San José State University
Department of Applied Data Science

DATA 225
Database Systems for Analytics

Spring 2023
Section 22
Instructor: Ron Mak

Assignment #2

Assigned:	Monday, February 6
Due:	Monday, February 13 at 5:30 pm
	Team assignment, 100 points max

Create and load database tables with CSV data

The purpose of this assignment is to give you practice using Python code in a Jupyter notebook to create and load some database tables. You will perform ETL with CSV data.

Search the internet for one or more interesting datasets in the form of CSV files. A good search strategy is to Google "datasets for analysis". You can mix and match datasets.

Once you've identified some candidate data:

- Design MySQL database tables hold the data. Use Python code to create the tables and to perform the ETL. Limit the size of each table to 100 rows or fewer.
- Create tables that support the following relationships (at least one example of each relationship):
  - One-to-one
  - One-to-many
  - Many-to-many

For each table:

- Create it in a Jupyter notebook. You can have one notebook that creates all the tables, or a separate notebook for each table.
- Identify the primary key, and if it has any, the foreign key(s). These can be described in comments of your Jupyter notebooks.
- Extract, transform, and load (ETL) data with the Python code of your notebook(s).
- Display in a dataframe the first 25 rows (or fewer if the table is smaller than 25 rows) by making a query from Python. An example SQL command:

```
SELECT * FROM my_table LIMIT 25
```

- Perform at least 5 SQL queries from Python that join your tables with one-to-one, one-to-many, and many-to-many relationships. Use the examples at http://ies-ads-classdb.sjsu.edu/dbclass/ (Show SQL) as guides.

What to submit

- URL(s) of your CSV data source(s).
- An SQL dump (export) of your loaded database that we can use to recreate it.
- The `.ini` configuration file for your database.
- Your notebook(s) with instructions on how to run them if it's not obvious. Include the output cells.

TIP: Clear all the output, restart the kernel, and run all the cells before submitting a notebook.

Submit one or more zip files containing the above to Assignment #2 in Canvas — *only one submission per project team*.

Rubric

Criteria	Max points
• Database tables that demonstrate one-to-one, one-to-many, and many-to-many relationships with primary keys and any foreign keys identified.	• 20

• One or more Jupyter notebooks that perform:
  ○ Table creation and data insertion or loading.
  ○ Queries that display up to 25 rows of each table.
  ○ At least 5 queries that join your tables with one-to-one, one-to-many, and many-to-many relationships and produce results.
	• 70
	
	  ○ 15
	  ○ 15
	  ○ 40

• Your CSV data source(s), an SQL dump of your database, and its configuration file.	• 10