Clinical Data Management
Individual Coursework 1
2022

Overview and Instructions

Description

This coursework constitutes your first piece of assessment for the Clinical Data Management module. Its purpose is to improve your research abilities, test some of the MySQL skills that you have learned so far throughout the module and assess those same skills. For some of the questions, you will need to use commands that we haven't discussed in class. This is where your research skills come into hand.

What you will need

- MySQL server and MySQL workbench installed on your computer.
- Access to the internet to do some research.
- Your lecture notes and the practicals
- Data provided in the Data folder of the coursework. Please see the Data section below for further details.

What you will need to submit

To complete the assessment you must submit:

1. All of your answers written in a report (word document or pdf) with the corresponding MySQL code that allowed you to compute the answer to the question. Imagine that you are submitting this report to your stakeholders. It needs to look professional, have relevant sections, titles and so on. Marks will be awarded for report organisation and presentation.

2. One MySQL script containing the code to your answers. Please use the comment functionality to organise your code into different sections. Marks will be awarded for code organisation and code readibility.

Please zip your files into a folder titled *CDM_CW1_your_student_number* and upload your zipped folder rather than individual files. So if your student number is 012345 then it would be *CDM_CW1_012345*.

Important Note - Terms and Conditions

This coursework constitutes an **individual** piece of assessment. For some of the questions you will need to do research online to discover how to perform certain tasks and this is expected. You must ensure that the answers are strictly yours and that there is no copying and pasting from websites or from each other. Imperial College London has strict policy on academic misconduct. Please see the course handbook, which can be found on the course blackboard page, to understand what constitutes the different forms of academic misconduct.

How you will be assessed

You will be assessed against a specific rubric which awards marks for each correct answer, legibility of your MySQL code, adherence to basic MySQL conventions (as explained in the lectures) and report organisation and presentation.

Data

The data for this coursework is provided in the Coursework Resources Folder. It is a transformed subset of the data that you have already encountered in Week 1 of Translational Data Science - the London Paris study. You can revisit the materials from TDS to get a refresher on the data. The data is formed of three csv files:

- country_ids.csv - contains the country names with the corresponding country ids;
- sample_characteristics.csv - contains 15 characteristics of 1000 individuals; and,
- sample_genotypes.csv - contains the genotype data (5 SNPs) and the sample_id of 500 of the individuals represented in sample_characteristics.csv.

QUESTIONS

CREATING A DATABASE, TABLES AND IMPORTING DATA

Question 1

Create a database called `London_Paris_DB`. Provide the instructions or code to create the database. Explain how you discovered how to to do this. This can include screen-shots or commands. Please acknowledge all relevant sources.

Question 2

Using the relevant MySQL commands, create three "empty" tables called:

- country_ids;
- sample_characteristics; and,
- sample_genotypes.

Some tips for coding of the variables:

1. Code the attributes with the data types displayed in the table below:

CHAR(20)	INTEGER	DOUBLE
sample_id	SNP1	a_sta_smoke
country	SNP2	a_quit_smok
center	SNP3	age_recr
dodiag	SNP4	n_cigret_lifetime
dob	sex	
dor	cc_status	
dod	smoke_status	
country	vit_status	
country_name		

2. Do not forget to add all relevant primary keys and foreign keys.

Note that to specify foreign keys you might need to add a variation of the code we saw in the lectures as follows:

```
FOREIGN KEY (attribute_name) REFERENCES table_name(attribute_name);
```

Question 3

Use MySQL Workbench - not a MySQL command - to import **all three** csv files into their respective tables.

Tip for importing the data: When importing the data make sure to use the encoding UTF-16.

Notes for answer submission

For this answer please provide in your report, a sequence of instructions that allowed you to import the data into the relation sample_genotypes. You can use screen shots to help explain your answers in more detail. Note that you only need to provide the instructions for the sample_genotypes table.

EXPLORING THE DATA

Question 4

Create an ER diagram of the data schema and describe the difference between: 1) a red coloured key and a yellow colored key; 2) a full-colored diamond and an empty colored diamond. Please only present the ER diagram as your answer along with a short paragraph explaining what the different symbols represent.

Question 5

Use relevant MySQL queries to determine how many records are there in the following relations:

5.1 sample_characteristics

5.2 sample_genotypes relation

5.3 country_ids table

Question 6

6.1. Explain what the following code does:

```sql
SELECT
    SUM(CASE WHEN Attribute_1 IS NULL then 1 ELSE 0 END) as Attribute_1,
FROM table1
```

6.2. Which of the attributes in sample_characteristics contain NULL values.

6.3. How many NULLS do each of those columns - the ones you identified in 6.2 - contain?

QUERYING THE DATA

Question 7

7.1 Using a join of your choice, write one MySQL query that determines, for each country, what the average age of recruitment (`age_recr`) is, rounded up to two decimal places. Your query should return at least the following three columns:

- `country;`
- `country_name`
- `avg_age_country`

Once you have the results, please fill in the below table:

country	country_name	avg_age_country
1		
2		
3		
4		
5		
6		
7		
8		
9		
B		

7.2 Using only one MySQL query, determine the name of the country that has the largest number of samples. Order your results by descending order. Your query should return at least the following three columns:

- `country;`
- `country_name`
- `counts`

7.3 Which center has the least amount of samples?

7.4 Write a MySQL query of your choice to determine what is the country name of the sample with the youngest age at recruitment (`age_recr`).

7.5 Assuming that `1` represents female and `2` represents male, determine the name of the country that has the largest amount of male samples.

7.6 Which country has the largest number of distinct centers supplying samples to it? And how many distinct centers does it have?

7.7 Which country-center pair has the largest amount of samples?

CREATING NEW DATA FROM OLD DATA

Question 8

8.1 Using the tables currently available in your London_Paris_DB, write a MySQL query that creates a table called `sample_char_genotypes` and that contains, for the 500 samples for which there are genotype data, the following attributes:

- sample_id
- dob
- sex
- cc_status
- age_recr
- country
- country_name
- SNP1
- SNP2
- SNP3
- SNP4
- SNP5.

8.2 Has the newly created table inherited any of the primary keys or any of the foreign keys from the tables used in its construction? Use the functionalities of MySQL or those of MySQL Workbench to provide an answer to this question and, in one or two sentences, explain how you investigated this.

8.3 Once you have created the table export it to a csv file called `sample_char_genotypes.csv`. Please provide simple instructions on how you achieved this.

CONNECTING R TO MYSQL

Question 9

9.1 Explain how you can connect your MySQL server to `R` so that you can read data directly into `R`. Please provide instructions and/or code that you used to do this.

9.2 Connect `R` to MySQL using your instructions and provide the code that allows you to bring in the table `sample_genotypes` into `R`.

GOOD LUCK!