Assignment 7 - CT5102 (20 marks)

Relational data with dplyr

The aim of this assignment is to use the package dplyr to process related data. The idea is to explore correlations between electricity demand (`eirgrid17`) and recorded temperature from selected weather stations, as stored in `observations`.

First, load the following required libraries.

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

The first task is to preapre a new tibble that has an additional column (Season) for three weather stations, MACE HEAD, DUBLIN AIRPORT and SherkinIsland. Make use of the function `case_when` to calculate the new Season column. Assume Winter is November, December and January; Spring is February, March and April; Summer is May, June and July; and Autumn is August, September and October.

Here is a snapshot of the weather data generated.

```
obs
```

```
## # A tibble: 26,280 x 13
##    station    year month   day  hour date                 rain  temp  rhum   msl
##    <chr>     <dbl> <dbl> <int> <int> <dttm>              <dbl> <dbl> <dbl> <dbl>
##  1 DUBLIN A~  2017     1     1     0 2017-01-01 00:00:00   0.9   5.3    91 1020.
##  2 DUBLIN A~  2017     1     1     1 2017-01-01 01:00:00   0.2   4.9    95 1020.
##  3 DUBLIN A~  2017     1     1     2 2017-01-01 02:00:00   0.1   5      92 1020.
##  4 DUBLIN A~  2017     1     1     3 2017-01-01 03:00:00   0     4.2    90 1020.
##  5 DUBLIN A~  2017     1     1     4 2017-01-01 04:00:00   0     3.6    88 1020.
##  6 DUBLIN A~  2017     1     1     5 2017-01-01 05:00:00   0     2.8    89 1020.
##  7 DUBLIN A~  2017     1     1     6 2017-01-01 06:00:00   0     1.7    91 1020.
##  8 DUBLIN A~  2017     1     1     7 2017-01-01 07:00:00   0     1.6    91 1021
##  9 DUBLIN A~  2017     1     1     8 2017-01-01 08:00:00   0     2      89 1022.
## 10 DUBLIN A~  2017     1     1     9 2017-01-01 09:00:00   0     2.6    84 1023.
## # ... with 26,270 more rows, and 3 more variables: wdsp <dbl>, wddir <dbl>,
## #   Season <chr>
## # i Use ‘print(n = ...)‘ to see more rows, and ‘colnames()‘ to see all variable names
```

```
glimpse(obs)
```

```
## Rows: 26,280
## Columns: 13
## $ station <chr> "DUBLIN AIRPORT", "DUBLIN AIRPORT", "DUBLIN AIRPORT", "DUBLIN ~
## $ year    <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 20~
## $ month   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ day     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ hour    <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ date    <dttm> 2017-01-01 00:00:00, 2017-01-01 01:00:00, 2017-01-01 02:00:00~
## $ rain    <dbl> 0.9, 0.2, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.~
## $ temp    <dbl> 5.3, 4.9, 5.0, 4.2, 3.6, 2.8, 1.7, 1.6, 2.0, 2.6, 3.0, 3.6, 4.~
## $ rhum    <dbl> 91, 95, 92, 90, 88, 89, 91, 91, 89, 84, 84, 80, 76, 75, 73, 72~
## $ msl     <dbl> 1019.9, 1019.7, 1019.8, 1020.2, 1020.2, 1020.4, 1020.4, 1021.0~
## $ wdsp    <dbl> 12, 8, 8, 12, 11, 12, 13, 13, 13, 13, 11, 12, 13, 16, 14, 15, ~
## $ wddir   <dbl> 340, 310, 310, 330, 330, 330, 330, 330, 330, 340, 350, 350, 35~
## $ Season  <chr> "Winter", "Winter", "Winter", "Winter", "Winter", "Winter", "W~
```

Next, extract energy data from `eirgrid17` that records the average hourly demand from both regions of the island, `IEDemand` and `NIDemand`.

```
## ‘summarise()‘ has grouped output by ’year’, ’month’, ’day’. You can override
## using the ‘.groups‘ argument.
```

Here is a snapshot of the energy data generated.

```
ener
```

```
## # A tibble: 8,759 x 7
##     year month   day  hour    IE    NI CheckObs
##    <dbl> <dbl> <int> <int> <dbl> <dbl>    <int>
##  1  2017     1     1     0 2833.  763.        4
##  2  2017     1     1     1 2617.  732.        4
##  3  2017     1     1     2 2427.  675.        4
##  4  2017     1     1     3 2295.  625.        4
##  5  2017     1     1     4 2223.  598.        4
##  6  2017     1     1     5 2180.  583.        4
##  7  2017     1     1     6 2218.  606.        4
##  8  2017     1     1     7 2265.  646.        4
##  9  2017     1     1     8 2277.  692.        4
## 10  2017     1     1     9 2444.  757.        4
## # ... with 8,749 more rows
## # i Use ‘print(n = ...)‘ to see more rows
```

```
glimpse(ener)
```

```
## Rows: 8,759
## Columns: 7
## $ year     <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2~
## $ month    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ day      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ hour     <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
## $ IE       <dbl> 2832.695, 2616.740, 2426.577, 2294.968, 2222.948, 2179.637, 2~
## $ NI       <dbl> 762.5170, 731.8795, 675.1053, 624.5440, 598.3955, 583.1503, 6~
## $ CheckObs <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
```

Next, after setting the seed value to 100, left-join the datasets and sample 10% of the records.

Here is a snapshot of the sampled table.

```
ds
```

```
## # A tibble: 2,628 x 16
##     year month   day  hour    IE    NI Check~1 station date                 rain
##    <dbl> <dbl> <int> <int> <dbl> <dbl>   <int> <chr>   <dttm>              <dbl>
##  1  2017    10     8     4 2216.  581.       4 DUBLIN~ 2017-10-08 04:00:00   0
##  2  2017     8    23    13 3561. 1039.       4 Sherki~ 2017-08-23 13:00:00   0
##  3  2017     2    17    15 3763. 1177.       4 DUBLIN~ 2017-02-17 15:00:00   0
##  4  2017     2    21     7 3287. 1109.       4 Sherki~ 2017-02-21 07:00:00   0.1
##  5  2017    10    12     9 3641. 1121.       4 MACE H~ 2017-10-12 09:00:00   0
##  6  2017    12     4     2 2567.  650.       4 Sherki~ 2017-12-04 02:00:00   0
##  7  2017     2    12     9 3088.  940.       4 DUBLIN~ 2017-02-12 09:00:00   0
##  8  2017     6     5     0 2374.  645.       4 MACE H~ 2017-06-05 00:00:00   0
##  9  2017     4    24    19 3509.  963.       4 MACE H~ 2017-04-24 19:00:00   0
## 10  2017    12    14    21 4111. 1179.       4 MACE H~ 2017-12-14 21:00:00   0
## # ... with 2,618 more rows, 6 more variables: temp <dbl>, rhum <dbl>,
## #   msl <dbl>, wdsp <dbl>, wddir <dbl>, Season <chr>, and abbreviated variable
## #   name 1: CheckObs
## # i Use ‘print(n = ...)‘ to see more rows, and ‘colnames()‘ to see all variable names
```

```
glimpse(ds)
```

```
## Rows: 2,628
## Columns: 16
## $ year     <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2~
## $ month    <dbl> 10, 8, 2, 2, 10, 12, 2, 6, 4, 12, 9, 12, 8, 6, 7, 11, 11, 6, ~
## $ day      <int> 8, 23, 17, 21, 12, 4, 12, 5, 24, 14, 13, 29, 28, 3, 6, 22, 24~
## $ hour     <int> 4, 13, 15, 7, 9, 2, 9, 0, 19, 21, 5, 18, 0, 21, 6, 11, 5, 5, ~
## $ IE       <dbl> 2216.132, 3561.375, 3762.565, 3286.770, 3640.680, 2566.970, 3~
## $ NI       <dbl> 581.2073, 1038.9145, 1176.8927, 1109.3122, 1120.6975, 649.948~
## $ CheckObs <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ station  <chr> "DUBLIN AIRPORT", "SherkinIsland", "DUBLIN AIRPORT", "Sherkin~
## $ date     <dttm> 2017-10-08 04:00:00, 2017-08-23 13:00:00, 2017-02-17 15:00:0~
## $ rain     <dbl> 0.0, 0.0, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0~
## $ temp     <dbl> 11.6, 16.0, 11.9, 10.3, 13.1, 8.9, 3.5, 12.4, 5.7, 7.0, 11.8,~
## $ rhum     <dbl> 96, 85, 80, 96, 84, 81, 73, 90, 65, 84, 73, 72, 94, 74, 98, 1~
## $ msl      <dbl> 1020.1, 1014.2, 1021.7, 1019.3, 1012.5, 1036.2, 1029.6, 999.5~
## $ wdsp     <dbl> 7, 10, 13, 17, 18, 3, 18, 21, 18, 18, 23, 17, 19, 17, 4, 11, ~
## $ wddir    <dbl> 270, 240, 150, 250, 200, 290, 70, 230, 30, 350, 280, 250, 210~
## $ Season   <chr> "Autumn", "Autumn", "Spring", "Spring", "Autumn", "Winter", "~
```

Reduce the number of columns in the table

```
ds
```

```
## # A tibble: 2,628 x 6
##    station        month  temp Season    IE    NI
##    <chr>          <dbl> <dbl> <chr>  <dbl> <dbl>
##  1 DUBLIN AIRPORT    10  11.6 Autumn 2216.  581.
##  2 SherkinIsland      8  16   Autumn 3561. 1039.
##  3 DUBLIN AIRPORT     2  11.9 Spring 3763. 1177.
##  4 SherkinIsland      2  10.3 Spring 3287. 1109.
##  5 MACE HEAD         10  13.1 Autumn 3641. 1121.
##  6 SherkinIsland     12   8.9 Winter 2567.  650.
##  7 DUBLIN AIRPORT     2   3.5 Spring 3088.  940.
##  8 MACE HEAD          6  12.4 Summer 2374.  645.
##  9 MACE HEAD          4   5.7 Spring 3509.  963.
## 10 MACE HEAD         12   7   Winter 4111. 1179.
## # ... with 2,618 more rows
## # i Use ‘print(n = ...)‘ to see more rows
```

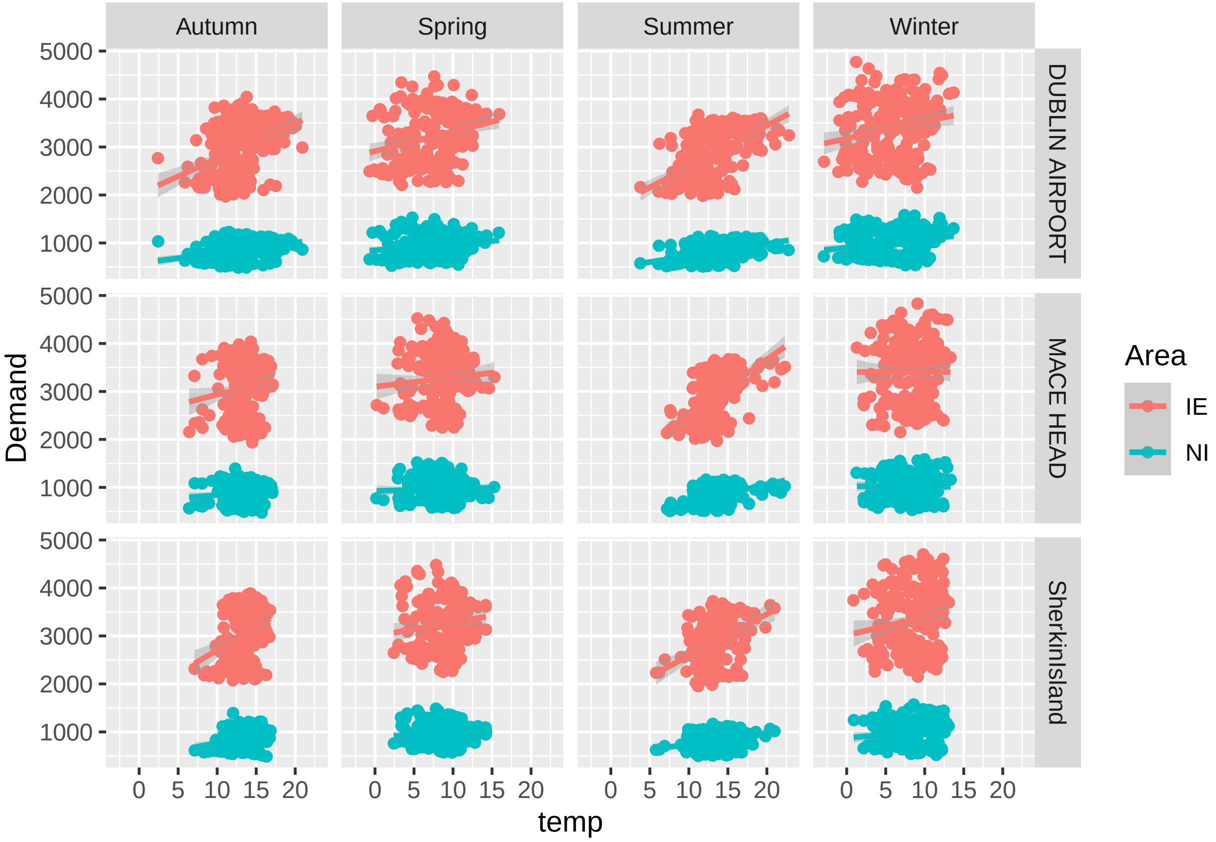
Create a new dataset `ds1` fromn `ds`, using `tidyr::pivot_longer` to generate the following.

```
ds1
```

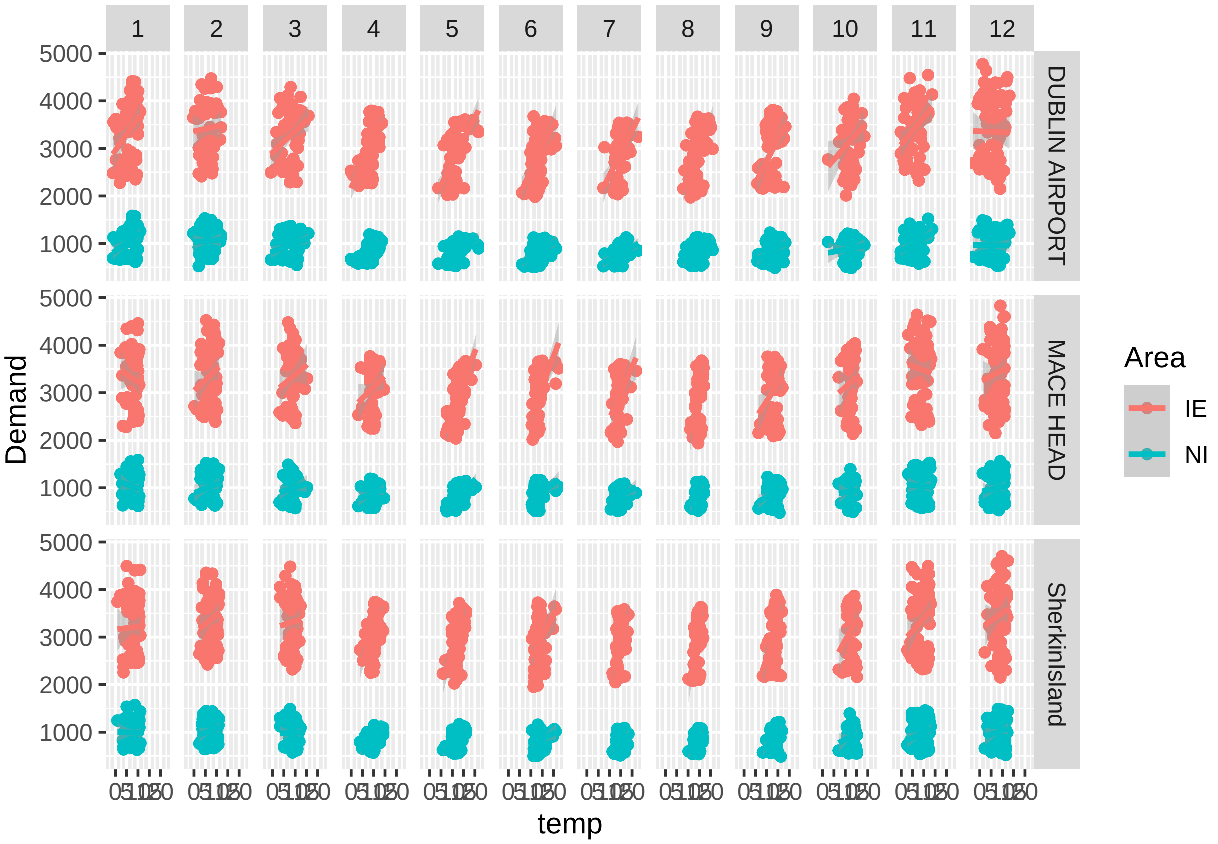
```
## # A tibble: 5,256 x 6
##    station        month  temp Season Area  Demand
##    <chr>          <dbl> <dbl> <chr>  <chr>  <dbl>
##  1 DUBLIN AIRPORT    10  11.6 Autumn IE     2216.
##  2 DUBLIN AIRPORT    10  11.6 Autumn NI      581.
##  3 SherkinIsland      8  16   Autumn IE     3561.
##  4 SherkinIsland      8  16   Autumn NI     1039.
##  5 DUBLIN AIRPORT     2  11.9 Spring IE     3763.
##  6 DUBLIN AIRPORT     2  11.9 Spring NI     1177.
##  7 SherkinIsland      2  10.3 Spring IE     3287.
##  8 SherkinIsland      2  10.3 Spring NI     1109.
##  9 MACE HEAD         10  13.1 Autumn IE     3641.
## 10 MACE HEAD         10  13.1 Autumn NI     1121.
## # ... with 5,246 more rows
## # i Use ‘print(n = ...)‘ to see more rows
```

Plot the following two graphs from the tiblle `ds1`.

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
## 'geom_smooth()' using formula 'y ~ x'
```



Generate the following summary tables from the dataset `ds1`.

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

```
slice(cor_season,1:nrow(cor_season))
```

```
## # A tibble: 12 x 5
##    station        Season Corr_IE  Corr_NI     Diff
##    <chr>          <chr>    <dbl>    <dbl>    <dbl>
##  1 DUBLIN AIRPORT Autumn 0.387    0.309    0.0776
##  2 DUBLIN AIRPORT Spring 0.261    0.195    0.0662
##  3 DUBLIN AIRPORT Summer 0.555    0.464    0.0917
##  4 DUBLIN AIRPORT Winter 0.193    0.229   -0.0359
##  5 MACE HEAD      Autumn 0.154    0.127    0.0262
##  6 MACE HEAD      Spring 0.0881   0.0451   0.0430
##  7 MACE HEAD      Summer 0.533    0.454    0.0790
##  8 MACE HEAD      Winter 0.00168 -0.00542  0.00710
##  9 SherkinIsland  Autumn 0.295    0.259    0.0361
## 10 SherkinIsland  Spring 0.127    0.0727   0.0538
## 11 SherkinIsland  Summer 0.345    0.258    0.0871
## 12 SherkinIsland  Winter 0.157    0.134    0.0221
```

```
slice(cor_month,1:nrow(cor_month))
```

```
## # A tibble: 36 x 5
##    station        month Corr_IE Corr_NI    Diff
##    <chr>          <dbl>   <dbl>   <dbl>   <dbl>
##  1 DUBLIN AIRPORT     1  0.343   0.388  -0.0448
##  2 DUBLIN AIRPORT     2  0.0588  0.0450  0.0138
##  3 DUBLIN AIRPORT     3  0.310   0.286   0.0243
##  4 DUBLIN AIRPORT     4  0.665   0.621   0.0435
##  5 DUBLIN AIRPORT     5  0.642   0.600   0.0424
##  6 DUBLIN AIRPORT     6  0.564   0.476   0.0881
##  7 DUBLIN AIRPORT     7  0.542   0.452   0.0896
##  8 DUBLIN AIRPORT     8  0.630   0.533   0.0970
##  9 DUBLIN AIRPORT     9  0.463   0.374   0.0893
## 10 DUBLIN AIRPORT    10  0.242   0.111   0.131
## # ... with 26 more rows
## # i Use ‘print(n = ...)‘ to see more rows
```