Milestone 6 due Friday, 10/19.

Perform the following modeling tasks to improve the data quality of *dataset1*.

1. For each table in *dataset1*:
   - split the table if it contains multiple *entity types*. There should be one entity type per table.
   - identify a *primary key* (PK) for the table.
   - check for the presence of duplicate rows based on the PK.
   - remove any duplicate rows found in the table.

2. For each *child* table in *dataset1*:
   - if the table's *foreign key* (FK) points to a PK, check for the presence of orphan rows in the child table.
   - remove any orphan rows found.

3. For each table in *dataset1*:
   - if the table has a column of type *string* and it stores *integer*, numeric, *date* or timestamp values, convert the column to the appropriate type using BQ's **CAST** function.

4. Create an ERD v2 that denotes:
   - field names, data types, and keys (PK, FK) for each entity.
   - relationships between entities.

5. Verify queries, views, and charts:
   - re-run queries and views developed for Milestones 3 - 5.
   - fix any broken queries or views and update the appropriate .sql file with code fix.
   - open Data Studio report and fix any broken charts on your dashboard.

CS 327E Milestone 6 Rubric
Due Date: 10/19/18

For `dataset1`, all tables should have an identified primary key. Values in the primary key should have no duplicates. String fields, if able to be casted to a more fitting type, should be. In addition, identify all entity types in your tables and split additional entity types into their own tables. **-40** no primary keys identified from ERD **-20** marked primary keys contain duplicates **-10** each string field containing only `INTEGER, NUMERIC, DATE, or TIMESTAMP` not cast, up to **-40** **-10** each additional entity type in each table, up to **-40**	40
For `dataset1`, all child tables should have an identified foreign key. **-30** no foreign keys identified on child tables in ERD **-20** relation is incorrect **-15** orphaned rows contained in child table	30
An ERD should be pushed that contains all detailed information for the fields in `dataset1`. Note that credit from other parts of the assignment may rely on this part. **-30** `./ERD-dataset1-v2.pdf` not found in repository **-10** missing field types **-10** missing field names **-10** missing field keys **-5** incorrect keys marked	30
Fix all broken SQL statements from previous milestones 3-5. Make sure each statement runs properly. Save them into the same files, replacing the broken statements with their fixed counterparts. Fix broken DataStudio charts as well. Push the new dashboard screenshots onto GitHub, replacing the broken ones with the fixed counterparts. **-5** each erroneous SQL query or malformed data chart, up to **-20**	

`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**