CS 327E Milestone 2, due Sunday, 02/09.

1. Finalize your choice of primary dataset (aka `dataset1`). Your dataset should be made up of multiple CSV files from the same source and the data elements in these files should be related via a primary key to foreign key relationship. If you've made any changes to your selection since Milestone 1, update your `DATASETS.txt` file so that it reflects the current state.

2. Create a bucket in Google Cloud Storage (GCS) and a folder for your `dataset1`. Upload the files for `dataset1` into this folder. If you need help with this step, refer to this guide.

3. Create a new Jupyter notebook. This notebook should be named `<source>_ingest.ipynb` where is the source of your `dataset1` (e.g. musicbrainz, fda, bls, etc.)

4. Implement the following logic in your Jupyter notebook:

    - Make a BQ dataset for storing the staging tables for `dataset1`. The dataset should be named `<source>_staging`.

    - Import the CSV files from GCS into your new dataset in BQ. Ensure that you import each file into its own table.

    - Check that each table was loaded correctly. Visually inspect the contents of each table by selecting a few sample records.

    - Write some simple queries to explore the data in the tables. You can start off by running these exploratory queries in the BQ Console. Once you have identified a few interesting queries, add them to your notebook. You should have at least 1 query per table. The queries should have a `WHERE` clause and `ORDER BY` clause.

    - Add a short comment above each SQL statement to describe the query.

CS 327E Milestone 2 Rubric

Due Date: 02/09/20

Finalize selection of primary dataset for project. `dataset1` should be described in a file named `DATASETS.txt` (named exactly like so, no extensions) and each dataset should meet the following criteria: - Be available to you as multiple CSV files - Contain multiple related files connected via a primary to foreign key **-20** no `DATASETS.txt` file found **-20** dataset is made up of only one file, or no files connect via a primary to foreign key	20
Jupyter notebook `<source>_ingest.ipynb` containing the ingestion pipeline for `dataset1`, as described in the outline. **-40** no Jupyter notebook found in group's repository **-30** no dataset present in group's BQ project **-10** incorrect naming convention for Jupyter notebook **-10** incorrect naming convention for BQ dataset **-10** inconsistent naming conventions across tables **-15/10** each missing table load **-10/5** each missing sample records from table	40
Jupyter notebook `<source>_ingest.ipynb` containing SQL queries for `dataset1`, as described in the outline. **-40** no queries found in Jupyter notebook **-20** no queries use a `WHERE` clause **-20** no queries use an `ORDER BY` clause **-20/15** each un-queried table, depending number of tables **-10** each missing comment **-5** each incorrect comment, or comment too similar to query	40

`submission.json` submitted into Canvas. Your project **will not** be graded without this submission. The file should have the following schema:
 	
`{`
`"commit-id": "your most recent commit ID from Github",`
`"project-id": "your project ID from GCP"`
`}`
 	
Example:
 	
`{`
`"commit-id": "dab96492ac7d906368ac9c7a17cb0dbd670923d9",`
`"project-id": "some-project-id"`
`}`
	Required

Total Credit:	**100**