# Lab 2: SQL DDL & DML

**Due Date:**

Monday, October 14th, 5:00pm

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

## Lab Assignment

Use your course MySQL account for this lab (and all future labs.) You will find access instructions on PolyLearn.

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 100 different airports.
2. **BAKERY** This OLTP (on-line transaction processing) dataset represents one month of 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).
3. **INN** This OLTP dataset holds one year of hotel reservation data for a small Bed & Breakfast that features 10 uniquely named rooms. The dataset specifies for each reservation the checkin and checkout dates, the name of the person making the reservation and the number of adults and children who occupied the room for the duration of the reservation.
4. **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.
5. ***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 related 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. Create these tables using SQL DDL. *Choose appropriate data types and constraints.* Populate the tables with approximately 50-100 records in total (counting 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

## Important Guidelines

1. Column names are your choice. However, you must use a *consistent* naming scheme throughout this lab.
2. Choose *appropriate* ANSI SQL data types for each column.
3. You must properly detect and declare **all** constraints, including primary key, candidate key (SQL's `UNIQUE` column constraint), and referential integrity/foreign key constraints.
4. For the standard datasets – BAKERY, AIRLINES and INN – table names must match CSV files of the dataset one-to-one.
5. Each standard 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. All string values are enclosed in single quotes.
6. You are welcome (and encouraged) to explore the use of scripts 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 PolyLearn a total of 11 files as described below, in a single zip or tar/gzip archive.

1. `AIRLINES-setup.sql` : DDL statements (`CREATE TABLE` / `ALTER TABLE`) to define the structure and constraints for all tables in the AIRLINES dataset.
2. `AIRLINES-populate.sql` : `INSERT` statements to populate all AIRLINES tables.
3. `BAKERY-setup.sql` *(as above)*
4. `BAKERY-populate.sql` *(as above)*
5. `INN-setup.sql` *(as above)*
6. `INN-populate.sql` *(as above)*
7. `KATZENJAMMER-setup.sql` *(as above)*
8. `KATZENJAMMER-populate.sql` *(as above)*
9. `CUSTOM-setup.sql` *(as above)*
10. `CUSTOM-populate.sql` *(SQL INSERT statements only, do not include raw CSV/XML/Excel/etc. data files)*
11. `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* questions you might ask based on the dataset you chose