DSO 545: Statistical Computing and Data Visualization

Midterm 02

Spring 2016

Instructions

- This is an open book exam. You are allowed to use your notes as well as online sources
- You can't communicate with ANY PERSON in or outside the class during the exam period
- If you are asked to create an exact copy of some graph, make sure your graph look exactly the same. Pay attention to axises names, legend name/ position, order of bars etc.
- If you save your result in a variable, make sure to print it out.
- Use the file `lastname_firstname.rmd` as a template for your solution, and make sure to save this file under your name
- Answer all questions below
- Don't change the data file names
- Submit both the RMD and PDF file to blackboard
- You have 80 mins to finish this exam
- If you are well prepared, you should be able to finish the exam in 80 minutes.
- Suggested **max** times to spend on each case:

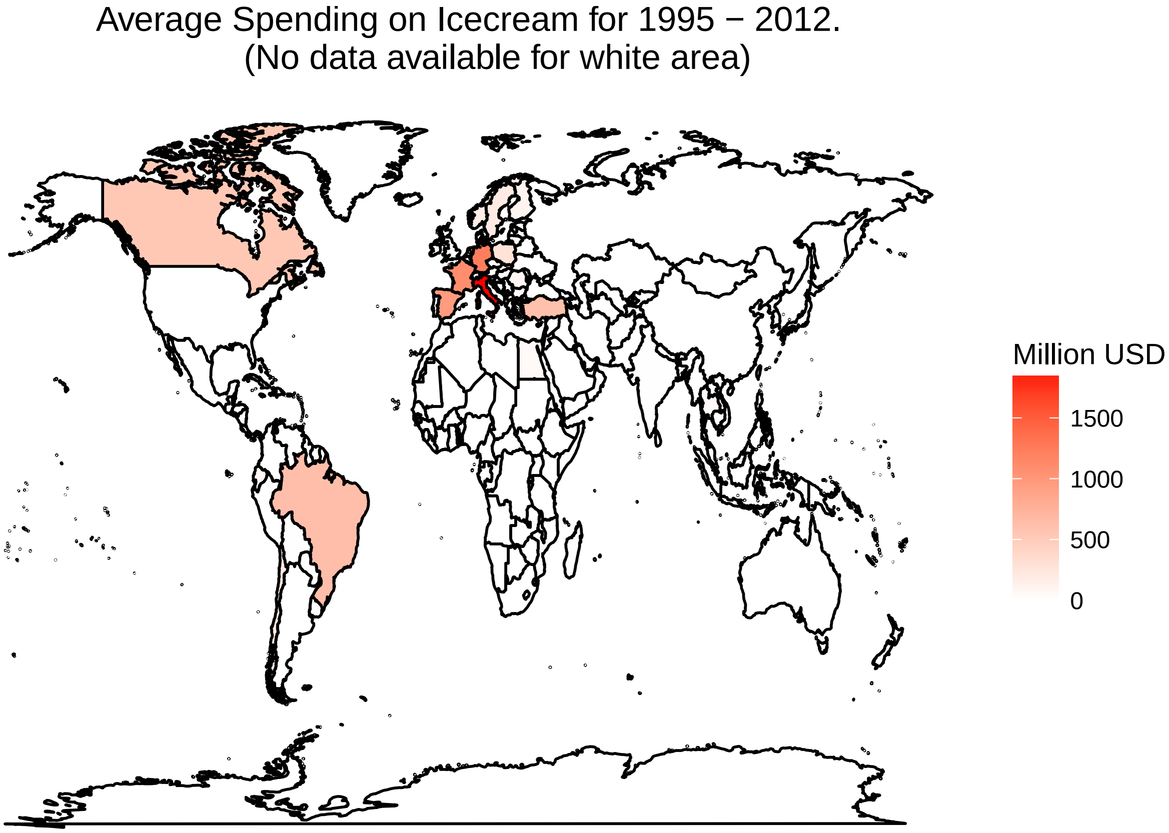
Download	Case 01	Case 02	Case 03	Upload
5 mins	25 mins	30 mins	15 mins	5 mins

Case 01: There is always Room for Icecream!

The United Nations Industrial Commodity Statistics Database provides annual statistics on the production of major industrial commodities by country. Data are provided in terms of physical quantities as well as monetary value. The online database covers the years 1995 to 2012 (Source: http://data.un.org/Data.aspx?d=ICS&f=cmID:22970-0). You can find the dataset in `icecream.csv` on blackboard.

1. (1.5 points) Aggregate the dataset to find the average spending on icecream for all listed countries over the given years.

2. (3 points) Create a chloropleth map that shows the average spending on icecream for the listed countires over the given years. Your map would look as follows:


Average Spending on Icecream for 1995 − 2012.
(No data available for white area)
Million USD
1500
1000
500
0

Case 02: MLS Players Compensation

Since football season is not here yet, you might want to enjoy some soccer for the rest of the summer. Major League Soccer (MLS) is a professional soccer league representing the sport's highest level in both the United States and Canada. In this case, we will be scrapping the compensations for the players coming to MLS.

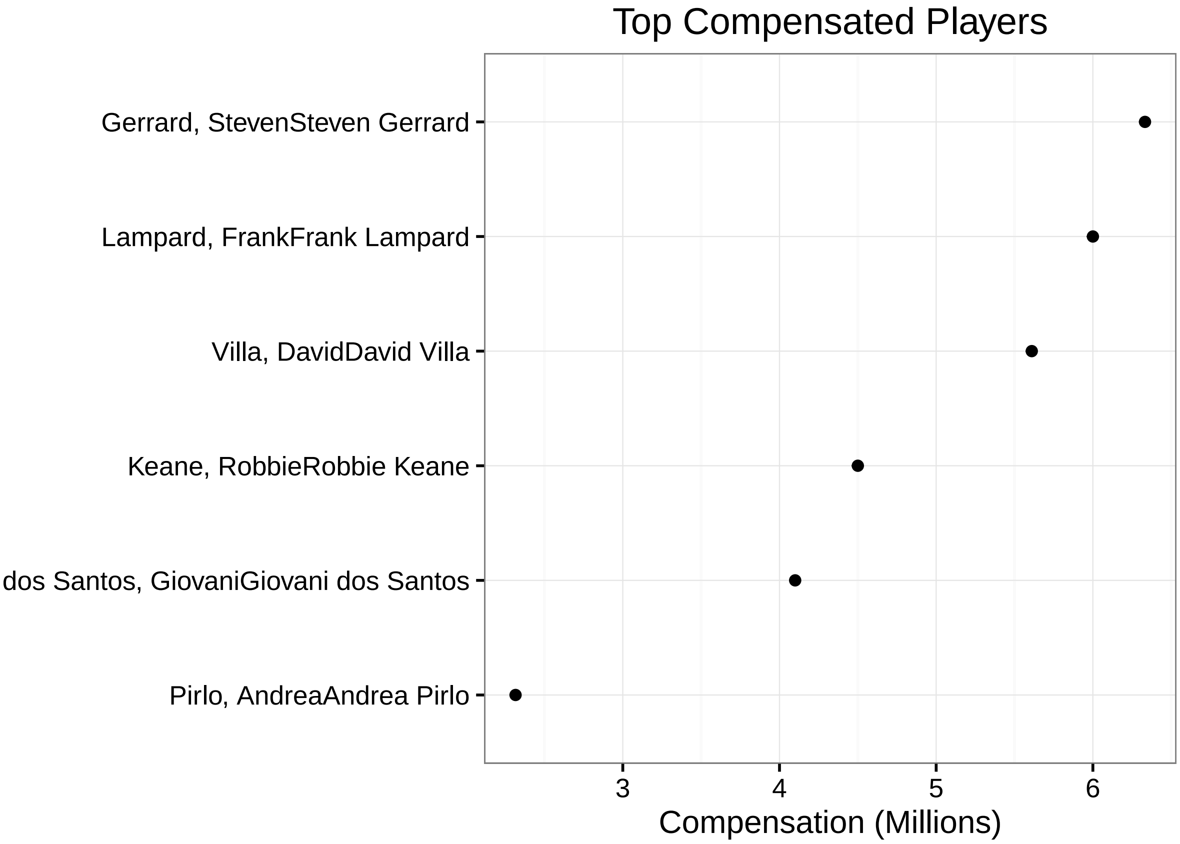
The wikipedia page (`http://en.wikipedia.org/wiki/Designated_Player_Rule`) has a table of those compensations!

1. (1 point) Use `rvest` R package to scrape the data table. Save it to `players`. The xpath for the table is: `//*[@id="mw-content-text"]/table[1]`

2. (1 point) Clean the column with compensation information. Change the column type to `numeric`, and rename it `Compensation`. (Hint: The dollar sign is a wild character, so you need to escape it in your code!)

3. (1 point) Create a subset of `players` called `NYLAplayers`, which only contains records of players currently playing for `New York City FC` or `LA Galaxy`, and order your subset by `Compensation` in decreasing order. Do you know any of these big names? Let's go down to Stubhub Center to watch them live! (Hint: You might need to rename the columns first!)

4. (1 point) Visualize the `NYLAplayers` compensation as follows:



Case 03: How much does Joey Owe Chandler in Friends TV Show?

(1.5 points) Have you seen Friends? In season 8 episode 22 of Friends, Joey is figuring out how much money he owes Chandler for rent, acting lessons, dance lessons, head shots, etc. How much did Joey owe Chandler?

The text files `friends.txt` summarizes a conversation between Joey and Chandler. Use `stringr` R package and regular expressions to help with the math. **What is the total amount that Joey owes Chandler**?

Hints:

1. The amounts have dollar signs
2. The $ sign is a wild character
3. Remember that most of `stringr` functions return a list. So in order to access the elements in a list, you need to use double square brackets. e.g. amount[[1]] returns the first element in a list.

```r
# read the text file as follows
library(stringr)
fileName <- "friends.txt"
text <- readChar(fileName, file.info(fileName)$size)
```