Final Project

DAB501 - Section 1 and 2

*27 April 2022*

# Student Information

Name (print): Ravi ID: 0795606

Signature: *Ravi*

# Academic Integrity

*\*Replace the underscores below with your name acknowledging that you have read and understood the statement in the context of St. Clair College’s Academic Integrity policies.\**

I, Ravi, hereby state that I have not communicated with or gained information in any way from any person or resource that would violate the College’s academic integrity policies, and that all work is my own.

# Instructions

1. Download the data file entitled ***starbucks.csv***
2. Download the template file entitled ***final\_project\_template(1).Rmd***
3. All of your answers should go in the appropriate section of the template file
4. After completing the project, convert the .Rmd file to a PDF or Word document
5. Upload the PDF or Word document to Blackboard
6. Project is not graded unless you fill in your name, ID, and sign this paper in the appropriate places and submit with other required documents

# Data Set Description

This data set contains a subset of data collected on food items sold at a typical Starbucks coffee shop. It contains the following variables:

* ***calories*** represents the number of calories in a particular food item (This is the variable to be predicted)
* ***fat*** represents the fat content of a particular food item, measured in grams
* ***carb*** represents the carbohydrate content of a particular food item, measured in grams

# Modeling

For each pair of explanatory/response variables (allocated marks are **per** pair of variables):

1. Identify the explanatory variable. **(1 mark)**
2. Identify the response variable. **(1 mark)**
3. Create a linear regression model and display the full output of the model. **(3 marks)**
4. Using the variables noted in #1 and #2 above and the results of #3, write the equation for your model. **(4 marks)**
5. Explain what the intercept means in the context of the data. **(2 marks)**
6. Is the intercept a useful/meaningful value in the context of our data? If yes, explain. If not, explain what purpose it serves. **(1 mark)**
7. Explain what the slope means in the context of the data. **(3 marks)**

# Model Assessment

1. Which metric can you use to choose between the two models you just created? **(1 mark)**
2. Explain what this metric means and why it is good for comparing models. **(3 marks)**
3. According to this metric, which model is the best of the two you created? Why? **(2 marks)**

# Model Diagnostics

***NOTE****: Even if you think your model fails one of the diagnostic tests, please provide an answer for ALL diagnostics. Your answers below should contain* ***three*** *plots.*

Now that you have chosen your best model, you need to determine if it is reliable or not:

1. Create two new data columns based on your best model: predicted values for your response variable and the corresponding residuals. **(4 marks)**
2. Create a plot to check the assumption of ***linearity***. State whether or not this condition is met and explain your reasoning. **(6 marks)**
3. Create a plot to check the assumption of ***nearly normal residuals***. State whether or not this condition is met and explain your reasoning. **(6 marks)**
4. Create a plot to check the assumption of ***constant variability***. State whether or not this condition is met and explain your reasoning. **(6 marks)**

# Conclusion

1. Based on the results of the “Model Diagnostics” section above, what can you conclude about your model? **(2 marks)**