CSE 8A - Fall 2023

Project 1 - Data Science

Checkpoint Deadline: End of tutor hours Thursday 11/2
Project Deadline: 10pm Thursday 11/9

Overview

In this project, you will work individually to find a dataset on Kaggle that is meaningful to you, develop questions to ask of the data, analyze the data using Python code (with Copilot help), and answer those questions. The goal is for this activity to showcase your creativity and be valuable to you personally. Truly, we want you to have fun with this!

You'll have a checkpoint due on Thursday 11/2 where you will talk through your plans with a tutor who will help you scope the project well. For the project deadline, you'll submit your code, a diagram of the functions you wrote for the project, and a video explaining the project.

Your Tasks

Dataset

Kaggle.com has a wealth of datasets available publicly for download. You'll want to find a dataset that is interesting to you, has enough data that you can ask interesting questions of it, and the data you need should be stored in a *single* csv file. Kaggle lets you preview CSV files so you can get a good idea of what's in the data quickly.

Questions from the Data

As you are searching for a good dataset, you should be thinking about what questions you'd want answered. Suppose you find a dataset on world CO2 emissions, you might ask the following questions: which continents are the largest contributors, how CO2 emissions have changed over time, and which countries contribute the most CO2 per capita. Similarly, assuming you find a dataset on US prison parole data, you might ask: how do parole rates vary by gender and race overall, how do parole rates vary by the crime committed, and for a group of crimes, how do parole rates vary by gender and race. Another example, assuming you find a dataset on video game reviews and sales data: are game ratings and reviews correlated, what game has the most sales but the lowest reviews (and what are some exemplar comments in the reviews for that game), and which publisher produces games with the highest reviews vs. which publisher produces games with the lowest reviews (assuming at least a certain amount of sales). Have fun finding something interesting to you!

You may have really interesting questions, but can't find a dataset that has the data you need to answer those questions. So your search through the datasets and your questions will adapt dynamically - you ask a question, find some data, realize the data can answer a different question, etc.

Problem Decomposition

Once you've found the dataset and the questions you want to answer, you'll need to break up the overall problem into multiple functions. Maybe one function finds the average of a column, maybe another returns a subset of the data matching some criteria. You'll ultimately need to break up the overall problem into small enough tasks that Copilot can help you write the code for those steps. Once you've completed the problem decomposition, we'll expect you to create a figure like Figure 7.3 in the textbook. Please expand that figure to show the full function signatures and a brief description of each function.

Coding and Testing your Solution

In conjunction with the task of problem decomposition, you'll be authoring the functions (optionally with help from Copilot) to complete the tasks. Be sure to have a plan on how you'll test each function, likely using a smaller dataset rather than your full dataset.

Checkpoint - Due at end of tutor hours on Thursday 11/2 (5%)

Your meeting with the tutor will last roughly 5-10 minutes. This is an opportunity for you to get feedback on your plans for the project and for us to make sure you are on track to complete the project on time. The checkpoint is worth 5% of your project grade. Completing it before Tuesday 10/31 will yield an additional 1% bonus toward the project. A calendar with all the tutor hours can be found on the Course Calendar here.

What to have done before your meeting with the tutor:

Before meeting with the tutor, have the following done:

1. Have your dataset selected
2. Have your question of the data prepared
3. Have code that can open the csv file you are using and perform some task on the data
    a. Tasks could be simply printing elements of the data, finding the average of particular columns, etc.

What to expect from your meeting with the tutor:

You will request your meeting with the tutor either online or in-person using **Autograder** based on the Tutor Lab Hours posted on the Class Calendar. (Professor and TA Office Hours will not be used for checkpoints and will be reserved for other questions you may have about the class. In your meeting, the tutor will ask you about your dataset, the questions you've prepared, and ask you to show them your code and the dataset itself. If you have each item done, you'll receive full credit and the tutor will check you off as completing your checkpoint. The tutor will also offer you feedback about the difficulty of the task and you can ask them for advice on how to analyze the data.

Note that tutor hours are limited. Waiting until Thursday to ask for your checkpoint risks not getting your checkpoint completed in time. There are no late checkpoints, so not completing your checkpoint by the end of the last tutor hours on Thursday will result in a 0 for your checkpoint. We are happy to still discuss your project with you, etc., after the checkpoint deadline but no credit will be given.

Project Submission - due 10pm Thursday 11/9

You will be turning in 3 parts to Canvas for your project submission: your code, the diagram of your function hierarchy, and a link to a video of you explaining your project. See details below:

Code submission (20%)

You will upload all the code and your dataset to Canvas for grading. **The tutor grading your submission needs to be able to download your code and run it and have the code work as expected.** All questions you've asked and answered with your project need to be answered in the output of your project in a readable way. For example, you might output:

What is the gap in wages between people with and without a college degree?

The average wage for someone with a college degree from 1976-2022 is: $34.77
The average wage for someone with a high school degree from 1976-2022 is: $20.88

If your code outputs a graph, that's fine as well. Be sure the graph is clearly labeled (each axis/legend should be understandable) so we can interpret the results.

Problem Decomposition (20%)

You will upload an image summarizing your problem decomposition to Canvas. Specifically, how did you take your large problem and break it into functions that helped to solve the larger problems. Your image should be an expanded version of the example in Figure 7.3. Each function should have its inputs and outputs included and a brief description of the function.

Explanation Video (55%)

You will upload a link to a video of you explaining the project on Canvas. Your video should include the following:

Video quality/Details:

- Record the video with both your face and your presentation/code.
- Please either give us a link to the video in a **google drive folder** (that is public to anyone with the link) or to a private **Youtube** video.

The 5 minute video should contain:

- (1 min) You briefly describing your dataset and the questions you are asking of the data
- (1 min) You talking through why you decomposed the problem into the functions you did
- (3 min) A detailed walkthrough of one of your functions where you explain how it works.

Submitting Files

You will be submitting 3 things to Canvas

1. A zip file containing your Python code and your csv file(s)
2. A PNG or JPEG image of your project decomposition
3. A text file that contains the link to your video

Academic Integrity

You can ask for help from the instructional staff and Copilot. You should not discuss your project with other students in the class. We expect the questions you ask of your data and the code you write to be your own (the code can be aided by Copilot). Use of any existing analysis online is forbidden (as is using the project of a classmate) and turning in that code will result in your work being submitted for an Academic Integrity Violation.

Please see the class Academic Integrity Agreement for more details.