CA2.1

The data for this assignment is contained in the file "CA2-G1.xls". This file contains the following variables:

Timestamp – Time and date of reading

Batch 1, Batch 2, … – pH readings for several batches of product.

Temperature (^{0}C) – Temperature measured in degrees Celsius

Pressure (bar) – Pressure measured in bar

Conductivity (mS/cm) – Conductivity measured in mS/cm

Concentration (mg/mL) – Concentration of product measured in mg/mL

Feed – denotes which type of feed is used (1 or 2)

Hose – denotes which hose was used (A, B, C, or D)

1. Using the readxl package or otherwise, read this file into R and generate a data frame.

2. Convert the timestamp to a POSIXct time/date variable in R using the appropriate function.

3. The pH readings for each batch are given in different columns, with missing values throughout the dataframe. Change this dataframe so that it contains only one column (with no missing values) describing the pH reading and an extra column denoting the batch; i.e. simplify the dataframe so that all pH values are in one column, with another column showing the batch number. You may find the functions tidyr::pivot_longer and tidyr::drop_na useful here.

4. Ensure that all dataframe variables are in the correct format and no errors have occurred in steps 2 or 3.

5. Create a new column in this dataframe that shows the time interval in minutes between successive readings. Hence, create another new column that gives the time in minutes since the first data point; this is essentially how long the vessel has been active.

6. Use the dplyr::group_by and dplyr::summarise functions to make a table showing the following summary statistics for each batch: maximum pH, minimum pH, mean pH, median pH, standard deviation of pH and the number of datapoints for the batch.

Continued overleaf

7. Create a timeseries plot of pH, i.e. a line plot with pH on the y-axis and either the Date or time active on the x-axis. Use different colours to distinguish the batches in the plot. This plot should also have clear titles & axes labels.

8. Extend the code from Q7 to create an animated timeseries plot of pH, where the timeseries is revealed over several seconds. The plot should have an appropriate title and labels etc.

9. Create appropriate boxplots showing the effect of Feed on both pH and Concentration. Similarly, create appropriate violin plots showing the effect of Hose on pH and concentration. Give a brief comment on each of these plots indicating what they show.

10. Investigate the GGally::ggpairs function and use this to create a plot showing the effect of all numeric variables (pH, Temperature, Pressure Conductivity, Concentration) on each other. Do any of these variables have significant relationships with each other or show any patterns? For those that do (if any), use separate scatterplots with appropriate smoothers to highlight the relationship(s) or pattern(s).