8STT108 – Big Data Statistics Analytical Tools
Lab #1
Descriptive statistics

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 file Lab1_Buffalo is provided to the students as a spreadsheet.

This file contains data on the Cattles for each region in United Kingdom. Data includes the regions, the number of female and male cattles, and the data is available from 2005 to 2010.

You work for a data analyst company and your supervisor asks you to present the results using descriptive statistics, since he does not have time to analyze the data himself.

1. Load the dataset in python. Think about re-arranging the data in the spreadsheet beforehand, it could be useful.
2. For all regions of UK, present descriptive statistics for female, male and total separately. Statistics you should use are: measures of dispersion and measures of central tendency.
3. Plot a boxplot for female, male and total separately.
4. Analyze the results obtained in 2. and 3. Discuss and conclude on the results based on the boxplots and the descriptive statistics. Is there a difference between female and male, more specifically?
5. Draw a boxplot for the total of each cattles for each region. Analyze the results. Does a particular region stand out, and if so, why?