Final Project Milestone 1, due Thursday, 11/05.

1. Make your dataset selections for the Final Project. Please refer to the dataset listings slide for details on which datasets are approved for this project. Create a file called `DATASETS.txt` and with your selection. Include the following details in this file:
    - URL(s) to the data you downloaded for each dataset.
    - Interesting entities and attributes you noticed in the data.
    - 3-5 lines of sample data for each entity of interest.
    - Very important: explain what insights you hope to gain from exploring the data. If you don't know *why* you want to look at this data, you should stop now and think about your objectives or look for a different pair of datasets for this project.

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 our guide for steps. Note: you don't need to copy your datasets to GitHub.

3. Create a Jupyter notebook and name it `milestone1.ipynb`. Implement the following tasks from your notebook.

4. Create a BQ dataset for each of your datasets. Name your BQ dataset `<source>_staging` where `<source>` is the source of your data (e.g. fda, bls, noaa, imdb, etc.).

5. Import the CSV files for each of dataset into BQ:
    - Ensure each file is imported as its own table
    - Ensure each table is created in the appropriate dataset based on its source
    - Use schema auto-detection if possible, otherwise specify schema
    - Use consistent naming convention across tables

6. Write some SQL queries to explore the data:
    - Come up with at least 10 queries, 5 per dataset.
    - Each query should include at least 3 clauses from this list: JOIN, WHERE, GROUP BY, HAVING, ORDER BY, LIMIT.
    - Each query should be preceded by a Markdown comment that explains its function.
    - Queries should also include joins across datasets if possible. If not possible, add an explanation to `DATASETS.txt` on what conversions you plan to perform in order to implement cross-dataset joins at a later date.

CS 327E Final Project Milestone 1 Rubric
Due Date: 11/05/20

Primary and secondary datasets chosen from approved list should be described in a file named `DATASETS.txt` (named exactly like so, no extensions). **-30** no `DATASETS.txt` file found in repository **-10** missing interesting entities and/or attributes **-10** missing sample data on entities of interest **-10** missing explanations (objectives and necessary conversions for implementing cross-dataset joins)	30
Import your selected datasets into BigQuery (BQ) **-30** no datasets present in BQ **-7** for each dataset named incorrectly **-7** for each missing table in BQ **-5** for each table loaded incorrectly from `milestone1.ipynb` (missing records, missing columns, load command errors) **-5** inconsistent naming convention across tables	30
Write 10 SQL queries that explore the data. Each query should use 3/6 clauses. Each query should be preceded by a Markdown comment that explains its function. **-40** queries missing from `milestone1.ipynb`: **-20** no queries use the `WHERE` clause **-20** no queries use a `JOIN` clause **-20** no queries use a `GROUP BY` clause **-20** no queries use a `HAVING` clause **-20** no queries use the `ORDER BY` clause **-20** no queries use a `LIMIT` clause **-5** each incorrect comment, or comment too similar to query	40
`DATASETS.txt` and `milestone1.ipynb` pushed to your group's private repo on GitHub. Your project **will not** be graded without this submission.	**Required**

`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**