Week-2-3-TASK-2.R

shadankhan

2024-08-13

```r
#loading the library dplyr
library(dplyr)
```

```
## 
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```

```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```r
#loading the library stringr
library(stringr)

#we have skipped the first 11 lines beacuse it includes the metadata
data<- read.csv('DATA.csv',skip=11)

#we use head to see some of the data
head(data)
```

```
##                                                  Statistic.Element      January
## 1          Mean maximum temperature (°C) for years 1970 to 2024         26.6
## 2               Highest temperature (°C) for years 1970 to 2024         46.0
## 3         Date of Highest temperature  °C  for years 1970 to 2024  25 Jan  2019
## 4           Lowest maximum temperature (°C) for years 1970 to 2024         13.9
## 5 Date of Lowest maximum temperature  °C  for years 1970 to 2024  05 Jan  1991
## 6        Decile 1 maximum temperature (°C) for years 1970 to 2022         19.7
##        February        March        April         May         June        July
## 1          26.6         24.3         20.4        16.7         13.7        13.2
## 2          46.8         40.8         34.5        27.0         21.8        22.7
## 3 07 Feb  2009 08 Mar  1991 10 Apr  2005 07 May  2002 08 Jun  2005 30 Jul  1975
## 4          13.5         12.7         11.7         8.0          6.2         5.7
## 5 02 Feb  2005 29 Mar  1973 07 Apr  1995 31 May  1977 19 Jun  1975 03 Jul  1984
## 6          19.8         18.2         15.5        13.3         11.1        10.6
##        August    September      October     November     December      Annual
## 1          14.5         16.8         19.5        22.1         24.6        19.9
## 2          25.6         30.2         36.0        41.6         44.6        46.8
```

```
## 3 29 Aug  1982 23 Sep  2017 12 Oct  2006 21 Nov  2019 20 Dec  2019 07 Feb  2009
## 4          6.5          8.2          10.4         11.6         13.0          5.7
## 5 16 Aug  1970 05 Sep  1995 16 Oct  1974 15 Nov  2006 04 Dec  1995 03 Jul  1984
## 6         11.5         12.8          14.4         15.9         18.2
##   Number.of.Years Start.Year End.Year
## 1              54       1970     2024
## 2              54       1970     2024
## 3             N/A       1970     2024
## 4              54       1970     2024
## 5             N/A       1970     2024
## 6              47       1970     2022
```

```r
# Convert 'Start Year' and 'End Year' columns to numeric after removing non-numeric characters
data <- data %>%
  mutate(
    Start.Year = as.numeric(gsub("[^0-9]", "", Start.Year)),
    End.Year = as.numeric(gsub("[^0-9]", "", End.Year))
  )

# Calculate 'Number of Years' if it is NA, using 'End Year' - 'Start Year'
data <- data %>%
  mutate(Number.of.Years = ifelse(Number.of.Years == "N/A", End.Year - Start.Year, Number.of.Years))

# Clean the 'Statistic Element' column by removing text after "for years"
data <- data %>%
  mutate(Statistic.Element = str_remove(Statistic.Element, "for years.*"))

# Create a metadata table containing 'Statistic Element', 'Start Year', 'End Year', 'Number of Years',
metadata <- data %>%
  select(Statistic.Element, Start.Year, End.Year, Number.of.Years, Annual)

# Remove metadata columns from the original data
data <- data %>%
  select(-Start.Year, -End.Year, -Number.of.Years, -Annual)

# We first separate the date data and keep all the data pertaining to dates together.
date_data <- data %>%
  filter(grepl("date", Statistic.Element, ignore.case = TRUE))

# we then remove date data from the original table for more cleaning.
# Remove date rows from the original table
data <- data %>%
  filter(!grepl("date", Statistic.Element, ignore.case = TRUE))

# now we extract rows in rainfall metrics
rainfall_data <- data %>%
  filter(grepl("rainfall|rain", Statistic.Element, ignore.case = TRUE))

# now we have to extract rows in sunshine and solar exposure metrics
sunshine_data <- data %>%
  filter(grepl("sunshine|solar", Statistic.Element, ignore.case = TRUE))

# now we extract The temperature data
temperature_data <- data %>%
```

```r
  filter(grepl("temperature|°C", Statistic.Element, ignore.case = TRUE))

# now we extract rows into other metrics like humidity, cloud cover, and evaporation
other_metrics_data <- data %>%
  filter(grepl("humidity|cloud|evaporation|dew", Statistic.Element, ignore.case = TRUE))

# now we extract rows into wind speed metrics
wind_data <- data %>%
  filter(grepl("wind|gust", Statistic.Element, ignore.case = TRUE))

# The Process for rainfall_data

rainfall_data <- rainfall_data %>% rename(Months = `Statistic.Element`) %>% t() %>%
  as.data.frame(stringsAsFactors = FALSE)
head(rainfall_data[,1:4])
```

```
##                     V1                  V2                  V3
## Months    Mean rainfall (mm)  Highest rainfall (mm)  Lowest rainfall (mm)
## January              43.2                106.4                1.6
## February             39.0                200.6                1.0
## March                37.1                142.2                2.4
## April                45.4                141.6                4.8
## May                  40.0                155.5                8.0
##                                  V4
## Months    Decile 1 monthly rainfall (mm)
## January              10.0
## February              4.9
## March                10.5
## April                13.3
## May                  13.1
```

```r
# The Process for other_metrics_data
other_metrics_data <- other_metrics_data %>% rename(Months = `Statistic.Element`) %>% t() %>%
  as.data.frame(stringsAsFactors = FALSE)
head(other_metrics_data[,1:4])
```

```
##                     V1                  V2
## Months    Mean number of cloudy days  Mean daily evaporation (mm)
## January              13.1                8.1
## February             10.3                7.1
## March                13.7                5.7
## April                14.8                3.8
## May                  17.5                2.5
##                                  V3                  V4
## Months    Mean 9am dew point temperature (°C)  Mean 9am relative humidity (%)
## January              10.7                65
## February             11.6                69
## March                10.5                70
## April                 8.8                72
## May                   7.7                79
```

```r
# The Process for sunshine_data
sunshine_data <- sunshine_data %>% rename(Months = `Statistic.Element`) %>% t() %>%
  as.data.frame(stringsAsFactors = FALSE)
head(sunshine_data[,1:2])
```

```
##                                     V1                                  V2
## Months    Mean daily sunshine (hours)  Mean daily solar exposure (MJ/(m*m))
## January                           8.7                                23.9
## February                          8.1                                20.9
## March                             7.2                                16.2
## April                             5.9                                11.2
## May                               4.6                                 7.8
```

```r
head(temperature_data)
```

```
##                  Statistic.Element January February March April  May June
## 1      Mean maximum temperature (°C)     26.6     26.6  24.3  20.4 16.7 13.7
## 2            Highest temperature (°C)     46.0     46.8  40.8  34.5 27.0 21.8
## 3    Lowest maximum temperature (°C)     13.9     13.5  12.7  11.7  8.0  6.2
## 4 Decile 1 maximum temperature (°C)     19.7     19.8  18.2  15.5 13.3 11.1
## 5 Decile 9 maximum temperature (°C)     35.6     34.8  31.5  26.1 20.8 16.3
## 6       Mean number of days >= 30 °C      8.7      8.3   5.0   0.4  0.0  0.0
##   July August September October November December
## 1 13.2   14.5      16.8    19.5     22.1     24.6
## 2 22.7   25.6      30.2    36.0     41.6     44.6
## 3  5.7    6.5       8.2    10.4     11.6     13.0
## 4 10.6   11.5      12.8    14.4     15.9     18.2
## 5 15.6   17.9      21.2    25.5     29.6     32.7
## 6  0.0    0.0       0.0     1.0      3.0      6.1
```

```r
#The Process for temperature data
temperature_data <- temperature_data %>% rename(Months = Statistic.Element)
temperature_data<-t(temperature_data)
temperature_data<-as.data.frame(temperature_data)
head(temperature_data[,1:4])
```

```
##                                     V1                               V2
## Months    Mean maximum temperature (°C)  Highest temperature (°C)
## January                           26.6                          46.0
## February                          26.6                          46.8
## March                             24.3                          40.8
## April                             20.4                          34.5
## May                               16.7                          27.0
##                                     V3                                    V4
## Months    Lowest maximum temperature (°C)  Decile 1 maximum temperature (°C)
## January                           13.9                                 19.7
## February                          13.5                                 19.8
## March                             12.7                                 18.2
## April                             11.7                                 15.5
## May                                8.0                                 13.3
```

```
#The Process for wind_data
wind_data <- wind_data %>%
  rename(Months = `Statistic.Element`) %>% t() %>%
  as.data.frame(stringsAsFactors = FALSE)
head(wind_data[,1:4])
```

```
##                                V1                              V2
## Months    Mean daily wind run (km)  Maximum wind gust speed (km/h)
## January                       454                             137
## February                      439                             122
## March                         426                             113
## April                         401                             107
## May                           432                             108
##                                V3                              V4
## Months    Mean 9am wind speed (km/h)  Mean 3pm wind speed (km/h)
## January                      18.5                            22.3
## February                     17.0                            21.2
## March                        16.9                            20.6
## April                        16.7                            19.9
## May                          17.2                            19.7
```