ASSIGNMENT

Exploratory Data Analysis and Essential Statistics Using R

13 October 2023

Maria Anastasiadi (m.anastasiadi@cranfield.ac.uk)

Assignment Brief

The assignment is a compilation of tasks designed to test your understanding of the matter covered in the lectures and practicals. You must employ **R** throughout.

Submit the R code in a separate file. If you use external sources of information please be sure to supply references.

A total of **100 marks** are available.

Deliverables:

Deliverable I: Analysis Report outlining and discussing your results: [90 marks]

Your report will be marked according to the successful completion of the tasks set below and the quality of your report. The marks will be allocated as follows:

Task 1 [40 marks].

Asparagus is a highly perishable seasonal vegetable, which has increased in popularity over the past years due to its unique taste and nutritional value. The UK harvest season for asparagus typically lasts from April to June, which means that the UK market heavily depends on imported asparagus to cover consumer demand during the rest of the year. In an effort to extend the UK season, a sophisticated storage system using Dynamic Controlled Atmosphere (DCA) was tested. DCA manipulates the O_2 and CO_2 concentrations inside the storage boxes while the metabolic rates of asparagus are simultaneously monitored *in situ*. Air stored asparagus were used as control to assess the efficacy of DCA. Sampling took place at two different time

points; one week after the beginning of storage and four weeks after the beginning of storage. The quality parameters assessed included non-structural carbohydrates (fructose, sucrose, glucose), total soluble solids (TSS), moisture loss and the phytohormone abscisic acid (ABA) and its metabolites (7-OH-ABA, DPA, PA). ABA is a phytohormone related to many biological processes in plants including senescence and response to dehydration. ABA levels tend to drop rapidly following harvest, due to conversion to its inactive metabolites, but the trend can be slowed down due to abiotic stress, such as dehydration.

The different quality parameters (variables) have been expressed **in different units** in the provided csv.

You have been supplied with a csv file called **A_SC.csv**. The first column of this file includes the code name of each sample (AIR1:AIR6 and DCA1:DCA6 are the biological replicates for the control and DCA respectively after 1 week in storage, while AIR7:AIR12 and DCA7:DCA12 are the biological replicates for the control and DCA respectively after 4 weeks in storage).

The 2nd column in the csv file contains information regarding the treatment (AIR or DCA) and the 3rd column contains information regarding the time of storage in days (8 and 28).

This task requires you to perform exploratory data analysis and statistical inference to decide whether the DCA performs better than the control in retaining the quality parameters of asparagus for longer. More specifically you are required to do the following:

1) Load the dataset into the R environment (call it **ASP**) and perform quality control and initial exploratory data analysis using descriptive statistics. Create histograms to check the distribution for each variable and comment on the patterns you observe. Check if there are any missing values, and take appropriate action.
2) Create a new column containing the sum of all sugars (call it **sum_sugar**) and one containing the sum of ABA and its metabolites (call it **ABA_metab**). Add the new variables in the dataframe and call the new dataframe **ASPext**. Caution: Do not replace the dataframe **ASP**, keep both **ASP and ASPext**).
3) Use boxplots to visualise the following variables: **sum_sugar, ABA_metab, TSS, moisture_loss**. For each variable, include 4 boxplots in total (in the same graph) to visualise the distribution of values for each time/treatment combinations. Comment on the plots. Are there any suspected outliers? Make a note of suspected outliers.

4) Perform **two-way anova** to infer the effect of **Treatment** and **Time** and their interaction on each of the following variables: **sum_sugar, ABA_metab, TSS, moisture_loss**. Perform a Tukey test and make a list of the statistically significant effects and/or interactions.
5) Create bar plots for the variables: **sum_sugar, ABA_metab, TSS, moisture_loss**, showing side-by-side the concentrations for AIR and DCA samples at 8 and 28 days respectively. Add standard error bars and make use of symbols to denote significant differences (if any) based on the anova summary and Tukey test results.
6) Perform PCA for the original dataframe (**ASP**). Visualise the **PCA scores plot and biplot for PC1 and PC2** and print the %variance captured by each component in the axis labels. Colour the samples according to the groups they belong to (Treatment and Time) so you have 4 groups in total. Display a legend on the side of the plot with the different groups. What assumptions can you make by looking at the generated plots?
7) Create barplots of the loadings for the first two PCs and comment on which variables are more important for each PC.
8) Visualise the results in heatmaps/dendrograms and comment on the separation of the different treatments/timepoints.
9) Perform k-means clustering on a) the raw data and b) on auto-scaled data and look at the differences in the results. Has k-means successfully clustered together samples belonging to the same group?

Task 2 [15 marks].

You have been supplied with a csv file called "**Mean_Central_Eng_Temp.csv**" which is a subset of the Hadley Centre Central England Temperature (HadCET) dataset. The CET dataset is the longest instrumental record of temperature in the world. The mean, minimum and maximum datasets are updated monthly, with data for a month usually available by the 3rd of the next month. The mean daily data series begins in 1772 and the mean monthly data in 1659. Mean maximum and minimum daily and monthly data are also available, beginning in 1878.

These daily and monthly temperatures are representative of a roughly triangular area of the United Kingdom enclosed by Lancashire, London and Bristol. The monthly series, which begins in 1659, is the longest available instrumental record of temperature in the world. Since

1974 the data have been adjusted to allow for urban warming: currently a correction of -0.2 °C is applied to mean temperatures.

This task requires you to:

a) Import the csv file Mean_Central_Eng_Temp.csv into the R workspace. This dataset contains the Mean Central England Temperature (Degrees Celsius) from 1659 to 2022. Plot the raw data for the column "Annual" which contains the yearly average temperatures and try to identify general trends.

b) Perform **Loess** smoothing of the raw annual data to uncover the real trend in the time series. You can make use of the function **smooth_vec()** in the library **timetk**. Choose different parameters, so as to 1) perform smoothing equivalent roughly to a 10-year moving average and 2) smoothing to see the overall long-trend across all years. What are your observations? (Tip: create a plot with the raw and smoothed data ovelayed).

c) Calculate the 1961-1990 mean annual temperature. Then create a new column in the dataset and subtract the 1961-1990 mean from the annual mean temperature for all the years. Call the column "**Diff**". This is done in order to identify any anomalies relative to the 1961-1990 mean. Plot the data from the new column against the year and colour the negative values blue and the positive red. Then apply loess smoothing again roughly equivalent to a 10-year moving average. Plot the raw and smoothed data on the same plot. Describe the results and attempt to draw conclusions. Make use of literature references, if possible, to support your conclusions.

Task 3 [20 marks]

Cell division is a process by which a cell (parent) divides into two cells (daughter cells). Let us suppose that **waiting times** between cell divisions follow an **Exponential distribution** with mean θ. We wish to estimate the **rate θ** at which cell divisions occur, i.e., **the average number of cell divisions per minute**.

You have been supplied with a file named **"event_times.txt"**, which contains the cumulative time it takes for a cell to undergo 134 consecutive divisions (in minutes).

This task requires you to:

a) Import the data set into the R environment and store the values in a vector. Check the length of the vector and store this value in a variable. Plot the data to see how division times change over successive divisions. (The x axis should be the division time for each division, the y axis is meaningless and should be kept constant). What do you observe?

b) Create a new vector with the waiting times between cell divisions.

c) Create a histogram to visualise the distribution of waiting times. Do they look like they follow an exponential distribution?

d) Perform Bayesian analysis to estimate the **rate θ** of cell divisions. Consider a continuous **Gamma(s, r)** prior distribution for θ. Select parameters **s=5** and **r=15** and create a plot to visualise the prior distribution including 95% CI. Comment on the plot you created. For example, does the prior look informative etc?

e) Next, find the likelihood distribution for θ and calculate the Maximum Likelihood Estimator, i.e. the value of θ that maximises the likelihood function.

f) Finally find the posterior distribution of θ and create a plot with all the distributions (prior, scaled likelihood and posterior) on the same graph. Add the 95% CI for the posterior.

g) Now repeat the analysis but this time choose a **Gamma(15, 0.5)** distribution for your prior. Compare the results with the previous analysis. What are your conclusions?

Assignment Report Quality [15 marks]

Your analysis report needs to include

a) A comprehensive description and justification of the process you followed to perform the analysis for the tasks outlined below and the analysis results you acquired. Include the most important figures/tables in the main body and make use of an appendix to include supplementary information.
b) An interpretation of the results for each task, attempting to derive some general conclusions. You can use references from the literature to support your conclusions.

Deliverable II:

R script for this report [10 marks]:

Submit a single R script for all the tasks in this assignment. Ensure your R script produces all the results and graphs mentioned in your report and the process is clearly outlined with comments **[10 marks]**.

Submission:

The analysis report and the analysis scripts are to be submitted via the normal route on Canvas.

The assignment report must be written in either: MS Office Word (Windows/MAC) or LibreOffice (Ubuntu Linux). Export your report as a pdf and include it in the zip file as well.

The assignment report, scripts/functions and any accompanying data, image and results (text) files must be archived into one compressed file (e.g. ZIP or TAR) and include your name and student number in the filename.

Deadline:

Full-time students

The archived files must be uploaded on Canvas by: Saturday 21st October 23:59.

Part-time students

The archived files must be uploaded on Canvas by: Saturday 4th November 23:59.