Task -3

1. Why is data cleaning important in real-time data processing?

- Real-time systems depend on **accurate, consistent, and reliable data**.
- Without cleaning, issues like **missing values, inconsistent formats, and incorrect data types** can cause wrong decisions, system crashes, or misleading analytics.
- Cleaning ensures that **streaming dashboards, machine learning models, and automated alerts** work with **trustworthy input**.

  ☞ In short: **clean data = accurate insights + stable real-time pipelines.**

---

2. What are pipeline artifacts and how are they used in DevOps workflows?

- **Pipeline artifacts** are the **files produced by a pipeline run** (e.g., processed CSVs, logs, build outputs).
- In **Azure DevOps**, they allow sharing outputs between pipeline jobs and stages.
- Example: one stage generates `clean_sales_data.csv`, and another stage (like deployment) can **download and use it directly**.
- They also provide traceability, since artifacts can be **downloaded later for debugging, testing, or reporting**.

  ☞ Think of artifacts as the **handoff mechanism** inside DevOps pipelines.

---

3. How would you modify the pipeline to store the cleaned data into Azure Blob Storage?

To extend the pipeline for **Azure Blob Storage integration**:

1. **Install Azure SDK** in the pipeline (`pip install azure-storage-blob`).
2. Add **environment variables** for credentials in Azure DevOps → Pipeline → Variables:
   - `AZURE_STORAGE_ACCOUNT_NAME`
   - `AZURE_STORAGE_ACCOUNT_KEY`
   - `AZURE_CONTAINER_NAME`
3. Update `data_processing.py` to include an **upload step** using `BlobServiceClient`.