Practice Graded Assignment: Enhance a data science workflow and mitigate risks with the help of Microsoft Copilot

**AI Grading**

**Project goal**

In this project, you will choose a scenario, gather reference data for the scenario, and use Microsoft Copilot to supplement and/or augment your data. Then, you’ll apply your data to a data task analysis task. Finally, you’ll prepare a technical report that will evaluate Generative AI (GenAI) solutions for specific data science tasks while addressing data security and privacy concerns.

**Project scenario**

Let’s start with context. Imagine you are a promising data scientist. Based on a chosen scenario, you’ll gather your reference data and then will integrate cutting-edge generative AI into your workflow. Microsoft Copilot, a powerful AI assistant, is the tool you will use to streamline processes, supplement your data, and discover new insights. However, the potential risks of AI in data science, such as bias, misinformation, and data privacy breaches, must be carefully considered and mitigated. Your challenge is to demonstrate Copilot's benefits while ensuring ethical and responsible AI usage.

**Objective**

This project aims to produce a comprehensive portfolio showcasing your mastery of Microsoft Copilot in data science. Your portfolio should demonstrate your ability to:

* Utilize Copilot across various stages of the data science pipeline.
* Apply generative AI models (GANs, VAEs, or Transformers) to solve real-world data science problems.
* Identify and mitigate potential data security and privacy risks associated with generative AI.
* Communicate your findings, methodology, and ethical considerations effectively.

Add your Word document responses to your downloaded and saved template: [Create your Copilot productivity portfolio template.](https://docs.google.com/presentation/d/1Iqhw3211Kq-ov2FK8drnxZHPciaOZgdq/edit#slide=id.p1)

Your project will be evaluated on the following:

* **Copilot proficiency:** How effectively did you use Copilot throughout the project?
* **Technical knowledge:** Do you strongly understand generative AI and data science?
* **Problem-solving skills:** Did you effectively identify challenges and apply AI solutions with Copilot's help?
* **Ethical awareness:** Did you consider and address the potential ethical implications of generative AI? What steps did you take to anonymize data or avoid using sensitive data?
* **Communication skills:** Are your project objectives, methodology, results, and implications communicated in both written and (optional) oral formats?

Select one of the following project scenarios or use them as inspiration to design your own:

* **Customer churn prediction:** Use Copilot to enhance customer churn analysis and apply a generative AI model to synthesize customer profiles or behaviors.
* **Image generation or enhancement:** Employ Copilot to streamline image processing tasks and use a GAN to generate realistic images or enhance existing ones.
* **Text generation or summarization:** Leverage Copilot for text preprocessing and apply a Transformer model for text generation or summarization tasks.
* **Anomaly detection:** Combine Copilot's data exploration capabilities with a generative model to identify anomalies in datasets (e.g., fraud detection, network intrusion).
* **Recommendation systems:** Use Copilot to optimize feature engineering and apply a generative model to personalize product or content recommendations.

**Create a synthetic dataset with GitHub Copilot by following these steps:**

1. **Define the use case and data structure**
   1. Clarify the purpose and task: Utilize Copilot's code suggestions to craft explicit comments that outline the specific purpose of your synthetic dataset and the task it will be used for.
   2. Gather your reference data: Engage Copilot, using targeted prompts, to create the reference data for your chosen scenario.
   3. Define the data structure and variables: With Copilot's assistance, define the structure of your reference data (the real-world data you aim to emulate). Specify whether it's tabular, image-based, text-based, time-series, or another format. Describe the variables involved (e.g., numerical, categorical, ordinal).
2. **Data exploration and statistical analysis**
   1. Analyze the reference data: Leverage Copilot's code completion to explore your reference data. Copilot can recommend libraries and functions to:
      1. Understand the distributions of various variables.
      2. Identify relationships and correlations between variables.
      3. Detect any biases or outliers present in the data.
   2. Calculate statistical measures: Calculate key statistical measures such as mean, median, standard deviation, correlations, and covariances with Copilot's help to understand your reference data better.
3. **Select a synthetic data generation model**
   1. Choose the right model: Copilot can provide code snippets for various synthetic data generation methods tailored to your data structure, use case, and complexity.
   2. Explore Copilot's suggestions: Depending on your specific needs, consider Copilot's suggestions for statistical models, machine learning models (like VAEs or GANs), agent-based models, or rule-based models.
4. **Model training and validation**
   1. Accelerate model development: Copilot can expedite model training and validation by offering code suggestions for:
      1. Data preprocessing
      2. Model architecture
      3. Hyperparameter tuning
      4. Evaluation metrics
   2. Compare synthetic and reference data: Utilize Copilot's assistance to compare the generated synthetic data to your reference data using statistical tests and visual inspection to ensure they align closely.
5. **Generate synthetic data**
   1. Generate samples: Copilot can aid in writing code to generate synthetic data samples using the trained model or chosen method.
   2. Preserve data characteristics: Ensure that the generated data maintains the statistical properties, relationships, and any privacy requirements of your reference data with Copilot's code recommendations.
6. **Post-processing and evaluation**
   1. Refine and prepare the data: Copilot can suggest code for post-processing tasks like data cleaning, formatting, or anonymization to ensure the synthetic data is ready for use.
   2. Evaluate the quality: Assess the quality of the synthetic dataset by comparing its performance on your intended task to that of the reference data, with Copilot's help in writing evaluation code.

The ability to generate synthetic data that accurately reflects real-world scenarios or fulfills specific requirements is invaluable. Let’s see how to use your synthetic dataset.

* **Utilize the synthetic data** Now that you've generated your synthetic dataset, it's time to put it to use. This part will focus on how you'll employ the synthetic data for your intended task, such as training a machine learning model, testing an algorithm, or conducting research.
* **Outline the analysis task** Describe the specific analysis or task you'll perform using the synthetic data. Below, we have the three most common analysis tasks and decisions. Think of the task descriptions as high-level summaries to guide your overall approach. They help you frame the purpose of your analysis and provide context for the code you'll be writing with Copilot's assistance.
  + **Example 1: Training a machine learning model** Task description: "In this part, we will utilize the generated synthetic dataset to train a classification model for predicting customer churn. We'll explore various algorithms, tune hyperparameters, and evaluate model performance using appropriate metrics."
  + **Example 2: Testing an algorithm** Task description: "We will employ the synthetic dataset to rigorously test the performance and robustness of our newly developed anomaly detection algorithm. We'll assess its ability to identify outliers and unusual patterns in the data under different scenarios."
  + **Example 3: Data augmentation** Task description: "We'll leverage the synthetic data to augment our existing dataset for training an image recognition model. This will help improve the model's generalization capabilities and reduce overfitting."
* **Leverage Copilot for code** Continue to leverage Copilot's code suggestions as you write the code to execute your analysis or task. Copilot can assist with data loading, preprocessing, model building (if applicable), and result interpretation.
* **Evaluate and interpret results** Analyze the results obtained from using the synthetic data. Based on your findings, draw conclusions and insights.
* **Task definition:** Clearly define the data science problem and its significance.
* **Copilot integration:** Detail how Copilot was used at every project stage (ideation, code generation, model building, evaluation).
* **Generative AI solution:** Explain the selected model, its implementation, and Copilot's contribution.
* **Risk mitigation strategy:** Outline a plan to address potential data security and privacy risks using insights and recommendations from Copilot.