CS982: Big Data Technologies
CS989: Big Data Fundamentals

COURSEWORK

AIM OF THE ASSIGNMENT

To provide deeper understanding of appropriate methodological approaches to processing and analysing noisy data; and to encourage appreciation of the challenges involved in data analysis.

LEARNING OUTCOMES

Understanding of the fundamentals of Python to enable the use of various big data technologies; Understand how classical statistical techniques are applied in modern data analysis; Understanding of the potential application of data analysis tools for various problems and appreciate their limitations; Understanding of the challenges and complexity of data analysis.

THE BRIEF

Provide a brief report on analysis of an open dataset. Example datasets are available on the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php) or on Kaggle website (https://www.kaggle.com/datasets) for example. You can also select any other dataset from other sources, but make sure that the dataset is public and that you have the right to access and analyse the dataset and to share the results. However, you cannot select a dataset that comes packaged with Scikit-Learn (https://scikit-learn.org/stable/datasets/toy_dataset.html) or Seaborn (https://github.com/mwaskom/seaborn-data), and you cannot select a dataset that we have worked on during Lab sessions. You can focus your report on one aspect of the dataset or multiple aspects, the main objective is to find some interesting questions or problems to answer.

The following criteria will be used when marking your assignment:

- Identification and description of key challenge(s) or problem(s) to be addressed   10%
- Introduction to the dataset   10%
- The challenge(s)/problem(s) is (are) to be addressed using the following
    - Summary statistics (including figures) for data being analysed   20%
    - Description, rationale, application and findings from one unsupervised analysis method   20%
    - Description, rationale, application and findings from one supervised analysis method   20%
- Reflection on methods used for analysis   10%
- Structure presentation, and proper citation of references   10%

SUBMISSION

The report to be submitted should be 3000 words (+/- 10%) excluding front cover, table of content, list of figure / tables, and appendices. The document must be in pdf format. All code used to the analysis is to also be submitted, if not submitted the submission will be considered incomplete and a late penalty will be applied until all components of the assessment are submitted; More details will be available on the submission page on MyPlace. Both the code and the report should be submitted using MyPlace; no submission will be accepted in any different way. Any extensions should be requested in advance of the submission deadline, with a valid reason. Assessments submitted after the deadline without an approved extension will be subject to penalties on a sliding percentage scale: 10% for the first 24hrs, and 5% for each additional day. Penalties will be applied to late submitted assessments up until four days, and assessments submitted after four days of the deadline will receive a mark of zero.

DUE: 12:00 noon, Monday November 8th, 2021