Data Preprocessing in R

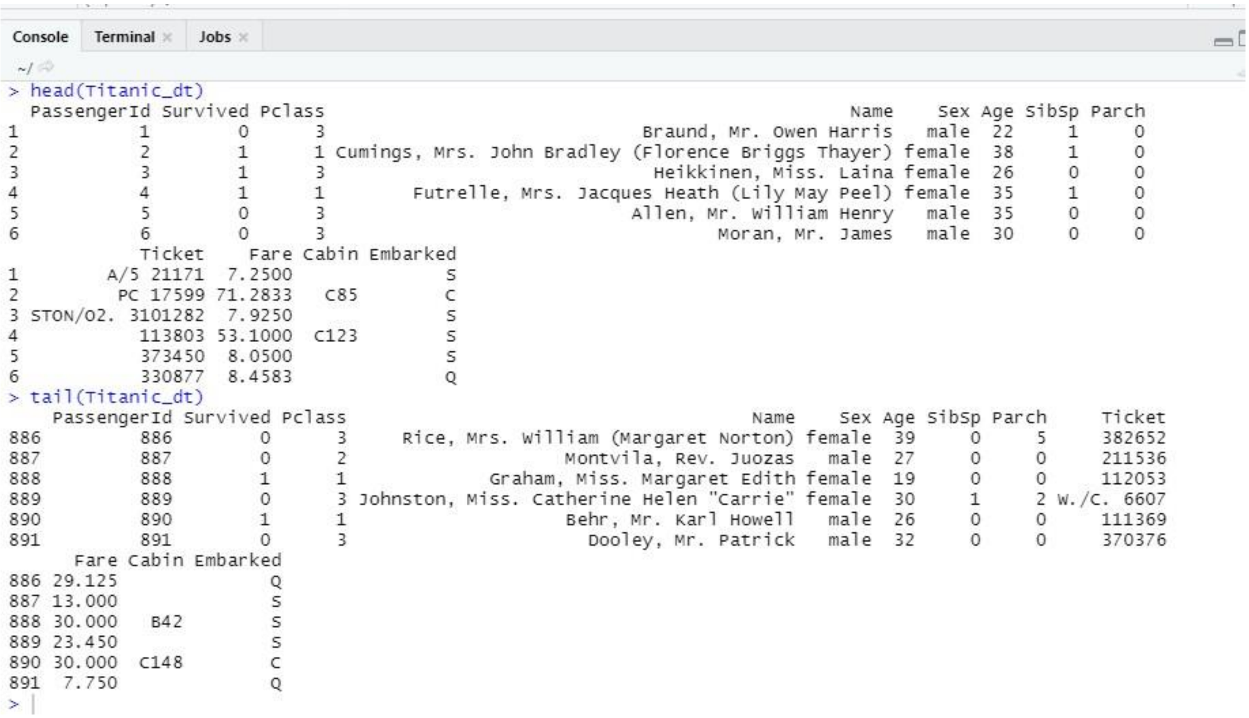
Use following data for this exercise:

titanic_df<-read.csv("D:/MCA_R/Titanic.csv")

marks**<-c(22,NA,45,30,NA,50,20)**

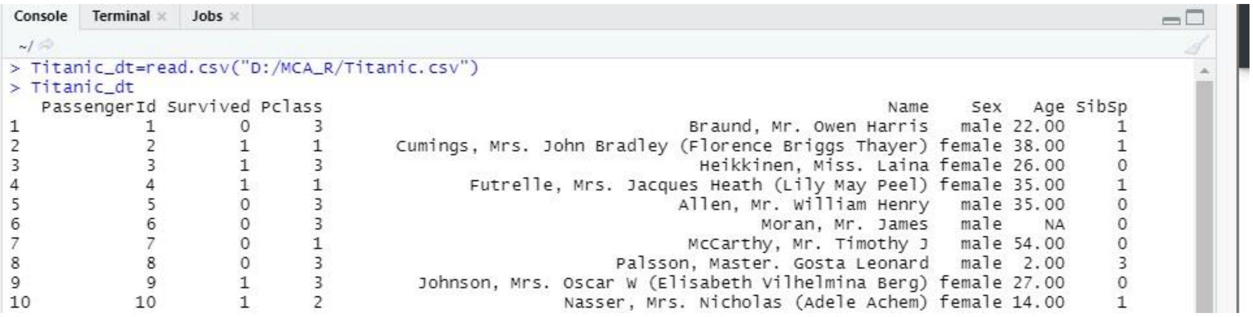
1.Naming and renaming variables, adding a new variable.

1. Load titanic data in R environment and 1) Display first 5 rows 2) Display last 5 rows



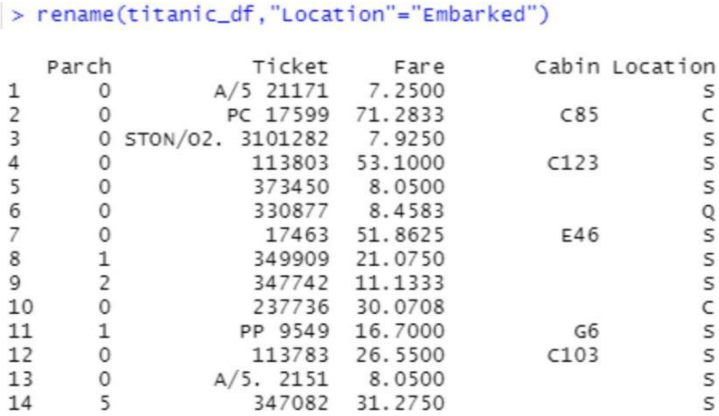
```
> head(Titanic_dt)
  PassengerId Survived Pclass                                                Name    Sex Age SibSp Parch
1           1        0      3                             Braund, Mr. Owen Harris   male  22     1     0
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
3           3        1      3                              Heikkinen, Miss. Laina female  26     0     0
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
5           5        0      3                            Allen, Mr. William Henry   male  35     0     0
6           6        0      3                                    Moran, Mr. James   male  30     0     0
            Ticket    Fare Cabin Embarked
1        A/5 21171  7.2500             S
2         PC 17599 71.2833   C85        C
3 STON/O2. 3101282  7.9250             S
4           113803 53.1000  C123        S
5           373450  8.0500             S
6           330877  8.4583             Q
> tail(Titanic_dt)
    PassengerId Survived Pclass                                     Name    Sex Age SibSp Parch     Ticket
886         886        0      3     Rice, Mrs. William (Margaret Norton) female  39     0     5     382652
887         887        0      2                    Montvila, Rev. Juozas   male  27     0     0     211536
888         888        1      1             Graham, Miss. Margaret Edith female  19     0     0     112053
889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female  30     1     2 W./C. 6607
890         890        1      1                    Behr, Mr. Karl Howell   male  26     0     0     111369
891         891        0      3                      Dooley, Mr. Patrick   male  32     0     0     370376
      Fare Cabin Embarked
886 29.125             Q
887 13.000             S
888 30.000   B42        S
889 23.450             S
890 30.000  C148        C
891  7.750             Q
> |
```

2. Display the first 5 columns of the titanic dataset.



```
> Titanic_dt=read.csv("D:/MCA_R/Titanic.csv")
> Titanic_dt
   PassengerId Survived Pclass                                                Name    Sex   Age SibSp
1            1        0      3                             Braund, Mr. Owen Harris   male 22.00     1
2            2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.00     1
3            3        1      3                              Heikkinen, Miss. Laina female 26.00     0
4            4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.00     1
5            5        0      3                            Allen, Mr. William Henry   male 35.00     0
6            6        0      3                                    Moran, Mr. James   male    NA     0
7            7        0      1                             McCarthy, Mr. Timothy J   male 54.00     0
8            8        0      3                      Palsson, Master. Gosta Leonard   male  2.00     3
9            9        1      3   Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female 27.00     0
10          10        1      2                 Nasser, Mrs. Nicholas (Adele Achem) female 14.00     1
```

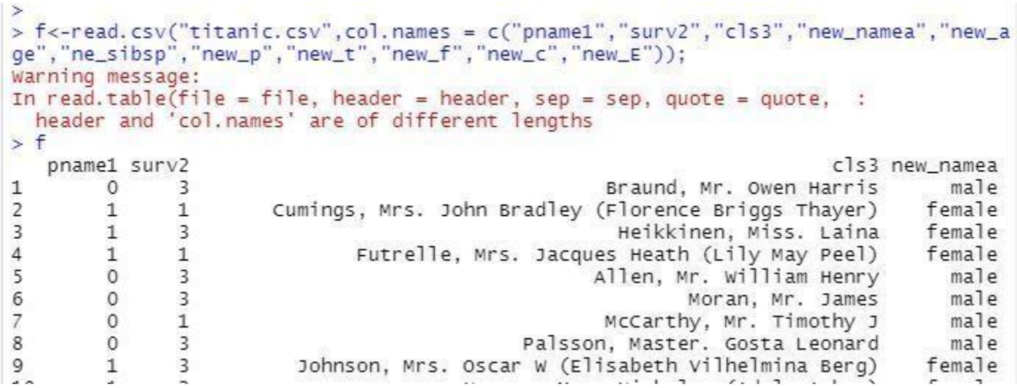
3. Rename the column Embarked with name Location of titanic dataframe.



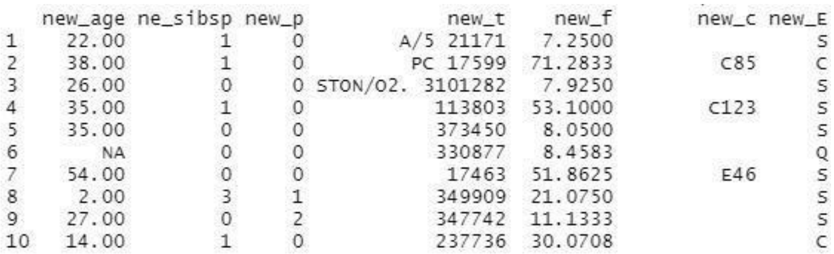
```
> rename(titanic_df,"Location"="Embarked")

   Parch           Ticket    Fare       Cabin Location
1      0        A/5 21171  7.2500                    S
2      0         PC 17599 71.2833         C85        C
3      0 STON/O2. 3101282  7.9250                    S
4      0           113803 53.1000        C123        S
5      0           373450  8.0500                    S
6      0           330877  8.4583                    Q
7      0            17463 51.8625         E46        S
8      1           349909 21.0750                    S
9      2           347742 11.1333                    S
10     0           237736 30.0708                    C
11     1          PP 9549 16.7000          G6        S
12     0           113783 26.5500        C103        S
13     0        A/5. 2151  8.0500                    S
14     5           347082 31.2750                    S
```

4. Load titanic data with user defined column name.

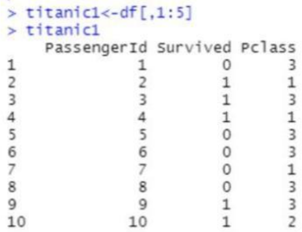


```
>
> f<-read.csv("titanic.csv",col.names = c("pname1","surv2","cls3","new_namea","new_a
ge","ne_sibsp","new_p","new_t","new_f","new_c","new_E"));
Warning message:
In read.table(file = file, header = header, sep = sep, quote = quote,  :
  header and 'col.names' are of different lengths
> f
   pname1 surv2                                                cls3 new_namea
1       0     3                             Braund, Mr. Owen Harris      male
2       1     1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)    female
3       1     3                              Heikkinen, Miss. Laina    female
4       1     1        Futrelle, Mrs. Jacques Heath (Lily May Peel)    female
5       0     3                            Allen, Mr. William Henry      male
6       0     3                                    Moran, Mr. James      male
7       0     1                             McCarthy, Mr. Timothy J      male
8       0     3                      Palsson, Master. Gosta Leonard      male
9       1     3  Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)    female
```

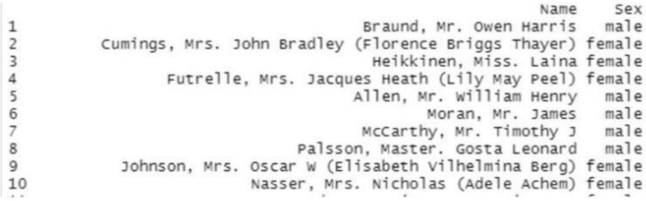


```
   new_age ne_sibsp new_p            new_t   new_f       new_c new_E
1    22.00        1     0        A/5 21171  7.2500                 S
2    38.00        1     0         PC 17599 71.2833         C85     C
3    26.00        0     0 STON/O2. 3101282  7.9250                 S
4    35.00        1     0           113803 53.1000        C123     S
5    35.00        0     0           373450  8.0500                 S
6       NA        0     0           330877  8.4583                 Q
7    54.00        0     0            17463 51.8625         E46     S
8     2.00        3     1           349909 21.0750                 S
9    27.00        0     2           347742 11.1333                 S
10   14.00        1     0           237736 30.0708                 C
```

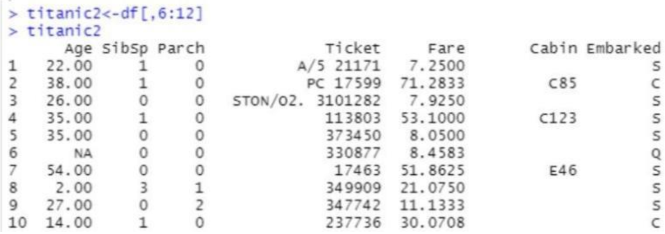
5. Load first 5 column data in dataframe titanic1 and rest of the columns in titanic2 and merge these two dataframe in titanic3



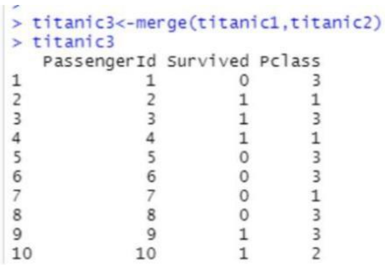
```
> titanic1<-df[,1:5]
> titanic1
   PassengerId Survived Pclass
1            1        0      3
2            2        1      1
3            3        1      3
4            4        1      1
5            5        0      3
6            6        0      3
7            7        0      1
8            8        0      3
9            9        1      3
10          10        1      2
```



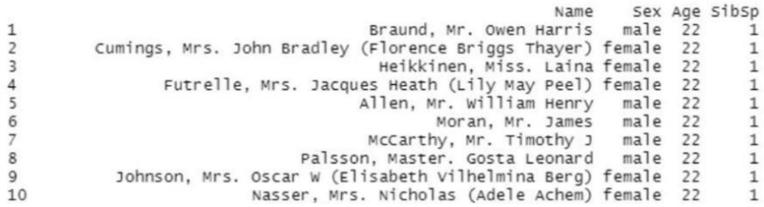
```
                                                  Name    Sex
1                              Braund, Mr. Owen Harris   male
2        Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
3                               Heikkinen, Miss. Laina female
4              Futrelle, Mrs. Jacques Heath (Lily May Peel) female
5                             Allen, Mr. William Henry   male
6                                     Moran, Mr. James   male
7                              McCarthy, Mr. Timothy J   male
8                       Palsson, Master. Gosta Leonard   male
9          Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female
10                     Nasser, Mrs. Nicholas (Adele Achem) female
```



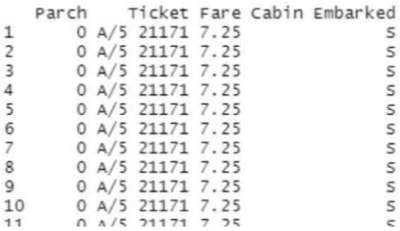
```
> titanic2<-df[,6:12]
> titanic2
     Age SibSp Parch           Ticket    Fare       Cabin Embarked
1  22.00     1     0        A/5 21171  7.2500                    S
2  38.00     1     0         PC 17599 71.2833         C85        C
3  26.00     0     0 STON/O2. 3101282  7.9250                    S
4  35.00     1     0           113803 53.1000        C123        S
5  35.00     0     0           373450  8.0500                    S
6     NA     0     0           330877  8.4583                    Q
7  54.00     0     0            17463 51.8625         E46        S
8   2.00     3     1           349909 21.0750                    S
9  27.00     0     2           347742 11.1333                    S
10 14.00     1     0           237736 30.0708                    C
```



```
> titanic3<-merge(titanic1,titanic2)
> titanic3
   PassengerId Survived Pclass
1            1        0      3
2            2        1      1
3            3        1      3
4            4        1      1
5            5        0      3
6            6        0      3
7            7        0      1
8            8        0      3
9            9        1      3
10          10        1      2
```



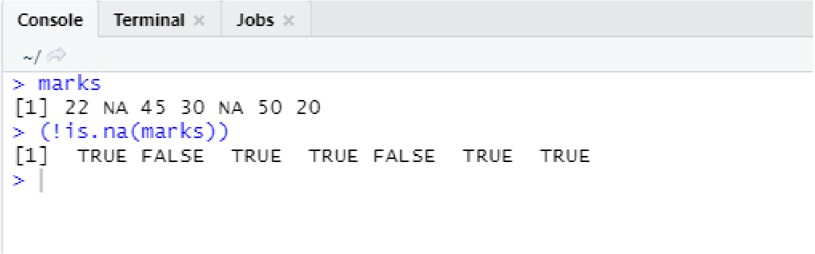
```
                                                  Name    Sex Age SibSp
1                              Braund, Mr. Owen Harris   male  22     1
2       Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  22     1
3                               Heikkinen, Miss. Laina female  22     1
4             Futrelle, Mrs. Jacques Heath (Lily May Peel) female  22     1
5                             Allen, Mr. William Henry   male  22     1
6                                     Moran, Mr. James   male  22     1
7                              McCarthy, Mr. Timothy J   male  22     1
8                       Palsson, Master. Gosta Leonard   male  22     1
9         Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female  22     1
10                     Nasser, Mrs. Nicholas (Adele Achem) female  22     1
```



```
   Parch    Ticket Fare Cabin Embarked
1      0 A/5 21171 7.25              S
2      0 A/5 21171 7.25              S
3      0 A/5 21171 7.25              S
4      0 A/5 21171 7.25              S
5      0 A/5 21171 7.25              S
6      0 A/5 21171 7.25              S
7      0 A/5 21171 7.25              S
8      0 A/5 21171 7.25              S
9      0 A/5 21171 7.25              S
10     0 A/5 21171 7.25              S
11     0 A/5 21171 7.25              S
```

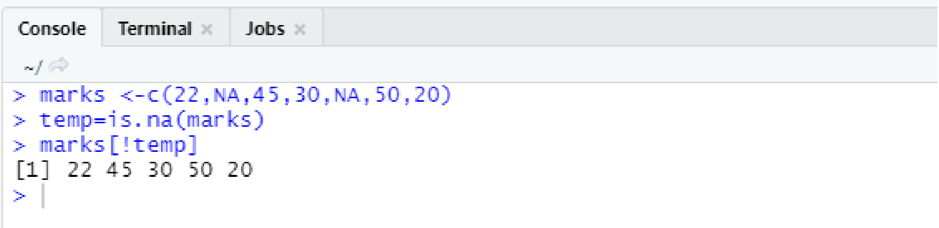
2. Dealing with Missing Data

1. Missing data are represented by NA values in R, and so we wish to check how many NA elements there are in the marks vector. Also calculate how many non NA elements are there in the vector.



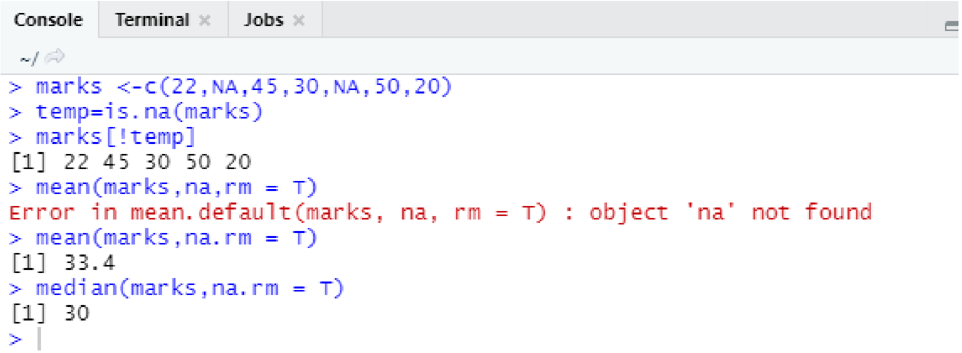
```
> marks
[1] 22 NA 45 30 NA 50 20
> (!is.na(marks))
[1]  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE
> 
```

2. Display vector marks with values that are not NA.



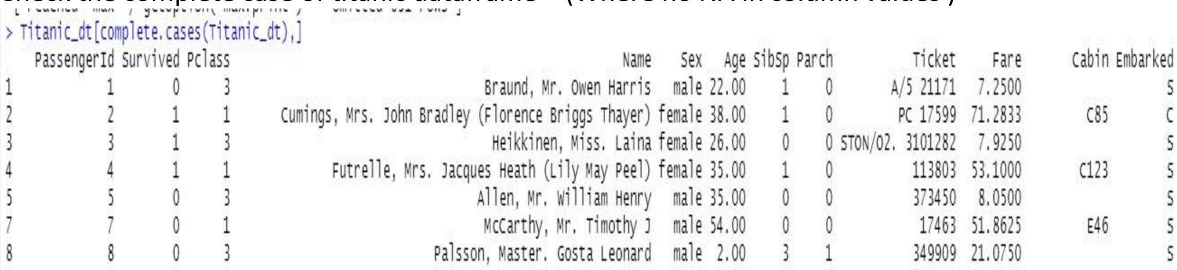
```
> marks <-c(22,NA,45,30,NA,50,20)
> temp=is.na(marks)
> marks[!temp]
[1] 22 45 30 50 20
> 
```

3. Calculate mean and median of given marks vector.

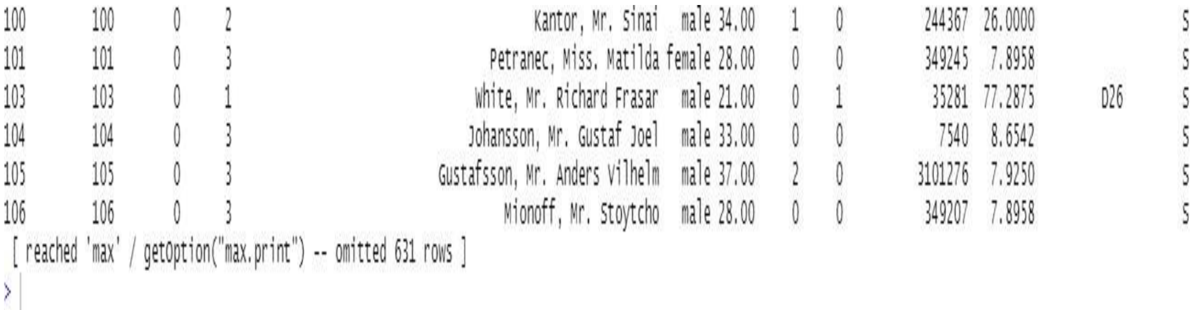


```
> marks <-c(22,NA,45,30,NA,50,20)
> temp=is.na(marks)
> marks[!temp]
[1] 22 45 30 50 20
> mean(marks,na,rm = T)
Error in mean.default(marks, na, rm = T) : object 'na' not found
> mean(marks,na.rm = T)
[1] 33.4
> median(marks,na.rm = T)
[1] 30
> 
```

4. Check the complete case of titanic dataframe – (Where no NA in column values)

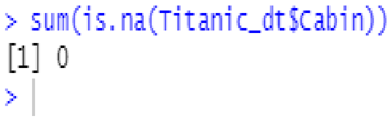


```
> Titanic_dt[complete.cases(Titanic_dt),]
    PassengerId Survived Pclass                                                Name    Sex   Age SibSp Parch           Ticket    Fare     Cabin Embarked
1             1        0      3                             Braund, Mr. Owen Harris   male 22.00     1     0        A/5 21171  7.2500                  S
2             2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.00     1     0         PC 17599 71.2833       C85        C
3             3        1      3                              Heikkinen, Miss. Laina female 26.00     0     0 STON/O2. 3101282  7.9250                  S
4             4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.00     1     0           113803 53.1000      C123        S
5             5        0      3                            Allen, Mr. William Henry   male 35.00     0     0           373450  8.0500                  S
7             7        0      1                             McCarthy, Mr. Timothy J   male 54.00     0     0            17463 51.8625       E46        S
8             8        0      3                      Palsson, Master. Gosta Leonard   male  2.00     3     1           349909 21.0750                  S
```



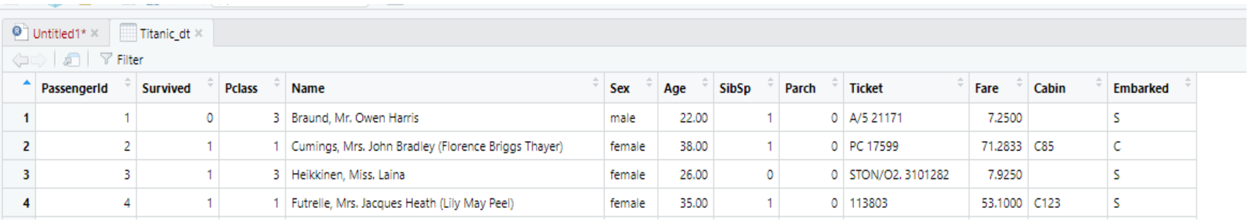
```
100         100        0      2                                   Kantor, Mr. Sinai   male 34.00     1     0           244367 26.0000                  S
101         101        0      3                             Petranec, Miss. Matilda female 28.00     0     0           349245  7.8958                  S
103         103        0      1                           White, Mr. Richard Frasar   male 21.00     0     1            35281 77.2875       D26        S
104         104        0      3                          Johansson, Mr. Gustaf Joel   male 33.00     0     0             7540  8.6542                  S
105         105        0      3                      Gustafsson, Mr. Anders Vilhelm   male 37.00     2     0          3101276  7.9250                  S
106         106        0      3                               Mionoff, Mr. Stoytcho   male 28.00     0     0           349207  7.8958                  S
 [ reached 'max' / getOption("max.print") -- omitted 631 rows ]
> 
```

5. Check the total missing values of the cabin column of the titanic dataframe without using the complete.cases function.

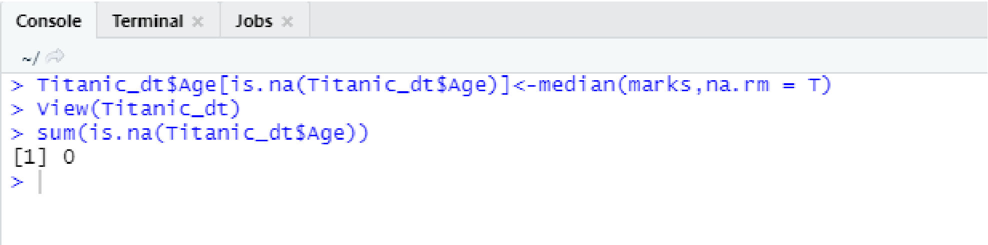


```
> sum(is.na(Titanic_dt$Cabin))
[1] 0
> 
```

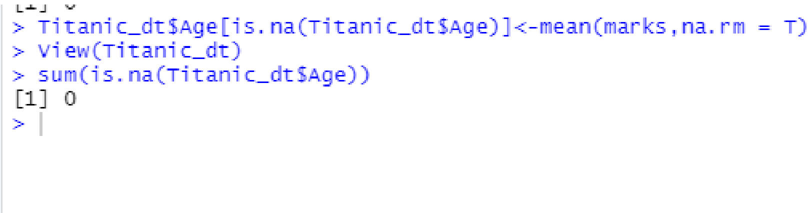
6. Replace missing value of age column with 1) mean ii) median



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S



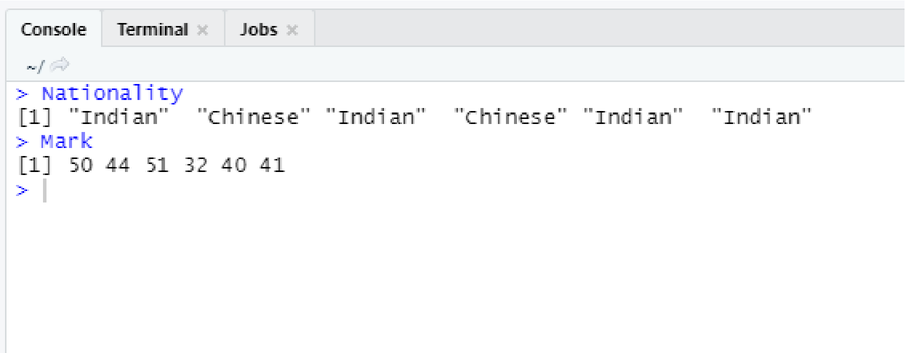
```
> Titanic_dt$Age[is.na(Titanic_dt$Age)]<-median(marks,na.rm = T)
> View(Titanic_dt)
> sum(is.na(Titanic_dt$Age))
[1] 0
> 
```



```
> Titanic_dt$Age[is.na(Titanic_dt$Age)]<-mean(marks,na.rm = T)
> View(Titanic_dt)
> sum(is.na(Titanic_dt$Age))
[1] 0
> 
```

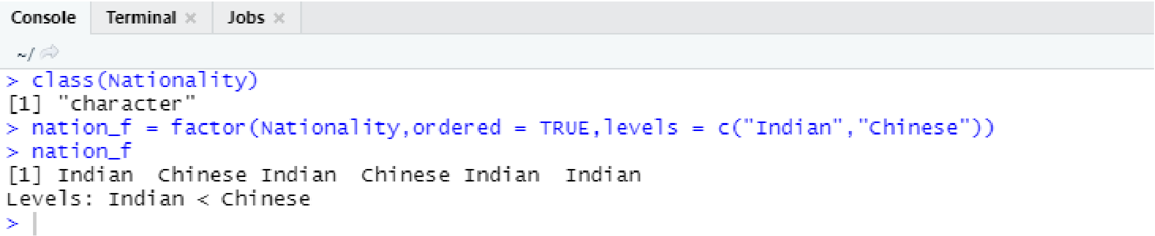
3. Dealing with categorical data.

1. Create category **Nationality** vector ("Indian", "Chinese", "Indian", "Chinese", "Indian", "Indian") and **Mark** vector (50, 44, 51, 32, 40, 41)



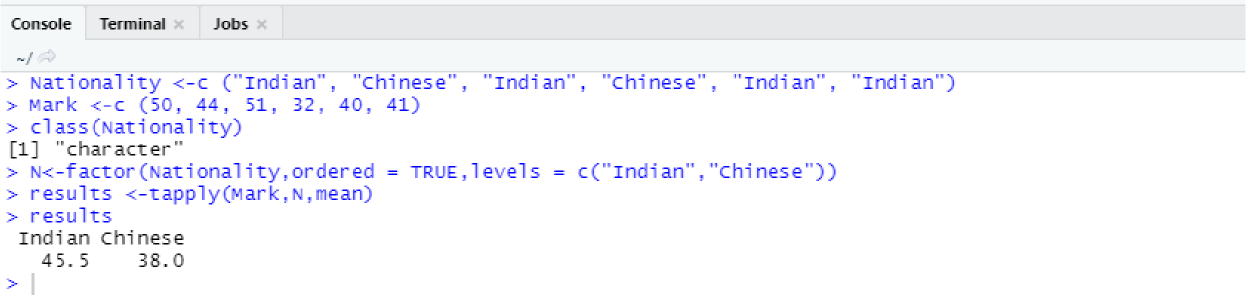
```
> Nationality
[1] "Indian"  "Chinese" "Indian"  "Chinese" "Indian"  "Indian"
> Mark
[1] 50 44 51 32 40 41
> 
```

2. Check the class of nationality vector and convert it into factor



```
> class(Nationality)
[1] "character"
> nation_f = factor(Nationality,ordered = TRUE,levels = c("Indian","Chinese"))
> nation_f
[1] Indian  Chinese Indian  Chinese Indian  Indian
Levels: Indian < Chinese
> 
```

3. Display Category wise average **Mark** using above vector data **Nationality** and **Mark** (Hint: tapply function)



```
> Nationality <-c ("Indian", "Chinese", "Indian", "Chinese", "Indian", "Indian")
> Mark <-c (50, 44, 51, 32, 40, 41)
> class(Nationality)
[1] "character"
> N<-factor(Nationality,ordered = TRUE,levels = c("Indian","Chinese"))
> results <-tapply(Mark,N,mean)
> results
 Indian Chinese 
   45.5    38.0 
> |
```