Exercise 1:

Data Types and Structures:

- What are the common data types used in data science libraries like NumPy and Pandas?
- How do these data types differ from those used in general-purpose programming languages?
- How can understanding data types help you choose the appropriate data structures and algorithms for your data science tasks?
- Can you describe a specific instance where a data type mismatch caused issues in your project?
- What strategies do you use to optimize memory usage and performance when working with large datasets and complex data structures?
- Are there any specific data types that you find particularly challenging to work with? If so, why?
- How do you ensure that the data types you choose are compatible with the algorithms and tools you are using?
- What are some common challenges in working with data structures in data science projects?

Exercise 2:

In next statements, identify the data challenge: data sparsity, data dimensionality, or data resolution.

- An e-commerce platform is attempting to understand customer journeys by analyzing interactions on its website. However, most customers only interact with a small subset of available features.
- A healthcare analytics team is working with patient data, including medical records, genetic information, and lifestyle factors. However, they notice that some features have a high number of missing values
- A credit card company is developing a system to detect fraudulent transactions automatically. However, instances of fraud are rare in the dataset, and most transactions are legitimate.
- An image recognition system is designed to identify objects in high-resolution images. However, in some cases, the images captured have low contrast and limited details, especially in low light conditions.
- A social media analytics company is analyzing user interactions on a platform with billions of users. They are finding it challenging to process and extract meaningful insights from such massive datasets.
- A government agency is analyzing population data across different regions. However, some regions have sparse population data, leading to challenges in making accurate demographic predictions

Exercise 3:

<u>General Data Collection Methods:</u>

- What are the primary methods for collecting data?
- How do you choose the appropriate data collection method for a specific research question?

<u>Quantitative Data Collection:</u>

- What are the main methods for collecting quantitative data?
- How do you design a good survey or questionnaire?
- What are the key considerations for conducting experiments?
- What are the challenges of collecting quantitative data from large populations?

<u>Qualitative Data Collection:</u>

- What are the main methods for collecting qualitative data?
- How do you develop effective interview questions?
- What are the key considerations for conducting focus groups?
- What are the challenges of analyzing qualitative data?

Exercise 3:

<u>Parallel Processing:</u>

- How does parallel processing improve the efficiency of big data processing?
- What are the key challenges in designing and implementing parallel algorithms for big data?

<u>Batch Processing:</u>

- When is batch processing more suitable than streaming processing?
- How can batch processing be optimized for large datasets and complex workflows?

Streaming Processing:

- What are the key challenges in processing real-time data streams at scale?
- How can streaming processing be used for anomaly detection and real-time analytics?

Exercise 4

Data integration techniques:

- What are the key differences between exact matching and partial matching for identifying duplicate records?
- How can you determine the appropriate threshold for correlation coefficients when analyzing attribute correlations to identify redundancy?

Semantic Data Integration:

- How does semantic data integration differ from traditional data integration? What are the key components of an ontology-based approach to semantic data integration?
- How can ontologies be used to establish correspondences between concepts and relationships across different schemas?

Scheme Matching:

- What are the challenges of scheme matching in the presence of heterogeneous data sources and different terminologies?
- What are the trade-offs between manual and automated scheme matching techniques?

Bias in DATA:

- How can biases in data be detected and mitigated? What are some common sources of bias in datasets?
- What are the potential consequences of using biased data in data science models? How can bias impact the fairness and accuracy of outcomes?
- How can bias be addressed in algorithms and models? Are there specific techniques or approaches to mitigate bias?

Exercise 5:

You are given two datasets from different e-commerce platforms. Each dataset contains product information, but they have different structures and terminologies for similar data. Your task is to integrate the datasets semantically to create a unified view of the product catalog:

DataSet1: ProductID | Prod_Name | Category | Cost_USD | Stock

DataSet2: Item_ID | Item_Description | Item_Type | Price_in_Dollars | Quantity_Available

- Identify Semantic Similarities:
  - Map equivalent fields from both datasets (e.g., ProductID and Item_ID, Prod_Name and Item_Description, etc.).
- Create a Unified Schema:
  - Develop a single schema that represents both datasets. For example: Unified_ProductID | Name | Category | Price | Stock_Level
- Data Transformation:
  - Convert both datasets to the unified schema. Make sure to transform values where necessary (e.g., converting price to a uniform currency if needed).
- Data Merging:
  - Merge the two datasets into a single, unified dataset based on the mapped fields.
- Handling Conflicts:
  - If there are inconsistencies between datasets (e.g., different prices for the same product), define rules to resolve the conflicts.