba	Latin America and Caribbean	Caribbean	9100.00	11257013	111198.91	101.23	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
724	ESP	Spain	Europe	Western Europe	28536.42	44878945	505283.96	88.82	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
250	FRA	France	Europe	Western Europe	31624.73	61707072	547106.71	112.79	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
826	GBR	United Kingdom	Europe	Western Europe	33716.77	61001341	247168.89	246.80	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
288	GHA	Ghana	Sub-Saharan Africa	Western Africa	1290.32	22870966	231729.96	98.70	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
300	GRC	Greece	Europe	Western Europe	26928.23	11193366	131890.98	84.87	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
320	GTM	Guatemala	Latin America and Caribbean	Meso America	4333.36	13348222	108523.47	123.00	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
360	IDN	Indonesia	East Asia and the Pacific	South East Asia	3504.21	225630065	1897811.61	118.89	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
364	IRN	Iran	Middle East and North Africa	South Asia	10345.55	71021039	1590351.27	44.66	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
376	ISR	Israel	Middle East and North Africa	Western Europe	24824.21	7180100	21878.11	328.19	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
380	ITA	Italy	Europe	Western Europe	28681.57	59374701	299286.24	198.39	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
404	KEN	Kenya	Sub-Saharan Africa	Eastern Africa	1456.45	37530726	579617.08	64.75	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
458	MYS	Malaysia	East Asia and the Pacific	South East Asia	12766.19	26549518	330798.79	80.26	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
586	PAK	Pakistan	South Asia	South Asia	2357.26	162481399	785319.81	206.90	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
591	PAN	Panama	Latin America and Caribbean	Meso America	10756.84	3340605	74515.22	44.83	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
620	PRT	Portugal	Europe	Western Europe	21168.74	10608335	91426.39	116.03	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
792	TUR	Turkey	Eastern Europe and Central Asia	Central Europe	11967.68	73003736	768693.62	94.97	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
834	TZA	Tanzania	Sub-Saharan Africa	Southern Africa	1118.13	41276209	891021.57	46.32	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
840	USA	United States of America	North America	North America	43102.28	301290000	9210753.38	32.71	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
704	VNM	Viet Nam	East Asia and the Pacific	South East Asia	2455.18	85154900	328819.59	258.97	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252
710	ZAF	South Africa	Sub-Saharan Africa	Southern Africa	9224.22	47850700	1217643.11	39.30	0	⋯	0	-4.60517	-4.60517	-4.60517	0	2.397895	2.397895	1.252

```
1 EPI_No_Density <- na.omit(filterdf_density)
```

```
1 hist(EPI_No_Density$EPI)
```




```
1 hist(EPI$EPI)
```




```
1 #EXERCISE: how to filter on EPI_regions or GEO_subregion?
2 EPI_South_Asia <- EPI[EPI$GEO_subregion=="South Asia"]
3 head(EPI_South_Asia)
```

A data.frame: 6 × 18

	EPI	AIR_H	AIR_E	WATSTR_pt	FORGRO_pt	MTI_pt	NOX_raw	PACOV_raw	MPAEEZ_raw	WSI_w	MPAEEZ_w	FORCOV_w	ACSAT_t	OZONE_t	NOX_t	NMVOC_ttr	GHGCAP_ttr	CO2KWH_ttr
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
9	61.0	63.21	48.24	23.10284	78.0540	78.64179	0.753744	3.899200	0.182754	0	0.16784562	-0.4	100	0	0.01	-4.60517	1.252763	2.302585
29	63.4	90.18	39.31	71.66523	83.5036	100.00000	2.275718	9.400000	0.874877	0	0.62854306	-0.6	100	0	0.01	-4.60517	1.252763	2.302585
35	66.4	97.37	25.27	76.66305	100.0000	0.00000	6.835122	7.766200	0.514395	0	0.41501602	0.0	100	0	0.01	-4.60517	1.252763	2.302585
37	73.3	74.40	42.21	31.73881	100.0000	100.00000	1.265852	6.700397	0.028248	0	0.02785638	0.4	100	0	0.01	-4.60517	1.252763	2.302585
38	49.0	40.07	30.19	27.89280	100.0000	100.00000	2.362702	8.568200	0.264059	0	0.23432797	2.2	100	0	0.01	-4.60517	1.252763	2.302585
48	78.1	97.37	40.66	19.13183	100.0000	100.00000	2.493357	6.586000	0.606140	0	0.47383379	2.2	100	0	0.01	-4.60517	1.252763	2.302585

```
1
```

```
1 ##########THIS MARKS THE BEGINNING OF THE WATER TREATMENT DATASET#######
2 Water=read.csv("water-treatment.csv")
3 Water=Water[-1,]
```

```
1 Water_cleaned <- na.omit(Water)
```

```
1 head(Water_cleaned)
```

A data.frame: 6 × 39

	DATE	Q.E	ZN.E	PH.E	DBO.E	DQO.E	SS.E	SSV.E	SED.E	COND.E	…	COND.S	RD.DBO.P	RD.SS.P	RD.SED.P	RD.DBO.S	RD.DQO.S	RD.DBO.G	RD.DQO.G	RD.SS.G	RD.SED.G
	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	…	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
2	D-2/3/90	39024	3.00	7.7	?	443	214	69.2	6.5	2660	…	2590	?	60.7	94.8	?	80.8	?	79.5	92.1	100
3	D-4/3/90	32229	5.00	7.6	?	528	186	69.9	3.4	1666	…	1888	?	58.2	95.6	?	52.9	?	75.8	88.7	98.5
4	D-5/3/90	35023	3.50	7.9	205	588	192	65.6	4.5	2430	…	1840	33.1	64.2	95.3	87.3	72.3	90.2	82.3	89.6	100
5	D-6/3/90	36924	1.50	8.0	242	496	176	64.8	4.0	2110	…	2120	?	62.7	95.6	?	71.0	92.1	78.2	87.5	99.5
6	D-7/3/90	38572	3.00	7.8	202	372	186	68.8	4.5	1644	…	1764	?	59.7	96.5	86.7	78.3	90.1	73.1	84.9	100
7	D-8/3/90	41115	6.00	7.8	?	552	262	64.1	5.0	1603	…	1703	?	61.9	93.8	89.1	79.8	?	86.2	90.1	99.0

```
1 summary(Water_cleaned)
```

```
      DATE               Q.E               ZN.E               PH.E       
 Length:526         Length:526         Length:526         Min.   :6.90  
 Class :character   Class :character   Class :character   1st Qu.:7.60  
 Mode  :character   Mode  :character   Mode  :character   Median :7.80  
                                                          Mean   :7.81  
                                                          3rd Qu.:8.00  
                                                          Max.   :8.70  
    DBO.E              DQO.E              SS.E               SSV.E        
 Length:526         Length:526         Length:526         Length:526        
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  


    SED.E               COND.E          PH.P         DBO.P          
 Length:526         Min.   : 651   Min.   :7.30   Length:526        
 Class :character   1st Qu.:1200   1st Qu.:7.70   Class :character  
 Mode  :character   Median :1406   Median :7.80   Mode  :character  
                    Mean   :1477   Mean   :7.83                     
                    3rd Qu.:1672   3rd Qu.:8.00                     
                    Max.   :3230   Max.   :8.50                     
      SS.P          SSV.P              SED.P              COND.P       
 Min.   : 104   Length:526         Length:526         Min.   : 646  
 1st Qu.: 184   Class :character   Class :character   1st Qu.:1217  
```

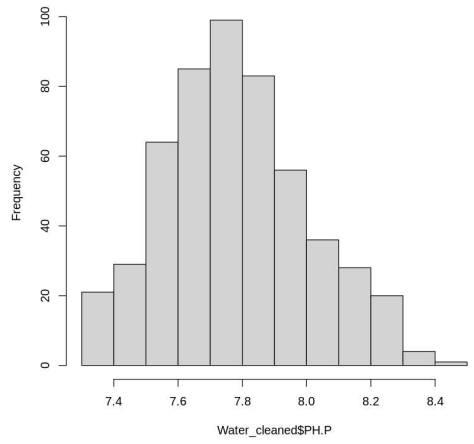
```
1 fivenum(Water_cleaned$PH.P)
```

7.3 · 7.7 · 7.8 · 8 · 8.5

```
1 ############AFTER REMOVING THE NA. WE NOTICE THAT THE FIVENUM BY DEFAULT DOES NOT CONSIDER THE NA WHILE COMPUTING THE STATISTICS
2 fivenum(Water_cleaned$PH.P,na.rm=T)
```

7.3 · 7.7 · 7.8 · 8 · 8.5

```
1 #######WATER WITH PH VALUES BETWEEN 7.7 AND 7.8 ARE THE HIGHEST######
2 hist(Water_cleaned$PH.P)
```




```
 Length:526         Length:526         Length:526        
```

```
1 stem(Water_cleaned$PH.P)
```

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

  73 | 00000
  74 | 0000000000000000
  75 | 00000000000000000000000000000
  76 | 0000000000000000000000000000000000000000000000000000000000000000000000000
  77 | 00000000000000000000000000000000000000000000000000000000000000000000000000000+5
  78 | 000000000000000000000000000000000000000000000000000000000000000000000000000000+19
  79 | 0000000000000000000000000000000000000000000000000000000000000000000000000000000+3
  80 | 000000000000000000000000000000000000000000000000000000000000000000
  81 | 00000000000000000000000000000000000
  82 | 0000000000000000000000000000000
  83 | 00000000000000000000
  84 | 0000
  85 | 0
```

```
1 #STEM PLOT WITH MORE DEPTH
2 stem(Water_cleaned$PH.P,scale=0.2)
```

```
  The decimal point is at the |

  7 | 333334444444444444444
  7 | 555555555555555555555555555566666666666666666666666666666666666666666+280
  8 | 000000000000000000000000000000000000000000000000000000000000111111111111+64
  8 | 5
```

```
1 class(Water_cleaned$COND.S)
```

'character'

```
1 class(Water_cleaned$SS.D)
```

'character'

```
1
```

```
1 class(Water_cleaned$PH.P)
```

'numeric'

```
1
2
3
```

```
1 ########WHY ARE WE GETTING THIS ERROR##########################3
2 PH_P<- Water_cleaned$PH.P
3 PH_S<- Water_cleaned$PH.S
4 boxplot(PH_P,PH_S, names = c("PH_P", "PH_S"), main="Box plot of different variables ")
```

```
  Error in x[floor(d)] + x[ceiling(d)]: non-numeric argument to binary operator
  Traceback:

  1. boxplot(PH_P, PH_S, names = c("PH_P", "PH_S"), main = "Box plot of different variables ")
  2. boxplot.default(PH_P, PH_S, names = c("PH_P", "PH_S"), main = "Box plot of different variables ")
  3. boxplot.stats(unclass(groups[[i]]), range)
  4. stats::fivenum(x, na.rm = TRUE)
```

EXPLAIN ERROR

```
1 #########ANSWER IS THE PRESENCE OF THESE SPECIAL CHARACTERS IN THE DATASET WITH A QUESTION MARK. These must be eliminated.
2 class(Water_cleaned$PH.S)
```

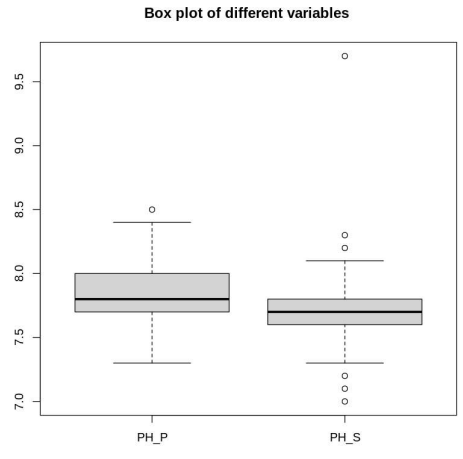
'character'

```
1 # Remove question marks from the 'Value' column
2 Water_cleaned$PH.S <- gsub("\\?", "", Water_cleaned$PH.S)
3
```

```
# Convert the 'Value' column back to numeric (if needed)
Water_cleaned$PH.S <- as.numeric(Water_cleaned$PH.S)
class(Water_cleaned$PH.S)

```

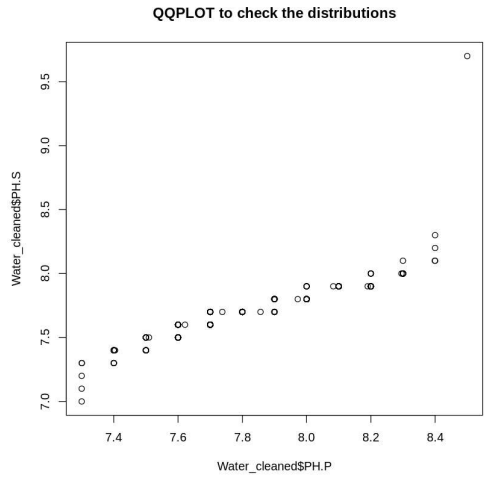
'numeric'

```
########NOW WE DONT GET THAT ERROR. WE ALSO SEE THAT THESE TWO VARIABLES HAVE DIFFERENT DISTRIBUTIONS AND BOTH HAVE C
PH_P<- Water_cleaned$PH.P
PH_S<- Water_cleaned$PH.S
boxplot(PH_P,PH_S, names = c("PH_P", "PH_S"), main="Box plot of different variables ")
```

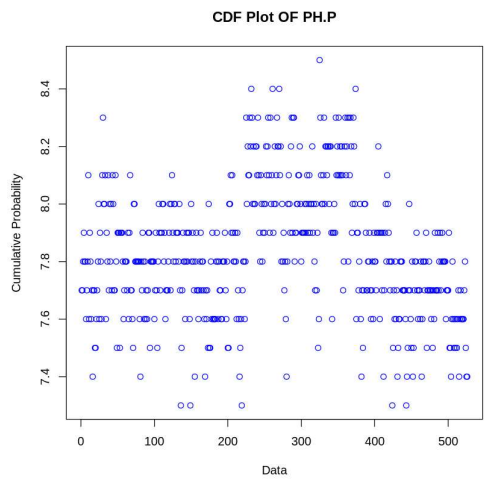



```
#If both sets of quantiles came from the same distribution, then the points on the plot should roughly form a straight diagonal line.
```

```
#######WE SEE THAT BOTH ARE FROM DIFFERENT DISTRIBUTIONS#################
qqplot(Water_cleaned$PH.P, Water_cleaned$PH.S,main="QQPLOT to check the distributions")
```



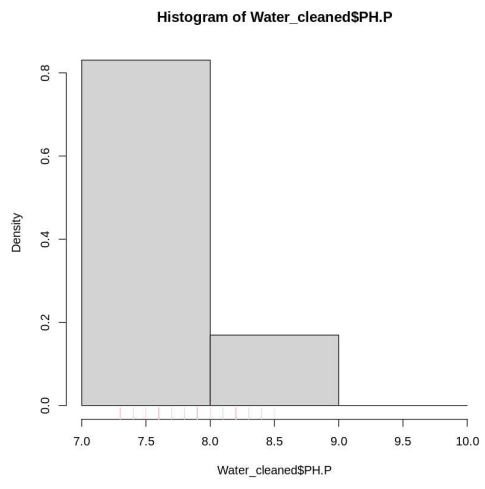

```
#CDF plot will show the cumulative distribution of the data, allowing to understand how the data values are
#spread across the entire range and what proportion of data falls below or equal to specific values.
#########we see that majority of the values for this variable are at 7.8
plot(Water_cleaned$PH.P, main = "CDF Plot OF PH.P", xlab = "Data", ylab = "Cumulative Probability", col = "blue")
```



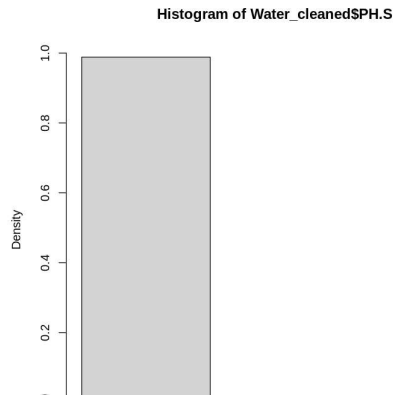

```
data_range <- range(Water_cleaned$PH.P)
print(data_range)
```

```
[1] 7.3 8.5
```

```
########WE NEED TO GET THE RANGE OF VALUES BEFORE WE PLOT THE RUG PLOT###########

hist(Water_cleaned$PH.P, seq(7., 10., 1.0), prob = TRUE)
rug(Water_cleaned$PH.P,col="Pink")
```




```
hist(Water_cleaned$PH.S, seq(7., 10., 1.0), prob = TRUE)
rug(Water_cleaned$PH.S,col="Pink")
```




```
1 ################FILTERING#################
2 #FILTER OUTPUT PH VALUE:KEEP ONLY HIGHER PH VALUES
```

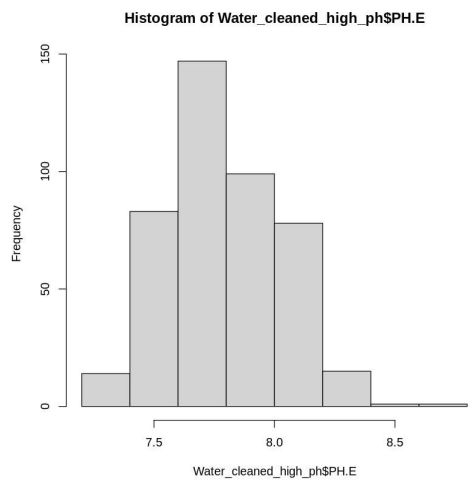
```
1 filterdf_density <- Water_cleaned%>% filter(PH.S>7.5)
2
3 # Print filtered data frame
4 filterdf_density
```

A data.frame: 438 × 39

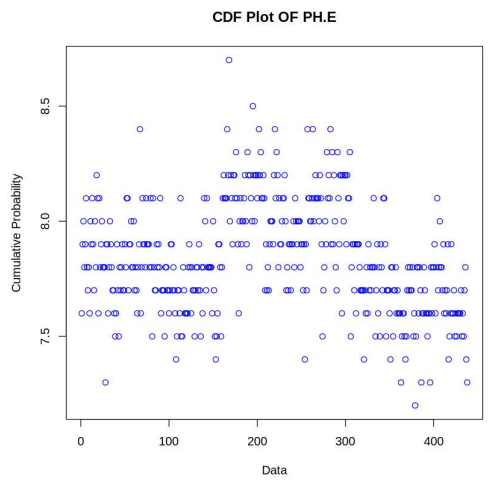
	DATE	Q.E	ZN.E	PH.E	DBO.E	DQO.E	SS.E	SSV.E	SED.E	COND.E	…	COND.S	RD.DBO.P	RD.SS.P	RD.SED.P	RD.DBO.S
	<chr>	<chr>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	…	<chr>	<chr>	<chr>	<chr>	<chr>
1	D-4/3/90	32229	5.00	7.6	?	528	186	69.9	3.4	1666	…	1888	?	58.2	95.6	?
2	D-5/3/90	35023	3.50	7.9	205	588	192	65.6	4.5	2430	…	1840	33.1	64.2	95.3	87.3
3	D-6/3/90	36924	1.50	8.0	242	496	176	64.8	4.0	2110	…	2120	?	62.7	95.6	?
4	D-12/3/90	39246	2.00	7.8	172	506	200	69.0	5.0	1865	…	1616	21.2	59.7	90.8	88.4
5	D-13/3/90	42393	0.70	7.9	189	478	230	67.0	5.5	1410	…	1575	0.6	45.8	92.0	11.6
6	D-16/3/90	40376	?	8.1	204	333	174	67.8	3.0	2390	…	2550	41.1	50.0	84.0	76.5
7	D-19/3/90	43830	1.5	7.8	177	512	214	58.9	5.5	1605	…	1110	28.0	72.7	92.7	78.8
8	D-26/3/90	47665	1.20	7.7	156	405	200	74.0	4.0	1856	…	1706	28.1	50.0	91.4	85.2
9	D-27/3/90	44314	3.00	7.8	155	389	308	49.4	6.0	1927	…	1869	52.0	64.9	90.8	87.6
10	D-28/3/90	40841	1.00	7.6	179	389	168	69.0	3.5	1240	…	1416	26.7	66.2	95.0	89.2
11	D-29/3/90	41157	3.00	8.0	145	398	192	66.7	4.5	2240	…	2290	34.3	65.0	94.2	89.3
12	D-30/3/90	40078	1.40	7.9	198	464	228	64.9	4.6	1431	…	1475	27.2	67.6	97.3	90.4
13	D-5/2/90	37312	1.00	8.1	205	492	192	70.8	4.0	1454	…	1275	?	33.0	94.5	85.3
14	D-9/2/90	36332	3.50	7.9	120	455	184	67.4	4.0	1224	…	1420	18.0	52.1	96.4	85.7

```
1 Water_cleaned_high_ph <- na.omit(filterdf_density)
```

```
1 hist(Water_cleaned_high_ph$PH.E)
```




```
1 plot(Water_cleaned_high_ph$PH.E, main = "CDF Plot OF PH.E", xlab = "Data", ylab = "Cumulative Probability", col = "bl
```



15	29/10/91															
16	D-30/10/91	31524	1.60	7.9	?	478	204	64.7	6.0	1798	…	1568	?	43.9	65.3	?
17	D-1/8/91	29834	3.00	7.4	160	348	194	61.9	3.0	1720	…	1772	29.1	53.5	92.0	85.7
18	D-2/8/91	28492	2.60	7.5	124	281	172	66.3	3.0	1520	…	1549	34.2	47.7	99.0	84.4
19	D-5/8/91	29719	0.20	7.6	133	284	186	71.0	5.0	1114	…	1100	55.1	73.2	96.0	82.0
20	D-6/8/91	29741	0.45	7.9	151	316	196	64.3	2.5	948	…	951	44.8	64.5	95.0	82.2
21	D-7/8/91	29027	0.40	7.6	136	328	186	67.7	3.0	899	…	898	?	64.1	94.0	84.9
22	D-8/8/91	30211	0.50	7.6	114	521	506	44.3	7.5	866	…	884	48.7	86.9	98.8	81.0
23	D-9/8/91	30848	0.20	7.7	142	376	144	70.8	3.0	940	…	947	?	47.6	97.3	?
24	D-11/8/91	17527	0.55	7.5	150	171	172	37.2	1.4	732	…	728	62.8	77.7	90.0	73.8
25	D-12/8/91	33331	0.23	7.6	92	233	234	37.6	1.4	829	…	929	36.9	43.6	86.7	87.7
26	D-13/8/91	27998	0.62	7.5	138	268	154	66.2	1.7	890	…	858	38.1	41.6	86.7	87.7