002-Working-on-real-data

March 19, 2024

1 TP 02 - About Data Frames - 2/2

- Written by Alexandre Gazagnes
- Last update: 2024-02-01
- Based on https://www.w3schools.com/r/default.asp

1.1 Load the dataset

[ ]:
```
# Url to the csv file
url = "https://raw.githubusercontent.com/AlexandreGazagnes/
 ↪Unilassalle-Public-Ressources/main/2a-statistical-tests/02-session/database.
 ↪csv"
```

[ ]:
```
# Local file
fn = "/home/alex/Downloads/my_awsome_db.csv"
```

[ ]:
```
# Load the csv file
df = read.csv(url, header=TRUE, sep=",", dec=".")
```

1.2 Data Discovery

[ ]:
```
# head of the csv file
head(df, 5)  # first 2 rows
```

[ ]:
```
# str of the csv file
str(df)  # compact display of the structure of an R object
```

[ ]:
```
# is.numeric
is.numeric(df$Taille) # TRUE
is.numeric(df$Nom) # TRUE
```

[ ]:
```
# min and max
min(df$Poids) # 1.5
max(df$Poids) # 2.5
min(df$Taille) # 1.5
max(df$Taille) # 2.5
max(df$Taille, na.rm = TRUE) # 2.5
```

[ ]:
```r
# Mr Blois
df[df$Nom == "Blois",]
```

[ ]:
```r
# Try this
"Coucou" %in% df$Nom
```

[ ]:
```r
# Advanced selection
df[df$Sexe == "M" & ( (df$Quart == "Q4") |(df$Quart == "Q1")), ]
```

[ ]:
```r
# count the number of elements in each category
table(df$Sexe)
```

[ ]:
```r
# ask for the shape of the dataframe
dim(df)  # 4 rows and 2 columns
```

[ ]:
```r
# ask for the summary of the dataframe
summary(df)
```

1.3 Answers

[ ]:
```r
# Calculate the means of the estimates of the teacher's weight (poids) in group␣
 ↪D1 (then D2).
mean(df[df$Groupe == "D1",]$Poids, na.rm = TRUE)
```

[ ]:
```r
# Calculate the average of the first random numbers (Nombre1) of the students␣
 ↪in group D2?
mean(df[df$Groupe == "D2",]$Nombre1, na.rm = TRUE)
```

[ ]:
```r
# Estimate the variance of the size (Taille) of the teacher, we will take all␣
 ↪the students.
var(df$Taille, na.rm = TRUE)
```

[ ]:
```r
# Calculate the median of the sample consisting of the first random number␣
 ↪(Nombre1)  between 0 and 99 given by the students (Men=M).
median(df[df$Sexe == "M",]$Nombre1, na.rm = TRUE)
```

[ ]:
```r
# Calculate the deciles of the sample consisting of the first random number␣
 ↪(Nombre1)  between 0 and 99 given by the students (Men).
quantile(df[df$Sexe == "M",]$Nombre1, na.rm = TRUE)
```

[ ]:
```r
# Calculate the coefficients of variation, skewness and kurtosis of the␣
 ↪teacher's heights (taille) of all students.
cv(df$Taille, na.rm = TRUE)
skewness(df$Taille, na.rm = TRUE)
kurtosis(df$Taille, na.rm = TRUE)
```

[ ]:
```
# Summarize the series by these main characteristics (function summary).
summary(df)
```

[ ]:
```
# 9) Display the division of the series made up of the second random number␣
↪(Nombre2)  between 0 and 99 given by the students (Men =M and Women=M)␣
↪according to the following breakdown:
# ]-1 ; 9], ]9 ; 19], ]19 ; 29], ]29 ; 39], ]39 ; 49], ]49 ; 59], ]59 ; 69],␣
↪]69 ; 79], ]79 ; 89], ]89 ; 99].
```

[ ]:
```
# Create the breaks
breaks = seq(-1, 99, 10)
```

[ ]:
```
# Cut the data
df$cut = cut(df$Nombre2, breaks = breaks)
```

[ ]:
```
# Count the number of elements in each category
table(df$cut)
```

[ ]:
```
# Display the distribution of the series made up of the second random number␣
↪(Nombre2)  between 0 and 99 given by the
# students
hist(df$Nombre2, breaks = breaks, main = "Histogramme de la variable Nombre2",␣
↪xlab = "Nombre2", ylab = "Fréquence")
```

[ ]:
```
# Give the contingency table for the factors Group (Demi) and Sex.
table(df$Demi, df$Sexe)
```

[ ]:
```
# Calculate the covariance between the first random number (Nombre1) and the␣
↪second (Nombre2).
cov(df$Nombre1, df$Nombre2, use = "complete.obs")
```

[ ]:
```
# Calculate the correlation coefficient between the first random number␣
↪(Nombre1)  and the second (Nombre2).
cor(df$Nombre1, df$Nombre2, use = "complete.obs")
```

[ ]:
```
# Calculate the correlation coefficient between the first random number␣
↪(Nombre1)  and the second (Nombre2).
cor(df$Poids, df$Taille, use = "complete.obs")
```

[ ]:
```
# Sort the dataframe by the column "Poids"
df[order(df$Poids),]
```

[ ]:
```
# Sort the dataframe by the column "Poids" in decreasing order
df[order(df$Poids, decreasing = TRUE),]
```

[ ]:
```
# Sort the dataframe by the column "Poids" and "Taille"
df[order(df$Poids, df$Taille),]
```

[ ]:
```r
# Sort the dataframe by the column "Poids" and "Taille" in decreasing order
df[order(df$Poids, df$Taille, decreasing = TRUE),]
```