Big Data Tools for Managers

Dept. of MBA, Siddaganga Institute of Technology-Tumkur

Working with Titanic Dataset

The files we just opened are available on the data page for the Titanic competition on Kaggle. That page also has a data dictionary, which explains the various columns that make up the data set. Below are the descriptions contained in that data dictionary:

Download Titanic data

Note: This link take you in another page and then click on File menu -> Save Page As -> Open file window and click on Save

- **PassengerID**— A column added by Kaggle to identify each row and make sublessons easier
- **Survived**— Whether the passenger survived or not and the value we are predicting (0=No, 1=Yes)
- **Pclass**— The class of the ticket the passenger purchased (1=1st, 2=2nd, 3=3rd)
- **Name**- The name of passenger's
- **Sex**— The passenger's sex
- **Age**— The passenger's age in years
- **SibSp**— The number of siblings or spouses the passenger had aboard the Titanic
- **Parch**— The number of parents or children the passenger had aboard the Titanic
- **Ticket**— The passenger's ticket number
- **Fare**— The fare the passenger paid
- **Cabin**— The passenger's cabin number
- **Embarked**— The port where the passenger embarked (C=Cherbourg, Q=Queenstown, S=Southampton)

Write R Code for below questions.

1. Read titanic dataset in R
2. Get the dimension of Titanic dataset
3. Display column names of dataset
4. View data in Excel like screen
5. Get Quick summary for all the columns
6. Indentify the Null(Missing) values for dataset
7. Get the passanger details which is has age 0.42year?
8. How many male and female are on Titanic?
9. What percentage of male and female are on Titanic?
10. Display the Female rows from the Dataset
11. Find out oldest Female in Passengers
12. Find out maximum Ticket Fare
13. Display the distribution of Fare variable in Histogram
14. Display the distribution of Age variable in Histogram
15. How many passengers got survived according to dataset
16. Display in Pie chart, How many passengers got survived according to dataset
17. Count the number of Passengers based on Pclass(1=1st, 2=2nd, 3=3rd)
18. Display in barplot the number of Passengers based on Pclass(1=1st, 2=2nd, 3=3rd)
19. Fill the color in Age distribution in Histogram
20. Sort the titanic dataset based on Age of passenger elder to younger

Note : This analysis is based on very few variables, you may have to practice yourself with remaing variables to get more hands-on with R & Data

In [ ]:
```r
#1. Read titanic data in R
data <- read.csv("titanic.csv")
```

In [ ]:
```r
#2. Get the dimension of Titanic dataset
dim(data)
```

In [ ]:
```r
#3. Display column names of dataset
colnames(data)
```

In [ ]:
```r
#4. View data in Excel like screen
View(data)
```

In [ ]:
```r
#5. Get Quick summary for all the columns
summary(data)
```

In [ ]:
```r
#6. Indentify the Null(Missing) values for dataset
colSums(is.na(data))
# Age columns has missing values ~177 rows
```

In [ ]:
```r
#7. Describe the Age column
summary(data$Age)
# Passangers are from 0.42 to 80years old
```

In [ ]:
```r
#8. Get the passanger details which is has age 0.42year?
subset(data, Age==0.42)
```

In [ ]:
```r
#9. How many male and female are on Titanic?
table(data$Sex)
# There are 314 Female and 577 Male
```

In [ ]:
```r
#10. What percentage of male and female are on Titanic?
gender_freq <- table(data$Sex)
prop.table(gender_freq) * 100

# There are 35% of Female and 65% are Male onboaded to Titanic
```

In [ ]:
```r
#11. Display the Female rows from the Dataset
subset(data, Sex=="female")
```

In [ ]:
```r
#12 Find out oldest Female in Passengers
temp <- subset(data, Sex=="female")
max_age <- max(temp$Age, na.rm=TRUE) # Maximum Age for Female dataset, na.rm=TRUE ignore the null values from data

subset(temp, Age==max_age)
# There are two oldest passesange in Female and age is 63
```

In [ ]:
```r
#12 Find out maximum Ticket Fare
max(data$Fare)
# The Maximum ticket fare is $512.32
```

In [ ]:
```r
#13. Display the distribution of Fare variable in Histogram
hist(data$Fare, main="Distribution of Ticket Fare")
```

In [ ]:
```r
#13. Display the distribution of Age variable in Histogram
hist(data$Age, main="Distribution of Age")
```

In [ ]:
```r
#14. How many passengers got survived according to dataset
table(data$Survived)
# There are 342 Passengers got survived
```

In [ ]:
```r
#15. Display in Pie chart, How many passengers got survived according to dataset
res= table(data$Survived)
pie(res,
    main="Survived/Not-Survived Passenger")
```

In [ ]:
```r
#16. Based on gender who survived more
table(data$Sex, data$Survived)
# Female are survived more ~233
```

In [ ]:
```r
#17. Count the number of Passengers based on Pclass(1=1st, 2=2nd, 3=3rd)
table(data$Pclass)
```

In [ ]:
```r
#18. Display in barplot the number of Passengers based on Pclass(1=1st, 2=2nd, 3=3rd)
res = table(data$Pclass)
barplot(res,
        main="Number of Passengers based on Ticket Class",
        col=c("green","blue","red")
        )

# we can give colours using col= parameters in all(pie, barplot, histogram) the graphs.
```

In [ ]:
```r
# 19. Fill the color in Age distribution in Histogram
hist(data$Age,
    main="Distribution of Age with colour",
    col="yellow")
```

In [ ]:
```r
#20. Sort the titanic dataset based on Age of passenger elder to younger
data[order(data$Age, decreasing=TRUE),]
```