8STT108 – Big Data Statistics Analytical Tools

Lab #2

Pre-processing

Fall 2024 **Professor: Sara Séguin**

Python is required for the lab assignment. Use any relevant libraries such as: pandas, numpy, statistics, matplotlib.

The goal of the lab is for the students to experiment with the libraries, therefore, you should read any required documentation to understand the tools.

- The spreadsheet **Olympics** is used for this part. The initial file is given to the students.

As seen in class, this dataset has many problems.

1. Import the data set in Python.
2. Using the methods seen in class, clean the dataset by explaining step-by-step what you have done to improve the quality of the dataset.
3. Attach a copy of the clean data set to the report.
4. Plot any relevant graph that could help visualize this data set without scrolling through the file. Explain your choices for the graphs and the data they show.

- The spreadsheet **Boston Housing** is used for this part.

1. Import the data set in Python.
2. Use any multivariate tools to visualize the data and explain your choices, and present a data analysis. For example, pairwise scatter plots of heatmaps could be used.