Practical 01
Data Exploration

Data Exploration in SQLite

We will be using SQLite for this exercise; you can do this on your own machine or at https://sqliteonline.com/.

1. Download the file **titanic.db** file using this link: https://github.com/PaulHancock/COMP5009_pracs/raw/main/data/titanic.db
2. Load the database into your sqlite3 browser of choice.
3. Create a query which will SELECT all columns from the *manifest* table, and LIMIT the results to the 10 rows
4. Examine each of the column names and content and determine an appropriate data type for each.
5. Look at the database schema and compare the data types with those that you suggested above.
6. Determine whether there are any missing values in this data set
    a. Note which columns have data which IS NULL
    b. COUNT the number of entries which IS NULL and which IS NOT NULL
7. For each of the numeric data columns create a query which *aggregates* the data to find:
    a. The minimum value
    b. The maximum value
    c. The average value
    d. The sum of all values
8. Create a query that will return all passengers with a name that is LIKE "<something>Dr.<something>"
    a. Modify this query to COUNT number of doctors and GROUP the results BY sex
9. Create a query that will return the average ticket cost
    a. Modify this query to show this average GROUPed BY the different classes of ticket
    b. Further modify this query to GROUP BY the *embarkation* port.

Remember to record your work in a logbook so that you can refer to it later in the course.

Useful references:

Software Carpentry intro to SQL: https://swcarpentry.github.io/sql-novice-survey/

Short data description: https://www.kaggle.com/c/titanic/data?select=train.csv

SQLite documentation: https://www.sqlite.org/docs.html