DATA WAREHOUSING & DATA MINING

10.1.23
—

ABHIJIT MISHRA

2020BCS0094

CSE - 321

Goals

1. Implement the given questions

Note

I have used built in software to upload this code on RPubs, code with output is printed in a neat and clean way.

http://rpubs.com/Panda_250/990930

Code

#2020BCS0094

question 1

z <- c(0.1, 0.6, 33.8, 1.9, 9.6, 4.3, 33.7, 0.3, 0.0, 0.1) # Inputting values into a vector

z # Printing the vector

mean(z) # Mean of the vector z

sum(z) # sum of all elements of z

max(z) # maximum element in vector z

s <- z[z>20] # forming a subset of z such that z>20

s # printing vector

mean(z[z>4]) # finding mean with the given condition

k <- z[z==0 | z==6] # forming a subset with 'OR' of 2 condition

k # printing the vector k

question 2

```
# reading the data into the vectors
length = c(2.1, 3.4, 2.5, 2.7, 2.9)
radius = c(0.3, 0.5, 0.6, 0.9, 1.1)

# correlation between lengths and diameters
cor(length,radius,method = "pearson")

# volume of each cylinder
volume = length * radius * radius
volume

# mean, standard deviation, and coefficient of variation
mean(volume)
sd(volume)
cof_of_var <- sd(volume)/mean(volume) *100
cof_of_var



# Assume your measurements are in centimeters. Recalculate the
# volumes so that their units are in cubic millimeters.
volume_changed_units = volume * 1000
mean(volume_changed_units)
sd(volume_changed_units)
new_cof_of_var = sd(volume_changed_units)/mean(volume_changed_units) * 100
new_cof_of_var

# question 3

# reading input
```

```r
x = c(1,2,5,9,11)
y = c(2,5,1,0,23)

# intersection, set difference, union of 2 vectors
print(intersect(x,y))
print(setdiff(x,y))
print(setdiff(y,x))

# Difference between union and c(x,y)
print(union(x,y))
print(c(x,y))
# there may be repeated values in c(x,y) but union only counts every unique value only
once.

# question 4

# Construct a matrix with 10 columns and 10 rows, all filled with random
# numbers between 0 and 1
m = matrix( runif(100,0,1), nrow = 10)
m
# Calculate the row means  of this matrix. and std of rowmeans
rowMeans(m)
sd(rowMeans(m))

# Remake the above matrix with 100 columns, and 10 rows.
m_new = matrix(runif(1000,0,1), nrow = 10)
m_new
colMeans(m_new)
hist(colMeans(m_new))
```

```r
# question 5

# setting the working dir and reading the data

print(getwd())
setwd("C:/Users/ASUS/OneDrive/Desktop/Study-Material/IIIT-Kottayam/SEM-6/Data warehousing and mining ICS 321/Lab/Lab - 1")
data <- read.csv("cereal.csv")
print(data)

# inspecting the dataset
print(is.data.frame(data))
print(ncol(data))
print(nrow(data))
print(head(data))
print(tail(data))
print(max(data$protein))
print(subset(data,rating==max(rating)))

# adding new column
data$totalcarb <- data$carbo + data$sugars
data

# finding all the rows in the data such that type = 'HOT'
hot=subset(data,data$type=="H")

# counting the number of distinct elements in a col
library(dplyr)
n_distinct(data$mfr)
```

```r
# find the subset whose mfr value == k
kell = subset(data, data$mfr=="K")
kell

# subset satisfying both the conditions
dob = subset(data, data$calories<80 & data$vitamins>20)
dob

# showing only selected cols for the given data set
sug = subset(data,data$sugars>1)
sug
data.frame(sug$name,sug$calories,sug$vitamins)

# writing the kell data into a csv file
write.csv(kell,"C:/Users/ASUS/OneDrive/Desktop/Study-Material/IIIT-Kottayam/SEM-6/Data
warehousing and mining ICS 321/Lab/Lab - 1/kell.csv",row.names = FALSE)

# renaming a column
colnames(data)[colnames(data)=="mfr"] = "Producer"
data

#2020BCS0094
#Abhijit Mishra
```