Student Name: Han Weiyu
Matriculation Number: G2301835L

Story-telling with Graphics and Visualizations (BS6203)

Assignment 2

---

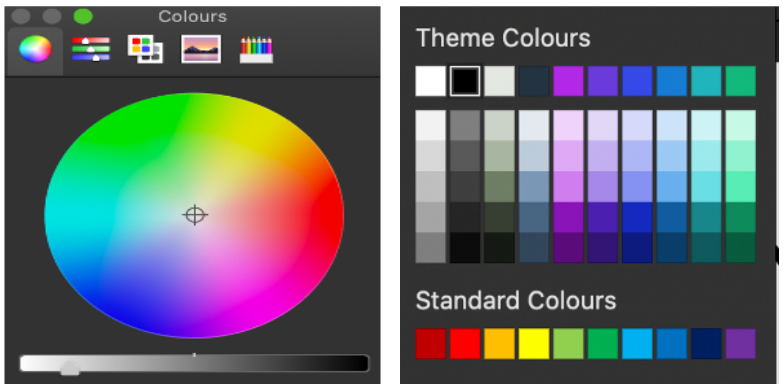
1. **Question 1**
   How do you think simultaneous contrast and metamerism can affect the effective representation of continuous variables using colour?

   Simultaneous contrast means putting different colors around the same color may cause the same color looks different. Metamerism means the different environment light could affect the same color to be sensed differently. Thus, Both the simultaneous contrast and metamerism could have a negative effect on the effective representation of continuous variables using colour.

   When comes to representation of continuous variables using colour, a lot of similar color are used to present very close values. Simultaneous contrast could easily confound the specific value. For example, there are two points that have same values in the heatmap, which means their colors are the same as well. However, the first point is surrounded by those points that quite differ from its value and color. And the other one is surrounded by those points that are quite similar with its value and color. Affected by simultaneous contrast, the viewer could be told they are different colors, which might confound the result.

   This fuzziness also occurs when observers compare results in different environments. For example, two similar continuous values are represented using color. It is certain that their colors are also similar. The similar color could be easily told the difference in the lab, but is hard to differ in the bedroom, which definitely affect the effective representation.

2. **Question 2**
   These are the theme colours you can find right in MS PowerPoint. Looking at both rows and columns, outline where the concepts of saturation, hues and lightness are (you can use the colour wheel on the left to help you.

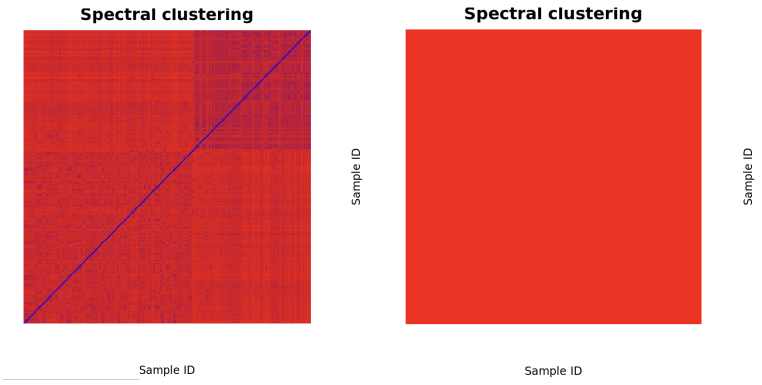


In the right side graphic, the rows represents how the same hues become different in different saturation. Thus, different rows refer to different saturation. Similarly, both the upside 'theme colours' and the downside 'standard colors' have different hues from left to the right in one row. In other words, different columns represent different hues. Unfortunately, lightness cannot be easily told in the right graphic.

Similarly, the colour wheel tells us a lot. Each dot on the edge of the wheel represents a different hue, and the distance of the middle dot from the centre of the wheel represents a different degree of saturation. Lightness refers to the proportion of black colour added to a colour, and the drag bar underneath the colour wheel can represent the concept of lightness.

3. **Question 3**
Upload the three datasets attached with this assignment, into the "Microbiome submission" tab of the webtool accessible at https://integrative-microbiomics.ntu.edu.sg and click on "Next" followed by "Merge". This will take you to "Cluster visualisation" tab. Click on "Similarity Plot". You will see a heat-map appear, play around by changing different options such as colours and data-transformation. Relating the concepts learnt in the class, which options of the heatmap do you think represents the clusters in the data, best and worst? Discuss the advantages and disadvantages of heatmaps? Can you come-up with an alternative representation of this data?

The colors may represents the clusters in the data best and worst. As announced by the website, color 1 indicates lower similarity and color 2 indicates higher similarity. Clustering brings like next to like items to reveal patterns in the data. In other words, similarity represented by colors is the basis of the clustering. Therefore, the choice of more contrasting colours facilitates the reader to better distinguish between the different clusters. In my opinion, the combination of setting colour 1 as red and colour 2 as blue is the best. Because of its high contrast, the difference can be distinguished at a glance. And setting any two of the same colour is the worst choice. Because all the information is covered up and no difference can be seen. The contrast diagram is shown below.

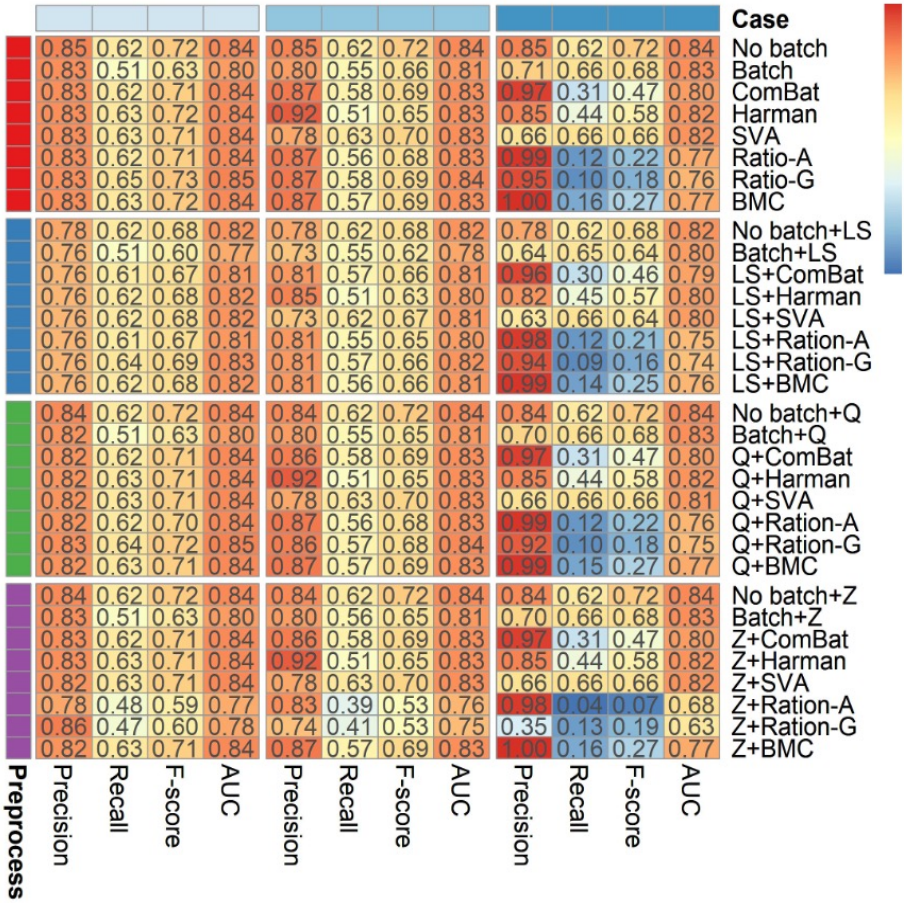


(c) Red and blue (d) Red and red

The advantages of heatmap are mainly about simple and easy to compare. When representing high-throughput data, heatmap is a suitable method because it reduce the information contained by data

to two or three hues' colors, which is likely for viewers to identifying high-density information contained by high-throughput data. Besides, the change of hues also shows the viewers the similarities, clusters and differences simply. However, there are also disadvantages of heatmaps. The details of information has been hidden and it is difficult to tell the meanings of a cluster and the basis on which it plotted so. It could be considered as a trade-off.

4. **Question 4**
This below figure represents a table showing 3 case scenarios on the columns (where balanced is an easy setup, and severely unbalanced is a challenging setup). Each column is evaluated on 4 metrics, precision, recall, f-score and AUC. You do not need to know what these are, but you only need to know only that the closer to 1, the better. The pre-process describes the use of none (not using any), scaling, quantile and z-score data transformation procedssures before using a batch effect removal algorithm such as Combat, Harman, SVA, Ratio-A and Ratio-G. The no batch and Batch rows refer to the situations where no batch effects exist, and where batch effects were introduced into the data. Which retinal variables are used here? And how does one go about interpreting this graph?



Colors and shape are the specific retinal variables used in this diagram. There are some hues represented in the right side bar, red, white, blue and etc. These different colors are used to show the continuous value changing from 0 to 1(probably, it is not clear in the diagram, which may cause confounding). Besides showing quantitative values, color are also used to show the categorical cases and transformation procedssures. For example, light grey blue and dark grey blue represent bal-

anced and unbalanced(samely, not announced in the diagram and caused confusing). Shape are also used in the diagram to help better comparision. Different cases and transformation procedssures have rectangles away from each other to clarify each group.

The graph shows how different cases and transformation procedssures groups evaluated on 4 metrics. There are three different colour blocks on the top of the diagram representing different case scenarios, and four color blocks on the left representing the use of none, scaling, quantile and z-score data transformation procedssure. Accordingly, the whole large matrix is divided into 12 small matrices, each of which has the scores of precision, recall, F-score and AUC after different batch effect removal algorithm inside the small matrix. Closer to red means closer to 1, the higher this rating is, similarly, loser to blue means closer to 0, the lower this rating is. According to the picture, different cases perform significantly better in precision and AUC than in recall and F-score. Severely unbalanced cases perform especially poorly in recall and F-score. ComBat, Ratio-A, Ratio-G, and BMC algorithms will obtain higher precision and lower recall and F-score in severely unbalanced cases.