Milestone 9 due Friday, 11/15.

Part 1:

- Find a second dataset (aka `dataset2`) in CSV format that meets our dataset requirements.
- Add a description of your dataset to the existing `DATASETS.txt` file.
- Create a new folder in your Cloud Storage bucket and upload the dataset files to this folder.

Part 2:

In the following section, `<source>` refers to the source of your data (e.g. fda, bls, etc.).

Create a Jupyter notebook `<source>_notebook.ipynb` with the following logic:

- Create a new dataset in BQ for storing the staging tables for `dataset2`. The dataset should be named `<source>_staging`.
- Import the dataset files from GCS into your new dataset in BQ. Ensure that you import each file into its own table.
- Verify that each table was loaded correctly by doing a `select count(*)` from each one.
- Create a new dataset in BQ for storing the modeled tables. The modeled dataset should be named `<source>_modeled` .
- Create modeled tables by applying the design principles from Milestone 4.
- Each modeled table should have a primary key. Check for any primary key violations and deduplicate the records in SQL if possible. Otherwise, make a note of the table which doesn't contain a valid primary key. You will need to deduplicate this table with Beam in the next milestone.
- Check for any referential integrity violations between any parent and child tables.

Part 3:

1. Update your ERD to include the modeled tables in `dataset2`. Be sure to denote in the diagram the relationships between the tables within `dataset2` as well as **across** the two datasets. Name your updated ERD file `ERD-dataset2-modeled.pdf`.

2. Think of 3 interesting queries that span your primary and secondary datasets. These queries should use a join to combine the data from `dataset1` and `dataset2`. In addition, these queries should require some prior data transformation process that

cleanses, enriches or deduplicates the data (e.g. name or address standardization). The required transformations will be done through Apache Beam in the next milestone.

For each of your 3 queries:

- Briefly describe the expected results from the query and what SQL operations the query will use to produce those results (1-2 sentences).
- Briefly describe what type(s) of data transforms are required to successfully implement the query (1-2 sentences).

Create a file `CROSS-DATASETS.txt` and add your descriptions and explanations to this file.

CS 327E Milestone 9 Rubric
Due Date: 11/15/19

Part 1 - Edit the file `./DATASETS.txt` to include information on your new dataset. **-10** no description of new dataset in `DATASETS.txt`	10
Part 2 - Create a Jupyter notebook `<source>_notebook.ipynb` containing the ingestion and modeling pipeline, as described in the outline. **-60** `<source>_notebook.ipynb` is missing. **-30** datasets `<source>_staging` or `<source>_modeled` not present **-10** each missing staging or modeled table **-10** inconsistent naming conventions across tables **-10** each non-merged entity type, table with multiple entity types, or un-unioned tables containing the same data (i.e tables representing the same data across different years). **-10** each string field in modeled tables containing only `INTEGER`, `NUMERIC`, `DATE`, or `TIMESTAMP` not cast, up to **-40**	60
Part 3 - Create a new ERD titled `ERD-dataset2-modeled.pdf` which also includes data in your new dataset. Diagram their relationships as you have in previous milestones - this does include adding potential relationships between tables from both datasets. **-15** `ERD-dataset2-modeled.pdf` is missing. **-5** each incorrectly labeled keys **-5** each incorrect relationship **-5** each incorrectly labeled data type Create a file `./CROSS-DATASETS.txt` containing query and transformation information for 3 queries, as described in the outline. Keep in mind that you do not actually have to *write* the query, just a description of one and transformations required to make the query work. **-15** `./CROSS-DATASETS.txt` does not exist **-15** for each missing pair of query description and required transformation(s) description, up to **-15**	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**