Lab 2: SQL DDL & DML

This is an individual lab. Each student must complete all work independently.

Lab Assignment

In this lab you will create and populate relational tables for several datasets using SQL's Data Definition Language (DDL) and Data Manipulation Language (DML).

Datasets

1. **AIRLINES** This graph dataset stores information about a number of airlines and the flights these airlines operate between different airports.
2. **KATZENJAMMER** The database contains information about the musical career of a pop band named Kaztenjammer from Norway. The band consists of four members, each of whom sings and plays a variety of instruments. The dataset details each band member's contribution to each song recorded and performed by the band.
3. **BAKERY** This OLTP (on-line transaction processing) dataset represents sales at a small bakery. The dataset captures the notions of a transaction (a single purchase) and market baskets (each purchase may contain more than one item).
4. ***CUSTOM*** This dataset is up to you to create! Find an interesting, non-commercial, non-copyrighted dataset on the internet that may be mapped to no fewer than 3 tables. You may use a small slice of a large dataset but you must define at least 3 tables that fit together in a cohesive way, including appropriate foreign key constraints. Create your custom tables using SQL DDL. Choose appropriate data types and enforce all relevant constraints. Populate the tables with at least 25 records in total (across all tables.) Possible sources of data include:
   (a) https://github.com/awesomedata/awesome-public-datasets
   (b) https://datasource.kapsarc.org
   (c) https://www.re3data.org/
   (d) https://www.kaggle.com/datasets
   (e) https://catalog.data.gov/dataset
   (f) https://www.usaspending.gov

Guidelines

1. Choose appropriate ANSI SQL data types for each column.
2. You must properly detect and declare **all** constraints, including primary key, candidate key (SQL's `UNIQUE` column constraint), and referential integrity/foreign key constraints.
3. Each provided dataset comes with a README file which describes data files included in the dataset and briefly explains the meaning of the dataset. Before starting your work on a dataset, **please carefully review the README file and make sure you understand the structure of the dataset.** All data files in each dataset are stored in CSV (comma-separated values) format.
4. You are encouraged to use scripts (Python, `sed` / `awk`, etc.) to generate `INSERT` statements. You do not need to submit your scripts with this lab. Submit only the SQL DDL and DML you produce.

Submission Instructions

Submit via Canvas a total of 9 files as described below, in a single zip or tar/gzip archive.

1. `BAKERY-setup.sql` : DDL statements (`CREATE TABLE` / `ALTER TABLE`) to define the structure and constraints for all tables in the BAKERY dataset.
2. `BAKERY-populate.sql` : `INSERT` statements to populate all BAKERY tables.
3. `AIRLINES-setup.sql` *(as above)*
4. `AIRLINES-populate.sql` *(as above)*
5. `KATZENJAMMER-setup.sql` *(as above)*
6. `KATZENJAMMER-populate.sql` *(as above)*
7. `CUSTOM-setup.sql` *(as above)*
8. `CUSTOM-populate.sql` *(SQL INSERT statements only, do not include raw CSV/XML/Excel/etc. data files)*
9. `CUSTOM-detail.txt` : Discuss the following in several paragraphs:
   (a) Source of the data (URL, name of the person or organization who produced the data)
   (b) A brief description of the tables you defined and the relationships between them.
   (c) Any mapping challenges you may have encountered.
   (d) Three *non-trivial* information requests / queries that can be answered using the dataset you chose