Zeru-Zhou-project02

September 1, 2021

1 Project 2 – Zeru Zhou

TA Help: NA

Collaboration: NA

- get some help from piazza questions
- get help from the videos provided by Dr. Ward

1.1 Question 1

```
[3]: stations <- read.csv("/depot/datamine/data/whin/stations.csv")
```

```
[2]: weather <- read.csv("/depot/datamine/data/whin/weather.csv")
```

```
[2]: head(stations)
```

A data.frame: 6 x 4

id	name	latitude	longitude
<int>	<chr>	<dbl>	<dbl>
142	WHIN052-MONT004	40.10483	-86.86619
143	WHIN053-PULA005	40.98224	-86.38542
151	WHIN059-CASS006	40.84436	-86.18173
20	WHIN020-FOUN001	40.27096	-87.14860
144	WHIN054-WHIT007	40.53722	-86.95342
163	WHIN072-FOUN005	40.16179	-87.35246

```
[14]: head(weather)
```

A data.frame: 6 x 26

station_id	latitude	longitude	name	observation_time	tempera
<int>	<dbl>	<dbl>	<chr>	<chr>	<dbl>
1	40.93894	-86.47418	WHIN001-PULA001	2019-07-10T04:00:00Z	70
1	40.93894	-86.47418	WHIN001-PULA001	2019-07-10T04:15:00Z	69
1	40.93894	-86.47418	WHIN001-PULA001	2019-07-11T04:00:00Z	76
1	40.93894	-86.47418	WHIN001-PULA001	2019-07-11T04:15:00Z	76
1	40.93894	-86.47418	WHIN001-PULA001	2019-07-11T04:30:00Z	76
1	40.93894	-86.47418	WHIN001-PULA001	2019-07-11T04:45:00Z	75

```
[3]: tail(stations)
```

A data.frame: 6 x 4

	id <int>	name <chr>	latitude <dbl>	longitude <dbl>
173	31	WHIN031-CASS005 EXT	40.78383	-86.33381
174	35	WHIN035-CASS004 Ivy Tech	40.73612	-86.35604
175	36	WHIN036-TIPP004	40.29861	-86.90033
176	41	WHIN041-TIPP006 Cumberland Gardens	40.46325	-86.91867
177	42	WHIN042-CARR002	40.54233	-86.48150
178	44	Pedestrian Bridge	40.41936	-86.89753

[15]:
```
tail(weather)
```

A data.frame: 6 x 26

	station_id <int>	latitude <dbl>	longitude <dbl>	name <chr>	observation_time <chr>
999995	171	40.2968	-87.39029	WHIN038E-WARR004	2021-07-10T06:15:00
999996	171	40.2968	-87.39029	WHIN038E-WARR004	2021-07-10T06:30:00
999997	171	40.2968	-87.39029	WHIN038E-WARR004	2021-07-10T06:45:00
999998	171	40.2968	-87.39029	WHIN038E-WARR004	2021-07-10T07:00:00
999999	171	40.2968	-87.39029	WHIN038E-WARR004	2021-07-10T07:15:00
1000000	171	40.2968	-87.39029	WHIN038E-WARR004	2021-07-10T07:30:00

[4]:
```
str(stations)
```

```
'data.frame':   178 obs. of  4 variables:
 $ id       : int  142 143 151 20 144 163 166 145 153 164 …
 $ name     : chr  "WHIN052-MONT004" "WHIN053-PULA005" "WHIN059-CASS006"
"WHIN020-FOUN001" …
 $ latitude : num  40.1 41 40.8 40.3 40.5 …
 $ longitude: num  -86.9 -86.4 -86.2 -87.1 -87 …
```

[20]:
```
str(weather)
```

```
'data.frame':   1000000 obs. of  26 variables:
 $ station_id                : int  1 1 1 1 1 1 1 1 1 1 …
 $ latitude                  : num  40.9 40.9 40.9 40.9 40.9 …
 $ longitude                 : num  -86.5 -86.5 -86.5 -86.5 -86.5 …
 $ name                      : chr  "WHIN001-PULA001" "WHIN001-PULA001"
"WHIN001-PULA001" "WHIN001-PULA001" …
 $ observation_time          : chr  "2019-07-10T04:00:00Z"
"2019-07-10T04:15:00Z" "2019-07-11T04:00:00Z" "2019-07-11T04:15:00Z" …
 $ temperature               : num  70 69 76 76 76 75 75 74 74 74 …
 $ temperature_high          : num  71 70 77 76 76 76 75 75 74 74 …
 $ temperature_low           : num  70 69 76 76 76 75 75 74 74 74 …
 $ humidity                  : num  83 84 76 77 77 79 80 81 81 81 …
 $ solar_radiation           : num  NA NA NA NA NA NA NA NA NA NA …
 $ solar_radiation_high      : num  NA NA NA NA NA NA NA NA NA NA …
 $ rain                      : num  0 0 0 0 0 0 0 0 0 0 …
 $ rain_inches_last_hour     : num  0 0 0 0 0 0 0 0 0 0 …
 $ wind_speed_mph            : num  0 1 2 2 2 2 1 2 2 3 …
 $ wind_direction_degrees     : num  NA 248 202 202 225 …
 $ wind_gust_speed_mph        : num  3 3 4 4 4 3 3 4 4 4 …
 $ wind_gust_direction_degrees: num  248 248 202 202 202 …
 $ pressure                   : num  30.1 30 29.9 29.9 29.9 …
 $ soil_temp_1                : num  77 76 80 80 80 79 79 79 79 79 …
 $ soil_temp_2                : num  78 78 80 80 80 80 79 79 79 79 …
 $ soil_temp_3                : num  76 76 78 78 78 77 77 77 77 77 …
 $ soil_temp_4                : num  74 74 75 75 75 75 75 75 75 75 …
 $ soil_moist_1               : num  24 24 31 31 32 31 32 32 32 32 …
 $ soil_moist_2               : num  24 25 30 31 31 31 31 31 31 31 …
 $ soil_moist_3               : num  10 10 12 12 12 12 12 12 12 12 …
 $ soil_moist_4               : num  9 9 10 10 10 10 10 10 10 10 …
```

[5]:
```
names(stations)
```

1. 'id' 2. 'name' 3. 'latitude' 4. 'longitude'

[17]:
```
names(weather)
```

1. 'station_id' 2. 'latitude' 3. 'longitude' 4. 'name' 5. 'observation_time' 6. 'temperature' 7. 'temperature_high' 8. 'temperature_low' 9. 'humidity' 10. 'solar_radiation' 11. 'solar_radiation_high' 12. 'rain' 13. 'rain_inches_last_hour' 14. 'wind_speed_mph' 15. 'wind_direction_degrees' 16. 'wind_gust_speed_mph' 17. 'wind_gust_direction_degrees' 18. 'pressure' 19. 'soil_temp_1' 20. 'soil_temp_2' 21. 'soil_temp_3' 22. 'soil_temp_4' 23. 'soil_moist_1' 24. 'soil_moist_2' 25. 'soil_moist_3' 26. 'soil_moist_4'

[6]:
```
dim(stations)
```

1. 178 2. 4

[18]:
```
dim(weather)
```

1. 1000000 2. 26

[7]:
```
summary(stations$id)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   54.25   98.50   99.29  142.75  197.00 
```

[19]:
```
summary(weather$rain)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
  0.00000   0.00000   0.00000   0.08923   0.00000 101.00000 
```

Code and outputs are listed above, including read.csv function and some functions like head(), tail(), dim(), summary(), str(), and names(). Answering questions: The dimension of dataset “stations” is 178 rows and 4 columns. The dimension of dataset “weather” is 1000000 rows and 26 columns. The first 5 rows are listed above in the code: head(stations) and head(weather). The column names are displayed above in the code: names(stations) and names(weather).

1.2 Question 2

[11]:
```
temp <- weather$temperature
```

[12]:
```
head(temp)
```

1. 70 2. 69 3. 76 4. 76 5. 76 6. 75

[13]:
```
temp[100]
```

63

[14]:
```
tail(temp)
```

1. 64 2. 64 3. 64 4. 64 5. 64 6. 64

[15]:
```
typeof(temp)
```

'double'

[16]:
```
class(temp)
```

'numeric'

Code and output are displayed above. The first value in the vector temp is 70; the 100th value is 63; the last value is 64. The type of data in the vector is Double data type. The class of data is numeric.

1.3 Question 3

[6]:
```
temp100 <-␣
 ↪head(weather$rain_inches_last_hour,n=100)+tail(weather$rain_inches_last_hour,n=100)
```

One line code is above code, since I see we do not need to print temp100 on piazza, we do not have output through this one line code. It only add them together and form a new vector.

1.4 Question 4

[5]:
```
Sub <- subset(weather, station_id == 20)
```

[6]:
```
hot_temps <- Sub$temperature[Sub$temperature >= 85]
```

[7]:
```
length(hot_temps)
```

909

[8]:
```
head(hot_temps)
```

1. <NA> 2. 85 3. 85 4. 86 5. 87 6. 87

[9]:
```
cold_temps <- Sub$temperature[Sub$temperature <= 40]
```

```
[10]: length(cold_temps)
```

20627

```
[11]: head(cold_temps)
```

1. <NA> 2. 40 3. 39 4. 39 5. 38 6. 38

```
[13]: head(hot_temps+cold_temps)
```

```
Warning message in hot_temps + cold_temps:
"longer object length is not a multiple of shorter object length"
```

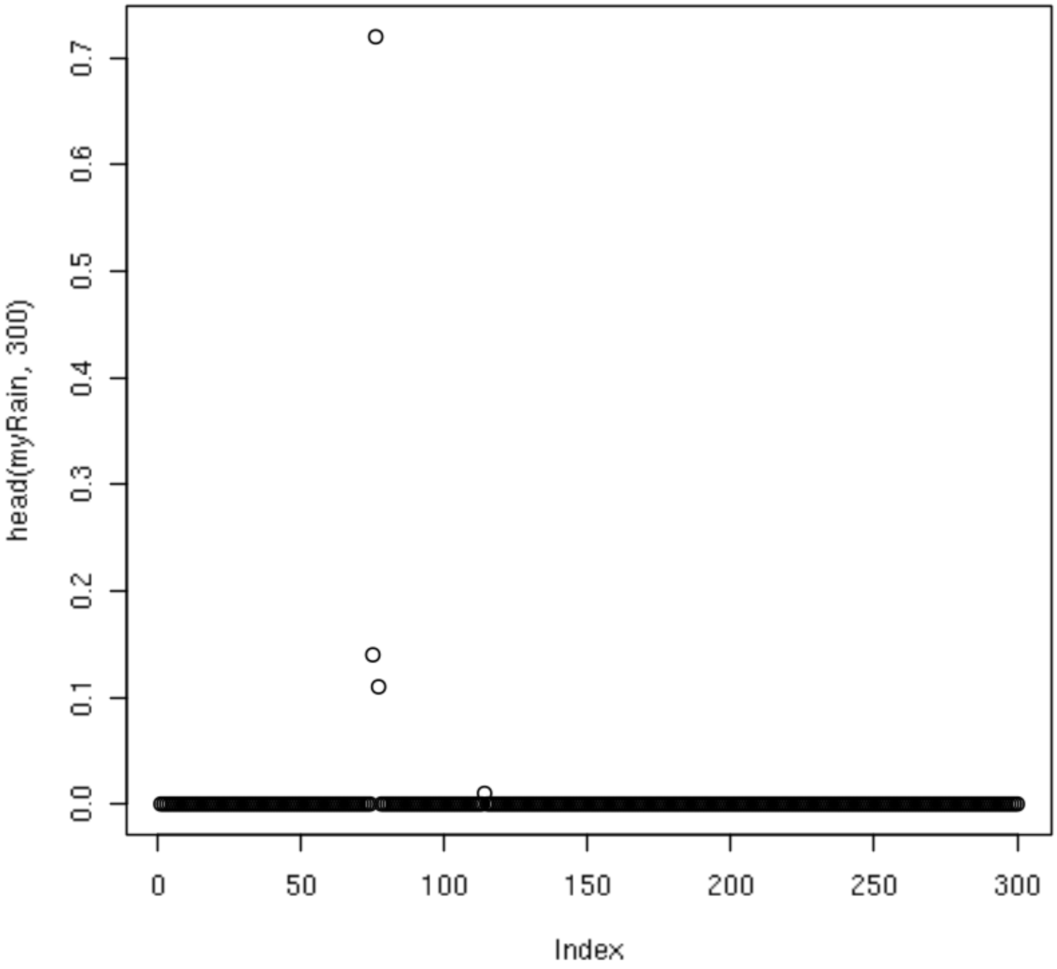
1. <NA> 2. 125 3. 124 4. 125 5. 125 6. 125

Hot_temps and cold_temps are created above. There are 909 elements in hot_temps and 20627 elements in cold_temps. If I add them together, an error occurs : “longer object length is not a multiple of shorter object length”. This is because when two vector are added, the shorter one would be recycled until it matches to the length of the longer vector.

1.5 Question 5

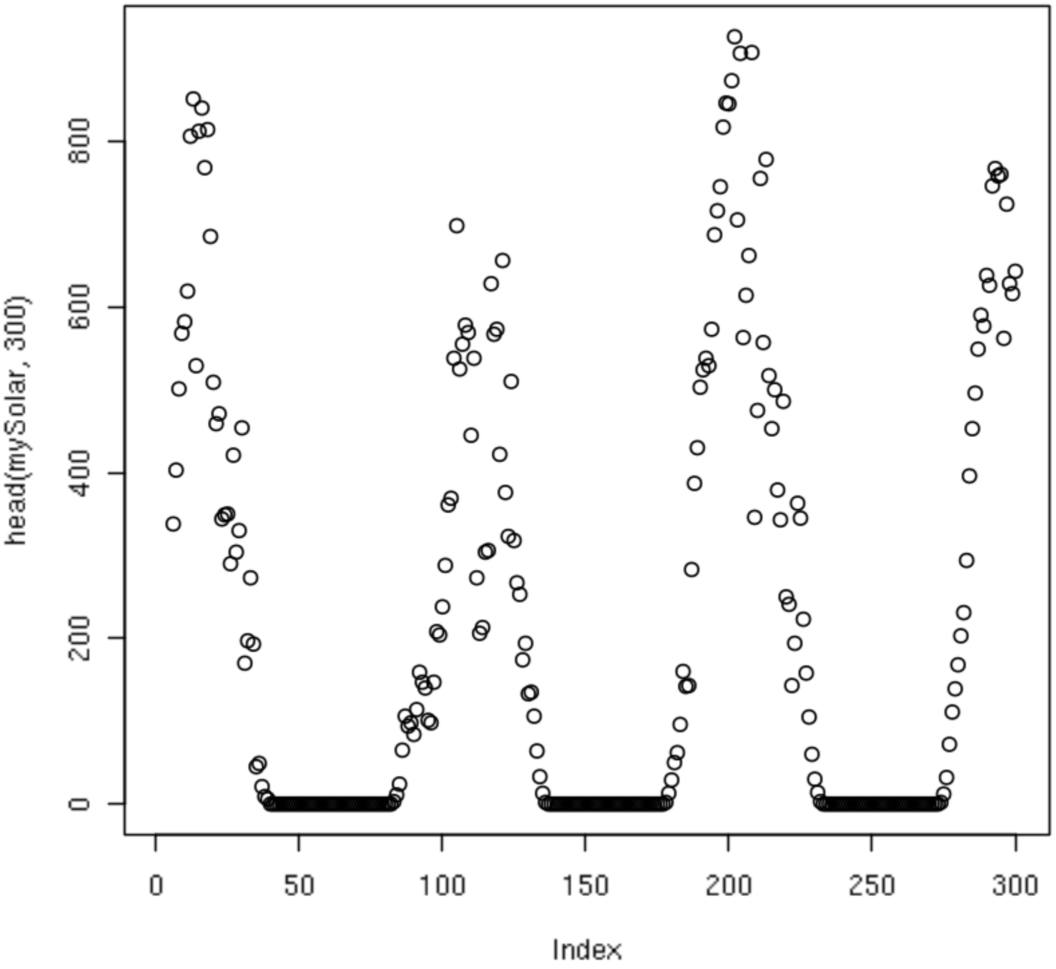
```
[14]: myRain <- weather$rain[weather$station_id == 20]
```

```
[23]: plot(head(myRain,300))
```



[20]:
```
mySolar <- weather$solar_radiation[weather$station_id == 20]
```

[22]:
```
plot(head(mySolar,300))
```



I tried to plot on station_id=20 and column rain and solar_radiation seperately. For plot for column rain, the pattern is that rain does not deviate from 0 with the change in index. There are only few cases that rain is not 0, and I think they could be outliers. For plot for column solar_radiation, the pattern is fluctuating with index. Going down first, and remain at solar_radiation=0 for around 50 indexs, then going up and repeat this procedure for many times as index move forward.

1.6 Question 6

Plot 3 is my favorite graphic. This is because it is easily to discern the trend of each station ID seperately and not getting confounded like the dot plot with color above(plot 2). I think one way to improve the graphic 3 is to let it describe more data. This graphic only include data from 2019-07 to 2020-12 but we can definitely include more! One thing interesting is that I find that the graphic messed up at Date 2020-03 to 2020-07. Maybe there are someway to avoid this phenomenon.

1.7 Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

> As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together – We are Purdue.