Data bases 1
2025–2026
M1

Project guidelines

The idea of the project for the Databases 1 course is to create a database from a chosen data set. In particular, it will be necessary to detail and discuss the choices made in terms of : modelling and database population. You will also be asked to present some significant queries and realistic modifications to the database. The project must be carried out in pairs. In addition, a written report must be delivered, with all the codes and data needed to reproduce the work performed. All this material must be delivered by 8/12/2025 (via Moodle) and then an oral presentation of the project will be given on 10/12/2025 (the time available for each pair and the exact time will be communicated once the pairs are defined).
Obviously, the project code must be done in Python. The code in Python must be well commented. For the report and slides, you can use the tools of your choice.

For the project, please send an e-mail by 19/11 stating with whom you will do the project.

1 Data for your project

You can choose to make your project about what you like. There is a lot of data available online. Choose the one you think is suitable. However, think that you have to model a database with several tables and then run queries. So you need different data potentially taken from one csv (particularly large) or several csv (perhaps from different sources) that you can combine in your project. Attention : you must provide the link or source of your data ! It is therefore pointless to take something that was already a structured database. In this first part, the aim is to be creative and demonstrate that you can identify data that can be structured in a relational db.

An example of data you can use is given here.

— Data on bridges and cities in California : https://www.kaggle.com/datasets/camnugent/california-housing-feature-engineering/data?select=cal_populations_city.csv and https://www.fhwa.dot.gov/bridge/nbi/ascii.cfm

If you decide to use exactly this data, know that you will be evaluated not as highly on this first part of the project as your colleagues who will have chosen something different.

If necessary, if you find a db that interests you but is too large, you can decide to take only a subset of the rows.

We will have some time dedicated to the project in class, so please feel free to discuss your project idea for feedback.

2 Modelling the database

In this part of the project, you must decide and argue how you will structure the data you have in a relational db. You will have to describe the existing tables, the choices made on attributes, the constraints imposed in the db and so on. Obviously also build on what you have seen in class on CDM and LDM. The aim is to be able to justify and defend the choices made, as well as to explain them.

3 Population of tables

This is the part where you will potentially have to do the most research. Here, you must decide how to handle any problems (missing data, inconsistent data, merging of different databases,...) that you may encounter once you go to populate your db. An example might be if you decide to have a db on crimes in some American states (for example) and you find census data from different states and want to merge them into your project. Perhaps you have some codes that do not match ? Some redundant information ? Some missing ? How you handle them is up to you, explain the choices you make. Obviously, by searching for data integration online, you may find several hints. As far as this project is concerned, once again, let us look for correct reasoning that does not alter or change too much the data we are entering and which we will then interrogate. Again, refer to what we saw in class and explain the process and the choices made.

4 Querying your database

In the last part, go on to propose queries that make sense in your db. For each one, explain what you want to achieve, how and if there are alternative solutions. You must obviously present queries of increasing difficulty. Also propose some changes to be made to the data or the structure of your db (as might happen in real life). Of course, you can even expect that new data will become available and that it will be necessary to integrate them into the db. Then discuss other queries that could then be made later and whether those already proposed remain valid or need to be modified.

At this point, you will have realised, the aim of the project is to simulate a person coming to you with some raw data and asking you to create a relational db to meet their needs. This must of course take into account possible modifications that may become necessary over time.

Good work !