Week 1 Assignment

Q1 a. What are the types of variable (quantitative / qualitative) and levels of measurement (nominal / ordinal /interval / ratio) for PassengerId, and Age?

PassengerID : quantitative and nominal

Age: quantitative and ratio

Q1b. Which variable has the most missing observations?

Cabin is the most missing observation with a total of 687.

Q2. Impute missing observations for Age, SibSp, and Parch with the column median (ordinal, interval, or ratio), or the column mode. To do so, use something like this: *mydata$Age[is.na(mydata$Age)]=median(mydata$Age, na.rm=TRUE)* .

NA in Age were placed with 28 years old.

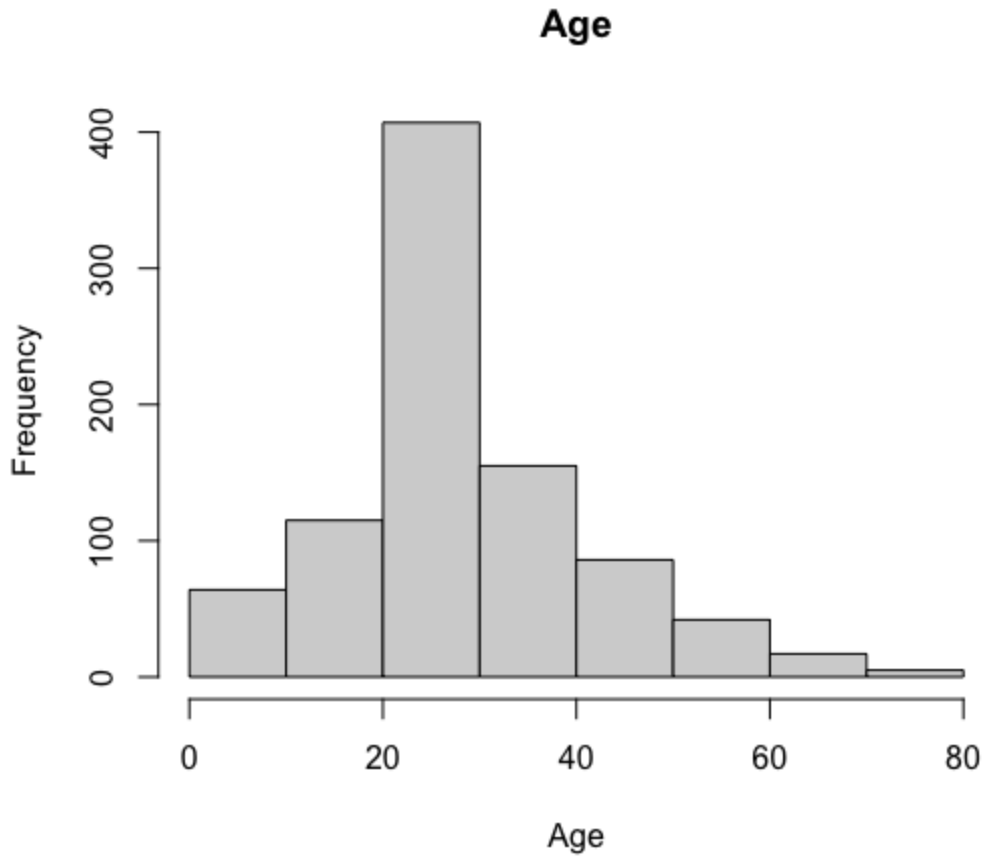
Q3. Install the *psych* package in R: *install.packages('pscyh')*. Invoke the package using *library(psych)*. Then provide descriptive statistics for Age, SibSp, and Parch (e.g., *describe(mydata$Age*).

> describe(titanic$Age)

vars n mean sd median trimmed mad min max range skew kurtosis se

X1 1 891 29.36 13.02 28 28.83 8.9 0.42 80 79.58 0.51 0.97 0.44

> hist(titanic$Age, xlab = "Age", main = "Age")

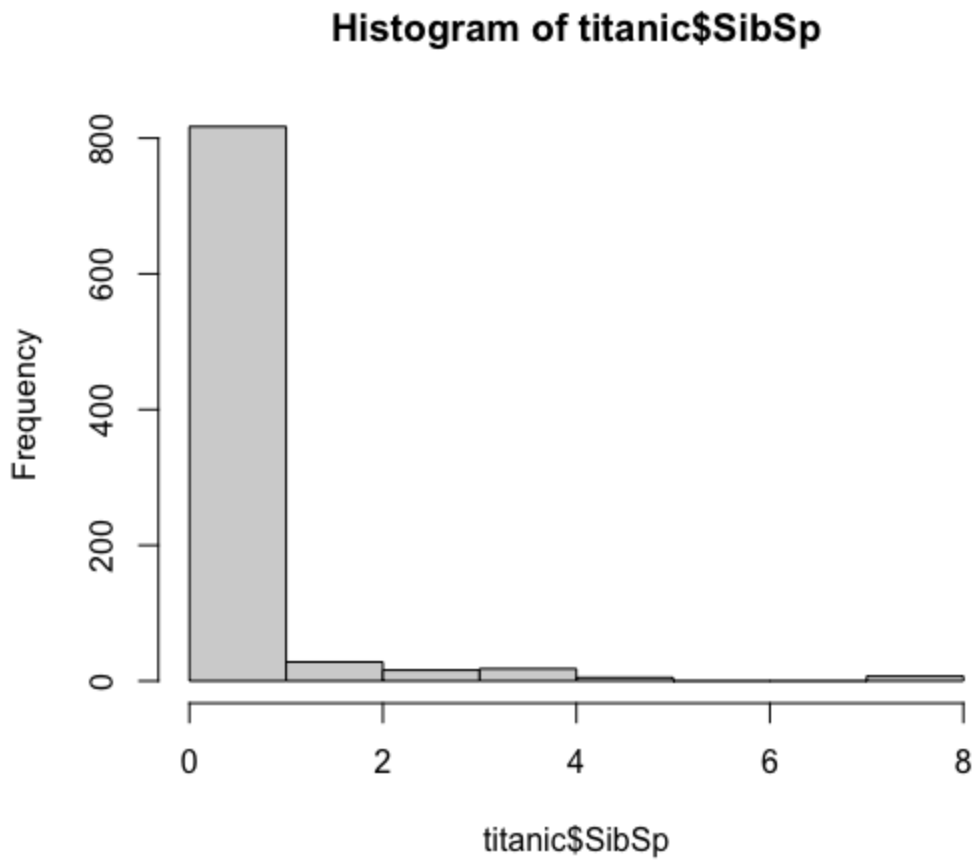


> describe(titanic$SibSp)

vars n mean sd median trimmed mad min max range skew kurtosis se

X1 1 891 0.52 1.1 0 0.27 0 0 8 8 3.68 17.73 0.04

> hist(titanic$SibSp)

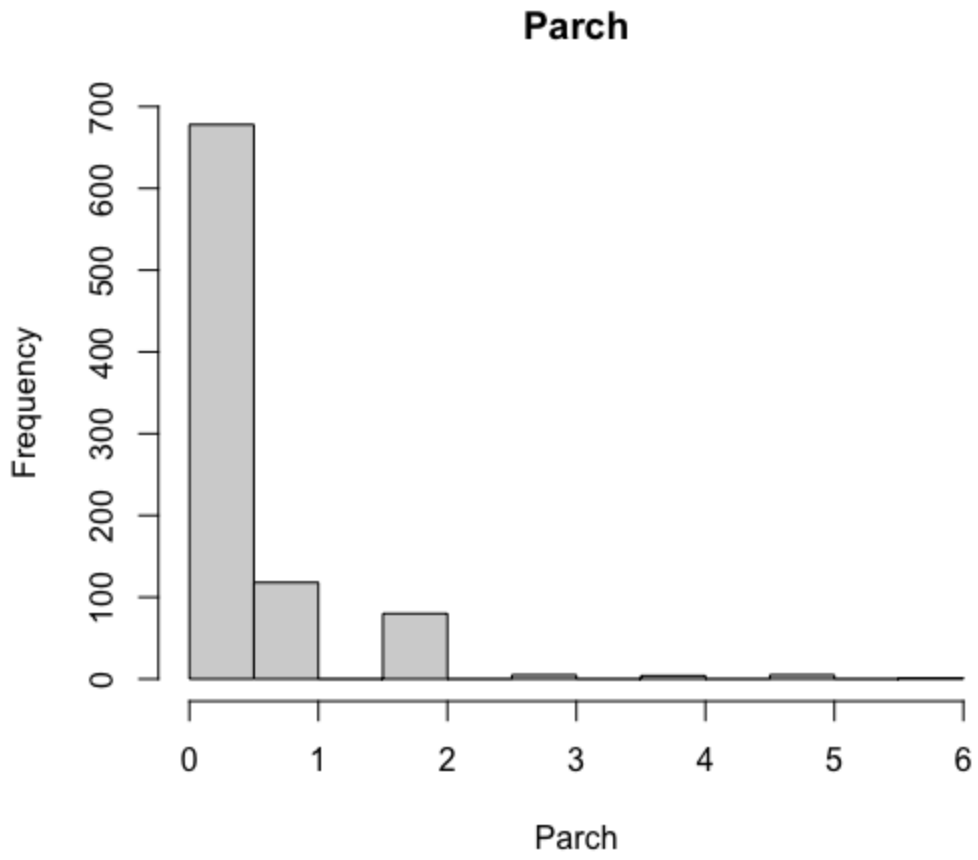


> describe(titanic$Parch)

vars n mean sd median trimmed mad min max range skew kurtosis se

X1 1 891 0.38 0.81 0 0.18 0 0 6 6 2.74 9.69 0.03

> hist(titanic$Parch, xlab = "Parch", main = "Parch")

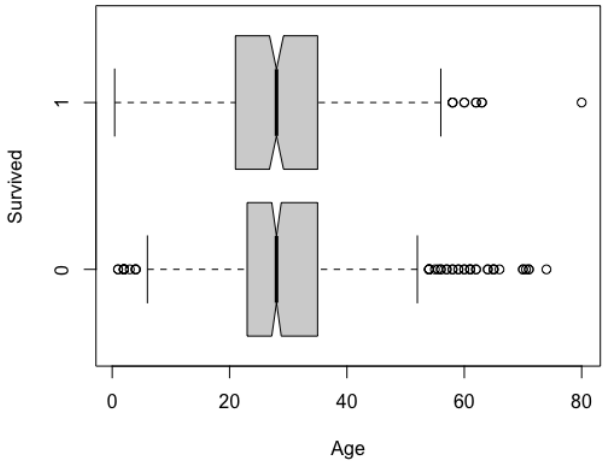


Q4. Provide a cross-tabulation of Survived and Sex (e.g., table(mydata$Survived, mydata$Sex). What do you notice?

	female	male
0	81	468
1	233	109

I noticed that the number of females who survived the incident is more than males. It could be that females' safety was prioritized before males during the incident. With this information, we are not able to determine the number of children and elderly persons who survived the incident.

Q5. Provide notched boxplots for Survived and Age (e.g., boxplot(mydata$Age~mydata$Survived, notch=TRUE, horizontal=T). What do you notice?



With this graph, I noticed that there is an overlap around the median age of 28. The reason could be due to the missing observation of age in the dataset. A lot of the elderly did not manage to survive the incident and there's four children did not as well.