Workshop Handout Synthetic Data for Research: Epistemic and Practical Directions

Tsehaye Haidemariam, PhD

Prepared for:

Generative AI for Social Science Research

Norwegian Business School Workshop | 2025

Handout (For Participants)

Part I: Step-by-Step Guide to Generating Synthetic Data Using GPT-4

1. Define the Objective

- Decide on the goal and format of your synthetic data. Is it survey responses, interviews, dialogue, etc.?
- Example use case: training NLP models or simulating qualitative research data.

2. Create a Schema or Template

- Define the structure your data should follow.
- *Example JSON structure:*

{"age": 29,

"gender": "Female",

"response": "The program has improved access to clean energy." }

3. Write a Clear Prompt

- GPT-4 needs specific and structured prompts.
- *Example:* “Generate 10 synthetic interview responses from urban dwellers about climate change. Include name, age, city, and a quote.”

4. Generate the Data

- Use GPT-4 (via ChatGPT or OpenAI API) and paste your prompt.
- Review for clarity and logical consistency.

5. Post-Process the Data

- Clean and convert output into formats like CSV or JSON.
- Tokenize or annotate if needed for machine learning.

6. Validate the Data

- Check for logical errors (e.g., age 5 with college degree).
- Ensure variety and avoid stereotypes.

7. Apply in Research

- Use synthetic data in NLP tasks, simulations, or educational settings.

8. Document the Process

- Save prompt versions and outputs.
- Explain your generation and validation steps.

Part II: Step-by-Step Guide to Analysing and Visualizing Synthetic Data Using GPT-4

(*This **Guide** assumes you are using GPT-4 through **ChatGPT (Pro)** or via the **API**, optionally with a Python coding environment (like Jupyter or Google Colab*).

Step 1: Generate Your Synthetic Data

Use a structured prompt via ChatGPT or OpenAI API.

Example Prompt:

Generate 20 synthetic records of survey responses. Each record should include:

- Name
- Age
- Occupation
- Country
- Satisfaction Score (1-5)
- Free-text comment
  Return the output as a JSON or CSV-like table.

Step 2: Ask GPT-4 to Convert It into Code

Once GPT-4 returns the data:

- Say:

“Can you write a Python script to convert this data into a pandas DataFrame and calculate the average satisfaction by occupation?”

GPT-4 will provide something like:

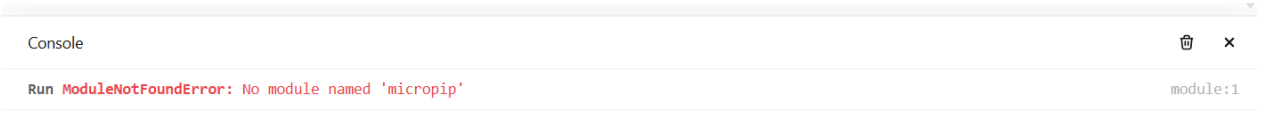
```python
import pandas as pd

# Paste your generated data here (or load from file)
data = [
    {"name": "Anna", "age": 32, "occupation": "Teacher", "country": "USA", "satisfaction": 4,
    {"name": "Liam", "age": 29, "occupation": "Engineer", "country": "Canada", "satisfaction"
    ...
]

df = pd.DataFrame(data)
print(df.groupby("occupation")["satisfaction"].mean())
```

Next click “edit”: GPT-4 will provide something like:

Next click “Run”: If you have the right software installed in your local machine, GPT4 will provide you with the result you requested. If you do not have the required software, for example, micropip which is a lightweight package installer for Pyodide and other projects that use Pyodide, GPT-4 will provide something like:



What the error means is that you need to install micropip in your local machine.

Next you have two choices: Either you install micropip in your local machine or ask GPT4 to provide you with an alternative solution that does not require micropip or use a one go alternative (see the Bonus Section)

Step 3: Ask GPT-4 to Create Visualizations

Just say:

“Plot the average satisfaction by occupation using matplotlib or seaborn.”

GPT-4 might return:

python Copy Edit

```python
import seaborn as sns
import matplotlib.pyplot as plt

avg_satisfaction = df.groupby("occupation")["satisfaction"].mean().reset_index()

sns.barplot(data=avg_satisfaction, x="occupation", y="satisfaction")
plt.title("Average Satisfaction by Occupation")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

You can also request:

- Pie charts (e.g., occupation distribution)
- Histograms (e.g., age)
- Word clouds (from free-text comments)
- Sentiment analysis (using TextBlob or GPT-4)

Step 4: Deeper Analysis with GPT-4’s Help

Ask GPT-4 to:

- Detect outliers

- Segment data by country
- Perform sentiment scoring from comments
- Suggest clustering (e.g., K-means) for patterns

Step 5: Export Your Results

Once analysis is done, ask:

"How can I export this DataFrame as a CSV file?"

GPT-4 will suggest:

```python
df.to_csv("synthetic_data_analysis.csv", index=False)
```

Bonus: Automate the Loop

Want to regenerate, analyse, and visualize in one go? Ask GPT-4:

"Write a full script that generates synthetic survey data, analyses average satisfaction, and visualizes results in one go."

To use GPT-4 with the OpenAI API to Generate Synthetic Data

- Go to https://platform.openai.com/
- Sign in with your OpenAI account.
- Navigate to API keys under your account settings.
- Click Create new secret key and copy it safely (you will need this in your script).