

HW1

Name: Naga Satya Silpa Annadevara

Student ID: A20517818

Use the data set in case study 2 and utilize R to answer the following questions.

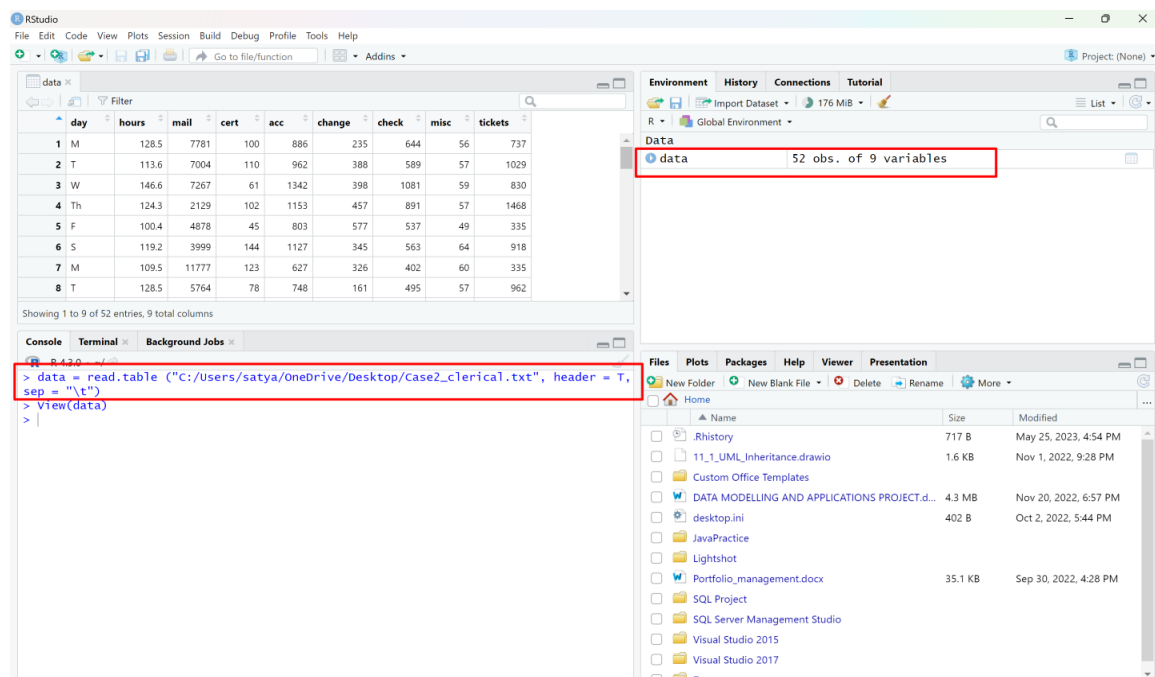
Note: you need to take snapshots of your R coding & outputs, also use your texts to answer the following questions (if necessary). Just like the example given in our slides. Upload a single PDF document as your submission.

1. Load dataset into R and show the column names, example of values, and the size of the data [10]

ANSWER:

Loading data set:

- Case study 2(Case2_clerical.txt) is loaded in R Studio by using the command:
- `data = read.table ("C:/Users/satya/OneDrive/Desktop/Case2_clerical.txt", header = T, sep = "\t")`
- **Note:** I used separator as “\t” , because I noticed that the separator in the text file you have given is not coma(,).
- The separator between the columns in the text file is space. So, we need to use \t as separator.



Show the column names:

The screenshot shows the RStudio interface. In the Console, the following commands have been executed:

```
> data = read.table ("C:/Users/satya/OneDrive/Desktop/Case2_clerical.txt", header = T, sep = "\t")
> View(data)
>
>
>
> names(data)
[1] "day"      "hours"    "mail"     "cert"     "acc"      "change"   "check"
[8] "misc"     "tickets"
```

The output of `names(data)` is displayed in the Console, showing the column names of the data frame. The Environment pane on the right shows the data frame 'data' with 52 observations and 9 variables.

Example of values:

The example of values here are : "M" "T"

128, 114etc.,

The screenshot shows the RStudio interface. In the Console, the following commands have been executed:

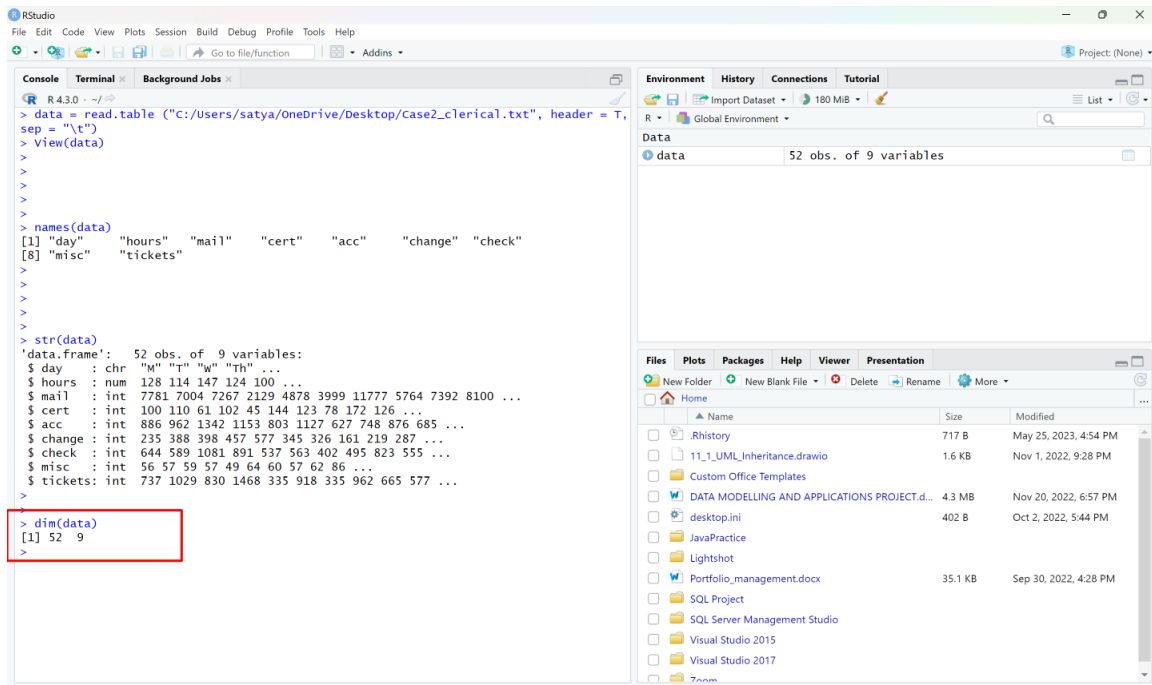
```
> data = read.table ("C:/Users/satya/OneDrive/Desktop/Case2_clerical.txt", header = T, sep = "\t")
> View(data)
>
>
>
> names(data)
[1] "day"      "hours"    "mail"     "cert"     "acc"      "change"   "check"
[8] "misc"     "tickets"
```

The output of `names(data)` is displayed in the Console. Below it, the `str(data)` command has been executed, showing the structure of the data frame:

```
> str(data)
'data.frame':   52 obs. of  9 variables:
 $ day   : chr  "M" "T" "W" "Th" ...
 $ hours : num  128 114 147 124 100 ...
 $ mail  : int   7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert  : int   100 110 61 102 45 144 123 78 172 126 ...
 $ acc   : int   886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change: int   235 388 398 457 577 345 326 161 219 287 ...
 $ check : int   644 589 1081 891 537 563 402 495 823 555 ...
 $ misc  : int    56 57 59 57 49 64 60 57 62 86 ...
 $ tickets: int   737 1029 830 1468 335 918 335 962 665 577 ...
```

The Environment pane on the right shows the data frame 'data' with 52 observations and 9 variables.

Size of the data: Size of the data can be known by using `dim` function. There are 52 observations/records & 9 variables/columns.



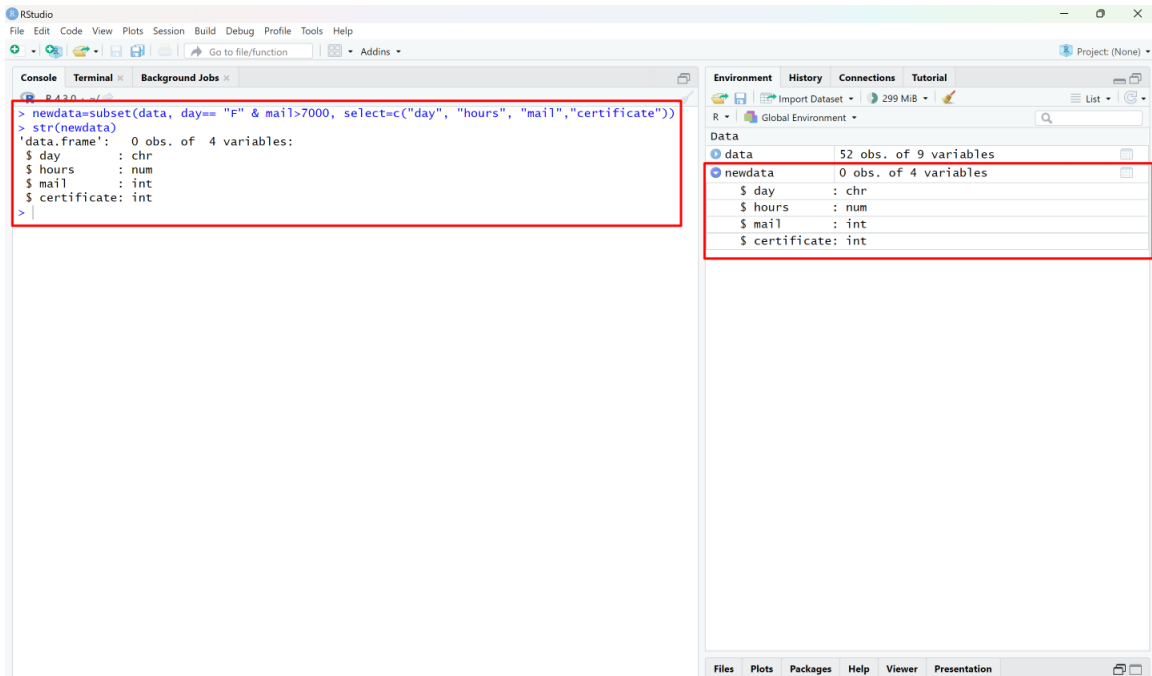
The screenshot shows the RStudio interface. The console on the left displays the following R code and its output:

```
> data = read.table("C:/Users/satya/OneDrive/Desktop/case2_clerical.txt", header = T, sep = "\t")
> View(data)
>
>
>
> names(data)
[1] "day"      "hours"    "mail"     "cert"     "acc"      "change"   "check"
[8] "misc"     "tickets"
>
> str(data)
'data.frame': 52 obs. of 9 variables:
 $ day : chr "M" "T" "W" "Th" ...
 $ hours : num 128 114 147 124 100 ...
 $ mail : int 7781 7004 7267 2129 4878 3999 11777 5764 7392 8100 ...
 $ cert : int 100 110 61 102 45 144 123 78 172 126 ...
 $ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
 $ change : int 235 388 398 457 577 345 326 161 219 287 ...
 $ check : int 644 589 1081 891 537 563 402 495 823 555 ...
 $ misc : int 56 57 59 57 49 64 60 57 62 86 ...
 $ tickets : int 737 1029 830 1468 335 918 335 962 665 577 ...
>
> dim(data)
[1] 52 9
>
```

The Environment pane on the right shows a variable named 'data' with 52 observations and 9 variables.

- Return a list of records with columns <day, hours, mail, cert>, where day is Friday, and the number of mails is larger than 7000 [10]

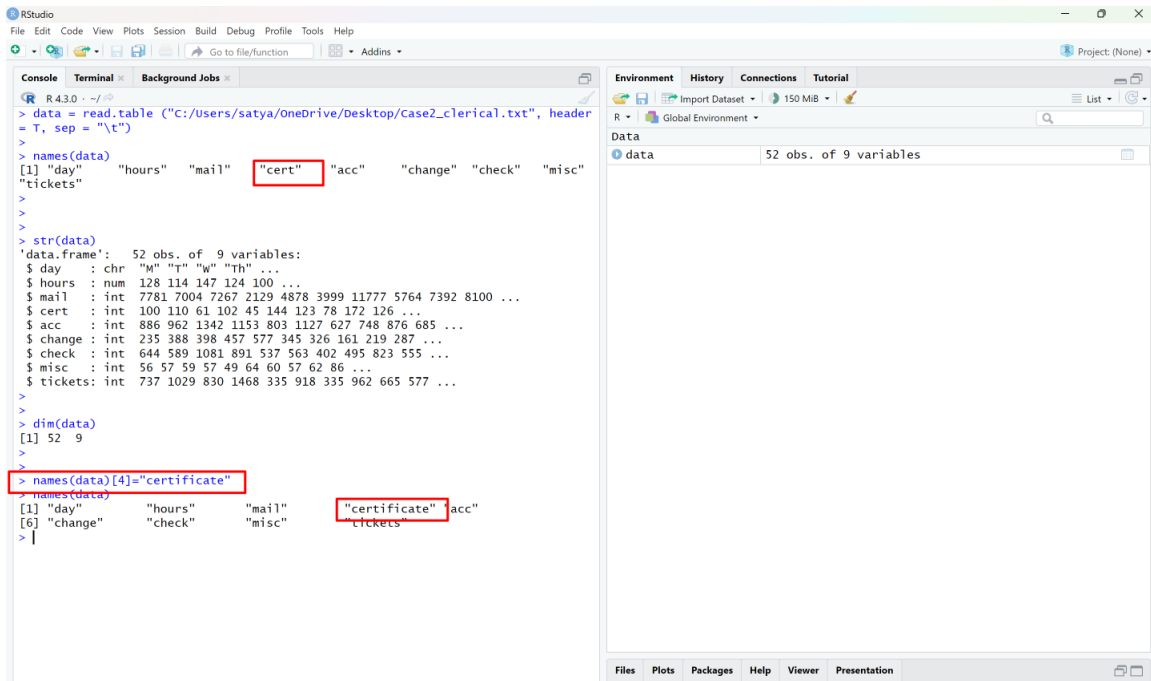
ANSWER: There are zero observations for the above constraints. Used `Subset` function.



3. Change the column name from “cert” to “certificate” [5]

ANSWER:

The column name changed from "cert" to "certificate" using index:



4. Use descriptive statistics to understand the column “day”. More specifically, return class frequency, and class relative frequency. Also visualize this column by using bar graph (use class relative frequency as y-axis) and pie chart [20]

ANSWER: Installed and loaded library ‘plyr’ to use and call the functions in it and to calculate the Class frequency and Class relative Frequency.

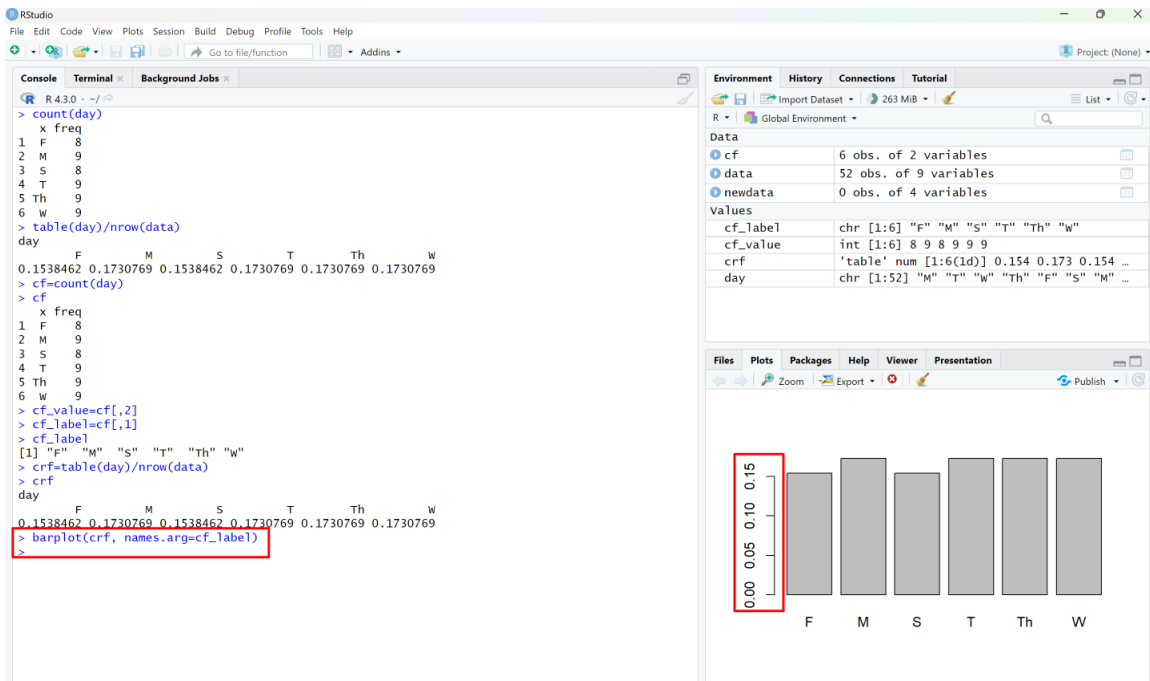
Class frequency (CF) of the column ‘day’ using **count** function & CRF is also calculated using R. Here is the screenshot.

```
R 4.3.0 - ~/
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Adding
Console Terminal Background Jobs
$ acc : int 886 962 1342 1153 803 1127 627 748 876 685 ...
$ change : int 235 388 398 457 577 345 326 161 219 287 ...
$ check : int 644 589 1081 891 537 563 402 495 823 555 ...
$ misc : int 56 57 59 57 49 64 60 57 62 86 ...
$ tickets : int 737 1029 830 1468 335 918 335 962 665 577 ...
> options(warn = 0)
> install.packages("plyr")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/satya/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/plyr_1.8.8.zip'
Content type 'application/zip' length 1162905 bytes (1.1 MB)
downloaded 1.1 MB
package 'plyr' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
C:/Users/satya/AppData/Local/Temp/Rtmpv9AOZ/downloaded_packages
> library(plyr)
> day
[1] "M" "S" "W" "Th" "F" "S" "M" "T" "W" "F" "S" "M" "T" "W" "Th"
[17] "F" "S" "M" "T" "Th" "F" "S" "M" "T" "W" "Th" "F" "S" "M" "T"
[33] "W" "Th" "F" "S" "M" "T" "W" "Th" "F" "S" "M" "T" "W" "Th"
[49] "M" "T" "W" "Th"
> count(day)
  x freq
1 F 8
2 M 9
3 S 8
4 T 9
5 Th 9
6 W 9
> table(day)/nrow(data)
      day
      F      M      S      T      Th      W
0.1538462 0.1730769 0.1538462 0.1730769 0.1730769 0.1730769
> |
```

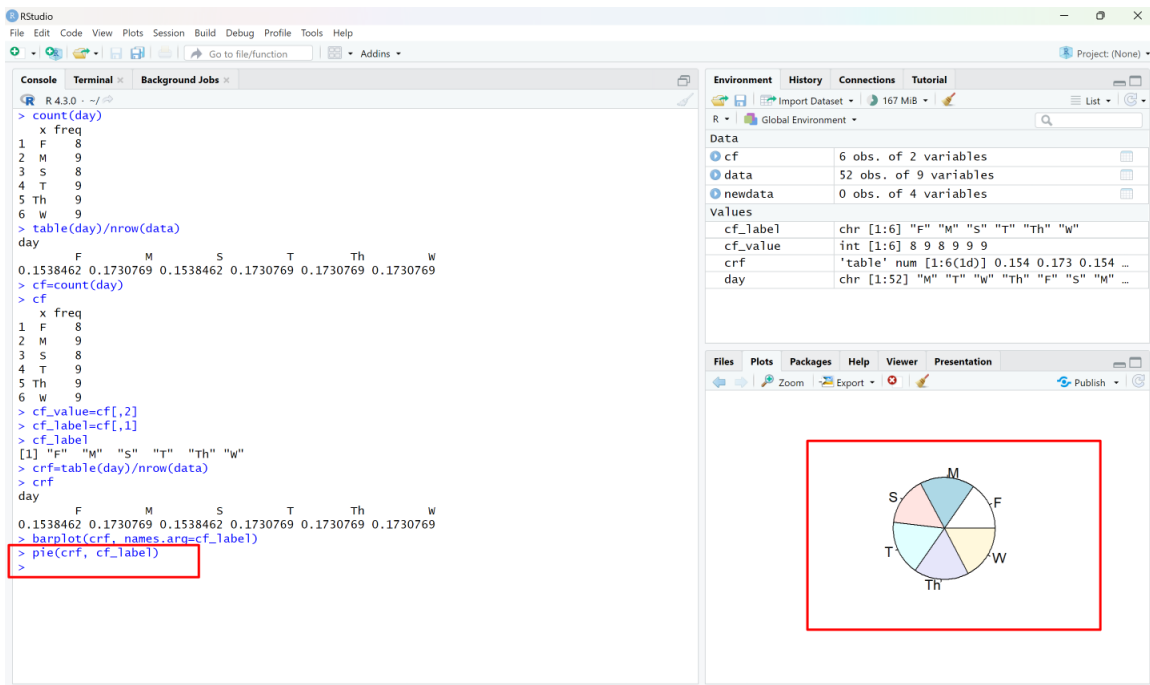
Class relative frequency of the column ‘day’: (CRF=CF/n) (manual calculation by using formula)

- CRF of F = $CF/n = 8/52 = 0.153846$
- CRF of M = $CF/n = 9/52 = 0.1730769$
- CRF of S = $CF/n = 8/52 = 0.153846$
- CRF of T = $CF/n = 9/52 = 0.1730769$
- CRF of Th = $CF/n = 9/52 = 0.1730769$
- CRF of W = $CF/n = 9/52 = 0.1730769$

Visualizing the 'day' column by bar-graph using class relative frequency on y-axis:



Visualizing the 'day' column by pie-chart using class relative frequency:



5. Use descriptive statistics to understand the column “certificate”. More specifically, we want to get q1, q2, q3, average value, variance value. Also, visualize this variable by using histogram, and interpret your histogram [20]

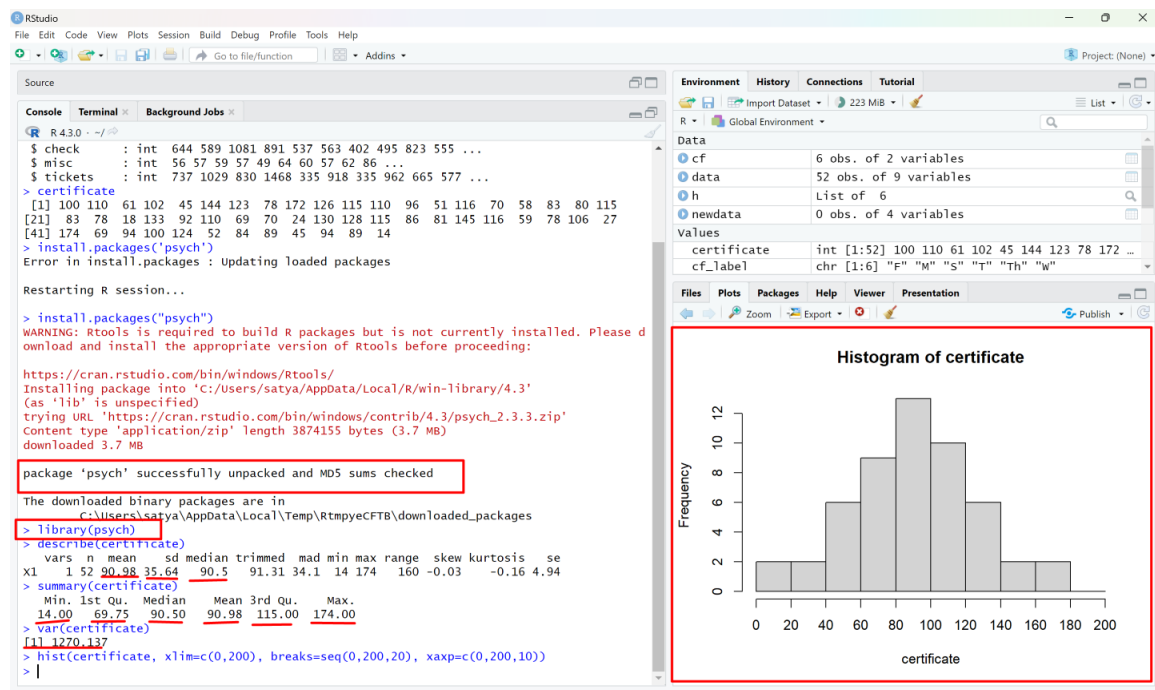
ANSWER:

To calculate the descriptive statistics asked, library ‘psych’ is installed and loaded. Q1, Q2, Q3, average value and variance value can be calculated by using 2 functions.

1. describe()
2. summary()
3. var()

The variable ‘certificate’ can be visualized by histogram by using hist() function.

Here is the screenshot of the r coding calculating the q1, q2, q3, average value(mean) and variance value and the histogram of the variable ‘certificate’:



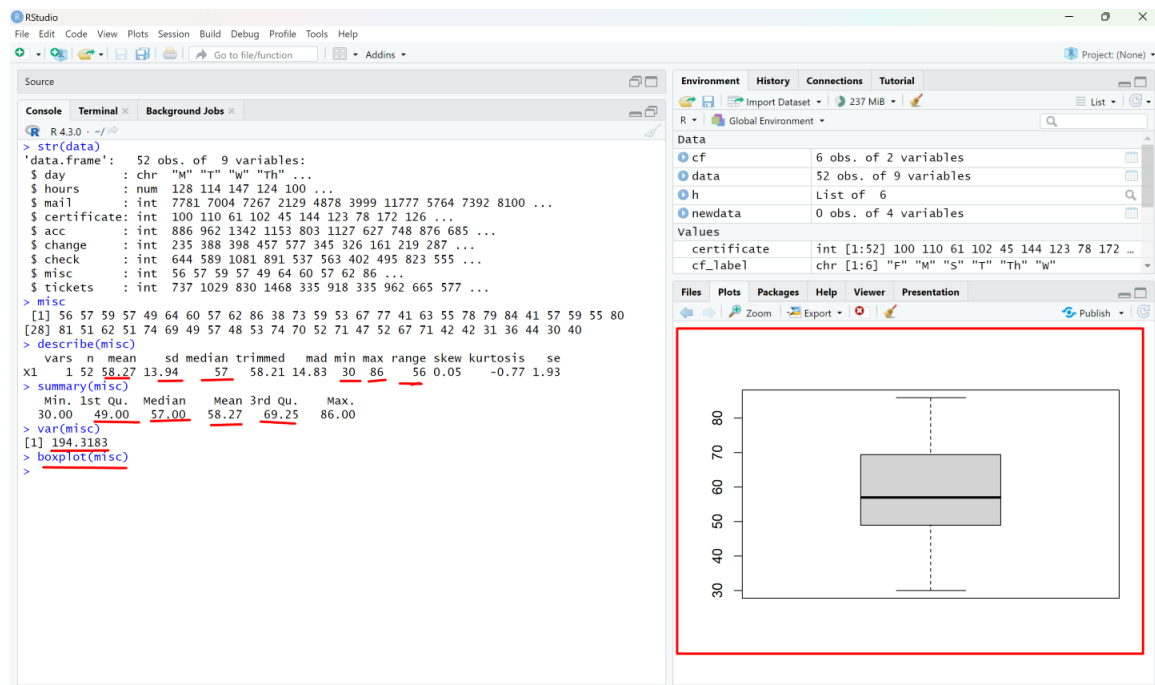
Interpretation of the histogram of certificate: (Analyzing the shape by skewness and outliers)

- **Distribution:** The distribution is “Normal distribution/ Symmetric”. There is no skew in data. Hence the data is evenly distributed around the center.
- **Variance:** Variance is nothing but the variation or the spread of data. The variance is large in this data.
Calculation of variance. = 1270.137

- **Standard deviation:** 35.64. (it's the square root of variance)
- **Potential Outliers:** There are no potential outliers in this distribution of data.
- **Minimum value:** 14
- **Maximum value:** 174
- **1st Quartile(Q1):** 69.75
- **2nd Quartile(Q2):** 90.50 (Note: Q2 is also called Median)
- **3rd Quartile(Q3):** 115
- **Range:** 160 (max-min)
- **Mean/ Average value of the data:** 90.98.
- **Median:** 90.50 (Note: Median is also called Q2)

6. Use descriptive statistics to understand the column "misc". Visualize it by using boxplot, and interpret your boxplot [20]

ANSWER:



Interpretation of the boxplot of the variable 'misc':

- **The distribution:** The box-plot follows "Normal distribution/ Symmetric". Hence the data is evenly distributed around the center. The median is in the middle.
 - **Skewness:** There is no skewness in the boxplot. Because the median is exactly in the middle. So, it follows Normal distribution.
 - **Potential Outliers:** There are no potential outliers.
 - **The variance.:** Variance is nothing but the variation or the spread of data. The variance is small in the boxplot. The data is less
- Calculation of variance: 194.3183

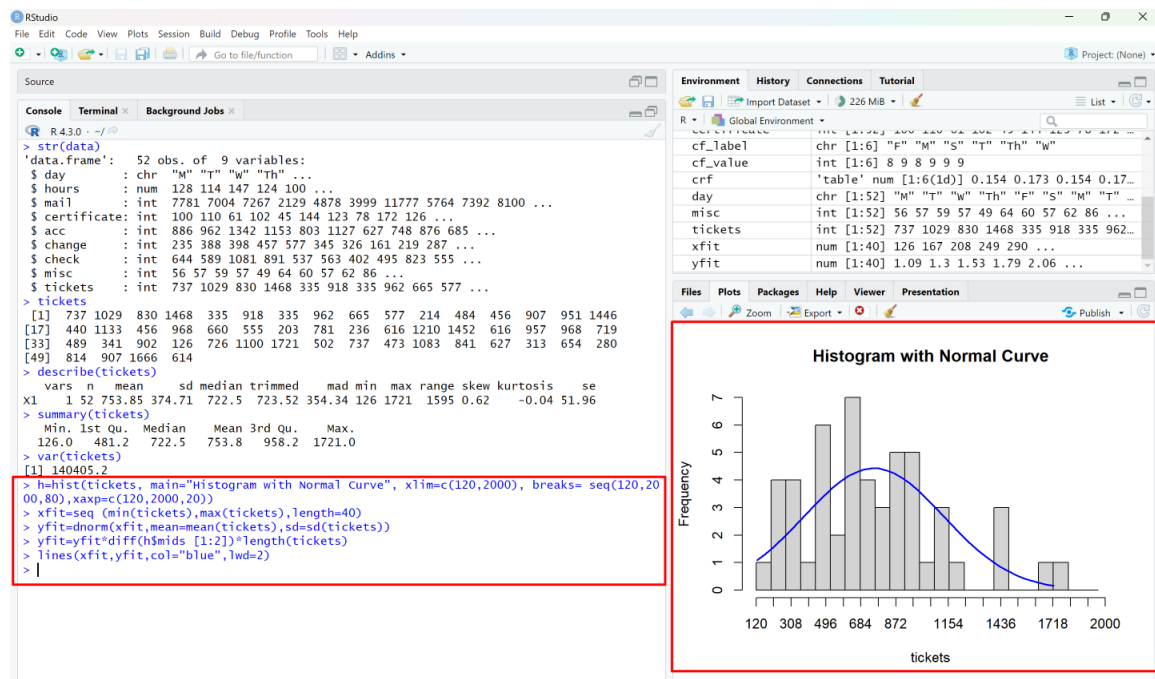
- **Standard deviation:** 13.94 (it's the square root of variance)
- **Minimum value:** 30
- **Maximum value:** 86
- **1st Quartile(Q1):** 49
- **2nd Quartile(Q2):** 57 (Note: Q2 is also called Median)
- **3rd Quartile(Q3):** 69.25
- **Range:** (maximum value – minimum value) = 56
- **Mean value/ Average value of the data:** 58.27
- **Median value:** 57 (Note: Median is also called Q2)

7. Visualize the column “tickets” by using probability curve, interpret it [15]

ANSWER:

To draw the probability curve, first we need to draw the histogram of the variable ‘tickets’ by using ‘hist’ function.

The process is as follows shown in the screenshot:



Interpretation of the distribution curve:

- **The distribution:** Its almost Normal/Symmetric distribution where the data is evenly distributed around the center. But the distribution is slightly skewed towards right.
✓ Variable tickets follows normal distribution: tickets ~ N(mean, variance)
- **Skewness:** There is a slight rightly skewed/positive skewed distribution

- **Potential outliers:** There is 1 outlier towards the maximum value (1721)
 - The maximum value could be the outlier here, according to the formula: $Q3 + 1.5(IQR)$ which calculates the upper boundary. Any data point above upper boundary / lower boundary falls under outliers. (where IQR = Interquartile Range)
 - Here the upper boundary is 1673.7 according to the formula.
 - So, the maximum value is 1721 above the upper boundary 1673.7, which is considered as an outlier in this distribution.
- **Variance:** Variance is nothing but the variation or the spread of data. The variance is large in this data.
 - Calculation of variance: 140405.2
- **Standard deviation:** 374.71 (it's the square root of variance)
- **Minimum value:** 126
- **Maximum value:** 1721
- **Range:** 1595 (max-min)
- **1st Quartile(Q1):** 481.2
- **2nd Quartile(Q2):** 722.5 (Note: Q2 is also called Median)
- **3rd Quartile(Q3):** 958.2
- **Mean value/ Average value of the data:** 753.8
- **Median value:** 722.5 (Note: Median is also called Q2)