Assignment – 1
Data Wrangling and Exploration

Due on Sunday, 18 April (23:55)

Motivation

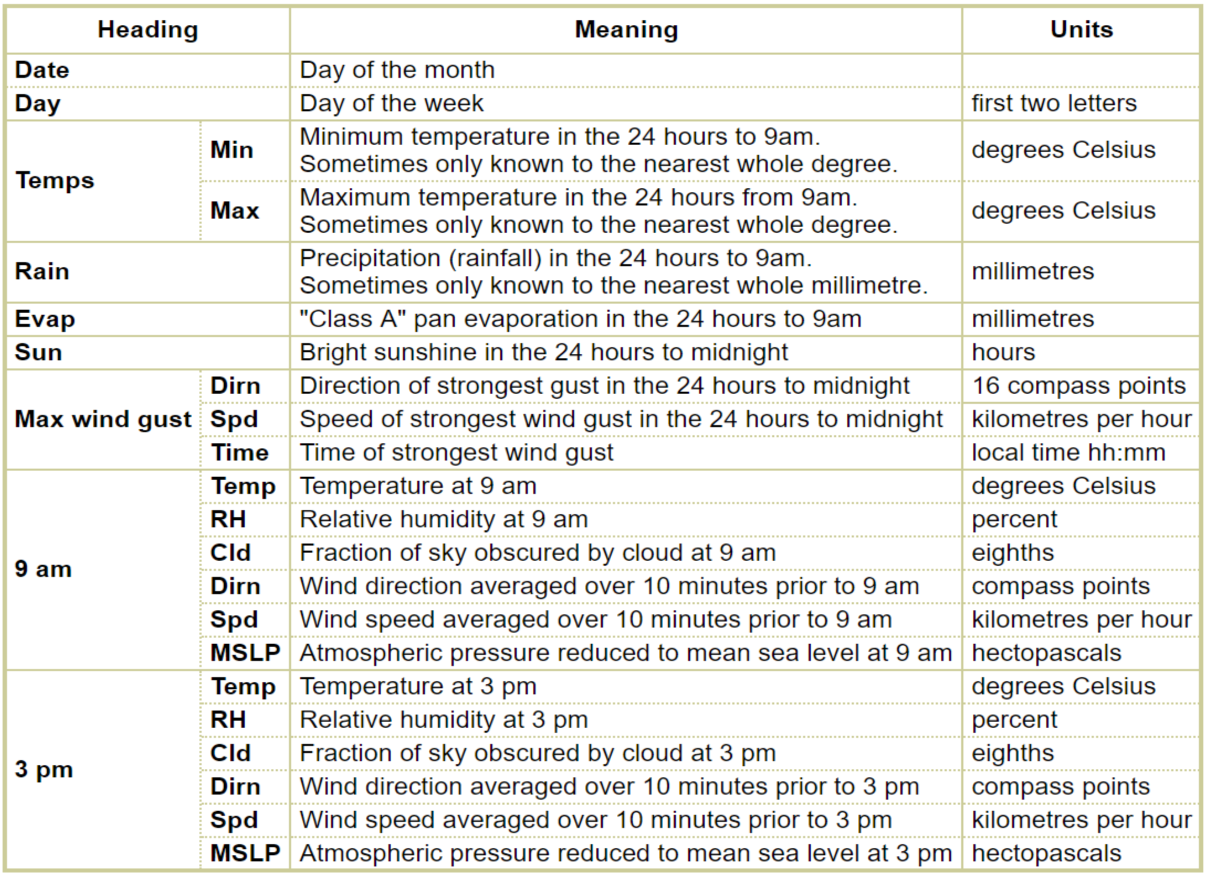
The purpose of this assignment is to assess your skills on reading multiple data files, merging them into a single data frame, applying different cleaning and wrangling steps and then getting the data ready for the modelling.

Data Description

The observations in the attached CSV files have been taken from the Bureau of Meteorology's "real time" system. These observations provide some details about the weather in the Australian Capital territory for 19 months. Most of the data are generated automatically. Some quality checking has been performed, but it is still possible for erroneous values to appear. Sometimes, when the daily maximum and minimum temperatures, rainfall or evaporation are missing, the next value given has been accumulated over several days rather than the normal one day.

There are 19 comma-separated data files provided with this assignment. These data are for the months from August 2018 to February 2020. The variables reported in each file are described in Table 1.

Table 1 Column Meanings



Heading		Meaning	Units
Date		Day of the month	
Day		Day of the week	first two letters
Temps	**Min**	Minimum temperature in the 24 hours to 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
	Max	Maximum temperature in the 24 hours from 9am. Sometimes only known to the nearest whole degree.	degrees Celsius
Rain		Precipitation (rainfall) in the 24 hours to 9am. Sometimes only known to the nearest whole millimetre.	millimetres
Evap		"Class A" pan evaporation in the 24 hours to 9am	millimetres
Sun		Bright sunshine in the 24 hours to midnight	hours
Max wind gust	**Dirn**	Direction of strongest gust in the 24 hours to midnight	16 compass points
	Spd	Speed of strongest wind gust in the 24 hours to midnight	kilometres per hour
	Time	Time of strongest wind gust	local time hh:mm
9 am	**Temp**	Temperature at 9 am	degrees Celsius
	RH	Relative humidity at 9 am	percent
	Cld	Fraction of sky obscured by cloud at 9 am	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 9 am	compass points
	Spd	Wind speed averaged over 10 minutes prior to 9 am	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 9 am	hectopascals
3 pm	**Temp**	Temperature at 3 pm	degrees Celsius
	RH	Relative humidity at 3 pm	percent
	Cld	Fraction of sky obscured by cloud at 3 pm	eighths
	Dirn	Wind direction averaged over 10 minutes prior to 3 pm	compass points
	Spd	Wind speed averaged over 10 minutes prior to 3 pm	kilometres per hour
	MSLP	Atmospheric pressure reduced to mean sea level at 3 pm	hectopascals

Copyright of the Data

Tasks

Part A, Reading (For UG and G students) (15 marks)

First, you have 19 CSV files and you need to write R code to:

1- Load these files into your working directory, one by one. (8 marks)

2- Concatenate all the data of these files into one data frame. You may use loop statement to achieve that. It is recommended to use the functions from the `tidyverse` package to read and to import these files. However, using any of the other functions are also fine. (5 marks)

3- Check for problems while loading and parsing the data. (2 marks)

Please note that, appendix-I shows a template code for the structure and the steps of part-A. You can follow it or create your own code structure.

Part B – Preparing (For UG and G students) (15 marks)

Write code to do the following tasks:

1- Remove the variables, which have no data at all (*i.e.* all the records in these variables are NAs) (2 marks)

2- Drop the variables, which have few data (i.e. NAs values are more than 90% of number of records in these variables). (2 marks)

3- Change the column names to have no spaces between the words and replace these spaces with underscore the '_' character. (2 marks)

4- Change the type of the column called "Date" from character to Date data type. (2 marks)

5- Add two new columns for the month and year of the data in each file, you may extract the contents of this column from the "Date" column. Please note that the data are collected for 19 months across 3 years (2018, 2019 and 2020). (3 marks)

6- Change the type of the "Month" and "Year" columns from Character to Ordinal with levels as the number of months in a year (i.e. 12) and number of years (3). (2 marks)

7- For all the numeric columns, replace the remaining NAs with the median of the values in the column, if exist. (2 marks)

Part C – Analysing (For UG students only) (20 marks)

Write code to do the following tasks:

1. Printing the summary (i.e. minimum, median, mean, maximum) of each of the following variables: (3 marks)
   a. `Minimum_temperature`,
   b. `9am_Temperature`,
   c. ` Speed_of_maximum_wind_gust_(km/h)`
2. Extracting the average of minimum temperature per month and year (3 marks)
3. Extracting the average of speed of maximum wind gust by direction of maximum wind gust (3 marks)
4. Which month was dry, if any, (i.e. no rainfall at all)? And in which year? (3 marks)
5. What about the humidity, which month in the ACT has the highest humidity level in 2019? (3 marks)
6. For 2019, extract the minimum, maximum and average temperature, wind speed and humidity per month and per quarter in 2019 only. (3 marks)
7. Plot the histograms/bar-charts for each variable of the previous question. (2 marks)

Part C – Analysing (For G students only) (20 marks)

Write code to do the following tasks:

1. Printing the summary (i.e. minimum, median, mean, maximum) of each of the following variables: (2.5 marks)
   a. `Minimum_temperature`,
   b. `Maximum_temperature`,
   c. `9am_Temperature`,
   d. `3pm_Temperature` and
   e. ` Speed_of_maximum_wind_gust_(km/h)`.
2. Extracting the average of minimum temperature per month and year (2 marks)
3. Extracting the average of maximum temperature per month and year (2 marks)
4. Extracting the average of speed of maximum wind gust per direction of maximum wind gust (2 marks)
5. Which month has the highest rain fall quantity? And in which year? (2 marks)
6. Which months were dry, if any, (i.e. no rainfall at all)? And in which year? (2 marks)
7. What about the humidity, which month in the ACT has the highest humidity level in the 2019 year? (2.5 marks)
8. For 2019, extract the minimum, maximum and average temperature, wind speed and humidity per month and per quarter in 2019 only. (3 marks)
9. Plot the histograms/bar-charts for each variable of the previous question. (2 marks)

Part D – Insights (For UG students only) (5 marks)

As a data scientist, you need to practise extracting insights and valuable information from the analysis you conduct on the data. This can be done by raising some questions that can be answered by doing this analysis. Questions such as, "Based on the weather analysis, what is the best time of the year, in which you would recommend people living outside ACT to come and to visit it?"

Can you list at least **2** questions that can be answered by running analysis on this data set?

Part D – Insights (For G students only) (5 marks)

As a data scientist, you need to practise extracting insights and valuable information from the analysis you conduct on the data. This can be done by raising some questions that can be answered by doing this analysis. Questions such as, "Based on the weather analysis, what is the best time of the year, in which you would recommend people living outside ACT to come and to visit it?"

Can you list at least **4** questions that can be answered by running analysis on this data set?

Deliverables

You are required to submit a compressed (e.g. ZIP) file to Canvas with the following two files:

1- Single R file with the code for the three first parts; Part A, Part B, and Part C.

2- A PDF document with the output on the Console window of running your code part by part.

3- A PDF document with the questions that you have generated for part D.

Please follow the following structure to name the submitted zip file:

[studentID_lastname_assignment1.zip]

Replace the studentID with your university ID and surname with your lastname.

Appendix-I: Code Template

This template is just an example of the code structure; you may change it completely to do the above mentioned tasks.

```
# Unit name and Id
# Student name and Id
# Description of what this code id for.

setwd("replace with the path of your working directory")

#################### Part A ####################
# data files (you need to specify the paths of the CSV files (e.g. relative
or absolute) )
files <- c("data/201808.csv",
           ...,
           ...,
           ...,
           ...,
           ...,
           ...,
           ...,
           ...,
           ...,
           ...,
           ...,
           ...
           )

# install tidyverse if it is not installed
...

# load the tidyverse library
...
# read the files one by one and append the new rows, you may skip the first
7 rows as they are meta-deta
for (i in 1:length(files)){
  ...
}

# inspect the structure of data object

# you may view the data with R studio viewer
view(data)

# check for problems
...

# assert that there is NO problems
assertthat::assert_that(nrow(problems(data)) == 0,
                        msg="There is still problem/s, which you need to
fix first")
# print data dimensions
...
# Extract the completed records only (i.e. the records without at least one
NA)
...
#################### Part B ####################
...
#################### Part C ####################
...
```

Appendix-II: Assessment Criteria

To understand how your assignment will be marked and to know the points that you need to consider while doing the assignment, please have a look to the "*marking_guide.pdf*" file that is attached on Canvas with this assignment.