Chaudhary_week_5b

Jyoti Chaudhary

October 30, 2016

PROBLEM 1

3. Consider the following three websites, Medal count from Rio Olympics http://www.cbssports.com/olympics/news/2016-rio-olympics-medal-tracker/ (http://www.cbssports.com/olympics/news/2016-rio-olympics-medal-tracker/) Three letter contry codes: https://en.wikipedia.org/wiki/List_of_IOC_country_codes (https://en.wikipedia.org/wiki/List_of_IOC_country_codes) (Note: I originally posted the page http://www.nationsonline.org/oneworld/country_code_list.htm (http://www.nationsonline.org/oneworld/country_code_list.htm) , but that has ISO codes, not IOC codes) Populations of countries: https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations) (https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations))

a. Please create a dataset with variables equal to the country name, the UN continental region, the population (2016), and the number of gold, silver, bronze, and total medals from the Rio Olympics.

```
# Extracting olympic medal table from the webpage

url1 <- "http://www.cbssports.com/olympics/news/2016-rio-olympics-medal-tracker/"

olympic_medal <- url1 %>% read_html() %>% html_nodes("table") %>%
  html_table(fill = TRUE)

olympic_medal <- data.frame(olympic_medal)

colnames(olympic_medal) <- c("COUNTRY", "GOLD", "SILVER", "BRONZE", "TOTAL")

# Extracting Country Codes table from webpage. 3 Tables extracted and combined in one

url2 <- "https://en.wikipedia.org/wiki/List_of_IOC_country_codes"

Country_code <- url2 %>% read_html() %>% html_nodes("table") %>% .[1:3] %>%
  html_table(fill = TRUE)


code_df1 <- data.frame(Country_code[[1]]$Code, Country_code[[1]]$`Nation (NOC)`)
code_df2 <- data.frame(Country_code[[2]]$Code, Country_code[[2]]$`Nation/Team`)
code_df3 <- data.frame(Country_code[[3]]$Code, Country_code[[3]]$`Nation (NOC)`)

country_code <- rbind(as.matrix(code_df1),as.matrix(code_df2),as.matrix(code_df3))
country_code <- data.frame(country_code)
colnames(country_code) <- c("CODE", "NATION")


# Extracting country population table from webpage.


url3 <- "https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)"

population.list <- url3 %>% read_html() %>% html_nodes("table")  %>% .[1:1] %>%
  html_table(fill = TRUE)

population.df <- population.list[[1]]

names(population.df) <- make.names(names(population.df))

# Join first 2 URL data on country code
df2_1 <- left_join(olympic_medal,country_code,by=c("COUNTRY" = "CODE"))



# Join df2_1 and 3rd URL data on country name and create a new data frame df_2_1_3 that has colu
mns for country name, the UN continental region, the population (2016), and the number of gold,
 silver, bronze, and total medals from the Rio Olympics.

population.df$Country.or.area <- str_trim(population.df$Country.or.area)
df2_1$NATION <- str_trim(df2_1$NATION)

df2_1_3 <- left_join(df2_1, population.df, by=c("NATION" = "Country.or.area")) %>%
```

```
          filter(!is.na(Population..1.July.2016..2.)) %>%
          select(NATION, UN.continental.region.1., Population..1.July.2016..2., GOLD, SILVER, B
RONZE, TOTAL)

colnames(df2_1_3) <- c("NATION", "REGION", "POP16", "GOLD", "SILVER", "BRONZE", "TOTAL")

head(df2_1_3)
```

```
##          NATION   REGION         POP16 GOLD SILVER BRONZE TOTAL
## 1 United States Americas   324,118,787   46     37     38   121
## 2         China     Asia 1,382,323,332   26     18     26    70
## 3        Russia   Europe   143,439,832   19     18     19    56
## 4       Germany   Europe    80,682,351   17     10     15    42
## 5         Japan     Asia   126,323,715   12      8     21    41
## 6        France   Europe    64,668,129   10     18     14    42
```

b. Grouping by region, please fit a linear regression of total medals on population and create a small dataset with only the variables of region and the R^2 from the regression.

```
lreg <- group_by(df2_1_3, REGION) %>%
        do (ftreg = lm(TOTAL ~ POP16, data = .))

reg_dataset = glance(lreg, ftreg)[c("REGION","r.squared")]

head(reg_dataset)
```

```
## Source: local data frame [5 x 2]
## Groups: REGION [5]
##
##     REGION r.squared
##      <chr>     <dbl>
## 1   Africa         1
## 2 Americas         1
## 3     Asia         1
## 4   Europe         1
## 5  Oceania         1
```

c. Create a long (or tall) version of the dataset in (a) with a variable indicating the type of medal and a variable showing the number of medals for the country of that type.

```
tall_dataset <- gather(df2_1_3, GOLD:TOTAL, key="medal_type", value = "count")
head(tall_dataset)
```

```
##          NATION   REGION         POP16 medal_type count
## 1 United States Americas   324,118,787       GOLD    46
## 2         China     Asia 1,382,323,332       GOLD    26
## 3        Russia   Europe   143,439,832       GOLD    19
## 4       Germany   Europe    80,682,351       GOLD    17
## 5         Japan     Asia   126,323,715       GOLD    12
## 6        France   Europe    64,668,129       GOLD    10
```