

EXP1.The intervals and corresponding frequencies are as follows. age frequency

1-5. 200

5-15 450

15-20 300

20-50 1500

50-80 700

80-110 44

Compute an approximate median value for the data

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for calculating the median of grouped data.
- Environment:** Displays the global environment with variables: marks, math_scor..., max_val, mean_age, median_age, median_cl..., and median_va....
- Console:** Shows the execution of the R script, resulting in the output [1] 32.94.

```
1 intervals <- c("1-5", "5-15", "15-20", "20-50", "50-80", "80-110")
2 frequencies <- c(200, 450, 300, 1500, 700, 44)
3 cumulative_freq <- cumsum(frequencies)
4 N <- sum(frequencies)
5 median_class_index <- which(cumulative_freq >= N/2)[1]
6 L <- c(1, 5, 15, 20, 50, 80)[median_class_index]
7 CF <- ifelse(median_class_index == 1, 0, cumulative_freq[median_class_index - 1])
8 f <- frequencies[median_class_index]
9 h <- c(4, 10, 5, 30, 30, 30)[median_class_index]
10 median_value <- L + (((N/2) - CF) / f) * h
11 print(median_value)
```

Console output:

```
> source("C:/Users/sri/Downloads/1.R")
[1] 32.94
>
```

EXP 2. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

- (a) What is the mean of the data? What is the median?
- (b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
- (c) What is the midrange of the data?
- (d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for calculating mean, median, mode, midrange, and quartiles from a vector named 'ages'.
- Environment:** Displays the results of the calculations in the Global Environment.
- Console:** Shows the execution of the code and the resulting values for the statistical measures.

R Code in Source Editor:

```
4 mean_age <- mean(ages)
5 median_age <- median(ages)
6
7 # Mode
8 mode_age <- as.numeric(names(sort(table(ages), decreasing = TRUE)[1]))
9
10 # Midrange
11 midrange <- (min(ages) + max(ages)) / 2
12
13 # Q1 & Q3
14 Q1 <- quantile(ages, 0.25)
15 Q3 <- quantile(ages, 0.75)
16
17
```

Environment Panel:

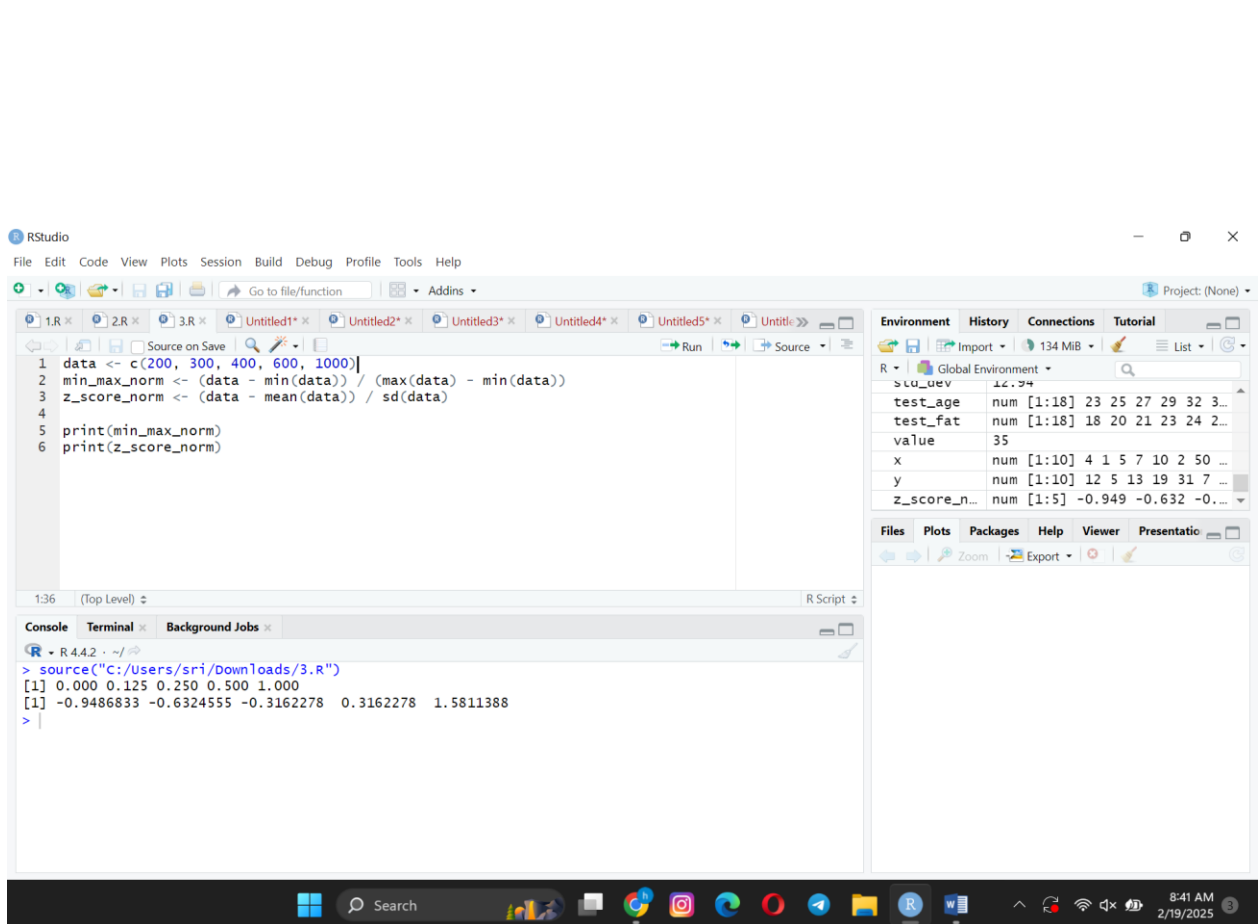
Variable	Value
mean_age	29.96296
median_age	25
mode_age	25
midrange	41.5
Q1	20.5
Q3	35

Console Output:

```
> source("C:/Users/sri/Downloads/2.R")
25% 75%
29.96296 25.00000 25.00000 41.50000 20.50000 35.00000
>
```

EXP 3.Data Preprocessing :Reduction and Transformation

Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000 (a) min-max normalization by setting min = 0 and max = 1 (b) z-score normalization



The screenshot displays the RStudio interface. The script editor on the left contains the following R code:

```
1 data <- c(200, 300, 400, 600, 1000)
2 min_max_norm <- (data - min(data)) / (max(data) - min(data))
3 z_score_norm <- (data - mean(data)) / sd(data)
4
5 print(min_max_norm)
6 print(z_score_norm)
```

The console at the bottom shows the output of the code:

```
R - R 4.4.2 - ~/
> source("C:/Users/sri/Downloads/3.R")
[1] 0.000 0.125 0.250 0.500 1.000
[1] -0.9486833 -0.6324555 -0.3162278  0.3162278  1.5811388
>
```

The Environment pane on the right shows the objects created in the Global Environment:

Object	Class	Value
test_age	num	[1:18] 23 25 27 29 32 3...
test_fat	num	[1:18] 18 20 21 23 24 2...
value	num	35
x	num	[1:10] 4 1 5 7 10 2 50 ...
y	num	[1:10] 12 5 13 19 31 7 ...
z_score_n...	num	[1:5] -0.949 -0.632 -0....

EXP 4.Data:11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71, 72,73,75

- a) Smoothing by bin mean
- b) Smoothing by bin median
- c) Smoothing by bin boundaries

The screenshot shows the RStudio interface with a script editor, environment pane, and console. The script defines a vector `data_exp4` and performs bin smoothing using `lapply` and `unlist`. The console displays the resulting smoothed data.

```
1 data_exp4 <- c(11, 13, 13, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 45, 71, 72, 73, 75)
2 bins <- split(sort(data_exp4), cut(sort(data_exp4), 4, labels = FALSE))
3 bin_means <- sapply(bins, mean)
4 bin_mean_smoothing <- unlist(lapply(bins, function(b) rep(mean(b), length(b))))
5 bin_medians <- sapply(bins, median)
6 bin_median_smoothing <- unlist(lapply(bins, function(b) rep(median(b), length(b))))
7 bin_boundaries_smoothing <- unlist(lapply(bins, function(b) ifelse(abs(b - min(b)) < abs(b - m
8 print(bin_mean_smoothing)
9 print(bin_median_smoothing)
10 print(bin_boundaries_smoothing)
```

Console Output:

```
R - R 4.4.2 - ~/>
18.20 18.20 18.20 18.20 18.20 18.20 18.20 18.20 18.20 18.20 18.20 18.20 18.20 18.20
21 22 31 32 33 41 42 43 44
35.00 35.00 45.00 45.00 45.00 72.75 72.75 72.75 72.75
11 12 13 14 15 16 17 18 19 110 111 112 113 114 115 21 22 31
20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 35.0 35.0 45.0
32 33 41 42 43 44
45.0 45.0 72.5 72.5 72.5 72.5
11 12 13 14 15 16 17 18 19 110 111 112 113 114 115 21 22 31 32 33 41 42 43
11 11 11 11 11 24 24 24 24 24 24 24 24 24 24 30 40 45 45 45 71 71 75
44
75
>
```

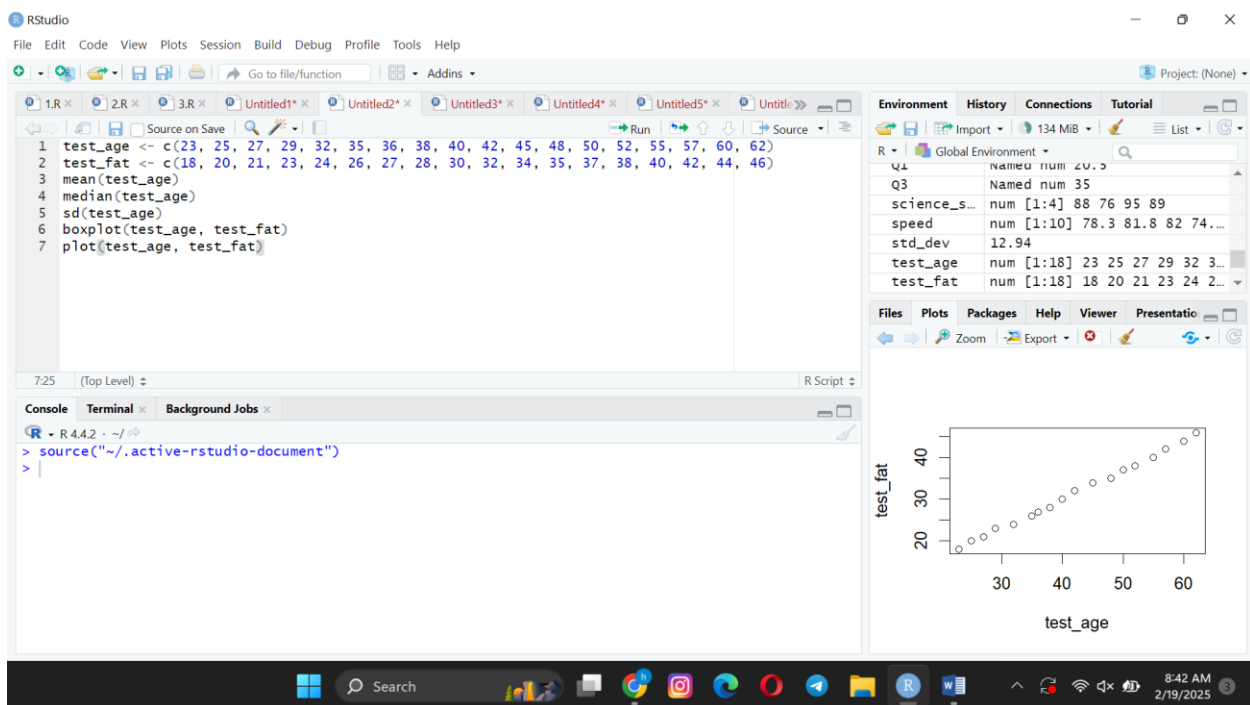
Environment Pane:

Object	Class	Value
data_exp4	num [1:24]	11 13 13 15 15 1...
decimal_s...	0.35	
english_s...	num [1:4]	90 85 80 87
f	1500	
frequenci...	num [1:6]	200 450 300 1500 ...
h	30	
intervals	chr [1:6]	"1-5" "5-15" "15-...

5. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

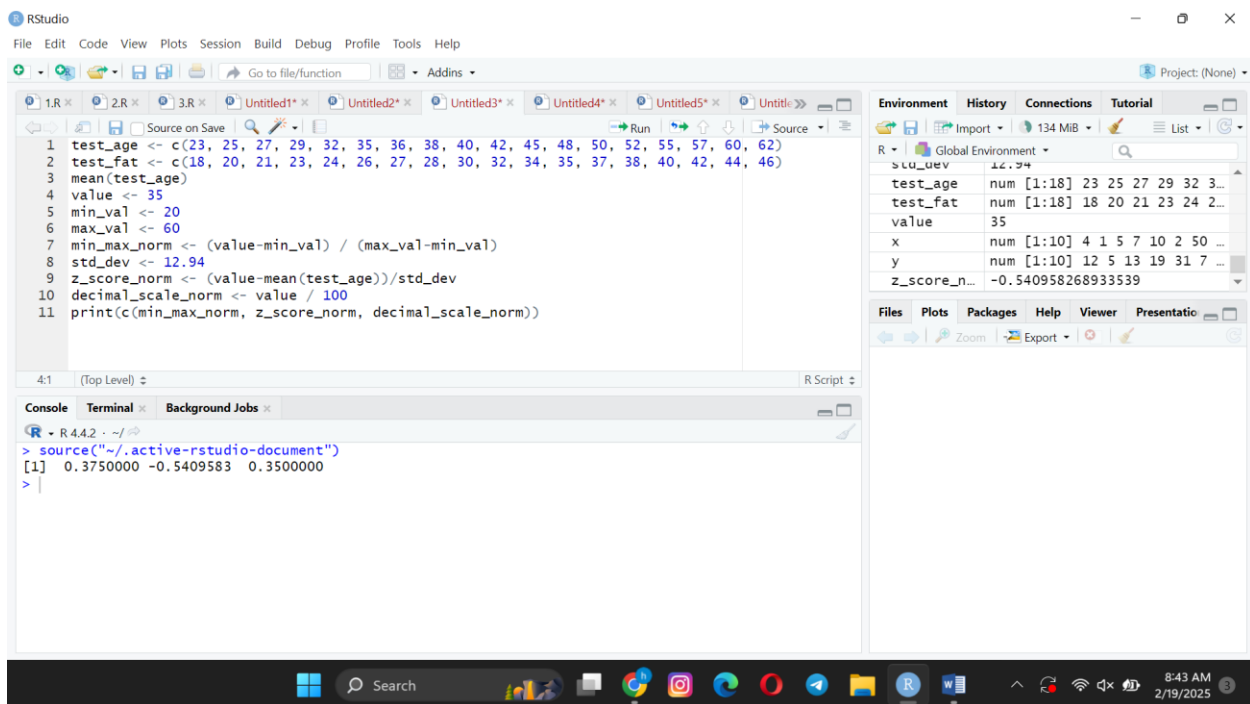
age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

- Calculate the mean, median, and standard deviation of age and %fat.
- Draw the boxplots for age and %fat.
- Draw a scatter plot and a q-q plot based on these two variables.



6. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

- (i) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
- (ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.
- (iii) Use normalization by decimal scaling to transform the value 35 for age. Perform the above functions using R – tool



The screenshot shows the RStudio interface. The script editor contains the following R code:

```
1 test_age <- c(23, 25, 27, 29, 32, 35, 36, 38, 40, 42, 45, 48, 50, 52, 55, 57, 60, 62)
2 test_fat <- c(18, 20, 21, 23, 24, 26, 27, 28, 30, 32, 34, 35, 37, 38, 40, 42, 44, 46)
3 mean(test_age)
4 value <- 35
5 min_val <- 20
6 max_val <- 60
7 min_max_norm <- (value-min_val) / (max_val-min_val)
8 std_dev <- 12.94
9 z_score_norm <- (value-mean(test_age))/std_dev
10 decimal_scale_norm <- value / 100
11 print(c(min_max_norm, z_score_norm, decimal_scale_norm))
```

The console shows the output of the code:

```
> source("~/active-rstudio-document")
[1] 0.3750000 -0.5409583 0.3500000
>
```

The Environment pane on the right shows the following objects:

Object	Class	Value
std_dev	num	12.94
test_age	num [1:18]	23 25 27 29 32 35 36 38 40 42 45 48 50 52 55 57 60 62
test_fat	num [1:18]	18 20 21 23 24 26 27 28 30 32 34 35 37 38 40 42 44 46
value	num	35
x	num [1:10]	4 1 5 7 10 2 50 ...
y	num [1:10]	12 5 13 19 31 7 ...
z_score_norm	num	-0.540958268933539

7. The following values are the number of pencils available in the different boxes. Create a vector and find out the mean, median and mode values of set of pencils in the given data.

Box1 Box2 Box3 Box4 Box5 Box6 Box7 Box8 Box9 Box 10

9 25 23 12 11 6 7 8 9 10

The screenshot shows the RStudio environment with the following components:

- Source Editor:** Contains the following R code:

```
1 pencils <- c(9, 25, 23, 12, 11, 6, 7, 8, 9, 10)
2 mean(pencils)
3 median(pencils)
4 mode_pencils <- as.numeric(names(sort(table(pencils), decreasing = TRUE)[1]))
5 print(mode_pencils)
```
- Environment Pane:** Displays the objects in the global environment:

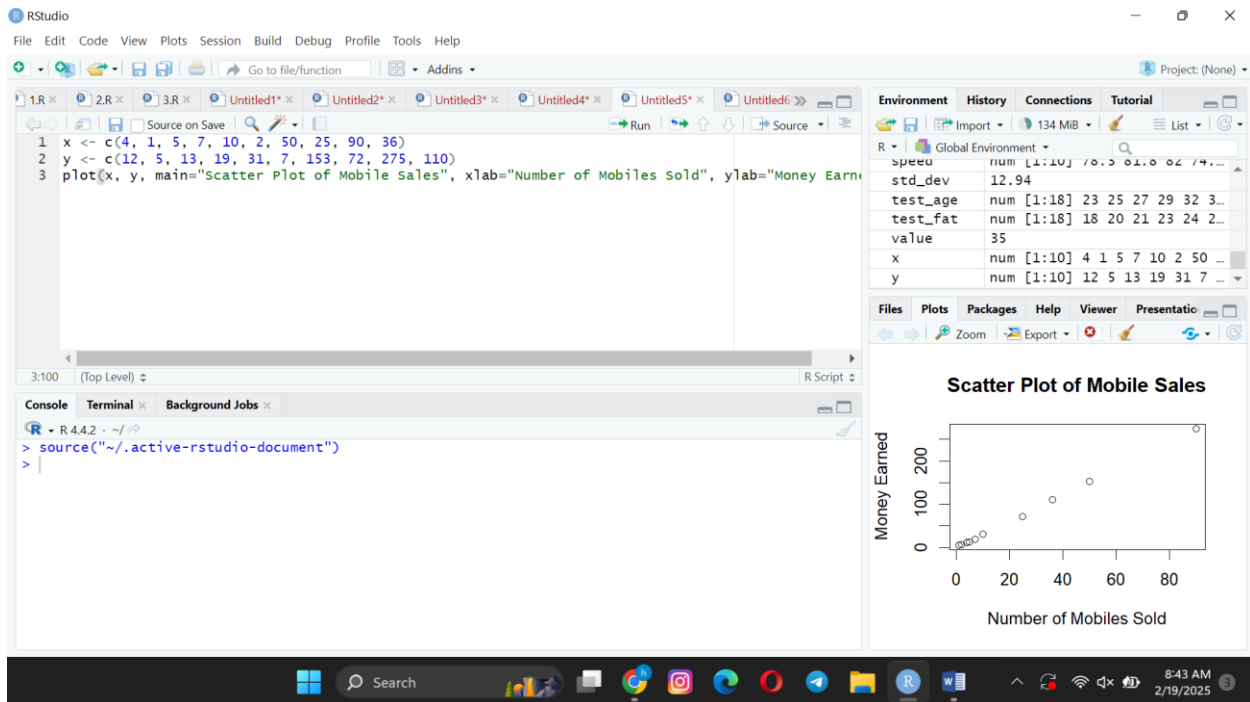
Object	Class	Value
pencils	num [1:10]	9 25 23 12 11 6 ...
product_n...	chr [1:5]	"Laptop" "Smartph...
product_p...	num [1:5]	1000 500 300 150 ...
product_q...	num [1:5]	50 150 75 200 100
Q1	Named num	20.5
Q3	Named num	35
science_s	num [1:41]	88 76 95 89
- Console:** Shows the output of the executed code:

```
> source("~/active-rstudio-document")
[1] 9
>
```

8.the following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.

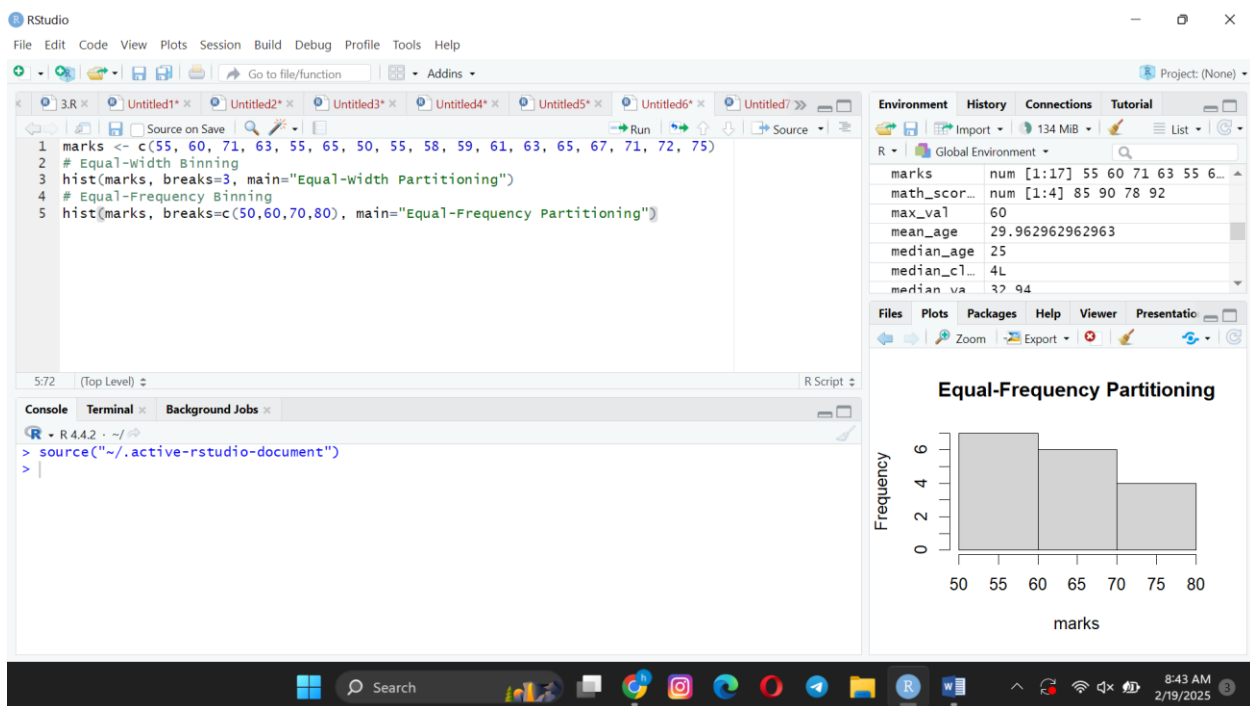
x :4 1 5 7 10 2 50 25 90 36

y :12 5 13 19 31 7 153 72 275 110



9. Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55, 58, 59, 61, 63, 65, 67, 71, 72, 75. Partition them into three bins by each of the following methods. Plot the data points using histogram.

(a) equal-frequency (equi-depth) partitioning (b) equal-width partitioning



10. Suppose that the speed car is mentioned in different driving style.

Regular 78.3 81.8 82 74.2 83.4 84.5 82.9 77.5 80.9 70.6 Speed

Calculate the Inter quantile and standard deviation of the given data

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code to calculate the first quartile (Q1), third quartile (Q3), and interquartile range (IQR) for a vector named 'speed'.
- Console:** Displays the output of the R script, showing the values for Q1, Q3, and IQR.
- Environment:** Lists the objects in the global environment, including 'product_c...', 'product_p...', 'product_q...', 'Q1', 'Q3', 'science_s...', and 'speed'.

```
1 speed <- c(78.3, 81.8, 82, 74.2, 83.4, 84.5, 82.9, 77.5, 80.9, 70.6)
2 Q1 <- quantile(speed, 0.25)
3 Q3 <- quantile(speed, 0.75)
4 IQR_value <- IQR(speed)
5 cat("Q1 (First Quartile):", Q1, "\n")
6 cat("Q3 (Third Quartile):", Q3, "\n")
7 cat("Interquartile Range (IQR):", IQR_value, "\n")
8
```

Console Output:

```
> source("~/active-rstudio-document")
Q1 (First Quartile): 77.7
Q3 (Third Quartile): 82.675
Interquartile Range (IQR): 4.975
>
```

Environment:

Object	Class	Attributes
product_c...	chr [1:5]	Laptop Smartphn...
product_p...	num [1:5]	1000 500 300 150 ...
product_q...	num [1:5]	50 150 75 200 100
Q1	Named num	77.7
Q3	Named num	82.7
science_s...	num [1:4]	88 76 95 89
speed	num [1:10]	78.3 81.8 82 74...

11. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

The screenshot shows the RStudio interface. The script editor contains the following R code:

```
1 age <- c(13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70)
2
3 Q1 <- quantile(age, 0.25)
4 Q3 <- quantile(age, 0.75)
5 print(Q1)
6 print(Q3)
7
```

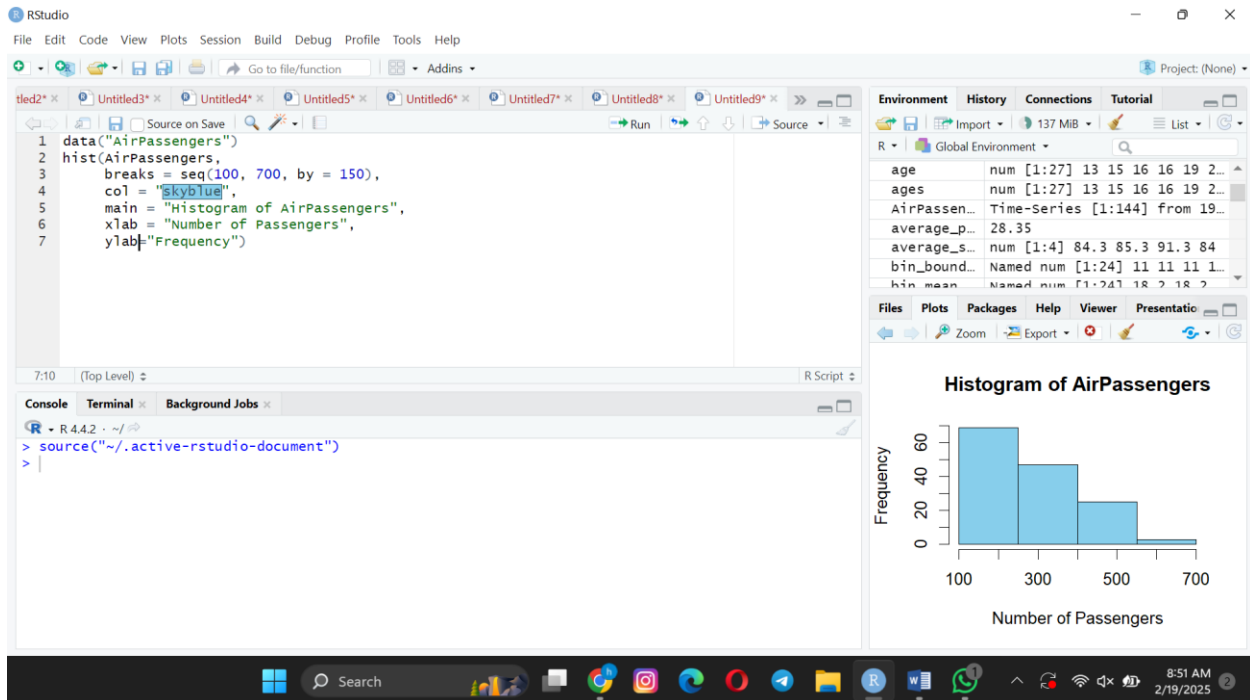
The Environment pane on the right shows the following values:

Variable	Class	Dimensions	Values
age	num	[1:27]	13 15 16 16 19 20 20 21 22 22 25 25 25 25 30 33 33 35 35 35 35 36 40 45 46 52 70
ages	num	[1:27]	13 15 16 16 19 20 20 21 22 22 25 25 25 25 30 33 33 35 35 35 35 36 40 45 46 52 70
average_p...			28.35
average_s...	num	[1:4]	84.3 85.3 91.3 84
bin_bound...	Named num	[1:24]	11 11
bin_masn	Named num	[1:24]	18 18

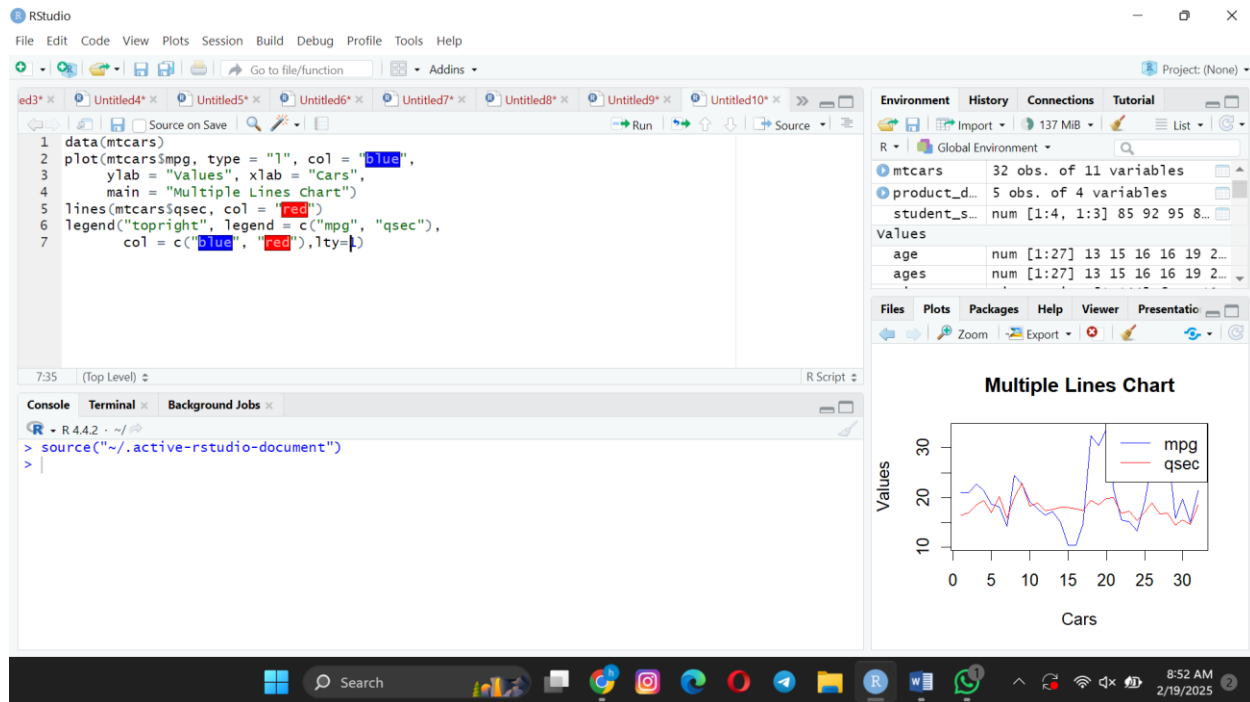
The Console pane shows the output of the R script:

```
> source("~/active-rstudio-document")
25%
20.5
75%
35
>
```

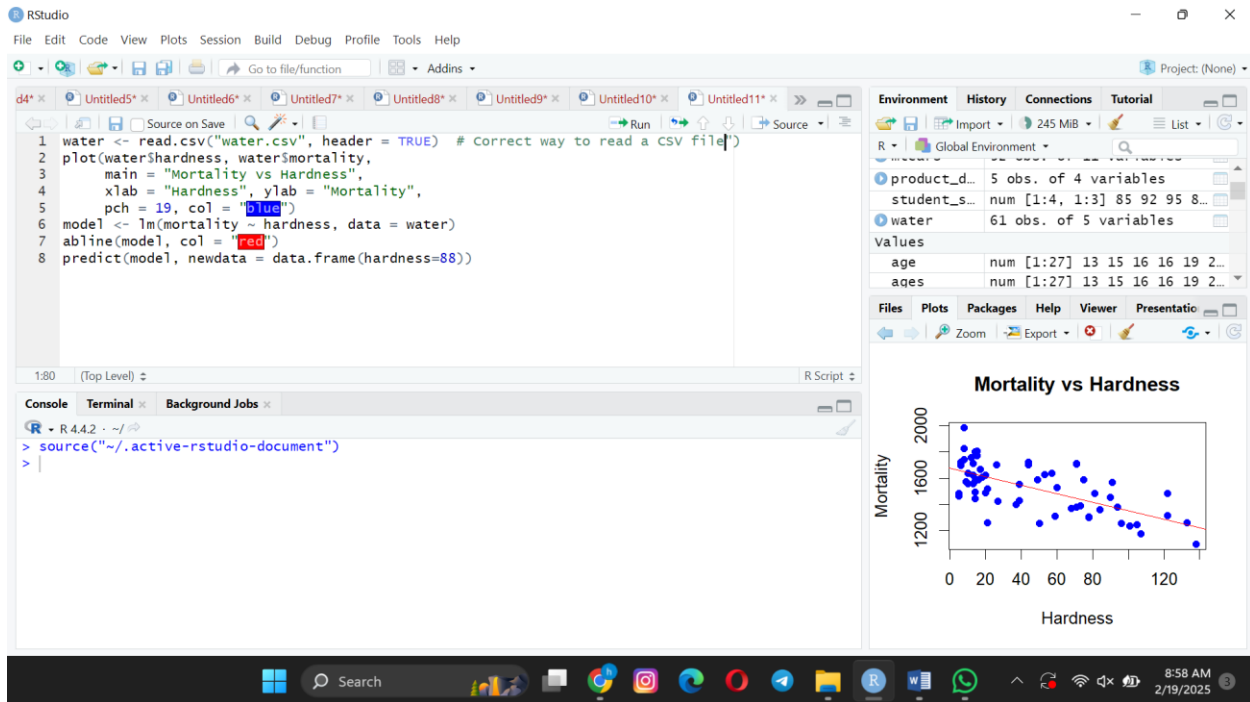
12 Make a histogram for the “AirPassengers” dataset, start at 100 on the x-axis, and from values 200 to 700, make the bins 150 wide



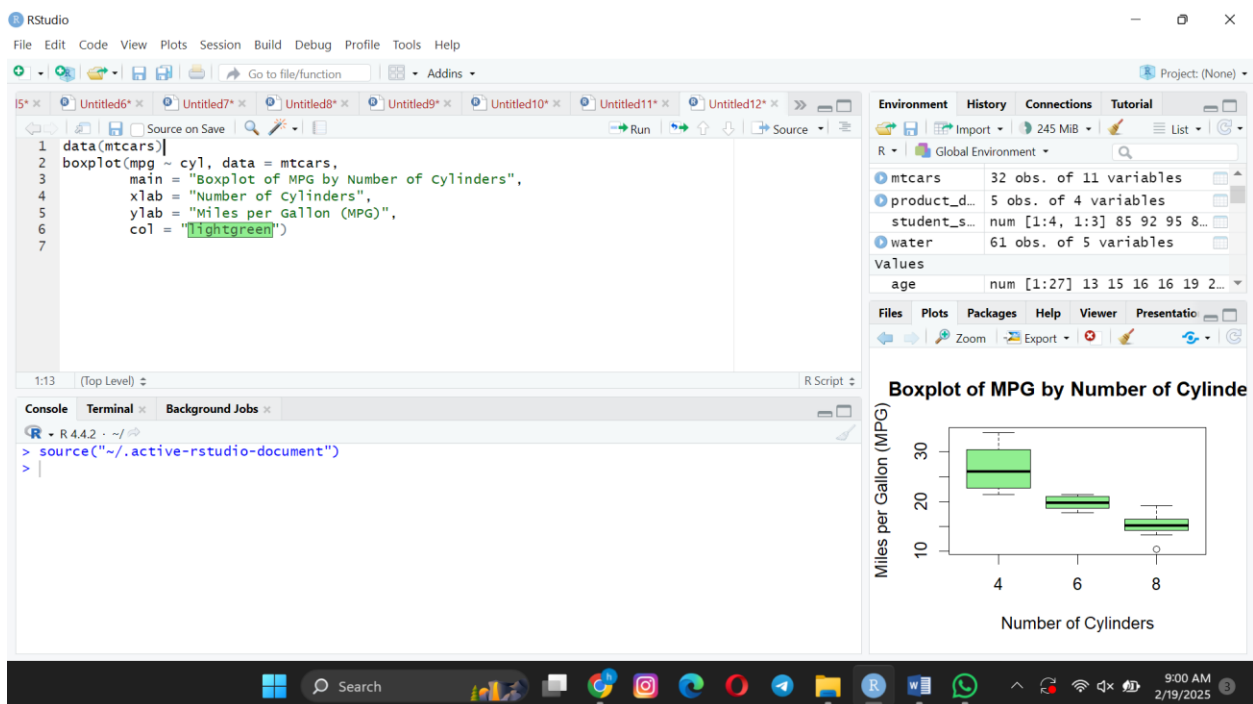
13 Obtain Multiple Lines in Line Chart using a single Plot Function in R. Use attributes “mpg” and “qsec” of the dataset “mtcars



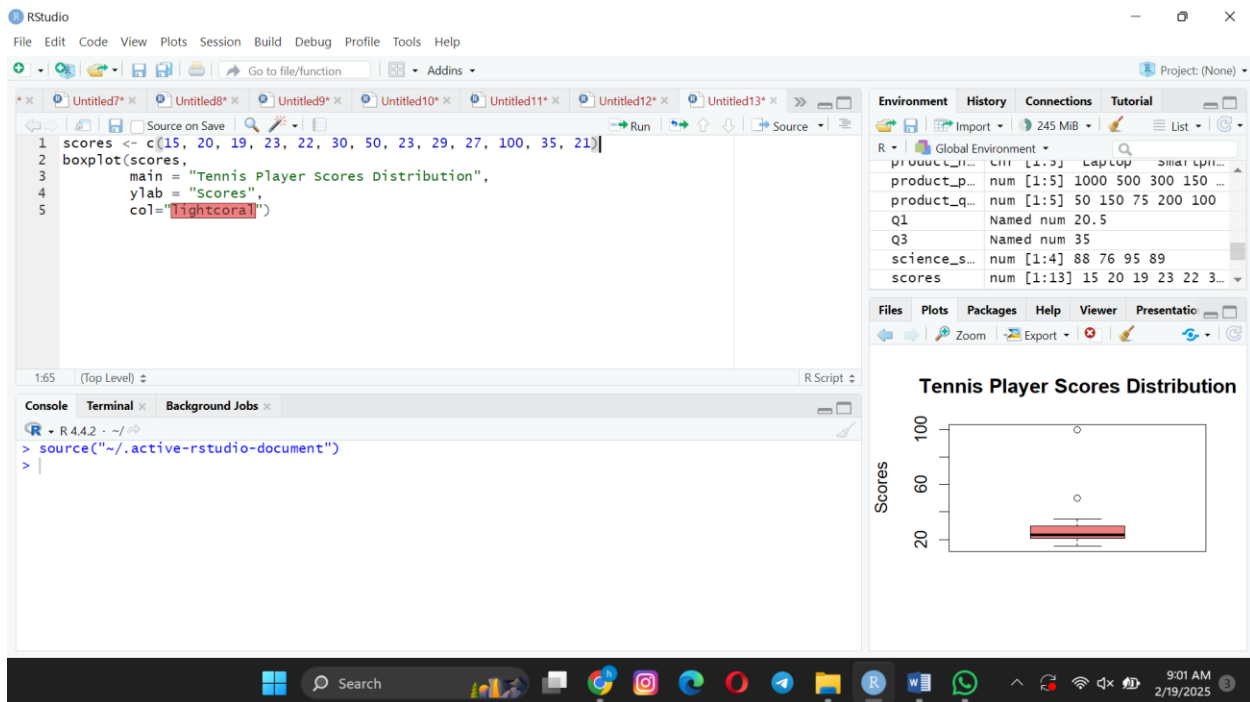
14 Download the Dataset "water" From R dataset Link. Find out whether there is a linear relation between attributes "mortality" and "hardness" by plot function. Fit the Data into the Linear Regression model. Predict the mortality for the hardness=88.



15 Create a Boxplot graph for the relation between "mpg"(miles per galloon) and "cyl"(number of Cylinders) for the dataset "mtcars" available in R Environment.



16. Assume the Tennis coach wants to determine if any of his team players are scoring outliers. To visualize the distribution of points scored by his players, then how can he decide to develop the box plot? Give suitable example using Boxplot visualization technique.



17 Implement using R language in which age group of people are affected by blood pressure based on the diabetes dataset show it using scatterplot and bar chart (that is BloodPressure vs Age using dataset “diabetes.csv”)

