Assignment 4: Using the gglot2 package for bioprocess data visualisation

CHEN40770: Data Science For Biopharmaceutical Manufacturing

21/10/2020

Contents

1 Introduction

In our practical session we learned how to plot data using the ggplot2 and dplyr packages in R.

For this assignment you will again use the bioprocess dataset from a simulated penicillin fermentation process created by Dr. Stephen Goldrick at Univerisity College London. For more informaton see www.industrialpenicillinsimulation.com

Load the data as follows:

```r
library(chen40770data1)
```

Load the tidyverse packages as follows:

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## Warning: package 'dplyr' was built under R version 4.0.2
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.0.2
```

2 Submitting your assignment

To submit this assignment create an R script in your project folder in R Studio Cloud. The R script should be properly commented and adhere to the tidverse style guide at all times. At the top of the R script please include your **name** and **student number**.

Important Remember to save your work.

3 Assignment Requirements

Requirement 1: All questions must be answered using **ggplot2** and **dplyr**

Requirement 2: All plots must have an appropiate **title** as well as correct **axes labels**.

Requirement 3: All plots must be saved to a **tiff** file with **300dpi**. The filename must **clearly indicate the associated question**.

4 Questions

Total: 10 Marks

1. Create a **scatter** plot of Dissovled oxegen concentration (x-axis) versus Oxegen offgas (%)(y-axis) for **Batch ID 43** on **Day 7**. Include a **loess** smoothed line to show the relationship. (2 marks)
   Hint: dplyr **filter** can be used to select the required data

2. Create a **line** plot showing the **average substrate concentration** for the **3 control strategies and defective batches** from **day 3 to day 7** (2 marks)
   Hint: dplyr **filter** can be used to select the required data

3. Use a **density plot** to show the distrubtion of **offline NH3 concentration** for **each control strategy and defective batch** between **110 and 150 hours** of the fermentation. Create **4 subplots** *or* **control the transperancy** of the plots to clearly show the result. (2 marks)
   Hint: dplyr **filter** can be used to select the required data

4. Create a **barplot** to show the **median Offline Biomass concentration (g L^-1)** for each day except for **Day 12 and Day 13**. Only show the **raman** and **defect** batches. (2 marks)
   Hint: You can **chain** two dplyr **filter** statments to achieve your answer. Remember there is **NAs**, you will need to exclude them when calculating the median.

5. The dataset contains a variable called **Fault flag**. This variable denotes if the fermentation was successful (Fault flag = 0) or defective (Fault flag = 1). Create a **boxplot** showing **Penicillin concentration** for correct and defective batches at **Day 11**. Determine if the difference between the two groups is **statistically significant** using a **t-test**. To achieve full marks you will need to move the p-value label on the plot to a position that does not overlap the boxplot. (2 Marks)
   Hint: to see how to move the pvalue label see the help for stat_compare_means