American International University – Bangladesh

Department of Computer Science & Engineering

Project Title: Apply data preparation steps (which can be applied) and do the univariate data exploration for the given dataset (Titanic – modified).

Submitted by-	Submitted to-
Name: Md. Shohaibur Rahman ID: 20-42424-1 Section: C Summer 2022-2023B.Sc. CSE	Name: DR. ABDUS SALAM

Dataset Description:

The Titanic dataset is a comprehensive dataset that provides information about passengers on board the RMS Titanic. It consists of data related to 891 passengers. But for this project we will use dataset from 250 passengers includes various attributes such as their gender, age, sibling, parch, ticket fare, embarked, class, survival status etc. The dataset offers insights into the passengers' demographics, ticket details, and important factors that may have influenced their survival. The target variable, "Survived," indicates whether a passenger survived the Titanic disaster or not. This dataset serves as a valuable resource for analyzing and understanding the demographics and factors associated with the survival of passengers on the Titanic.

Attributes:

1. Gender: The gender of the passenger, classified as male or female.
2. Age: The age of the passenger in years.
3. SibSp: The number of siblings or spouses the passenger had aboard the Titanic.
4. Parch: The number of parents or children the passenger had aboard the Titanic.
5. Fare: The fare or ticket price paid by the passenger.
6. Embarked: The port of embarkation for the passenger, classified as C (Cherbourg), Q (Queenstown), or S (Southampton).
7. Class: The passenger class, indicating their socio-economic status, classified as 1 (First class), 2 (Second class), or 3 (Third class).
8. Who: Indicates whether the passenger is an adult or a child.
9. Alone: Specifies if the passenger was traveling alone or with family members, classified as yes or no.
10. Survived: Indicates whether the passenger survived the Titanic disaster or not, classified as 0 (No) or 1 (Yes).

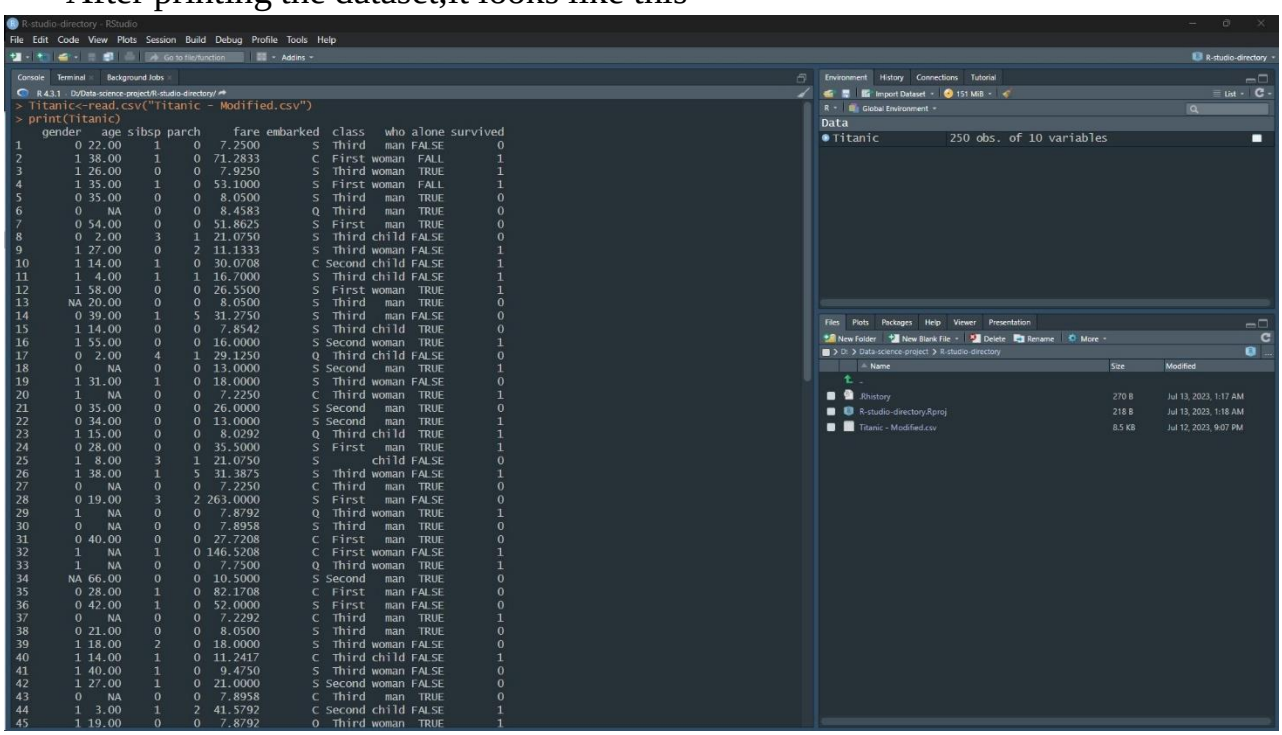
Purpose: The Titanic dataset aims to analyze the factors influencing passenger survival aboard the Titanic. It provides information on attributes such as gender, age, family relations, fare, embark, class, and survival status. The Titanic dataset enables the exploration of patterns and insights to understand the demographic and situational factors associated with higher chances of surviving the tragic Titanic event.

Project Overview:

Data pre-processing plays a critical role in data analysis by transforming raw data into a structured format suitable for analysis by computers and machine learning algorithms. Raw data often contains errors and inconsistencies that need to be addressed before it can be effectively utilized. Additionally, univariate exploration focuses on analyzing individual variables within a dataset, without considering their interrelationships. In the given dataset, it is evident that data cleaning and pre-processing are necessary before proceeding with any analysis. By performing these steps, we can ensure that the dataset is prepared for further analysis and insights can be extracted accurately

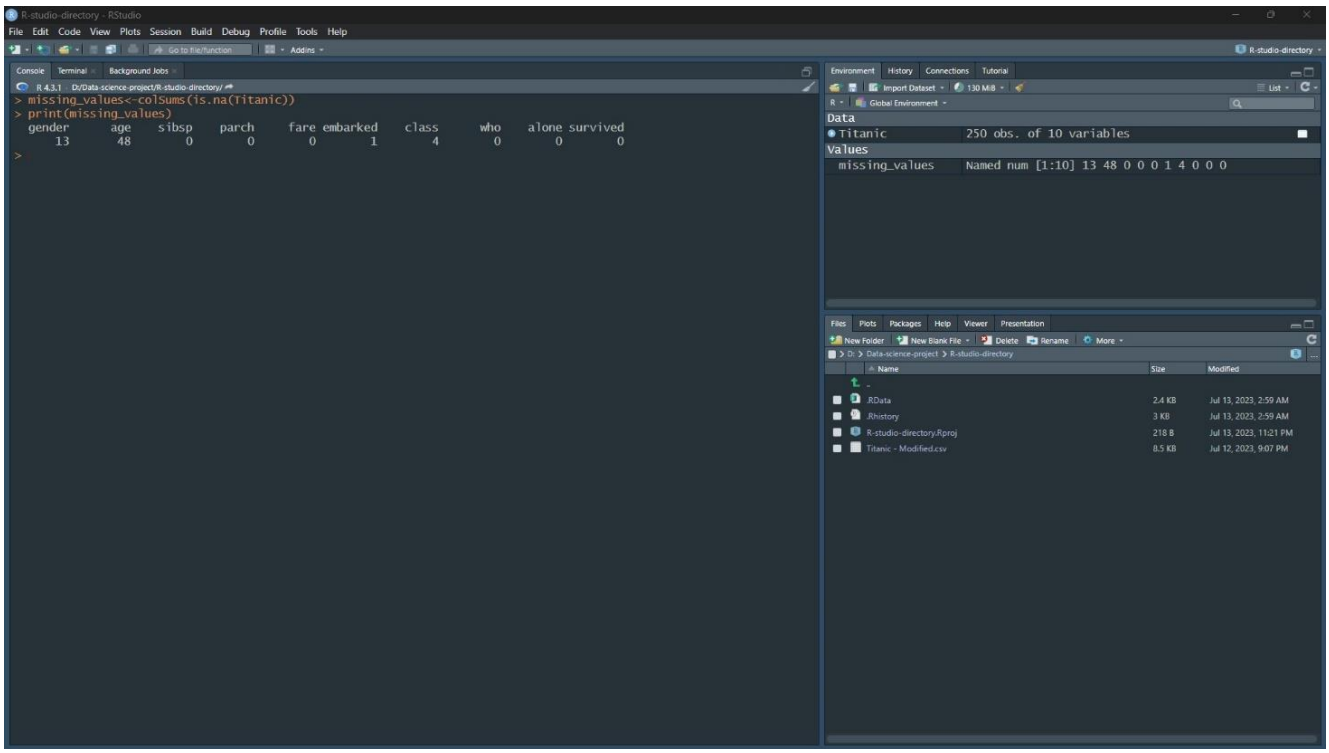
Data pre-processing:

1. **Importing the Dataset:** The dataset is located in a file called Titanic -modified .csv in the current working directory. To begin data pre-processing using R, the first stepis to import the dataset. Once imported, the Titanic -modified .csv file is transformed into anR data frame and stored in a variable named "Titanic". After printing the dataset,it looks like this-



2. Dealing with Missing Values:

2.1 We can see from the dataset that, there is some missing value (NA) and blank present in column name- gender[13], age[48], embarked[1], and class[4]. We can find out the missing values in this way:



Before, the dataset looked like this-

R-studio-directory - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Console Terminal Background Jobs

R 4.3.1 · D:/Data-science-project/R-studio-directory/

```
> print(Titanic)
    gender  age sibsp parch     fare embarked  class   who alone survived
1        0   22     1     0   7.2500        S  Third   man FALSE        0
2        1   38     1     0  71.2833        C  First woman  FALL        1
3        1   26     0     0   7.9250        S  Third woman  TRUE        1
4        1   35     1     0  53.1000        S  First woman  FALL        1
5        0   35     0     0   8.0500        S  Third   man  TRUE        0
6        0 <NA>     0     0   8.4583        Q  Third   man  TRUE        0
7        0   54     0     0  51.8625        S  First   man  TRUE        0
8        0    2     3     1  21.0750        S  Third child FALSE        0
9        1   27     0     2  11.1333        S  Third woman FALSE        1
10       1   14     1     0  30.0708        C Second child FALSE        1
11       1    4     1     1  16.7000        S  Third child FALSE        1
12       1   58     0     0  26.5500        S  First woman  TRUE        1
13      NA   20     0     0   8.0500        S  Third   man  TRUE        0
14       0   39     1     5  31.2750        S  Third   man FALSE        0
15       1   14     0     0   7.8542        S  Third child  TRUE        0
16       1   55     0     0  16.0000        S Second woman  TRUE        1
17       0    2     4     1  29.1250        Q  Third child FALSE        0
18       0 <NA>     0     0  13.0000        S Second   man  TRUE        1
19       1   31     1     0  18.0000        S  Third woman FALSE        0
20       1 <NA>     0     0   7.2250        C  Third woman  TRUE        1
21       0   35     0     0  26.0000        S Second   man  TRUE        0
22       0   34     0     0  13.0000        S Second   man  TRUE        1
23       1   15     0     0   8.0292        Q  Third child  TRUE        1
24       0   28     0     0  35.5000        S  First   man  TRUE        1
25       1    8     3     1  21.0750        S   <NA> child FALSE        0
26       1   38     1     5  31.3875        S  Third woman FALSE        1
27       0 <NA>     0     0   7.2250        C  Third   man  TRUE        0
28       0   19     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     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     0     0  10.5000        S Second   man  TRUE        0
35       0   28     1     0  82.1708        C  First   man FALSE        0
36       0   42     1     0  52.0000        S  First   man FALSE        0
37       0 <NA>     0     0   7.2292        C  Third   man  TRUE        1
38       0   21     0     0   8.0500        S  Third   man  TRUE        0
39       1   18     2     0  18.0000        S  Third woman FALSE        0
40       1   14     1     0  11.2417        C  Third child FALSE        1
41       1   40     1     0   9.4750        S  Third woman FALSE        0
42       1   27     1     0  21.0000        S Second woman FALSE        0
43       0 <NA>     0     0   7.8958        C  Third   man  TRUE        0
44       1    3     1     2  41.5792        C Second child FALSE        1
45       1   19     0     0   7.8792        Q  Third woman  TRUE        1
46       0 <NA>     0     0   8.0500        S  Third   man  TRUE        0
```

R-studio-directory - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Background Jobs

R 4.3.1 · D:/Data-science-project/R-studio-directory/

45	1	19	0	0	7.8792	Q	Third	woman	TRUE	1
46	0	<NA>	0	0	8.0500	S	Third	man	TRUE	0
47	0	<NA>	1	0	15.5000	Q	Third	man	FALSE	0
48	1	<NA>	0	0	7.7500	Q	Third	woman	TRUE	1
49	0	<NA>	2	0	21.6792	C	Third	man	FALSE	0
50	1	18	1	0	17.8000	S	Third	woman	FALSE	0
51	0	7	4	1	39.6875	S	Third	child	FALSE	0
52	NA	21	0	0	7.8000	S	Third	man	TRUE	0
53	1	49	1	0	76.7292	C	First	woman	FALSE	1
54	1	29	1	0	26.0000	S	Second	woman	FALSE	1
55	0	65	0	1	61.9792	C	First	man	FALSE	0
56	NA	<NA>	0	0	35.5000	S	First	man	TRUE	1
57	1	21	0	0	10.5000	S	Second	woman	TRUE	1
58	0	28.5	0	0	7.2292	C	Third	man	TRUE	0
59	1	5	1	2	27.7500	S	Second	child	FALSE	1
60	0	11	5	2	46.9000	S	Third	child	FALSE	0
61	0	22	0	0	7.2292	C	Third	man	TRUE	0
62	1	38	0	0	80.0000	<NA>	First	woman	TRUE	1
63	0	45	1	0	83.4750	S	First	man	FALSE	0
64	0	4	3	2	27.9000	S	Third	child	FALSE	0
65	0	<NA>	0	0	27.7208	C	First	man	TRUE	0
66	0	<NA>	1	1	15.2458	C	Third	man	FALSE	1
67	1	29	0	0	10.5000	S	Second	woman	TRUE	1
68	0	19	0	0	8.1583	S	Third	man	TRUE	0
69	1	17	4	2	7.9250	S	Third	woman	FALSE	1
70	0	26	2	0	8.6625	S	Third	man	FALSE	0
71	0	32	0	0	10.5000	S	Second	man	TRUE	0
72	1	16	5	2	46.9000	S	Third	woman	FALSE	0
73	0	21	0	0	73.5000	S	Second	man	TRUE	0
74	0	26	1	0	14.4542	C	Third	man	FALSE	0
75	0	32	0	0	56.4958	S	Third	man	TRUE	1
76	0	25	0	0	7.6500	S	Third	man	TRUE	0
77	NA	<NA>	0	0	7.8958	S	Third	man	TRUE	0
78	0	<NA>	0	0	8.0500	S	Third	man	TRUE	0
79	0	0.83	0	2	29.0000	S	Second	child	FALSE	1
80	1	30	0	0	12.4750	S	Third	woman	TRUE	1
81	0	22	0	0	9.0000	S	Third	man	TRUE	0
82	0	29	0	0	9.5000	S	Third	man	TRUE	1
83	1	<NA>	0	0	7.7875	Q	Third	woman	TRUE	1
84	0	28	0	0	47.1000	S	First	man	TRUE	0
85	1	17	0	0	10.5000	S	Second	woman	TRUE	1
86	1	33	3	0	15.8500	S	Third	woman	FALSE	1
87	0	16	1	3	34.3750	S	Third	man	FALSE	0
88	0	<NA>	0	0	8.0500	S	Third	man	TRUE	0
89	1	23	3	2	263.0000	S	First	woman	FALSE	1
90	0	24	0	0	8.0500	S	Third	man	TRUE	0
91	0	29	0	0	8.0500	S	Third	man	TRUE	0
92	0	20	0	0	7.8542	S	Third	man	TRUE	0

2.2 Now, as “gender” and “age” columns are in the numerical format we can replace the missing value with the “mean value” of those columns. R code for replacing missing value by the mean, Now, the dataset looks like this-

```
> Titanic$gender <- ifelse(is.na(Titanic$gender), mean(Titanic$gender, na.rm = TRUE), Titanic$gender)
>
> Titanic$age <- ifelse(is.na(Titanic$age), mean(Titanic$age, na.rm = TRUE), Titanic$age)
>
> print(Titanic)
       gender      age sibsp parch     fare embarked  class   who alone survived
1   0.0000000 22.00000     1     0   7.2500        S  Third   man FALSE        0
2   1.0000000 38.00000     1     0  71.2833        C  First woman  FALL        1
3   1.0000000 26.00000     0     0   7.9250        S  Third woman  TRUE        1
4   1.0000000 35.00000     1     0  53.1000        S  First woman  FALL        1
5   0.0000000 35.00000     0     0   8.0500        S  Third   man  TRUE        0
6   0.0000000 33.32837     0     0   8.4583        Q  Third   man  TRUE        0
7   0.0000000 54.00000     0     0  51.8625        S  First   man  TRUE        0
8   0.0000000  2.00000     3     1  21.0750        S  Third child FALSE        0
9   1.0000000 27.00000     0     2  11.1333        S  Third woman FALSE        1
10  1.0000000 14.00000     1     0  30.0708        C Second child FALSE        1
11  1.0000000  4.00000     1     1  16.7000        S  Third child FALSE        1
12  1.0000000 58.00000     0     0  26.5500        S  First woman  TRUE        1
13  0.3628692 20.00000     0     0   8.0500        S  Third   man  TRUE        0
14  0.0000000 39.00000     1     5  31.2750        S  Third   man FALSE        0
15  1.0000000 14.00000     0     0   7.8542        S  Third child  TRUE        0
16  1.0000000 55.00000     0     0  16.0000        S Second woman  TRUE        1
17  0.0000000  2.00000     4     1  29.1250        Q  Third child FALSE        0
18  0.0000000 33.32837     0     0  13.0000        S Second   man  TRUE        1
19  1.0000000 31.00000     1     0  18.0000        S  Third woman FALSE        0
20  1.0000000 33.32837     0     0   7.2250        C  Third woman  TRUE        1
21  0.0000000 35.00000     0     0  26.0000        S Second   man  TRUE        0
22  0.0000000 34.00000     0     0  13.0000        S Second   man  TRUE        1
23  1.0000000 15.00000     0     0   8.0292        Q  Third child  TRUE        1
24  0.0000000 28.00000     0     0  35.5000        S  First   man  TRUE        1
25  1.0000000  8.00000     3     1  21.0750        S        child FALSE        0
26  1.0000000 38.00000     1     5  31.3875        S  Third woman FALSE        1
27  0.0000000 33.32837     0     0   7.2250        C  Third   man  TRUE        0
28  0.0000000 19.00000     3     2 263.0000        S  First   man FALSE        0
29  1.0000000 33.32837     0     0   7.8792        Q  Third woman  TRUE        1
30  0.0000000 33.32837     0     0   7.8958        S  Third   man  TRUE        0
31  0.0000000 40.00000     0     0  27.7208        C  First   man  TRUE        0
32  1.0000000 33.32837     1     0 146.5208        C  First woman FALSE        1
33  1.0000000 33.32837     0     0   7.7500        Q  Third woman  TRUE        1
34  0.3628692 66.00000     0     0  10.5000        S Second   man  TRUE        0
35  0.0000000 28.00000     1     0  82.1708        C  First   man FALSE        0
36  0.0000000 42.00000     1     0  52.0000        S  First   man FALSE        0
37  0.0000000 33.32837     0     0   7.2292        C  Third   man  TRUE        1
38  0.0000000 21.00000     0     0   8.0500        S  Third   man  TRUE        0
39  1.0000000 18.00000     2     0  18.0000        S  Third woman FALSE        0
40  1.0000000 14.00000     1     0  11.2417        C  Third child FALSE        1
41  1.0000000 40.00000     1     0   9.4750        S  Third woman FALSE        0
42  1.0000000 27.00000     1     0  21.0000        S Second woman FALSE        0
```

R-studio-directory - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Console | Terminal | Background Jobs

R 4.3.1 · D:/Data-science-project/R-studio-directory/

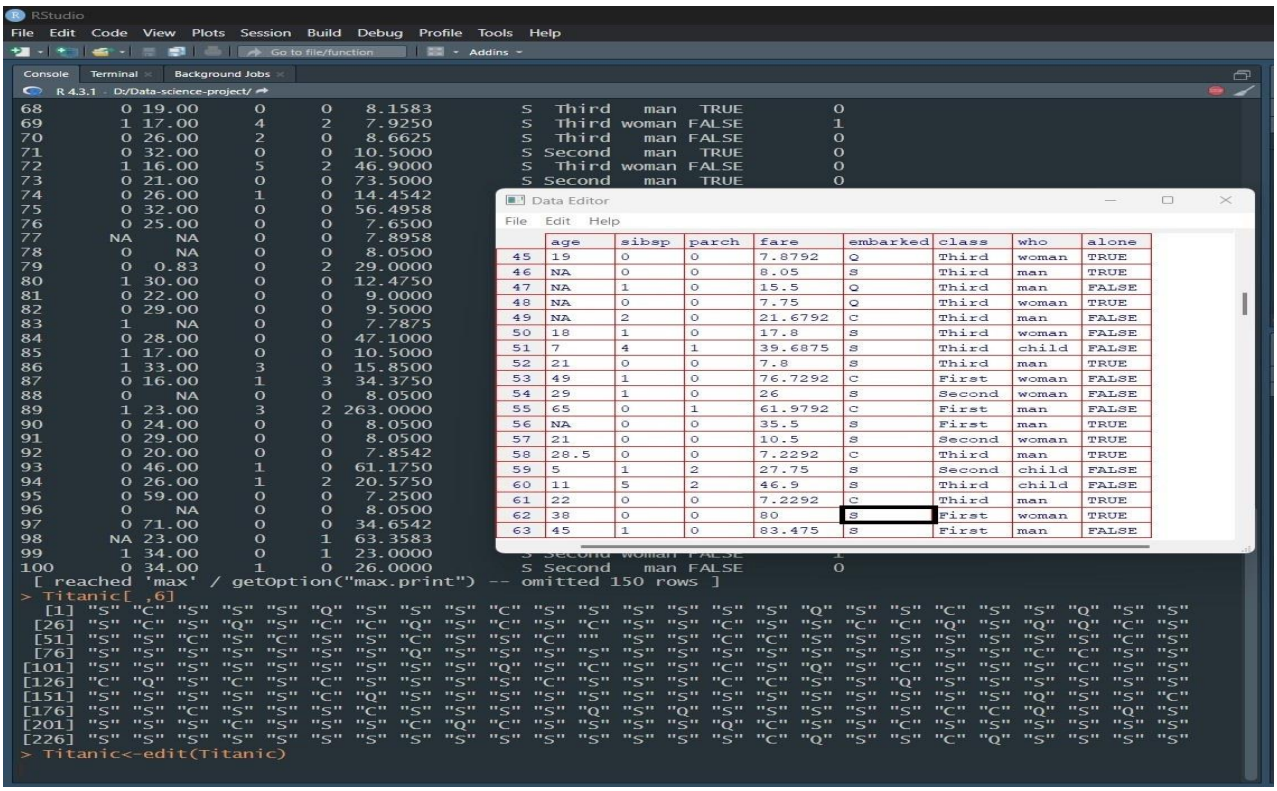
44	1.0000000	3.00000	1	2	41.5792	C	Second	child	FALSE	1
45	1.0000000	19.00000	0	0	7.8792	Q	Third	woman	TRUE	1
46	0.0000000	33.32837	0	0	8.0500	S	Third	man	TRUE	0
47	0.0000000	33.32837	1	0	15.5000	Q	Third	man	FALSE	0
48	1.0000000	33.32837	0	0	7.7500	Q	Third	woman	TRUE	1
49	0.0000000	33.32837	2	0	21.6792	C	Third	man	FALSE	0
50	1.0000000	18.00000	1	0	17.8000	S	Third	woman	FALSE	0
51	0.0000000	7.00000	4	1	39.6875	S	Third	child	FALSE	0
52	0.3628692	21.00000	0	0	7.8000	S	Third	man	TRUE	0
53	1.0000000	49.00000	1	0	76.7292	C	First	woman	FALSE	1
54	1.0000000	29.00000	1	0	26.0000	S	Second	woman	FALSE	1
55	0.0000000	65.00000	0	1	61.9792	C	First	man	FALSE	0
56	0.3628692	33.32837	0	0	35.5000	S	First	man	TRUE	1
57	1.0000000	21.00000	0	0	10.5000	S	Second	woman	TRUE	1
58	0.0000000	28.50000	0	0	7.2292	C	Third	man	TRUE	0
59	1.0000000	5.00000	1	2	27.7500	S	Second	child	FALSE	1
60	0.0000000	11.00000	5	2	46.9000	S	Third	child	FALSE	0
61	0.0000000	22.00000	0	0	7.2292	C	Third	man	TRUE	0
62	1.0000000	38.00000	0	0	80.0000	<NA>	First	woman	TRUE	1
63	0.0000000	45.00000	1	0	83.4750	S	First	man	FALSE	0
64	0.0000000	4.00000	3	2	27.9000	S	Third	child	FALSE	0
65	0.0000000	33.32837	0	0	27.7208	C	First	man	TRUE	0
66	0.0000000	33.32837	1	1	15.2458	C	Third	man	FALSE	1
67	1.0000000	29.00000	0	0	10.5000	S	Second	woman	TRUE	1
68	0.0000000	19.00000	0	0	8.1583	S	Third	man	TRUE	0
69	1.0000000	17.00000	4	2	7.9250	S	Third	woman	FALSE	1
70	0.0000000	26.00000	2	0	8.6625	S	Third	man	FALSE	0
71	0.0000000	32.00000	0	0	10.5000	S	Second	man	TRUE	0
72	1.0000000	16.00000	5	2	46.9000	S	Third	woman	FALSE	0
73	0.0000000	21.00000	0	0	73.5000	S	Second	man	TRUE	0
74	0.0000000	26.00000	1	0	14.4542	C	Third	man	FALSE	0
75	0.0000000	32.00000	0	0	56.4958	S	Third	man	TRUE	1
76	0.0000000	25.00000	0	0	7.6500	S	Third	man	TRUE	0
77	0.3628692	33.32837	0	0	7.8958	S	Third	man	TRUE	0
78	0.0000000	33.32837	0	0	8.0500	S	Third	man	TRUE	0
79	0.0000000	0.83000	0	2	29.0000	S	Second	child	FALSE	1
80	1.0000000	30.00000	0	0	12.4750	S	Third	woman	TRUE	1
81	0.0000000	22.00000	0	0	9.0000	S	Third	man	TRUE	0
82	0.0000000	29.00000	0	0	9.5000	S	Third	man	TRUE	1
83	1.0000000	33.32837	0	0	7.7875	Q	Third	woman	TRUE	1
84	0.0000000	28.00000	0	0	47.1000	S	First	man	TRUE	0
85	1.0000000	17.00000	0	0	10.5000	S	Second	woman	TRUE	1
86	1.0000000	33.00000	3	0	15.8500	S	Third	woman	FALSE	1
87	0.0000000	16.00000	1	3	34.3750	S	Third	man	FALSE	0
88	0.0000000	33.32837	0	0	8.0500	S	Third	man	TRUE	0
89	1.0000000	23.00000	3	2	263.0000	S	First	woman	FALSE	1
90	0.0000000	24.00000	0	0	8.0500	S	Third	man	TRUE	0
91	0.0000000	29.00000	0	0	8.0500	S	Third	man	TRUE	0

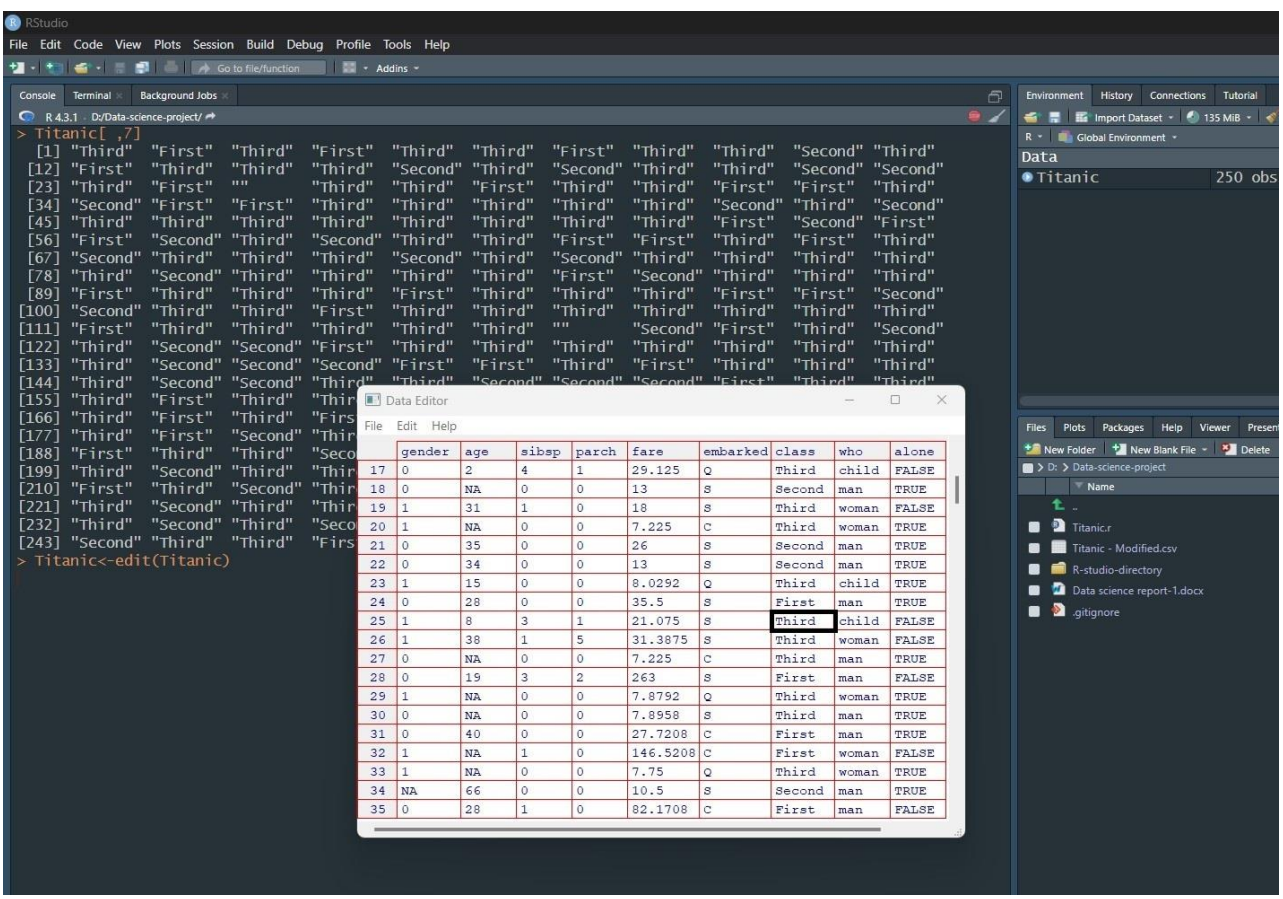
2.3 Here we can see that in the “embarked” and “class” column, some values are missing.

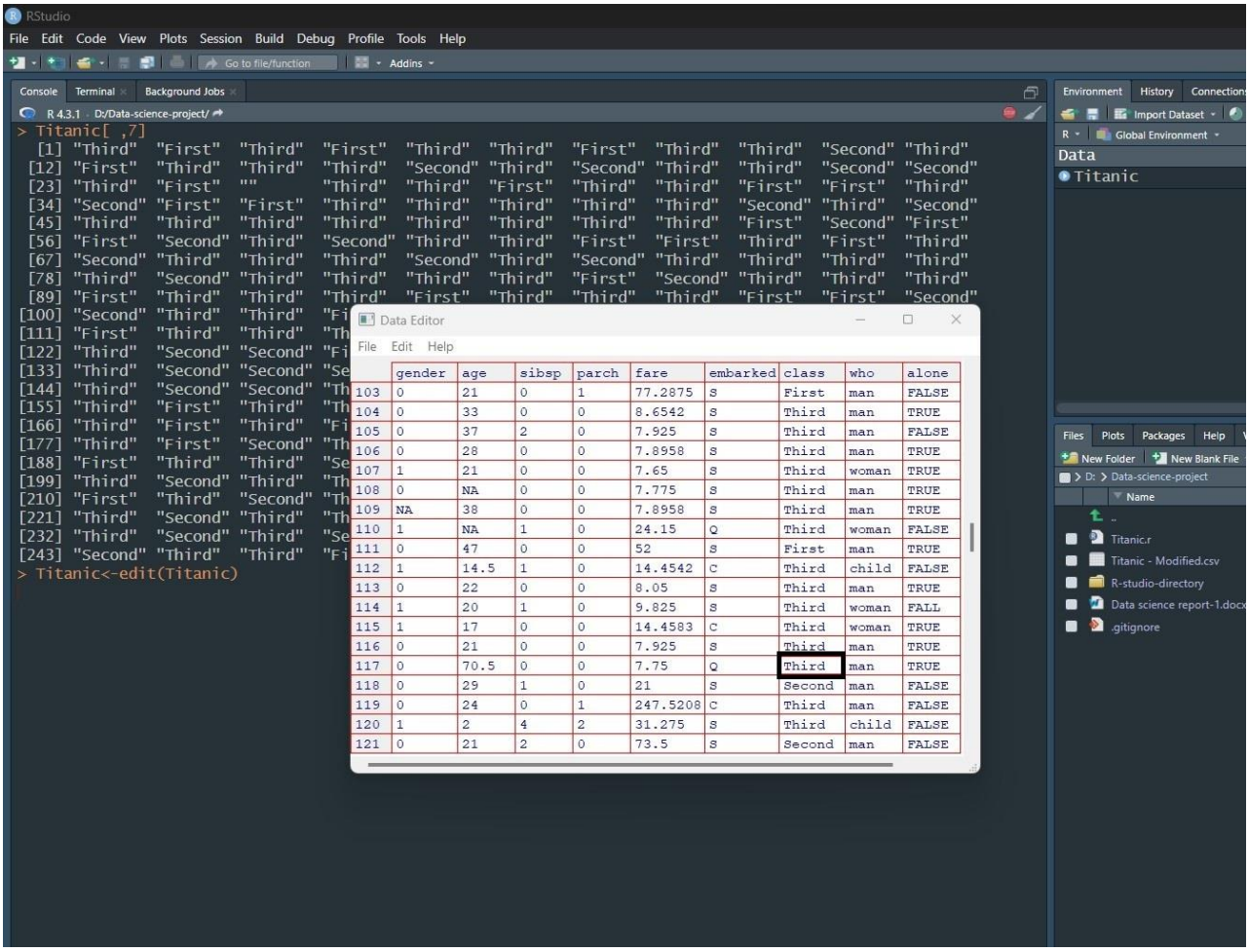
```
> Titanic[ ,6]
  [1] "S" "C" "S" "S" "S" "Q" "S" "S" "S" "C" "S" "S" "S" "S" "S" "S" "Q" "S" "S" "C" "S" "S" "Q" "S" "S"
 [26] "S" "C" "S" "Q" "S" "C" "C" "Q" "S" "C" "S" "C" "S" "S" "C" "S" "S" "C" "C" "Q" "S" "Q" "Q" "C" "S"
 [51] "S" "S" "C" "S" "C" "S" "S" "C" "S" "S" "C" "" "S" "S" "C" "C" "S" "S" "S" "S" "S" "S" "S" "C" "S"
 [76] "S" "S" "S" "S" "S" "S" "S" "Q" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "C" "C" "S" "S"
[101] "S" "S" "S" "S" "S" "S" "S" "S" "S" "Q" "S" "C" "S" "S" "C" "S" "Q" "S" "C" "S" "S" "S" "C" "S" "S"
[126] "C" "Q" "S" "C" "S" "C" "S" "S" "S" "S" "C" "S" "S" "S" "C" "C" "S" "S" "Q" "S" "S" "S" "S" "S" "S"
[151] "S" "S" "S" "S" "S" "C" "Q" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "Q" "S" "S" "C"
[176] "S" "S" "C" "S" "S" "S" "C" "S" "S" "S" "S" "Q" "S" "Q" "S" "S" "S" "S" "S" "S" "C" "C" "Q" "S" "Q" "S"
[201] "S" "S" "S" "C" "S" "S" "S" "C" "Q" "C" "S" "S" "S" "S" "Q" "C" "S" "S" "C" "S" "S" "S" "S" "S" "S"
[226] "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "S" "C" "Q" "S" "S" "C" "Q" "S" "S" "S" "S"
>
```

```
> Titanic[ ,7]
  [1] "Third"  "First"  "Third"  "First"  "Third"  "Third"  "First"  "Third"  "Third"  "Second" "Third"
 [12] "First"  "Third"  "Third"  "Third"  "Second" "Third"  "Second" "Third"  "Third"  "Second" "Second"
 [23] "Third"  "First"  ""       "Third"  "Third"  "First"  "Third"  "Third"  "First"  "First"  "Third"
 [34] "Second" "First"  "First"  "Third"  "Third"  "Third"  "Third"  "Third"  "Second" "Third"  "Second"
 [45] "Third"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"  "First"  "Second" "First"
 [56] "First"  "Second" "Third"  "Second" "Third"  "Third"  "First"  "First"  "Third"  "First"  "Third"
 [67] "Second" "Third"  "Third"  "Third"  "Second" "Third"  "Second" "Third"  "Third"  "Third"  "Third"
 [78] "Third"  "Second" "Third"  "Third"  "Third"  "Third"  "First"  "Second" "Third"  "Third"  "Third"
 [89] "First"  "Third"  "Third"  "Third"  "First"  "Third"  "Third"  "Third"  "First"  "First"  "Second"
[100] "Second" "Third"  "Third"  "First"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"
[111] "First"  "Third"  "Third"  "Third"  "Third"  "Third"  ""       "Second" "First"  "Third"  "Second"
[122] "Third"  "Second" "Second" "First"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"
[133] "Third"  "Second" "Second" "Second" "First"  "First"  "Third"  "First"  "Third"  "Third"  "Third"
[144] "Third"  "Second" "Second" "Third"  "Third"  "Second" "Second" "Second" "First"  "Third"  "Third"
[155] "Third"  "First"  "Third"  "Third"  "Third"  "Third"  "Third"  "Second" "Third"  "Third"  "Third"
[166] "Third"  "First"  "Third"  "First"  "Third"  "First"  "Third"  "Third"  "Third"  "First"  "Third"
[177] "Third"  "First"  "Second" "Third"  "Third"  "Second" "Third"  "Second" "Third"  "First"  "Third"
[188] "First"  "Third"  "Third"  "Second" "Second" ""       "Second" "First"  "First"  "Third"  "Third"
[199] "Third"  "Second" "Third"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"  "Third"
[210] "First"  "Third"  "Second" "Third"  "Second" "Third"  "First"  "Third"  "Second" "First"  "Second"
[221] "Third"  "Second" "Third"  "Third"  "First"  "Third"  "Second" "Third"  "Second" ""       "First"
[232] "Third"  "Second" "Third"  "Second" "Third"  "Second" "Second" "Second" "Second" "Third"  "Third"
[243] "Second" "Third"  "Third"  "First"  "Third"  "Second" "First"  "Second"
>
```

As the values are not numeric but “string”, we can solve the issue by putting the most common value of the column this way-

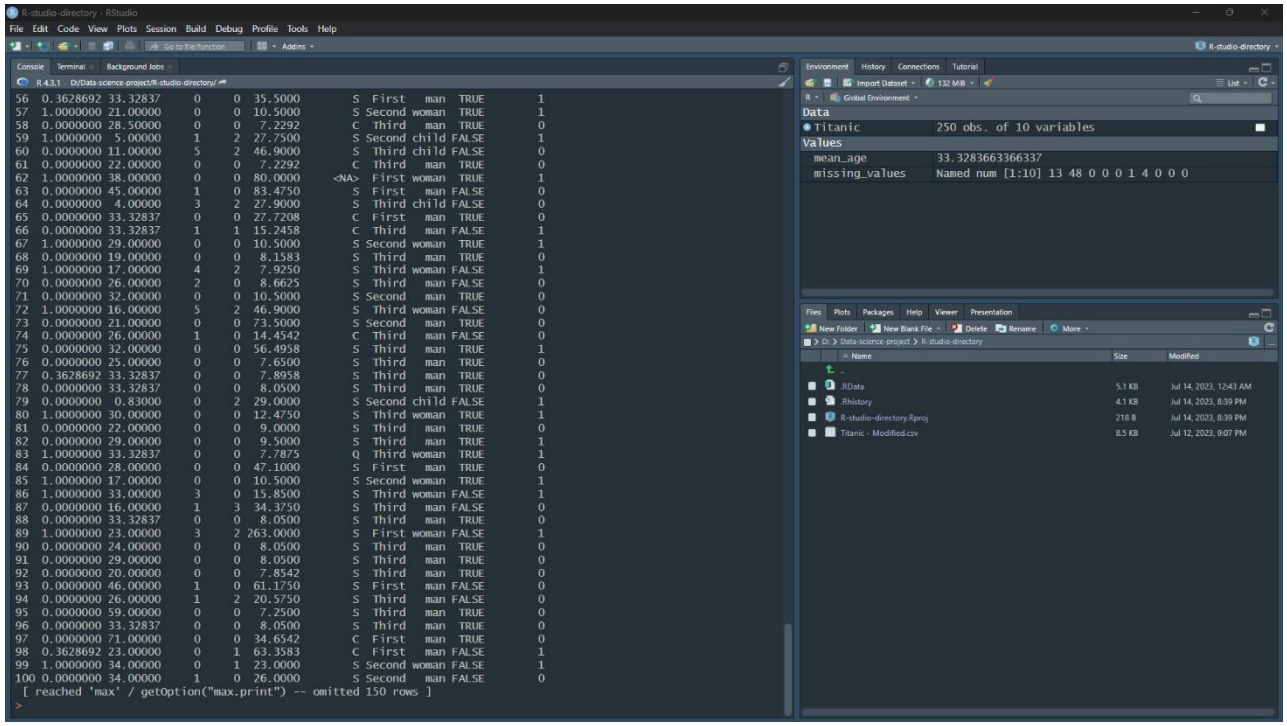






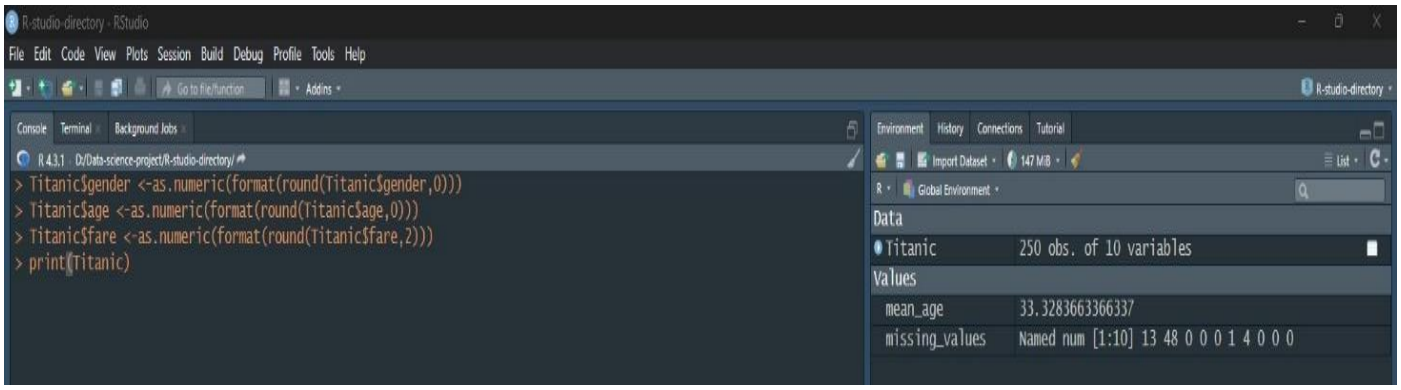
	gender	age	sibsp	parch	fare	embarked	class	who	alone
103	0	21	0	1	77.2875	S	First	man	FALSE
104	0	33	0	0	8.6542	S	Third	man	TRUE
105	0	37	2	0	7.925	S	Third	man	FALSE
106	0	28	0	0	7.8958	S	Third	man	TRUE
107	1	21	0	0	7.65	S	Third	woman	TRUE
108	0	NA	0	0	7.775	S	Third	man	TRUE
109	NA	38	0	0	7.8958	S	Third	man	TRUE
110	1	NA	1	0	24.15	Q	Third	woman	FALSE
111	0	47	0	0	52	S	First	man	TRUE
112	1	14.5	1	0	14.4542	C	Third	child	FALSE
113	0	22	0	0	8.05	S	Third	man	TRUE
114	1	20	1	0	9.825	S	Third	woman	FALL
115	1	17	0	0	14.4583	C	Third	woman	TRUE
116	0	21	0	0	7.925	S	Third	man	TRUE
117	0	70.5	0	0	7.75	Q	Third	man	TRUE
118	0	29	1	0	21	S	Second	man	FALSE
119	0	24	0	1	247.5208	C	Third	man	FALSE
120	1	2	4	2	31.275	S	Third	child	FALSE
121	0	21	2	0	73.5	S	Second	man	FALSE

3. Dealing with Data types and Conversion:

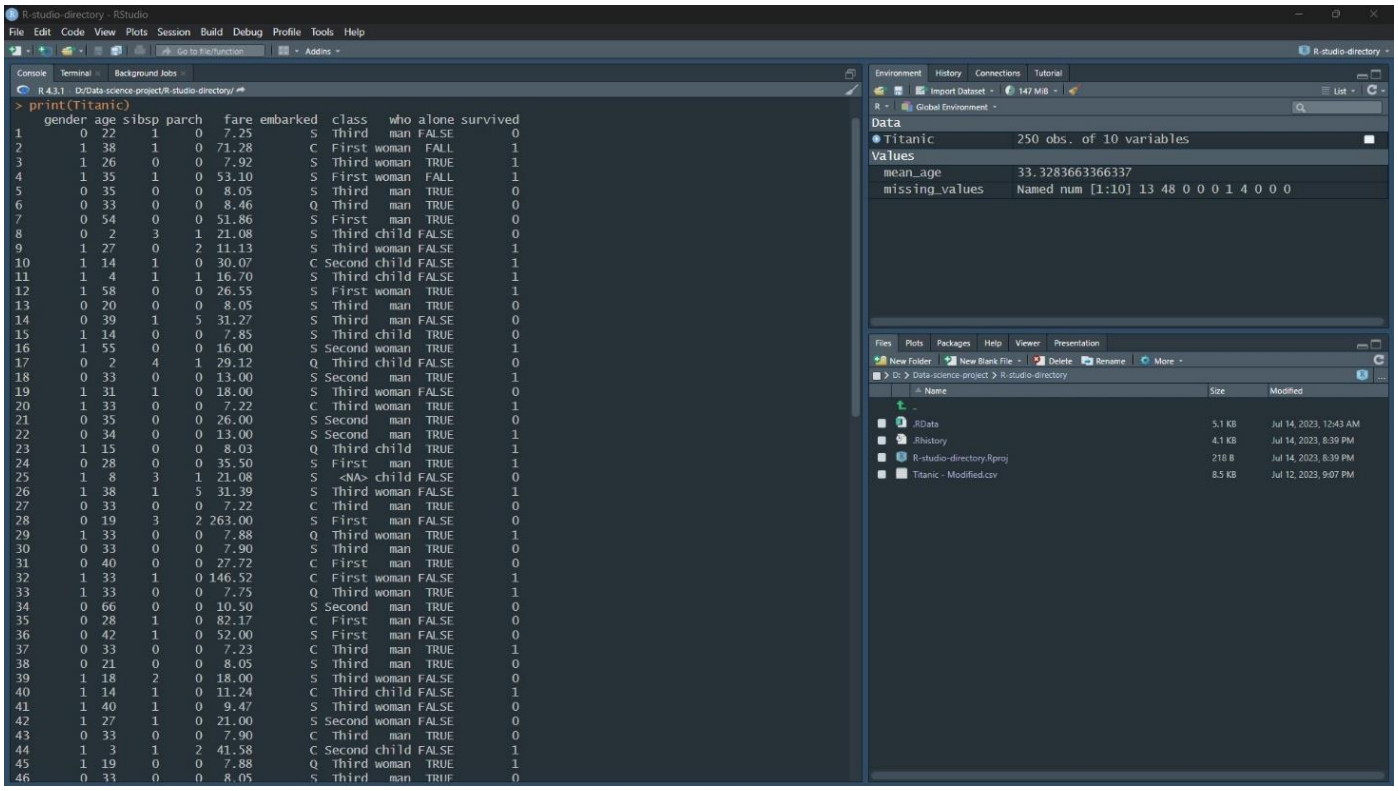


Now we can see that the maximum column contains decimal places in the data after dealing with null values in those columns. As we are not interested in having decimal places for those columns, we will round it up.

We can round those variables in this way-



Datset with round figure-



4. Dealing with Outliers:

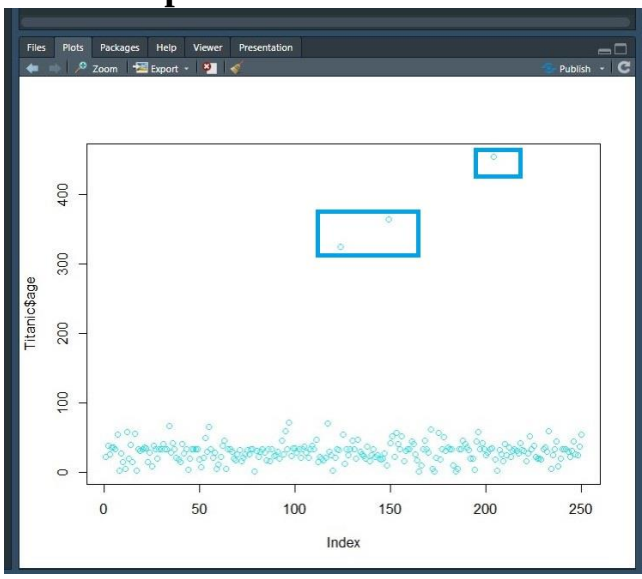
Here, we can check whether there are any outliers or not. So, to find it out we can use this way-

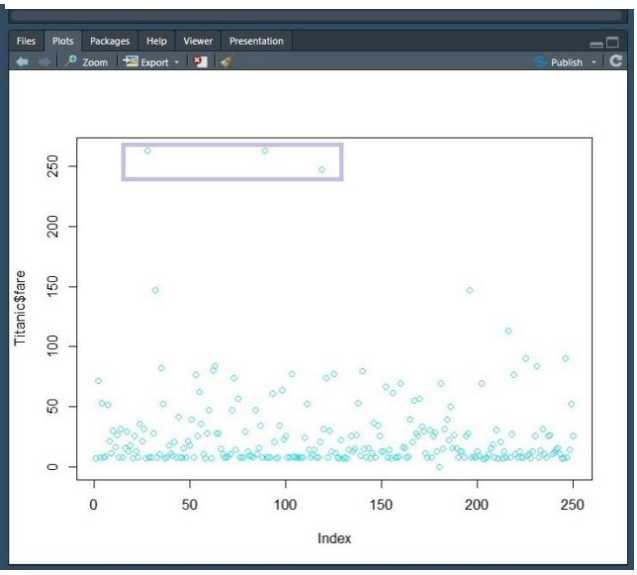
R code-

```
R-studio-directory - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R 4.3.1 · D:/Data-science-project/R-studio-directory/
> plot(Titanic$age,col=5)
>
```

```
R-studio-directory - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R 4.3.1 · D:/Data-science-project/R-studio-directory/
> plot(Titanic$fare,col=5)
>
```

Output-

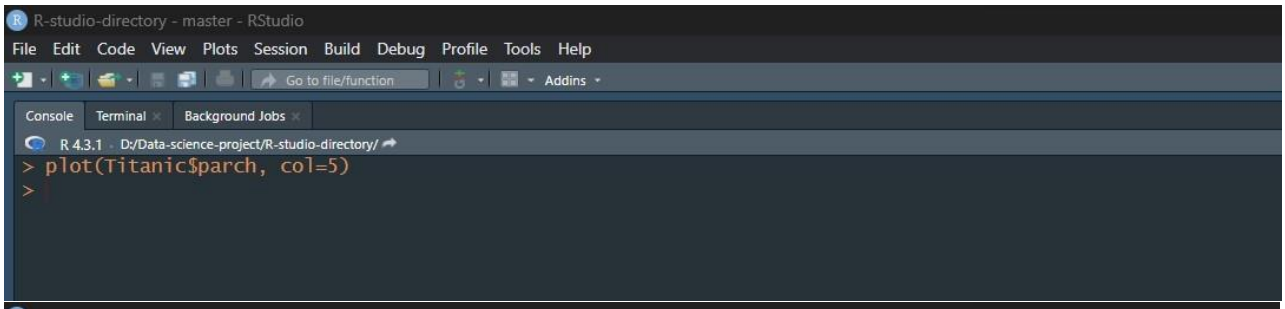


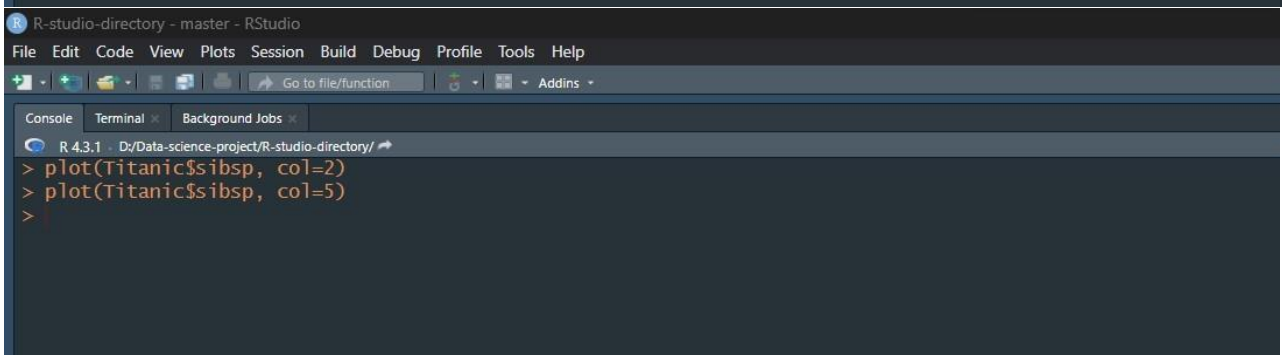


So, here we can see that in column $age and $fare , there are three points or data which are quite different from the rest of the points. So, this can be called Outliers.

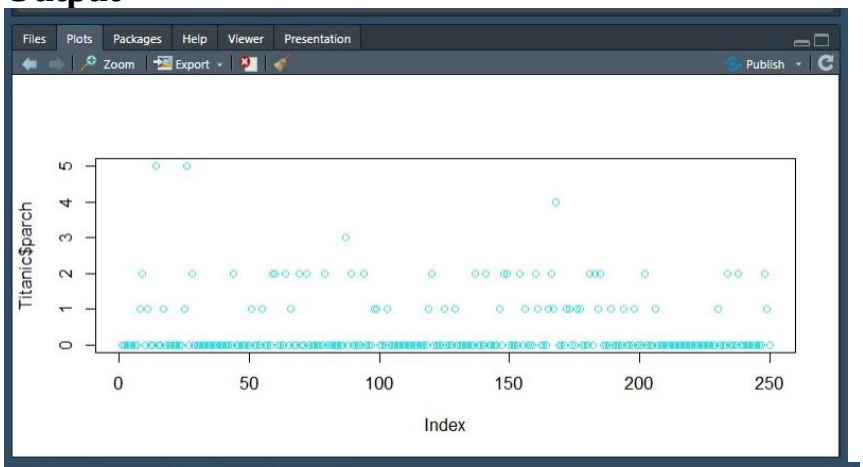
But for other attributes- $parch and $sibsp we didn't find any outliers because some values may not usual compared to most but those result can be possible. Given below-

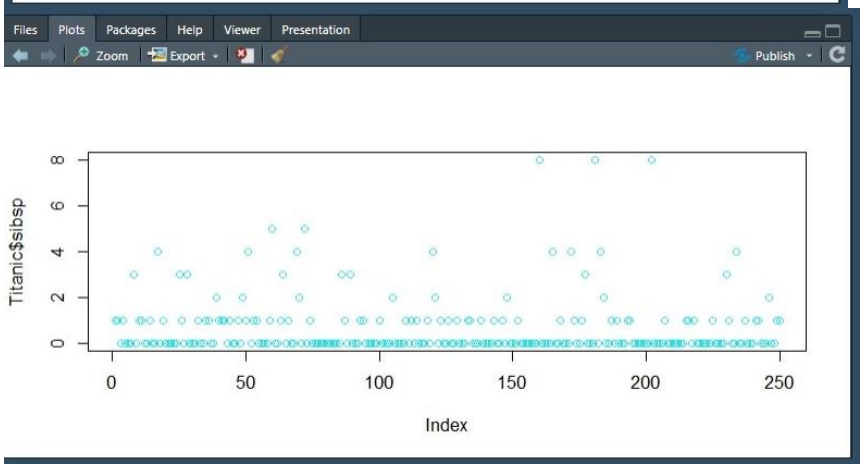
R code-





Output-





Representation of Data:

Using barplot we can evaluate some insights from dataset such as-

R code-

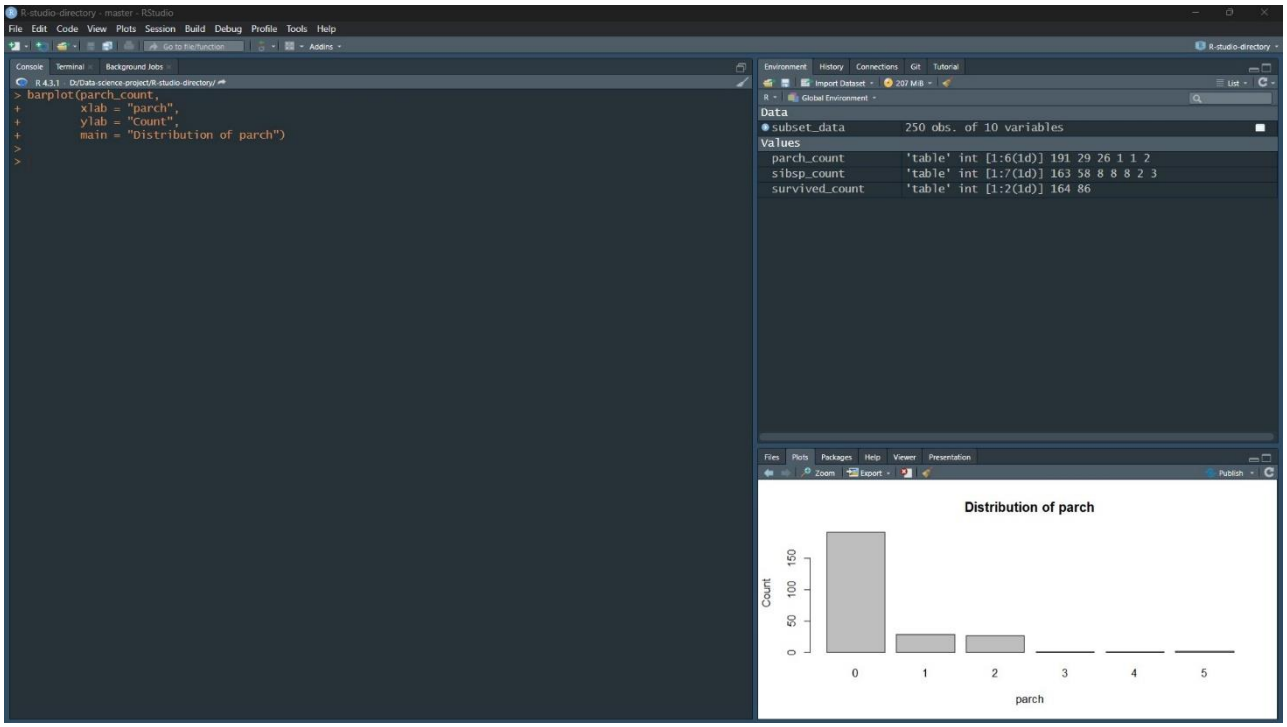
```
> gender_counts <-table(Titanic$gender)
> barplot(gender_counts,names.arg = c("0","1"),xlab = "gender index", ylab = "gender counts")
>
```

Output-

Here “0” represents male and “1” female in genders , from which we can extract above 150 amongst the passenger were male and below 100 but above 50 were female amongst 250 passengers dataset.

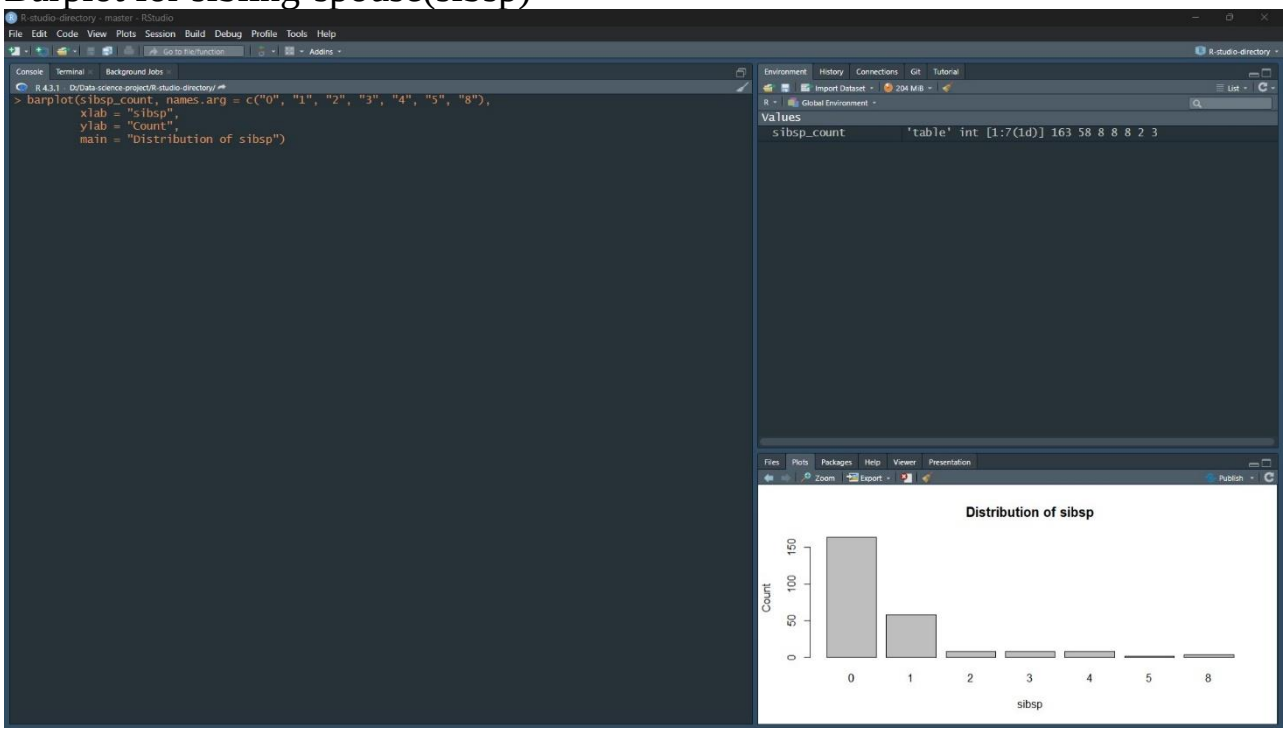
Using this method we can get barplot and can discuss what it represent for other attributes as well.

Barplot for parch -



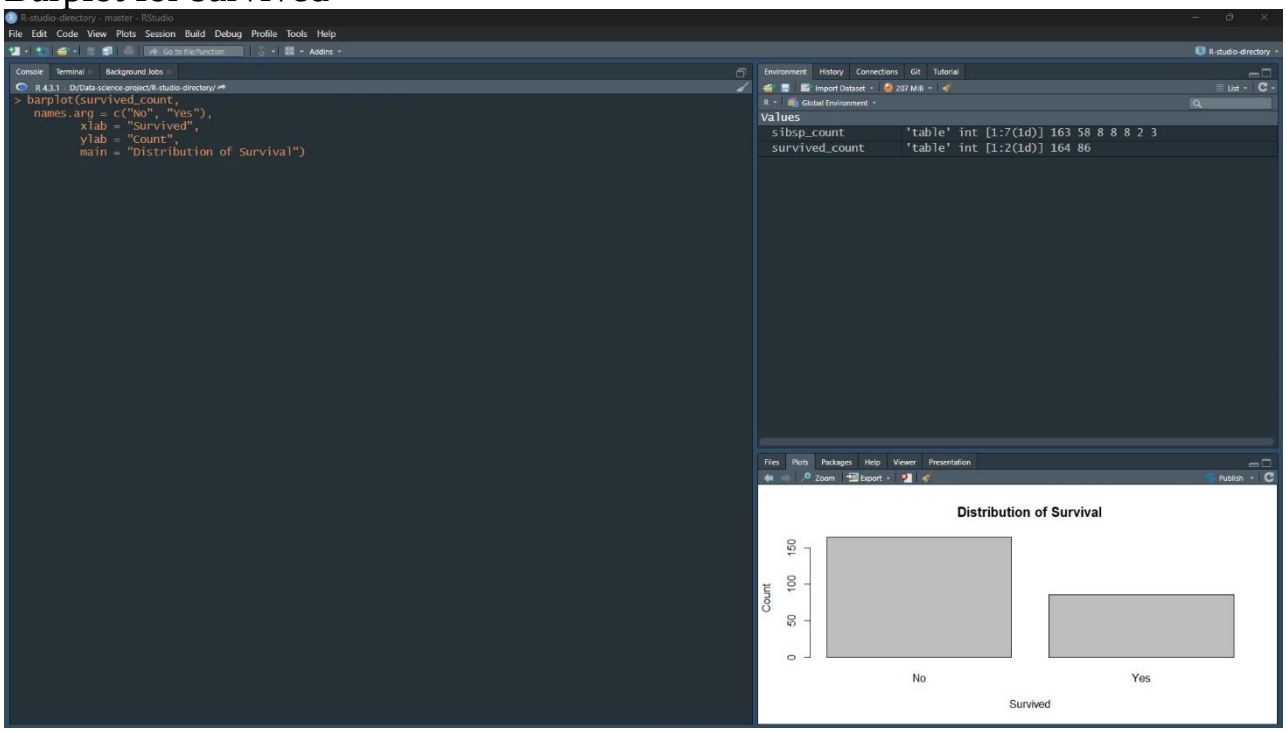
From the barplot, we see more than 150 passengers had not any parents or children amongst the 250 passenger dataset. Less than 50 passenger had 1 and 2 parents or children. Less than 50 passengers had from 3 up-to 5 parents and children which might represent with passengers whole family.

Barplot for sibling-spouse(sibsp)-



From the barplot, we see more than 150 passengers had not any siblings or spouse amongst the 250 passenger dataset. Around 50 passenger had 1 sibling or spouse. Less than 50 passengers had from 2 up-to 8 siblings and spouse.

Barplot for survived-



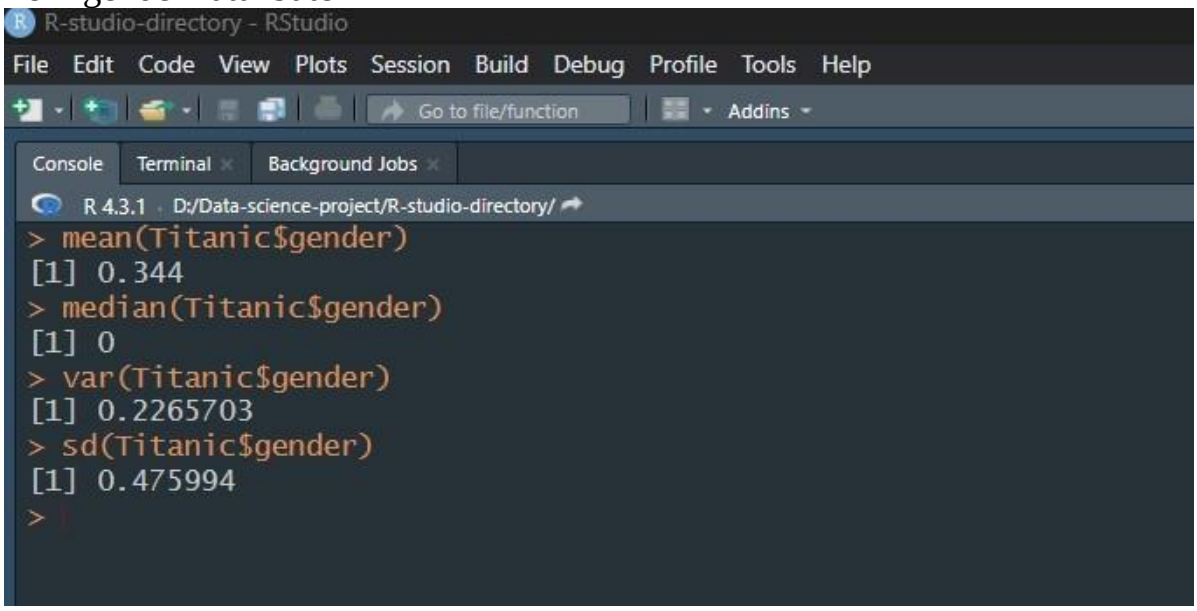
From the barplot, we see around more than 50 and less than 100 passengers have survived but the not survival rate is higher than 150 passengers amongst the given 250 passengers dataset.

Univariate data exploration:

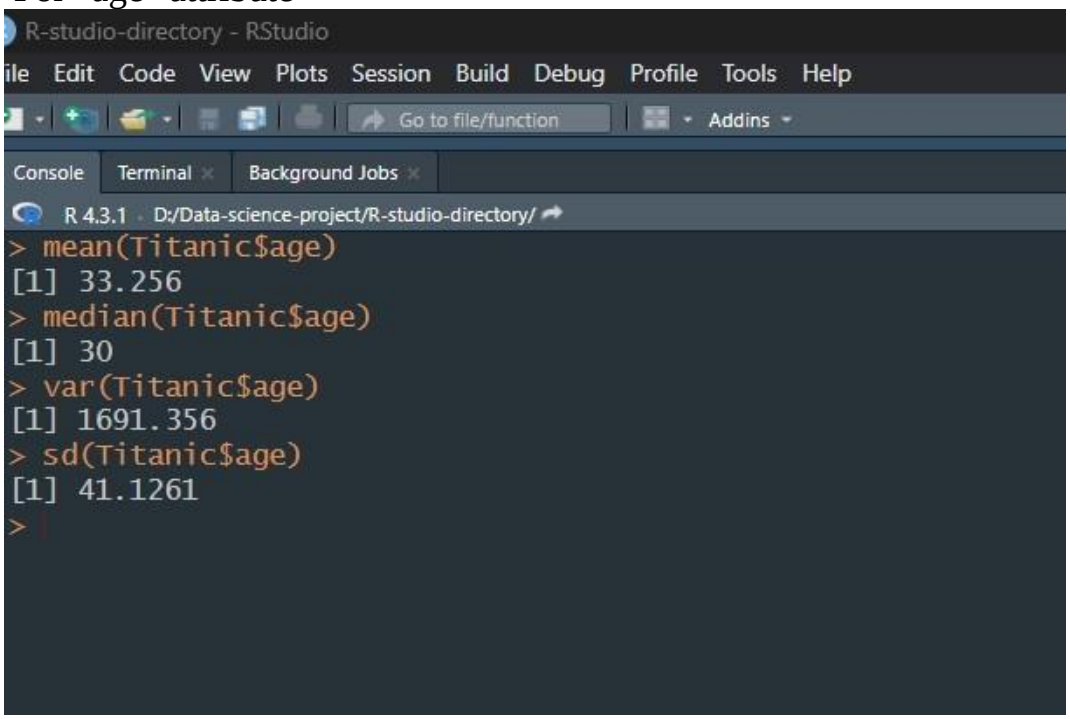
Univariate exploration in data science involves analyzing individual variables in a dataset one at a time, without considering the relationship between variables. This type of analysis is useful for gaining a basic understanding of the distribution, central tendency, and variability of a variable.

5. So, here we can find out the exploration of the given attribute-

For “gender” attribute



For “age” attribute



For “sibsp” attribute

```
Console   Terminal   Background Jobs
R 4.3.1 · D:/Data-science-project/
> mean(Titanic$sibsp)
[1] 0.656
>
> median(Titanic$sibsp)
[1] 0
> var(Titanic$sibsp)
[1] 1.704482
> sd(Titanic$sibsp)
[1] 1.305558
>
```

For “parch” attribute

```
Console   Terminal   Background Jobs
R 4.3.1 · D:/Data-science-project/
> mean(Titanic$parch)
[1] 0.392
>
> median(Titanic$parch)
[1] 0
> var(Titanic$parch)
[1] 0.6810602
> sd(Titanic$parch)
[1] 0.8252637
>
```

For “fare” attribute

```
R-studio-directory - RStudio
File  Edit  Code  View  Plots  Session  Build  Debug  Profile  Tools  Help
Go to file/function   Addins
Console   Terminal   Background Jobs
R 4.3.1 · D:/Data-science-project/R-studio-directory/
> mean(Titanic$fare)
[1] 26.58748
> median(Titanic$fare)
[1] 13.975
> var(Titanic$fare)
[1] 1212.548
> sd(Titanic$fare)
[1] 34.82166
>
```

For “survived” attribute

```
Console   Terminal   Background Jobs
R 4.3.1 · D:/Data-science-project/
> mean(Titanic$survived)
[1] 0.344
>
> median(Titanic$survived)
[1] 0
> var(Titanic$survived)
[1] 0.2265703
> sd(Titanic$survived)
[1] 0.475994
>
```

So we can summarize from above attributes-

Attributes	Gender	Age	fare	parch	survived
Mean	0.344	33.256	26.587	0.392	0.344
Median	0	30	13.975	0	0
Variance	0.226	1691.365	1212.548	0.6841	0.226
Standard deviation	0.476	41.126	34.821	0.825	0.476

6. Standard deviation of each attribute (gender, age, sibsp, parch, fare & survived)

```
R-studio-directory - RStudio
File  Edit  Code  View  Plots  Session  Build  Debug  Profile  Tools  Help
Go to file/function   Addins
Console   Terminal   Background Jobs
R 4.3.1 · D:/Data-science-project/R-studio-directory/
> Titanic%>% summarise_if(is.numeric,sd)
    gender     age    sibsp     parch     fare survived
1 0.475994 41.1261 1.305558 0.8252637 34.82166 0.475994
>
```

Here, we calculated the standard deviation of each numerical attribute.

Discussion & Conclusion:

At the beginning of the project, we were given a dataset that was totally messy. Null values, missing values, and outliers were present in this dataset. The dataset was like this-

R-studio-directory - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

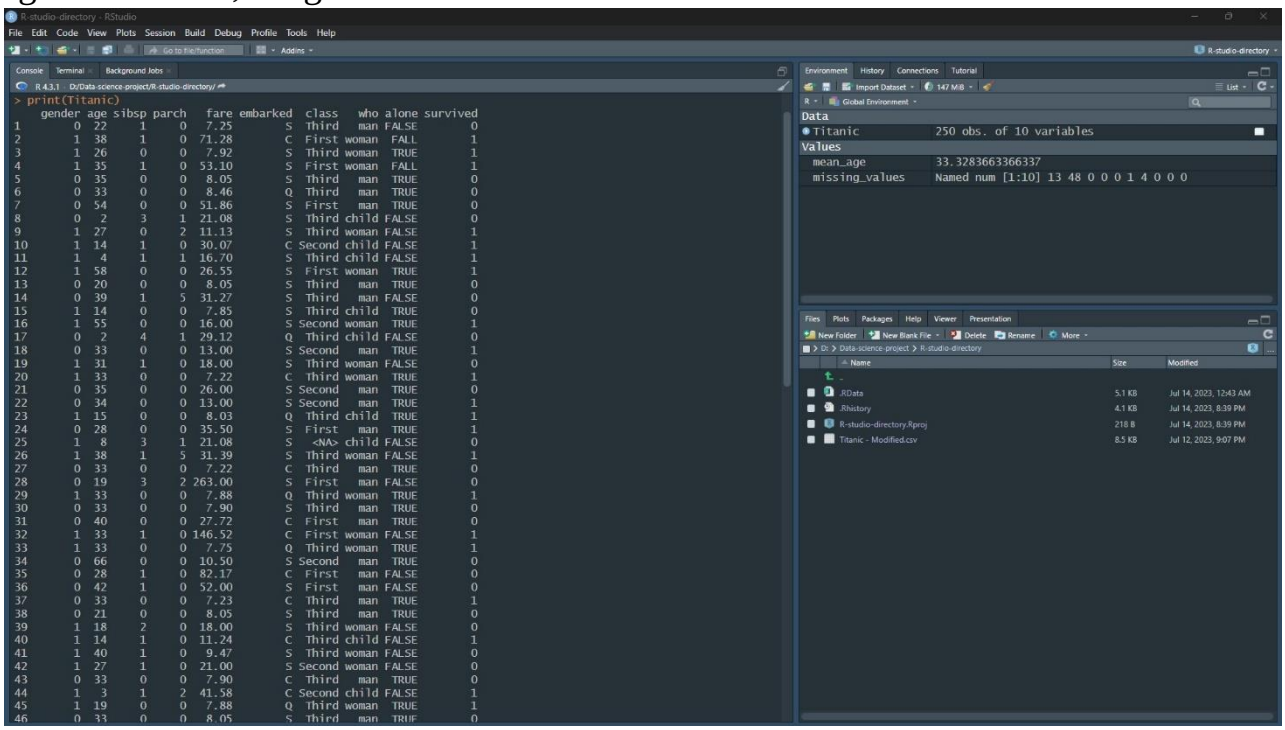
Go to file/function Addins

Console Terminal Background Jobs

R 4.3.1 · D:/Data-science-project/R-studio-directory/

```
> print(Titanic)
    gender  age sibsp parch     fare embarked  class   who alone survived
1        0   22     1     0   7.2500        S  Third   man FALSE        0
2        1   38     1     0  71.2833        C  First woman  FALL        1
3        1   26     0     0   7.9250        S  Third woman  TRUE        1
4        1   35     1     0  53.1000        S  First woman  FALL        1
5        0   35     0     0   8.0500        S  Third   man  TRUE        0
6        0 <NA>     0     0   8.4583        Q  Third   man  TRUE        0
7        0   54     0     0  51.8625        S  First   man  TRUE        0
8        0    2     3     1  21.0750        S  Third child FALSE        0
9        1   27     0     2  11.1333        S  Third woman FALSE        1
10       1   14     1     0  30.0708        C Second child FALSE        1
11       1    4     1     1  16.7000        S  Third child FALSE        1
12       1   58     0     0  26.5500        S  First woman  TRUE        1
13      NA   20     0     0   8.0500        S  Third   man  TRUE        0
14       0   39     1     5  31.2750        S  Third   man FALSE        0
15       1   14     0     0   7.8542        S  Third child  TRUE        0
16       1   55     0     0  16.0000        S Second woman  TRUE        1
17       0    2     4     1  29.1250        Q  Third child FALSE        0
18       0 <NA>     0     0  13.0000        S Second   man  TRUE        1
19       1   31     1     0  18.0000        S  Third woman FALSE        0
20       1 <NA>     0     0   7.2250        C  Third woman  TRUE        1
21       0   35     0     0  26.0000        S Second   man  TRUE        0
22       0   34     0     0  13.0000        S Second   man  TRUE        1
23       1   15     0     0   8.0292        Q  Third child  TRUE        1
24       0   28     0     0  35.5000        S  First   man  TRUE        1
25       1    8     3     1  21.0750        S   <NA> child FALSE        0
26       1   38     1     5  31.3875        S  Third woman FALSE        1
27       0 <NA>     0     0   7.2250        C  Third   man  TRUE        0
28       0   19     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     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     0     0  10.5000        S Second   man  TRUE        0
35       0   28     1     0  82.1708        C  First   man FALSE        0
36       0   42     1     0  52.0000        S  First   man FALSE        0
37       0 <NA>     0     0   7.2292        C  Third   man  TRUE        1
38       0   21     0     0   8.0500        S  Third   man  TRUE        0
39       1   18     2     0  18.0000        S  Third woman FALSE        0
40       1   14     1     0  11.2417        C  Third child FALSE        1
41       1   40     1     0   9.4750        S  Third woman FALSE        0
42       1   27     1     0  21.0000        S Second woman FALSE        0
43       0 <NA>     0     0   7.8958        C  Third   man  TRUE        0
44       1    3     1     2  41.5792        C Second child FALSE        1
45       1   19     0     0   7.8792        Q  Third woman  TRUE        1
46       0 <NA>     0     0   8.0500        S  Third   man  TRUE        0
```

After Applying data preparation steps and the univariate data exploration for the given data set., we got the dataset looks like this-



Now, we can use this clean, pre-processed dataset for further use.