

CS145: Data Management and Data Systems

Stanford University, Fall 2018

Project 3: Querying, Visualizing, Predicting -- The Full Data Cycle

20% of Course Grade

Proposal Due Date: Friday, November 9th, 11:59PM

Project Due Date: Friday, November 30th, 11:59PM

Overview

Welcome to the final CS145 project! You've gotten a good amount of experience with SQL by now. Congrats! In the first project you learned how to navigate a reasonably complicated dataset and extract facts from it. In the second project, you learned how to think about the design and tradeoffs of different schema designs for real datasets -- you also got experience visualizing and reasoning about the information in these datasets (and you probably realized that real-world data is not as nice and structured as we'd like!).

In this assignment, you will use the tools you've learned throughout the quarter to follow your own explorations on a topic of your choice. You will pick your own dataset and come up with one or more interesting questions that you want to explore, just like how you began to explore what makes a Git repo popular using the GitHub dataset for Project 2. You will finish by using machine learning to make predictions related to your question.

This project may be done alone or in pairs.

For this project, we will also require submission of a very short proposal (see **Task A** for more details), detailing your choice of dataset and question as well as your group. The intent of the proposal is for CAs to give feedback on whether or not your dataset and question is appropriate for this project; it is worth 5% of the project grade and must be submitted by **Friday, November 9th** at 11:59 PM. **No late days can be used for this proposal.**

Note: Though this assignment can be done in pairs, late days are applied individually. If a pair submits late, any member without late days will receive a zero.

Task A: Project Proposal (5%)

This proposal is not meant to be a long assignment; you should not need to write much more than two or three small paragraphs. To submit your proposal, go to the *Project 3 - Proposal* assignment on [Gradescope](#) and fill out the questions directly on Gradescope. The proposal will ask about what dataset you plan to use and what question(s) you'd like to tackle regarding that dataset.

The proposal is worth 5% of your project grade and is due on Friday, November 9th. Don't worry too much about it; as long as you turn it in and fill out the form with reasonable detail, you will receive full credit. After you turn it in, the CA's will quickly go through the proposals to ensure that the dataset and question you choose to explore are rich and complex enough for this project; by Tuesday, November 13th, you will have feedback on your proposal. We expect that in most cases, it will be a simple go-ahead; for some, you may be asked to choose a dataset with greater size or complexity or to choose a question with more depth. In other cases, we may advise you to decrease the scope of your project.

You are allowed to change your choice of dataset and question after you submit your proposal; however, you will have to take initiative and come to OH if you want to verify that your new choices are appropriate for this project.

Task B: Introduction to Machine Learning (10%)

For this project, you will be using BigQuery's machine learning features to make predictions related to your question of choice. Do not worry if you've never studied machine learning before! We have created a Colab notebook, `project3_MLwarmup.ipynb`, that will guide you through the fundamentals of what you need to know to complete this project. You can access the notebook from the course website, **make a copy of it in your own drive**, and begin the assignment.

Read the notebook and fill out the questions where specified.

This notebook will be worth 10% of your project grade.

Task C: Your Own Data Cycle (85%)

For the main portion of Project 3, you will be exploring a question that interests you on a BigQuery dataset of your choice. Please create a Colab notebook containing all of your work. You will use what you've learned from Project 1 and Project 2 to begin your explorations on the dataset and your question. Once you've done your explorations, you will use machine learning to make predictions relevant to your interesting question.

As an example, take the GitHub dataset from Project 2. The question we were trying to answer there was what factors impact the popularity of a GitHub repository. Given this question, once we finish our explorations and decide what features were important, we can use machine learning to predict the popularity of various repositories based on those features.

This part of the project is largely open-ended; the specific explorations you choose to do are up to you to decide. However, we expect your project will contain at least the following sections:

- **Analysis of your dataset (30%)**
 - Comment on how the dataset you chose is organized. Some questions you may think about are: Is there redundant data? What are the relationships between tables? If there is only a single table, what are the keys of the table? Are there keys (in the FD sense, or maybe in the OKV sense) at all? Are the tables normalized in some way (e.g. BCNF) or are they denormalized? What are the tradeoffs between the design of the dataset you chose as it is and other alternatives? Why do you think the original authors of the dataset chose this particular structure?
 - Your analysis should demonstrate that you understand the structure of the data you are working with.
 - You may supplement your analysis with concrete examples to corroborate your statements. For example, your analysis could benefit from having an E/R diagram to reference in some cases.
- **Exploring your questions, with appropriate visualizations (40%)**
 - Use SQL in your Colab notebook to gather information about your dataset -- you should use your plotting library of choice to visualize and understand your data. Ask and answer quantitative questions that revolve around your central questions. Your exploration should both shed light on the questions that you've asked yourself and provide insights for the feature engineering you will use when you generate predictions.
 - In this section, conclude with an analysis and summary of your observations. What features of your data seem especially prominent or related to your question? Are there anomalies? What trends do you see? Have you answered your questions?
 - Your analysis in this section should demonstrate that you have a reasonable quantitative understanding of your dataset. By the end of this section, you should be ready to use your newly-gained domain knowledge to generate predictions.
- **Predictions based on your explorations (20%)**
 - Use BigQuery to make predictions related to the questions you set out to ask. In the GitHub example in Project 2, the prediction task would be, given that I have a good set of features about a GitHub repo, can I predict how popular that repo is?
 - Evaluate your model in BigQuery. Comment on the performance of your model. Is it good? Is it bad? In either case, why do you think your model does well or badly?
 - Finally, use your model to make predictions **on data besides that which you trained (i.e. created your model) on.**
- **Conclusion (10%)**
 - What have you learned? What conclusions have you made or been unable to make about your dataset and why? What is obvious, and what did you not expect to see? Support your statements with the charts or predictions you generated. If you had more time, what other data exploration would you pursue?

Note that we are generally more concerned with the depth with which you've leveraged visualizations and BigQuery to gain insight into your questions than the quality of your machine learning predictions. This is an open-ended project -- there is no right or wrong answer, only quality of exploration and inquiry.

This part will be worth 85% of the project grade.

Honor Code

As in all Stanford classes, you are expected to follow the Stanford Honor Code. For example, the following activities are prohibited and will be treated as Honor Code violations (this is not intended to be a complete list of Honor Code violations):

- Submitting code that you did not write personally, with the exception of project code written by your partner.
- Consulting pre-existing solutions for problem sets and projects (such as solutions posted on the Internet).
- Posting your solutions on the Internet or making them available to other students in any form.

You are allowed to discuss general approaches and issues with other students in the class besides your team members. It's also fine to give other students help finding bugs if they are stuck, or to answer general questions. But, any code you write must be written by you and your partner, from scratch, without consulting existing solutions. We reserve the right to use computer software such as MOSS to analyze material that you submit in order to detect duplication with other students or existing solutions.

A general way to think about this is that if a particular activity significantly short-circuits the learning process (it saves you time but reduces the amount you learn and/or figure out on your own), or if it misrepresents your capabilities or accomplishments, then it is probably an Honor Code violation.

Submission Instructions

Proposal Submission:

Please directly answer the proposal questions in the *Project 3 - Proposal* assignment on Gradescope.

Project Submission:

1. Download the Machine Learning Warmup Colab notebook as an iPython notebook - you can do this by going to *File > Download .ipynb*. Submit it to the *Project 3 - ML iPython* assignment on Gradescope.
2. Print out a PDF of the Machine Learning Warmup 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"*. You may need to adjust the paper size and scale to ensure all your graphs are readable. Make sure you've closed the table of contents sidebar before you print so we can easily see your work and output. Submit the PDF to the *Project 3 - ML PDF* assignment on Gradescope.
3. Download your personal exploration Colab notebook as an iPython notebook. Submit the iPython notebook to the *Project 3 - iPython* assignment on Gradescope.
4. Print out a PDF of your personal exploration Colab notebook, **making sure that you have run all cells first**. Submit the PDF file of your project to the *Project 3 - PDF* assignment on Gradescope.
5. For each of your submissions on Gradescope, **make sure to add your project partner as a group**. You can do this after submitting by clicking *View or Edit Group* underneath your name and searching for your partner.

In total, for this project you should be submitting to four different assignments on Gradescope; two for iPython files, and two for PDF files.

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.

Notes on Colab-to-PDF Problems:

We are aware that printing to a PDF in a Colab notebook will often result in cut off charts and tables. If you can't get a good printout of your Colab notebook by tuning page size and scale, add a public view only link to your notebook at the top of your PDF submission. We'll cross-check anything not fully visible in your PDF with the actual Colab notebook you share with us.