Assignment 6 - CT5102 Transforming data with dplyr

Chin Zhe Jing 22221970

zhejing

2022-10-20

```
library(aimsir17)
library(ggplot2)
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
```

a new tibble that is a summary of the total rainfall and the average temperature for each day (and for each station)

```
s_data <- observations |>
  group_by(station, day, month) |>
  summarize(TotalRain = sum(rain, na.rm=T),
            AvrTemp = mean(temp, na.rm=T))
```

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

```
s_data
```

```
## # A tibble: 9,125 x 5
## # Groups:   station, day [775]
##    station   day month TotalRain AvrTemp
##    <chr>   <int> <dbl>     <dbl>   <dbl>
##  1 ATHENRY     1     1       0.2    3.51
##  2 ATHENRY     1     2       1      6.25
##  3 ATHENRY     1     3       7.6    4.52
##  4 ATHENRY     1     4       0.3    8.35
##  5 ATHENRY     1     5       0      9.64
```

```
##  6 ATHENRY     1     6       4.4   13.4
##  7 ATHENRY     1     7       0     13.4
##  8 ATHENRY     1     8       0.3   14.1
##  9 ATHENRY     1     9       0.1   10.9
## 10 ATHENRY     1    10       5.2   13.3
## # ... with 9,115 more rows
```

```
glimpse(s_data)
```

```
## Rows: 9,125
## Columns: 5
## Groups: station, day [775]
## $ station   <chr> "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "ATHE~
## $ day       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ month     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5, 6, 7, ~
## $ TotalRain <dbl> 0.2, 1.0, 7.6, 0.3, 0.0, 4.4, 0.0, 0.3, 0.1, 5.2, 0.3, 0.0, ~
## $ AvrTemp   <dbl> 3.5125000, 6.2458333, 4.5208333, 8.3458333, 9.6375000, 13.37~
```

to calculate the daily changes in temperature and rainfall, for each station

```
s_data_diff <- ungroup(s_data) |>
  arrange(station, month, day) |>
  group_by(station) |>
  mutate(RainDiff = TotalRain - lag(TotalRain),
         AbsRainDiff = abs(TotalRain - lag(TotalRain)),
         MeanTempDiff = AvrTemp - lag(AvrTemp),
         AbsMeanTempDiff = abs(AvrTemp - lag(AvrTemp)))

s_data_diff
```

```
## # A tibble: 9,125 x 9
## # Groups:   station [25]
##    station   day month TotalRain AvrTemp RainDiff AbsRainDiff MeanTemp~1 AbsMe~2
##    <chr>   <int> <dbl>     <dbl>   <dbl>    <dbl>       <dbl>      <dbl>   <dbl>
##  1 ATHENRY     1     1       0.2   3.51    NA          NA         NA      NA
##  2 ATHENRY     2     1       0     0.679   -0.2         0.2       -2.83    2.83
##  3 ATHENRY     3     1       0     3.75     0           0          3.07    3.07
##  4 ATHENRY     4     1       0     5.13     0           0          1.38    1.38
##  5 ATHENRY     5     1       0.1   6.85     0.1         0.1        1.71    1.71
##  6 ATHENRY     6     1      18    10.0     17.9        17.9        3.18    3.18
##  7 ATHENRY     7     1       1.4   9.28   -16.6        16.6       -0.746   0.746
##  8 ATHENRY     8     1       1.2   9.76    -0.200       0.200      0.475   0.475
##  9 ATHENRY     9     1       5.4   6.99     4.2         4.2       -2.77    2.77
## 10 ATHENRY    10     1       0.7   9.15    -4.7         4.7        2.16    2.16
## # ... with 9,115 more rows, and abbreviated variable names 1: MeanTempDiff,
## #   2: AbsMeanTempDiff
```

```
glimpse(s_data_diff)
```

```
## Rows: 9,125
## Columns: 9
## Groups: station [25]
```

```
## $ station         <chr> "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY",~
## $ day             <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ month           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TotalRain       <dbl> 0.2, 0.0, 0.0, 0.0, 0.1, 18.0, 1.4, 1.2, 5.4, 0.7, 0.8~
## $ AvrTemp         <dbl> 3.5125000, 0.6791667, 3.7500000, 5.1333333, 6.8458333,~
## $ RainDiff        <dbl> NA, -0.2, 0.0, 0.0, 0.1, 17.9, -16.6, -0.2, 4.2, -4.7,~
## $ AbsRainDiff     <dbl> NA, 0.2, 0.0, 0.0, 0.1, 17.9, 16.6, 0.2, 4.2, 4.7, 0.1~
## $ MeanTempDiff    <dbl> NA, -2.8333333, 3.0708333, 1.3833333, 1.7125000, 3.183~
## $ AbsMeanTempDiff <dbl> NA, 2.8333333, 3.0708333, 1.3833333, 1.7125000, 3.1833~
```

```
print(n=30, s_data_diff |> filter(station == "ATHENRY", month == 5, day == 31))
```

```
## # A tibble: 1 x 9
## # Groups:   station [1]
##   station   day month TotalRain AvrTemp RainDiff AbsRainDiff MeanTempD~1 AbsMe~2
##   <chr>   <int> <dbl>     <dbl>   <dbl>    <dbl>       <dbl>       <dbl>   <dbl>
## 1 ATHENRY    31     5         0    13.8     -0.8         0.8       0.325   0.325
## # ... with abbreviated variable names 1: MeanTempDiff, 2: AbsMeanTempDiff
```

```
arrange(s_data_diff,desc(AbsRainDiff)) |> slice(1:5)
```

```
## # A tibble: 125 x 9
## # Groups:   station [25]
##    station      day month TotalRain AvrTemp RainDiff AbsRainDiff MeanT~1 AbsMe~2
##    <chr>      <int> <dbl>     <dbl>   <dbl>    <dbl>       <dbl>   <dbl>   <dbl>
##  1 ATHENRY       22     7       0   15.1       -41.7        41.7  3.10    3.10
##  2 ATHENRY       21     7      41.7 12.0        33.4        33.4  0.175   0.175
##  3 ATHENRY       24    12      31.2 11.1        30.5        30.5  0.521   0.521
##  4 ATHENRY       10    12      24.7  0.0375     24.7        24.7 -2.02    2.02
##  5 ATHENRY       11    12       0.4  1.18      -24.3        24.3  1.14    1.14
##  6 BALLYHAISE    19     7      22.5 15.2        22.5        22.5 -2.95    2.95
##  7 BALLYHAISE    27     9      21   12.6        21          21    0.0667  0.0667
##  8 BALLYHAISE    22    11      21.8  6.20       20.7        20.7 -5.83    5.83
##  9 BALLYHAISE    20     9      20.5 12.5        20.5        20.5  0.567   0.567
## 10 BALLYHAISE    23    11       1.9  3.22      -19.9        19.9 -2.98    2.98
## # ... with 115 more rows, and abbreviated variable names 1: MeanTempDiff,
## #   2: AbsMeanTempDiff
```

```
arrange(s_data_diff,desc(AbsMeanTempDiff)) |> slice(1:5)
```

```
## # A tibble: 125 x 9
## # Groups:   station [25]
##    station      day month TotalRain AvrTemp RainDiff AbsRainDiff MeanT~1 AbsMe~2
##    <chr>      <int> <dbl>     <dbl>   <dbl>    <dbl>       <dbl>   <dbl>   <dbl>
##  1 ATHENRY       19    12       0.2   11.0       0.2         0.2    6.57    6.57
##  2 ATHENRY        7    12      10.5    5.25      1           1     -6.05    6.05
##  3 ATHENRY        3     9       2.1   17.0     -14.4        14.4    5.81    5.81
##  4 ATHENRY       22    11      19.8    6.48     19.4        19.4   -5.65    5.65
##  5 ATHENRY       25    12       9      5.48    -22.2        22.2   -5.62    5.62
##  6 BALLYHAISE    20    11       6.2   11.6       2.8         2.8    6.72    6.72
##  7 BALLYHAISE     6     1       3.7    9.69      3           3      6.38    6.38
##  8 BALLYHAISE    22    11      21.8    6.20     20.7        20.7   -5.83    5.83
```

```
##  9 BALLYHAISE    24     1       0.8    9.33      0.8         0.8    5.65    5.65
## 10 BALLYHAISE     7    12       8.6    5.02      4.5         4.5   -5.61    5.61
## # ... with 115 more rows, and abbreviated variable names 1: MeanTempDiff,
## #   2: AbsMeanTempDiff
```

a new output tibble out which generates the following monthly summaries for each weather station (average, standard deviation, minumim and maxiumm).

```
out <- s_data_diff |>
  group_by(station, month) |>
  summarize(AvrDiffTemp = mean(MeanTempDiff, na.rm=T),
            SDDiffTemp = sd(MeanTempDiff, na.rm=T),
            MinDiffTemp = min(MeanTempDiff, na.rm=T),
            MaxiffTemp = max(MeanTempDiff, na.rm=T),
            AvrDiffRain = mean(RainDiff, na.rm=T),
            SDDiffRain = sd(RainDiff, na.rm=T),
            MinDiffRain = min(RainDiff, na.rm=T),
            MaxDiffRain = max(RainDiff, na.rm=T))
```

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

```
out
```

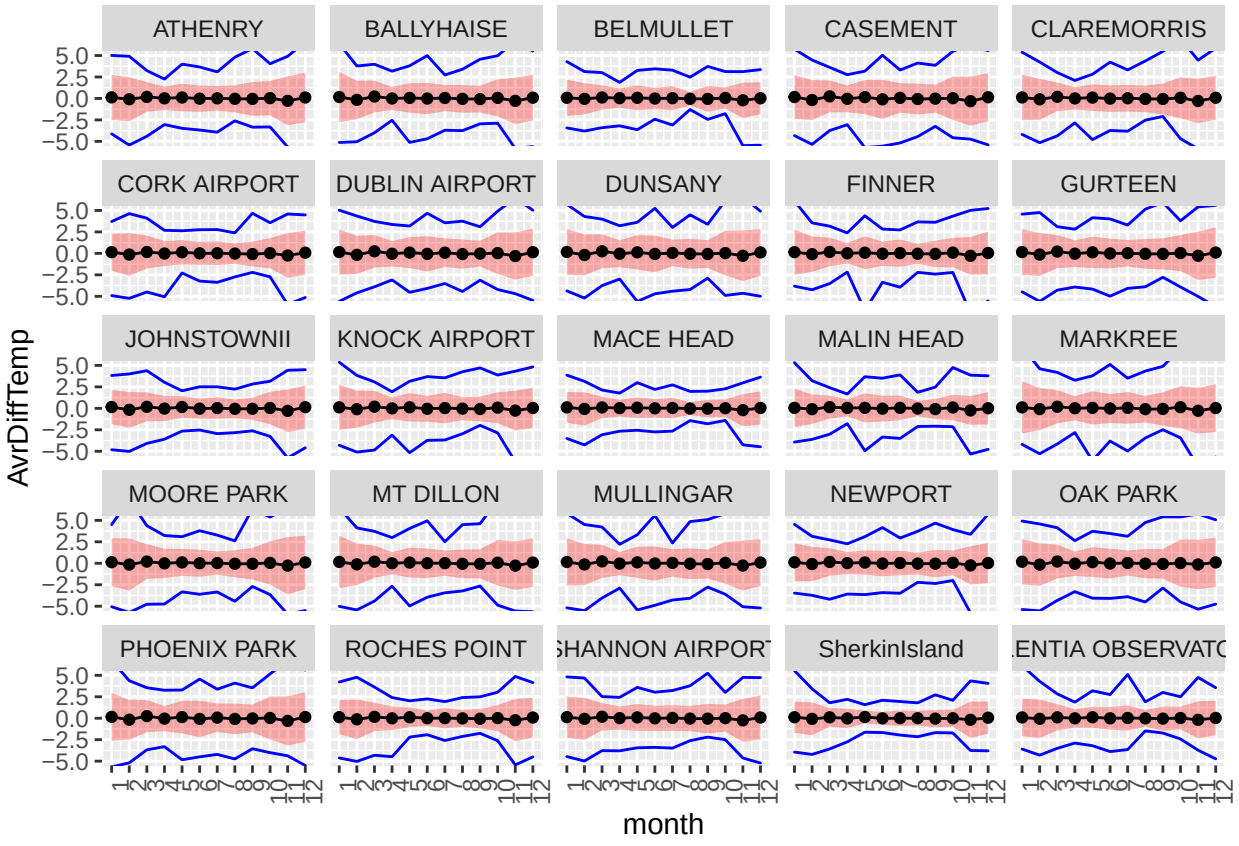
```
## # A tibble: 300 x 10
## # Groups:   station [25]
##    station month AvrDiffTemp SDDiffT~1 MinDi~2 Maxif~3 AvrDiff~4 SDDif~5 MinDi~6
##    <chr>   <dbl>       <dbl>     <dbl>   <dbl>   <dbl>     <dbl>   <dbl>   <dbl>
##  1 ATHENRY     1      0.121       2.62   -4.12    5.01  3.33e- 3    4.82   -16.6
##  2 ATHENRY     2     -0.0929      2.57   -5.43    4.91  1.43e- 1    6.09   -17.2
##  3 ATHENRY     3      0.177       1.73   -4.41    3.24  6.45e- 2    7.68   -21.4
##  4 ATHENRY     4      0.0131      1.36   -3.05    2.25 -2.03e- 1    1.73    -6
##  5 ATHENRY     5      0.109       1.61   -3.48    4    -6.45e- 3    4.37   -13.7
##  6 ATHENRY     6     -0.0232      1.53   -3.68    3.66  6.66e-17    5.31   -10.1
##  7 ATHENRY     7      0.0140      1.43   -3.93    3.10  2.74e- 1   10.5    -41.7
##  8 ATHENRY     8     -0.0431      1.58   -2.60    4.8  -2.74e- 1    5.50   -13.9
##  9 ATHENRY     9     -0.0535      1.96   -3.34    5.81  8   e- 2    6.99   -14.4
## 10 ATHENRY    10      0.0348      2.00   -3.32    4.04 -5.81e- 2    8.53   -21.8
## # ... with 290 more rows, 1 more variable: MaxDiffRain <dbl>, and abbreviated
## #   variable names 1: SDDiffTemp, 2: MinDiffTemp, 3: MaxiffTemp,
## #   4: AvrDiffRain, 5: SDDiffRain, 6: MinDiffRain
```

```
glimpse(out)
```

```
## Rows: 300
## Columns: 10
## Groups: station [25]
## $ station     <chr> "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "AT~
## $ month       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5, 6, 7~
## $ AvrDiffTemp <dbl> 0.120555556, -0.092857143, 0.177016129, 0.013055556, 0.109~
## $ SDDiffTemp  <dbl> 2.623227, 2.567652, 1.734485, 1.357669, 1.609070, 1.533679~
## $ MinDiffTemp <dbl> -4.125000, -5.429167, -4.408333, -3.045833, -3.479167, -3.~
```

```
## $ MaxiffTemp  <dbl> 5.008333, 4.908333, 3.241667, 2.250000, 4.000000, 3.662500~
## $ AvrDiffRain <dbl> 3.333333e-03, 1.428571e-01, 6.451613e-02, -2.033333e-01, -~
## $ SDDiffRain  <dbl> 4.823612, 6.086197, 7.682341, 1.728959, 4.365695, 5.311276~
## $ MinDiffRain <dbl> -16.6, -17.2, -21.4, -6.0, -13.7, -10.1, -41.7, -13.9, -14~
## $ MaxDiffRain <dbl> 17.9, 18.9, 18.6, 3.2, 12.2, 16.3, 33.4, 11.1, 18.6, 21.4,~
```

```
ggplot(out, aes(x = month,
                y = AvrDiffTemp)) +
  geom_ribbon(aes(ymin = MinDiffTemp,
                  ymax = MaxiffTemp),
              alpha=0,
              colour="blue") +
  geom_ribbon(aes(ymin = AvrDiffTemp - SDDiffTemp,
                  ymax = AvrDiffTemp + SDDiffTemp),
              alpha=0.3,
              fill = "red") +
  geom_line(colour="black") +
  geom_point() +
  facet_wrap(~station) +
  coord_cartesian(ylim = c(-5,5)) +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  theme(axis.text.x=element_text(angle=90,hjust=1))
```




```
ggplot(out, aes(x = month,
                y = AvrDiffRain)) +
```

```
  geom_ribbon(aes(ymin = MinDiffRain,
                  ymax = MaxDiffRain),
              alpha=0,
              colour="red") +
  geom_ribbon(aes(ymin = AvrDiffRain - SDDiffRain,
                  ymax = AvrDiffRain + SDDiffRain),
              alpha=0.3,
              fill = "blue") +
  geom_line(colour="black") +
  geom_point() +
  facet_wrap(~station) +
  coord_cartesian(ylim = c(-20,20)) +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  theme(axis.text.x=element_text(angle=90,hjust=1))
```