Fall 2023 - CSE-6040 | Extra Credit Project

Assignment Due: Thursday, Dec 7th @ 1159 UTC
Peer-grading Due: Wednesday, Dec 13th @ 1159 UTC

This document outlines the requirements for the end of semester **peer-graded extra credit assignment**. There are two parts: (1) you create a Jupyter notebook on a topic of your choosing (but see below), and (2) you grade three notebooks submitted by your peers. Your notebook (Part 1) is due on Dec 7th, 1159 UTC, while the peer-grading (Part 2) must be completed by Dec 13th, 1159 UTC. These deadlines are firm and there will be absolutely no extensions. If you miss **either deadline** you will receive no extra credit points. The assignment will be submitted via Vocareum, but unlike homework assignments, the design is up to you.

Assignment: Pick a topic that interests you. Find some data related to that topic and come up with some questions you might answer using that data. Create a Jupyter notebook that processes the data and implements an analysis to address your questions. We expect your notebook will need to clean the data, analyze it using tools you've learned in this class, and produce some outputs that will help you show what you've learned from the data. Remember that your outputs may disprove your original theory or be inconclusive, a situation that is very common in analytics; this result is okay as long as you demonstrate a solid effort to answer your questions. See your task

Ambiguity: Aside from the requirements below, this assignment is designed to be open-ended so that you can showcase what you've learned in this class. This activity is very similar to what you will do in the real-world and is essentially how we create notebooks and tests in this class.

Points: This assignment will give you up to 3 **percentage points** towards your final grade.[1] The base requirements are strict and must be met with **no exceptions**. Grading is not designed to make it hard to get full credit—it's designed to make sure you're putting in enough effort to get something out of this assignment (by doing it yourself, but also by learning and sharing with your peers during grading). See grading

Plagiarism: Your work must be original and must be your own. You can copy any dataset you like, but everything else must be something you come up with on your own. You are encouraged to look online for ideas and inspiration, but you must build the notebook on your own (see FAQ for more information). *Simply put, copying someone else's project or stealing large amounts of code will result in 0 points being awarded, and a report to OSI for academic misconduct.*

Career advice: This is a very good starter file to include on your Github and resume. This assignment will allow you to demonstrate your coding, data analysis, and presentation ability. This is also similar to many of the Practicum projects that OMSA students complete as part of the degree. If you are unsure how to do this, make a public piazza post.

---

[1] That is, your course grade will be calculated without the extra credit, then at the end you will receive 0-3 points based on peer evaluations of your extra credit. Ex: if your course grade is 78% and you got the max score on your peer evaluations, your course grade will be 78% + (100% * 3%) = 81%.

Minimum Requirements:

The following files must be included and **submitted in Vocareum by the due date to receive <u>any</u> credit**:

- A dataset (.csv file or format readable by your notebook) of no more than 10 MB in size. Multiple, smaller datasets are acceptable, but the combined size must still be under 10 MB.
- A single Jupyter notebook (.ipynb file)

Expanded Requirements:

Your Jupyter notebook should do the following:

- Read in the dataset you included in Vocareum
- Clean data (confirm no missing data, imputes/deletes as needed)
    - This may not be necessary if your dataset is already "clean", but you should show some quick analysis or stats to show how you confirmed this
- Analyze data using a combination of tools (Python packages, analytics methodologies, etc.)
- Generates useful outputs including visualizations
- Your notebook should only depend on packages available in Vocareum
    - This means that **import** is fine, but **install** will not work. It may work on your end, but it won't work when your peers are grading, therefore you will end up getting a 0 as that will generate an error.
    - If you miss this requirement or install a package anyway, the only way to return your environment back to normal is to **reset assignment**. This will delete your work so save it locally before reset.
- **Your code must run without an internet connection:**
    - The packages you use **must not rely on connecting to the internet to be used**. The main packages you come across that do this are NLP packages and web scraping packages. If you want to use those then you will need to figure out how to download the data to your environment so your graders can use it. For webpages this would be simple, for NLP packages it's a bit harder (we wouldn't recommend spending time on this).
    - **The process is then**: use package to download data, save data in environment (as CSV, or for NLP it may be other filetypes), change your code to rely on these Vocareum-stored files, then remove any code that connects to the internet. **This is not hard to do, but risky if you're unsure what's going on. If you have issues with this then we recommend doing something different altogether. We won't be able to help resolve this.**

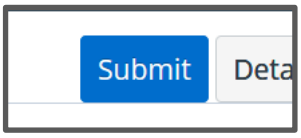
While doing that, it needs to also (*these are all requirements*):

- **Run without generating any errors (you will receive 0 points if it does not run error-free)**
    - This includes any error, no matter the severity. If the peer graders do "run all" and it stops on any cell, that counts as an error. Even if it's a cell you forgot to delete or is not crucial to the analysis.

    - Warnings (**usually a red box with text in Jupyter**) **are OK**, but you should make sure you understand them. There are some situations in which the red warning will cause you to realize that you made an error in your logic.
- Be properly documented (not line by line, but logically by code chunks)
- Be organized logically so peers and others can look at it from top to bottom and understand what you're doing
- Include small blocks of text throughout that will serve as your **write-up**. These blocks should cover:
    - The problem you decided to investigate
    - How you are going to analyze data to answer that question
    - The results of your analysis, including how your visualizations are relevant and useful
    - Suggestions for future research (including improvements to what you did)

Submitting your work:

- Locate the assignment "Extra Credit" in Canvas. It's located in Modules near the bottom
- Enter the Vocareum assignment and upload everything needed to run your project in there (this would include: data, notebook, and any other files). There is already a template notebook in there which you may use or delete and replace with your own.
- Run everything to double-check that it runs. **The 'submit' feature will not notify you if it doesn't run, so you will need to check yourself.** This means you should 'restart & run all', and then confirm everything runs, no errors are generated, the kernel doesn't crash, and visualizations appear correctly.
- **Hit 'submit' at the top right**

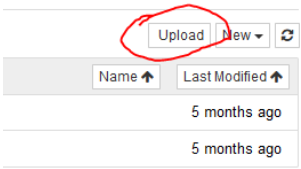


- After the submission deadline, you will be able to go into the assignment and do your peer-grading, which is due by the deadline listed on the top of this document. If this doesn't work initially, please wait until ~12 hours after the submission deadline closes to see if it works. If it's still not visible, make a private piazza post and we'll investigate
- Reminder: Your work must be submitted through Vocareum before the deadline. We will not accept late submissions or submissions via email, piazza, or any other method even if they're on time.

Uploading your dataset(s):

The dataset(s) needs to be uploaded into the same directory as your notebook. Note the requirement is 10MB, regardless of if you upload 1 file or 5 files. I encourage you to keep it simple and keep the number of datasets small so your peers can easily understand the structure.

To upload, navigate to the folder view in Vocareum (click the Jupyter logo at the top left), then click upload in the upper right corner.



Note on accessing the file once you upload it: You should instead be referring to the file directly, so something like `pd.read_csv('sample_data.csv')` should natively work.

Do not use `os.getcwd()` **to find the file location and access it**. Vocareum stores all your data in unique directories, so when you use that the path will work for you, but not your peer graders. Simply refer to the file using the one line of code above.

Grading:

This assignment will be **peer-graded** by three of your peers. Your score will be the average of those three. Requests for manual regrades by the course teaching staff will be granted only under extraordinary circumstances. Getting 1 point from one peer and 3 points from another peer does not warrant a regrade on its own; this is the reason for taking the average.

Please see the three tables below that break down the grading criteria. While the rubric is a little complex, its subsections are somewhat subjective, so the grading is **up to you**, except for the pass/fail criteria. Please give your fellow classmates a fair grade for the work they submitted, but also consider putting some time in to give them real feedback on their work. Give the level of feedback you'd like to receive.

Note: You must complete your three peer grade reviews in order to receive your extra credit points. If you fail to do them before the deadline, you will receive no credit. This means you must enter a score AND provide feedback. If you leave little to no commentary we may withhold your points. This should not be hard, it should only take you a few extra minutes to leave some comments and provide some tips or even ask questions. It's unfair to your classmates if you simply drop a score and leave a five word reply.

We realize people are doing this for the grade boost, but we'd like you to get something extra out of it as well. Good feedback will allow your fellow classmates to learn from the work they did and apply it in the future. The more ambitious students may find this a good opportunity to revise their project and post it on their Github, which is highly encouraged!

Pass/Fail Criteria

Criteria	Pass/Fail
Data file(s) and Jupyter notebook file are present	• **0** for the entire project if all the criteria are not met. • If all criteria pass, use the rubric below to determine final score
Jupyter notebook runs without error **Warnings are allowed	

Grading Rubric

Criteria	0 points	1 point	2 points	3 points
Data selection	Data is attached but is not relevant to the question being asked		Data is relevant and is substantial enough to analyze	
Analysis	Analysis is limited and does little to gain insights from the data		Analysis is thorough and explained. The analysis answers the question(s) asked	
Code	Code runs but is lacking documentation or is highly unoptimized		Code is functional and well-documented. Optimization is a +, but not necessary	
Visualizations: Tables, Graphs, etc.	No outputs produced	Some tables, graphs, or other visual outputs created	Visualizations are easy to understand and contribute to the analysis. All charts and tables have titles and other components needed to understand them.	

Final Score Calculation

If any pass/fail criteria fails:	0
If all pass/fail criteria passes:	Sum of subsection scores from rubric, divided by 4, **rounded up** to the nearest whole number. **Possible scores are: [0,1,2,3]**

FAQ's:

Can you provide some example projects?
The past and current midterms closely resemble the style we're looking for, but that level of complexity is not necessary (but is doable!). In particular, look at how each one starts out with some data, goes through some cleaning and analysis, and then outputs something, whether it's a specific output or just well-documented insights.

How long should I spend on this?
We can't answer that for you, but it's reasonable to expect this to take 10-15 hours, depending on how comfortable you are with the material we've covered in the class so far. We're not asking you to do anything new, just use things you've learned in the class.

Can the deadline be extended for this?
Unfortunately, no. We need to leave enough time for the teaching staff to calculate and input grades before the Georgia Tech grade submissions deadline.

I had a great idea and a good plan, but once I did the analysis the results weren't clear.
That's ok! As long as you explain all of that in your notebook, you will still be eligible for credit. A well-documented project and breakdown of your findings (or lack thereof) is still useful. Academic research papers and real business projects often end with this result, it's just part of the process.

The dataset I want to use can't be publicly shared.
Please choose a different project or dataset to use. Your peers need to be able to run your code against the data to see the results. You can investigate using a smaller piece of the dataset, or redacting some of it, but we recommend you don't spend much time on that step. You'd be better served by finding a new dataset and focusing your time on analyzing it and building your project around that.

Can I create the notebook locally and then upload it?
Yes, but your code must be executable within the Vocareum/Jupyter environment. The file will need to correctly link to the data file once uploaded, and you'll need to make sure that any package differences don't affect your work. We won't be able to help you debug this, so you will need to make sure this causes 0 issues before the deadline.

Will there be partial credit?
Grading will be done by your peers and may result in any number score in the range [0,3]. Everyone should be able to get either 2-3 points with enough effort.

Can I collaborate with another student?
Yes, but each student will need to submit their own work. You can share a dataset and work together to brainstorm methodologies, but your analysis must be your own. Please also name the student you worked with in the top block of your notebook.

How do I access and submit my peer reviews?
The peer review process is accessed through the same link in your LMS (Canvas or EdX). The day after the assignment closes (usually by noon ET the following day), you simply click on that link again and it will take you to an interface to view your peer submissions and submit grades + feedback.

Can my code import the data from a website instead of storing it in Vocareum?
No, the data must be uploaded into the Vocareum environment for peer grading.

One of my peers gave me a 0/3 and I think I deserve higher, can you regrade it?
No, there will be no re-grades.

Can the TA's take a look at my project before I submit it?
You may make a Piazza post to share the project with us and ask for some quick ideas, but given the time period we can't guarantee a high-quality or quick reply. Please also note that we'll provide our opinions, but the assignment is graded by your *peers*, so we're not a guaranteed source for a high score.

The question I have wasn't answered in this document.
Here's how to get an answer:

- If it's a **general question**: ask as a follow-up in the Extra Credit Piazza post
- If you want to brainstorm with others and get suggestions: create a public piazza post or ask in Slack. You can work with others to refine your ideas or get advice, just make sure the work in your notebook is your own.
- If it's specific to your project: create a private post but **please use your judgement**—we can't guarantee that we'll respond in a timely fashion, and while we would like to help everyone out with tips on their projects, we simply don't have the ability due to end of semester activities. We would prefer that you continue making progress on the project while waiting for a reply from us.