STAT346: Statistical Data Science I

HW#3 – Due: Saturday, October 31, 2020 by 6 p.m.

October 14, 2020

Instruction: Answer to the following questions and write your report using R Markdown. You should submit two files, through KU Black Board system (https://kulms.korea.ac.kr), which should have the following naming format:

- `stat346_hw3_your_id.rmd`
- `stat346_hw3_your_id.pdf` or `stat346_hw3_your_id.docx`

---

1. The `macleish` package contains weather data collected every 10 minutes in 2015 from two weather stations in Whately, MA, USA.

```
library(macleish)
head(whately_2015)
```

```
## # A tibble: 6 x 8
##   when                temperature wind_speed wind_dir rel_humidity pressure
##   <dttm>                    <dbl>      <dbl>    <dbl>        <dbl>    <int>
## 1 2015-01-01 00:00:00       -9.32       1.40     225.         54.6      985
## 2 2015-01-01 00:10:00       -9.46       1.51     248.         55.4      985
## 3 2015-01-01 00:20:00       -9.44       1.62     258.         56.2      985
## 4 2015-01-01 00:30:00       -9.3        1.14     244.         56.4      985
## 5 2015-01-01 00:40:00       -9.32       1.22     238.         56.9      984
## 6 2015-01-01 00:50:00       -9.34       1.09     242.         57.2      984
## # ... with 2 more variables: solar_radiation <dbl>, rainfall <int>
```

   Use `ggplot2` to create a data graphic that displays the average temperature over each 10-minute interval (`temperature`) as a function of time (`when`).

---

2. The `storms` data in the `nasaweather` package contains information about tropical cyclone tracks through the Atlantic Ocean, Caribbean Sea and Gulf of Mexico from 1995 to 2005.

```
library(nasaweather)
head(storms)
```

```
## # A tibble: 6 x 11
##   name    year month   day  hour   lat  long pressure  wind type         seasday
##   <chr>  <int> <int> <int> <int> <dbl> <dbl>    <int> <int> <chr>          <int>
```

```
## 1 Allis~  1995     6     3     0  17.4 -84.3     1005    30 Tropical De~       3
## 2 Allis~  1995     6     3     6  18.3 -84.9     1004    30 Tropical De~       3
## 3 Allis~  1995     6     3    12  19.3 -85.7     1003    35 Tropical St~       3
## 4 Allis~  1995     6     3    18  20.6 -85.8     1001    40 Tropical St~       3
## 5 Allis~  1995     6     4     0  22   -86       997    50 Tropical St~       4
## 6 Allis~  1995     6     4     6  23.3 -86.3     995    60 Tropical St~       4
```

a. Using `storms` data, create a scatterplot between `wind` and `pressure`, with color being used to distinguish the `type` of storm.

b. Use the `geom_path()` function to plot the path of each tropical storm in the `storms` data table. Use color to distinguish the storms from one another, and use facetting to plot each year in its own panel.

---

3. Two teams, A and B, are playing a seven game series. Team A is better than team B and has a $p > 0.5$ chance of winning each game. Given a value p, the probability of winning the series for the underdog team B can be computed with the following function based on a Monte Carlo simulation:

```
prob_win <- function(p){
  B <- 10000
  result <- replicate(B, {
     b_win <- sample(c(1,0), 7, replace = TRUE, prob = c(1-p, p))
     sum(b_win)>=4
  })
  mean(result)
}
```

a. Use the function `sapply` to compute the probability, call it `Pr`, of winning for `p <- seq(0.5, 0.95, 0.025)`. Then plot the result.

b. Repeat the exercise above, but now keep the probability fixed at `p <- 0.75` and compute the probability for different series lengths: best of 1 game, 3 games, 5 games,... Specifically, `N <- seq(1, 25, 2)`. Hint: use this function:

```
  prob_win <- function(N, p=0.75){
B <- 10000
result <- replicate(B, {
   b_win <- sample(c(1,0), N, replace = TRUE, prob =   c(1-p, p))
   sum(b_win)>=(N+1)/2
  })
  mean(result)
}
```

---

4. Introduction to Data Science – Exercise 15.7: #1-6 (link).
5. Introduction to Data Science – Exercise 18.6: #1-3 (link).
6. Introduction to Data Science – Exercise 19.6: #1-5 (link).