

ex - 1c

October 31, 2025

```
[ ]: import pandas as pd
import json
from xml.etree import ElementTree as ET

def structured_example():
    data = {
        'id': [1,2,3],
        'name': ['Alice','Bob','Carol'],
        'age': [29, 34, 23]
    }
    df = pd.DataFrame(data)
    print("Structured data (pandas DataFrame):")
    print(df)
    df.to_csv('structured_example.csv', index=False)
    print("Saved to structured_example.csv")

def unstructured_example():
    docs = [
        "Today I attended a data science meetup and learned about transformers.",
        ↪ "Error: Connection refused at 2025-10-31 10:12:00 - service X failed.",
        "Image: binary data (not text) - e.g. photos, audio transcripts"
    ]
    print("\nUnstructured data (plain text documents):")
    for i, doc in enumerate(docs,1):
        print(f"Doc {i}: {doc}")
    with open('unstructured_example.txt','w',encoding='utf-8') as f:
        for d in docs:
            f.write(d + "\n")

def semi_structured_example():
    items = [
        {"id":1, "name":"Alice", "skills":["python","sql"]},
        {"id":2, "name":"Bob", "contact":{"email":"bob@example.com","phone":
        ↪ "12345"}},
        {"id":3, "name":"Carol", "notes":"Prefers remote"}
    ]
```

```

print("\nSemi-structured data (JSON-like):")
print(json.dumps(items, indent=2))
with open('semi_structured_example.json', 'w', encoding='utf-8') as f:
    json.dump(items, f, indent=2)

def xml_example():
    root = ET.Element('employees')
    e1 = ET.SubElement(root, 'employee', attrib={'id': '1'})
    ET.SubElement(e1, 'name').text = 'Alice'
    ET.SubElement(e1, 'role').text = 'Data Scientist'
    tree = ET.ElementTree(root)
    tree.write('semi_structured_example.xml', encoding='utf-8',
    ↪xml_declaration=True)
    print("\nWrote semi_structured_example.xml (XML is semi-structured)")

if __name__ == "__main__":
    structured_example()
    unstructured_example()
    semi_structured_example()
    xml_example()

    print("\nCharacteristics summary:")
    print("- Structured: rigid schema, easy to query (e.g., SQL tables, CSV).")
    print("- Unstructured: no predefined schema (text, images), needs parsing/
    ↪NLP/vision.")
    print("- Semi-structured: tags/keys but not rigid (JSON, XML, logs with key:
    ↪value).")

```