CS 327E Milestone 2 due Thursday, 11/19.

The goals of this milestone are to derive a data model from your raw data and continue exploring your data. Perform the following data modeling tasks to accomplish these goals.

1. Create a new Jupyter notebook named `milestone2.ipynb`.

2. Create a new BQ dataset to store your modeled tables. The dataset should be named `<source>_refined` where source is the source of your data (e.g. imdb, airbnb, noaa, etc.).

3. Create your modeled tables by applying the design principles we learned for relational databases (refer to slide 9 from 09/04 lecture for details):
    - split staging tables that contain more than one entity into separate tables.
    - join staging tables that store different attributes belonging to the same entity.
    - union staging tables that store distinct records belonging to the same entity.
    - identify a candidate primary key (PK) for each modeled table.
    - check for records with the same primary key and remove unwanted ones.
    - identify parent-child relationships between tables.
    - check for referential integrity violations between parent-child tables.
    - remove any child records which violate referential integrity.

4. For each field in your modeled tables, choose a primitive data type that most precisely represents its domain of values:
    - if the field is of type `STRING` and it stores `INTEGER`, `NUMERIC`, `DATE` or `TIMESTAMP` values, cast its type to the most fitting type.
    - if the field is of type `INTEGER` and it stores a `DATE` or `TIMESTAMP` value, cast its type to the most fitting type.
    - if the field is of type `TIMESTAMP` and the values it stores are of type `DATE` (i.e. the time component is not being used), cast its type to `DATE`.
    - Convert from one data type to another using the CAST function. If the CAST function returns an error due to non-conforming characters, use one or more STRING functions to parse and reformat the data. Refer to covid_19_modeled.ipynb to see a working example.

5. Create an ERD in Lucidchart of your data model:
    - The diagram should capture your modeled tables across both `refined` datasets
    - The diagram should include the collection of fields (names and types) for each entity.
    - The diagram should specify a primary key for each entity.
    - The diagram should specify a foreign key on each child entity.
    - Download your diagram and name it `final_project_data_model.pdf`.

6. Continue to explore your data by writing SQL queries on your modeled tables:
    - Come up with 5 new queries, at least 3 of which should contain a subquery and at least 2 of which should contain an aggregation.
    - Each query should also include at least 2 clauses from this list: JOIN, WHERE, GROUP BY, HAVING, ORDER BY, LIMIT.
    - Precede each query with a Markdown comment that describes its function.

7. Create data visualizations:
    - Create a new BQ dataset for storing your reporting views. Name the dataset `insights`.
    - Choose 2 of your most interesting queries from the previous section.
    - Create a view for each one with a descriptive name and prefixed with `v_` (e.g. `v_Highest_Nominated_Movies`).
    - Open Data Studio
    - Create a Data Source that accesses each view. You'll need one Data Source per view.
    - Create a chart in Data Studio that visualizes the data in a compelling way.
    - Add both charts to a single Data Studio report (aka dashboard).
    - Download your dashboard as a pdf and name it `dashboard-v1.pdf`.

CS 327E Milestone 2 Rubric
Due Date: 11/19/20

Create a data model consisting of modeled tables from your raw data. Identify entities in your raw tables, split additional entities into their own tables, join tables belonging to the same entity, and union all tables that share the same fields. All modeled tables should have a valid primary key. All child modeled tables should have a valid foreign key. String fields, if able to be casted to a more fitting type, should be (e.g. Ints, Dates, etc.) via BQ functions. **-40** `milestone2.ipynb` not found in repository **-20** `<source>_refined` for each dataset not found in BQ project **-10** for each non-merged entity, table with multiple entities, or un-unioned tables containing the same data (i.e tables representing the same data across different years). **-10** for each table in refined dataset without a primary key (identified in ERD and supported by code) **-7** for each marked primary keys which contains duplicates **-7** for each child table in refined dataset without a foreign key (identified in ERD and supported by code) **-5** for each child table with foreign key violations **-5** each string field containing only `INTEGER, NUMERIC, DATE,` or `TIMESTAMP` not cast, up to **-20**	40
An ERD which contains detailed information on your modeled tables. **-30** `./final_project_data_model.pdf` not found in repository **-10** missing entities **-5** for each missing key (primary and foreign) or incorrect keys marked **-5** for each entity with missing fields (names and types) **-5** for each missing relationship between entities	30
Write 5 SQL queries that explore the data. 3/5 queries with a subquery and 2/5 queries with an aggregate function. Each query should be preceded by a Markdown comment that explains its function. **-20** queries missing from `milestone2.ipynb`: **-5** for each query missing a subquery or incorrect subquery **-5** for each query missing an aggregate function or incorrect aggregate **-3** for each query not using at least 2 clauses from: JOIN, WHERE, GROUP BY, HAVING, ORDER BY, LIMIT **-2** each incorrect comment, or comment too similar to query	20
Create data visualizations in Data Studio. Visualizations should represent the results from two BQ views.	10

The visualization file should contain 2 charts made from Data Studio, with a relevant title for each one describing the data. **-10** `./dashboard-v1.pdf` not found in repository **-5** each missing chart -**5** each chart created from a BQ table instead of a BQ view **-2** each missing title	
`milestone2.ipynb, final_project_data_model.pdf`, and `dashboard-v1.pdf` 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**