Assignment – 1
Data Wrangling and Exploration

Motivation

The purpose of this assignment is to assess your skills on reading multiple data files, merging these files into a single data frame, understanding these data by 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 the last year. 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 13 comma-separated data files provided with this assignment. These data are for the months from August 2018 to August 2019. 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 (15 marks)

First, you have 13 csv files and you need to write R code to

1- Load these into your working directory, one by one.
2- Concatenate all of the records 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` library to read and to load the files. However, using any of the other functions are also fine.
3- Check for problems while loading and parsing the data.

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 (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- Drop the variables, which have few data (i.e. NAs values are more than 90% of number of records in these variables).
3- Change the column names to have no spaces between the words and replace these spaces with underscore the '_' character.
4- Change the type of the column called "Date" from character to Date data type.
5- Add a new column and name it "Month", you may extract the contents of this column from the "Date" column.
6- Change the type of "Month" column from Character to Ordinal with levels as the number of months (i.e. 13)
7- For all of the numeric columns, replace the remaining NAs with the median value of the values in the column.

Part C – Analysing (15 marks)

Write code to do the following tasks:

1. Show the summary (i.e. min, 1st Qu., median, mean, 3rd Qu., max) of each of the following variables:
    a. ` Minimum_temperature`,
    b. `Maximum_temperature`,
    c. `9am_Temperature`,
    d. `3pm_Temperature` and
    e. ` Speed_of_maximum_wind_gust_(km/h)`.
2. Extract the mean of minimum temperature by month
3. Extract the mean of maximum temperature by month
4. Extract the mean of speed of maximum wind gust by direction of maximum wind gust
5. Which month has the highest rain fall quantity?
6. Which months were dry, if any, (i.e. no rainfall at all)?
7. What about the humidity, which month in the ACT has the highest humidity level in the last year?

Part D – Insights (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 that you recommend people living outside ACT to come and to visit it?"

Can you list at least **three** 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 questions that you have generated for part D.

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.