Assignment 7 - CT5102 Relational data with dplyr

Chin Zhe Jing 22221970

zhejing

2022-10-27

```
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
```

```
library(tidyr)

glimpse(observations)
```

```
## Rows: 219,000
## Columns: 12
## $ station <chr> "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "ATHENRY", "ATHENR~
## $ 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> 0.0, 0.0, 0.0, 0.1, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.~
## $ temp    <dbl> 5.2, 4.7, 4.2, 3.5, 3.2, 2.1, 2.0, 1.7, 1.0, 1.1, 3.0, 4.3, 5.~
## $ rhum    <dbl> 89, 89, 90, 87, 89, 91, 89, 89, 91, 91, 84, 78, 75, 72, 72, 71~
## $ msl     <dbl> 1021.9, 1022.0, 1022.1, 1022.5, 1022.7, 1023.3, 1023.5, 1024.4~
## $ wdsp    <dbl> 8, 9, 8, 9, 8, 8, 7, 7, 7, 8, 9, 12, 11, 12, 11, 11, 11, 6, 8,~
## $ wddir   <dbl> 320, 320, 320, 330, 330, 330, 330, 340, 330, 330, 320, 350, 36~
```

```
obs <- observations |>
  filter(station %in% c("MACE HEAD", "DUBLIN AIRPORT", "SherkinIsland")) |>
  mutate(Season = case_when(
    month %in% c(11, 12, 1) ~ "Winter",
```

```
    month %in% c(2, 3, 4)  ~ "Spring",
    month %in% c(5, 6, 7)  ~ "Summer",
    month %in% c(8, 9, 10)  ~ "Autumn"))

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>
```

```
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~
```

```
glimpse(eirgrid17)
```

```
## Rows: 35,040
## Columns: 15
## $ year              <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 201~
## $ month             <dbl> 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, ~
## $ hour              <int> 0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, ~
## $ minute            <int> 0, 15, 30, 45, 0, 15, 30, 45, 0, 15, 30, 45, 0, 15,~
## $ date              <dttm> 2017-01-01 00:00:00, 2017-01-01 00:15:00, 2017-01-~
## $ NIGeneration      <dbl> 889.005, 922.234, 908.122, 918.802, 882.441, 848.86~
## $ NIDemand          <dbl> 775.931, 770.233, 761.186, 742.718, 749.238, 742.45~
```

```
## $ NIWindAvailability <dbl> 175.065, 182.866, 169.796, 167.501, 174.094, 189.92~
## $ NIWindGeneration   <dbl> 198.202, 207.765, 193.103, 190.757, 195.790, 212.95~
## $ IEGeneration       <dbl> 3288.57, 3282.12, 3224.27, 3171.27, 3190.28, 3184.6~
## $ IEDemand           <dbl> 2921.44, 2884.19, 2806.38, 2718.77, 2682.91, 2649.8~
## $ IEWindAvailability <dbl> 1064.79, 965.60, 915.35, 895.38, 1028.03, 1144.17, ~
## $ IEWindGeneration   <dbl> 1044.72, 957.74, 900.46, 870.81, 998.31, 1119.12, 1~
## $ SNSP               <chr> "28.4%", "26.4%", "25.2%", "24.7%", "27.9%", "31.4%~
```

```
ener <- eirgrid17 |>
  group_by(year, month, day, hour) |>
  summarise(IE = mean(IEDemand, na.rm=T),
            NI = mean(NIDemand, na.rm=T),
            CheckObs = n())
```

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

```
ener
```

```
## # A tibble: 8,759 x 7
## # Groups:   year, month, day [365]
##     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
```

```
glimpse(ener)
```

```
## Rows: 8,759
## Columns: 7
## Groups: year, month, day [365]
## $ 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~
```

```
set.seed(100)

ds <- left_join(ener, obs) |>
  ungroup() |>
  sample_frac(0.1)
```

```
## Joining, by = c("year", "month", "day", "hour")
```

```
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
```

```
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", "~
```

```
ds <- ds |>
  select(station, month, temp, Season, IE, NI)

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
```

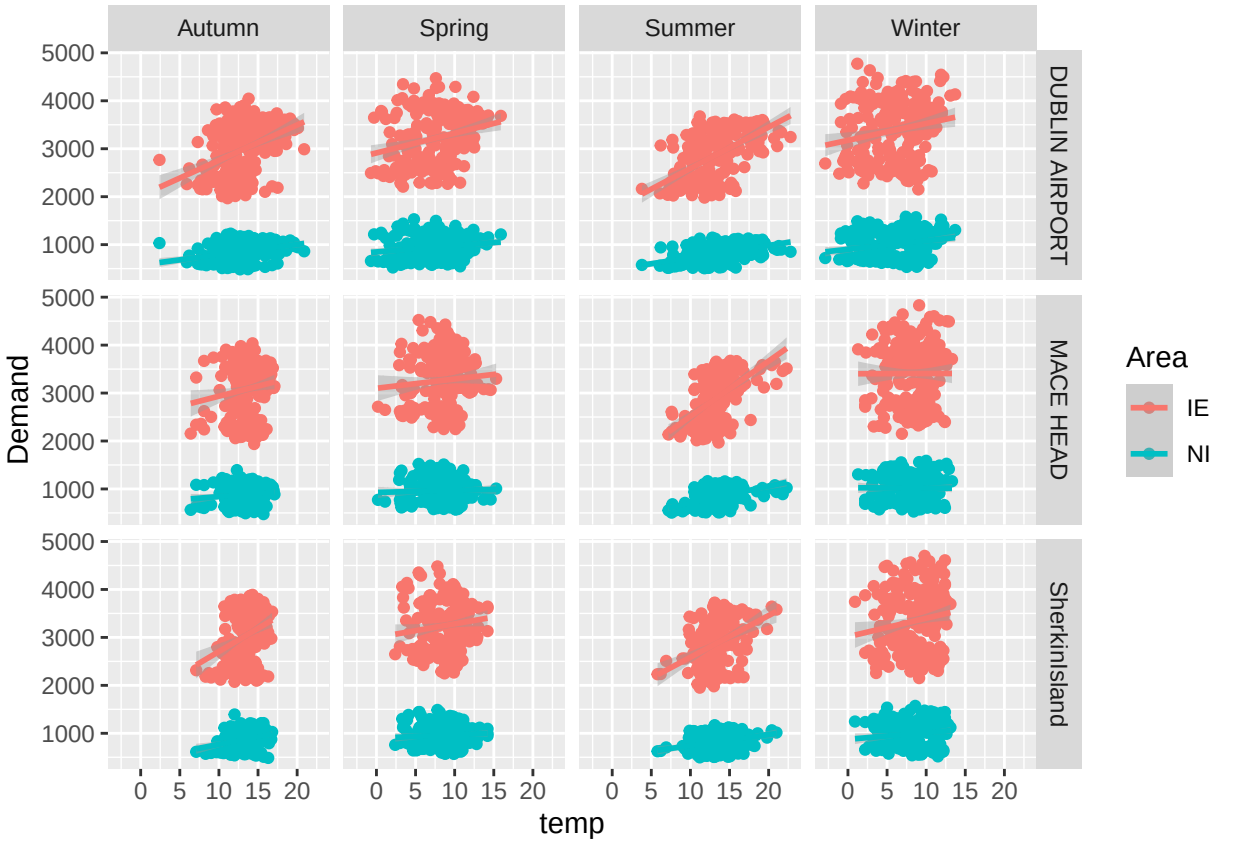
```
ds1 <- ds |>
  pivot_longer(-(station:Season),
               names_to="Area",
               values_to='Demand')

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
```

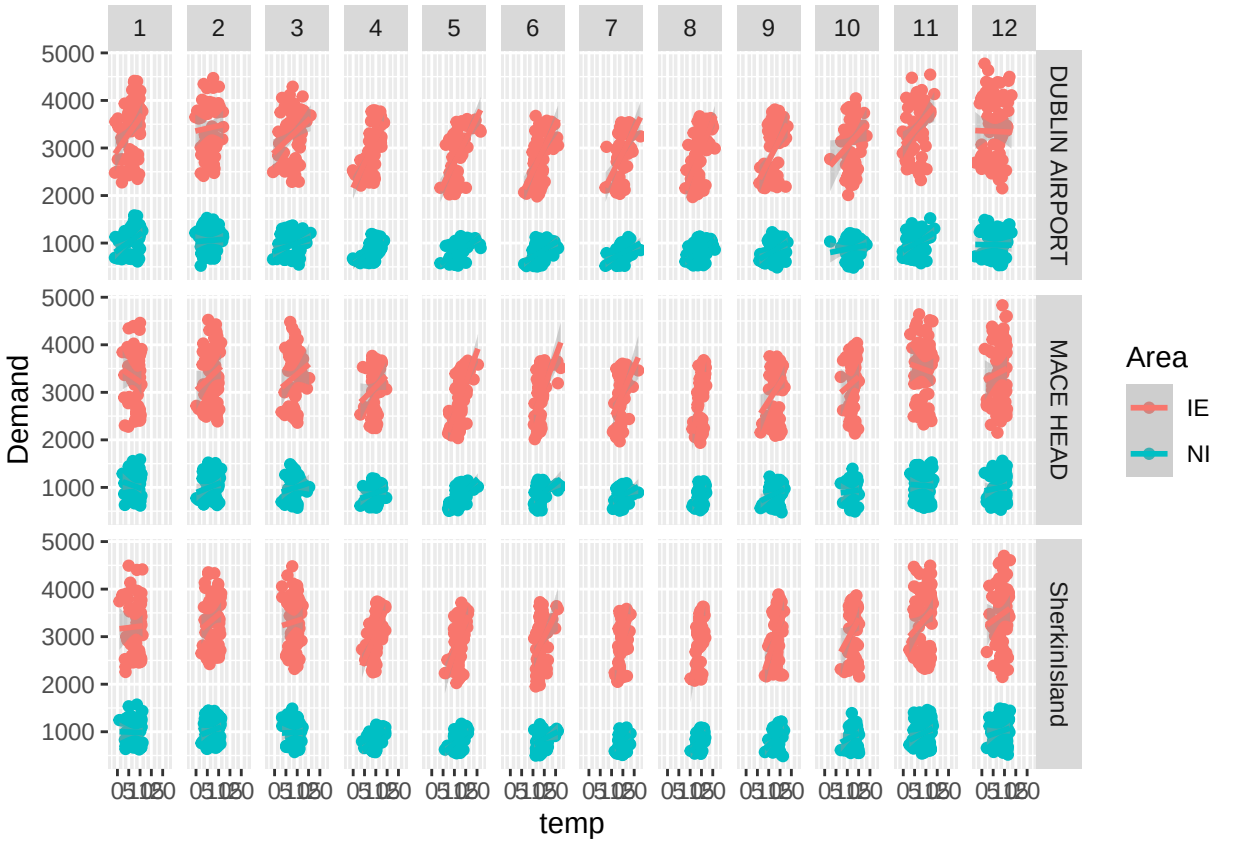
```
ggplot(ds1, aes(x=temp, y=Demand, colour=Area))+
  geom_point()+geom_smooth(method='lm')+
  facet_grid(station~Season)
```

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




```
ggplot(ds1, aes(x=temp, y=Demand, colour=Area))+
  geom_point()+geom_smooth(method='lm')+
  facet_grid(station~month)
```

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




```
cor_season <- ds1 |>
  group_by(station, Season, Area) |>
  summarise(corr = cor(Demand, temp))
```

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

```
cor_season <- cor_season |> ungroup() |>
  pivot_wider(c('station', 'Season'),
              names_from = Area,
              names_glue = "Corr_{Area}",
              values_from = corr) |>
    mutate(Diff = Corr_IE - Corr_NI)

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
```

```
cor_month <- ds1 |>
  group_by(station, month, Area) |>
  summarise(corr = cor(Demand, temp))
```

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

```
cor_month <- cor_month |> ungroup() |>
  pivot_wider(c('station', 'month'),
              names_from = Area,
              names_glue = "Corr_{Area}",
              values_from = corr) |>
  mutate(Diff = Corr_IE - Corr_NI)

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
```