Milestone 2, due Friday, 09/20.

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 each dataset. Upload the files for each dataset into their folder. Refer to guide for steps.

3. Create a dataset in BigQuery (BQ) for your `dataset1`. Name your dataset `<source>_staging` where `<source>` is the source of your data (e.g. fda, bls, etc.). Refer to guide for steps.

4. Import the files from GCS into BQ as separate BQ tables:
   - Ensure each table is created in the appropriate dataset
   - Use schema auto detection for quicker import
   - Use consistent naming convention for all tables
   - Refer to guide for steps
   - Switch to the Classic UI when troubleshooting data load issues

5. In the BQ Console, write some simple queries to explore the data:
   - Keep track of the SQL from the most interesting queries.
   - Copy the SQL queries into a `simple-queries.sql` file.
   - Add a short comment above each SQL statement to describe the query. Comments should begin with a `"--"` (e.g. `--this is a legal comment in SQL`).
   - You should have at least 1 query per table. The queries should have a `WHERE` clause and `ORDER BY` clause.

CS 327E Milestone 2 Rubric

Due Date: 09/20/19

Import datasets into BigQuery **-15/10** each missing table **-30** no datasets present **-10** inconsistent naming conventions across tables	30
Include SQL queries that explores the BigQuery data, and explain what the queries do. Each query should be preceded by a comment that explains what it does. All queries should be written and saved in `./simple-queries.sql` in your group's repository. **-40** `./simple-queries.sql` not found in repository **-20** no queries use the `WHERE` clause **-20** no queries use the `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
Finalize datasets from Milestone 1. Dataset 1 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 **-30** no `DATASETS.txt` file found **-20** dataset is made up of only one file, or no files connect via a primary to foreign key	30

`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**