**DSTA Coursework 1: Dataset dimensionality analysis**

# Data Science Tech. and App. – BUCI057H7 2020-21

Coursework description, guidelines and marking scheme

Version 4: the coding part has been clarified.

## 1. Introduction

This assignment is an integral part of this module and contributes 10% to the overall mark.

The aim is to **[to make you start exploring the Kaggle platform and analyse an open datasets, starting with the dimensionalities it involves, possible secondary analyses and in general its data management side].**

By doing this coursework you will get experience with **[dimensionality analysis. Coding is not strictly required and an essay suffices; however it could be useful to use Python to draw histograms and scatterplots (for which the baseline code has been presented in the lab) to be included in the presentation, and/or computing statistics or ranges that don't already appear in the Kaggle description. Also centering and covariance matrix might be deemed useful, please see resource v on the DSTA web page for an introduction].**

The assignment is further explained in Sections 2 below. Section 3 of this document gives you an example of how to structure your report and explains the marking scheme. Section 4 presents the deadlines and submission instructions. Section 5 explains the penalties for late submissions, and Section 6 explains how the College deals with plagiarism. Section 7 and Section 8 provide additional information on learning resources and referencing.

## 2. Description of the work

[As a preview of Coursework II: the second assessment will ask you to perform some light coding, i.e., 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 based on your graduation project please note it in your essay.

When making your choice please also consider the following: as Coursework II will ask you to perform Principal Component Analysis, it is important that the Kaggle dataset you select has numerical (i.e., non-categorical dimensions) dimensions. "Temperature in London" is an example of a numerical dimension, with values normally in the range of 0 to 30 Celsius degrees. Conversely, "Eye color" is a categorical dimension with values normally in the set {black, brown, green, blue, grey}.

**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? + variable description(continuous, discrete)
+ existing literature

**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 distributed. Describe and comment those dimensions (e.g., <u>range of the dimension</u>, quality of the data, possible data quality/integrity issues) in your essay.]
+ balanced? / missing values?
+ comment on selected variables and the target
+ data quality

## 3. Your submission and marking scheme

**[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.**
**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, e.g., LibreOffice 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 12/4, 1.15 line spacing or higher, justified alignment).**
**].**

**[Marks will be awarded incrementally, i.e., we start from 0 and we add points, up to 100, as we read. We will add points five by five:**
**Technical detail, insights: [0..30]**
**Completeness of the analysis: [0..20]**
**Quality of the graphical/tabular part [0..15]**
**Style, conciseness: [0..15].**
**Overall perceived quality of the presentation: [0..10]**
**Choice of a challenging dataset: [0..10]**
**].**

## 4. Deadlines and submission instructions

**Submission is only through Moodle and consists of [a file containing the essay with its graphical displays and tables].**

4.1. Detailed submission instructions and deadlines.

You should upload on Moodle the completed assignment by **[the Moodle automated deadlines: submission and late submission]** (this is Moodle time, i.e., GMT/London local time).

It is your responsibility to ensure that files transferred from your own machines are in the correct format and that any programs execute as intended on Department's systems prior to the submission date.

**In the unlikely case that Moodle does not offer a checkbox for the approval of the "Academic Declaration" then the following procedure should be carried out:**

Each piece of submitted work MUST also have a page entitled "Academic Declaration" by the author that certifies that the author has read and understood the sections of plagiarism in the document

http://www.bbk.ac.uk/mybirkbeck/services/rules/Assessment%20Offences.pdf

which describes College's Policy on assessment offences. Confirm that the work is your own, with the work of others fully acknowledged. Submissions must also be accompanied by a declaration giving us permission to submit your report to the plagiarism testing database that the College is using.

**Reports without a Declaration form are not considered as completed assignments and are not marked.**

**The Academic Declaration should read as follows[1]: "I have read and understood the sections of plagiarism in the College Policy on assessment offences and confirm that the work is my own, with the work of others clearly acknowledged. I give my permission to submit my report to the plagiarism testing database that the College is using and test it using plagiarism detection software, search engines or meta-searching software."**

You should note that all original material is retained by the Department for reference by internal and external examiners when moderating and standardising the overall marks after the end of the module.

## 5. Late coursework

**Please notice that with the start of the 2020-21 academic year a new set of college rules concerning late coursework submissions entered into force.**

**It is our policy to accept and mark late submissions of coursework. You do not need to negotiate new deadlines and there is no need to obtain prior consent of the module lecturer.**

We will accept and mark late items of coursework up to the deadline for late submission. Therefore the **last day the system will accept a late submission for this module is [please refer to the Moodle setup]** (this is Moodle time, i.e., GMT/London local time). **This is the absolute cut-off deadline for coursework submission.**

---

[1] Please notice the double quotes to make it invisible to Turnitin.

Even if the personal circumstances that prevented you from submitting the coursework by the last day are extreme, **the Department will not accept coursework after this date**. We will, naturally, be very sympathetic, and the MSc Programme Director will be happy to discuss ways in which you can proceed with your studies, but please do not ask us to accept coursework after this date; we will not be able to as there is a College-wide procedure for managing late submissions and extenuating circumstances in student assessment. As soon as you know that you will not be able to meet the deadline, it will be useful for you to inform the module lecturer. They will be able to advise you on how best to proceed. Another person to speak to, particularly if the problem is serious, is the MSc Programme Director. You will then have the opportunity to discuss various options as to how best to continue your studies.

Further details concerning the rules and regulations regarding all matters concerning assessment (which naturally includes coursework), you should consult College Regulations at

http://www.bbk.ac.uk/mybirkbeck/services/rules.

**Please see the programme handbook for the rules governing Late Submissions and consideration of Mitigating Circumstances and the Policy for Mitigating Circumstances at the College's website**

http://www.bbk.ac.uk/mybirkbeck/services/rules.

## 6. Plagiarism

The College defines plagiarism as "copying a whole or substantial parts of a paper from a source text (e.g. a web site, journal article, book or encyclopaedia), without proper acknowledgement; paraphrasing of another's piece of work closely, with minor changes but with the essential meaning, form and/or progression of ideas maintained; piecing together sections of the work of others into a new whole; procuring a paper from a company or essay bank (including Internet sites); submitting another student's work, with or without that student's knowledge; submitting a paper written by someone else (e.g. a peer or relative), and passing it off as one's own; representing a piece of joint or group work as one's own".

The College considers plagiarism a serious offence, and as such it warrants disciplinary action. This is particularly important in assessed pieces of work where the plagiarism goes so far as to dishonestly claim credit for ideas that have been taken from someone else.

Each piece of submitted work MUST have an "Academic Declaration" form signed by the student which certifies that the students have read and understood the sections of plagiarism in the College Regulation and confirm that the work is their own, with the work of others fully acknowledged. Submissions must be also accompanied by a declaration giving us permission to submit coursework to a plagiarism testing database that the College is subscribed.

**If you submit work without acknowledgement or reference of other students (or other people), then this is one of the most serious forms of plagiarism**. When you wish to include material that is not the result of your own efforts alone, **you should make a reference to their contribution, just as if that were a published piece of work**. You should put a clear acknowledgement (either in the text itself, or as a footnote) identifying the students that you have worked with, and the contribution that they have made to your submission.

## 7. Referencing

References include the full bibliographic information about the source, such as the author(s)'s name(s), date of publication, title of work, place of publication, and publisher. This information is usually given in the section called Reference List or Bibliography at the end of the text. The key principle is that you should give enough information to allow another person to find the source for themselves.

Here are some examples using the Harvard referencing system:

[when you are referring to a book]
Lewin, K., 1951. Field Theory in Social Science. New York: Harper & Row.

[when you are referring to a chapter in a book, where 'ed.' means editor, and 'edn.' means 'edition']
Piaget, J., 1970. Piaget's theory. In: P. Smith, ed., Handbook of child psychology. 3rd edn. New York: Wiley, 1970, pp. 34-76.

[when you are referring to a journal article]
Holmqvist, M., 2003. A Dynamic Model of Intra- and Interorganizational Learning. Organization Studies, 24(1), 95-123.

[when you are referring to a webpage]
W3C, Web Accessibility Guidelines and Techniques, available online at http://www.w3.org/WAI/guid-tech.html. Last accessed 12/02/2015.

**Independent of their type (e.g. book, article, webpage), all references are included at the end of a document in alphabetical order** starting from the author's name as in the example above.

## 8. Useful resources

Here are some resources on plagiarism, study skills, time management and referencing that can help you to better manage your project and avoid plagiarism.

### *On Plagiarism*
https://owl.purdue.edu/owl/teacher_and_tutor_resources/preventing_plagiarism/index.html
- https://ori.hhs.gov/avoiding-plagiarism-self-plagiarism-and-other-questionable-writing-practices-guide-ethical-writing
- https://moodle.bbk.ac.uk/mod/url/view.php?id=38983
- https://moodle.bbk.ac.uk/mod/url/view.php?id=70384

### *On Referencing Systems*
Harvard guide to citing references:
- http://www.open.ac.uk/libraryservices/documents/Harvard_citation_hlp.pdf
- https://library.aru.ac.uk/referencing/harvard.htm

### *On Study Skills*
- http://www.bbk.ac.uk/student-services/learning-development