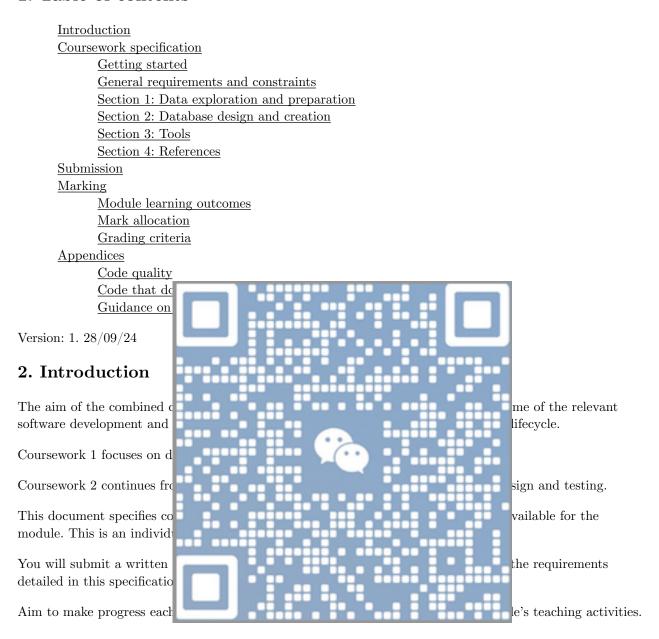
# COMP0035 Coursework 01 2024 Coursework specification

## 1. Table of contents



# 3. Coursework specification

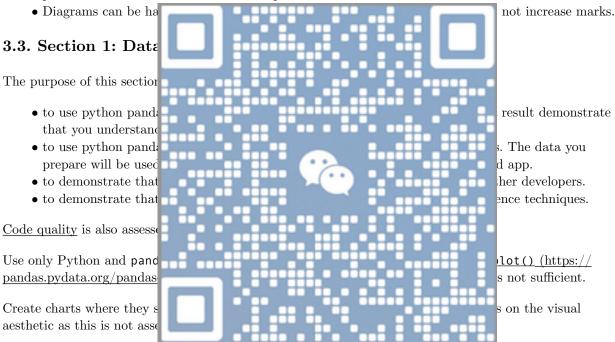
#### 3.1. Getting started

- 1. Select a dataset using the 'group' selection task in Moodle Week 1 (https://moodle.ucl.ac.uk/mod/choicegroup/view.php?id=6089982). Each 'group' option is associated with a data set. 'Group selection' is a Moodle term for the type of task, the coursework is **individual**.
- 2. Accept a GitHub classroom assignment. This creates the repository. Instructions are also given in Tutorial 1.
  - $1. \ Login \ to \ GitHub.com.$
  - 2. Click on the GitHub classroom link (https://classroom.github.com/a/zqVIaThf)
  - 3. Accept the assignment.
  - $4.\ \mbox{If prompted},$  accept to join the comp0035-ucl organisation.

3. Download the dataset for your group choice and add it to your repository. Use the links in Moodle (Resources > Datasets). For files > 25MB use <u>GitHub large file storage (https://docs.github.com/en/repositories/working-with-files/managing-large-files/about-git-large-file-storage)</u>.

#### 3.2. General requirements and constraints

- Compile all written work into a single report in either PDF or Markdown format. Name the document coursework1.
- The report supports the code and techniques used in the coursework. It is not an essay, be succinct. There are no word limits.
- Demonstrate regular use of source code control using GitHub. Create the repository using the GitHub classroom assignment. Keep the repository private. Keep the repository in the ucl-comp0035 organisation.
- You must use the data set allocated to you on Moodle.
- This is an individual coursework. Do not collude with other students using the same data set.
- Use of code AI tools is permitted when writing code. <u>UCL recommends using Microsoft Copilot (https://liveuclac.sharepoint.com/sites/Office365/SitePages/Bing-Enterprise-Chat.aspx)</u> using your UCL credentials. This must be stated in the 'References' section.
- Use relevant techniques from the course, or from data science and/or software engineering processes. Provide references for techniques not included in the course material.



You may need to prepare the data in order to complete the exploration and hence your code may not neatly split between 1.1 and 1.2. This is OK, the code structure does not need to exactly match the report structure.

#### 3.3.1. Section 1.1 Data exploration

- 1. Code: Write python code to **explore** and describe the data structure and content. Including, but not limited to, size, attributes and their data types, statistics, distribution of the data, etc. Consider potential data quality issues.
- 2. Report: Describe the results of your exploration of the data. Do not include the code in the report.

#### 3.3.2. Section 1.2 Data preparation

1. Report: Briefly describe a target audience and state at least 3 questions that they might be interested to explore using the data. This defines the purpose for which you will prepare the data.

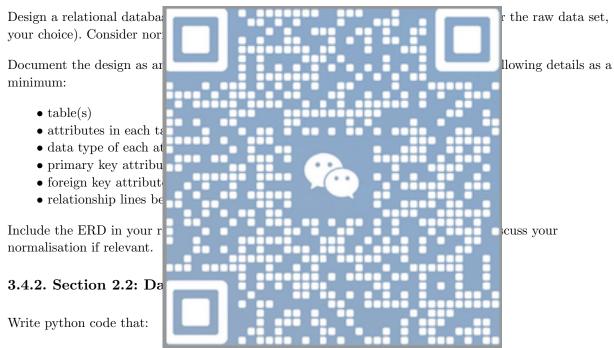
- 2. Code: Write python code to **prepare** the data such that it can be used to try to answer the questions for the audience described in step 1. Aim to have sufficient data, and avoid unnecessary data. The prepared data should be in a format that can be read into one or more pandas dataframes from a file (.csv or .xlsx). If relevant, address any data quality issues identified in section 1.1.
- 3. Report: Explain how you ensured the data is relevant for the purpose.
- 4. Include the original and prepared versions of your data set files in your repository.

# 3.4. Section 2: Database design and creation

The purpose of this section is:

- to demonstrate that you understand the structure of a relational database and the principles of normalisation by designing an appropriate database and drawing this as an entity relationship diagram (ERD).
- to demonstrate that you can write Python code to create an SQLite database based on the ERD. The database you create can be used in COMP0034 coursework in a data driven web application.

# 3.4.1. Section 2.1: Database design



- creates a database structure based on the ERD for an SQLite database file.
- takes the data from the dataset file and saves it to the SQLite database file. Note: do not create a database that requires a server such as MySQL or PostgresSQL.

The quality of the code is assessed.

Use relevant Python packages, i.e. pandas and sqlite3.

## 3.5. Section 3: Tools

The purpose of this section is to demonstrate appropriate and effective use of relevant software engineering tools.

#### 3.5.1. Section 3.1 Environment management

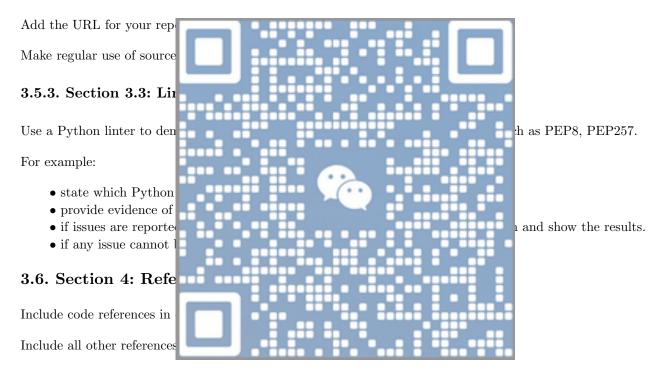
Provide relevant files and instructions that allow the marker to set up and run your code in a Python virtual environment. They will use pip and setuptools with the commands:

```
pip install -r requirements.txt
pip install -e .
```

As a minimum, edit the files that were provided in the starter code of the repository:

- requirements.txt: list the packages used in your code
- pyproject.toml: provide basic project details and code package location
- README.md: provide instructions to install and run your code for the data preparation and the database creation

#### 3.5.2. Section 3.2: Source code control



## 3.6.1. Section 4.1 Reference use of AI

State either that you used AI, or state that you did not.

If you used AI, include the details <u>stated in the UCL guidance (https://library-guides.ucl.ac.uk/referencing-plagiarism/acknowledging-AI#s-lg-box-wrapper-19164308)</u>.

#### 3.6.2. Section 4.1 Dataset attribution

Comply with any license condition required for your data set (given in the data set link in Moodle > Resources > Data sets).

Each license is different and tells you what has to be cited; e.g. see open government licence v3 (https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/). Typically, but not always, 'attribution' is required: i.e. include a statement listing who owns the data and its location.

# 4. Submission

Refer to Moodle > Assessment for the deadline date and time.

Submit your work on Moodle in the assignment submission. The submission states the upload format: .zip for the code (and report if in markdown) plus .pdf for the report (if not in markdown).

GitHub is **not** an acceptable alternative for submission, though its facility to download the code files as zip may be useful to you.

Make sure all files are in the submission. URLs linking to external files cannot be marked as they could be changed after the submission time. The only exception is where the original data files are too large to upload to Moodle - in this exceptional situation list url(s) to the data files in your report or the README.md instead.

Do not include your .venv folder in the zip file, this creates unnecessarily large zip files.

Table: Submission checklist

Section	Report	Code files
1. Data exploration and preparation	Descript explana	Python code to explore/describe the data.
2. Database design and creation	Entity Relatior Diagran	
3. Tools	Source c control: GitHub Linting	.txt,
4. References	Stateme use. Data se attribut Other re used.	es.

# 5. Marking

# 5.1. Module learning outcomes

The module's published learning outcomes that are assessed in this coursework are indicated in the table.