A screenshot of a computer

Description automatically generated

Creates a data frame df 1 with columns Name,State, Sales and 6 rows.

A screenshot of a computer

Description automatically generated

Computes the summary of sales by states

A screenshot of a computer

Description automatically generated

Library(dplyr) loads library dplyr which provides a set of verbs for data maniipulation

A screenshot of a computer

Description automatically generated

Line df1 %>% group\_by(State) %>% summarise(sum\_sales = sum(Sales))

Uses library dplyr to group summary of sales by states.

#(a) Find the size of the dataf rame.How many rows,how many columns?

print(paste("The number of rows is ",

nrow(worldCupMatchesDF),

", Number of columns is ",

ncol(worldCupMatchesDF)

))

A screenshot of a computer

Description automatically generated

#(b) Uses summary function to report the statistical summary of your data.

print(summary(worldCupMatchesDF))

A screenshot of a computer

Description automatically generated

#(c) Find how many unique locations olympics were held at.

uniqueloc = unique(na.omit(worldCupMatchesDF$City))

print(paste("The count of unique locations is: ",

length(uniqueloc)

))

A screenshot of a computer

Description automatically generated

#(d) Find the average attendance.

print(paste("The average attendence for matches is: ",mean(worldCupMatchesDF$Attendance,na.rm=TRUE)))

A screenshot of a computer

Description automatically generated

#(e) For each Home Team,what is the total number of goals scored?(Hint:Please refer to question1)

worldCupMatchesDF %>% group\_by(Home.Team.Name) %>% summarise(Home.Team.Goals = sum(Home.Team.Goals)) %>% show()

A screenshot of a computer

Description automatically generated

#(f) What is the average number of attendees for each year? Is there a trend or pattern in the data in that sense?

worldCupMatchesDF %>% group\_by(Year) %>% summarise(Avg.Attendance = mean(Attendance, na.rm =TRUE)) %>% show()

print(median(worldCupMatchesDF$Attendance, na.rm = TRUE))

The average attendees seem to have rose from 1934 to 1966 then held close to the median of 41579 and in 2006,2010,2014 has been above the median.

A screenshot of a computer

Description automatically generated

#(a) Find how many Alzheimers patients there are in the dataset.(Hint:Pleaserefertoquestion1)

summary(metabolitesDF)

alzheimersCount = nrow(metabolitesDF[metabolitesDF$Label == "Alzheimer", ])

print(paste("Count of patients with Alzheimer is: ", alzheimersCount))

A screenshot of a computer

Description automatically generated

#(b) Determine the number of missing values for each column.(Hint:is.na())

print(colSums(is.na(metabolitesDF)))

A screenshot of a computer

Description automatically generated

#(c) Remove the rows which has missing value for the Dopamine column and assign the result to a new data frame. (Hint: is.na( ) )

metabolitesMissingDopamineDF =metabolitesDF[!is.na(metabolitesDF$Dopamine), ]

print(metabolitesMissingDopamineDF)

A screenshot of a computer

Description automatically generated

#(d) In the newdataframe,replace the missing values in the c4-OH-Pro column with the median value of the same column. (Hint: there is median( ) function.)

metabolitesMissingDopamineDF %>% mutate(c4.OH.Pro = ifelse(is.na(c4.OH.Pro), median(c4.OH.Pro, na.rm = T),c4.OH.Pro)) %>% show()

A screenshot of a computer

Description automatically generated

#(e) (Optional)Drop columns which have more than 25% missing values.

#(Hint:when you slice your data frame,you can use -c(.., ..., ...) where ... represent one column name)

metabolitesDF = metabolitesDF[, which((!colMeans(is.na(metabolitesDF))) \* 100 >= 25)]

print(metabolitesDF)

A screenshot of a computer

Description automatically generated