Data Science Techniques and Applications, 2018-19

Coursework I: Dataset analysis

The goal of this coursework is to make you explore the Kaggle[1] platform and analyse some open datasets from the point of view of long-term data management, applying the concepts seen in class. No coding is required and an essay suffices.
As a preview of Coursework II: the second assessment will ask you to perform some light coding, e.g., Principal component analysis, on the dataset you study in this coursework.
Your Work, and the writing of your essay shall be organized in phases as follows.

Phase 1.
Subscribe to Kaggle, preferably using your academic email address (i.e., from the dcs.bbk.ac.uk or the bbk.ac.uk domains).
Search the available datasets and challenges. Select a dataset which concerns an application domain for which you have some expertise/interest. For instance, if you like watching TV series you can find a dataset on AV consumption,
Another reason for selecting one of the Kaggle datasets is relevance to the topic of your MSc graduation project. Should you select a dataset on the basis of your graduation project please note it in your essay.

Phase 2.
Write a brief description of the dataset: what's inside? Who and when collected the data? What Kaggle challenges are proposed for this dataset?

Phase 3.
Dimensional analysis: write down the main aggregate measures of the dataset: number of data points, number of dimensions. Select a small number of dimensions that you consider the key to understanding how data is distributied. Describe and comment those dimensions (e.g., range of the dimension, quality of the data, possible data quality/integrity issues) in your essay.

Important dates: Please refer to the course web page on Moodle for the web submission procedure and for the deadlines.

Submission: **For this coursework, length and format of the essay are free.**
Please use your judgement on the right amount of data and length of presentation for a technical description of a dataset and a possible repository. In this instructor's opinion, two pages could suffice for relatively plain datasets, e.g., the one on wine reviews seen in class.

1 https://en.wikipedia.org/wiki/Kaggle

Please edit your essay using mainstream formats (e..g., HTML, PDF, ODT, TEX, MD) that support your presentation style. Please also be advised that heavy Microsoft Office formats (e.g., DOCX, PPTX) could cause anomalies when displayed by Libre Office[2] so please refrain from using them. As a courtesy, please use an easy-to-read style similar to that of this document (Times New Roman font or similar, size 12 or bigger, 1.15 line spacing or higher, justified alignment).

Plagiarism: please be advised that Moodle deploys a state-of-the-art plagiarism detection software[3] to evaluate coursework submissions against both Web sources and other submissions, past and present. Each submission will be scored for originality; submissions with low originality might be discarded or penalized.
It is however possible to insert quotations by using appropriate typographic style and providing the reference:

> *this phrase is an example of a typographic style for citation and reference: it will be discounted by Turnitin analysis.*

Please make use of a formalized citation system and report articles and books you refer to, e.g. [Narayanan et al., 2011] and [Shenoy et al., 2011].

References

[Shenoy et al., 2011]
G. G. Shenoy, M. A. Wagle, A. Shaikh, 2017
Kaggle Competition: Expedia Hotel Recommendations
https://arxiv.org/abs/1703.02915

[Narayanan et al., 2011]
A. Narayanan, E. Shi, B. I. P. Rubinstein, 2011.
Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge
Proc. of the 2011 Int'l Joint Conference on Neural Networks.
https://arxiv.org/abs/1102.4374

2 https://www.libreoffice.org/
3 https://turnitin.com/gateway/index.html