Assignment 6 - CT5102 (30 marks)

Transforming data with dplyr

The aim of this assignment is to use the package dplyr to transform data. The dataset centres on teh `observations` tibble in `aimsir17`, and involves an analysis of the change in two weather variables - temperature and rainfall - over time.

First, load the following required libraries.

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

The first task is to preapre a new tibble that is a summary of the total rainfall and the average temperature for each day (and for each station).

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

Here is a snapshot of the data generated.

```
s_data
```

```
## # A tibble: 9,125 x 5
##    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
## # i Use `print(n = ...)` to see more rows
```

```
glimpse(s_data)
```

```
## Rows: 9,125
## Columns: 5
## $ 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~
```

Next, perform a set of differencing operatiopns on the data, to calculate the daily changes in temperature and rainfall, for each station. Check to see that the differeence operation is valid, for example, the difference between two observations on two successive days for a given weather station should be checked. The `dplyr` function `lag()` can be used to get a previous value of an observation. Note that the `group_by()` function can also be used for a mutate operation.

Here is a snapshot of the data generated.

```
s_data_diff
```

```
## # A tibble: 9,125 x 9
##    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      -0.2         0.2      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
## # i Use 'print(n = ...)' to see more rows
```

```
glimpse(s_data_diff)
```

```
## Rows: 9,125
## Columns: 9
## $ 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~
```

As a check, the following values should be displayed.

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

```
## # A tibble: 5 x 9
##   station           day month TotalRain AvrTemp RainDiff AbsRa~1 MeanT~2 AbsMe~3
##   <chr>           <int> <dbl>     <dbl>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>
```

```
## 1 MALIN HEAD         22     8      77.7   17.4      76.6    76.6    1.15    1.15
## 2 MALIN HEAD         23     8       1.4   15.3     -76.3    76.3   -2.12    2.12
## 3 DUBLIN AIRPORT     23    11       0.2    4.02    -51.7    51.7   -4.12    4.12
## 4 DUBLIN AIRPORT     22    11      51.9    8.14     51.5    51.5   -4.69    4.69
## 5 SHANNON AIRPORT    22     7       0.1   15.1     -46.2    46.2    3.23    3.23
## # ... with abbreviated variable names 1: AbsRainDiff, 2: MeanTempDiff,
## #   3: AbsMeanTempDiff
```

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

```
## # A tibble: 5 x 9
##   station      day month TotalRain AvrTemp RainDiff AbsRainDiff MeanTe~1 AbsMe~2
##   <chr>      <int> <dbl>     <dbl>   <dbl>    <dbl>       <dbl>    <dbl>   <dbl>
## 1 MARKREE       19    12       0.1   11.2       0           0       7.54    7.54
## 2 MOORE PARK     6     2      18.6    7.72     18.4        18.4     7.40    7.40
## 3 MT DILLON     28    10       0.5   12.8       0.4         0.4     7.35    7.35
## 4 MARKREE       24     1       1.4    9.67      1.4         1.4     7.19    7.19
## 5 MARKREE       28    10       0.8   12.7       0.8         0.8     7.00    7.00
## # ... with abbreviated variable names 1: MeanTempDiff, 2: AbsMeanTempDiff
```

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

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

```
out
```

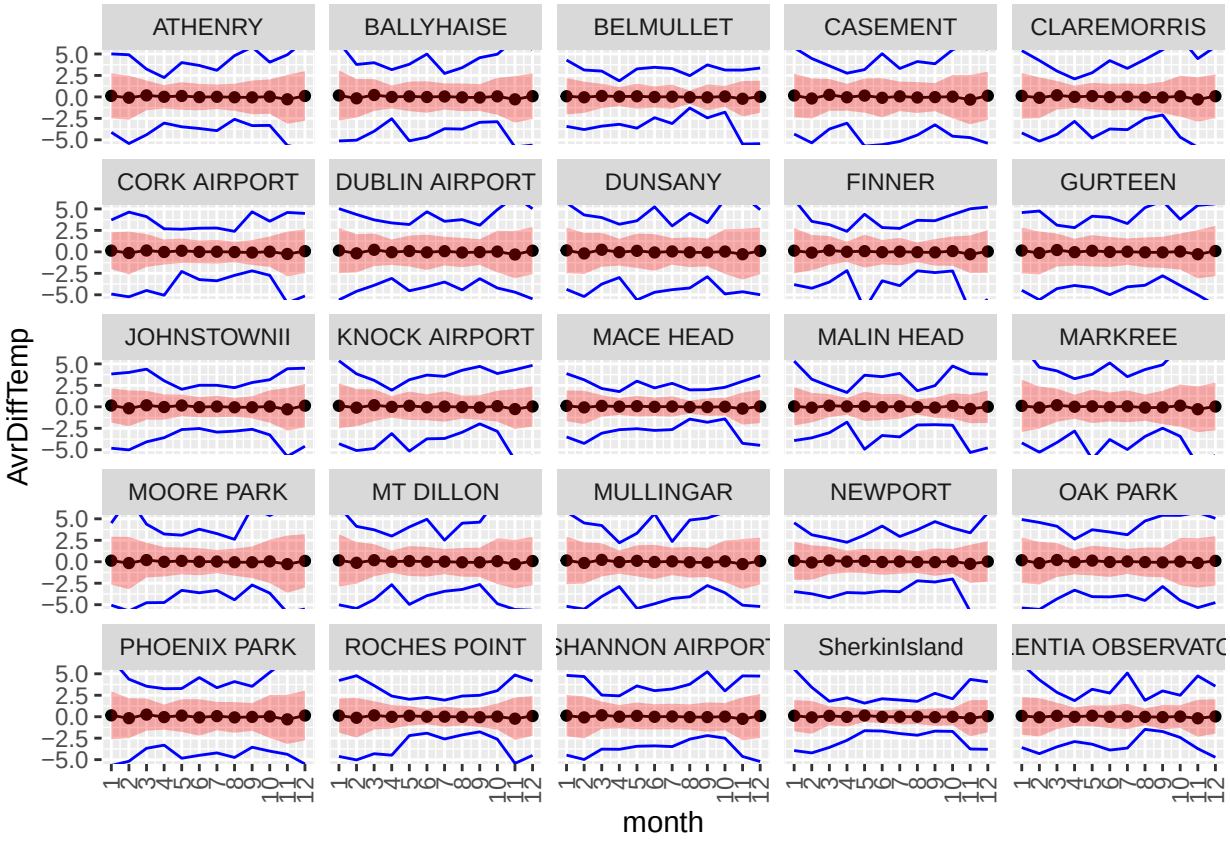
```
## # A tibble: 300 x 10
##    station month AvrDiffTemp SDDiffT~1 MinDi~2 MaxDi~3 AvrDiff~4 SDDif~5 MinDi~6
##    <chr>   <dbl>       <dbl>     <dbl>   <dbl>   <dbl>     <dbl>   <dbl>   <dbl>
##  1 ATHENRY     1      0.121       2.62   -4.12    5.01  3.33e- 3    4.82   -16.6
##  2 ATHENRY     2     -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  4.45e-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  8.00e- 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: MaxDiffTemp,
## #   4: AvrDiffRain, 5: SDDiffRain, 6: MinDiffRain
## # i Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

```
glimpse(out)
```

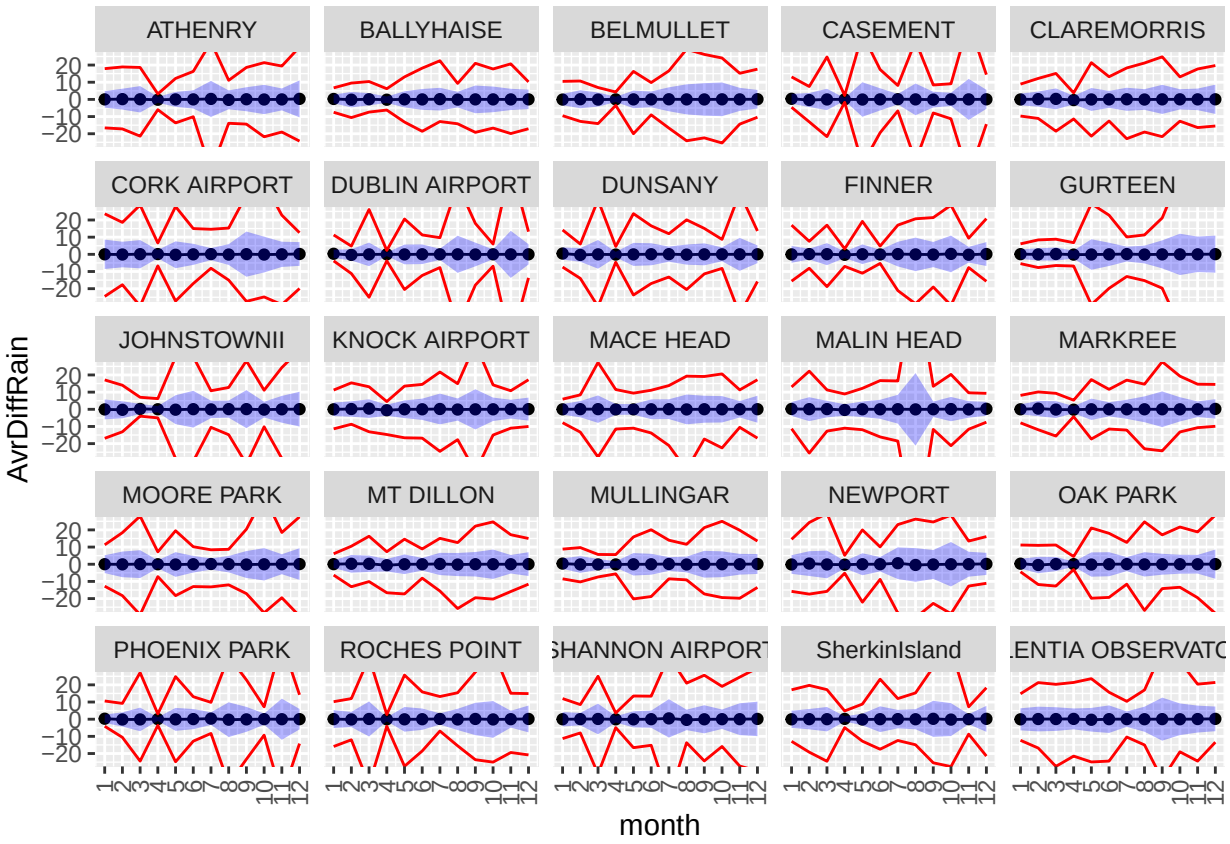
```
## Rows: 300
## Columns: 10
## $ station     <chr> "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "AT~
```

```
## $ 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.~
## $ MaxDiffTemp <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,~
```

Generate the following graph (based on temperature) that shows the mean, the standard deviation (`geom_ribbon()`) and the min and max for each month (blue line). Note that the output is constrained to show the range -5 to 5, use the function `coord_cartesian()` for this.



Generate the following graph (based on rainfall) that shows the mean, the standard deviation (`geom_ribbon()`) and the min and max for each month (red line). Note that the output is constrained to show the range -5 to 5, use the function `coord_cartesian()` for this.


ATHENRY
BALLYHAISE
BELMULLET
CASEMENT
CLAREMORRIS
CORK AIRPORT
DUBLIN AIRPORT
DUNSANY
FINNER
GURTEEN
JOHNSTOWNII
KNOCK AIRPORT
MACE HEAD
MALIN HEAD
MARKREE
MOORE PARK
MT DILLON
MULLINGAR
NEWPORT
OAK PARK
PHOENIX PARK
ROCHES POINT
SHANNON AIRPORT
SherkinIsland
LENTIA OBSERVATO
AvrDiffRain
20
10
0
−10
−20
1 2 3 4 5 6 7 8 9 10 11 12
month