Extra Credit

Alice Ding

2023-02-09

Overview

```
weather_data <- read.csv("london_weather.csv")
glimpse(weather_data)
```

```
## Rows: 15,341
## Columns: 10
## $ date             <int> 19790101, 19790102, 19790103, 19790104, 19790105, 197~
## $ cloud_cover      <dbl> 2, 6, 5, 8, 6, 5, 8, 8, 4, 7, 1, 3, 1, 7, NA, 8, 8, 8~
## $ sunshine         <dbl> 7.0, 1.7, 0.0, 0.0, 2.0, 3.8, 0.0, 0.1, 5.8, 1.9, 6.8~
## $ global_radiation <dbl> 52, 27, 13, 13, 29, 39, 13, 15, 50, 30, 55, 54, 57, 1~
## $ max_temp         <dbl> 2.3, 1.6, 1.3, -0.3, 5.6, 8.3, 8.5, 5.8, 5.2, 4.9, 2.~
## $ mean_temp        <dbl> -4.1, -2.6, -2.8, -2.6, -0.8, -0.5, 1.5, 6.9, 3.7, 3.~
## $ min_temp         <dbl> -7.5, -7.5, -7.2, -6.5, -1.4, -6.6, -5.3, 5.3, 1.6, 1~
## $ precipitation    <dbl> 0.4, 0.0, 0.0, 0.0, 0.0, 0.7, 5.2, 0.8, 7.2, 2.1, 2.3~
## $ pressure         <dbl> 101900, 102530, 102050, 100840, 102250, 102780, 10252~
## $ snow_depth       <dbl> 9, 8, 4, 2, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0,~
```

The dataset I've chosen to use for this assignment is London weather data taken from Kaggle: https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data

The description from the source:

`The dataset featured below was created by reconciling measurements from requests of individual weather attributes provided by the European Climate Assessment (ECA). The measurements of this particular dataset were recorded by a weather station near Heathrow airport in London, UK.`

The dataset seems to go from 1979 all the way to 2020 at a daily cadence and has values for cloud cover, sunshine, radiation, temperature, precipitation, pressure, and snow depth. Here are the definitions for each field:

- date: recorded date of measurement - (int)
- cloud_cover: cloud cover measurement in oktas - (float)
- sunshine: sunshine measurement in hours (hrs) - (float)
- global_radiation: irradiance measurement in Watt per square meter (W/m2) - (float)
- max_temp: maximum temperature recorded in degrees Celsius (°C) - (float)
- mean_temp: mean temperature in degrees Celsius (°C) - (float)
- min_temp: minimum temperature recorded in degrees Celsius (°C) - (float)
- precipitation: precipitation measurement in millimeters (mm) - (float)
- pressure: pressure measurement in Pascals (Pa) - (float)
- snow_depth: snow depth measurement in centimeters (cm) - (float)

Cleaning Up the Data

Upon first glance, I'd like to reformat the `date` column so it's more readable.

```
weather_data$date <- anydate(weather_data$date)
glimpse(weather_data)
```

```
## Rows: 15,341
## Columns: 10
## $ date             <date> 1979-01-01, 1979-01-02, 1979-01-03, 1979-01-04, 1979~
## $ cloud_cover      <dbl> 2, 6, 5, 8, 6, 5, 8, 8, 4, 7, 1, 3, 1, 7, NA, 8, 8, 8~
## $ sunshine         <dbl> 7.0, 1.7, 0.0, 0.0, 2.0, 3.8, 0.0, 0.1, 5.8, 1.9, 6.8~
## $ global_radiation <dbl> 52, 27, 13, 13, 29, 39, 13, 15, 50, 30, 55, 54, 57, 1~
## $ max_temp         <dbl> 2.3, 1.6, 1.3, -0.3, 5.6, 8.3, 8.5, 5.8, 5.2, 4.9, 2.~
## $ mean_temp        <dbl> -4.1, -2.6, -2.8, -2.6, -0.8, -0.5, 1.5, 6.9, 3.7, 3.~
## $ min_temp         <dbl> -7.5, -7.5, -7.2, -6.5, -1.4, -6.6, -5.3, 5.3, 1.6, 1~
## $ precipitation    <dbl> 0.4, 0.0, 0.0, 0.0, 0.0, 0.7, 5.2, 0.8, 7.2, 2.1, 2.3~
## $ pressure         <dbl> 101900, 102530, 102050, 100840, 102250, 102780, 10252~
## $ snow_depth       <dbl> 9, 8, 4, 2, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0,~
```

Looks good!

Window Functions

I'll be using the `global_radiation` and `mean_temp` columns for the window function calculations.

Year to Date

To add a YTD column here, I'll use `dplyr`.

```
## YTD for mean_temp
ytd_mean_temp <- weather_data |>
  mutate(year = year(date), day = date(date)) |>
  group_by(year) |>
  summarise(ytd_mean_temp = cummean(mean_temp)) |>
  ungroup()
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
weather_data$ytd_mean_temp <- ytd_mean_temp$ytd_mean_temp

## YTD for global_radiation
ytd_global_radiation <- weather_data |>
  mutate(year = year(date), day = date(date)) |>
  group_by(year) |>
  summarise(ytd_global_radiation = cummean(global_radiation)) |>
  ungroup()
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
weather_data$ytd_global_radiation <- ytd_global_radiation$ytd_global_radiation

## snapshot of data
filter(weather_data
       , (yday(date) > 363 | yday(date) < 8)
       & (year(date) == 1979 | year(date) == 1980))  |>
  select(date
         , mean_temp
         , ytd_mean_temp
         , global_radiation
         , ytd_global_radiation
         )
```

```
##          date mean_temp ytd_mean_temp global_radiation ytd_global_radiation
## 1  1979-01-01      -4.1    -4.1000000               52             52.00000
## 2  1979-01-02      -2.6    -3.3500000               27             39.50000
## 3  1979-01-03      -2.8    -3.1666667               13             30.66667
## 4  1979-01-04      -2.6    -3.0250000               13             26.25000
## 5  1979-01-05      -0.8    -2.5800000               29             26.80000
## 6  1979-01-06      -0.5    -2.2333333               39             28.83333
## 7  1979-01-07       1.5    -1.7000000               13             26.57143
## 8  1979-12-30       1.8    10.0101648               29            112.77747
## 9  1979-12-31       1.4     9.9865753               45            112.59178
## 10 1980-01-01      -1.8    -1.8000000               14             14.00000
## 11 1980-01-02      -1.8    -1.8000000               51             32.50000
## 12 1980-01-03       2.6    -0.3333333               13             26.00000
## 13 1980-01-04       5.2     1.0500000               39             29.25000
## 14 1980-01-05       4.8     1.8000000               23             28.00000
## 15 1980-01-06       5.8     2.4666667               36             29.33333
## 16 1980-01-07       5.2     2.8571429               21             28.14286
## 17 1980-12-29       6.2    10.3774725               12            115.26374
## 18 1980-12-30       9.0    10.3736986               13            114.98356
## 19 1980-12-31       9.2    10.3704918               13            114.70492
```

After checking each of these calculations, it looks like both `ytd` columns seem to be taking a cumulative mean for each day and all the days before it and then the column resets every year. We can see this by looking at how the first 4 days of the year average day-by-day, then how 2000-01-01 resets the mean back to the first day when it hits a new year.

Rolling 6 Day Average

I'll also be using `dplyr` for this and I'll also assume 6 day moving average means the average of the past 5 days and the current day. I also won't group this by year as we just want to compare a smaller period of time independent of date aggregation. This means that the only rows with less than 6 days worth of data will be the first 6 days in the dataset and the last 6 as well.

```
## Rolling 6 day average for mean_temp
rolling_mean_temp <- weather_data |>
  mutate(temp1 = lag(mean_temp, 0),
         temp2 = lag(mean_temp, 1),
         temp3 = lag(mean_temp, 2),
         temp4 = lag(mean_temp, 3),
```

```
        temp5 = lag(mean_temp, 4),
        temp6 = lag(mean_temp, 5)) %>%
  summarise(
    rolling_mean_temp = rowMeans(
      cbind(temp1, temp2, temp3, temp4, temp5, temp6
            ), na.rm = TRUE)
    )

weather_data$rolling_average_mean_temp <- rolling_mean_temp$rolling_mean_temp

## Rolling 6 day average for global_radiation
rolling_global_radiation <- weather_data |>
  mutate(temp1 = lag(global_radiation, 0),
        temp2 = lag(global_radiation, 1),
        temp3 = lag(global_radiation, 2),
        temp4 = lag(global_radiation, 3),
        temp5 = lag(global_radiation, 4),
        temp6 = lag(global_radiation, 5)) %>%
  summarise(
    rolling_global_radiation = rowMeans(
      cbind(temp1, temp2, temp3, temp4, temp5, temp6
            ), na.rm = TRUE)
    )

weather_data$rolling_average_global_radiation <- rolling_global_radiation$rolling_global_radiation

## snapshot of data
filter(weather_data
       , (yday(date) > 363 | yday(date) < 8)
       & (year(date) == 1979 | year(date) == 1980)) |>
  select(date
         , mean_temp
         , rolling_average_mean_temp
         , ytd_mean_temp
         , global_radiation
         , rolling_average_global_radiation
         , ytd_global_radiation
         )
```

```
##          date mean_temp rolling_average_mean_temp ytd_mean_temp
## 1  1979-01-01      -4.1                -4.100000    -4.1000000
## 2  1979-01-02      -2.6                -3.350000    -3.3500000
## 3  1979-01-03      -2.8                -3.166667    -3.1666667
## 4  1979-01-04      -2.6                -3.025000    -3.0250000
## 5  1979-01-05      -0.8                -2.580000    -2.5800000
## 6  1979-01-06      -0.5                -2.233333    -2.2333333
## 7  1979-01-07       1.5                -1.300000    -1.7000000
## 8  1979-12-30       1.8                 3.166667    10.0101648
## 9  1979-12-31       1.4                 3.333333     9.9865753
## 10 1980-01-01      -1.8                 2.733333    -1.8000000
## 11 1980-01-02      -1.8                 1.283333    -1.8000000
## 12 1980-01-03       2.6                 1.033333    -0.3333333
## 13 1980-01-04       5.2                 1.233333     1.0500000
```

```
## 14 1980-01-05       4.8                 1.733333     1.8000000
## 15 1980-01-06       5.8                 2.466667     2.4666667
## 16 1980-01-07       5.2                 3.633333     2.8571429
## 17 1980-12-29       6.2                 5.583333    10.3774725
## 18 1980-12-30       9.0                 5.250000    10.3736986
## 19 1980-12-31       9.2                 5.483333    10.3704918
##    global_radiation rolling_average_global_radiation ytd_global_radiation
## 1                52                         52.00000             52.00000
## 2                27                         39.50000             39.50000
## 3                13                         30.66667             30.66667
## 4                13                         26.25000             26.25000
## 5                29                         26.80000             26.80000
## 6                39                         28.83333             28.83333
## 7                13                         22.33333             26.57143
## 8                29                         29.33333            112.77747
## 9                45                         29.50000            112.59178
## 10               14                         28.83333             14.00000
## 11               51                         35.33333             32.50000
## 12               13                         32.83333             26.00000
## 13               39                         31.83333             29.25000
## 14               23                         30.83333             28.00000
## 15               36                         29.33333             29.33333
## 16               21                         30.50000             28.14286
## 17               12                         27.66667            115.26374
## 18               13                         27.83333            114.98356
## 19               13                         23.33333            114.70492
```

Just as a gut check, comparing these to the YTD calcs makes sense – the first 6 days match up perfectly, but after that it begins to diverge as the 6 day rolling average drops days. We also see how this doesn't reset after the year ends – it keeps going!

Conclusion

In terms of usefulness, the rolling average numbers likely are more helpful for temperature considering London is a city that experiences multiple seasons and ranges of temperature, therefore a year to date average doesn't really tell you what it's actually like to live there or what the temperature is on any day. A more useful approach would likely be a quarter to date average as seasons typically follow a quarterly schedule, at least for places like London (Q1 = Spring, Q2 = Summer, etc.).

The radiation numbers are interesting as this value seems to fluctuate a lot more from day-to-day. Upon further research, it seems that irradiance "is a measurement of solar power and is defined as the rate at which solar energy falls onto a surface." (click for source). I would say that the rolling average numbers are likely more helpful here as well than the year to date average given how much it differs and also the strength of the sun changes due to the intensity of the sun changing depending on whether it's summer vs. winter.

Further analysis could include looking at the `sunshine` field and seeing if there's a correlation between this and `global_radiation`, as well as just seeing how the different numbers relate to one another such as `sunshine` and `mean_temp`. Just based off of intuition, I would suspect that a higher `sunshine` value would mean higher `mean_temp`.

Overall, I think I've just scratched the surface with a dataset like this and it could also be interesting to see it replicated for other cities or regions around the world. If the data were to back even further too, we could see how much climate change is affecting each season to verify if it's actually getting warmer over time.