Python coursework specification for 2021

Data Analytics ECS648U/ ECS784U/ ECS784P
Revised on 21/01/2021 by Anthony Constantinou and Syed Rafee.

1. Important Dates

- Release date: Tuesday **5th February** 2021 at 10:00 AM.
- Submission deadline: Wednesday, **7th April** 2021 at 10:00 AM.
- Late submission deadline (cumulative penalty applies): Within 7 days after deadline.

General information:

i. When submitting coursework online you receive an automated e-mail as proof of submission. Turnitin receipt does not constitute proof of submission. Some students will sometimes upload their coursework and not hit the submit button. Make sure you fully complete the submission process.

ii. A penalty will be applied automatically by the system for late submissions.
   a. Your lecturer cannot remove the penalty!
   b. Penalties can only be challenged via submission of an Extenuating Circumstances (EC) form which can be found on your Student Support page. All the information you need to know is on that page; including how to submit an EC claim along with the deadline dates and full guidelines.
   c. If you submit an EC form, your case will be reviewed by a panel and the panel will make a decision on the penalty and inform the Module Organiser.

iii. If you miss both the submission deadline and the late submission deadline, you will automatically receive a score of 0. Extensions can only be granted through approval of an EC claim.

iv. Submissions via e-mail are not accepted.

v. It is recommended by the School that we set the deadline at 10:00 AM. Do not wait until the very last moment to submit the coursework.

vi. Your submission should be a single PDF file.

vii. For more details on submission regulations, please refer to your relevant handbook.

2. Coursework overview

The coursework is based on the Python lectures and labs and can be completed individually or in a group of up to three students. You are free to choose or collate your own dataset and apply two data analytic techniques to a real-world problem of your choice. In brief, you will:

i. Decide whether to form a group or not,
ii. Agree on the application area,
iii. Investigate the are and prepare the Introduction along with a short literature review,
iv. Collect data,
v. Clean and pre-process the data (if necessary),
vi. Apply two data analytic methods to the data,
vii. Present and discuss results,
viii. Draw conclusions,
ix. Finalise the report covering all of the above (see Section 4 for marking criteria).

You should address a data-related problem in your professional field or a field you are interested in. If you are motivated in the subject matter the project will be more fun for you and you will likely produce a better report. The same applies to the data analytic methods; i.e., you are free to apply and test the methods of your choice from those covered in the Python labs or in the Python lectures.

Please note:

i. This module is available to students with and without computer science background. The labs provide step-by-step tutorials on machine learning with Python. Anyone should be able to follow these tutorials to analyse some data irrespective of previous academic background.
ii. Projects can be done individually or in groups of up to three people. If you form a group, it is always a good idea include people who have different skills so they can be assigned to different parts of the project (e.g., a coder, an analyst, a person to do the literature review, etc).
    a. Some students will not be able to join a group due to work commitments or other reasons. This is not a problem since the coursework can be completed individually.
    b. Please do not send an e-mail asking if it is acceptable to form a group of more than three people; it is *not* acceptable.

3. Deliverables

The coursework deliverable takes the form of a mini conference paper. The report shall have a maximum length of 10 pages excluding References and Appendices. Font size should not be lower than 11 and Page margins should not be lower than 2 (these restrictions do not apply to the References and Appendices).

Reports should be written with a technical audience in mind. It should be concise and clear, adopting the same style you would use in writing a scientific report. Some of the components your report **may** include:

i. Problem statement and hypothesis.
ii. Description of your dataset and how it was obtained.
iii. Description of any data pre-processing steps you took.
iv. What you have learned from exploring the data, including visualisations.
v. How you chose which features to use in your analysis.
vi. Details of your modelling process, including how you selected your data analytic methods as well as the model through validation.
vii. Your challenges and successes.
viii. Key findings.
ix. Possible extensions or business applications of your project.

You should also create Appendices to your report that include:

i. **Code:** Sample commented Python scripts you have used to develop your project. The appendix code should not include code from the libraries – only some of the main commands you have used to call the Python libraries.
ii. **Data:** a sample of the data you used for your project, along with pointers to your data sources.

Note that while there is no page limit for References and Appendix, these sections should not include references, sample code and sample data. You should *not* add main text material to the appendices, such as results, as these will be out of the 10-page limit and will *not* be marked.

4. Marking criteria

The table below lists the criteria we will take into consideration in assessing your project report.

Criterion	Part of report	Evidenced by (at least)
#1	Introduction to the project and background information	Problem statement and hypothesis; project aims; concepts communicated; clarity.
#2	Literature review	Subject placed in the context of literature; a minimum of 6 references to journal papers, conference papers, or books (web references do not count in the 6 required).
#3	Data management	Data source; description of data; any pre-processing steps; any feature selection methods.
#4	Methods and/or methodology	Description of the method/s used; justification of the methods used.
#5	Analysis, testing, results	Includes possible cross-validation or any other approach to assess predictive accuracy; documentation of testing; analysis of the strengths and weaknesses.
#6	Completeness of the aims	Have the project aims been delivered?
#7	Concluding remarks	Discuss limitations and achievements; possible future improvements or directions; conclusion based on results.
#8	Appendices	Sample code presented clearly with comments (excludes code in libraries!), and a sample of the dataset used.
#9	Quality of report	Clarity; organisation; quality of the writing; quality and clarity of tables and figures; ease of understanding of the presentation of ideas.

Please note:

i. We do realise that each project is different. The marking scheme shown above is similar to the marking criteria we use for project dissertations and apply to any project.

ii. Marks will be adjusted for group size relative to overall project quality and effort as determined by the Lecturer.

5. Submission requirements for Groups

If you decide to form a group, EACH group member MUST submit the same report on QM+. The report submitted must be identical for ALL group members, apart from the cover page shown below.

A sample cover page is shown below. The cover page provides the option to each group member to specify their subjective opinion on the level of effort associated with each group member in terms of overall project effort. For example, if you are in a group of three people and everyone agrees that the contribution by each member was *roughly* similar (it will never be equivalent!), then the first page should look something like this:

Project title

Name 1 Surname 1

Group size: 3

Table of individual contribution by each member of my group, based on my subjective opinion:

Student	**Effort**
Name 1 Surname 1	33%
Name 2 Surname 2	33%
Name 3 Surname 3	33%

Please note:

i. The name of the person submitting the document should be provided both below the title as well as in the table of contribution. The names of ALL project members must be provided in the table of contribution as shown above.

ii. If everyone in the group agrees that the level of effort was adequate by all members of your group, then all group members should provide identical levels of effort for each of their group members.

iii. In cases where you report effort similar to 60-40 or 40-30-30, we **may** or **may not** take any action in adjusting the marks.

iv. In cases where you report effort similar to 30-70 or 60-20-20, and there is **a clear disagreement** between group members, we will invite ALL group members to discuss this further prior to any mark adjustment.

v. In cases where you report effort similar to 30-70 or 60-20-20 and there is a **general agreement** between group members, we may adjust marks without discussing this with the group.

vi. You should not claim on your cover page a 50-50 effort and later send an e-mail claiming a different level of effort. We will only consider the level of effort indicated on the cover page of your report.

vii. If one group member fails to submit their report before the deadline, they will still receive an automatic penalty by the system. The penalty will only apply to the individual group member. This happened last year and the student was penalised by the School despite the lecturers supporting the case of the student – so **be aware!**

6. Examples of project plan and timetable

This project is released in Week 2 and lasts for a total of around 8.5 weeks – up to week 11. To ensure the coursework runs smoothly, be careful not to deviate much from the following timetable:

Weeks 2 and 3: Determine whether to form a group, Question and Dataset

What is the question you hope to answer? What data are you planning to use to answer that question? What do you know about the data so far? Why did you choose this topic?

Example: We are planning to predict passenger survival on the Titanic. We have Kaggle's Titanic dataset with 10 passenger characteristics. We know that many of the fields have missing values that some of the text fields are messy and will require cleaning, and that about 38% of the passengers in the training set survive. We chose this topic because we are interested in the history of the Titanic.

Weeks 4, 5 and 6: Topic and data analytic methods

You may discover during your data exploration that you do not have the data necessary to answer your project's question. You may decide to change the research question to address in the project. You should aim to finalise any changes as soon as possible.

Our advice is to spend your time during the first few weeks wisely doing some research on the data sources and the data analytic methods covered in the labs, depending on the problem you are trying to address. Researching appropriate dataset and determining what data analytic method to use in the project is also a part of your coursework. Various data sources links are provided for reference at the end of this document.

Weeks 7 and 8: Data Exploration and Analysis

What data have you gathered, and how did you gather it? What steps have you taken to explore the data? Which areas of the data have you cleaned, and which areas still need cleaning? What insights have you gained from your exploration? Will you be able to answer your question with these data, or do you need to gather more data (or adjust your question)? How might you use modelling to answer your question?

Example: We have created visualisations and numeric summaries to explore how survivability differs by passenger characteristic, and it appears that gender and class have a large role in determining survivability. We estimated missing values for age using the titles provided in the Name column. We created features to represent "spouse on board" and "child on board" by further analysing names. We think that the fare and ticket columns might be useful for predicting survival, but we still need to clean those columns.

We have analysed the differences between the training and testing sets and found that the average fare was slightly higher in the testing set. Since we are predicting a binary outcome, we plan to use a classification method to make our predictions.

Weeks 9 and 10: Have produced your first draft (no, you do not need to submit the draft!)

At a minimum, this should include a) literature review and background information on your selected topic, b) narrative of what you have done so far, ideally in a format similar to a short dissertation, c) visualisations of the results, d) appendix with code and comments that explain the code.

Week 10 to 11: Finalise draft and submit the coursework.

7. Data sources

Using public data is the most common choice. If you have access to private data, that is also an option, though you will have to be careful about what results you can release. Some sources of publicly available data are listed below (you don`t have to use these sources).

- **UK Covid Data**
  https://coronavirus.data.gov.uk/
  Official UK COVID data

- **Data.gov**
  http://data.gov
  This is the resource for most government-related data.

- **Socrata**
  http://www.socrata.com/resources/
  Socrata is a good place to explore government-related data. Furthermore, it provides some visualization tools for exploring data.

- **US Census Bureau**
  http://www.census.gov/data.html
  This site provides information about US citizens covering population data, geographic data, and education.

- **UN3ta**
  https://data.un.org/
  UN data is an Internet-based data service which brings UN statistical databases.

- **European Union Open Data Portal**
  http://open-data.europa.eu/en/data/
  This site provides a lot of data from European Union institutions.

- **Data.gov.uk**
  http://data.gov.uk/
  This site of the UK Government includes the British National Bibliography: metadata on all UK books and publications since 1950.

- **The CIA World Factbook**
  https://www.cia.gov/library/publications/the-world-factbook/
  This site of the Central Intelligence Agency provides a lot of information on history, population, economy, government, infrastructure, and military of 267 countries.

- **Health Data**
  Healthdata.gov
  https://www.healthdata.gov/
  This site provides medical data about epidemiology and population statistics.

- **NHS Health and Social Care Information Centre**
  http://www.hscic.gov.uk/home
  Health datasets from the UK National Health Service.

- **Social Data**
  Facebook Graph
  https://developers.facebook.com/docs/graph-api
  Facebook provides this API which allows you to query the huge amount of information that users are sharing with the world.

- **Topsy**
  http://topsy.com/
  Topsy provides a searchable database of public tweets going back to 2006 as well as several tools to analyze the conversations.

- **Google Trends**
  http://www.google.com/trends/explore
  Statistics on search volume (as a proportion of total search) for any given term, since 2004.

- **Likebutton**
  http://likebutton.com/
  Mines Facebook's public data--globally and from your own network--to give an overview of what people "Like" at the moment.

- **Amazon Web Services public datasets**
  http://aws.amazon.com/datasets
  The public data sets on Amazon Web Services provide a centralized repository of public data sets. An interesting dataset is the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information. Also a NASA database of satellite imagery of Earth is available.

- **DBPedia**
  http://wiki.dbpedia.org
  Wikipedia contains millions of pieces of data, structured and unstructured, on every subject. DBPedia is an ambitious project to catalogue and create a public, freely distributable database allowing anyone to analyze this data.

- **Freebase**
  http://www.freebase.com/
  This community database provides information about several topics, with over 45 million entries.

- **Gapminder**
  http://www.gapminder.org/data/
  This site provides data coming from the World Health Organization and World Bank covering economic, medical, and social statistics from around the world.

- **Google Finance**
  https://www.google.com/finance
  Forty years' worth of stock market data, updated in real time.

- **National Climatic Data Center**
  http://www.ncdc.noaa.gov/data-access/quick-links#loc-clim
  Huge collection of environmental, meteorological, and climate data sets from the US National Climatic Data Center. The world's largest archive of weather data.

- **WeatherBase**
  http://www.weatherbase.com/
  This site provides climate averages, forecasts, and current conditions for over 40,000 cities worldwide.

- **Wunderground**
  http://www.wunderground.com/
  This site provides climatic data from satellites and weather stations, allowing you to get all information about the temperature, wind, and other climatic measurements.

- **Football datasets**
  http://www.football-data.co.uk/
  This site provides historical data for football matches around the world.

- **Pro-Football-Reference**
  http://www.pro-football-reference.com/
  This site provides data about football and several other sports.

- **New York Times**
  http://developer.nytimes.com/docs
  Searchable, indexed archive of news articles going back to 1851.

- **Google Books Ngrams**
  http://storage.googleapis.com/books/ngrams/books/datasetsv2.html
  This source searches and analyses the full text of any of the millions of books digitized as part of the Google Books project.

- **Million Song Data Set**
  http://aws.amazon.com/datasets/6468931156960467
  Metadata on over a million songs and pieces of music. Part of Amazon Web Services.