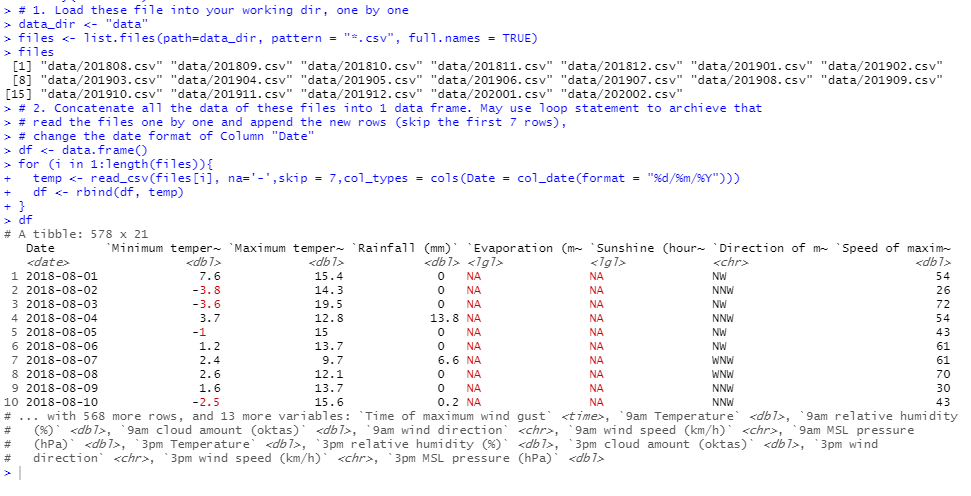
**Introduction to Data Scientist 11372 (UG)**

**Assignment 1 – Data Wrangling and Exploration**

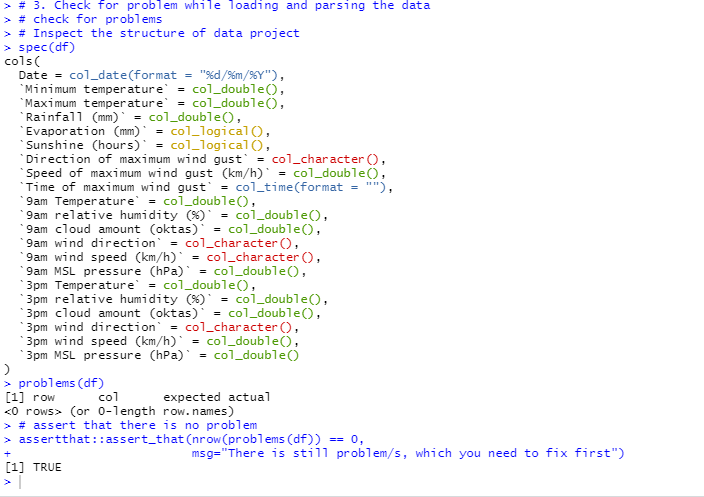
Tuan Anh (Vincent) Nguyen – u3196825

# Part A:

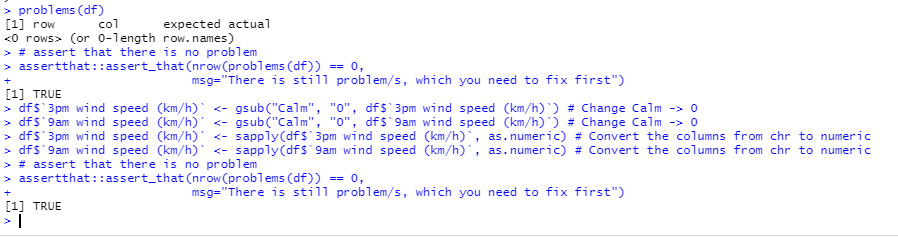
Q1 + Q2:



Q3:



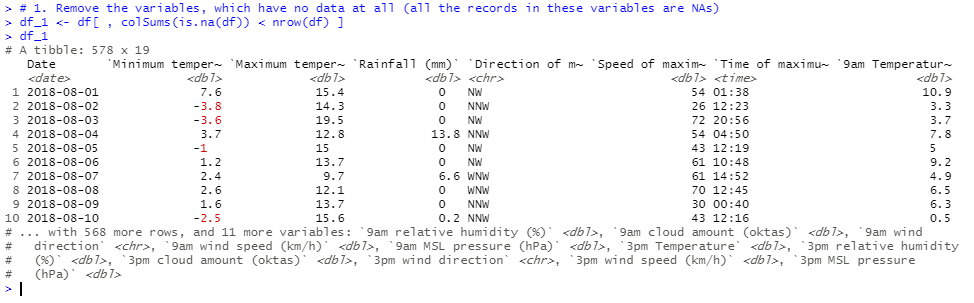
Note: It’s been quite weird after I just my code a little bit it did not shown any problems with the input data, but I know it is still there. As `9 am wind speed` supposed to be numeric not char.



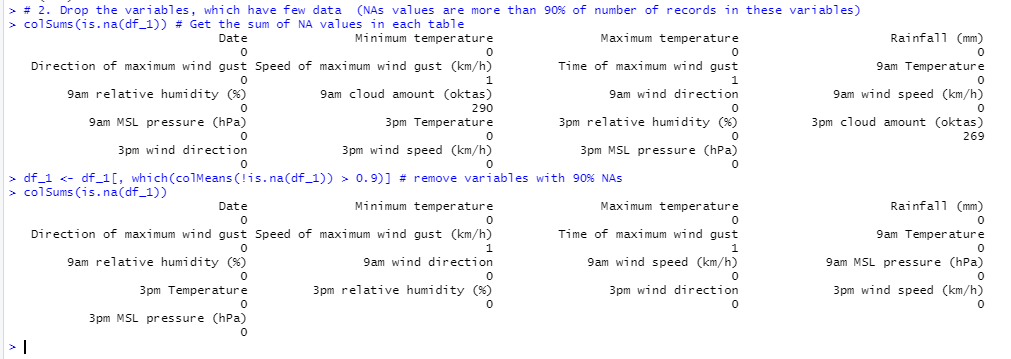
This code used to change all “Calm in these 2 columns as 0, according to the reference document from BOM, ref: <http://www.bom.gov.au/marine/knowledge-centre/reference/wind.shtml>

# Part B:

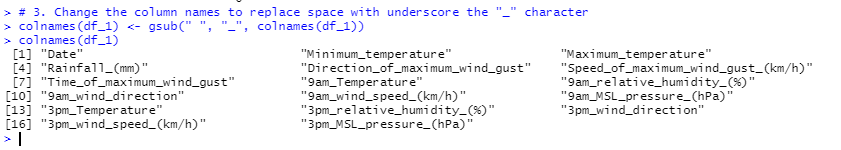
Q1: Remove from 21 columns down to 19 columns.



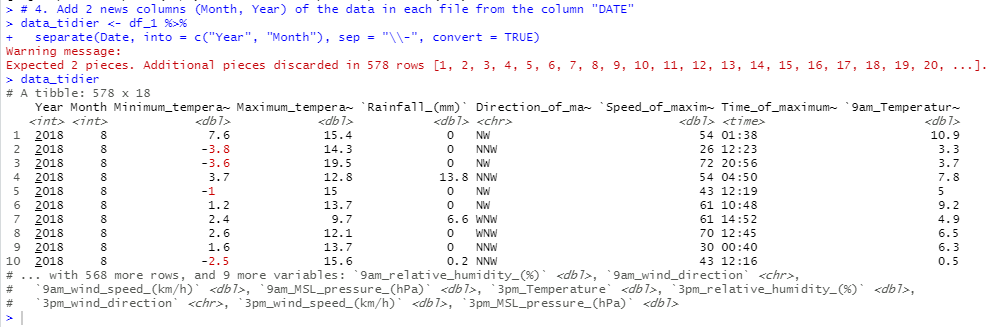
Q2:



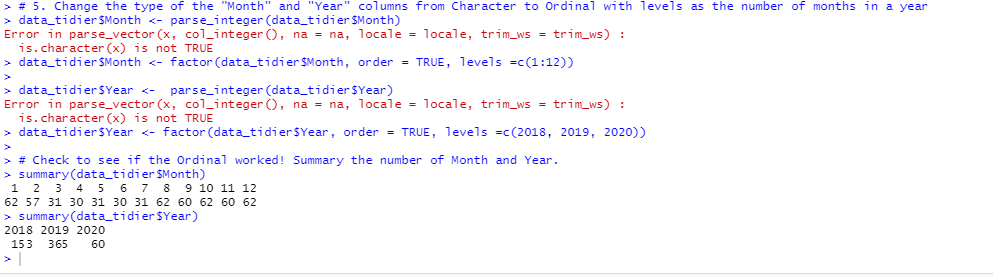
Q3:



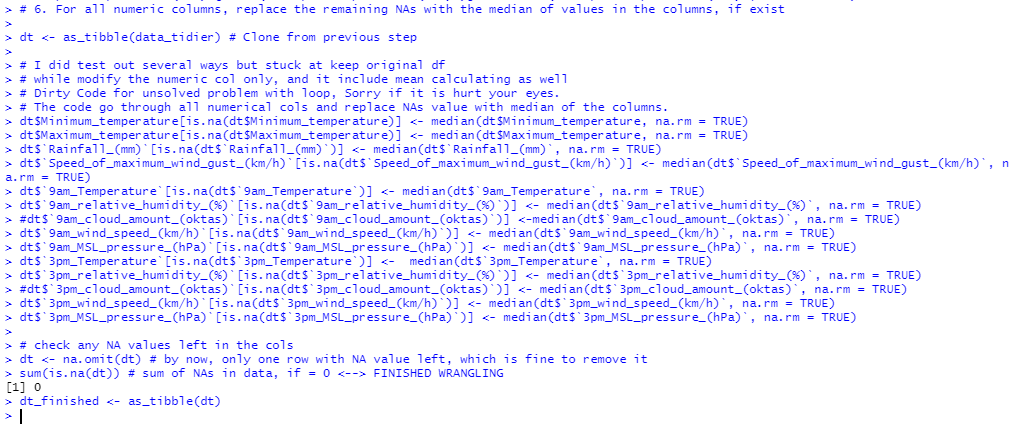
Q4 + 5: Note (The Date Columns already been converted to Date while read\_csv step already)



Q6:



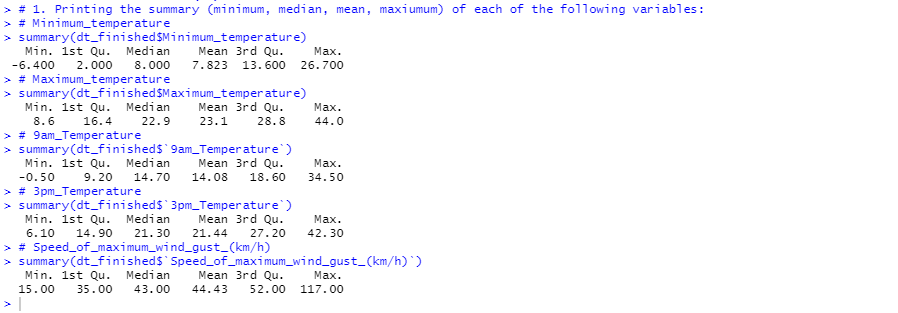
Q7:



Quick note for 6: As for Q7 I supposed to have a loop, but for some weird reasons I could not make it work after several hours attempt, so I decided to do it manually to move forward as it took me almost a day to find a way looping through all numeric columns and replace NAs value with median of the columns.

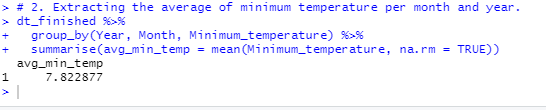
# Part C:

Q1:

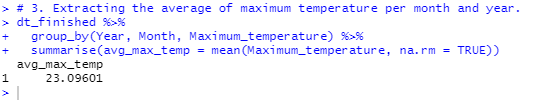


Note: I saw some example have to write the code separately but I found this way more efficient, not sure it is alright to use this interface?

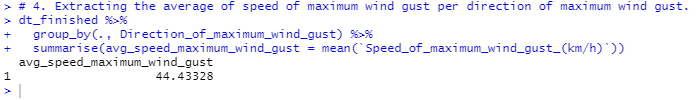
Q2:



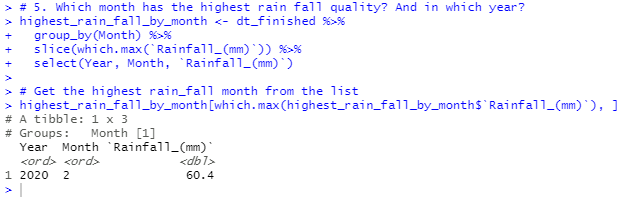
Q3:



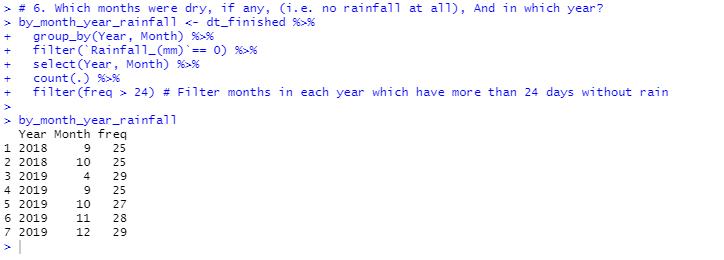
Q4:



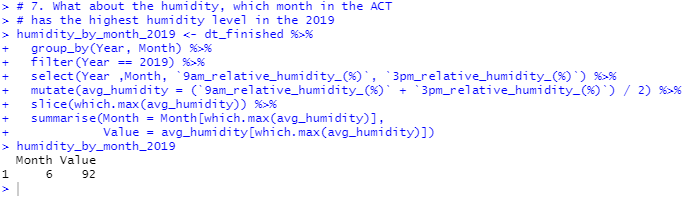
Q5:



Q6:



Q7:



For this part, I group the year and month then filter to only 2019. Then get the average humidity between 9am and 3pm to then get the maximum of each month then summarise and get the month with most humidity averaging.

Q8:

Note: The code here quite straight forward to get the Min, Max and Average of Temperature, Wind Speed and Humidity per month and quarter so I first have to create a new Quarter table with Ordinal levels of 4 for easier filter and summarise. However, there was 2 small problems with Wind speed and Humidity as it has 2 columns of data for 9am and 3pm so I have to calculate the mean to get the min and max per month and quarter. Which can be quite messy if try to find the Min and Max from each different row within a Month or Year.

The code was really redundant, I have to admit as by the time I going to submit this, I don’t have enough time to write function and loop through this to make it less eye-killer.





