Zoo and Stringr

Kasra Pourang

6/18/2021

You are a Data Analyst working for the UN and you have been asked to look into the international rainfall data but unfortunately the data is messy and needs a lot of manipulating

```
library(dplyr)
library(zoo)
library(stringr)

data = read.csv("data.csv")
incorrect_data = read.csv("incorrect_data.csv")
head(data, n = 10)
```

1. Load both data sets

```
##           Country.or.Area             Year     Value                Unit
## 1                 Ireland             2008  98949.70 million cubic metres
## 2                 Algeria The year of 2007 100000.00 million cubic metres
## 3                 Algeria             2001  80000.00 million cubic metres
## 4                 Georgia             2015  72390.42 million cubic metres
## 5                  Israel             1990   6200.00 million cubic metres
## 6   Saint kitts and nevis             2007    138.80 million cubic metres
## 7                 Georgia             2013  82259.94 million cubic metres
## 8  Bosnia and herzegovina             2014  75610.38 million cubic metres
## 9                  Gambia             2002   6353.00 million cubic metres
## 10                 Panama             2005 228325.42 million cubic metres
```

```
head(incorrect_data, n = 10)
```

```
##    Country.or.Area Year      Value                 Unit
## 1        Mauritius 2012 15060.4783 million cubic metres
## 2        Mauritius 2011 16009.5079 million cubic metres
## 3        Mauritius 2010 17570.8825 million cubic metres
## 4        Mauritius 2009  2269.1284 million cubic metres
## 5        Mauritius 2008  8708.1784 million cubic metres
## 6        Mauritius 2007 10128.1771 million cubic metres
## 7        Mauritius 2006  7253.9663 million cubic metres
## 8        Mauritius 2005 15833.4878 million cubic metres
## 9        Mauritius 2004   637.6626 million cubic metres
## 10       Mauritius 2003  5669.7103 million cubic metres
```

```
str(data)
```

2. The data in ‘incorrect_data’ are false readings which are also in the main data set. Remove these entries from the main data set

```
## ’data.frame':    1618 obs. of  4 variables:
##  $ Country.or.Area: chr  "Ireland" "Algeria" "Algeria" "Georgia" ...
##  $ Year           : chr  "2008" "The year of 2007" "2001" "2015" ...
##  $ Value          : num  98950 100000 80000 72390 6200 ...
##  $ Unit           : chr  "million cubic metres" "million cubic metres" "million cubic metres" "milli
```

```
str(incorrect_data)
```

```
## ’data.frame':    42 obs. of  4 variables:
##  $ Country.or.Area: chr  "Mauritius" "Mauritius" "Mauritius" "Mauritius" ...
##  $ Year           : int  2012 2011 2010 2009 2008 2007 2006 2005 2004 2003 ...
##  $ Value          : num  15060 16010 17571 2269 8708 ...
##  $ Unit           : chr  "million cubic metres" "million cubic metres" "million cubic metres" "milli
```

```
incorrect_data$Year = as.character(incorrect_data$Year)
data$Year = as.character(data$Year)
data = data %>%
  anti_join(incorrect_data)
```

```
## Joining, by = c("Country.or.Area", "Year", "Value", "Unit")
```

```
head(data, n = 10)
```

```
##           Country.or.Area             Year     Value                Unit
## 1                 Ireland             2008  98949.70 million cubic metres
## 2                 Algeria The year of 2007 100000.00 million cubic metres
## 3                 Algeria             2001  80000.00 million cubic metres
## 4                 Georgia             2015  72390.42 million cubic metres
## 5                  Israel             1990   6200.00 million cubic metres
## 6   Saint kitts and nevis             2007    138.80 million cubic metres
## 7                 Georgia             2013  82259.94 million cubic metres
## 8  Bosnia and herzegovina             2014  75610.38 million cubic metres
## 9                  Gambia             2002   6353.00 million cubic metres
## 10                 Panama             2005 228325.42 million cubic metres
```

```
data = data %>%
  mutate(Country.or.Area = str_to_title(Country.or.Area))
head(data, n = 10)
```

3. Ensure all data is in title format with each word starting with a capital letter (e.g. ‘New Zealand’ not ‘New zealand’)

```
##           Country.or.Area             Year     Value                Unit
## 1                 Ireland             2008  98949.70 million cubic metres
## 2                 Algeria The year of 2007 100000.00 million cubic metres
## 3                 Algeria             2001  80000.00 million cubic metres
## 4                 Georgia             2015  72390.42 million cubic metres
## 5                  Israel             1990   6200.00 million cubic metres
## 6   Saint Kitts And Nevis             2007    138.80 million cubic metres
## 7                 Georgia             2013  82259.94 million cubic metres
## 8  Bosnia And Herzegovina             2014  75610.38 million cubic metres
## 9                  Gambia             2002   6353.00 million cubic metres
## 10                 Panama             2005 228325.42 million cubic metres
```

```
data = data %>%
  mutate(Country.or.Area = str_replace(Country.or.Area, 'Of', 'of')) %>%
  mutate(Country.or.Area = str_replace(Country.or.Area, 'And', 'and'))
head(data, n = 10)
```

4. Replace ‘Of’ with ‘of’ and ‘And’ with ‘and’

```
##           Country.or.Area             Year     Value                Unit
## 1                 Ireland             2008  98949.70 million cubic metres
## 2                 Algeria The year of 2007 100000.00 million cubic metres
## 3                 Algeria             2001  80000.00 million cubic metres
## 4                 Georgia             2015  72390.42 million cubic metres
## 5                  Israel             1990   6200.00 million cubic metres
## 6   Saint Kitts and Nevis             2007    138.80 million cubic metres
## 7                 Georgia             2013  82259.94 million cubic metres
## 8  Bosnia and Herzegovina             2014  75610.38 million cubic metres
## 9                  Gambia             2002   6353.00 million cubic metres
## 10                 Panama             2005 228325.42 million cubic metres
```

```
data = data %>%
  mutate(Year = str_extract(Year, '[0-9]+'))
head(data, n = 10)
```

5. Extract the year from the ‘Year’ column

```
##           Country.or.Area Year     Value                Unit
## 1                 Ireland 2008  98949.70 million cubic metres
## 2                 Algeria 2007 100000.00 million cubic metres
## 3                 Algeria 2001  80000.00 million cubic metres
## 4                 Georgia 2015  72390.42 million cubic metres
## 5                  Israel 1990   6200.00 million cubic metres
## 6   Saint Kitts and Nevis 2007    138.80 million cubic metres
## 7                 Georgia 2013  82259.94 million cubic metres
## 8  Bosnia and Herzegovina 2014  75610.38 million cubic metres
## 9                  Gambia 2002   6353.00 million cubic metres
## 10                 Panama 2005 228325.42 million cubic metres
```

```r
data$Year = as.numeric(data$Year)
head(data, n = 10)
```

6. Format the new year variable as numeric

```
##           Country.or.Area Year     Value                 Unit
## 1                 Ireland 2008  98949.70 million cubic metres
## 2                 Algeria 2007 100000.00 million cubic metres
## 3                 Algeria 2001  80000.00 million cubic metres
## 4                 Georgia 2015  72390.42 million cubic metres
## 5                  Israel 1990   6200.00 million cubic metres
## 6   Saint Kitts and Nevis 2007    138.80 million cubic metres
## 7                 Georgia 2013  82259.94 million cubic metres
## 8  Bosnia and Herzegovina 2014  75610.38 million cubic metres
## 9                  Gambia 2002   6353.00 million cubic metres
## 10                 Panama 2005 228325.42 million cubic metres
```

```r
data = data %>%
  arrange(Country.or.Area, Year)
head(data, n = 10)
```

7. Sort data by country and year

```
##    Country.or.Area Year Value                 Unit
## 1          Albania 1990 28385 million cubic metres
## 2          Albania 1995 40311 million cubic metres
## 3          Albania 1999 38284 million cubic metres
## 4          Albania 2000 30683 million cubic metres
## 5          Albania 2001 30491 million cubic metres
## 6          Albania 2002 35883 million cubic metres
## 7          Albania 2003 27893 million cubic metres
## 8          Albania 2004 42787 million cubic metres
## 9          Albania 2005 42840 million cubic metres
## 10         Albania 2006 32380 million cubic metres
```

```r
data = data %>%
  group_by(Country.or.Area) %>%
  mutate(avg_value_l3yr = rollmean(Value, 3, fill = NA, align = 'left'))
head(data, n = 10)
```

8. Add the 3 year rolling mean for each country

```
## # A tibble: 10 x 5
## # Groups:   Country.or.Area [1]
##    Country.or.Area  Year Value Unit                 avg_value_l3yr
##    <chr>           <dbl> <dbl> <chr>                         <dbl>
##   1 Albania          1990 28385 million cubic metres         35660
##   2 Albania          1995 40311 million cubic metres         36426
##   3 Albania          1999 38284 million cubic metres         33153.
##   4 Albania          2000 30683 million cubic metres         32352.
##   5 Albania          2001 30491 million cubic metres         31422.
##   6 Albania          2002 35883 million cubic metres         35521
##   7 Albania          2003 27893 million cubic metres         37840
##   8 Albania          2004 42787 million cubic metres         39336.
##   9 Albania          2005 42840 million cubic metres         35395.
##  10 Albania          2006 32380 million cubic metres         30934
```

```
head(data %>%
       filter(Year == 2012) %>%
       arrange(desc(avg_value_l3yr)))
```

9. Which country had the highest 3 year rolling mean rainfall in 2012?

```
## # A tibble: 6 x 5
## # Groups:   Country.or.Area [6]
##   Country.or.Area  Year    Value Unit                 avg_value_l3yr
##   <chr>           <dbl>    <dbl> <chr>                         <dbl>
## 1 China            2012 6515000  million cubic metres       6346433.
## 2 Indonesia        2012 4463718. million cubic metres       4688279.
## 3 Malaysia         2012  891220. million cubic metres        867027.
## 4 Paraguay         2012  569654. million cubic metres        605805.
## 5 Panama           2012  185971. million cubic metres        180888.
## 6 Iceland          2012  168420  million cubic metres        178652
```