CS145: Data Management and Data Systems

Stanford University, Fall 2018

Project 2: Visualizing Data
10% of Course Grade

Due Date: Monday, November 5th, 11:59PM

Overview

In this project you will explore two datasets using Colaboratory. Colaboratory is like a Jupyter notebook[1], but it has collaboration and integrations with BigQuery built into it. You will follow along with Colaboratory notebooks we provide and use SQL/Python and standard visualization libraries to explore and visualize World Bank data and GitHub data.

Note that this part is intended to prepare you for the last part of the course project where you will be running through an entire data cycle of querying, visualizing, and predicting on a dataset of your choosing. This is an individual project, but getting used to using Colab will aid you in Project 3 when you can work in groups.

Note: This project is at least twice as long as the previous one! You will have almost three weeks to complete the assignment, but please start early.

Task A: Get Setup With Colaboratory

Here is an overview of Colaboratory features and a brief guide for using BigQuery through Colaboratory. Before proceeding, make sure you have read and understood these support documents. To open a new notebook in Colab, you can go to *File* > *Upload notebook* and choose the file either from your computer or from Google Drive. You can also make a copy of an existing Colab noteboook by going to *File* > *Save a Copy in Drive ...* . Colab notebooks can be saved just like any other file to your own Google Drive account.

Note: You have to be careful with your BigQuery credits when running cells on your Colaboratory notebook, as Colab will not tell you how much data your query will use. We advise you to check your queries in the BigQuery interface first to see how much data it will consume; keep in mind that a query using about 5GB will cost ~ 2.5 cents. You can also set a byte limit to your queries if you are using the Classic UI; we suggest having a limit of 35GB.

[1] Jupyter notebooks are a standard tool for data scientists. They allow you to create and share documents that contain "cells" with runnable Python code as well as equations, visualizations, and text.

Task B: Exploring World Bank Data

For the first half of this project, you will be exploring the World Bank dataset in the provided notebook `project2_world_bank.ipynb`. You can access the notebook from the course website, **make a copy of it in your own drive**, and begin the assignment.

This notebook consists of two sections, each focusing on a different part of the exploration process. In the first part, you will investigate and design your own schema for the dataset. In the second part, you'll learn to create visualizations to help you understand and answer questions about the data.

Task C: Exploring GitHub Data

For the second half of this project, you will be exploring a subset of the GitHub public dataset in the provided notebook `project2_github.ipynb`. Now that you have some familiarity with exploring datasets in Colab and creating visualizations, you will be trying to answer a larger question: what features impact the popularity of a GitHub repository?

This notebook will also be split into two sections. In the first section you will get familiar with the GitHub dataset we provide by translating its existing schemas into an E/R diagram; you will then use that E/R diagram to design an alternate schema. The second half will be focused on creating visualizations and analyzing them to understand how certain features of a GitHub repository correlate with its popularity.

At the bottom of this notebook there are two extra credit questions. These are entirely optional and do not have to be completed, but we recommend you take a look if you have time, as they are relevant to what you will be doing in the final project.

Just like with the World Bank Colab notebook, you can access the GitHub notebook from the course website but you should make a copy of it in your own Drive before working on it.

Honor Code

This assignment is to be done individually. We encourage students to form study groups to complete the assignment, but the solutions to each assignment must be written independently. Be sure to list your collaborators on each part of the assignment at the top of the corresponding Colaboratory notebook.

We take the Honor Code seriously. Working in groups to work out a problem is OK, but the following would be considered honor code violations:

- Looking at the writeup or code of another student.
- Showing your writeup or code to another student.
- Discussing a problem in such detail that your solution is almost identical to another student's solution.
- Uploading your writeup or code to a public repository (where other students may be able to find it).

Submission Instructions

Once you have filled out both Colab notebooks completely, you are ready to submit. If you collaborated with others to discuss findings or generate queries, make sure to add their names and SUNet IDs to the cell at the top of the Colab notebooks.

To submit:

1. Download the World Bank Colab notebook as an iPython notebook - you can do this by going to *File > Download .ipynb*.
2. Print out a PDF of your Colab notebook, **making sure that you have run all cells first.** In Google Chrome, you can do this by going to *File > Print* and then choosing *"Save to PDF"*. Make sure you've closed the table of contents sidebar before you print so we can easily see your work and output.
3. Do steps 1-2 for the GitHub notebook.
4. Merge the two PDF files (World Bank and GitHub) into a single PDF and submit to the **Project 2 - PDF** assignment on Gradescope.
5. Submit the two iPython notebooks (World Bank and GitHub) to the **Project 2 - iPython** assignment on Gradescope.

Note: We reserve the right to deduct points from your project if you do not follow the submission instructions, if there are some cells which have not been run or whose output is not readable, or if you have assigned your pages to the questions on Gradescope incorrectly. **Please read through your PDF document before you submit it and *ensure that all answers are clearly visible.*** Please also leave yourself enough time to do the assignment/submission, and go over your assignment in Gradescope to make sure it is correct!

You may resubmit as many times as you like; however, only the latest submission and timestamp will be saved, and we will use your latest submission for grading your work and determining any late penalties that may apply. Submissions via email will not be accepted.