Name: Shaikh Inamul Hasan

Roll No: 100

rr1. 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.

Code & Output:

```
> head(TitanicCSV)
# A tibble: 6 × 12
  PassengerId Survived Pclass Name           Sex     Age SibSp Parch Ticket  Fare Cabin Embarked
        <dbl>    <dbl>  <dbl> <chr>          <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr> <chr>
1           1        0      3 Braund, Mr. O… male     22     1     0 A/5 2…  7.25 NA    S
2           2        1      1 Cumings, Mrs.… fema…    38     1     0 PC 17… 71.3  C85   C
3           3        1      3 Heikkinen, Mi… fema…    26     0     0 STON/…  7.92 NA    S
4           4        1      1 Futrelle, Mrs… fema…    35     1     0 113803 53.1  C123  S
5           5        0      3 Allen, Mr. Wi… male     35     0     0 373450  8.05 NA    S
6           6        0      3 Moran, Mr. Ja… male     NA     0     0 330877  8.46 NA    Q

> tail(TitanicCSV)
# A tibble: 6 × 12
  PassengerId Survived Pclass Name           Sex     Age SibSp Parch Ticket  Fare Cabin Embarked
        <dbl>    <dbl>  <dbl> <chr>          <chr> <dbl> <dbl> <dbl> <chr>  <dbl> <chr> <chr>
1         886        0      3 "Rice, Mrs. W… fema…    39     0     5 382652 29.1  NA    Q
2         887        0      2 "Montvila, Re… male     27     0     0 211536 13    NA    S
3         888        1      1 "Graham, Miss… fema…    19     0     0 112053 30    B42   S
4         889        0      3 "Johnston, Mi… fema…    NA     1     2 W./C.… 23.4  NA    S
5         890        1      1 "Behr, Mr. Ka… male     26     0     0 111369 30    C148  C
6         891        0      3 "Dooley, Mr. … male     32     0     0 370376  7.75 NA    Q
```

2. Display first 5 columns of titanic dataset.

Code & Output:

```
> head(df[, 1:5])
  PassengerId Survived Pclass                                                Name    Sex
1           1        0      3                             Braund, Mr. Owen Harris   male
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
3           3        1      3                              Heikkinen, Miss. Laina female
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female
5           5        0      3                            Allen, Mr. William Henry   male
6           6        0      3                                    Moran, Mr. James   male
```

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

Code & Output:

```
> library(dplyr)
> df <- df %>% rename(Location = Embarked)

      Age SibSp Parch           Ticket     Fare       Cabin Location
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
11   4.00     1     1          PP 9549  16.7000          G6       S
12  58.00     0     0           113783  26.5500        C103       S
13  20.00     0     0        A/5. 2151   8.0500                   S
14  39.00     1     5           347082  31.2750                   S
15  14.00     0     0           350406   7.8542                   S
```

4. Load test data without column name.

Code & Output:

```
> Test_df =read.table(file="test.csv", sep =",")
> Test_df
  V1                  V2   V3         V4
1  1 Shaikh Inamul Hasan 5000 02-02-2023
2  2                Inam 6000 04-04-2023
3  3        Inamul Hasan 7000 06-06-2023
```

5. Load test data with user defined column name.

Code & Output:

```
> Test_df=read.csv(file="test.csv", col.names=c("Sno","NAME","SAL","JoinDate"))
> Test_df
  Sno                NAME  SAL    JoinDate
1   1 Shaikh Inamul Hasan 5000 02-02-2023
2   2                Inam 6000 04-04-2023
3   3        Inamul Hasan 7000 06-06-2023
```

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

Code & Output:

```
> titanic1<- df[,1:5]
> titanic2<- df[,6:12]
> titanic3<-merge(df1,df2)
> titanic3
   PassengerId Survived Pclass                                                 Name    Sex
1            1        0      3                              Braund, Mr. Owen Harris   male
2            2        1      1    Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
3            3        1      3                               Heikkinen, Miss. Laina female
4            4        1      1         Futrelle, Mrs. Jacques Heath (Lily May Peel) female
5            5        0      3                             Allen, Mr. William Henry   male
6            6        0      3                                     Moran, Mr. James   male
7            7        0      1                              McCarthy, Mr. Timothy J   male
8            8        0      3                       Palsson, Master. Gosta Leonard   male
9            9        1      3      Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female
10          10        1      2                    Nasser, Mrs. Nicholas (Adele Achem) female
11          11        1      3                        Sandstrom, Miss. Marguerite Rut female
12          12        1      1                               Bonnell, Miss. Elizabeth female
13          13        0      3                       Saundercock, Mr. William Henry   male
14          14        0      3                          Andersson, Mr. Anders Johan   male
15          15        0      3                  Vestrom, Miss. Hulda Amanda Adolfina female
16          16        1      2                      Hewlett, Mrs. (Mary D Kingcome)  female
17          17        0      3                                  Rice, Master. Eugene   male
18          18        1      2                       Williams, Mr. Charles Eugene   male
19          19        0      3 Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) female
20          20        1      3                             Masselmani, Mrs. Fatima female
21          21        0      2                                Fynney, Mr. Joseph J   male
22          22        1      2                               Beesley, Mr. Lawrence   male
23          23        1      3                         McGowan, Miss. Anna "Annie" female
24          24        1      1                        Sloper, Mr. William Thompson   male
25          25        0      3                       Palsson, Miss. Torborg Danira female
26          26        1      3 Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson) female
27          27        0      3                              Emir, Mr. Farred Chehab   male
28          28        0      1                       Fortune, Mr. Charles Alexander   male
29          29        1      3                        O'Dwyer, Miss. Ellen "Nellie" female
30          30        0      3                                  Todoroff, Mr. Lalio   male
```

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.

Code & Output:

```
> marks <- c(90, 85, NA, 76, 92, NA, 88, 94, NA)
> NA_count <- sum(is.na(marks))
> marks <- c(90, 85, NA, 76, 92, NA, 88, 94, NA)
> NA_count <- sum(is.na(marks))
> non_NA_count <- length(marks) - NA_count
> print(NA_count)
[1] 3
> print(non_NA_count)
[1] 6
```

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

Code & Output:

```
> non_NA_marks <- marks[!is.na(marks)]
> print(non_NA_marks)
[1] 90 85 76 92 88 94
```

3. Calculate mean and median of given marks vector.

Code & Output:

```
> mean(marks,na.rm = T)
[1] 87.5
> median(marks,na.rm = T)
[1] 89
```

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

Code & Output:

```
> DF<- df[complete.cases(df), ]
> head(DF)
  PassengerId Survived Pclass                                                Name
1           1        0      3                             Braund, Mr. Owen Harris
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)
3           3        1      3                              Heikkinen, Miss. Laina
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel)
5           5        0      3                            Allen, Mr. William Henry
7           7        0      1                             McCarthy, Mr. Timothy J
     Sex Age SibSp Parch           Ticket    Fare Cabin Location
1   male  22     1     0        A/5 21171  7.2500              S
2 female  38     1     0         PC 17599 71.2833   C85        C
3 female  26     0     0 STON/O2. 3101282  7.9250              S
4 female  35     1     0           113803 53.1000  C123        S
5   male  35     0     0           373450  8.0500              S
7   male  54     0     0            17463 51.8625   E46        S
```

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

Code & Output:

```
> missing_cabin_count <- sum(is.na(df$Cabin))
> print(missing_cabin_count)
[1] 0
```

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

Code & Output:

Filter

	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

```
> mean_age <- mean(df$Age, na.rm = TRUE)
> df$Age[is.na(df$Age)] <- mean_age
> df
```

```
        Age SibSp Parch           Ticket     Fare       Cabin Location
1  22.00000     1     0        A/5 21171   7.2500                   S
2  38.00000     1     0         PC 17599  71.2833         C85        C
3  26.00000     0     0 STON/O2. 3101282   7.9250                   S
4  35.00000     1     0           113803  53.1000        C123        S
5  35.00000     0     0           373450   8.0500                   S
6  29.69912     0     0           330877   8.4583                   Q
```

```
> median_age <- median(df$Age, na.rm = TRUE)
> df$Age[is.na(df$Age)] <- median_age
> df
```

```
        Age SibSp Parch           Ticket     Fare       Cabin Location
1  22.00000     1     0        A/5 21171   7.2500                   S
2  38.00000     1     0         PC 17599  71.2833         C85        C
3  26.00000     0     0 STON/O2. 3101282   7.9250                   S
4  35.00000     1     0           113803  53.1000        C123        S
5  35.00000     0     0           373450   8.0500                   S
6  29.69912     0     0           330877   8.4583                   Q
```

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)

Code & Output:

```
> nationality <- c("Indian", "Chinese", "Indian", "Chinese", "Indian", "Indian")
> mark <- c(50, 44, 51, 32, 40, 41)
> data <- data.frame(Nationality = nationality, Mark = mark)
> data$Nationality <- as.factor(data$Nationality)
> print(data)
  Nationality Mark
1      Indian   50
2     Chinese   44
3      Indian   51
4     Chinese   32
5      Indian   40
6      Indian   41
```

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

Code & Output:

```
> class_before <- class(nationality)
> print(paste("Class before:", class_before))
[1] "Class before: character"
> nationality_factor <- as.factor(nationality)
> class_after <- class(nationality_factor)
> print(paste("Class after:", class_after))
[1] "Class after: factor"
```

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

Code & Output:

```
> data <- data.frame(Nationality = nationality, Mark = mark)
> average_marks <- tapply(data$Mark, data$Nationality, mean)
> print(average_marks)
Chinese  Indian
   38.0    45.5
```

4. Create a data frame from above vector Nationality and Mark and Create a factor corresponding to Mark with labels poor, average, good.

Code & Output:

```
> data <- data.frame(Nationality = nationality, Mark = mark)
> data$Mark_Category <- cut(data$Mark, breaks = c(-Inf, 40, 50, Inf), labels = c("poor", "average", "good"))
> print(data)
  Nationality Mark Mark_Category
1      Indian   50       average
2     Chinese   44       average
3      Indian   51          good
4     Chinese   32          poor
5      Indian   40          poor
6      Indian   41       average
```