8STT108 – Big Data Statistics Analytical Tools

Lab #3

Principal Component Analysis

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 **decathlete** is used for this assignment. These are results from the 10 events.

The variables are:

- Year
- Decathlete
- Run100m_pts – Points awarded for the 100m sprint
- LJ_pts – Points awarded for Long Jump
- SP_pts – Points awarded for Shot Put
- HJ_pts – Points awarded for High Jump
- Run400_pts – Points awarded for the 400m run
- H_pts – Points awarded for Hurdles
- DT_pts – Points awarded for Discus Throw
- PV_pts – Points awarded for Pole Vault
- JT_pts – Points awarded for Javelin Throw
- Run1500_pts – Points awarded for the 1500m run
- Overall – Overall points awarded

The goal of the PCA analysis will be to take the 10 events of these decathletes and see if we can reduce them to 2 or 3 components.

1. Upload the dataset in Python.
2. Using the methodology seen in class, perform a PCA analysis to see how many components you should keep for the regression. Report the methodology, results, and any relevant graphs. Discuss on the results.

3. Conduct a linear regression on the chosen components. If you can reduce the components, also do so, but explain why your remove any. Report the results of the regression and discuss on the results.