Mid Term Project
Introduction to Data Science
Topic: Titanic

Tonoy Chandra Sarker
ID: 20-43804-2
Section: C

Dataset Description:
This dataset appears to contain information about passengers on the Titanic, a famous ship that sank on its maiden voyage in 1912. Each row in the dataset represents a passenger and includes various attributes about them. Here is a short description of the columns:

1. gender: The gender of the passenger (male or female).
2. age: The age of the passenger.
3. sibsp: The number of siblings/spouses the passenger had aboard.
4. parch: The number of parents/children the passenger had aboard.
5. fare: The fare the passenger paid for the ticket.
6. embarked: The port of embarkation (S = Southampton, C = Cherbourg, Q = Queenstown).
7. class: The class of the ticket (First, Second, Third).
8. who: Represents the category of the passenger (man, woman, child).
9. alone: Indicates whether the passenger was traveling alone (TRUE or FALSE).
10. survived: Indicates whether the passenger survived the Titanic disaster (0 = No, 1 = Yes).

It seems the dataset has some missing values as indicated by blanks in certain cells. This data might be used to explore patterns related to passenger demographics and their survival outcomes during the Titanic's tragic event.

Import the data set as csv and print the data set:

```
mydata <- read.csv("C:/Titanic.csv", header = TRUE, sep = ",")
mydata
```

Output:

	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

Description :

Here is the code of import the dataset as csv file. It is the output of the dataset which is imported in RStudio.

To see the column name of the data set:

Code :

```
6  names(mydata)
```

Output:

```
> names(mydata)
 [1] "gender"   "age"      "sibsp"    "parch"    "fare"     "embarked" "class"    "who"      "alone"    "survived"
```

Description : In this code ,we can see the column name of the dataset. Here with this code can see the attributes names. The output of the name() function where we can see the attributes of the dataset.

Annotating datasets:

Code:

```
mydata$gender <- factor(mydata$gender,
                        levels = c(0,1),
                        labels = c("male", "female"))
mydata
```

.

Output:

```
    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
```

Description:

The gender column is converted from numeric (0 and 1) to a factor with labels "male" and "female".

Summary of the structure of data set:

Code:

```
8  str(mydata)
```

Output:

```
> str(mydata)
'data.frame':   250 obs. of  10 variables:
 $ gender  : int  0 1 1 1 0 0 0 0 1 1 ...
 $ 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 ...
> |
```

Description :

The structure of the dataset is displayed using str().

Descriptive Statistics Using summary() Function:

Code :

```
summary(mydata)
```

Output:

```
> summary(mydata)
     gender            age              sibsp            parch            fare           embarked            class          
 Min.   :0.0000   Min.   :  0.83   Min.   :0.000   Min.   :0.000   Min.   :  0.000   Length:250         Length:250        
 1st Qu.:0.0000   1st Qu.: 19.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:  8.034   Class :character   Class :character  
 Median :0.0000   Median : 27.00   Median :0.000   Median :0.000   Median : 13.977   Mode  :character   Mode  :character  
 Mean   :0.3629   Mean   : 33.33   Mean   :0.656   Mean   :0.392   Mean   : 26.588                                        
 3rd Qu.:1.0000   3rd Qu.: 37.00   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.: 29.094                                        
 Max.   :1.0000   Max.   :455.00   Max.   :8.000   Max.   :5.000   Max.   :263.000                                        
 NA's   :13       NA's   :48                                                                                              
     who               alone              survived      
 Length:250         Length:250         Min.   :0.000  
 Class :character   Class :character   1st Qu.:0.000  
 Mode  :character   Mode  :character   Median :0.000  
                                       Mean   :0.344  
                                       3rd Qu.:1.000  
                                       Max.   :1.000  
```

Description: Here is the code to see the descriptive Statistics. To see descriptive statistic, we use the summary() function. In the output here min, max, median, and mean are shown.

Summary in standard deviation:

Code:

```
library(dplyr)
mydata %>% summarise_if(is.numeric, sd)
```

.

Output:

```
> mydata %>% summarise_if(is.numeric, sd)
  age    sibsp     parch     fare survived
1  NA 1.305558 0.8252637 34.82165 0.475994
> |
```

Description :

The standard deviation of numeric columns in the dataset is calculated using the dplyr package.

Standard deviation of the values stored in a CSV file:

Code :

```
s<-mydata$fare
sd(s)
```

Output:

```
> sd(s)
[1] 34.82165
> |
```

Description:

The standard deviation of the "fare" column is directly calculated.

Finding Missing(null) values:

Code :

```
colSums(is.na(mydata))
```

Output:

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

Dealing with Missing Value:

Code :

```
mydata$age <- ifelse(is.na(mydata$age),mean(mydata$age, na.rm = TRUE),mydata$age)
mydata$parch <- ifelse(is.na(mydata$parch),mean(mydata$parch, na.rm = TRUE),mydata$parch)
mydata$fare <- ifelse(is.na(mydata$fare),mean(mydata$fare, na.rm = TRUE),mydata$fare)
mydata$gender <- ifelse(is.na(mydata$ gender  ),mean(mydata$ gender  , na.rm = TRUE),mydata$ gender  )
```

Output:

```
17  1.000000  2.00000     4     1  29.1250        Q  Third child FALSE        0
18  1.000000 33.32837     0     0  13.0000        S Second   man  TRUE        1
19  2.000000 31.00000     1     0  18.0000        S  Third woman FALSE        0
20  2.000000 33.32837     0     0   7.2250        C  Third woman  TRUE        1
21  1.000000 35.00000     0     0  26.0000        S Second   man  TRUE        0
22  1.000000 34.00000     0     0  13.0000        S Second   man  TRUE        1
23  2.000000 15.00000     0     0   8.0292        Q  Third child  TRUE        1
24  1.000000 28.00000     0     0  35.5000        S  First   man  TRUE        1
25  2.000000  8.00000     3     1  21.0750        S        child FALSE        0
26  2.000000 38.00000     1     5  31.3875        S  Third woman FALSE        1
27  1.000000 33.32837     0     0   7.2250        C  Third   man  TRUE        0
28  1.000000 19.00000     3     2 263.0000        S  First   man FALSE        0
29  2.000000 33.32837     0     0   7.8792        Q  Third woman  TRUE        1
30  1.000000 33.32837     0     0   7.8958        S  Third   man  TRUE        0
31  1.000000 40.00000     0     0  27.7208        C  First   man  TRUE        0
32  2.000000 33.32837     1     0 146.5208        C  First woman FALSE        1
33  2.000000 33.32837     0     0   7.7500        Q  Third woman  TRUE        1
34  1.362869 66.00000     0     0  10.5000        S Second   man  TRUE        0
35  1.000000 28.00000     1     0  82.1708        C  First   man FALSE        0
36  1.000000 42.00000     1     0  52.0000        S  First   man FALSE        0
37  1.000000 33.32837     0     0   7.2292        C  Third   man  TRUE        1
```

Description:

The code checks for missing values in each column using colSums(is.na(mydata)). Then, missing values in numeric columns ("age", "parch", "fare", and "gender") are replaced with their respective column means.

Find the specific row number of Missing Value:

Code :

```
which(is.na(mydata$gender))
```

Output:

```
> which(is.na(mydata$gender))
integer(0)
```

Description:

It finds the row numbers where "gender" has missing values (NAs).

Data Completeness:

Code :

```
completeness <- sapply(mydata, function(x) sum(!is.na(x)) / length(x))

print(completeness)
```

Output:

```
> completeness <- sapply(mydata, function(x) sum(!is.na(x)) / length(x))
>
> print(completeness)
  gender      age    sibsp    parch     fare embarked    class      who    alone survived
       1        1        1        1        1        1        1        1        1        1
> |
```

Description:

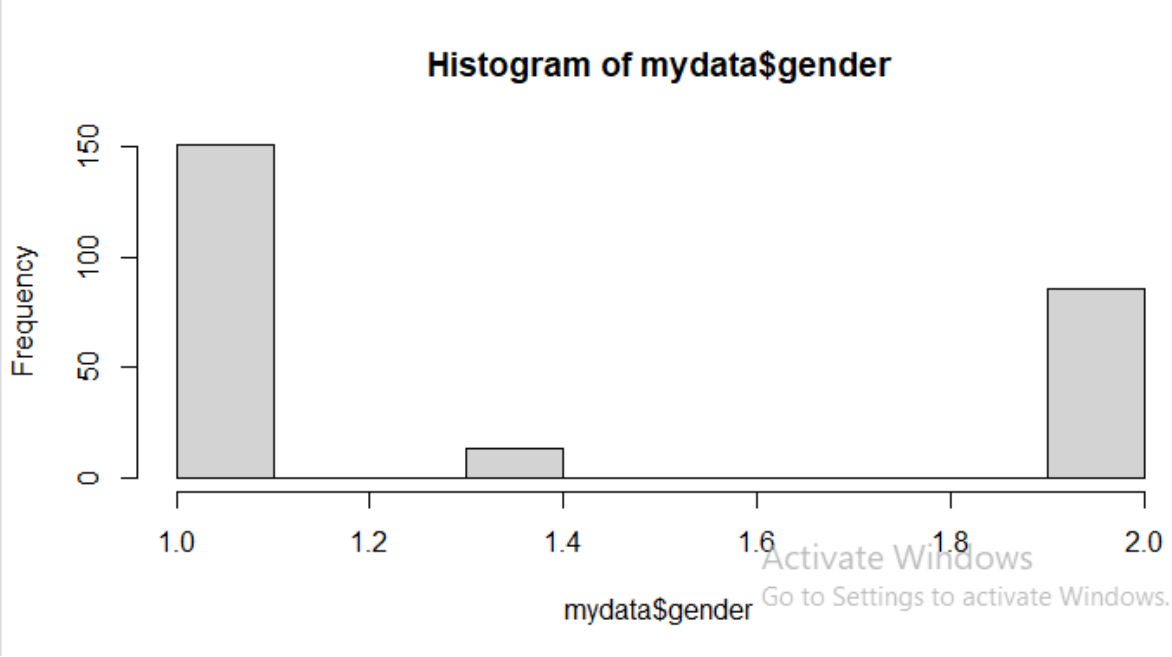
The completeness of each column is calculated, i.e., the proportion of non-missing values in each column.

Histogram:

Code :

```
hist(mydata$fare)
hist(mydata$gender)
```

Output:



Description:

Histograms for the "fare" and "gender" columns are plotted.

Univariate Exploration

For age attribute:

Code :

```
mean(mydata$age)
median(mydata$age)
var(mydata$age)
sd(mydata$age)
```

Output:

```
> mean(mydata$age)
[1] 33.32837
> median(mydata$age)
[1] 30
> var(mydata$age)
[1] 1691.317
> sd(mydata$age)
[1] 41.12562
```

For parch attribute:

Code :

```
mean(mydata$parch)
median(mydata$parch)
var(mydata$parch)
sd(mydata$parch)
```

Output:

```
> mean(mydata$parch)
[1] 0.392
> median(mydata$parch)
[1] 0
> var(mydata$parch)
[1] 0.6810602
> sd(mydata$parch)
[1] 0.8252637
```

Description:

Mean, median, variance, and standard deviation are calculated for "age" and "parch" attributes.