**Department of Computing**
**The University of the West Indies, Mona**

# COMP3162 – Data Science Principles

SEM 2 2020/2021

**Project Part 02 – Teamwork**
**(2 persons per team)**
[Weighted 15% of course marks]
**DUE: April 07, 2021**

1. **Tweet Data Analysis [20]**
   a. Merge the tweets collected from project 01 by each person in the team. [1]
   b. Remove all duplicate tweets in the newly merged set of tweets. A tweet is a duplicate if the text is exactly the same as the text in another tweet. In removing the duplicate tweets, it might be useful to keep the one that has the highest retweet count. [2]
   c. Explore the merged tweets and provide descriptive statistics. [3]
   d. What are the dominant emotions associated with beverages in any two locations? [4]
   e. What are the dominant emotions in the overall dataset? [2]
   f. What is the overall sentiment in tweets regarding "beverages" and "party or concert" (separately)? [4]
   g. Conduct ONE additional analysis of your choice to discover any further useful insights. [4]

2. **Collect, Explore, Prepare Structured Data[20 marks]**
   a. Download the datafile *consumer_pt02_2021.csv* from OurVLE
   b. Explore the data and provide details on all fields retrieved. You should ensure all features in the dataset (each column) are reviewed and summarized to verify things such as value ranges, missing values etc. Be sure to generate relevant graphical representations where necessary to demonstrate your review and decision making. [7]
   c. Fix noise, outlier and any other issues discovered (example: na values). You must provide discussion / explanation of all activities done and why each decision has been made. [8]
   d. Format/reformat the data as necessary. Please note that as you proceed through the project, you may need to do additional formatting to enable your analysis. [5]

3. **Structured Data Analysis/Modeling [35]**
   Write code to conduct analysis that will answer the questions below. You are encouraged to use tables/graphs where necessary to visualize results. Additionally, your code should be shown along with each question, the result and notes that explain the results.
   a. What is the average spend on beverages in each country? [3]
   b. Which country has the highest spending on beverages? [2]
   c. Which country consumes the most beverages? [2]
   d. What is the average profit from the sale of beverages in each country? [3]
   e. What has been the total revenue from beverages for each year since 2014? [5]
   f. Plot a time series graph showing change in overall revenues from beverages for the last six months (in the dataset). [4]
   g. What is the dominant sales channel for beverages? [2]

h. Determine whether beverages units sold is above the overall average for units sold for all other products. [3]

i. In which season (Spring, Summer, Autumn, Winter) does persons spend the most on beverages? [6]

j. Is there a correlation between the season and the units sold for beverages? Explain the result. [5]

4. **Recommendation:**

a. Based on your analysis of both the tweet data and structured data, what would you recommend to Hard Knocks and why? [5]

5. **BONUS – 10 marks**

a. Which features in the dataset can be used to predict the units sold for beverages?

WHAT TO SUBMIT:

Submit a pdf generated from ***rmarkdown*** with evidence of the work done. Again, please ensure you include all code, results, visualizations, and explanations. Submitting just code and the output of that code may be insufficient for many of the questions since the result may not directly answer the question and therefore requires discussion notes.