School of Mathematical Sciences MATH1041 2023/24

Assessed Coursework 1

This coursework is assessed and is worth 15% of the module mark. You should submit your solution via Moodle by 3pm 26th March 2024.

The data file banana.csv[1] contains data on the attributes of banana. Your task is to perform a statistical analysis of these data, and write up your analysis as a statistical report. The data file, and a template for your report, are available on the Moodle.

Each row contains the demeaned measures of the attributes of banana, namely “Size”, “Weight”, “Sweetness”, “Softness”, “Ripeness”, “Acidity”, “HarvestTime”, and “Quality”. Specifically, “Acidity” is measured in pH and “HarvestTime” is the time taken to grow bananas until they are harvested.

You are asked to conduct an exploratory analysis of these data. Present your findings in a statistical report.

Your analysis should include:

1. Analysis of the numerical characteristics of a banana.

   (a) Various suitable summary statistics for each of the variables “Size”, “Weight”, “Sweetness”, “Softness”, “Ripeness”, “Acidity”, and “HarvestTime” along with brief comments, especially about anything notable.

   (b) A matrix of normal QQ plot for each of the variables “Size”, “Weight”, “Sweetness”, “Softness”, “Ripeness”, and “Acidity”. Comment on the shapes of these distributions, what you see in the QQ plots, and any interesting features of the data.

   (c) A matrix of scatterplots showing pairwise relationships between the variables “Size”, “Weight”, “Sweetness”, “Softness”, “Ripeness”, and “Acidity”. Comment on the plot.

   (d) Histograms of “HarvestTime”, one for each of the classifications of banana under the variable “Quality”.

2. Analyse the association between the numerical characteristics and the quality of banana. Discuss if one can identify good bananas by their numerical characteristics.

   (a) For each of the variables “Size”, “Weight”, “Sweetness”, “Softness”, “Ripeness”, and “Acidity”, produce a pair of side-by-side boxplots, one for good and another for bad bananas. Comment on and make relevant comparison between what you observe in each of these graphs.

(b) Use various suitable summary statistics to support the conclusion that you make in part (a).

3. A short further investigation in an aspect of your choosing.

   It's up to you what you choose here. For example, you may consider whether findings are different if analysis is conducted separately for bananas of sizes above the mean measurement and those below the mean measurement. Think freely and creatively about what is worthy of investigating!

Your report should be a maximum of 1000 words. Please state the number of words used at the bottom of your report. There will be a penalty for exceeding the word limit.

Please include as an appendix the R code to produce the results in your report, but don't include R code in the main part of the report itself.

Hints and tips:

1. Be sure to read section 3 of the lecture notes which describes the requirements of a statistical report.

2. You need to do some literature review to support the study. Include at least two references of journal papers in Introduction. Besides, you need to include motivation of study and outline the flow of the report.

3. An approximate break-down of marks is as follows: Summary [5 marks], Introduction [10], Methods [5], Results and discussion [45, of which 15 are for the investigation of your choosing mentioned in point 3 above] and Conclusions [5]. An additional 5 marks are available for the overall clarity of the writing and presentation.

4. You will find the subset command very useful. Some examples:

   subset(dat, Quality == "Good") extracts the good bananas.

   subset(dat, Quality == "Good"& HarvestTime >= 0) extracts good bananas that are grown for more than or equal to the mean harvest time.

[1] https://www.kaggle.com