SCC.461: Coursework 3

Clement Lee

2021/2022

Before you start

- Read the instructions carefully. Failure to follow them will result in mark deduction.
- Submit both the knitted/generated file in pdf which shows your code and output, and the source code i.e. the .Rmd file (but changed to .txt for the purpose of submission). **Do not submit a word/html document, or a pdf printout of a word/html document**.
- The coursework is marked out of 20. One mark is allocated for submitting both the code and the generated document of the correct format.
- Include your 8-digit student ID (library card number) in the `author:` field at the top of your document when creating your knitted report.
- Deadline for coursework submission is **2021-11-08 (Monday) 9am (UK time)**.

Introduction

The Severe Storms Archive of the Australian bureau of Meteorology (http://www.bom.gov.au/australia/stormarchive) contains data relating to recorded Severe Thunderstorm and related events in Australia dating back to the 18th Century. In the file `Australian_severe_storms_1975-2015.csv` (on Moodle), all Australian storm events in a 40 year period from 28^{th} September 1975 to 28^{th} September 2015 are recorded. The data consists of a row for each severe weather event and 14 columns:

Table 1: List of variables in the Australian severe storm data set.

Variables	Description
`Event.ID`	Unique event ID
`Database`	Type of severe storm event
`ID`	A secondary ID variable
`Date.Time`	Date and time in the format `dd/mm/yyyy hh:mm`
`Nearest.town`	Nearest town to the event
`State`	Australian state in which the event was recorded
`Latitude`	Latitude of the event
`Longitude`	Longitude of the event
`Comments`, `X`, `X.1`, `X.2`, `X.3`, `X.4`	Six columns of comments

Australia has three times zones. Australian Eastern Standard Time (AEST), Australian Central Standard Time (ACST) and Australian Western Standard Time (AWST). These time zones are 10 hours, 9.5 hours and 8 hours ahead of Coordinated Universal Time (UTC), respectively.

- AEST applies to: New South Wales except Yancowinna County, which includes the city of Broken Hill; Victoria; Queensland; Tasmania; and the Australian Capital Territory. During daylight saving time, AEST becomes Australian Eastern Daylight Time (AEDT) and clocks are advanced to UTC+11. Note that Queensland does not observe daylight saving time.
- ACST applies to: South Australia; the Northern Territory; and Yancowinna County, which includes the city of Broken Hill, in New South Wales. During daylight saving time, ACST becomes Australian

Central Daylight Time (ACDT), and clocks are advanced to UTC +10:30. The Northern Territory does not observe daylight saving time.

- Finally, AWST applies to Western Australia.

Tasks

Download the data set, `Australian_severe_storms_1975-2015.csv`, from the Moodle page. Submit a compiled report that performs the following tasks. Unless specified otherwise, use the modified data frame obtained in Question i to answer Question $i + 1$.

1. Read in the data and print the dimensions of the Storm Events data frame. [**1 mark**]

2. Clean the data by removing the variable `ID` and also Waterspout events from the database. Print the dimensions of the cleaned data frame. Also print the first few rows **without the 6 columns of comments**, without creating an intermediate data frame. [**1 mark**]

3. Add a column to your data frame containing the time zone of each event using the following `OlsonNames()` classifications. [**3 marks**]

- Hint: Write a function to allocate a time zone based upon the input of `State` and `Nearest.town`. Use a `for` loop and the time zone allocation function to assign a time zone to each event.
Note: `Nearest.town` is only relevant to New South Wales, and records with "Broken Hill Airport", "Broken Hill Area" are all relevant to "Broken Hill". Also, some records might be in full UPPERCASE, and some records might **contain** "Broken Hill" (or any uppercase / lowercase variations) instead of being an exact match.

Table 2: List of `OlsonNames()` classifications of time zones.

State	Abbreviation	Time Zone
Queensland	QLD	Australia/Queensland
New South Wales (exc. Broken Hill)	NSW	Australia/NSW
Broken Hill, New South Wales	NSW (Broken Hill)	Australia/Broken_Hill
Victoria	VIC	Australia/Victoria
South Australia	SA	Australia/South
West Australia	WA	Australia/West
Tasmania	TAS	Australia/Tasmania
Northern Territory	NT	Australia/North
Australian Capital Territory	ACT	Australia/ACT

4. Parse the date, time and time zones from the necessary columns to create a new variable in the data frame which converts the time into UTC. You may need the function `lubridate::as_datetime()` and/or the use of loops. Print the first few rows of the resultant data frame, without the 6 columns of comments, again without creating an intermediate data frame. [**1 mark**]

5. Create new variables for the month and year of each event. Print the first few rows of the resultant data frame, without the 6 columns of comments and without creating an intermediate data frame. [**1 mark**]

6. After discarding Waterspout events there are five types of events left in the data; Rain, Hail, Lighting, Wind, and Tornado.

i) Create a new data frame which contains the total number of counts for each of the above type of events for each of the twelve months over the forty year period. You may need to use the function `dplyr::count()`.

ii) On a single plot, plot the total number of counts of each event against month. Use the R object `month.abb` for the labels of the months in the plot. [**3 marks**]

7. From the answer to Question 5, the 6 columns titled `Comments, X, X.1, X.2, X.3, X.4` consist of comments.

i) Combine the comments from these columns into a single column, named `All.comments`.

ii) Select the following columns to keep for further analysis, `Event.ID`, `Database`, `State`, `All.comments`, and the year variable you created.

iii) After which you should add the following command to your script: `print(sapply(DF, class))` where `DF` is the name of the data frame. [**3 marks**]

8. Now we use the answer to Question 7(ii) for further analysis.

i) Create an indicator variable which states whether or not a storm event has resulted in a flash flood. Make sure you sort out all terms relating to flash floods.

ii) Print a plot of the number of flash floods per year from 1975-2015. You may need to first create a vector or data frame to contain the number of flash floods per year. [**3 marks**]

9. For severe wind events often the wind speed is given. The wind speed is given in knots or km/h.

i) Extract all wind speeds both those in knots and km/h. Hint: Knots can be abbreviated by kts or kt. Also note that wind speed can be a single, double or triple digit number.

ii) Convert km/h wind speeds to knots (1 knot = 1.852 km/h) rounding the wind/speed to the nearest knot. Hint: It is helpful to work with a reduced data frame which includes only those observations with a wind speed recorded.

iii) Print a boxplot of the wind speeds recorded per state. [**4 marks**]