DSO 545: Statistical Computing and Data Visualization

Midterm 02

Spring 2016

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. Aggregate the dataset to find the average spending on icecream for all listed countries over the given years.

```
setwd("C:/Users/dan_9/Desktop/DSO 545/Final exam/FINAL")
icecream <- read.csv("icecream.csv")

#### INSERT YOUR CODE HERE
glimpse(icecream)

## Observations: 395
## Variables: 3
## $ Country.or.Area (fctr) Albania, Albania, Albania, Albania, Albania, ...
## $ Year            (int) 2010, 2009, 2007, 2006, 2005, 2004, 1999, 1998...
## $ USDinMillions   (dbl) 4.387297, 8.265061, 2.289117, 1.618701, 1.0683...

Q1p1 = icecream %>% group_by(Country.or.Area)  %>% summarise(MillionUSD =
mean(USDinMillions))

# FinalAnswer:
Q1p1

## Source: local data frame [45 x 2]
##
##    Country.or.Area MillionUSD
##             (fctr)      (dbl)
## 1          Albania   3.225439
## 2          Bolivia   5.487864
## 3           Brazil 637.807535
## 4         Bulgaria  32.672548
## 5           Canada 545.672454
## 6            Chile 208.901264
## 7           Cyprus  16.868139
## 8   Czech Republic  44.165245
## 9          Denmark  97.575150
## 10         Ecuador  64.974000
## ..             ...        ...
```

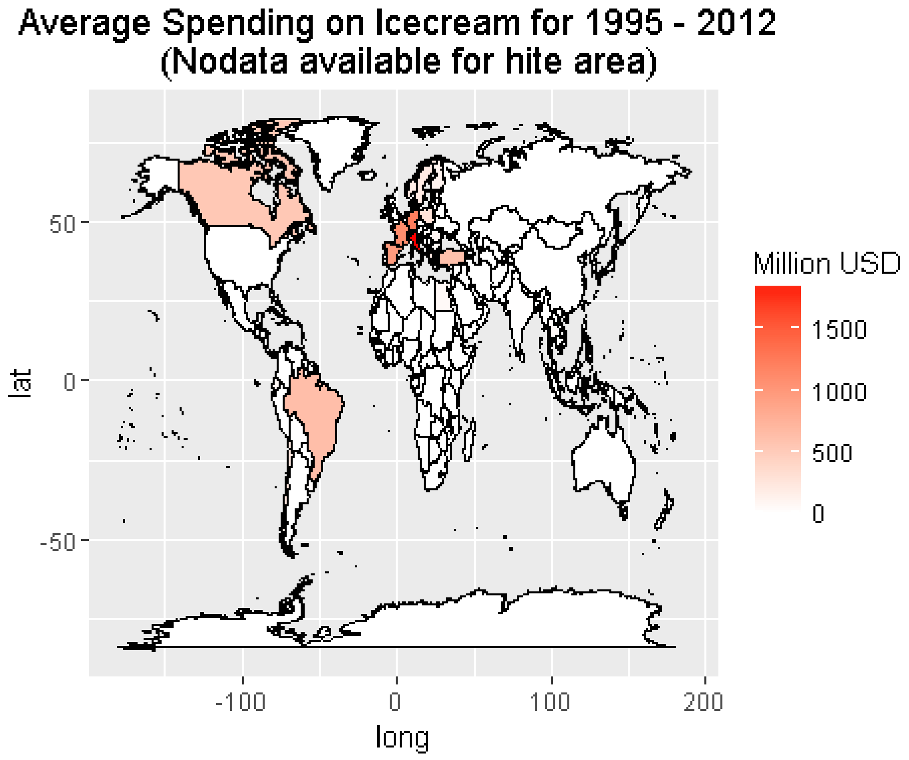
2. Create a chloropleth map that shows the average spending on icecream for the listed countires over the given years. Your map would look as follows:

```
#### INSERT YOUR CODE HERE
world_map = map_data("world")
world_map$region = tolower(world_map$region)
Q1p1$Country.or.Area = tolower(Q1p1$Country.or.Area)

glimpse(world_map)

## Observations: 99,338
## Variables: 6
## $ long      (dbl) -69.89912, -69.89571, -69.94219, -70.00415, -70.0661...
## $ lat       (dbl) 12.45200, 12.42300, 12.43853, 12.50049, 12.54697, 12...
## $ group     (dbl) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2...
## $ order     (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 1...
## $ region    (chr) "aruba", "aruba", "aruba", "aruba", "aruba", "aruba"...
## $ subregion (chr) NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...

countryIceCream = merge( world_map,Q1p1, by.x = "region", by.y =
"Country.or.Area", all.x = T)
countryIceCream = arrange(countryIceCream, group, order)

countryIceCream$MillionUSD[is.na(countryIceCream$MillionUSD)] = 0
table(countryIceCream$MillionUSD)

##
##                0      0.581282097      0.614546478 1.66615976428571
##            68594              202              266              233
##          1.98894     2.0429754595      3.225438587      5.487864443
##              170              225              113              418
## 5.49040877933333 10.1844749322222   13.50046640375          14.1224
##             1518              145              453               74
## 15.6613105444286 16.8681393883333      18.58058711 19.7378401192857
##              164              100              132              140
##     20.092385565      28.14098365    32.2847646575 32.6725479853333
##               96              195              143              179
##   34.59419668437 36.8537038666667 44.1652451183333 49.5903481663636
##              155              315              240              214
##   54.38829247125 57.6464216690909 57.7065138322222           64.974
##              186              253              578              347
## 86.7646670472727      97.57514984  110.27110395125     168.47210785
##              274              298              571              282
##     174.74600518     180.56938106 197.409124246471      208.9012642
##              923             1985              593             2006
## 273.885329123077   545.6724544875   615.6599711875 637.807534961538
##              316            11573              562             1885
## 971.278114533333 1094.87570298333 1228.36117975556     1887.4311595
##              448              605              568              601
```

```r
glimpse(countryIceCream)
```

```
## Observations: 99,338
## Variables: 7
## $ region     (chr) "aruba", "aruba", "aruba", "aruba", "aruba", "aruba...
## $ long       (dbl) -69.89912, -69.89571, -69.94219, -70.00415, -70.066...
## $ lat        (dbl) 12.45200, 12.42300, 12.43853, 12.50049, 12.54697, 1...
## $ group      (dbl) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, ...
## $ order      (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, ...
## $ subregion  (chr) NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ MillionUSD (dbl) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
```

```r
# FinalAnswer:
ggplot(countryIceCream, aes(x = long, y = lat, group = group, fill =
MillionUSD))+
  geom_polygon(color = "black") +
  scale_fill_continuous(low = "white", high = "red", name = "Million USD")+
  ggtitle("Average Spending on Icecream for 1995 - 2012 \n (Nodata available
for hite area)")
```


Average Spending on Icecream for 1995 - 2012
(Nodata available for hite area)
50
0
-50
lat
-100
0
100
200
long
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 games 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. 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]`

```
#### INSERT YOUR CODE HERE
urlC2 = "http://en.wikipedia.org/wiki/Designated_Player_Rule"
tbl = urlC2 %>%
  html() %>%
  html_nodes(xpath = '//*[@id="mw-content-text"]/table[1]' ) %>%
  html_table()
players = tbl[[1]]

# FinalAnswer:
players
```

```
##    Year signed                                          Player  Nation
## 1         2011                       Keane, RobbieRobbie Keane   Â IRL
## 2         2012             HiguaÃ-n, FedericoFederico HiguaÃ-n   Â ARG
## 3         2013                       Valeri, DiegoDiego Valeri   Â ARG
## 4         2014                       Laba, MatÃ-asMatÃ-as Laba   Â ARG
## 5         2013                     Dempsey, ClintClint Dempsey   Â USA
## 6         2013                           Diaz, MauroMauro Diaz   Â ARG
## 7         2014                              Gilberto, Gilberto   Â BRA
## 8         2014                 Bradley, MichaelMichael Bradley   Â USA
## 9         2014                         Edu, MauriceMaurice Edu   Â USA
## 10        2014                     Morales, PedroPedro Morales   Â CHI
## 11        2014                         Villa, DavidDavid Villa   Â ESP
## 12        2014                   Ridgewell, LiamLiam Ridgewell   Â ENG
## 13        2014                                    KakÃ¡, KakÃ¡   Â BRA
## 14        2014                   Piatti, IgnacioIgnacio Piatti   Â ARG
## 15        2014                         Besler, MattMatt Besler   Â USA
## 16        2014                         Zusi, GrahamGraham Zusi   Â USA
## 17        2014               Beasley, DaMarcusDaMarcus Beasley   Â USA
## 18        2014   PÃ©rez GarcÃ-a, MatÃ-asMatÃ-as PÃ©rez GarcÃ-a   Â ARG
## 19        2015 Wright-Phillips, BradleyBradley Wright-Phillips   Â ENG
## 20        2015                 Castillo, FabianFabian Castillo   Â COL
## 21        2015                     RÃ³chez, BryanBryan RÃ³chez   Â HON
## 22        2015                         Accam, DavidDavid Accam   Â GHA
## 23        2015                       Torres, ErickErick Torres   Â MEX
## 24        2015                   Rivero, OctavioOctavio Rivero   Â URU
## 25        2015                   Gerrard, StevenSteven Gerrard   Â ENG
## 26        2015                     Lampard, FrankFrank Lampard   Â ENG
## 27        2015                     Altidore, JozyJozy Altidore   Â USA
## 28        2015           Giovinco, SebastianSebastian Giovinco   Â ITA
## 29        2015           EspÃ-ndola, FabiÃ¡nFabiÃ¡n EspÃ-ndola   Â ARG
## 30        2015             Emeghara, InnocentInnocent Emeghara Â Â SUI
## 31        2015                       Rivas, CarlosCarlos Rivas   Â COL
## 32        2015                           Plata, JoaoJoao Plata   Â ECU
## 33        2015                         Doyle, KevinKevin Doyle   Â IRL
## 34        2015                       Pirlo, AndreaAndrea Pirlo   Â ITA
## 35        2015           dos Santos, GiovaniGiovani dos Santos   Â MEX
## 36        2015                       Melano, LucasLucas Melano   Â ARG
## 37        2015                   VerÃ³n, GonzaloGonzalo VerÃ³n   Â ARG
## 38        2015                     Valdez, NelsonNelson Valdez   Â PRY
## 39        2015     MartÃ-nez, Juan ManuelJuan Manuel MartÃ-nez   Â ARG
## 40        2015                     Drogba, DidierDidier Drogba   Â CIV
## 41        2016                     Dawkins, SimonSimon Dawkins   Â JAM
## 42        2016                   Movsisyan, YuraYura Movsisyan   Â ARM
## 43        2016                     Gruezo, CarlosCarlos Gruezo   Â ECU
## 44        2016                   Kouassi, XavierXavier Kouassi   Â CIV
## 45        2016                 Gashi, ShkÃ«lzenShkÃ«lzen Gashi   Â ALB
## 46        2016                           Kamara, KeiKei Kamara   Â SLE
## 47        2016               GonÃ§alves, JosÃ©JosÃ© GonÃ§alves   Â POR
##              Current club 2015 Guaranteed compensation [13]
## 1            LA Galaxy                        $4,500,000
## 2        Columbus Crew                        $1,175,000
## 3     Portland Timbers                          $550,000
## 4  Vancouver Whitecaps FC                       $325,000
## 5     Seattle Sounders FC                     $4,605,492
## 6            FC Dallas                          $442,400
## 7         Chicago Fire                        $1,144,922
## 8           Toronto FC                        $6,500,000
## 9    Philadelphia Union                         $768,750
## 10 Vancouver Whitecaps FC                     $1,410,900
## 11      New York City FC                      $5,610,000
## 12      Portland Timbers                      $1,000,000
## 13          Orlando City                      $7,167,500
## 14       Montreal Impact                        $400,000
## 15   Sporting Kansas City                       $683,250
## 16   Sporting Kansas City                       $682,102
## 17        Houston Dynamo                        $813,333
## 18   San Jose Earthquakes                       $240,000
## 19     New York Red Bulls                       $660,000
## 20             FC Dallas                        $160,000
## 21          Orlando City                        $279,500
## 22          Chicago Fire                        $720,938
## 23        Houston Dynamo                        $425,000
## 24 Vancouver Whitecaps FC                       $890,850
## 25             LA Galaxy                      $6,332,504
## 26      New York City FC                      $6,000,000
## 27            Toronto FC                      $4,750,000
## 28             Toronto FC                        $7,115,556
## 29            D.C. United                          $175,000
## 30   San Jose Earthquakes                        $1,040,000
## 31           Orlando City                           $60,000
## 32         Real Salt Lake                          $150,000
## 33        Colorado Rapids                        $1,170,000
## 34       New York City FC                        $2,315,694
## 35              LA Galaxy                        $4,100,008
## 36       Portland Timbers                          $799,992
## 37     New York Red Bulls                          $200,004
## 38    Seattle Sounders FC                        $1,215,000
## 39         Real Salt Lake                        $1,108,667
## 40        Montreal Impact                        $2,166,668
## 41   San Jose Earthquakes                              $n/a
## 42         Real Salt Lake                              $n/a
## 43              FC Dallas                              $n/a
## 44 New England Revolution                              $n/a
## 45        Colorado Rapids                              $n/a
## 46          Columbus Crew                              $n/a
## 47 New England Revolution                              $n/a
```

2. Clean the column with compensation information. Change the column type to numeric, and rename it Compensation. (Hint: The dollar sign is a wild character!)

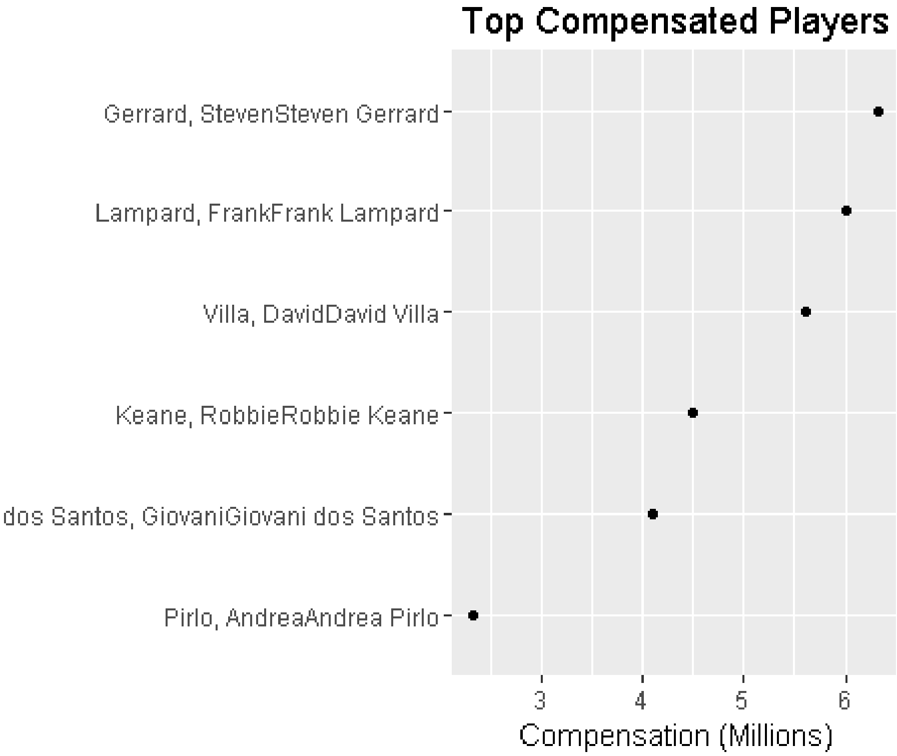
```
#### INSERT YOUR CODE HERE

glimpse(players)

## Observations: 47
## Variables: 5
## $ Year signed                       (int) 2011, 2012, 2013, 2014, 2013...
## $ Player                            (chr) "Keane, RobbieRobbie Keane",...
## $ Nation                            (chr) "Â IRL", "Â ARG", "Â ARG", "...
## $ Current club                      (chr) "LA Galaxy", "Columbus Crew"...
## $ 2015 Guaranteed compensation [13] (chr) "$4,500,000", "$1,175,000", ...

colnames(players)[5] = "Compensation"
players$Compensation =  str_replace_all(players$Compensation,"\\$","")
players$Compensation =  str_replace_all(players$Compensation,",","")
players$Compensation = as.numeric(players$Compensation)
C2Q2 = tbl_df(players)

# FinalAnswer:
C2Q2

## Source: local data frame [47 x 5]
##
##    Year signed                              Player Nation
##          (int)                               (chr)  (chr)
## 1         2011           Keane, RobbieRobbie Keane  Â IRL
## 2         2012 HiguaÃ-n, FedericoFederico HiguaÃ-n  Â ARG
## 3         2013           Valeri, DiegoDiego Valeri  Â ARG
## 4         2014           Laba, MatÃ-asMatÃ-as Laba  Â ARG
## 5         2013         Dempsey, ClintClint Dempsey  Â USA
## 6         2013               Diaz, MauroMauro Diaz  Â ARG
## 7         2014                  Gilberto, Gilberto  Â BRA
## 8         2014     Bradley, MichaelMichael Bradley  Â USA
## 9         2014             Edu, MauriceMaurice Edu  Â USA
## 10        2014         Morales, PedroPedro Morales  Â CHI
## ..         ...                                 ...    ...
## Variables not shown: Current club (chr), Compensation (dbl)
```

3. Create a subset of players called NYLAplayers, which only contains records of players currently play 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!)

```
#### INSERT YOUR CODE HERE

colnames(players)[4] = "CurrentClub"
NYLAplayers = players %>% filter(CurrentClub %in% c("LA Galaxy","New York
City FC")) %>% arrange(desc(Compensation))

# FinalAnswer:
NYLAplayers

##   Year signed                                Player Nation
## 1        2015         Gerrard, StevenSteven Gerrard  Â ENG
## 2        2015           Lampard, FrankFrank Lampard  Â ENG
## 3        2014               Villa, DavidDavid Villa  Â ESP
## 4        2011             Keane, RobbieRobbie Keane  Â IRL
## 5        2015 dos Santos, GiovaniGiovani dos Santos  Â MEX
## 6        2015             Pirlo, AndreaAndrea Pirlo  Â ITA
##        CurrentClub Compensation
## 1        LA Galaxy      6332504
## 2 New York City FC      6000000
## 3 New York City FC      5610000
## 4        LA Galaxy      4500000
## 5        LA Galaxy      4100008
## 6 New York City FC      2315694
```

4. Visualize the NYLAplayers compensation as follows:

```
#### INSERT YOUR CODE HERE

# FinalAnswer:
ggplot(NYLAplayers, aes(x = Compensation, y = reorder(Player,Compensation )))
+ geom_point()+
  xlab("Compensation (Millions)") +
  ylab("")+
  ggtitle("Top Compensated Players")+
  scale_x_continuous(breaks = seq(3000000,6000000,1000000), labels =c("3",
"4", "5", "6"))
```



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

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.

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

#### INSERT YOUR CODE HERE
```

```
str_locate_all(text, pattern = "\\$[0-9]*")

## [[1]]
##      start  end
## [1,]    59   63
## [2,]   196  200
## [3,]  1012 1017
## [4,]  1147 1151
## [5,]  1454 1458
## [6,]  2092 2097

Money = str_extract_all (text, pattern = "\\$[0-9]*")

MoneyNoDollarSign = as.numeric(str_replace(Money[[1]], "\\$",""))
total = sum(MoneyNoDollarSign)

# FinalAnswer:
total

## [1] 91760
```