Final Project Milestone 1, due Thursday, 04/15.

1. Make your dataset selection for the Final Project. Please refer to the approved datasets listing and selection guidelines from lecture slides.

2. Create a file called `DATASETS.txt` and the following details:
    - The names of the two datasets you chose (e.g. IMDB and Bollywood).
    - The URLs to the download sites for both datasets.
    - The interesting entities and attributes you noticed in the data.
    - Very important: explain what insights you hope to gain from exploring your two chosen datasets. If you don't know *why* you want to look at this data, you should stop and think about your objectives or look for a different pair of datasets for this project.

3. Create a bucket in Google Cloud Storage (GCS) with a folder for each dataset. Upload the files for each dataset into their respective folders. Refer to our guide for steps. Note: you should not add your datasets to your git repository.

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

5. 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.).

6. Import the CSV files for each dataset into their respective BQ dataset:
    - With the exception of the COVID daily reports, each CSV file should be imported into its own staging table in BQ
    - Use schema auto-detection if possible, otherwise specify a schema
    - Use STRING types when a stricter type causes parsing errors (e.g. DATE, NUMERIC, etc.)
    - Use consistent naming across your tables

7. Write some SQL queries to explore your BQ datasets:
    - Come up with at least 10 queries, ~5 per dataset.
    - Each query should include 4/6 of these clauses: JOIN, WHERE, GROUP BY, HAVING, ORDER BY, LIMIT.
    - Each query should be preceded by a Markdown comment that explains its function.
    - Queries should include joins across datasets if possible. If joins are not possible due to incompatible data formats, add a note to `DATASETS.txt` that explains which tables and fields you tried to join on and what types of transforms will be needed to implement cross-dataset joins at a later date.

CS 327E Final Project Milestone 1 Rubric
Due Date: 04/15/21

Primary and secondary datasets chosen from the approved list should be described in a file named `DATASETS.txt` (named exactly like so, no extensions). **-20** no `DATASETS.txt` file found in repository **-10** missing interesting entities and/or attributes **-10** missing explanations (objectives and necessary conversions for implementing cross-dataset joins)	20
Import your selected datasets into BigQuery (BQ) **-15** for each missing dataset in BQ **-3** for each dataset named incorrectly **-5** for each missing table in BQ **-5** for each table loaded incorrectly from `milestone1.ipynb` (missing records, missing columns, load errors) **-2** inconsistent naming convention across tables	30
Write 10 SQL queries that explore the data. Each query should use 4/6 clauses. **-5** for each missing query or query missing one or more required clauses **-3** for each query missing output in notebook **-2** for each incorrect or missing comment above query **-2** for omitting cross-dataset joins and missing explanation in `DATASETS.txt`	50
`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**