Data Analysis and Visualization Exercise 2.2: Data import

Felix Brechtmann, Matthias Heinig, Vicente Yépez

29 October 2019

1 Flat file questions

1.1 Question

Read the titanic file (titanic.csv) and figure out who was the oldest surviving passenger of the titanic accident? Hint: `?subset`

1.2 Question

For the next two questions we simulate files as strings. They can be read, as if they where files.

A csv file has numbers as column names in the first row. Which parameter of read.table() needs to be adjusted to read the column names as they are in the csv?

```
tmp_tidy_table <- "1_colname,2_colname,3_colname
  3,4,5
  a,b,c"
read.csv(text = tmp_tidy_table)
##   X1_colname X2_colname X3_colname
## 1          3          4          5
## 2          a          b          c
```

1.3 Question

How to read the following table to have the `identical()` information as in `tidy_txt_df` from question above?

```
tmp_messy_table <- "# This line is just useless info

  1_colname,2_colname,3_colname
  3,4,5

  a,b,c"
```

2 Excel questions

2.1 Question

Read only `Name`, `Type` and `Total` columns for only the first 10 pokemons of the pokemon.xlsx file. Hint: take a look at the file using Excel or any other spreadsheet application.

2.2 Question

Using the summer_olympic_medals.xlsx file, which athlete won most bronze medals?

```
oly_file <- file.path('extdata/summer_olympic_medals.xlsx')
```

2.3 Question

Are the columns `Gender` and `Event_gender` consistent? Find inconsistent gender entries.

2.4 Question

Which country won most medals? Which country has the highest ratio of silver medals? Use the data in the country summary sheet starting at row 147 of the summer_olympic_medals.xlsx file.

2.5 Question

Which countries did participate, but without winning medals? Assume, that all countries listed in the IOC COUNTRY CODES sheet participated.

3 SQL questions

3.1 Question

Connect to the `extdata/Northwind.sl3` SQLite data base (using the 'RSQLite' package). Inspect the data base tables using the 'dbListTables' and 'dbListFields' functions. Put together a SQL statement to retrieve a table that lists for all customers (name of the company, name of the contact person and city) all the products (name of the product) that they ordered. Execute the statement using 'dbGetQuery'. How many rows does this table have? Display the first 5 rows.

We provide the SQL statement here:

```
        "select customers.companyname, customers.contactname,
        customers.city, products.productname from customers inner join
        orders on customers.customerid = orders.customerid inner join
        `order details` on orders.orderid = `order details`.orderid inner
        join products on `order details`.productid = products.productid"
```

4 XML questions

4.1 Question

Load the XML document `plant_catalog.xml`. Use XPath and DOM functions to find out all unique element names in the document.

Get all plants of zone 4 and transform the data into an R list. Hint: ‘xmlToList’

4.2 Question

Read the HTML tables from the website https://www.skysports.com/premier-league-table into your workspace.

Which team is currently placed first in the premier league?

5 Prepare for next lecture

For data manipulation with the `data.table` package please read this intro.