Exercise 1.

You are analyzing a dataset with numerous features related to customer demographicsand purchasing behavior to identify patterns that can inform marketing strategies. You decide to apply Principal Component Analysis (PCA) to reduce the dimensionality of your data set.

a. Explain the main goals of PCA. What are the advantages of using PCA in data analysis?

b. Describe the steps involved in performing PCA. What are the key mathematical concepts behind PCA that facilitate dimensionality reduction?

c. After applying PCA, you notice that the first two principal components explain 85% of the variance in the data. Discuss what this indicates about the dataset and how it might affect your analysis.

d. Suppose you choose to visualize the data using the first two principal components. What are some important considerations to keep in mind when interpreting the results of this visualization?

Exercise 2

You are provided with a dataset containing information on individuals' lifestyle choices (e.g., exercise frequency, diet quality), socioeconomic status (e.g., income, education level), and health outcomes (e.g., BMI, incidence of chronic diseases).

a. Formulate a causal hypothesis related to the impact of exercise frequency on BMI. Clearly define the variables involved.

b. Choose a causal inference method to test your hypothesis. Justify your choice and outline the steps you would take to implement this method.

c. After analyzing the data, you find a statistically significant relationship between exercise frequency and BMI. Discuss what this result implies about causality, including any potential confounding variables that might affect the interpretation of your findings.

d. Based on your analysis, what recommendations would you provide to public health officials? Discuss how your findings could inform policy or intervention programs.

Exercise 1

a. Main goals of PCA & Advantages

Goals:

- **Dimensionality Reduction:** Reduce the number of variables while retaining as much information (variance) as possible.
- **Uncorrelated Components:** Transform correlated features into a set of uncorrelated principal components.
- **Identify Patterns:** Reveal underlying structure in data (latent factors).

Advantages:

- **Simplifies Models:** Reduces complexity → faster training and simpler interpretation.
- **Noise Reduction:** By discarding components with low variance, PCA may reduce noise.
- **Visualization:** Makes it possible to visualize high-dimensional data in 2D or 3D.
- **Preprocessing for ML:** Often improves performance of algorithms sensitive to multicollinearity.

b. Steps & Key Mathematical Concepts

Steps:

1. **Standardize Data** – Scale features to zero mean and unit variance to ensure comparability.
2. **Compute Covariance Matrix** – Captures how variables vary together.
3. **Eigen Decomposition (or SVD):** Find eigenvalues & eigenvectors of the covariance matrix.
4. **Sort Eigenvectors by Eigenvalues:** Largest eigenvalues → directions of greatest variance.
5. **Form Principal Components:** Each PC = linear combination of original features (using eigenvectors as coefficients).
6. **Project Data:** Transform original data onto the new principal component axes.

Key Concepts:

- **Variance Maximization:** PCA finds directions of maximal variance.
- **Orthogonality:** PCs are orthogonal (uncorrelated).
- **Eigenvectors/Eigenvalues:** Eigenvectors = directions (axes); eigenvalues = amount of variance captured.

c. Interpretation of "First two PCs explain 85% variance"

- **High Variance Capture:** 85% of the dataset's variability can be explained by just 2 components (huge reduction from original features).
- **Implication:** The dataset is highly redundant — many features convey similar information.

- **Effect on Analysis:**
  - Using two PCs is likely sufficient for capturing main patterns.
  - But 15% of variance is still unexplained, so some details may be lost.

d. Considerations when Visualizing First Two PCs

- **Information Loss:** Only 85% variance is shown — patterns from the remaining 15% are hidden.
- **Axes are Abstract:** PCs are linear combinations of features, not original variables; interpreting them requires examining loadings.
- **Cluster Interpretations:** Clusters seen in the plot may not correspond to actual customer segments in original space.
- **Scaling & Outliers:** Outliers can influence PCs heavily.

Exercise 2

a. Formulate a Causal Hypothesis

- **Hypothesis:** "Increased exercise frequency causally reduces Body Mass Index (BMI) among adults."
- **Variables:**
  - **Treatment (independent variable):** Exercise frequency (e.g., days/week).
  - **Outcome (dependent variable):** BMI.
  - **Covariates / Confounders:** Diet quality, income, education, age, genetics, etc.

b. Choose a Causal Inference Method

Method: Propensity Score Matching (PSM

Justification:

- Observational data (not randomized) → need to balance groups on confounders.
- PSM helps approximate randomized conditions by matching individuals with similar covariates but different exercise frequency.

Steps:

1. **Model Propensity Scores:** Use logistic regression to estimate probability of "high exercise" given covariates (age, diet, income, etc.).
2. **Match Individuals:** Pair high-exercise and low-exercise individuals with similar propensity scores.
3. **Compare Outcomes:** Compute average BMI differences between matched groups.
4. **Check Balance:** Ensure covariate distributions are similar after matching.

c. Interpreting Statistically Significant Relationship

- **Statistical Significance ≠ Causality:** Even after matching, unmeasured confounders may bias results.
- **Potential Confounders:** Diet quality, pre-existing health conditions, genetics, stress levels, medication, etc.
- **Implication:** Evidence is consistent with a causal effect, but not proof. True causality would require randomized controlled trial (RCT) or a strong natural experiment/instrument.

d. Recommendations for Public Health Officials

- **Intervention Suggestion:** Encourage regular physical activity as part of public health campaigns.
- **Complementary Strategies:** Pair exercise promotion with nutrition and lifestyle education, since diet also affects BMI.
- **Targeted Programs:** Focus on at-risk populations (e.g., low-income groups with limited access to exercise facilities).
- **Policy Impact:** Use findings to justify investments in public parks, subsidies for fitness programs, or employer-sponsored exercise initiatives.

Concise Takeaway Table

Exercise	Key Answer
1a	PCA reduces dimensions, removes correlation, simplifies analysis
1b	Standardize → Covariance → Eigenvalues/Vectors → PCs → Project data
1c	2 PCs = 85% variance → strong dimensionality reduction
1d	PCs are abstract; watch for lost variance and misinterpretation
2a	Hypothesis: More exercise → lower BMI (treatment: exercise, outcome: BMI)
2b	Use PSM (or another method) to control for confounding
2c	Significant correlation doesn't prove causation; confounders may remain
2d	Recommend integrated exercise & nutrition interventions