Dataframe\_questions

2024-04-17

# 1.Create a data frame named df\_students with the following columns: ID, Name, Age, and Grade. Populate it with data for 5 students. Perform the following operations: Add a new column Passed that indicates with TRUE or FALSE whether each student passed (assume a passing grade is at least 60). Select and print only the rows of students who are 18 years or older.  
df\_students <- data.frame(  
 ID = 1:5,  
 Name = c("Alice", "Bob", "Charlie", "David", "Emma"),  
 Age = c(20, 18, 19, 17, 21),  
 Grade = c(75, 80, 55, 90, 40)  
)  
  
# Add a new column Passed  
df\_students$Passed <- df\_students$Grade >= 60  
  
# Select and print only the rows of students who are 18 years or older  
older\_students <- df\_students[df\_students$Age >= 18, ]  
print(older\_students)

## ID Name Age Grade Passed  
## 1 1 Alice 20 75 TRUE  
## 2 2 Bob 18 80 TRUE  
## 3 3 Charlie 19 55 FALSE  
## 5 5 Emma 21 40 FALSE

# 2.Suppose you have two data frames, df\_A and df\_B... Merge these data frames  
df\_A <- data.frame(  
 ID = 1:3,  
 Name = c("Alice", "Bob", "Charlie"),  
 Age = c(20, 18, 19)  
   
)  
df\_B <- data.frame(  
 ID = 4:6,  
 Grade = c(75, 80, 55)  
)  
#merged\_df <- merge(df\_A, df\_B, by = "ID")  
merged\_df <- merge(df\_A, df\_B,by = "ID", all = TRUE)  
merged\_df

## ID Name Age Grade  
## 1 1 Alice 20 NA  
## 2 2 Bob 18 NA  
## 3 3 Charlie 19 NA  
## 4 4 <NA> NA 75  
## 5 5 <NA> NA 80  
## 6 6 <NA> NA 55

#3. Given a data frame df\_sales with columns Date, ProductID, Quantity, and Price...Calculate the total sales (Quantity \* Price) for each product  
# Create df\_sales data frame  
df\_sales <- data.frame(  
 Date = c("2024-04-01", "2024-04-01", "2024-04-02", "2024-04-02"),  
 ProductID = c(1, 2, 1, 2),  
 Quantity = c(10, 5, 8, 6),  
 Price = c(2.5, 3.0, 4.0, 3.5)  
)  
df\_sales$total\_sales <- df\_sales$Quantity \* df\_sales$Price  
df\_sales

## Date ProductID Quantity Price total\_sales  
## 1 2024-04-01 1 10 2.5 25  
## 2 2024-04-01 2 5 3.0 15  
## 3 2024-04-02 1 8 4.0 32  
## 4 2024-04-02 2 6 3.5 21

#4. For iris data set...Calculate the average Sepal Length for each Species  
data(iris)  
head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

# avg sepal length::  
avg\_sepal\_length <- aggregate(Sepal.Length ~ Species, data = iris, FUN = mean)  
avg\_sepal\_length

## Species Sepal.Length  
## 1 setosa 5.006  
## 2 versicolor 5.936  
## 3 virginica 6.588

# new column Sepal.Area:::  
iris$Sepal.Area <- iris$Sepal.Length \* iris$Sepal.Width  
iris$Sepal.Area

## [1] 17.85 14.70 15.04 14.26 18.00 21.06 15.64 17.00 12.76 15.19 19.98 16.32  
## [13] 14.40 12.90 23.20 25.08 21.06 17.85 21.66 19.38 18.36 18.87 16.56 16.83  
## [25] 16.32 15.00 17.00 18.20 17.68 15.04 14.88 18.36 21.32 23.10 15.19 16.00  
## [37] 19.25 17.64 13.20 17.34 17.50 10.35 14.08 17.50 19.38 14.40 19.38 14.72  
## [49] 19.61 16.50 22.40 20.48 21.39 12.65 18.20 15.96 20.79 11.76 19.14 14.04  
## [61] 10.00 17.70 13.20 17.69 16.24 20.77 16.80 15.66 13.64 14.00 18.88 17.08  
## [73] 15.75 17.08 18.56 19.80 19.04 20.10 17.40 14.82 13.20 13.20 15.66 16.20  
## [85] 16.20 20.40 20.77 14.49 16.80 13.75 14.30 18.30 15.08 11.50 15.12 17.10  
## [97] 16.53 17.98 12.75 15.96 20.79 15.66 21.30 18.27 19.50 22.80 12.25 21.17  
## [109] 16.75 25.92 20.80 17.28 20.40 14.25 16.24 20.48 19.50 29.26 20.02 13.20  
## [121] 22.08 15.68 21.56 17.01 22.11 23.04 17.36 18.30 17.92 21.60 20.72 30.02  
## [133] 17.92 17.64 15.86 23.10 21.42 19.84 18.00 21.39 20.77 21.39 15.66 21.76  
## [145] 22.11 20.10 15.75 19.50 21.08 17.70

head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species Sepal.Area  
## 1 5.1 3.5 1.4 0.2 setosa 17.85  
## 2 4.9 3.0 1.4 0.2 setosa 14.70  
## 3 4.7 3.2 1.3 0.2 setosa 15.04  
## 4 4.6 3.1 1.5 0.2 setosa 14.26  
## 5 5.0 3.6 1.4 0.2 setosa 18.00  
## 6 5.4 3.9 1.7 0.4 setosa 21.06

# For airquality data set...Calculate the average Ozone level for each month, excluding missing values..Determine the month with the highest average Ozone level  
data(airquality)  
head(airquality)

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6

avg\_ozone\_month <- aggregate(Ozone ~ Month, data = airquality, FUN = function(x) mean(x, na.rm = TRUE))  
avg\_ozone\_month

## Month Ozone  
## 1 5 23.61538  
## 2 6 29.44444  
## 3 7 59.11538  
## 4 8 59.96154  
## 5 9 31.44828

max\_avg\_ozone\_month <- avg\_ozone\_month[which.max(avg\_ozone\_month$Ozone), "Month"]  
max\_avg\_ozone\_month

## [1] 8