American International University of Bangladesh

Mid-Term Project Report

Introduction to Data Science

Section: C

Submitting to: **DR. ABDUS SALAM**
abdus.salam@aiub.edu
Faculty

Submitted by: Kakon, Khairul Islam (**20-42438-1**)

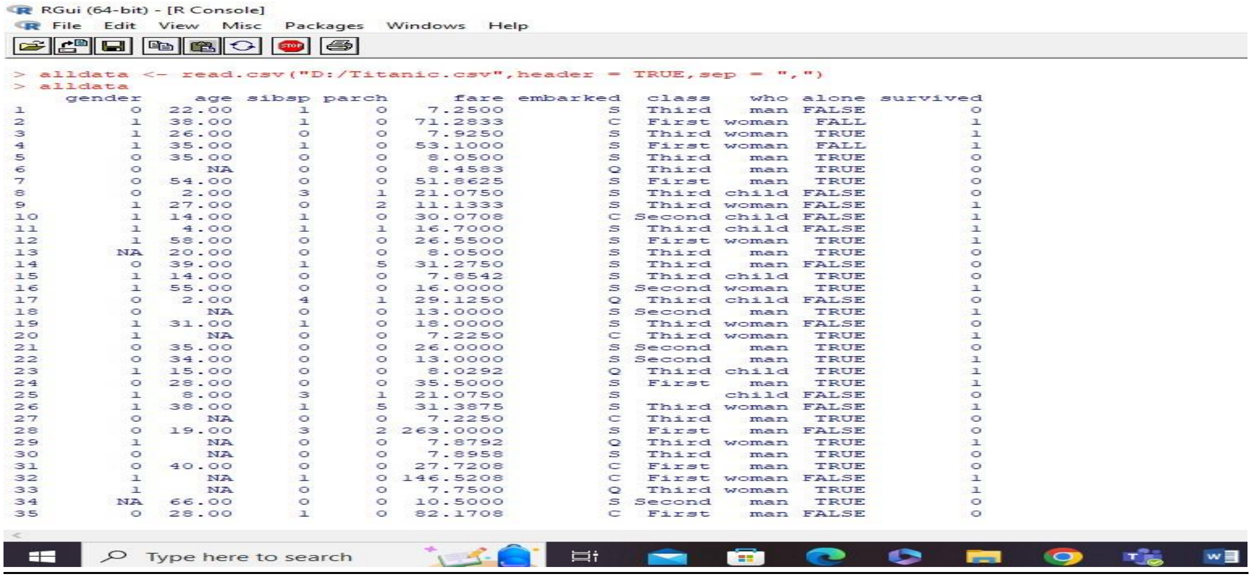
Summary

The Titanic dataset is a comprehensive and diverse collection of structured data that offers immense potential for research, analysis, and development within the [specific domain/subject].

Its size, coverage, and data types provide a rich resource for exploring various phenomena and building robust models.

The dataset includes various attributes for each passenger, their age, class, gender, etc. to predict if they would have survived or not. The dataset contains a total of 251 rows or instances, representing the passengers on board the Titanic. It provides valuable information for analyzing various aspects related to the survival rate of passengers, including factors such as passenger class, age, and gender.

Importing Titanic.csv file into R studio

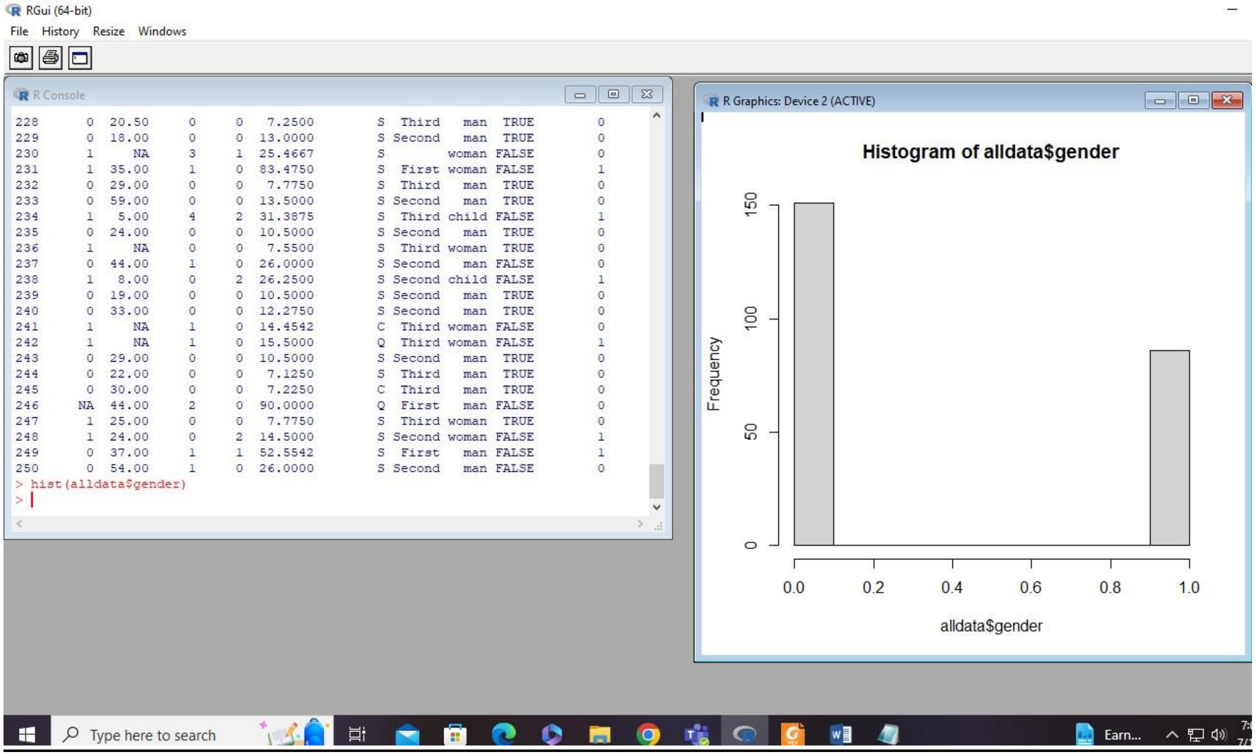

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
Type here to search

```
> alldata <- read.csv("D:/Titanic.csv",header = TRUE,sep = ",")
> alldata
    gender   age sibsp parch     fare embarked  class   who alone survived
1        0 22.00     1     0   7.2500        S  Third   man FALSE        0
2        1 38.00     1     0  71.2833        C  First woman  FALL        1
3        1 26.00     0     0   7.9250        S  Third woman  TRUE        1
4        1 35.00     1     0  53.1000        S  First woman  FALL        1
5        0 35.00     0     0   8.0500        S  Third   man  TRUE        0
6        0    NA     0     0   8.4583        Q  Third   man  TRUE        0
7        0 54.00     0     0  51.8625        S  First   man  TRUE        0
8        0  2.00     3     1  21.0750        S  Third child FALSE        0
9        1 27.00     0     2  11.1333        S  Third woman FALSE        1
10       1 14.00     1     0  30.0708        C Second child FALSE        1
11       1  4.00     1     1  16.7000        S  Third child FALSE        1
12       1 58.00     0     0  26.5500        S  First woman  TRUE        1
13      NA 20.00     0     0   8.0500        S  Third   man  TRUE        0
14       0 39.00     1     5  31.2750        S  Third   man FALSE        0
15       1 14.00     0     0   7.8542        S  Third child  TRUE        0
16       1 55.00     0     0  16.0000        S Second woman  TRUE        1
17       0  2.00     4     1  29.1250        Q  Third child FALSE        0
18       0    NA     0     0  13.0000        S Second   man  TRUE        1
19       1 31.00     1     0  18.0000        S  Third woman FALSE        0
20       1    NA     0     0   7.2250        C  Third woman  TRUE        1
21       0 35.00     0     0  26.0000        S Second   man  TRUE        0
22       0 34.00     0     0  13.0000        S Second   man  TRUE        1
23       1 15.00     0     0   8.0292        Q  Third child  TRUE        1
24       0 28.00     0     0  35.5000        S  First   man  TRUE        1
25       1  8.00     3     1  21.0750        S        child FALSE        0
26       1 38.00     1     5  31.3875        S  Third woman FALSE        1
27       0    NA     0     0   7.2250        C  Third   man  TRUE        0
28       0 19.00     3     2 263.0000        S  First   man FALSE        0
29       1    NA     0     0   7.8792        Q  Third woman  TRUE        1
30       0    NA     0     0   7.8958        S  Third   man  TRUE        0
31       0 40.00     0     0  27.7208        C  First   man  TRUE        0
32       1    NA     1     0 146.5208        C  First woman FALSE        1
33       1    NA     0     0   7.7500        Q  Third woman  TRUE        1
34      NA 66.00     0     0  10.5000        S Second   man  TRUE        0
35       0 28.00     1     0  82.1708        C  First   man FALSE        0
```

Here I have imported the dataset Titanic.csv using the “read.csv” command in the parameter I had provided the location of dataset Titanic.csv in my computer file.

Then I ensured the parameter of a header by giving true and making it in comma Delimited Text Data Set I used sep = “,” and to check all the data from the data set I wrote all data. The median is the value of the central point in the distribution.

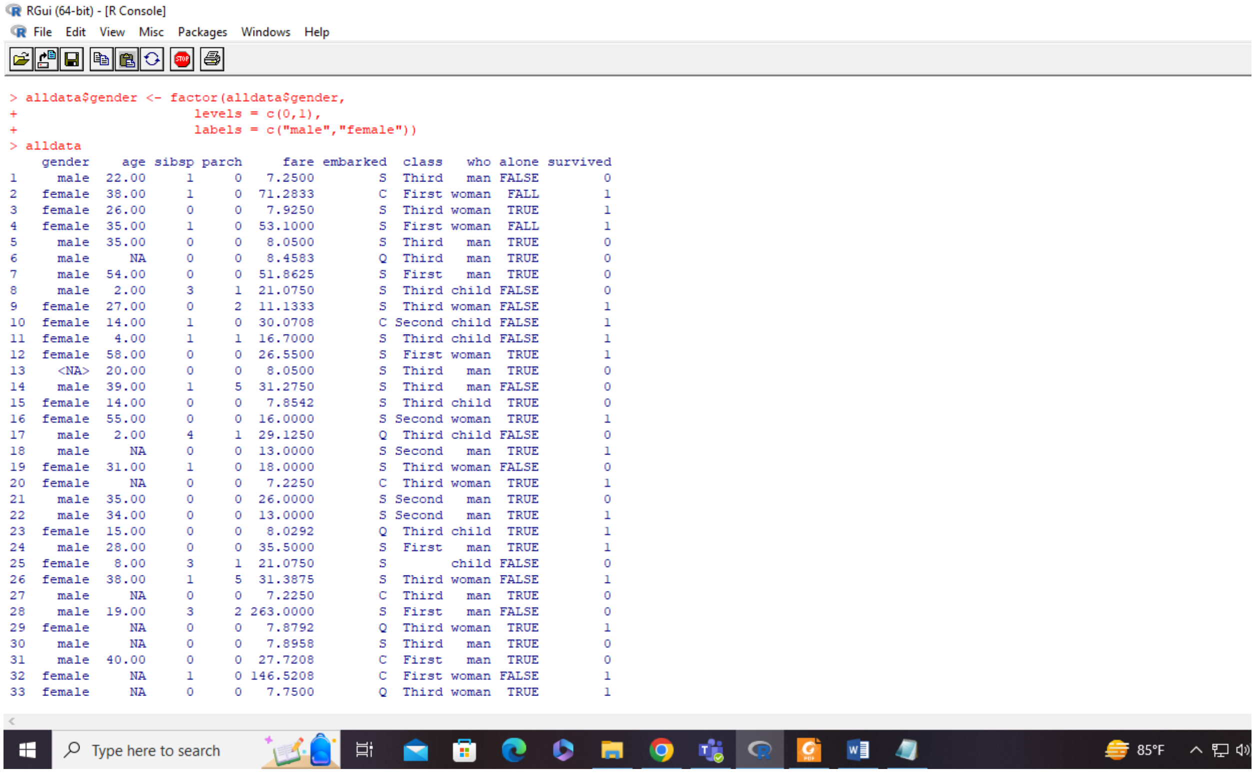
Histogram of Dataset



```
228      0  20.50     0     0   7.2500         S  Third   man  TRUE        0
229      0  18.00     0     0  13.0000         S Second   man  TRUE        0
230      1     NA     3     1  25.4667         S       woman FALSE        0
231      1  35.00     1     0  83.4750         S  First woman FALSE        1
232      0  29.00     0     0   7.7750         S  Third   man  TRUE        0
233      0  59.00     0     0  13.5000         S Second   man  TRUE        0
234      1   5.00     4     2  31.3875         S  Third child FALSE        1
235      0  24.00     0     0  10.5000         S Second   man  TRUE        0
236      1     NA     0     0   7.5500         S  Third woman  TRUE        0
237      0  44.00     1     0  26.0000         S Second   man FALSE        0
238      1   8.00     0     2  26.2500         S Second child FALSE        1
239      0  19.00     0     0  10.5000         S Second   man  TRUE        0
240      0  33.00     0     0  12.2750         S Second   man  TRUE        0
241      1     NA     1     0  14.4542         C  Third woman FALSE        0
242      1     NA     1     0  15.5000         Q  Third woman FALSE        1
243      0  29.00     0     0  10.5000         S Second   man  TRUE        0
244      0  22.00     0     0   7.1250         S  Third   man  TRUE        0
245      0  30.00     0     0   7.2250         C  Third   man  TRUE        0
246     NA  44.00     2     0  90.0000         Q  First   man FALSE        0
247      1  25.00     0     0   7.7750         S  Third woman  TRUE        0
248      1  24.00     0     2  14.5000         S Second woman FALSE        1
249      0  37.00     1     1  52.5542         S  First   man FALSE        1
250      0  54.00     1     0  26.0000         S Second   man FALSE        0
> hist(alldata$gender)
> 
```

Here, using hist(alldata$gender) command to get the histogram of values of gender attribute from Titanic.csv dataset.

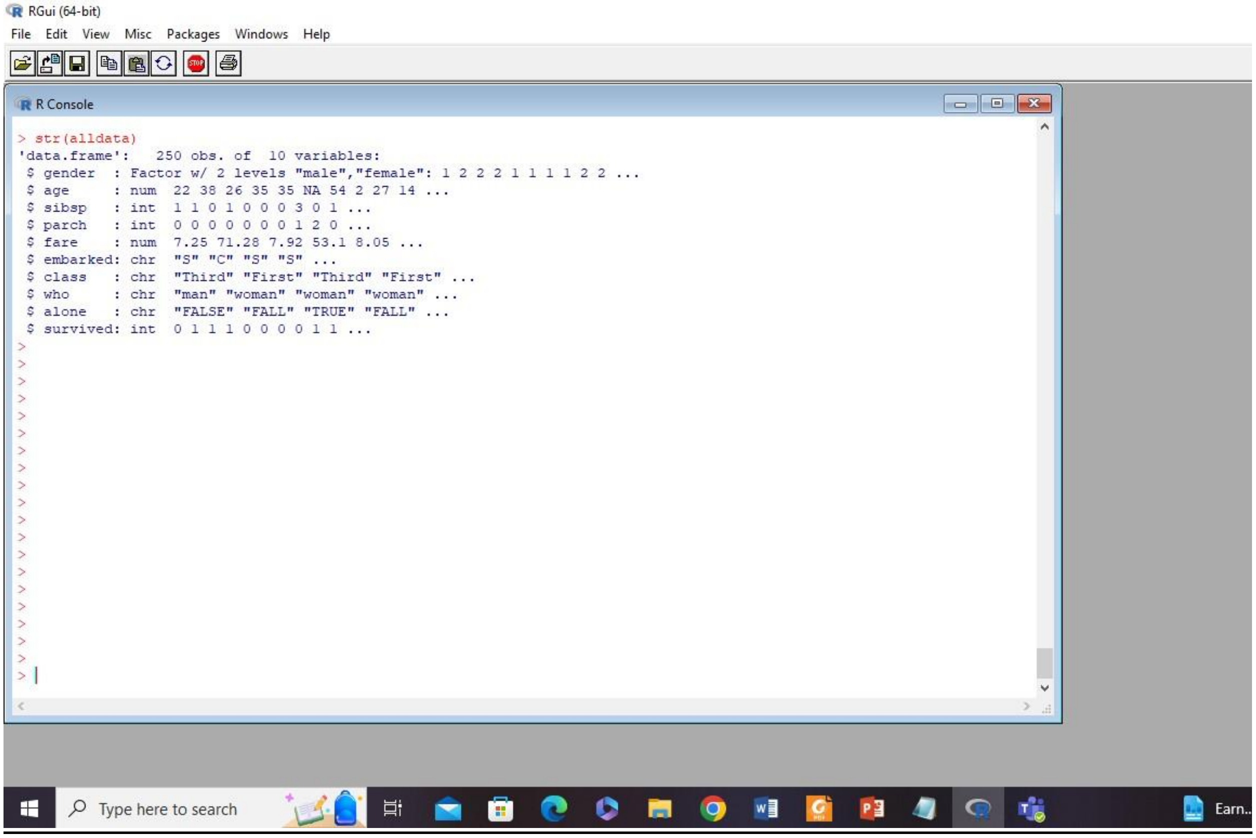
Annotating Dataset


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
Type here to search
85°F

```
> alldata$gender <- factor(alldata$gender,
+                      levels = c(0,1),
+                      labels = c("male","female"))
> alldata
   gender   age sibsp parch     fare embarked  class   who alone survived
1    male 22.00     1     0   7.2500        S  Third   man FALSE        0
2  female 38.00     1     0  71.2833        C  First woman  FALL        1
3  female 26.00     0     0   7.9250        S  Third woman  TRUE        1
4  female 35.00     1     0  53.1000        S  First woman  FALL        1
5    male 35.00     0     0   8.0500        S  Third   man  TRUE        0
6    male    NA     0     0   8.4583        Q  Third   man  TRUE        0
7    male 54.00     0     0  51.8625        S  First   man  TRUE        0
8    male  2.00     3     1  21.0750        S  Third child FALSE        0
9  female 27.00     0     2  11.1333        S  Third woman FALSE        1
10 female 14.00     1     0  30.0708        C Second child FALSE        1
11 female  4.00     1     1  16.7000        S  Third child FALSE        1
12 female 58.00     0     0  26.5500        S  First woman  TRUE        1
13   <NA> 20.00     0     0   8.0500        S  Third   man  TRUE        0
14   male 39.00     1     5  31.2750        S  Third   man FALSE        0
15 female 14.00     0     0   7.8542        S  Third child  TRUE        0
16 female 55.00     0     0  16.0000        S Second woman  TRUE        1
17   male  2.00     4     1  29.1250        Q  Third child FALSE        0
18   male    NA     0     0  13.0000        S Second   man  TRUE        1
19 female 31.00     1     0  18.0000        S  Third woman FALSE        0
20 female    NA     0     0   7.2250        C  Third woman  TRUE        1
21   male 35.00     0     0  26.0000        S Second   man  TRUE        0
22   male 34.00     0     0  13.0000        S Second   man  TRUE        1
23 female 15.00     0     0   8.0292        Q  Third child  TRUE        1
24   male 28.00     0     0  35.5000        S  First   man  TRUE        1
25 female  8.00     3     1  21.0750        S        child FALSE        0
26 female 38.00     1     5  31.3875        S  Third woman FALSE        1
27   male    NA     0     0   7.2250        C  Third   man  TRUE        0
28   male 19.00     3     2 263.0000        S  First   man FALSE        0
29 female    NA     0     0   7.8792        Q  Third woman  TRUE        1
30   male    NA     0     0   7.8958        S  Third   man  TRUE        0
31   male 40.00     0     0  27.7208        C  First   man  TRUE        0
32 female    NA     1     0 146.5208        C  First woman FALSE        1
33 female    NA     0     0   7.7500        Q  Third woman  TRUE        1
```

Here, annotating a dataset involves adding descriptive information, metadata, and contextual details to enhance the understanding, usability, and reliability of the data. By providing clear variable descriptions, data source information, preprocessing details, quality assessments, metadata, usage guidelines, and versioning information, dataset annotation ensures that users can effectively interpret, analyze, and make informed decisions based on the dataset. According to this code, the factor() function can be used to create value labels for categorical variables. Continuing for the above code example, say that I have a variable named gender, which is coded 0 for male and 1 for female.

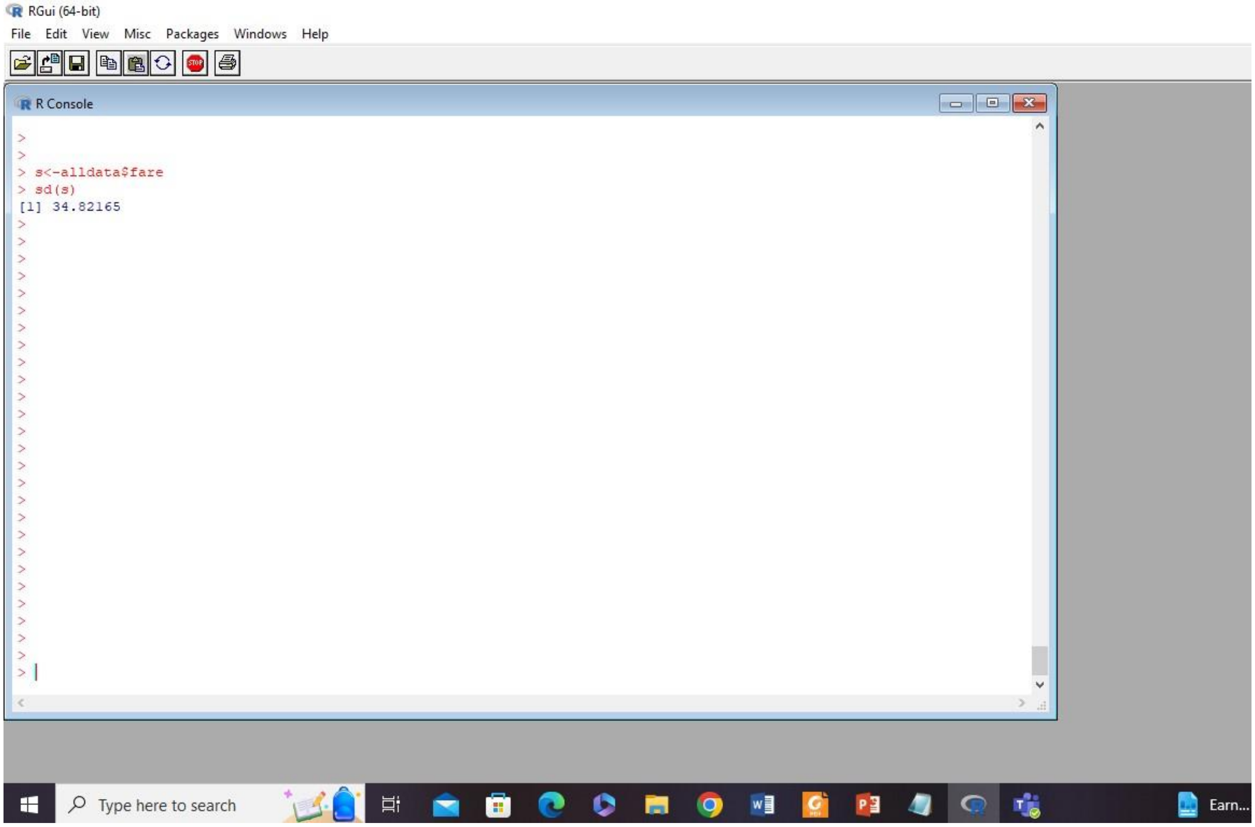
<u>Structure of Dataset</u>



```
> str(alldata)
'data.frame':   250 obs. of  10 variables:
 $ gender  : Factor w/ 2 levels "male","female": 1 2 2 2 1 1 1 1 2 2 ...
 $ age     : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ sibsp   : int  1 1 0 1 0 0 0 3 0 1 ...
 $ parch   : int  0 0 0 0 0 0 0 1 2 0 ...
 $ fare    : num  7.25 71.28 7.92 53.1 8.05 ...
 $ embarked: chr  "S" "C" "S" "S" ...
 $ class   : chr  "Third" "First" "Third" "First" ...
 $ who     : chr  "man" "woman" "woman" "woman" ...
 $ alone   : chr  "FALSE" "FALL" "TRUE" "FALL" ...
 $ survived: int  0 1 1 1 0 0 0 0 1 1 ...
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> |
```

Here, I have used ‘str()’ command which shows the summary of the dataset and It shows from 150 observations of 10 variables.

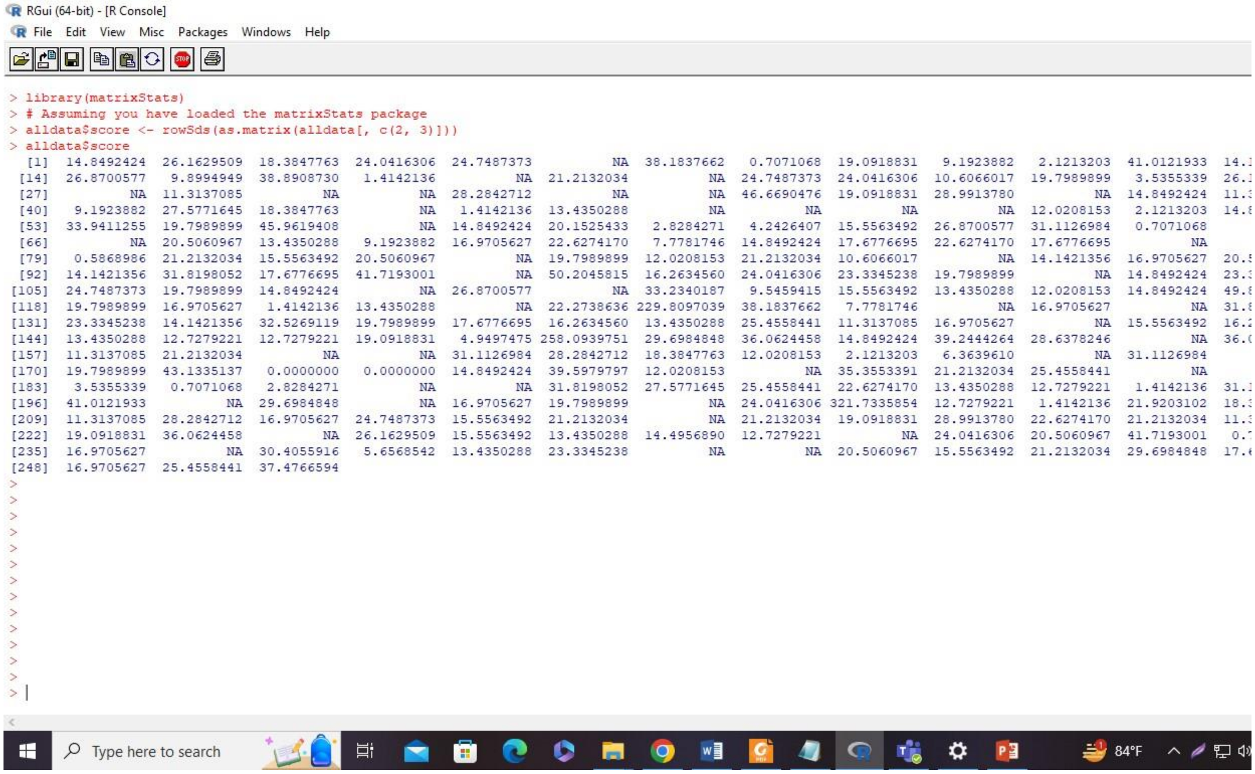
Standard Deviation



```
>
>
> s<-alldata$fare
> sd(s)
[1] 34.82165
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> 
```

A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean and high standard deviation indicates data are more spread out. Here, from the fare attribute data I have calculated the deviation value is 34.82165 more spread out. So, it is a high standard deviation

Raw wise standard deviation

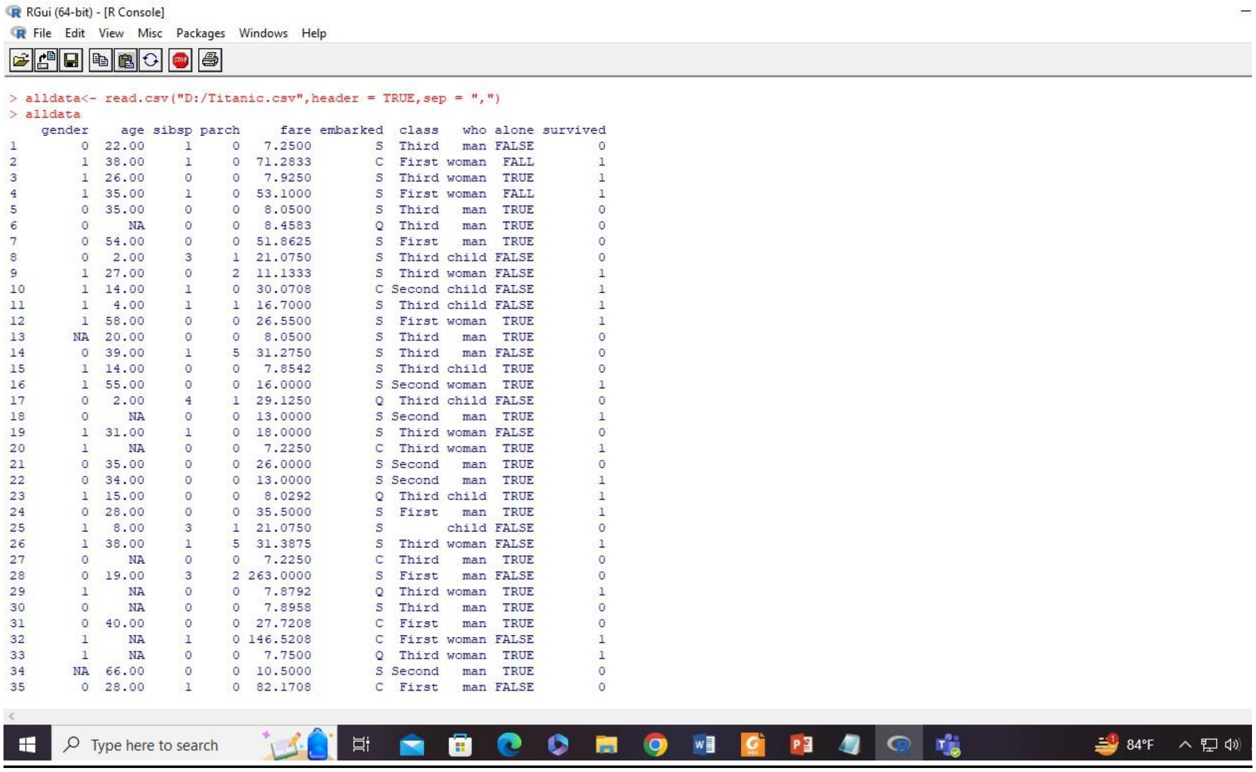

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
Type here to search
84°F

```
> library(matrixStats)
> # Assuming you have loaded the matrixStats package
> alldata$score <- rowSds(as.matrix(alldata[, c(2, 3)]))
> alldata$score
  [1]  14.8492424  26.1629509  18.3847763  24.0416306  24.7487373          NA  38.1837662   0.7071068  19.0918831   9.1923882   2.1213203  41.0121933  14.1
 [14]  26.8700577   9.8994949  38.8908730   1.4142136          NA  21.2132034          NA  24.7487373  24.0416306  10.6066017  19.7989899   3.5355339  26.1
 [27]          NA  11.3137085          NA          NA  28.2842712          NA          NA  46.6690476  19.0918831  28.9913780          NA  14.8492424  11.3
 [40]   9.1923882  27.5771645  18.3847763          NA   1.4142136  13.4350288          NA          NA          NA          NA  12.0208153   2.1213203  14.8
 [53]  33.9411255  19.7989899  45.9619408          NA  14.8492424  20.1525433   2.8284271   4.2426407  15.5563492  26.8700577  31.1126984   0.7071068
 [66]          NA  20.5060967  13.4350288   9.1923882  16.9705627  22.6274170   7.7781746  14.8492424  17.6776695  22.6274170  17.6776695          NA
 [79]   0.5868986  21.2132034  15.5563492  20.5060967          NA  19.7989899  12.0208153  21.2132034  10.6066017          NA  14.1421356  16.9705627  20.5
 [92]  14.1421356  31.8198052  17.6776695  41.7193001          NA  50.2045815  16.2634560  24.0416306  23.3345238  19.7989899          NA  14.8492424  23.3
[105]  24.7487373  19.7989899  14.8492424          NA  26.8700577          NA  33.2340187   9.5459415  15.5563492  13.4350288  12.0208153  14.8492424  49.8
[118]  19.7989899  16.9705627   1.4142136  13.4350288          NA  22.2738636 229.8097039  38.1837662   7.7781746          NA  16.9705627          NA  31.8
[131]  23.3345238  14.1421356  32.5269119  19.7989899  17.6776695  16.2634560  13.4350288  25.4558441  11.3137085  16.9705627          NA  15.5563492  16.2
[144]  13.4350288  12.7279221  12.7279221  19.0918831   4.9497475 258.0939751  29.6984848  36.0624458  14.8492424  39.2444264  28.6378246          NA  36.0
[157]  11.3137085  21.2132034          NA          NA  31.1126984  28.2842712  18.3847763  12.0208153   2.1213203   6.3639610          NA  31.1126984
[170]  19.7989899  43.1335137   0.0000000   0.0000000  14.8492424  39.5979797  12.0208153          NA  35.3553391  21.2132034  25.4558441          NA
[183]   3.5355339   0.7071068   2.8284271          NA          NA  31.8198052  27.5771645  25.4558441  22.6274170  13.4350288  12.7279221   1.4142136  31.1
[196]  41.0121933          NA  29.6984848          NA  16.9705627  19.7989899          NA  24.0416306 321.7335854  12.7279221   1.4142136  21.9203102  18.3
[209]  11.3137085  28.2842712  16.9705627  24.7487373  15.5563492  21.2132034          NA  21.2132034  19.0918831  28.9913780  22.6274170  21.2132034  11.3
[222]  19.0918831  36.0624458          NA  26.1629509  15.5563492  13.4350288  14.4956890  12.7279221          NA  24.0416306  20.5060967  41.7193001   0.7
[235]  16.9705627          NA  30.4055916   5.6568542  13.4350288  23.3345238          NA          NA  20.5060967  15.5563492  21.2132034  29.6984848  17.6
[248]  16.9705627  25.4558441  37.4766594
>
>
>
>
>
>
>
>
>
>
>
>
>
> |
```

The row-wise standard deviation in a dataset indicates the variability or spread of values within each row or observation of the dataset. It provides information about how much the individual values within a row deviate from the mean of that row.

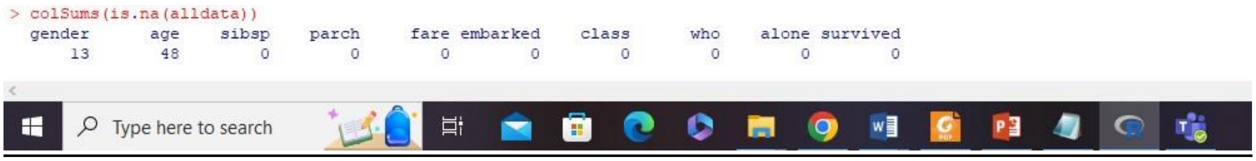
By using alldata$score <- rowSds(as.matrix(alldata[, c(2, 3)])) and > alldata$score command gives the standard deviation in row wise standard deviation.

Counting Missing Values



RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
Type here to search
84°F

```
> alldata<- read.csv("D:/Titanic.csv",header = TRUE,sep = ",")
> alldata
    gender    age sibsp parch      fare embarked  class   who alone survived
1        0  22.00     1     0    7.2500        S  Third   man FALSE        0
2        1  38.00     1     0   71.2833        C  First woman  FALL        1
3        1  26.00     0     0    7.9250        S  Third woman  TRUE        1
4        1  35.00     1     0   53.1000        S  First woman  FALL        1
5        0  35.00     0     0    8.0500        S  Third   man  TRUE        0
6        0     NA     0     0    8.4583        Q  Third   man  TRUE        0
7        0  54.00     0     0   51.8625        S  First   man  TRUE        0
8        0   2.00     3     1   21.0750        S  Third child FALSE        0
9        1  27.00     0     2   11.1333        S  Third woman FALSE        1
10       1  14.00     1     0   30.0708        C Second child FALSE        1
11       1   4.00     1     1   16.7000        S  Third child FALSE        1
12       1  58.00     0     0   26.5500        S  First woman  TRUE        1
13      NA  20.00     0     0    8.0500        S  Third   man  TRUE        0
14       0  39.00     1     5   31.2750        S  Third   man FALSE        0
15       1  14.00     0     0    7.8542        S  Third child  TRUE        0
16       1  55.00     0     0   16.0000        S Second woman  TRUE        1
17       0   2.00     4     1   29.1250        Q  Third child FALSE        0
18       0     NA     0     0   13.0000        S Second   man  TRUE        1
19       1  31.00     1     0   18.0000        S  Third woman FALSE        0
20       1     NA     0     0    7.2250        C  Third woman  TRUE        1
21       0  35.00     0     0   26.0000        S Second   man  TRUE        0
22       0  34.00     0     0   13.0000        S Second   man  TRUE        1
23       1  15.00     0     0    8.0292        Q  Third child  TRUE        1
24       0  28.00     0     0   35.5000        S  First   man  TRUE        1
25       1   8.00     3     1   21.0750        S        child FALSE        0
26       1  38.00     1     5   31.3875        S  Third woman FALSE        1
27       0     NA     0     0    7.2250        C  Third   man  TRUE        0
28       0  19.00     3     2  263.0000        S  First   man FALSE        0
29       1     NA     0     0    7.8792        Q  Third woman  TRUE        1
30       0     NA     0     0    7.8958        S  Third   man  TRUE        0
31       0  40.00     0     0   27.7208        C  First   man  TRUE        0
32       1     NA     1     0  146.5208        C  First woman FALSE        1
33       1     NA     0     0    7.7500        Q  Third woman  TRUE        1
34      NA  66.00     0     0   10.5000        S Second   man  TRUE        0
35       0  28.00     1     0   82.1708        C  First   man FALSE        0
```

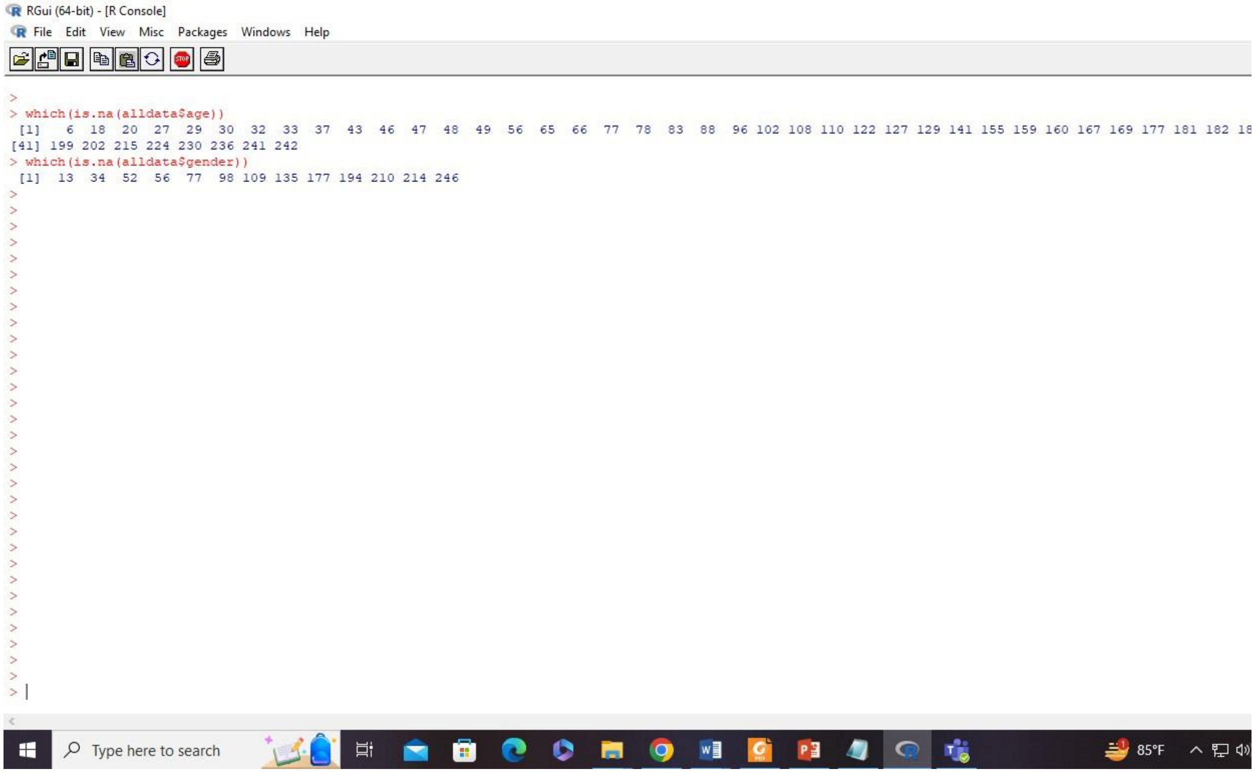


```
> colSums(is.na(alldata))
   gender       age     sibsp     parch      fare  embarked     class       who     alone  survived
       13        48         0         0         0         0         0         0         0         0
```

Here I have counted missing values on the dataset Titanic.csv. and found that gender has 13 and age is 48 missing values.

In R, the NA symbol is used to define the missing values and to represent impossible arithmetic operations (like dividing by zero) we use the NAN symbol which stands for “not a number”. In simple words, we can say that both NA or NAN symbols represent missing values in R.

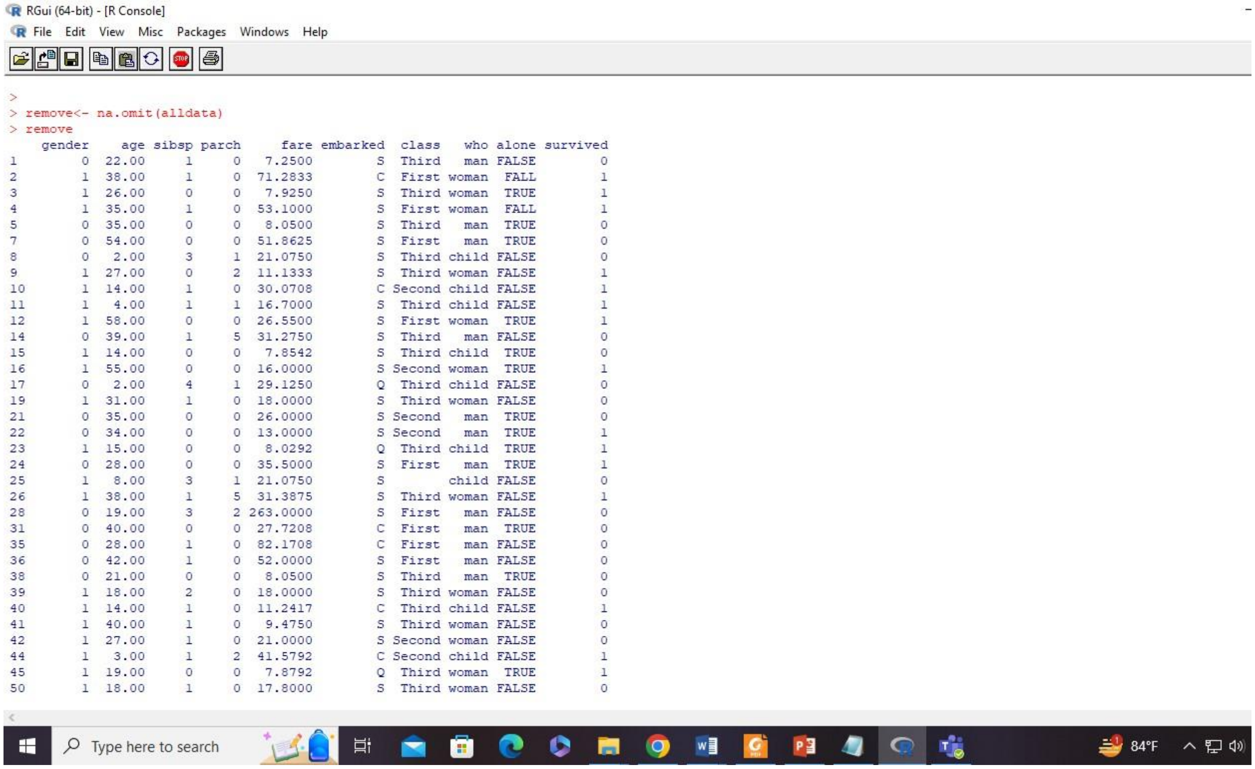
Finding Missing Values


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
Type here to search
85°F

```
>
> which(is.na(alldata$age))
 [1]   6  18  20  27  29  30  32  33  37  43  46  47  48  49  56  65  66  77  78  83  88  96 102 108 110 122 127 129 141 155 159 160 167 169 177 181 182 18
[41] 199 202 215 224 230 236 241 242
> which(is.na(alldata$gender))
 [1]  13  34  52  56  77  98 109 135 177 194 210 214 246
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```

After counting missing values I got that gender has 13 and age is 48 missing values. And finding those by using >which(is.na(alldata$age)) and using >which(is.na(alldata$gender)) command in r studio.

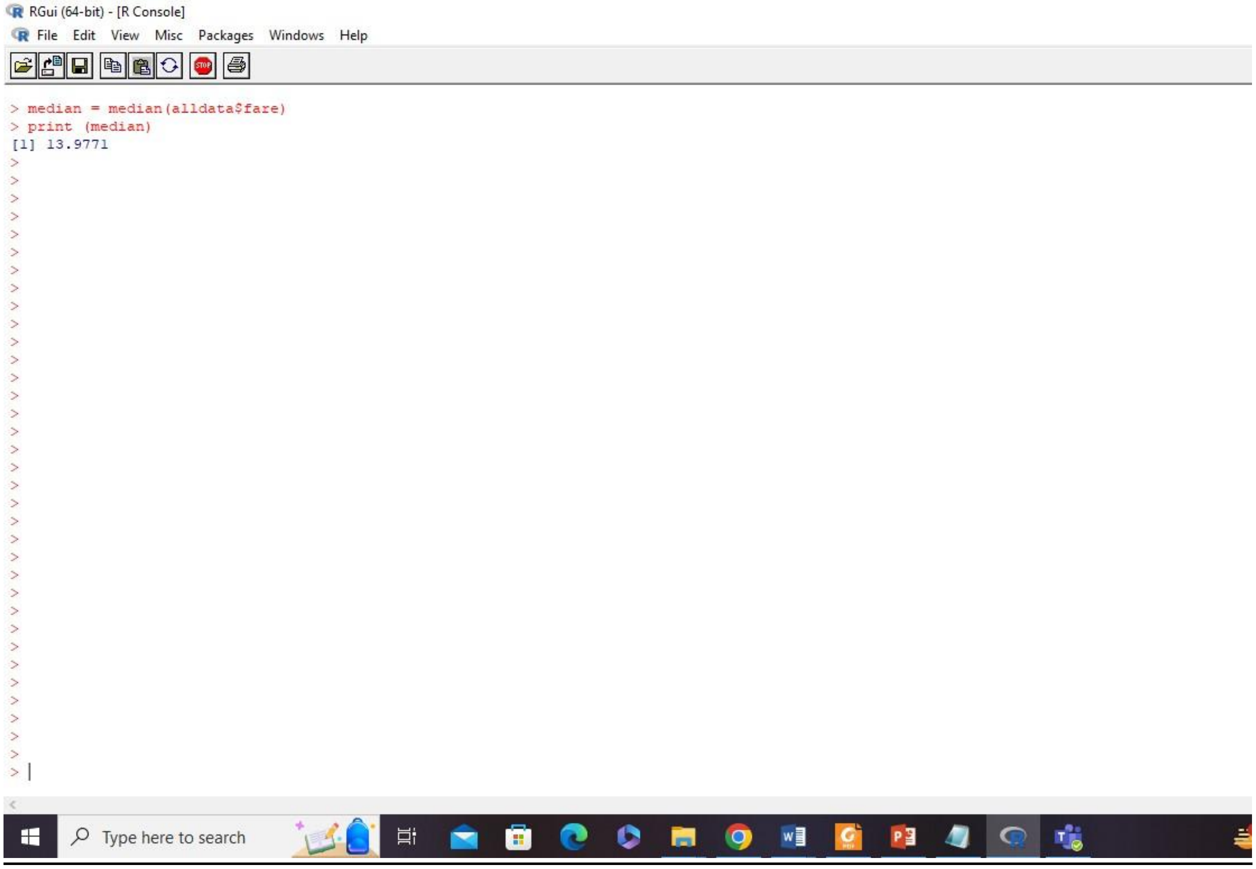
Removing Missing Values



```
>
> remove<- na.omit(alldata)
> remove
   gender   age sibsp parch     fare embarked  class   who alone survived
1       0 22.00     1     0   7.2500        S  Third   man FALSE        0
2       1 38.00     1     0  71.2833        C  First woman  FALL        1
3       1 26.00     0     0   7.9250        S  Third woman  TRUE        1
4       1 35.00     1     0  53.1000        S  First woman  FALL        1
5       0 35.00     0     0   8.0500        S  Third   man  TRUE        0
7       0 54.00     0     0  51.8625        S  First   man  TRUE        0
8       0  2.00     3     1  21.0750        S  Third child FALSE        0
9       1 27.00     0     2  11.1333        S  Third woman FALSE        1
10      1 14.00     1     0  30.0708        C Second child FALSE        1
11      1  4.00     1     1  16.7000        S  Third child FALSE        1
12      1 58.00     0     0  26.5500        S  First woman  TRUE        1
14      0 39.00     1     5  31.2750        S  Third   man FALSE        0
15      1 14.00     0     0   7.8542        S  Third child  TRUE        0
16      1 55.00     0     0  16.0000        S Second woman  TRUE        1
17      0  2.00     4     1  29.1250        Q  Third child FALSE        0
19      1 31.00     1     0  18.0000        S  Third woman FALSE        0
21      0 35.00     0     0  26.0000        S Second   man  TRUE        0
22      0 34.00     0     0  13.0000        S Second   man  TRUE        1
23      1 15.00     0     0   8.0292        Q  Third child  TRUE        1
24      0 28.00     0     0  35.5000        S  First   man  TRUE        1
25      1  8.00     3     1  21.0750        S        child FALSE        0
26      1 38.00     1     5  31.3875        S  Third woman FALSE        1
28      0 19.00     3     2 263.0000        S  First   man FALSE        0
31      0 40.00     0     0  27.7208        C  First   man  TRUE        0
35      0 28.00     1     0  82.1708        C  First   man FALSE        0
36      0 42.00     1     0  52.0000        S  First   man FALSE        0
38      0 21.00     0     0   8.0500        S  Third   man  TRUE        0
39      1 18.00     2     0  18.0000        S  Third woman FALSE        0
40      1 14.00     1     0  11.2417        C  Third child FALSE        1
41      1 40.00     1     0   9.4750        S  Third woman FALSE        0
42      1 27.00     1     0  21.0000        S Second woman FALSE        0
44      1  3.00     1     2  41.5792        C Second child FALSE        1
45      1 19.00     0     0   7.8792        Q  Third woman  TRUE        1
50      1 18.00     1     0  17.8000        S  Third woman FALSE        0
```

Here in R studio, I have used remove<-na. omit() command to remove all observations with missing data on ANY variable in the dataset, or use subset() to filter out cases that are missing on a subset of variables. Though there are many but I used this command to remove NA values in Titanic.csv file.

Median

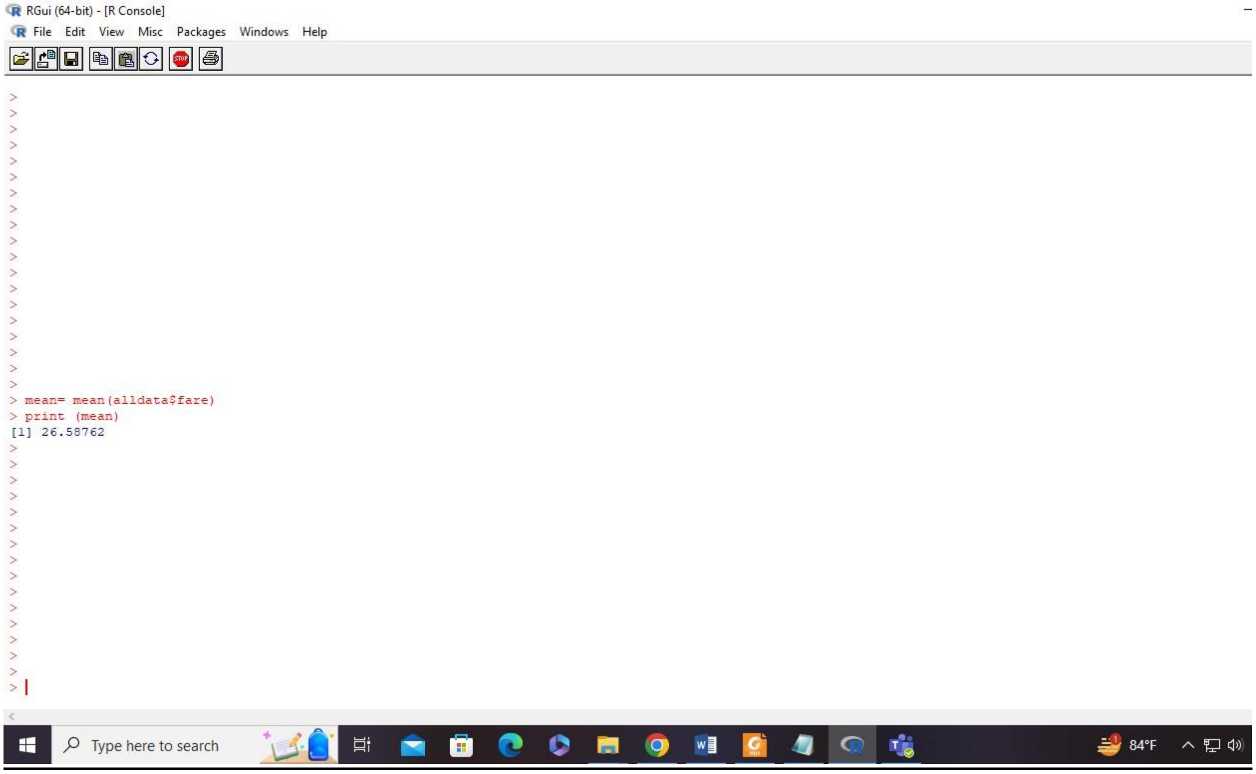


```
> median = median(alldata$fare)
> print (median)
[1] 13.9771
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> |
```

Here, I have used the median = median(alldata$fare) command to calculate the median value from the Titanic.csv file, and using print (median) I have the median value which is 13.9771 which indicates the central tendency of fare price from all the values from fare attribute

The median is useful because it is not affected by extreme values (outliers) to the same extent as the mean. It provides a robust measure of the center of the dataset, particularly in situations where the distribution is not symmetrical or when the data contains extreme values that could skew the mean.

Mean

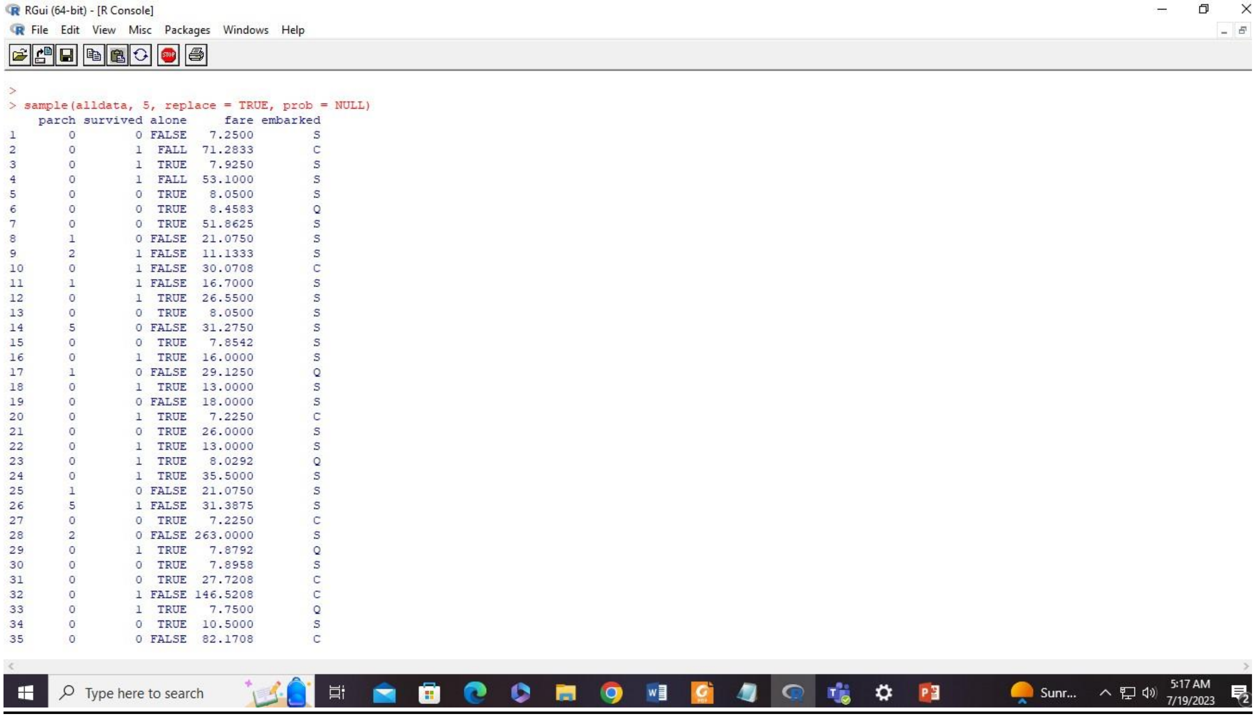

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
Type here to search
84°F

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> mean= mean(alldata$fare)
> print (mean)
[1] 26.58762
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```

The mean, also known as the average, is another measure of central tendency commonly used in statistics. Unlike the median, which represents the middle value of a dataset, the mean is the sum of all the values divided by the total number of observations.

From the values of the fare attribute, we can see that it is 26.58762 which is calculated from the mean value generating command mean= mean(alldata$fare) and printed that value with print (mean)

Sampling



RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
Type here to search
Sunr...
5:17 AM
7/19/2023

```
>
> sample(alldata, 5, replace = TRUE, prob = NULL)
    parch survived alone     fare embarked
1       0        0 FALSE   7.2500        S
2       0        1  FALL  71.2833        C
3       0        1  TRUE   7.9250        S
4       0        1  FALL  53.1000        S
5       0        0  TRUE   8.0500        S
6       0        0  TRUE   8.4583        Q
7       0        0  TRUE  51.8625        S
8       1        0 FALSE  21.0750        S
9       2        1 FALSE  11.1333        S
10      0        1 FALSE  30.0708        C
11      1        1 FALSE  16.7000        S
12      0        1  TRUE  26.5500        S
13      0        0  TRUE   8.0500        S
14      5        0 FALSE  31.2750        S
15      0        0  TRUE   7.8542        S
16      0        1  TRUE  16.0000        S
17      1        0 FALSE  29.1250        Q
18      0        1  TRUE  13.0000        S
19      0        0 FALSE  18.0000        S
20      0        1  TRUE   7.2250        C
21      0        0  TRUE  26.0000        S
22      0        1  TRUE  13.0000        S
23      0        1  TRUE   8.0292        Q
24      0        1  TRUE  35.5000        S
25      1        0 FALSE  21.0750        S
26      5        1 FALSE  31.3875        S
27      0        0  TRUE   7.2250        C
28      2        0 FALSE 263.0000        S
29      0        1  TRUE   7.8792        Q
30      0        0  TRUE   7.8958        S
31      0        0  TRUE  27.7208        C
32      0        1 FALSE 146.5208        C
33      0        1  TRUE   7.7500        Q
34      0        0  TRUE  10.5000        S
35      0        0 FALSE  82.1708        C
```

Sampling in a dataset refers to the process of selecting a subset of observations or data points from a larger population or dataset. The purpose of sampling is to obtain a representative sample that can provide insights or make inferences about the entire population. Thee code sample(alldata, 5, replace = TRUE, prob = NULL) randomly selects 5 observations from the titanic.csv dataset, allowing replacement (an observation can be chosen more than once), and each observation has an equal chance of being selected. The resulting output will be a sample of 5 observations from the Titanic.csv dataset.