RMIT UNIVERSITY

School of Science

ISYS1055/3412 (Practical) Database Concepts

Assessment 4: Database Design Project

Assessment type: Take-home assessment   **Word limit:** N/A

Draft Due Date: Sunday, 28 May at 23:59 (Melbourne Time) – Week 12 (otherwise no 4 mark bonus)

Final Due Date: Sunday, 11 June at 23:59 (Melbourne Time) – Week 14

Weighting: 35% (35 Marks)

Important Note: Students who submit the draft by the draft due date will receive up to 4 additional marks on top of their final assessment 4 grades.

Overview

This is a practical and real-world project that puts the knowledge you gained into practice. You are required to investigate and understand a publicly available dataset, design a conceptual model for storing the dataset in a relational database, apply normalisation techniques to improve the model, build the database according to your design and import the data into your database, and develop SQL queries in response to a set of requirements.

The objective of this assignment is to reinforce what you have learned in the whole course. Specifically, it involves how to build a simple application that connects to a database backend, running a simple relational schema.

Part A: Understanding the Data (0 Marks, Preliminary Work)

Part B: Designing the Database (10%)

- Task B.1 Produce an ER diagram for a relational database that will be able to store the given dataset.

Part C: Creating the Database and Importing Data (10%)

- Task C.1 Produce one SQL script file
- Task C.2 Create a database file and import the given dataset into your database.

Part D: Data Retrieval and Visualisation (15%)

- Task D.1-D.5 Produce one SQL query file that includes five SQL queries, produce a PDF file that includes the running result screenshot of five queries.
- Task D.1-D.5 Represent each query result as a graph and include this graph in the PDF file for query result.
- Task D.5 is for ISYS1055 students only; not required for ISYS3412 students.

Assessment criteria

This assessment will measure your ability to:

- Analyse the requirements outlined in the problem description

- Develop a conceptual model for the design of a database backend required for the system

- Use an industry-standard ER modelling tool to draw the ER model

- Use 7-step mapping process to create relational database schema

- Use normalisation process to evaluate the schema and make sure that all the relations are at least 3NF

- Create tables on SQLite Studio and populate them with data available from the specified sources.

- Write SQL statements required for CRUD (create, read, update and delete) operations on the database you built

- Develop your knowledge further to represent data in a meaningful way using data visualisation.

Course learning outcomes

This assessment is relevant to the following course learning outcomes:

CLO1	Describe the underlying theoretical basis of the relational database model and apply the theories into practice;
CLO2	Explain the main concepts for data modelling and characteristics of database systems.
CLO3	Develop a sound database design using conceptual modelling mechanisms such as entity-relationship diagrams.
CLO4	Develop a database based on a sound database design;
CLO5	Apply SQL as a programming language to define database schemas, update database contents, and to extract data from databases for specific users' information needs.
CLO6	Create then populate a normalised database based on a publicly available dataset, and visualise the results of queries against the dataset.

RMIT
UNIVERSITY

Assessment details

Part A: Understanding the Data

In this assignment, we are working with the publicly available dataset: **A Global Database of COVID-19 Vaccinations**. Further details about this dataset are available in the article available through the following URL: https://www.nature.com/articles/s41562-021-01122-8. The abstract of the article is as follows.

An effective rollout of vaccinations against COVID-19 offers the most promising prospect of bringing the pandemic to an end. We present the Our World in Data COVID-19 vaccination dataset, a global public dataset that tracks the scale and rate of the vaccine rollout across the world. This dataset is updated regularly and includes data on the total number of vaccinations administered, first and second doses administered, daily vaccination rates and population-adjusted coverage for all countries for which data are available (169 countries as of 7 April 2021). It will be maintained as the global vaccination campaign continues to progress. This resource aids policymakers and researchers in understanding the rate of current and potential vaccine rollout; the interactions with non-vaccination policy responses; the potential impact of vaccinations on pandemic outcomes such as transmission, morbidity and mortality; and global inequalities in vaccine access.

A live version of the vaccination dataset and documentation are available in a public GitHub repository at https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations. These data can be downloaded in CSV and JSON formats.

For the purposes of completing this assignment, we are only using the following files. You are required to review and analyse the dataset available in these files. You will find that reviewing the rest of the files, even if not listed below, will help you to form a better understanding about the big picture.

	FILE NAME	DESCRIPTION
1	locations.csv	Country names and the type of vaccines administered. Each line represents the last observation in a specific country. Refer to README.md for the details.
2	us_state_vaccinations.csv	History of observations for various locations in the US.
3	vaccinations-by-age-group.csv	History of observations for vaccinations of various age groups in each country.
4	vaccinations-by-manufacturer.csv	History of observations for various types of vaccines used in each country.
5	vaccinations.csv	Country-by-country data on global COVID-19 vaccinations. Each line represents an observation date. Refer to README.md for the details.
6	country_data/Australia.csv	Daily observations of vaccination in Australia.
7	country_data/United States.csv	Daily observations of vaccination in the US.
8	country_data/England.CSV	Daily observations of vaccination in England.
9	country_data/New Zealand.csv	Daily observations of vaccination in New Zealand.

Table 1: List of data files

To complete the tasks in the following sections, you are required to review and analyse the dataset that is available in the named files.

Part B: Designing the Database (10%)

Task B.1 Produce an ER diagram for a relational database that will be able to store the given dataset.

It is important to note that the given CSV files are not necessarily representing a good design for a relational database. It is your task to design a database that will adhere to good design principles that were taught throughout the course. This means your database schema will not match the structure of the CSV files and, therefore, you will require to manipulate the structure of the dataset (and not the data itself) to import it into your database. Importing the data is required to complete Task C.2.

The ER diagram must be produced by [Lucidchart](#) similar to the exercises that were completed in in the course. UML notation is expected and using other notations will not be acceptable. Including a high-quality image representing your model is important, which can be achieved using Export function of Lucidchart.

You are also required to transform the ER diagram into a database schema that will be used in the next part of the assignment.

Creating a good database design typically involves some database normalisation activities. You should document your normalisation activities and support them with good reasoning. This typically involves explaining what the initial design was, what the problem was, and what changes have been made to rectify the issue.

The expected outcome of completing this task is one PDF file named model.pdf containing the following sections.

1. Database ER diagram and, if needed, a reasonable set of assumptions.

2. Explanation of normalisation challenges and the resulting changes.

3. Database schema.

Part C: Creating the Database and Importing Data (10%)

Task C.1 Produce one SQL script file named database.sql. This script file requires all the SQL statements necessary to create all the database relations and their corresponding integrity constraints as per your proposed design in Part B. The script file must run without any errors in SQLite Studio and contain necessary commenting to separate various relations. Note that this script is not supposed to store any data into the relations.

The expected outcome of completing this task is one script file with the specific name of database.sql.

Task C.2 Create a database file named Vaccinations.db. Import the given dataset into your database.

To complete this task, you may need to change the format of the CSV files to match the attributes of your designed database. You can use a spreadsheet editor such as Microsoft Excel.

The next step is to *import* the spreadsheets into the database you create in SQLite Studio. To complete this task, use the menu option *Tools – Import* in SQLite.

The expected outcome of completing this task is one database file named Vaccinations.db, which must contain all the data that is stored in the CSV files named in Table 1.
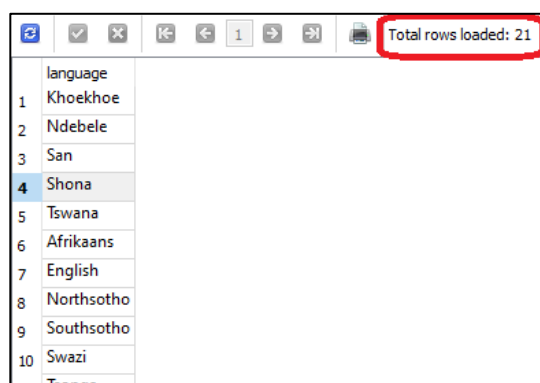
Part D: Data Retrieval and Visualisation (15%)

Now that you have created and populated a database, it is time to create some queries to investigate the data in various ways. In addition to writing the required queries, you are also asked to produce data visualisation for the results of your queries.

The tasks in this section represent the queries that must be supported. Each query must consist of one SQL statement. It would be acceptable to use several nested queries, combine several SELECT statements with various operators etc. However, it would not be acceptable to have multiple and separated queries for each task.

After you have written each query, you are expected to produce a data visualisation for each result set. You have the freedom to choose the tool for creating your visuals (e.g., Excel, Google Charts, Tableau) as well as the visualisation techniques (e.g., charts, plots, diagrams, maps). Completing this portion of the work will require that you understand the nature of the results of each query, undertake research to choose a visualisation tool you are comfortable with, decide about the best technique to visually represent each result set, and produce the visualisation. **Answers to tasks in Part D that are not supported by a visualisation can achieve up to 80% of the grade associated with each task.**

The expected outcome of completing this task is as follows.

1. One SQL script file named Queries.sql containing *all* the queries developed for the tasks in this section. It is important that you add comment lines to separate the queries and indicate which task they belong to. Note that valid SQL comments must not generate errors in SQLite Studio. The marker of your work will use this file to execute and test your queries.

1. A PDF file named QuerieResults.pdf containing the following elements for each task.

    a. The SQL query

    b. a snapshot of the first 10 results of your query. The snapshot must also show the total number of results retrieved by the query. A sample snapshot is provided below for your reference.

Figure 1: Sample results snapshot with total rows

    c. Data visualisation. This must be represented as a graph or chart that presents the results in a meaningful and easy to understand manner.

List of Tasks

Task D.1 List the **running total** of the number of people who have taken the vaccine **so far** for each observation day recorded in the dataset for each of all countries - this isn't the number of vaccines each day. DN should be the day number relative to the first day of vaccinations for each country (starting from day 0).

Each row in the result set must have the following structure.

(3 marks for ISYS1055; 4 marks for ISYS3412)

Country Name (CN)	Administered Vaccine on Day Number (DN)	Total Injected People

Figure 2: Column Headers in the Result Set for Task D.1

Task D.2 Find the countries with the biggest cumulative numbers of COVID-19 doses administered by each country (Note: the result may include multiple countries or a single country). Produces a result set containing the name of each country and the cumulative number of doses administered in that country.

Each row in the result set must have the following structure.

(3 marks for ISYS1055; 4 marks for ISYS3412)

Country	Cumulative Doses

Figure 3: Column Headers in the Result Set for Task D.2

Task D.3 Produce a list of vaccine types with the countries taking each vaccine type. For a vaccine type that has been taken in multiple countries, the result set is required to show several tuples reporting each country in a separate tuple.

Each row in the result set must have the following structure.

(3 marks for ISYS1055; 3 marks for ISYS3412)

Vaccine Type	Country

Figure 4: Column Headers in the Result Set for Task D.3

Task D.4 There are different data sources used to produce the dataset. Produce a report showing the largest total number of vaccines administered according to each data source (i.e., each unique URL). Order the result set by source name (URL).

Each row in the result set must have the following structure.

(3 marks for ISYS1055; 4 marks for ISYS3412)

Source Name (URL)	Largest total Administered Vaccines

Figure 5: Column Headers in the Result Set for Task D.4

Task D.5 How does various countries compare in the speed of their vaccine administration? Produce a report that lists all the observation months in 2022 and, for each month, list the total number of people *fully vaccinated* in each one of the 4 countries used in this assignment.
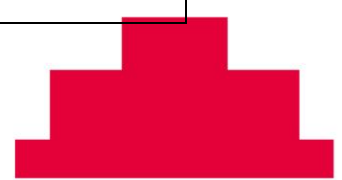
Each row in the result set must have the following structure.

(3 marks for ISYS1055. this question is for ISYS1055 only; not required for ISYS3412)

[Date, Australia, United States, England, New Zealand]

Date	Australia	United States	England	New Zealand

Figure 6: Column Headers in the Result Set for Task D.5

Timeline:

Consider the following timeline to help manage your time to successfully complete this Project

Timeline	Activities
Week 8-9 Part A	• Review this assignment, the assessment criteria & rubrics. • Familiarise yourself with the Harvard Referencing style. • Go to the website provided in the instructions for Part A to familiarise yourself with the dataset you are working with. - "A Global Database of COVID-19 Vaccinations". • Read the article for further details on this dataset. • Watch video/attend live session on how to get started, develop the ER model, create the relations, format the CSV file using Microsoft Excel and import the data into the database.
Week 10-11 Part B	• Work on Part B of the assessment. • Get up to date with Lucidchart to produce the ER diagram. • Create a PDF and create sections as mentioned in the assessment instructions for Part B.
Week 12 Part C	• Work on Part C of the assessment + submit draft (part A + B + partial C). • Create relations, format CSV files according to database design and import data.
Week 13 Part D	• Work on Part D of the assessment. • Download a data visualisation tool that you will be using e.g. Excel, Google Charts, Tableau.
Week 14	• Revise / fine-tune your assessment including the files required for submission.

Assessment Criteria

Your report will be assessed on the following criteria:

- Analyse the requirements outlined in the problem description
- Develop a conceptual model for the design of a database backend required for the system
- Use an industry-standard ER modelling tool to draw the ER model
- Use the 7-step mapping process to create relational database schema
- Use the normalisation process to evaluate the schema and ensure that all the relations are at least 3NF
- Create tables in SQLite Studio and populate them with data available from the specified sources
- Write SQL statements required for CRUD (create, read, update, and delete) operations on the database you built
- Develop your knowledge further to represent data in a meaningful way using data visualisation

Learn how to use rubrics for assessment in Canvas.

Submission Format

You are required to submit the files with the exact names as below.

1. Model.pdf
2. Database.sql
3. Vaccinations.db
4. Queries.sql
5. Queries.pdf

In the previous sections of the assignment, the expected content of each of the files is explained in detail.

Referencing guidelines

Use RMIT Harvard referencing style for this assessment.

You must acknowledge all the courses of information you have used in your assessments.

Refer to the RMIT Easy Cite referencing tool to see examples and tips on how to reference in the appropriated style. You can also refer to the library referencing page for more tools such as EndNote, referencing tutorials and referencing guides for printing.

Academic integrity and plagiarism

Academic integrity - When you submit work, it must be your own. Learn how to avoid plagiarism and cheating

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge, and ideas.

You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e., directly copied), summarised, paraphrased, discussed, or mentioned in your assessment through the appropriate referencing methods.

- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites.

If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct.

Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source

- Copyright material from the internet or databases

- Collusion between students

For further information on our policies and procedures, please refer to the University website.

Assessment declaration

When you submit work electronically, you agree to the assessment declaration.

Study support

Visit the 'Setting up for Success' module in the course to get help with referencing, writing skills, study skills, finding information, group work, and more. You can also get connected to one-on-one support with an Academic Skills Advisor, a librarian, or a peer mentor.