CSCI 2141 Winter 2025 Course Project – Find data, build a database, do interesting things

In this project you will design, create, and use a database to ask interesting questions of some data.

Importantly, the parts of the project that include code need to **work**. The TAs and I need to be able to execute your queries and modify them in interesting ways.

This is an overview document; details for each project part will be released well in advance of the due date.

Due Dates:

Submissions of the relevant project parts are to be made through Brightspace, with the following deadlines:

Project Part 1: Friday, February 7, 2025, at 11:59PM
Project Part 2: Friday, March 7, 2025, at 11:59PM
Project Part 3: Monday, April 7, 2025, at 11:59PM

Late submissions will be accepted with a 10% penalty up to 24 hours after the deadline. Submissions more than 24 hours after the deadline will not be accepted.

Datasets:

You can choose from one of the provided datasets or produce your own. But either way, your database must satisfy the following requirements:

Dimensions

- The final conceptual schema (Part 2) must contain at least four entities (i.e., you must create at least four tables);
- Each entity must contain several attributes of different types (character, integer, floating point, etc.);
- At least one of the tables must be derived from a real dataset (probably retrieved from an online source). Please see “Where Can I Find the Data?” below for elaboration on this.
- At least two of the tables must have a decent number (>100, potentially a lot more) of rows. Other tables can be smaller.

Data Use

You must have the rights to acquire, use, and submit the data for classwork. Ideally your data would fall under some type of Creative Commons license or equivalent; you can always check with us if you are unsure. If it matters, submission for course credit does not count as republication or redistribution for the purposes of licensing. You should at the very least

include a citation to your data, but please include the license information where possible. If you can't find any license information, please indicate so.

Where Can I Find Data?

Public, open-source datasets are in huge supply, so finding a relevant table should be very easy. I provide a few examples in the Appendix. Things get more difficult when you're trying to find multiple entities that connect naturally to one another in a conceptual model.

A movie database can provide a good example: you could have a conceptual model that includes the entities "MOVIE", "ACTOR", "GENRE", "COMPANY", and "COUNTRY". It's easy to imagine connections between these and to come up with queries that could make use of data across multiple tables. But public datasets with four or more nice tables are relatively rare.

If you can find a dataset with four or more tables, then great. But as long as the dataset you retrieve has at least one table (== one entity in the conceptual model), then you can **build the rest of your dataset to the project requirements** by generating your own data. This can be done in a couple of ways:

- Literally creating your own data: making tables (Excel sheets, tab-separated files, whatever) that add to your conceptual model. Maybe you downloaded a dataset of movies from Kaggle but you need to create some fake data about actors, genres, etc.

- Using our good friend generative AI to produce the necessary data. For example, I tried the prompt below and was able to recover a list of 50 actors with dates of birth, countries of origin, etc.

  > *"Please create a comma-separated list of actors, with a unique identifier, first and last names, dates of birth, and other potentially relevant information."*

- You **may** use ChatGPT or some other AI model to generate this kind of data for you

- You **may** ask it to generate some interesting columns for you, as I did above.

- You **may not** ask for help in designing the database (e.g., suggesting constraints or datatypes for attributes, providing models and schemas, writing SQL statements, or establishing relationships between tables via foreign keys). This is your job for the project; you may not seek help from other students or AI tools.

Modifications

You can use the dataset as provided / obtained, but you can also modify aspects of it to, for example, add difficult cases, create real or randomly generated attributes, or add information for missing cases you would like to include. You may also need to do some cleaning of the data before loading it into the database (or after). Please document any modifications you make for any part of the project.

Project Structure

The project is split into three parts outlined below:

Part 1 (5 points): Planning

- **Choose** a dataset of suitable complexity and size.
- **Describe** the dataset.
- **Create** a conceptual schema and an internal schema (by hand or in MySQL), marking possible keys, constraints, and data types.
- **Identify** some limitations of the initial schema.

Part 2 (10 points): Design and creation

- **Modify (if necessary)** the previous schemas to a higher normal form (probably 3NF);
- **Write out** the appropriate CREATE TABLE statements;
- **Run** basic queries that use SELECT and JOIN

Part 3 (10 points): Advanced queries and procedures

- **Construct** a range of queries, including table joins and procedural SQL.
- **Submit** the data with all table creation scripts.
- **Document** the structure and functionality of the database.

Our Expectations for Academic Integrity

This is an **individual project** and all submitted work must be your own. If you are uncertain about something, it is better to **ask** for clarification.

Things you **must** do:

- Cite your data sources (websites will almost always have information about how to cite their data). The citation style is less important (there are plenty of examples out there) than consistency.
- Write your own code and text.

Things you **may** do:

- Discuss ideas with your colleagues
- Seek input about data sources from others, including ChatGPT (for example, I asked it to recommend wikis that offer SQL data dumps)

Things you **must not** do:

- Submit **any** code or text that you did not create / write yourself
- Have someone else design your database for you
- Use data in ways that violate licensing agreements
- Use data or perform analyses that violate Dalhousie's Code of Conduct.

Be Creative!

Although we have certain specific expectations, this is an opportunity for you to explore datasets that are of interest to you and to deploy your skills in a realistic situation. Students often put excellent projects (class-based or outside of class) up on their Github pages, which can be a big differentiator when applying for jobs (as long as your code is good and well-documented).

Keep in mind that it doesn't need to be complicated; as long as you meet the minimum requirements the submission will be fine. We certainly like to see more-ambitious projects with much larger datasets, but this is not essential.

Appendix: Data Sources

There is a stunning array of open data sources on the Web; many countries, districts, and non- government organizations have their own open data portals. In many cases the data can be retrieved in .csv format through a direct download. Feel free to challenge yourself, but keep in mind that database performance will degrade as datasets get larger.

Meeting the four-table criterion will be challenging, and you need to consider how you might link disparate data sources (by country? By region? By person name?

There are many open map, image, etc. data sets out there, but these require skills that you will not acquire in this introductory course. I recommend you avoid these.

Statistics Canada 2021 Census Data -
https://www150.statcan.gc.ca/n1/en/type/data?portlet_levels=98P%2C98P10#tables

Navigate to any specific dataset, choose “Download options”, and choose “CSV Download entire table “…”. This will retrieve all data and metadata. You may want to restrict to a smaller subset of the data (province, etc.) to allow for database joins, etc. to take a reasonable amount of time.

Kaggle:

https://www.kaggle.com/datasets?fileType=csv&sizeEnd=100%2CMB&minUsabilityRating=8.00+or+higher

Kaggle (registration required) is a popular repository of datasets for training and testing machine-learning classifiers. It has an enormous number of datasets; the above REST-like link will filter for datasets with high usability scores with csv files. There is even a “dataset of datasets” .csv:
https://www.kaggle.com/datasets/jessevent/all-kaggle-datasets?select=kaggle_datasets.csv

Other suggestions:

Other interesting portals include the Halifax, Nova Scotia, and Canada data portals.

https://www.worldometers.info/sources/ has a **huge** list of data sources by countries that are parsed by the site.

Some lists of open data sources:

15 sources: https://careerfoundry.com/en/blog/data-analytics/open-data-sources/
50 sources: https://learn.g2.com/open-data-sources