VIT-AP UNIVERSITY, ANDHRA PRADESH

CSE4027– Data Analytics - Lab Sheet :9

Academic year: 2022-2023 **Branch/ Class:** B.Tech/M.Tech
Semester: Fall **Date: 07-11-22**
Faculty Name: Dr Syed Khasim **School:** SCOPE
Student name: M.Jaswanth **Reg. no.: 20bcd7171**

LAB 9

1. Read the “sample.txt” text file in R. Print the number of characters, number of digits, number of symbols and number of words in the text file.
   Code:

```
data <- read.delim(file = "Sample.txt",header=FALSE,sep=",")
print(data)
cat("\nNumber of characters:",nchar(data), "\n")
cat("\nNumber of digits:",nchar(gsub("\\D", "", data)), "\n")
cat("\nNumber of words:",lengths(gregexpr("\\W+", data)), "\n")
```

   output:

```
> data <- read.delim(file = "Sample.txt",header=FALSE,sep=",")
> print(data)

                                                                                                    V1
1 A wiki (/ËˆwÉªki/ (listen) WIK-ee) is an online hypertext publication collaboratively ed
ts own audience
2
y wiki software
1 [note 1]) is a Hawaiian word meaning quick.[3][6][7] This is a new Line.
> cat("\nNumber of characters:",nchar(data), "\n")

Number of characters: 205 171 221 128 208 128 331 80 139 140 67 126 190 54 25 195
> cat("\nNumber of digits:",nchar(gsub("\\D", "", data)), "\n")

Number of digits: 3 0 0 0 5 0 2 0 0 0 0 4 1 0 0 5
> cat("\nNumber of words:",lengths(gregexpr("\\W+", data)), "\n")

Number of words: 31 28 37 23 37 19 57 12 19 19 8 18 28 8 2 33
>
```

2. Append a new row to the “sample.txt” text file.
   Code:

```
nl <- "This is a new Line."
write(nl,file="Sample.txt",append=TRUE)
data <- read.delim(file = "Sample.txt",header=FALSE,sep=",")
print(data)
```

   output:

```
> nl <- "This is a new Line."
> write(nl,file="Sample.txt",append=TRUE)
> data <- read.delim(file = "Sample.txt",header=FALSE,sep=",")
> print(data)

                                                                                                              V1
1 A wiki (/ËˆwÉªki/ (listen) WIK-ee) is an online hypertext publication collaboratively edited and managed by i
ts own audience
2                                                                                                  Wikis are enabled b
y wiki software
3                                                                                                     The online encyc
lopedia project
4                                                                                                                 This
 is a new Line.
                                                                                                          V2
1  using a web browser. A typical wiki contains multiple pages for the subjects or scope of the project
2                                                       otherwise known as wiki engines. A wiki engine
3                                                                                            Wikipedia
4

              V3
1  and could be either open to the public or limited to use within an organization for maintaining its internal
 knowledge base.
2                                                                                           being a form of a content m
anagement system
3                                                                                                is the most popular wi
ki-based website
4
```

3. How to read this text file with missing values?

   Code:

   mydata <- read.delim(file = "Sample.txt", header=FALSE, na.strings=".")

   print(mydata)

   output:

```
> mydata <- read.delim(file = "Sample.txt", header=FALSE, na.strings=".")
> print(mydata)

                                                                                               V1
1

                                                                A wiki (/ËˆwÉªki/ (listen) WIK-ee) is an online h
ypertext publication collaboratively edited and managed by its own audience, using a web browser. A typical w
iki contains multiple pages for the subjects or scope of the project, and could be either open to the public
 or limited to use within an organization for maintaining its internal knowledge base.
2 Wikis are enabled by wiki software, otherwise known as wiki engines. A wiki engine, being a form of a conte
nt management system, differs from other web-based systems such as blog software, in that the content is crea
ted without any defined owner or leader, and wikis have little inherent structure, allowing structure to emer
ge according to the needs of the users.[1] Wiki engines usually allow content to be written using a simplifie
d markup language and sometimes edited with the help of a rich-text editor.[2] There are dozens of different
 wiki engines in use, both standalone and part of other software, such as bug tracking systems. Some wiki eng
ines are open-source, whereas others are proprietary. Some permit control over different functions (levels of
 access); for example, editing rights may permit changing, adding, or removing material. Others may permit ac
cess without enforcing access control. Other rules may be imposed to organize content.
3                                             The online encyclopedia project, Wikipedia, is the most popular wiki-ba
sed website, and is one of the most widely viewed sites in the world, having been ranked in the top twenty si
nce 2007.[3] Wikipedia is not a single wiki but rather a collection of hundreds of wikis, with each one perta
ining to a specific language. In addition to Wikipedia, there are hundreds of thousands of other wikis in us
e, both public and private, including wikis functioning as knowledge management resources, note-taking tools,
```

4. Read the Iris dataset from csv file and write into a xlsx file in R.

   Code:

   library(xlsx)

   df <- read.csv("Iris.csv")

   write.xlsx(df, file = "Iris.xlsx", sheetName = "Sheet1",col.names = TRUE,row.names = TRUE, append = FALSE)

   output:

```
19  library(xlsx)
20  df <- read.csv("Iris.csv")
21  write.xlsx(df, file = "Iris.xlsx", sheetName = "Sheet1",col.names = TRUE,row.names = TRUE, append = FALSE)
19:1  (Top Level)                                                                    R Script
Console  Terminal  Jobs
R 4.1.2 · C:/Users/Sashank K/Desktop/
d is one of the most widely viewed sites in the world, having been ranked in the top twenty since 2007.[3] Wikipedia is not
 a single wiki but rather a collection of hundreds of wikis, with each one pertaining to a specific language. In addition to
 Wikipedia, there are hundreds of thousands of other wikis in use, both public and private, including wikis functioning as k
nowledge management resources, note-taking tools, community websites, and intranets. The English-language Wikipedia has the
 largest collection of articles: as of February 2020, it has over 6 million articles. Ward Cunningham, the developer of the
 first wiki software, WikiWikiWeb, originally described wiki as the simplest online database that could possibly work.[4] Wi
ki (pronounced [wiki][note 1]) is a Hawaiian word meaning quick.[5][6][7]This is a new Line.
4
                                                                                This is a new Line.
> library(xlsx)
> df <- read.csv("Iris.csv")
> write.xlsx(df, file = "Iris.xlsx", sheetName = "Sheet1",col.names = TRUE,row.names = TRUE, append = FALSE)
>
```

5. Write the covid data set csv file dataset in the second sheet of the xlsx file created for the question4.
   Code:
   df2 <- read.csv("covid_19_clean_complete.csv")
   write.xlsx(df2, file = "Iris.xlsx", sheetName = "Sheet2",col.names = TRUE, row.names = TRUE, append = TRUE)output:

```
> df2 <- read.csv("covid_19_clean_complete.csv")
> write.xlsx(df2, file = "Iris.xlsx", sheetName = "Sheet2",col.names = TRUE, row.names = TRUE, append = TRUE)
Error in .jcall("RJavaTools", "Ljava/lang/Object;", "invokeMethod", cl,  :
  java.lang.OutOfMemoryError: Java heap space
>
```

6. Differentiate scan() and read_table using Iris data set.
   Code:
   data <- scan("Iris.csv",what = "character")
   print(head(data))

   df3 <- read.table('Iris.csv', header = TRUE, sep = ',')
   print(head(df3))
   output:

```
> data <- scan("Iris.csv",what = "character")
Read 153 items
> print(head(data))
[1] "Id,SepalLengthCm,SepalWidthCm,PetalLengthCm,PetalWidthCm,Weight"
[2] "in"
[3] "gm,Species,Season"
[4] "1,5.1,3.5,1.4,0.2,20,Iris-setosa,spring"
[5] "2,4.9,3,1.4,0.2,35,Iris-setosa,summer"
[6] "3,4.7,3.2,1.3,0.2,33,Iris-setosa,fall"
>
> df3 <- read.table('Iris.csv', header = TRUE,  sep = ',')
> print(head(df3))
  Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Weight.in.gm     Species Season
1  1           5.1          3.5           1.4          0.2           20 Iris-setosa spring
2  2           4.9          3.0           1.4          0.2           35 Iris-setosa summer
3  3           4.7          3.2           1.3          0.2           33 Iris-setosa   fall
4  4           4.6          3.1           1.5          0.2           27 Iris-setosa winter
5  5           5.0          3.6           1.4          0.2           41 Iris-setosa spring
6  6           5.4          3.9           1.7          0.4           17 Iris-setosa summer
>
```

7. Use iris dataset and plot the normal distribution on all the numerical columns
   a. dnorm()
   b. pnorm()
   c. qnorm()
   d. rnorm()

code:

```
df4 <- read.csv("Iris.csv")
library("dplyr")
df4 <- select_if(df4, is.numeric)
df4 <- subset(df4,select = c(2,3,4,5,6))
print(head(df4))

dvalues1 <- dnorm(df4$SepalLengthCm)
dvalues2 <- dnorm(df4$SepalWidthCm)
dvalues3 <- dnorm(df4$PetalLengthCm)
dvalues4 <- dnorm(df4$PetalWidthCm)
dvalues5 <- dnorm(df4$Weight.in.gm)
dvalues <- cbind(dvalues1,dvalues2,dvalues3,dvalues4,dvalues5)

for (i in 1:5) {
  plot(dvalues[,i],
       xaxt = "n",
       type = "l",
       main = "pdf of the Standard Normal",
       xlab= "Z-score")
}

pvalues1 <- pnorm(df4$SepalLengthCm)
pvalues2 <- pnorm(df4$SepalWidthCm)
pvalues3 <- pnorm(df4$PetalLengthCm)
pvalues4 <- pnorm(df4$PetalWidthCm)
pvalues5 <- pnorm(df4$Weight.in.gm)
pvalues <- cbind(pvalues1,pvalues2,pvalues3,pvalues4,pvalues5)

for (i in 1:5) {
  plot(pvalues[,i],
```

```
      xaxt = "n",
      type = "l",
      main = "cdf of the Standard Normal",
      xlab= "Quantiles",
      ylab="Probability Density")
}

rvalues1 <- rnorm(df4$SepalLengthCm,mean=70,sd=5)
rvalues2 <- rnorm(df4$SepalWidthCm,mean=70,sd=5)
rvalues3 <- rnorm(df4$PetalLengthCm,mean=70,sd=5)
rvalues4 <- rnorm(df4$PetalWidthCm,mean=70,sd=5)
rvalues5 <- rnorm(df4$Weight.in.gm,mean=70,sd=5)
rvalues <- cbind(rvalues1,rvalues2,rvalues3,rvalues4,rvalues5)

for (i in 1:5) {
  hist(rvalues[,i],
      breaks = 20)
}
```

Output: