CS 329E Project 0, due Thursday 01/25.

Part 1: Setup

1. Find another student in the class who wants to partner with you on the projects. If you need help finding a partner, please create a post on Ed and/or speak to the instructors.

2. Once you are partnered, you should choose a name for you and your partner's code repository. You will be given a repo in our GitHub organization (link). Your repo will be private to you, your partner, and the instructors.

3. Request a GitHub repo by following these steps:
    - Email your name, EID, and GitHub username for you and your partner
    - The requested name of your repo
    - Address the email to the Prof. and the two TAs
    - Copy your partner on the email
    - Use the email subject line: [CS 329E] Spring24 Repo Request

    Note: You cannot create your own GitHub repo. You must request a private repo under our GitHub organization by following the steps mentioned above.

4. If you are new to git or GitHub, go through some basic tutorials. You should install git on your laptop and learn the basic commands (e.g. git add, git commit, git push, etc.). We won't be spending any class time on git, this is something you'll need to learn on your own time.

5. Once we have created your repo, you and your partner should each receive an email invitation to join it. Accept the invitation and set up your git repository by following our guide. Remember to create a README file in your repo with you and your partner's full names, EIDs, and emails.

6. Create a GCP project for you and your partner by following our guide. Be sure to go through all the sections of the guide. Only one person per group needs to do this step. **You should end up with one GCP project per group.**

Part 2: Dataset

7. Download the BIRD training set to your laptop and uncompress its contents (you can read up on BIRD here). The training set uses about 40GB of space. If you don't have enough free space on your laptop, spin up a VM on GCP and download the file to the

VM. You can create a JupyterLab environment from the Vertex AI Workbench page just like we did in CS 327E. Use the `wget` command to download the file from JupyterLab.

8. The BIRD data is in SQLite files (*.sqlite). There is one SQLite database per dataset or domain. Download the SQLite binaries and start exploring the data.

   To open a SQLite file, navigate to the directory where the file is located and run:
   ```
   sqlite3 FILE_NAME
   ```

   For example, to open the file named `address.sqlite`:
   ```
   cd train/train_databases/address
   sqlite3 address.sqlite
   ```

   This puts you in the sqlite prompt and then you can type `.help` to see a list of the available commands:
   ```
   sqlite> .help
   ```

9. Explore the datasets available and choose one that you want to work with. If you choose more than one, they must be logically related. You can choose any of the datasets available except for the one called airline, which I'll be using for demos.

   Your dataset should have at least three related tables. For example, airlines, airports, and air carriers.

10. Once you have chosen your dataset, export each table to a CSV file. Here's an example of how to open airlines.sqlite, inspect the database schema, and export its Airlines table, to CSV:

    ```
    sqlite3 airline.sqlite
    sqlite> .schema
    sqlite> .headers on
    sqlite> .mode csv
    sqlite> .output Airlines.csv
    sqlite> select * from Airlines;
    sqlite> .quit
    ```

11. Create a GCS bucket to store your CSV files. Place the bucket in the us-central1 region and name it something descriptive (.e.g. airline_data). Note that you may need to add a suffix to the name because bucket names must be globally unique.

12. In your GCP bucket, create a folder named **raw** and upload the CSV files that you exported from SQLite into this folder.

13. Prepare an ERD in Lucidchart that represents your raw entities that are captured in the CSV files. Each file represents a different entity. It should look similar to this example.

    Please note that yours will only have the BIRD entities for now. However, in the next project (Project 1), you will add related entities by sourcing data from Faker and public APIs. For now, start thinking about additional entities that you want to add to your model and if you have extra time this week, start looking around for data sources. Come to office hours if you have questions and want to discuss further.

14. Prepare a data dictionary for each raw entity that's in your dataset. Your data dictionary should look similar to this example. Please note that yours will only have the BIRD dictionaries for now.

    Much of the information that's in the data dictionary is available through the BIRD training set. If you look in the **database_description** folder for the entity in question, it should contain a CSV file with the column names, descriptions, and formats. You'll need to create an additional column to capture the sample values (as highlighted in the example). This is important because we want to understand what type of information a table has just by looking at the data dictionary.

15. Add your ERD (erd-raw-v1.pdf) and data dictionary (data-dict-raw-v1.xlsx) to your git repo and publish to GitHub.

16. Create a submission.json file and upload it through Canvas by the deadline. Only one person per group needs to do this step.

CS 329E Project 0 Rubric
Due Date: 01/25/24

Create a `README.md` file in your team's private repository under our cs329e-spring2024 GitHub organization:
 	
• `README.md` should contain you and your partner's full names, EIDs, and emails in the following format (not including braces):
 	
`<your full name>, <your UT EID>, <your email>`
`<partner's full name>, <partner's EID>, <partner's email>`
 	
Example:
 	
`William Chia, wc1234, chiaw@example.com`
`Prithvi Chowhan, pc1234, chowp@example.com`
 	
-25 no private repository under the cs329e-spring2024 GitHub organization
-25 no README.md file, file incorrectly named, or incorrect info in file
	25

GCP Project is shared to cs329e.spring2024@gmail.com **-10** incorrect project name **-25** project not shared	25
Data dictionary is thorough and not missing any columns **-5** for each important entity missing **-10** missing sample values **-10** missing column_description (missing value description is acceptable) **-25** missing file in repository	25
ERD is accurate and demonstrates relationships between tables **-10** missing important links between keys **-25** missing file in repository	25

`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**