Individual Integrative Assignment (60%)

Introduction to the case: Data management for public safety

Ensuring public safety is the key goal at any Police Department (PD). Fighting crime and guaranteeing the safety of the city is a very difficult and challenging task. Yet, PDs are increasingly sharing their data and trying to enhance the power of data analytics to learn more about crime trends and improve public safety.

Imagine that you work at the Data Management Unit of a PD. You need to provide them with support to develop an efficient database to manage and analyze the crime data that will capture their needs. In addition, you are responsible to perform exploratory querying and reporting to address important questions about

i. Crime trends in relation to quantity, type of crimes and arrests
ii. How such crime trends vary in time, over certain periods of the year or years
iii. How such crime trends vary by location and/or type of locations
iv. How such crime trends vary per police units and districts

You are responsible to select and decide what kind of questions to prioritize. Those questions will define the scope your assignment. To achieve that, it is fundamental that you create a database using the provided data that will allow you to answer those questions by interrogating the data.

This assignment consists of a series of tasks (see all details in the following pages) and writing a final report including the outcomes of all tasks.

Important notes before you start

- Please, carefully read this document before starting with Task 1. It is important to have a clear overview of what to expect.
- Tasks relate to one or more topics covered on a given week of the course. Specific references of how a task relates to lectures/tutorials/weeks are indicated in the assignment file (see Table 1). Lectures might offer opportunities to further discuss the assignment by topic, as shown in the table.
- Ideally, tasks should be completed in a sequential order. Yet, you might and/or need to go back to some tasks and refine them at any time before submission.
- This assignment was designed to be feasible within the given time and it assumes students work on it on a weekly basis. You must complete the assignment on time (penalty is applicable for delays) and no exception will be made. ***Please note that Week 5 (deadline of the assignment) will be especially busy as other courses will also have other assignments due that week.***
- This assignment requires efficient time management and self-planning. You can choose your own pace in completing the tasks. Yet, it is highly recommended to work on tasks *vis-à-vis* the topics covered weekly. For suggestions on how to organize self-study, see 'Overview of activities and deadlines' on Canvas (Modules > Course Information).

Expectations

Students are expected to complete the assignment **individually** to the best of their own ability, without seeking or accepting the help of others or use resources that are not explicitly allowed nor to help others. All assignments will be checked for plagiarism upon submission. Please be aware that violation of these terms will be considered fraud.

What is / is not allowed for this assignment?

- Students ARE ALLOWED to consult and use (whether applicable) materials from lectures, video lectures, readings, textbooks, or Internet resources. If students include such materials in the report (e.g., as references) it is mandatory to cite them appropriately using APA style and including a bibliography at the end of the report.
- Students ARE NOT ALLOWED to collaborate in the assignment. Note that several different versions of the datasets were created so outcomes and procedure might vary depending on the assigned dataset.
- Students ARE NOT ALLOWED to communicate with others about how to solve a task, what SQL concepts and syntax are required, etc.. Students can ask clarification questions about the assignment via the Canvas discussion forum or in class; questions should be limited to general matters (and not solutions) that can be beneficial all students.

If there is suspect that students do not respect these requirements, action will be taken accordingly.

Software requirements for this assignment:

- **BD Browser for SQLite** (as used in the SQL tutorials and in class, see instructions for installing on Canvas). This will be useful for Tasks (1), 5, 6.
- A software for data exploration and cleaning for Tasks 1 and 4. There exists several options you can choose to perform these tasks and you can use them in combination. You can use **Excel** for basics procedures, and/or **Stata/R** for any basic to advance operation. In class, we will also learn how to use OpenRefine, an open-source, powerful tool to work with messy data.
- For Task 4 (Data Models), you might want to use an open **software to create the ERDs** (e.g., https://www.diagrams.net/)
- For Task 6 (reporting), besides using the implemented features in **DB Browser**, you can also create tables using other software (**Excel**). If you like (optional), you can also make use of visualizations/dashboards to report your results using **R/ Tableau/Power BI**. If you intend to use R, these two chapters from the Watson (2022)'s book might be useful for you:
  - on data visualizations: https://www.richardtwatson.com/open/Reader/_book/data-visualization-1.html
  - on dashboards: https://www.richardtwatson.com/open/Reader/_book/dashboards.html

Overview of provided data

For this assignment, you will work on public data. More specifically, on an adapted and sampled subset of the crime records database of the city of Chicago (for the original source and general information, see https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2). You will focus on crime records that occurred between 2017 to 2021. While the original dataset for the period 2017 – 2021 contains more than 1 million records, you will deal with a random sample of such dataset (~730,000 obs). This choice was made to make the assignment more feasible for individual work and to make all operations requiring software easier to run in any laptop (with computer memory available to do the operations of at least 2 GB).

The dataset has been prepared and manipulated accordingly in line with the assignment objectives. **Hence, note that the data you are working on does not fully resemble the original source, so please refrain from considering any outcome from this assignment as mirroring the truth.** In addition, while the variables (columns) included in the provided dataset are identical for all students, there exists several versions of the dataset that vary by record (row) level. This means that outcomes of querying and reporting will vary among students.

You can find an overview of the dataset and short explanation in Table 1.

Table 1. Overview of the dataset.

Variable	Short explanation (adopted from original source, see link above)
CrimeID	Unique identifier for the crime record.
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
Date	Date when the incident occurred. This is sometimes a best estimate.
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.
IUCR	The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e
Primary type	The primary description of the IUCR code.
Description	The secondary description of the IUCR code, a subcategory of the primary description.
Location Description	Description of the location where the incident occurred.
Arrest	Indicates whether an arrest was made.
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74
District	Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r
Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

Assignment tasks

Table 2 shows the list of tasks you have to perform in the assignment and related points. Full details of each task are provided separately below.

Table 2. Assignment tasks.

Task	Description	Refer to materials covered in:	Points
1	PLAN & EXPLORE: Define questions for the project and familiarize with the data	Lecture 1	10
2	DESIGN AND ORGANIZE: Identify and collect the necessary requirements for the database based on the questions, data and information needs.	Lecture 2	15
3	DATA MODELS: Translate the identified requirements into a relational database design: create the conceptual, logical and physical models (ERD models) and take care of normalization (to the extent to which normal forms are possible).	Lecture 2	20
4	DATA PROCESSING: Performing data quality checks	Lecture 3	15
5	IMPLEMENTATION: Implement your database using DB Browser and load data using the provided tables.	SQL basic commands + SQL Part 4	10
6	QUERYING AND REPORTING. Carry out queries on the implemented database to answer your questions. Report and briefly discuss your results.	Lecture 4; SQL Parts 4-6	30
Total			100

Task 1. PLAN & EXPLORE: Define questions for the project and familiarize with the data

Data management starts on Day 1 with a lot of planning. It is fundamental that you 'understand the business and the data': that is, to match the goal of your project with the needs of the CPD and the data you have available. This means that, for this task, you have to proceed in an iterative way between making a plan for your project (what questions do you want to focus on) and exploring the data.

For this task, first of all you have to make sure you understand the context and overview of the case. As explained above, your Data Management Unit's areas of interests are:

i. Crime trends in relation to quantity, type of crimes and arrests
ii. How such crime trends vary in time, over certain periods of the year or years
iii. How such crime trends vary by location and/or type of locations
iv. How such crime trends vary per police units and districts

You have to identify and formulate at least **3 questions** based on these areas. At least **1 of your questions** should address the **time** evolution aspect of the data.

Before formalizing your questions, explore the data and familiarize with it. Upload the provided CSV file in any software for data exploration (e.g., OpenRefine, R, Stata; you can even use DB Browser). It is important that you explore and fully understand the table schema, the columns/variables you are dealing with, and identify what data could help you in answering the identified questions.

Last, **write your introduction paragraph of your report and outline the problem definition and your questions.** Briefly explain **why the identified are relevant for the PD.**

Your questions should form a coherent and consistent framework that can translate your findings into relevant information for PD.

Task 2. DESIGN and ORGANIZE: Identify and collect the necessary requirements for the database based on the questions, data and information needs.

After understanding the business and the data, it is time to move to the design phase and think about data structures and identify the necessary requirements for the database. For this task, **write a section in your report where you address the following requirements questions**:

- What are the entities of importance to design this database?
- What variables should be included/excluded in line with your questions? Motivate your choice.
- What about **normalization procedures?** Can you already foresee potential issues that you must solve when developing the data models (Task 3) of the database? Important questions to answer:
  - Does the table conform to 1NF? If not, which columns should be changed and in what way in order to conform to 1NF? Which column(s) is/are the primary/foreign key(s)?
  - If you change the table, does it conform to 2NF? If not, what should be changed?
  - Next, does the table conform to 3NF? If not, what should be changed?

Task 3. DATA MODELS

Translate the identified requirements into a relational database: create the **conceptual, logical and physical models (ERD models)** and take care of **normalization issues** (to the extent to which normal forms are possible). Also, briefly discuss how your ERDs address the normal forms as identified in Task 2.

- Draw the **conceptual model** that contains the entities identified in Task 2. Describe all your entities. Every entity should have at least one relationship connecting it to another entity and all entities need to be linked to each other (even if just through a third entity). In addition, do not forget to include cardinalities (i.e., full relationship information). Explain the meaning of the relationships in the conceptual ERD using relationship sentences.
- Draw the **logical model** by choosing the appropriate primary keys (mark them with '(PK)", relevant attributes and data types. Briefly explain your choice of primary keys and be sure to define the meaning of each attribute.
- Draw the **physical model** to be implement in DB Browser SQLite by resolving all many-to-many relationships. In addition to marking the primary keys with "(PK)", mark all foreign keys with "(FK)".

TASK 4. DATA PROCESSING: Data quality checks

Performing data quality checks is one of the most important best practices in data management. In real life, data is always messy and needs to be cleaned before running any

analysis. Before you implement your database in DB Browser (Task 5), **perform a data quality check** on the original dataset following the key data quality dimensions studied in the course. You can use any software (or combination) for data cleaning and pre-processing (R, Stata, Excel, OpenRefine). In the report, **write a section in which you document and explain all operations**.

Export the final (clean) dataset and make sure you include all the necessary columns. The original file is Tab-separated. When exporting your cleaned file, you can keep this field separator or change it, but be mindful of the nature of the variables (some variables you might to keep in the dataset might include commas).

TASK 5. Implement your database in DB Browser and load (clean) data.

- Create a new database in DB browser and save it as a Project file as indicated ad the end of this document.
- Write SQL statements to create each table of your physical ERD, including correct data types and setting all needed constraints. Explain any choice you made about constraints. *Tip: when setting the constraints, in particular the FKs, the order of creating table is very important to avoid incurring in foreign key constraints. If you need to refresh how to set FK, you can check https://www.sqlitetutorial.net/sqlite-foreign-key/.*
- Populate the database tables using the clean data by writing appropriate SQL statements. When importing the clean dataset into DB Browser, make sure to be consistent in using the field separator and double check that datatypes are correct (sometimes types are set by default, but they might be wrong).

TASK 6. Carry out queries on the implemented database to answer your questions. Report and briefly discuss your results.

Using DB Browser, write **queries** to answer your questions. Some questions might require more than one query to be answered. All queries must be correctly and appropriately used for the goal. Write all the queries in separate "Execute SQL" **tabs** and label them with clear titles. Be sure to save all your queries saved also as a **txt.file** for backup and submit it together with the other deliverables on Canvas.

Important requirements for queries. Your set of queries **must** include:

- regular expressions, scalar functions and analytical functions (of which, at least 1 window function);
- date/time data;
- at least 2 JOIN operations;
- 1 view;
- 1 index: motivate why you created it and show proof that SQLite uses it;
- 1 trigger, you can choose between:
  - Create a trigger that prevents updates on columns that you consider important and cannot be changed.
  - Create a trigger that records operations into a log table (e.g., insert, update, delete data) if you plan to manipulate the database in any way after it is implemented.

In the report, **write a results section structured by question**.

- Briefly mention what SQLite queries you used to answer the questions (refer to the specific 'Execute SQL' tab, do not report the code in the report) and, when required, **motivate the choices you made**.
- **Report key outputs of your queries to provide evidence and answer the questions**. You can either use tables/screenshots from DB Browser or export the output and use other software for reporting and visualization. You can use the Plots function integrated in DB Browser; R, or other free DM visualization tools, such as Tableau or Power BI- limited user versions). SQL can be easily implemented in other environments, so feel free to explore your adventure there.
- Besides reporting outputs, **describe your findings and link them to the questions**.

Last, write a short **conclusion** of the obtained results so to provide an answer to your questions.

Outputs required at submission

The assignment requires the submission of the following items:

1) A **database project file (sqbpro)** as "Student Lastname_Name_student number_databaseproject" including all the needed data and tables as well as commented SQL Syntax. Syntax should be saved in the 'Execute SQL" tabs AND also into **separate txt file** for backup. Submitting these files is of extreme importance because, during grading, we will run all the queries you provided to check whether they are correct. If we miss something, you will lose points.
2) Final **report** as 'Student Lastname_Name_student number_report' either PDF or .doc format. Reports should not exceed **4000 words** (excluding references).

All outputs must be submitted on Canvas via the provided link by Sunday 2nd October (23:59). This deadline must be considered as ultimate date for submission. This implies:

- Late work policy. Students are expected to turn the assignment on time. There will be no exceptions. Late work is not accepted and will penalize the final grade. There will be an increasing penalty for every day past a missed deadline, as follows: 1 day after deadline: 25%; 2 days after deadline: 50%; 3 days after deadline: 75%. Example: if students submit within 1 day after the deadline, 7 becomes 7*0.75 = 5.25.
- Students are welcome to submit the completed assignment at any time/day before the deadline.

Grading

The grade is calculated based on the points received per task. All tasks will be examined against the expected answers and points will be given per task as explained in Table 2. 100 points equal to a grade of 10.