Appendix II - Process the Data

Jamie Cash

2023-07-05

Load the prepared data

Weather station data provided by the MET office is provided on their website in txt files. This data was prepared to create a clean dataset which is available here. Load this dataset.

```
data <- read.csv('met_station_readings.csv', colClasses =
                   c("character", "numeric", "numeric", "Date", "numeric", "numeric", "numeric"))
head(data)
```

```
##   station_name station_long station_lat year_month min_temp max_temp rainfall
## 1    Aberporth     -4.56999    52.13914 1941-01-01       NA       NA     74.7
## 2    Aberporth     -4.56999    52.13914 1941-02-01       NA       NA     69.1
## 3    Aberporth     -4.56999    52.13914 1941-03-01       NA       NA     76.2
## 4    Aberporth     -4.56999    52.13914 1941-04-01       NA       NA     33.7
## 5    Aberporth     -4.56999    52.13914 1941-05-01       NA       NA     51.3
## 6    Aberporth     -4.56999    52.13914 1941-06-01       NA       NA     25.7
```

Process the data

Seasonal temperatures by year

Summarise the data to create a dataset containing seasonal high, low, average max and average min temperature across all weather stations by year.

```
# Add season and year columns
data <- data %>%
  mutate(season =
    case_when(
      month(year_month) %in% c(12, 1, 2) ~ 'Winter',
      month(year_month) %in% c(3, 4, 5) ~ 'Spring',
      month(year_month) %in% c(6, 7, 8) ~ 'Summer',
      month(year_month) %in% c(9, 10, 11) ~ 'Autumn'
    )
  )

data$year = year(data$year_month)

# Group data by season and compute summary columns
seasonal <- data %>%
```

```
  group_by(year, season) %>%
  summarise(high = max(max_temp), low = min(min_temp),
            average_max = mean(max_temp, na.rm=TRUE),
            average_min = mean(min_temp, na.rm=TRUE), .groups='keep') %>%
  arrange(year, season)

# Drop NA rows
seasonal <- na.omit(seasonal)

# Save
write.csv(seasonal, "seasonal_summary.csv", row.names=FALSE)
head(seasonal)
```

```
## # A tibble: 6 x 6
## # Groups:   year, season [6]
##    year season  high   low average_max average_min
##   <dbl> <chr>  <dbl> <dbl>       <dbl>       <dbl>
## 1  1865 Autumn  22.6   3.1        15.6         7.5
## 2  1865 Summer  22.3  10.8        21.0        11.7
## 3  1866 Autumn  17     3.7        13.8         7.3
## 4  1866 Spring  15.7   1.7        12.3         4.13
## 5  1866 Summer  21.9  10.5        20.2        11.4
## 6  1866 Winter   9.9   0.9         8.82        2.87
```

Annual temperatures by station

Summarise the data to create a dataset containing annual high, low and total rainfall by station by year.

```
# Group data by year and station and compute summary columns
annual_station <- data %>%
  group_by(year, station_name, station_long, station_lat) %>%
  summarise(high = max(max_temp), low = min(min_temp),
            total_rainfall = sum(rainfall, na.rm=TRUE), .groups='keep') %>%
  arrange(year, station_name)

head(annual_station)
```

```
## # A tibble: 6 x 7
## # Groups:   year, station_name, station_long, station_lat [6]
##    year station_name station_long station_lat  high   low total_rainfall
##   <dbl> <chr>               <dbl>       <dbl> <dbl> <dbl>          <dbl>
## 1  1853 Armagh              -6.65        54.4  NA    NA             636.
## 2  1853 Oxford              -1.26        51.8  21.2  -1.8           692
## 3  1854 Armagh              -6.65        54.4  NA    NA             837.
## 4  1854 Oxford              -1.26        51.8  21.7   0.6           450.
## 5  1855 Armagh              -6.65        54.4  NA    NA             603.
## 6  1855 Oxford              -1.26        51.8  22.8  -4.5           640.
```

Different stations have different data start dates. Some also closed prior to 2023. We need to make these consistent as to not affect the mean calculations and remove stations that closed prior to 2023.

Find the latest first year for all stations.

```
first_years <- annual_station %>%
  group_by(station_name) %>%
  summarise(first_year = min(year)) %>%
  arrange(station_name)

max(first_years$first_year)
```

```
## [1] 1978
```

Remove rows before 1978

```
annual_station <- annual_station %>%
  filter(year >= 1978)

head(annual_station)
```

```
## # A tibble: 6 x 7
## # Groups:   year, station_name, station_long, station_lat [6]
##    year station_name        station_long station_lat  high   low total_rainfall
##   <dbl> <chr>                      <dbl>       <dbl> <dbl> <dbl>          <dbl>
## 1  1978 Aberporth                  -4.57        52.1  16.3   1.5           735.
## 2  1978 Armagh                     -6.65        54.4  18.1   0.4           774.
## 3  1978 Ballypatrick Forest        -6.15        55.2  15.6   0.5             0
## 4  1978 Bradford                   -1.77        53.8  17.7  -0.9           870.
## 5  1978 Braemar                    -3.40        57.0  16.6  -8.4           924.
## 6  1978 Camborne                   -5.33        50.2  17.5   5             380.
```

Get stations that closed prior ro 2023.

```
last_years <- annual_station %>%
  group_by(station_name) %>%
  summarise(last_year = max(year)) %>%
  arrange(station_name)

last_years %>%
  filter(last_year < 2023)
```

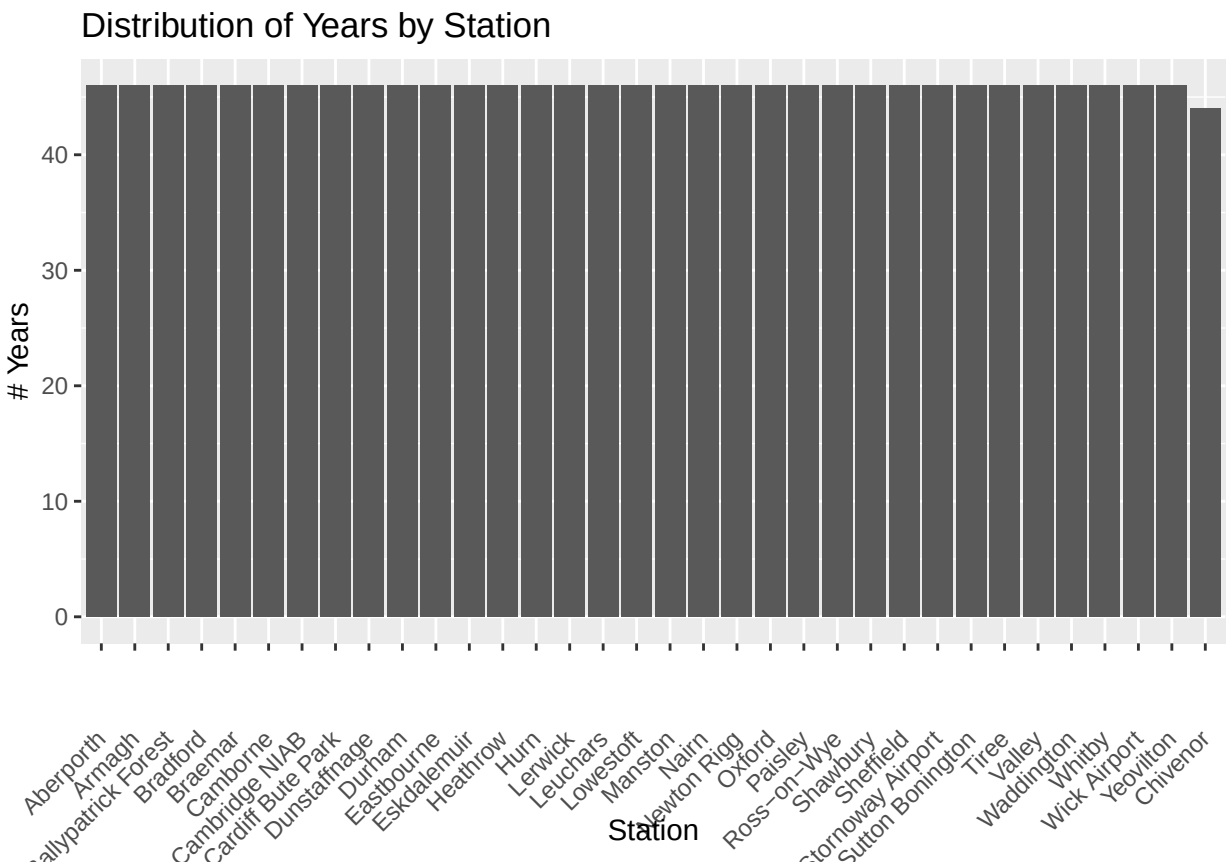
```
## # A tibble: 3 x 2
##   station_name last_year
##   <chr>            <dbl>
## 1 Cwmystwyth        2011
## 2 Ringway           2004
## 3 Southampton       2000
```

Remove them

```
annual_station <- annual_station %>%
  filter(!station_name %in% c('Cwmystwyth', 'Ringway', 'Southampton'))
```

We should now have a consistent number of years for every station. Check.

```
ggplot(data = annual_station) +
  geom_bar(mapping = aes(x = station_name)) +
  labs(title="Distribution of Years by Station",
       y = "# Years") +
  scale_x_discrete(name ="Station", limits=unique(annual_station$station_name)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```



We have an inconsistent number of years Chivenor. Check year.

```
annual_station %>%
  filter(station_name == 'Chivenor')
```

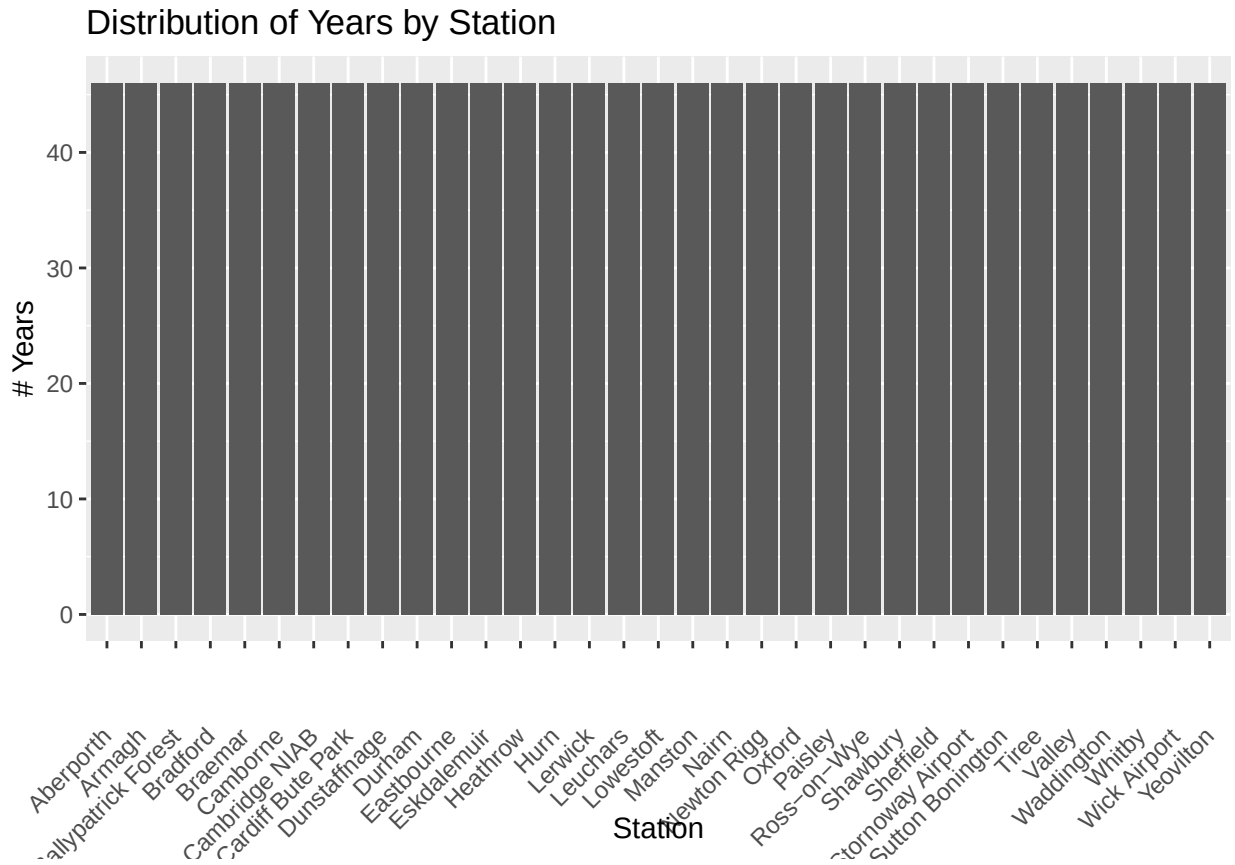
```
## # A tibble: 44 x 7
## # Groups:   year, station_name, station_long, station_lat [44]
##     year station_name station_long station_lat  high   low total_rainfall
##    <dbl> <chr>               <dbl>       <dbl> <dbl> <dbl>          <dbl>
##  1  1980 Chivenor            -4.15        51.1  NA    NA             446
##  2  1981 Chivenor            -4.15        51.1  NA    NA             954.
##  3  1982 Chivenor            -4.15        51.1  NA    NA             971.
##  4  1983 Chivenor            -4.15        51.1  NA    NA             833.
##  5  1984 Chivenor            -4.15        51.1  NA    NA             878.
##  6  1985 Chivenor            -4.15        51.1  NA    NA             863.
##  7  1986 Chivenor            -4.15        51.1  NA    NA             924.
##  8  1987 Chivenor            -4.15        51.1  20.4   0.3           808.
##  9  1988 Chivenor            -4.15        51.1  19.6   3.2           964.
```

```
## 10  1989 Chivenor            -4.15        51.1  23.1   2.8           863.
## # i 34 more rows
```

Chivenor doesnt have any data for 1978 and 1979, and has no temp readings for years between 1980 and 1986. Remove data for this station and reinspect distribution.

```
annual_station <- annual_station %>%
  filter(station_name != 'Chivenor')

ggplot(data = annual_station) +
  geom_bar(mapping = aes(x = station_name)) +
  labs(title="Distribution of Years by Station",
       y = "# Years") +
  scale_x_discrete(name ="Station", limits=unique(annual_station$station_name)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```



Looks good. Save the data.

```
# Save
write.csv(annual_station, "annual_station_summary.csv", row.names=FALSE)
head(annual_station)
```

```
## # A tibble: 6 x 7
## # Groups:   year, station_name, station_long, station_lat [6]
##    year station_name        station_long station_lat  high   low total_rainfall
```

```
##   <dbl> <chr>                    <dbl>       <dbl> <dbl> <dbl>          <dbl>
## 1  1978 Aberporth                -4.57        52.1  16.3   1.5           735.
## 2  1978 Armagh                   -6.65        54.4  18.1   0.4           774.
## 3  1978 Ballypatrick Forest      -6.15        55.2  15.6   0.5             0
## 4  1978 Bradford                 -1.77        53.8  17.7  -0.9           870.
## 5  1978 Braemar                  -3.40        57.0  16.6  -8.4           924.
## 6  1978 Camborne                 -5.33        50.2  17.5   5             380.
```