Data cleaning and preparation

Data cleaning and preparation (readiness)

Importing data into tables

Data cleaning and preparation (readiness)

- Preparing data for a database system is a time-consuming process
  - Data must be correct, complete and consistent
  - It requires time to clean and prepare the different tables in the data set
  - Data usually comes from other repositories and files
  - It's an investment

- For the project, you need to provide at least 20 records for table, so you can test your different queries with enough data (20 records is in the low range)

- Four things to take into consideration (next slides)

Importing data options

Data cleaning and preparation (readiness)

Four things to take into consideration:

1. Prepare your data in Excel (or similar tool) and generate .csv files for each table, then import/upload each table.
   - That's the usual way people prepare data (there are of course more advanced tools).
   - You may write insert statements but is a lot, not very efficient in terms of time (and potential errors)

2. Use data already available for music, movies, audiobooks on internet.
   - Sample data for music is provided from a music data website (see dataset on Carmen).
   - Do the same for the other media.

Importing data into tables

3. Not all the tables may need 20 records.
   - Some table, because of its nature, need only few records (.e.g. sex, race, artist type, etc)
4. Use services to generate random data
   - e.g. https://mockaroo.com/ , http://randat.com/ , https://www.onlinedatagenerator.com/, https://generatedata.com/
   - The latest version of SQLiteStudio has this option now (I haven't tested).

- Other websites to visit to consider for future ideas:
  - https://www.softwaretestinghelp.com/test-data-generation-tools/
  - https://www.onlinewebtoolkit.com/generatedata
  - https://download.cnet.com/GS-DataGenerator/3000-2092_4-10303373.html