**INSY 662: Data Mining and Visualization (Fall 2023)**

**Individual Assignment 5**

**Due by Wednesday, November 1, 2023 at 11:59pm on MyCourses**

* This is an individual assignment. The McGill Academic Integrity code is applied!
* You need to submit **TWO FILES to MyCourse**. One, **a Word file with your answers**. You can type your answers (highlighted places) directly on this assignment file. Two, **your Python code**.
* The goal of this assignment is to understand **Agglomerative and K-Means Clustering**.

In this assignment, we are going to use the Cereals dataset.

**Task 0.** We are going to develop clustering algorithms using the following variables: 'Calories', 'Protein', 'Fat', 'Fiber', 'Carbo', 'Sodium', 'Sugars', 'Potass', and 'Vitamins'. Drop any observations that have one or more missing value for these variables.

**Task 1.** First, perform agglomerative Clustering with complete linkage. Report the number of cereals in each cluster when the number of cluster is 2.

Number of cereals in the first cluster: 60

Number of cereals in the second cluster: 14

**Task 2.** Then, use the same set of variables above to perform K-Mean Clustering with k=2. Report the number of cereals in each cluster.

Number of cereals in the first cluster: 57

Number of cereals in the second cluster: 17

**Task 3.** With the results from K-Mean Clustering, intuitively explain the characteristics of cereals in each cluster.

Cereals in the first cluster are:

Less likely to be marketed as ‘healthy’ than in the second cluster. On average, this group of cereal contains less protein (not by much), higher fat, higher sodium, higher sugar content, less fiber, and less potassium. Though more vitamins.

Cereals in the second cluster are:

More reflective of cereal branded as healthy compared to those in the first cluster. Less calories, fat, sodium, sugar, and higher protein, fiber, potassium, and so on.

\*\*End of Assignment 5\*\*