Chapter 9: Làm việc với tập tin text, csv, excel

Exercise 1: Đọc tập tin và xử lý

- Cung cấp tập tin: SampleTextFile_10kb.txt
- Đọc tập tin
- In ra nội dung của phần tử đầu tiên trong tập tin
- In số phần tử của tập tin
- Mỗi phần tử có bao nhiêu ký tự?

Exercise 2: Đọc tập tin, xử lý, ghi nội dung vào tập tin

- Cung cấp tập tin: Table0.txt
- Đọc tập tin có cấu trúc dạng bảng. In nội dung tập tin

```
        V1 V2  V3 V4 V5
1     Alex 25 177 57  F
2    Lilly 31 163 69  F
3     Mark 23 190 83  M
4   Oliver 52 179 75  M
5   Martha 76 163 70  F
6    Lucas 49 183 83  M
7 Caroline 26 164 53  F
```

- Đặt tên cho các cột lần lượt là: 'Name', 'Age', 'Height', 'Weight','Sex'

```
        Name Age Height Weight Sex
1     Alex  25    177     57   F
2    Lilly  31    163     69   F
3     Mark  23    190     83   M
4   Oliver  52    179     75   M
5   Martha  76    163     70   F
6    Lucas  49    183     83   M
7 Caroline  26    164     53   F
```

- Đặt row.names là dữ liệu cột ‘Name’
- Xóa bỏ cột ‘Name’ thừa
- In nội dung

```
         Age Height Weight Sex
Alex      25    177     57   F
Lilly     31    163     69   F
Mark      23    190     83   M
Oliver    52    179     75   M
Martha    76    163     70   F
Lucas     49    183     83   M
Caroline  26    164     53   F
```

- Ghi nội dung này vào tập tin Table0_new.txt

Exercise 3: Đọc, ghi tập tin txt

- Cung cấp tập tin: Table6.txt
- Đọc tập tin có cấu trúc dạng bảng, có lấy header, bỏ đi các Strings as Factors, bỏ đi các comment => df. In head của df.
- Kiểm tra xem data có dữ liệu trùng không, nếu có thì xóa các dòng bị trùng
- In nội dung sau khi đã bỏ dòng trùng.

Name	Age	Height	Weight	Sex
Alex	25	177	57	F
Lilly	31	163	69	F
Mark	23	190	83	M
Oliver	52	179	75	M
Martha	76	163	70	F
Lucas	49	183	83	M
Caroline	26	164	53	F

- Đặt rownames là nội dung của cột Name, sau đó xóa bỏ cột Name. In kết quả.

	Age	Height	Weight	Sex
Alex	25	177	57	F
Lilly	31	163	69	F
Mark	23	190	83	M
Oliver	52	179	75	M
Martha	76	163	70	F
Lucas	49	183	83	M
Caroline	26	164	53	F

- Ghi lại nội dung sau khi đã chuẩn hóa vào tập tin Table6_new.txt

Exercise 4: Đọc, xử lý, ghi tập tin csv

- Cung cấp tập tin: states3.csv
- Đọc tập tin, row.names là cột state, chú ý dấu phân cách thập phân và dấu tách => df. In haed của df
- Từ df, tạo ra df1 có income > 4500 và population >10000

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Illinois	11197	5107	0.9	70.14	10.3	52.6	127	55748
New York	18076	4903	1.4	70.55	10.9	52.7	82	47831
Ohio	10735	4561	0.8	70.82	7.4	53.2	124	40975

- Ghi bộ dữ liệu này vào tập tin states_income.csv
- Đọc và hiển thị lại dữ liệu vừa ghi vào tập tin

Exercise 5: Đọc tập tin csv, xử lý, lưu vào tập tin xlsx

- Cung cấp tập tin: medals.csv
- Đọc nội dung => data. In head của data

Year	City	Sport	Discipline	NOC	Event	Event.gender	Medal
1924	Chamonix	Skating	Figure skating	AUT	individual	M	Silver
1924	Chamonix	Skating	Figure skating	AUT	individual	W	Gold
1924	Chamonix	Skating	Figure skating	AUT	pairs	X	Gold
1924	Chamonix	Bobsleigh	Bobsleigh	BEL	four-man	M	Bronze
1924	Chamonix	Ice Hockey	Ice Hockey	CAN	ice hockey	M	Gold
1924	Chamonix	Biathlon	Biathlon	FIN	military patrol	M	Silver

- Cho biết kiểu dữ liệu của data.
- Cho biết số dòng, số cột và tên các cột của data
- Tạo một bộ dữ liệu data_1932_skiing với Year = 1932 và Sport = "Skiing"
- Ghi df_sub vào file excel medals_new.xlsx, tên sheet là 1932_skiing, cho phép ghi tiếp.

Exercise 6: Đọc và xử lý tập tin xlsx

- Đọc nội dung tập tin đã tạo ở bài trên (medals_new.xlsx) => data. In head của data
- Hiệu chỉnh lại nội dung bằng cách bỏ cột NA
- Cho biết có bao nhiêu huy chương vàng trong danh sách này
- Chi tiết về các huy chương vàng này

	Year	City	Sport	Discipline	NOC	Event	Event.gender	Medal
2	1932	Lake Placid	Skiing	Cross Country S	FIN	50km	M	Gold
6	1932	Lake Placid	Skiing	Nordic Combined	NOR	individual	M	Gold
9	1932	Lake Placid	Skiing	Ski Jumping	NOR	K90 individual (70m)	M	Gold
11	1932	Lake Placid	Skiing	Cross Country S	SWE	18km	M	Gold

Gợi ý

Exercise 1: Đọc tập tin và xử lý

In [2]:
```
#SampleTextFile_10kb.txt
content <- readLines("Du_lieu/SampleTextFile_10kb.txt")
print("*** First element in content:")
print(content[1])
print(paste("*** Number of elements:", length(content)))
print("*** Number of characters in each element:")
```

```
[1] "*** First element in content:"
[1] "Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus condiment
um sagittis lacus, laoreet luctus ligula laoreet ut. Vestibulum ullamcorper acc
umsan velit vel vehicula. Proin tempor lacus arcu. Nunc at elit condimentum, se
mper nisi et, condimentum mi. In venenatis blandit nibh at sollicitudin. Vestib
ulum dapibus mauris at orci maximus pellentesque. Nullam id elementum ipsum. Su
spendisse cursus lobortis viverra. Proin et erat at mauris tincidunt porttitor
vitae ac dui."
[1] "*** Number of elements: 27"
[1] "*** Number of characters in each element:"
```

In [5]:
```
i<-1
while(i<=length(content)){
  print(paste("Element", i, "has", nchar(content[i]), "chars"))
  i<- i+1
}
```

```
[1] "Element 1 has 482 chars"
[1] "Element 2 has 0 chars"
[1] "Element 3 has 931 chars"
[1] "Element 4 has 0 chars"
[1] "Element 5 has 878 chars"
[1] "Element 6 has 0 chars"
[1] "Element 7 has 601 chars"
[1] "Element 8 has 0 chars"
[1] "Element 9 has 645 chars"
[1] "Element 10 has 0 chars"
[1] "Element 11 has 730 chars"
[1] "Element 12 has 0 chars"
[1] "Element 13 has 437 chars"
[1] "Element 14 has 0 chars"
[1] "Element 15 has 434 chars"
[1] "Element 16 has 0 chars"
[1] "Element 17 has 944 chars"
[1] "Element 18 has 0 chars"
[1] "Element 19 has 629 chars"
[1] "Element 20 has 0 chars"
[1] "Element 21 has 772 chars"
[1] "Element 22 has 0 chars"
[1] "Element 23 has 615 chars"
[1] "Element 24 has 0 chars"
[1] "Element 25 has 449 chars"
[1] "Element 26 has 0 chars"
[1] "Element 27 has 909 chars"
```

Exercise 2: Đọc tập tin, xử lý, ghi nội dung vào tập tin

In [8]:
```
df <- read.table("Du_lieu/Table0.txt")
print("Content of file:")
print(df)
```

```
[1] "Content of file:"
        V1 V2  V3 V4 V5
1     Alex 25 177 57  F
2    Lilly 31 163 69  F
3     Mark 23 190 83  M
4   Oliver 52 179 75  M
5   Martha 76 163 70  F
6    Lucas 49 183 83  M
7 Caroline 26 164 53  F
```

In [9]:
```
#dat ten cho cac cot
names(df) <- c('Name', 'Age', 'Height', 'Weight','Sex')
print(df)
```

```
      Name Age Height Weight Sex
1     Alex  25    177     57   F
2    Lilly  31    163     69   F
3     Mark  23    190     83   M
4   Oliver  52    179     75   M
5   Martha  76    163     70   F
6    Lucas  49    183     83   M
7 Caroline  26    164     53   F
```

In [10]:
```
#dat rowname la du lieu cot Name
row.names(df)<-df$Name
print(df)
```

```
              Name Age Height Weight Sex
Alex          Alex  25    177     57   F
Lilly        Lilly  31    163     69   F
Mark          Mark  23    190     83   M
Oliver      Oliver  52    179     75   M
Martha      Martha  76    163     70   F
Lucas        Lucas  49    183     83   M
Caroline  Caroline  26    164     53   F
```

In [11]:
```
#sau do bo di cot name thua
df$Name <- NULL
print("Content of data frame:")
print(df)
```

```
[1] "Content of data frame:"
         Age Height Weight Sex
Alex      25    177     57   F
Lilly     31    163     69   F
Mark      23    190     83   M
Oliver    52    179     75   M
Martha    76    163     70   F
Lucas     49    183     83   M
Caroline  26    164     53   F
```

In [12]:
```
# ghi noi dung tap tin
print("Write to file...")
write.table(df,file="Du_lieu/Table0_new.txt",sep="\t")
print("Write completely!")
```

```
[1] "Write to file..."
[1] "Write completely!"
```

In [28]:
```
df_new <- read.table("Du_lieu/Table0_new.txt")
print("Read saved file:")
print(df_new)
```

```
[1] "Read saved file:"
        Age Height Weight Sex
Alex     25    177     57   F
Lilly    31    163     69   F
Mark     23    190     83   M
Oliver   52    179     75   M
Martha   76    163     70   F
Lucas    49    183     83   M
Caroline 26    164     53   F
```

Exercise 3: Đọc, ghi tập tin txt

In [29]:
```
# doc du lieu
# co lay header
# bo di cac strings As Factors
# bo di cac comment
df <- read.table("Du_lieu/Table6.txt",
                 skip = 1,
                 header = TRUE,
                 flush = TRUE,
                 comment.char = "@",
                 stringsAsFactors = FALSE)
print("Content of file after reading:")
print(head(df))
```

```
[1] "Content of file after reading:"
    Name Age Height Weight Sex
1   Alex  25    177     57   F
2  Lilly  31    163     69   F
3   Mark  23    190     83   M
4 Oliver  52    179     75   M
5 Martha  76    163     70   F
6  Lucas  49    183     83   M
```

In [30]:
```
print("Before dropping: ")
print(paste("Rows:", nrow(df)))
```

```
[1] "Before dropping: "
[1] "Rows: 105"
```

In [31]:
```
library(tidyverse)
```

In [32]:
```
df = df %>% unique()
print("After dropping: ")
print(paste("Rows:", nrow(df)))
```

```
[1] "After dropping: "
[1] "Rows: 7"
```

In [33]:
```
df
```

Name	Age	Height	Weight	Sex
Alex	25	177	57	F
Lilly	31	163	69	F
Mark	23	190	83	M
Oliver	52	179	75	M
Martha	76	163	70	F
Lucas	49	183	83	M
Caroline	26	164	53	F

In [34]:
```
rownames(df) <- df$Name
df$Name <- NULL
df
```

	Age	Height	Weight	Sex
Alex	25	177	57	F
Lilly	31	163	69	F
Mark	23	190	83	M
Oliver	52	179	75	M
Martha	76	163	70	F
Lucas	49	183	83	M
Caroline	26	164	53	F

In [35]:
```
# ghi noi dung tap tin
print("Write to file...")
write.table(df,file="Du_lieu/Table6_new.txt",sep="\t")
print("Write completely!")
```

```
[1] "Write to file..."
[1] "Write completely!"
```

In [36]:
```
df_new <- read.table("Du_lieu/Table6_new.txt")
print("Read saved file:")
print(df_new)
```

```
[1] "Read saved file:"
         Age Height Weight Sex
Alex      25    177     57   F
Lilly     31    163     69   F
Mark      23    190     83   M
Oliver    52    179     75   M
Martha    76    163     70   F
Lucas     49    183     83   M
Caroline  26    164     53   F
```

Exercise 4: Đọc, xử lý, ghi tập tin csv

In [38]:
```
df <- read.csv("Du_lieu/states2.csv",
               row.names = 1,
               sep = ";",
               dec = ",")
print(head(df))
```

```
           Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
```

In [39]:
```
# tao ra bo du lieu ma income >4500 va population >10000
df1 <- subset(df, df$Income > 4500 & df$Population >10000)
print("Data set of income >4500 and population >10000:")
print(df1)
```

```
[1] "Data set of income >4500 and population >10000:"
           Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Illinois        11197   5107        0.9    70.14   10.3    52.6   127  55748
New York        18076   4903        1.4    70.55   10.9    52.7    82  47831
Ohio            10735   4561        0.8    70.82    7.4    53.2   124  40975
```

In [40]:
```
# luu bo du lieu nay vao tap tin states_income.csv
write.csv(df1, file = "Du_lieu/states_income.csv")
```

In [41]:
```
#doc du lieu trong file states_income.csv
dfa <- read.csv("Du_lieu/states_income.csv",row.names = 1)
print("Data in states_income.csv ")
print(dfa)
```

```
[1] "Data in states_income.csv "
           Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
California      21198   5114        1.1    71.71   10.3    62.6    20 156361
Illinois        11197   5107        0.9    70.14   10.3    52.6   127  55748
New York        18076   4903        1.4    70.55   10.9    52.7    82  47831
Ohio            10735   4561        0.8    70.82    7.4    53.2   124  40975
```

In [42]:
```
# solution 2
library(tidyverse)
```

In [43]:
```
df11 <- df %>% filter(Income > 4500, df$Population >10000)
df11
```

Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area
21198	5114	1.1	71.71	10.3	62.6	20	156361
11197	5107	0.9	70.14	10.3	52.6	127	55748
18076	4903	1.4	70.55	10.9	52.7	82	47831
10735	4561	0.8	70.82	7.4	53.2	124	40975

Exercise 5: Đọc tập tin csv, xử lý, lưu vào tập tin xlsx

In [44]:
```
# Load the library into R workspace.
library("xlsx")
```

```
Loading required package: rJava

Loading required package: xlsxjars
```

In [47]:
```
#doc tap tin school.csv
data <- read.csv("Du_lieu/medals.csv")
head(data)
```

Year	City	Sport	Discipline	NOC	Event	Event.gender	Medal
1924	Chamonix	Skating	Figure skating	AUT	individual	M	Silver
1924	Chamonix	Skating	Figure skating	AUT	individual	W	Gold
1924	Chamonix	Skating	Figure skating	AUT	pairs	X	Gold
1924	Chamonix	Bobsleigh	Bobsleigh	BEL	four-man	M	Bronze
1924	Chamonix	Ice Hockey	Ice Hockey	CAN	ice hockey	M	Gold
1924	Chamonix	Biathlon	Biathlon	FIN	military patrol	M	Silver

In [48]:
```
print(paste("Data type: ",class(data)))
print(paste("Number of rows:",ncol(data)))
print(paste("Number of cols:",nrow(data)))
print(paste("Name of columns", toString(names(data))))
```

```
[1] "Data type:  data.frame"
[1] "Number of rows: 8"
[1] "Number of cols: 2314"
[1] "Name of columns Year, City, Sport, Discipline, NOC, Event, Event.gender, M
edal"
```

In [49]:
```
# tao bo du lieu co year = 1932, Sport = 'Skiing'
data_1932_skiing <- subset(data, data$Year == 1932 & data$Sport=='Skiing')
print("Data set of year = 1932 and Sport = Skiing:")
data_1932_skiing
```

```
[1] "Data set of year = 1932 and Sport = Skiing:"
```

	Year	City	Sport	Discipline	NOC	Event	Event.gender	Medal
100	1932	Lake Placid	Skiing	Cross Country S	FIN	18km	M	Bronze
101	1932	Lake Placid	Skiing	Cross Country S	FIN	50km	M	Gold
102	1932	Lake Placid	Skiing	Cross Country S	FIN	50km	M	Silver
110	1932	Lake Placid	Skiing	Cross Country S	NOR	50km	M	Bronze
111	1932	Lake Placid	Skiing	Nordic Combined	NOR	individual	M	Bronze
112	1932	Lake Placid	Skiing	Nordic Combined	NOR	individual	M	Gold
113	1932	Lake Placid	Skiing	Nordic Combined	NOR	individual	M	Silver
114	1932	Lake Placid	Skiing	Ski Jumping	NOR	K90 individual (70m)	M	Bronze
115	1932	Lake Placid	Skiing	Ski Jumping	NOR	K90 individual (70m)	M	Gold
116	1932	Lake Placid	Skiing	Ski Jumping	NOR	K90 individual (70m)	M	Silver
119	1932	Lake Placid	Skiing	Cross Country S	SWE	18km	M	Gold
120	1932	Lake Placid	Skiing	Cross Country S	SWE	18km	M	Silver

In [50]:
```
# ghi bo du lieu nay vao file excel medals_new.xlsx
print("Write to file...")
write.xlsx(data_1932_skiing, file = "Du_lieu/medals_new.xlsx",
           sheetName = "1932_skiing", append = TRUE)
print("Finish!")
```

```
[1] "Write to file..."
[1] "Finish!"
```

Exercise 6: Đọc và xử lý tập tin xlsx

In [56]:
```
# Read the first worksheet in the file medals_1.xlsx
data_1932_skiing <- read.xlsx("Du_lieu/medals_new.xlsx", sheetName = "1932_Skiing
print("Content of file:")
head(data_1932_skiing)
```

```
[1] "Content of file:"
```

NA.	Year	City	Sport	Discipline	NOC	Event	Event.gender	Medal
100	1932	Lake Placid	Skiing	Cross Country S	FIN	18km	M	Bronze
101	1932	Lake Placid	Skiing	Cross Country S	FIN	50km	M	Gold
102	1932	Lake Placid	Skiing	Cross Country S	FIN	50km	M	Silver
110	1932	Lake Placid	Skiing	Cross Country S	NOR	50km	M	Bronze
111	1932	Lake Placid	Skiing	Nordic Combined	NOR	individual	M	Bronze
112	1932	Lake Placid	Skiing	Nordic Combined	NOR	individual	M	Gold

In [57]:
```
#xoa bo cot NA.
data_1932_skiing$NA. <- NULL
print("Data set now")
print(head(data_1932_skiing))
```

```
[1] "Data set now"
  Year        City  Sport      Discipline NOC      Event Event.gender  Medal
1 1932 Lake Placid Skiing Cross Country S FIN       18km             M Bronze
2 1932 Lake Placid Skiing Cross Country S FIN       50km             M   Gold
3 1932 Lake Placid Skiing Cross Country S FIN       50km             M Silver
4 1932 Lake Placid Skiing Cross Country S NOR       50km             M Bronze
5 1932 Lake Placid Skiing Nordic Combined NOR individual             M Bronze
6 1932 Lake Placid Skiing Nordic Combined NOR individual             M   Gold
```

In [58]:
```
#cho biet co bao nhieu huy chuong "Gold" trong du lieu nay
n <- sum(data_1932_skiing$Medal == "Gold")
print(paste("Number of Golden medals:", n))
```

```
[1] "Number of Golden medals: 4"
```

In [60]:
```
#thong tin chi tiet
data_golden_medals <- subset(data_1932_skiing,data_1932_skiing$Medal == "Gold")
print("Details:")
data_golden_medals
```

```
[1] "Details:"
```

	Year	City	Sport	Discipline	NOC	Event	Event.gender	Medal
2	1932	Lake Placid	Skiing	Cross Country S	FIN	50km	M	Gold
6	1932	Lake Placid	Skiing	Nordic Combined	NOR	individual	M	Gold
9	1932	Lake Placid	Skiing	Ski Jumping	NOR	K90 individual (70m)	M	Gold
11	1932	Lake Placid	Skiing	Cross Country S	SWE	18km	M	Gold