Assignment 8 - CT5102 Using purrr for a Data Science Workflow

Chin Zhe Jing 22221970

zhejing

2022-11-03

```
library(aimsir17)
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(ggplot2)
library(purrr)
library(tidyr)
library(ggpubr)
library(randomcoloR)
set.seed(200)

ener <- eirgrid17 |> group_by(year, month, day, hour) |>
  summarise(AvrWindGen = mean(IEWindGeneration))  |> ungroup()
```

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

```
ener
```

```
## # A tibble: 8,759 x 5
##     year month   day  hour AvrWindGen
##    <dbl> <dbl> <int> <int>      <dbl>
##  1  2017     1     1     0       943.
##  2  2017     1     1     1      1085.
##  3  2017     1     1     2      1284.
##  4  2017     1     1     3      1254.
##  5  2017     1     1     4      1277.
##  6  2017     1     1     5      1346.
```

```
##  7  2017     1     1     6      1423.
##  8  2017     1     1     7      1534.
##  9  2017     1     1     8      1542.
## 10  2017     1     1     9      1518.
## # ... with 8,749 more rows
```

```
glimpse(ener)
```

```
## Rows: 8,759
## Columns: 5
## $ year       <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017,~
## $ month      <dbl> 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,~
## $ hour       <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ AvrWindGen <dbl> 943.4325, 1084.9500, 1284.2800, 1253.7350, 1276.8525, 1345.~
```

```
ds <- ener |> left_join(observations, by = c("year", "month", "day", "hour"))
ds
```

```
## # A tibble: 218,975 x 13
##     year month   day  hour AvrWi~1 station date                rain  temp  rhum
##    <dbl> <dbl> <int> <int>   <dbl> <chr>   <dttm>              <dbl> <dbl> <dbl>
##  1  2017     1     1     0    943. ATHENRY 2017-01-01 00:00:00   0     5.2    89
##  2  2017     1     1     0    943. BALLYH~ 2017-01-01 00:00:00   0     4.4    94
##  3  2017     1     1     0    943. BELMUL~ 2017-01-01 00:00:00   0     5.2    79
##  4  2017     1     1     0    943. CASEME~ 2017-01-01 00:00:00   1.5   5.6    92
##  5  2017     1     1     0    943. CLAREM~ 2017-01-01 00:00:00   0     4.6    88
##  6  2017     1     1     0    943. CORK A~ 2017-01-01 00:00:00   1.4   8      94
##  7  2017     1     1     0    943. DUBLIN~ 2017-01-01 00:00:00   0.9   5.3    91
##  8  2017     1     1     0    943. DUNSANY 2017-01-01 00:00:00   0.5   5.2    99
##  9  2017     1     1     0    943. FINNER  2017-01-01 00:00:00   0     4.1    83
## 10  2017     1     1     0    943. GURTEEN 2017-01-01 00:00:00   0.3   5.4    91
## # ... with 218,965 more rows, 3 more variables: msl <dbl>, wdsp <dbl>,
## #   wddir <dbl>, and abbreviated variable name 1: AvrWindGen
```

```
glimpse(ds)
```

```
## Rows: 218,975
## Columns: 13
## $ year       <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017,~
## $ month      <dbl> 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,~
## $ hour       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ AvrWindGen <dbl> 943.4325, 943.4325, 943.4325, 943.4325, 943.4325, 943.4325,~
## $ station    <chr> "ATHENRY", "BALLYHAISE", "BELMULLET", "CASEMENT", "CLAREMOR~
## $ date       <dttm> 2017-01-01 00:00:00, 2017-01-01 00:00:00, 2017-01-01 00:00~
## $ rain       <dbl> 0.0, 0.0, 0.0, 1.5, 0.0, 1.4, 0.9, 0.5, 0.0, 0.3, 0.4, 0.1,~
## $ temp       <dbl> 5.2, 4.4, 5.2, 5.6, 4.6, 8.0, 5.3, 5.2, 4.1, 5.4, 9.8, 3.6,~
## $ rhum       <dbl> 89, 94, 79, 92, 88, 94, 91, 99, 83, 91, 94, 95, 88, 70, 84,~
## $ msl        <dbl> 1021.9, 1020.3, 1023.0, 1020.1, 1021.9, 1020.9, 1019.9, 102~
## $ wdsp       <dbl> 8, 5, 13, 8, 8, 11, 12, 6, 12, 7, 11, 8, 17, 19, NA, 9, 6, ~
## $ wddir      <dbl> 320, 310, 340, 330, 330, 230, 340, 340, 360, 330, 230, 330,~
```

```
sum(!complete.cases(ds))
```

```
## [1] 17755
```

```
ds <- ds[complete.cases(ds),] |> ungroup() |> sample_frac(0.01)
ds
```

```
## # A tibble: 2,012 x 13
##     year month   day  hour AvrWi~1 station date                rain  temp  rhum
##    <dbl> <dbl> <int> <int>   <dbl> <chr>   <dttm>              <dbl> <dbl> <dbl>
##  1  2017     7     8     2    33.9 FINNER  2017-07-08 02:00:00   0     9.9    77
##  2  2017    11    16    19   186.  CLAREM~ 2017-11-16 19:00:00   0     3.6    88
##  3  2017    12     4    12   223.  FINNER  2017-12-04 12:00:00   0.3   8.4    95
##  4  2017     5    27    17   858.  DUBLIN~ 2017-05-27 17:00:00   0    14.6    91
##  5  2017     7    15    21   754.  ATHENRY 2017-07-15 21:00:00   0    15.4    97
##  6  2017    11    11    17   663.  SHANNO~ 2017-11-11 17:00:00   1.3   8.8    99
##  7  2017    10     2     1  1351   VALENT~ 2017-10-02 01:00:00   0    13.3    74
##  8  2017     9     7     7   625.  MOORE ~ 2017-09-07 07:00:00   0    13.7    88
##  9  2017     5    31     6   368.  MULLIN~ 2017-05-31 06:00:00   0     9      90
## 10  2017     6    13    22   483.  GURTEEN 2017-06-13 22:00:00   0    12.5    86
## # ... with 2,002 more rows, 3 more variables: msl <dbl>, wdsp <dbl>,
## #   wddir <dbl>, and abbreviated variable name 1: AvrWindGen
```

```
glimpse(ds)
```

```
## Rows: 2,012
## Columns: 13
## $ year       <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017,~
## $ month      <dbl> 7, 11, 12, 5, 7, 11, 10, 9, 5, 6, 10, 5, 10, 3, 8, 3, 5, 7,~
## $ day        <int> 8, 16, 4, 27, 15, 11, 2, 7, 31, 13, 27, 19, 4, 8, 23, 27, 2~
## $ hour       <int> 2, 19, 12, 17, 21, 17, 1, 7, 6, 22, 12, 11, 6, 13, 6, 14, 1~
## $ AvrWindGen <dbl> 33.9425, 186.0475, 222.9575, 858.3500, 754.4050, 662.7375, ~
## $ station    <chr> "FINNER", "CLAREMORRIS", "FINNER", "DUBLIN AIRPORT", "ATHEN~
## $ date       <dttm> 2017-07-08 02:00:00, 2017-11-16 19:00:00, 2017-12-04 12:00~
## $ rain       <dbl> 0.0, 0.0, 0.3, 0.0, 0.0, 1.3, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,~
## $ temp       <dbl> 9.9, 3.6, 8.4, 14.6, 15.4, 8.8, 13.3, 13.7, 9.0, 12.5, 7.5,~
## $ rhum       <dbl> 77, 88, 95, 91, 97, 99, 74, 88, 90, 86, 100, 79, 81, 73, 97~
## $ msl        <dbl> 1019.5, 1028.8, 1034.8, 1009.9, 1018.8, 1012.7, 1013.1, 101~
## $ wdsp       <dbl> 4, 3, 12, 17, 6, 8, 19, 4, 4, 8, 2, 9, 7, 11, 2, 9, 4, 9, 7~
## $ wddir      <dbl> 190, 220, 250, 280, 250, 30, 270, 150, 80, 180, 180, 180, 2~
```

```
sum(!complete.cases(ds))
```

```
## [1] 0
```

```
ds_n <- ds |> group_by(station) |> nest()  |> ungroup()
ds_n
```

```
## # A tibble: 23 x 2
##    station              data
```

```
##    <chr>                <list>
##  1 FINNER               <tibble [96 x 12]>
##  2 CLAREMORRIS          <tibble [92 x 12]>
##  3 DUBLIN AIRPORT       <tibble [73 x 12]>
##  4 ATHENRY              <tibble [86 x 12]>
##  5 SHANNON AIRPORT      <tibble [94 x 12]>
##  6 VALENTIA OBSERVATORY <tibble [88 x 12]>
##  7 MOORE PARK           <tibble [98 x 12]>
##  8 MULLINGAR            <tibble [91 x 12]>
##  9 GURTEEN              <tibble [83 x 12]>
## 10 OAK PARK             <tibble [92 x 12]>
## # ... with 13 more rows
```

```
glimpse(ds_n)
```

```
## Rows: 23
## Columns: 2
## $ station <chr> "FINNER", "CLAREMORRIS", "DUBLIN AIRPORT", "ATHENRY", "SHANNON~
## $ data    <list> [<tbl_df[96 x 12]>], [<tbl_df[92 x 12]>], [<tbl_df[73 x 12]>]~
```

```
ds_n <- ds_n |> mutate(LM = map(data,
                        ~lm(AvrWindGen~wdsp, data = . )))
ds_n
```

```
## # A tibble: 23 x 3
##    station              data               LM
##    <chr>                <list>             <list>
##  1 FINNER               <tibble [96 x 12]> <lm>
##  2 CLAREMORRIS          <tibble [92 x 12]> <lm>
##  3 DUBLIN AIRPORT       <tibble [73 x 12]> <lm>
##  4 ATHENRY              <tibble [86 x 12]> <lm>
##  5 SHANNON AIRPORT      <tibble [94 x 12]> <lm>
##  6 VALENTIA OBSERVATORY <tibble [88 x 12]> <lm>
##  7 MOORE PARK           <tibble [98 x 12]> <lm>
##  8 MULLINGAR            <tibble [91 x 12]> <lm>
##  9 GURTEEN              <tibble [83 x 12]> <lm>
## 10 OAK PARK             <tibble [92 x 12]> <lm>
## # ... with 13 more rows
```

```
ds_n <- ds_n |>
  mutate(R_SQ = map_dbl(LM,
                        ~summary(.)$r.squared)) |>
  arrange(desc(R_SQ))
ds_n
```

```
## # A tibble: 23 x 4
##    station              data               LM      R_SQ
##    <chr>                <list>             <list> <dbl>
##  1 VALENTIA OBSERVATORY <tibble [88 x 12]> <lm>   0.690
##  2 KNOCK AIRPORT        <tibble [78 x 12]> <lm>   0.688
##  3 CORK AIRPORT         <tibble [88 x 12]> <lm>   0.665
##  4 SherkinIsland        <tibble [83 x 12]> <lm>   0.647
```

```
##  5 SHANNON AIRPORT      <tibble [94 x 12]> <lm>   0.644
##  6 MT DILLON            <tibble [88 x 12]> <lm>   0.630
##  7 ATHENRY              <tibble [86 x 12]> <lm>   0.616
##  8 CLAREMORRIS          <tibble [92 x 12]> <lm>   0.612
##  9 MULLINGAR            <tibble [91 x 12]> <lm>   0.603
## 10 ROCHES POINT         <tibble [86 x 12]> <lm>   0.583
## # ... with 13 more rows
```

```
ds_n <- ds_n |> mutate(Plots = map2(data, station,
                            ~ggplot(.x, aes(x=wdsp, y=AvrWindGen, colour=randomColor()))+
                              geom_point()+geom_smooth(colour=randomColor(luminosity = "light"))+
                              scale_colour_manual(values=randomColor(length(.y)))+
                              xlab("Speed")+ylab("Power")+
                              labs(title = .y)+
                              theme_classic()+theme(plot.title = element_text(size=6))
))
ds_n
```

```
## # A tibble: 23 x 5
##    station              data               LM      R_SQ Plots
##    <chr>                <list>             <list> <dbl> <list>
##  1 VALENTIA OBSERVATORY <tibble [88 x 12]> <lm>   0.690 <gg>
##  2 KNOCK AIRPORT        <tibble [78 x 12]> <lm>   0.688 <gg>
##  3 CORK AIRPORT         <tibble [88 x 12]> <lm>   0.665 <gg>
##  4 SherkinIsland        <tibble [83 x 12]> <lm>   0.647 <gg>
##  5 SHANNON AIRPORT      <tibble [94 x 12]> <lm>   0.644 <gg>
##  6 MT DILLON            <tibble [88 x 12]> <lm>   0.630 <gg>
##  7 ATHENRY              <tibble [86 x 12]> <lm>   0.616 <gg>
##  8 CLAREMORRIS          <tibble [92 x 12]> <lm>   0.612 <gg>
##  9 MULLINGAR            <tibble [91 x 12]> <lm>   0.603 <gg>
## 10 ROCHES POINT         <tibble [86 x 12]> <lm>   0.583 <gg>
## # ... with 13 more rows
```

```
ggarrange(plotlist = ds_n$Plots, legend = "none")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```