Exercise 1: Lecture questions:

Understanding EDA Concepts and Processes

- What are the main objectives of Exploratory Data Analysis (EDA)?
- Why is visualization considered an essential part of EDA?
- How do summary statistics help in understanding data distribution?

Outliers, Data Cleaning, and Preprocessing

- What challenges can outliers present, and how should they be handled during EDA?
- What is the role of data cleaning in the EDA process?

Detecting Patterns, Relationships, and Hypothesis Testing

- How does EDA help detect correlations and trends among variables?
- What are some examples of hypotheses that can be formulated through EDA?

Descriptive and Inferential Statistics

- What are the most common descriptive statistics used in EDA, and why are they useful?
- How do inferential statistics help confirm hypotheses generated during EDA?
- Can you explain the difference between summarizing data and testing hypotheses?

~~**Application and Practical Use Cases**~~

- ~~What is the role of hypothesis testing in data science projects?~~
- ~~Describe how an A/B test can be used to validate a hypothesis in EDA.~~
- ~~Why is smoothing important, and how does it help highlight trends in noisy data?~~

Outlier Detection Techniques

- What statistical methods are commonly used to detect outliers? Provide a brief explanation of each.
- How can Z-score and Interquartile Range (IQR) be used to identify outliers? What are the key differences between them?
- ~~What role do visualizations such as box plots, scatter plots, and histograms play in outlier detection? Which type of plot is most effective for detecting outliers in univariate data?~~

Mitigation Strategies for Outliers

- What is trimming, and when is it appropriate to use this strategy to handle outliers?
- How does applying transformations, such as a log transformation, reduce the influence of outliers in a dataset?
- What is imputation, and how can it be used to handle outliers without removing them from the data?
- In what situations would you prefer to use the median over the mean when imputing outliers? Why?
- What are the potential risks of ignoring outliers during the analysis process?

Exercise 2:

Given the next sample dataset with eight rows, answer the followed questions related to:

1. Z-score for outlier detection
2. IQR for outlier detection
3. Logarithmic transformation for mitigation

ID	Value
1	12
2	15
3	14
4	10
5	200
6	13
7	9
8	11

Part 1: Z-Score for Outlier Detection

- Calculate the mean and standard deviation of the Value column.
- Compute the Z-scores for each row in the dataset.
- Which values, if any, are considered outliers using a Z-score threshold of 3 (i.e., absolute value of Z-score > 3)?

Part 2: IQR for Outlier Detection

- Find the first quartile (Q1) and third quartile (Q3) of the Value column.
- Calculate the IQR (Q3 - Q1).
- Determine the lower and upper bounds for outlier detection (using 1.5 * IQR).
- Identify which values are considered outliers based on these IQR bounds.

Part 3: Logarithmic Transformation for Mitigation

- Apply a logarithmic transformation (using base 10) to the Value column.
- What effect does the transformation have on the outlier value (200)?
- Why is a logarithmic transformation helpful in reducing the impact of outliers?

Exercise 2: Given the data below, answer the following questions.

ID	Feature 1	Feature 2
1	12	15
2	14	18
3	13	16
4	200	5
5	10	11
6	12	14
7	9	10
8	13	17
9	300	8
10	11	12

Part 1: KNN for Outlier Detection

- Explain how KNN can be used to detect outliers.
- Calculate the Euclidean distance between point (ID = 1) and all other points in the dataset.
- Find the 2 nearest neighbors for each data point (using Euclidean distance).
- Use a threshold (e.g., mean + 2 * standard deviation of distances) to determine which points are outliers based on their distances to neighbors.
- Identify which points (if any) are potential outliers in the dataset.

Part 2: Mitigating Outliers Using KNN

- If point 9 (300, 8) is identified as an outlier, suggest how it could be mitigated using KNN-based imputation.
- What are the advantages and limitations of using KNN for outlier detection and mitigation?
- Discuss when KNN might be a better choice for outlier detection compared to Z-scores or IQR methods.

Exercise 4:

The dataset below is a table containing a set of data points (two features: Feature 1 and Feature 2). Assume the following DBSCAN parameters:

- **eps (ε)**: 1.0 and **min_samples**: 2

Point	**Feature 1**	**Feature 2**
A	1	2
B	1	2.5
C	1.5	2
D	10	10
E	1	1.5
F	2	2
G	1.1	2.1
H	0	0

Questions:

1. **Core, Border, and Noise Points**:
   - Identify and classify each point as a **Core Point**, **Border Point**, or **Noise Point** based on the provided parameters.
     - **Core Point**: A point with at least min_samples points (including itself) within the eps neighborhood.
     - **Border Point**: A point that is not a core point but is within the neighborhood of a core point.
     - **Noise Point**: A point that is neither a core point nor a border point.
2. **Identification of Outliers**:
   - Based on your classification, list which points are considered outliers. Provide reasoning for each classification.
3. **Parameter Impact**:
   - Discuss how changing the parameters eps and min_samples might affect the classification of the points.
     - What would happen if eps were increased to 2.0?
     - What would happen if min_samples were increased to 3?
4. **Scenario Analysis**:
   - If you were to run DBSCAN on a dataset with the following characteristics, would you expect it to perform well? Why or why not?
     - A dataset with a large number of outliers and varying densities of clusters.
     - A dataset with spherical clusters and minimal noise.