Case Study Rubric

Due Date:

TBD

Submission format:

- Upload link to GitHub repo to Canvas (or some other submission site)

Why am I doing this?

This is your opportunity to follow the steps provided by one of your peers to reproduce the results that they found in their projects. In Data Science, it is nice to be able to work with data pipelines and perform code analysis on the data, but it is also important to be able to reproduce and understand the information transmitted by others. This is the main reason we as data scientists perform the studies that we do, which is to inform others and to show them why it is important and how we did it. By following through with this case study, you will be able to train yourself on understanding your peers' workflows and hopefully learn something insightful about the topic of this case study.

What am I going to do?

You will read through this case study and follow each step-by-step instruction to reproduce the results of the individual before you. After completing all the steps listed in the **README** file in the GitHub repo, you will write a comprehensive 1-2 page report to demonstrate your understanding of this case study as well as share your personal insights on this case study.

Deliverables:

- Charts and Visualizations generated by the scripts
- 1-2 page report analyzing the results of this case study
- GitHub repo that contains the written report and the generated charts and visualizations

Tips for success:

- Read the instructions carefully. You do not want to miss steps, which could hinder the smooth progression of this case study.
- Try to understand the topic and process before going ahead and running the scripts. Just because you completed the case study, doesn't mean you understand it.
- Have patience. Downloading the data and training the model can take a while, so don't fret if you are not seeing any results immediately.

How will I know I have succeeded?

You will meet expectations on this case study when you follow the criteria in the rubric below.

Spec Category	Spec Details
Formatting	● GitHub repository ○ **Output** directory (should contain all outputs generated by the scripts) ○ **Scripts** directory (should contain all scripts used to generate the outputs) ○ **Report** PDF (1-2 pages)
Output	● This folder should contain all visualizations and evaluation metrics generated by your scripts. ○ Files should be pictures (.png, .jpeg, .jpg), not zipped
Scripts	● This folder should contain all scripts that were used in this case study. ○ If you made any changes to the scripts provided, explain your reasoning in the form comments next to the parts you changed/added.
Report	● This document should detail your findings after the completion of this case study. All of the following must be addressed: ○ What is the distribution of the data? (Is there an equal number of real images and fake images being analyzed?) ○ Do real images and fake images have the same or different overall brightness? If they do, is using brightness levels as a feature helpful in detecting fake facial images? If not, can we just assume fake images are naturally brighter/darker than real images? Justify your answer. ○ Do real images and fake images share the same average magnitude spectrum distribution? If yes, does this mean that it will be impossible for our trained CNN model to classify between real and fake images? (Hint: Average magnitude spectrum measures the amount of detail/texture in an image. Can CNN models still classify between two images of the same quality?) Justify your answer. ○ What are some constraints in training CNN models with images? (Hint: Think about the reasoning we moved to the Kaggle notebook environment for the second model training script.) ○ What are the results? (Did you get a model with relatively high performance → Accuracy ≥ 80%?) ○ What are some limitations of this trained CNN model? (What if we tested it on face images generated by a different AI model other than StyleGAN?) ○ Based on the Grad-CAM heatmaps, which part of the images does the trained CNN model primarily use to determine whether an image is real or fake? (They should all focus on a general area, which part is it and why?) ○ Write down any other insights you may have, as well as future improvements for this case study.