DASC509 Assignment 1 –

SQL coding exercise

|  |  |
| --- | --- |
| **Set Date:** | **18/10/2024 (9am)** |
| **Submission Date:** | **01/11/2024 (4pm)** |
| **Overall length:** | **1000 words + 10% (excluding code and tables)** |

# Assignment

Students will analyse a subset of synthesised data from CPRD structured as multiple tables in a database. The analyses will be completed by following the tasks below. The first task requires a short summary within a given word limit. For each other task, students should provide the code used to generate the results as well as providing their results in an appropriate manner (e.g. tables). The code provided should run without errors if entered into the pgAdmin environment.

## Aims

* To understand where the data have come from.
* To generate a type of data quality report.
* To be able to apply the knowledge from the SQL lectures/practicals to a new database.
* To present and communicate results clearly.
* To present code clearly and accessibly.

## The Task

1. Write a short summary about CPRD data, including the sources of the data and the content generally available. (200 words)
2. Create a list of all the tables in the database giving some basic information for each table:
   1. row counts.
   2. number of patients/individuals (where applicable).
3. Create summary tables describing the features of the patients in the dataset, including:
   1. Summary statistics for all numerical fields.
   2. Counts for all categorical or binary fields.
   3. Earliest and latest dates for all date fields.
4. Create a list of all the unique diagnoses which exist within the dataset.
5. Create a table of summary statistics based on the number of medications/treatments prescribed per patient prescribed in each year. The resulting table should have:
   1. Total number of medications prescribed each year.
   2. Total number of patients prescribed at least one medication each year.
   3. The range (minimum, maximum) of numbers of medications/treatments per patient for each year.
   4. The mean number of medications/treatments per patient per year.
   5. The median number of medications/treatments per patient per year.

### Guidelines for this task

* Writing:
  + Where requested you should write a concise response for each question detailing your findings. For most questions only tabulated results are required.
* Code:
  + You need to provide the code used for each question, preferably separate for each question.
  + **Code should be annotated to aid interpretation and to indicate which question it answers.**
  + Your code should be able to run on a computer with the database loaded into the pgAdmin postgreSQL environment.

## Assessment criteria

Marks will be apportioned as follows:

|  |  |
| --- | --- |
| **Marking Criteria** | **Available Mark** |
| Q1: Clearly written, informative summary of CPRD. | 10 |
| Q2: Accuracy of results and efficiency of code | 10 |
| Q3: Accurate results and efficiency of code | 20 |
| Q4: Accurate results and efficiency of code | 10 |
| Q5: Accurate results and efficiency of code | 20 |
| Presentation: Answers presented logically without spelling/grammar mistakes | 10 |
| Code: Code presented neatly and annotated clearly | 20 |
| **Total** | **100** |

### The usual guidance on preparation, submission and assessment of assignments applies - please refer to the MSc HDS Programme Handbook 2024/25

* The word-limits for MSc HDS assignments allow a +/-10% margin, and students must declare the word-count for each assignment. Students will have 5% deducted from their mark for not including a word count or for an incorrect word count. Students who exceed the word count will be penalised by the percentage over the word count. We will check the word count.
* The word-count excludes tables, graphs, line drawings, text-boxes, computer code and photographs, however these items must be ‘standalone’, i.e. have headings, captions, and legends that are carefully composed, succinct, informative and self-contained.  It is important to guide the reader through the main points of your argument/findings in the main commentary, mentioning tables/figures/text-boxes in passing to highlight/illustrate key issues/give further examples.
* Assignments should be presented double-spaced and with pages numbered.
* The front page should contain the module name, title of the assignment, date, word count and the student’s ID number at the bottom of the page as outlined in the assignment front sheet template. The student’s name should not appear on any page to facilitate blinding for marking purposes. Students are strongly advised to carefully proof read all material before submission, as marks will be deducted for badly presented work.
* Students are required to submit a copy to Canvas by 16:00 by the deadline stated. Assignments received after 16:00 will be treated as handed in a day late and will be subject to the 5% penalty according to the University's Code of Practice
* Students are advised to keep a copy of all submitted work.
* If students are unable to complete an assignment for unavoidable reasons, then permission in writing for an extension must be sought from the Module Lead or the Programme Director.

* If you have any queries about the assignment you should contact the Module Lead as soon as possible.

### The University’s standard penalties for the late submission of assessed work

The University’s standard penalties for the late submission of assessed work are:

* 5% of the total marks *available for the assessment* shall be deducted from the assessment mark for each working day after the submission date, up to a maximum of five working days (e.g. for work marked out of 100, five marks per day will be deducted; for work marked out of 20, one mark per day will be deducted); however, the mark will not be reduced below the pass mark for the assessment. Work assessed below the pass mark will not be penalised for late submission of up to five days.
* Work received more than five working days after the submission deadline will receive a mark of zero. In such circumstances, where a student is required to re-take the assessment, the re-assessment task must be different from the original assessment. Re-submission of the original piece of work is not permissible, except in the case of project work or dissertations.

(A working day is defined as a day when the University is open and staff would normally be available for work and thus also be available for contact by students).

Full information about the penalties for late submission of assessed work, including information about special circumstances and School/Departmental responsibilities, are available in section 6 of the University’s Code of Practice on Assessment:

## <http://www.liv.ac.uk/tqsd/code-of-practice-on-assessment/>

### Resubmissions

Students who submit work more than five working days after the submission date can resubmit an assessment but this will be counted as their second attempt. A combined mark of 50% is required to pass the module, provided that at least 40% is gained in each component. To calculate the overall module mark, the components are combined according to the weighting of each assignment. When an assignment has been failed, students are allowed to re-sit a failed assignment on one occasion only and this will be counted as their second attempt. The assignment title of the resubmitted work will normally differ to that of the original as stated in the University’s Code of Practice on Assessment.

Marks achieved through re-assessment will be capped at 50% for the purposes of calculating the overall average mark and determining classification for an award.