RMIT UNIVERSITY

School of Science

ISYS2095 Database Concepts

Assessment 3: Database Design Project

Assessment type: Take-home assessment		**Word limit:** N/A
Due Date: 21 August at 23:59 (Melbourne Time)		
Weighting: 35%, 35 Marks		

Overview

You are required to investigate and understand a publicly available dataset, design a conceptual model for storing the dataset in a relational database, build the database according to your design and host the data, 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: Part A: Understanding the Data (0 Marks, Preliminary Work)

- Part B: Designing the Database (10%)

- Part C: Creating the Database (10%)

- Part D: Data Retrieval (15%)

Assessment criteria

This assessment will measure your ability to:

- Analyse the requirements outlined in the problem description;

- Develop a conceptual model to assist you with the design of the 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 sources outlined above.
- write SQL statements required for CRUD (create, read, update and delete) operations on the database you built.

Course learning outcomes

This assessment is relevant to the following course learning outcomes:

CLO1	Describe various data modelling and database system technologies.
CLO2	Explain the main concepts for data modelling and characteristics of database systems.
CLO3	Identify issues with, compare and justify relational database designs using the functional dependency concepts.
CLO4	Apply SQL as a programming language to define database schemas and update database contents.

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.

	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/France.csv	Daily observations of vaccination in France.

| 9 | country_data/Israel.csv | Daily observations of vaccination in Israel. |

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 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.

The expected outcome of completing this task is an ER diagram produced by Lucidchart, which may also be accompanied with a reasonable set of assumptions. The ER diagram must be saved as a PDF file named model.pdf.

Part C: Creating the Database (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. The script file must run without any errors in SQLite Studio. 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.
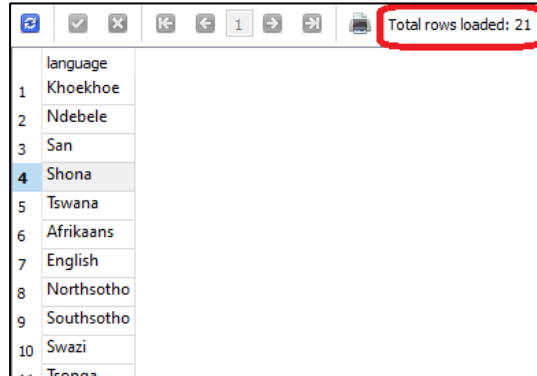
Part D: Data Retrieval (15%)

The following queries are to be supported. Each one of the queries below must be one SQL statement. It is fine to use several nested queries, link several SELECT statements with various operators etc. However, it is not acceptable to have multiple and separated queries for 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 <u>Queries.pdf</u> containing the query for each of the following tasks together with 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

Task D.1 For a given country (e.g., Afghanistan), list the total number of vaccines administered in each observation date recorded in the dataset.

Task D.2 Produces a result set containing cumulative number of COVID-19 doses administered by each country. That is, the name of each country and the cumulative number of doses administered in that country.

Task D.3 Produce a list of all countries with the type of vaccines (e.g., Oxford/AstraZeneca, Pfizer/BioNTech) administered in each country. For a country that has administered several types of vaccine, the result set is required to show several tuples reporting each type of vaccine in a separate tuple.

Task D.4 There are different sources of data used to produce the data set. Produce a report showing the total number of vaccines administered according to each data source.

Task D.5 How does various countries compare in the speed of their vaccine administration? Produce a report that lists all the observation dates and, for each date, list the total number of people *fully vaccinated* in each one of the 4 countries used in this assignment.

[Date, Australia, United States, France, Israel]

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 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.