# **Question 1:**

You have obtained the following dataset from a government organisation. The dataset contains socio-economic indicators across different regions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Region** | **GDP**  **per capita (USD)** | **Unemployment Rate (%)** | **Education Expenditure per capita (USD)** | **Healthcare Expenditure per capita (USD)** | **Crime Rate (per 1000 inhabitants)** |
| A | 25000 | 5 | 1500 | 2000 | 10 |
| B | 22000 | 7 | 1200 | 1800 | 15 |
| C | 28000 | 4 | 1800 | 2200 | 8 |
| D | 20000 | 9 | 1000 | 1500 | 20 |
| E | 30000 | 3 | 2000 | 2500 | 5 |
| F | 26000 | 6 | 1600 | 2100 | 12 |
| G | 24000 | 8 | 1300 | 1900 | 18 |
| H | 27000 | 5 | 1700 | 2300 | 9 |
| I | 23000 | 4 | 1400 | 1700 | 14 |
| J | 29000 | 2 | 1900 | 2400 | 7 |

1. Run the Principal Component Analysis on this dataset and interpret the result.
2. Based on your result what policy you can propose to the government?

# **Question 2:**

A large online retailer, E-Commerce Giant, wants to understand its customers' preferences and behaviors better to tailor its marketing strategies and product offerings. The company has collected a dataset containing information about customer demographics, purchase history, and website interactions.

# Data:

The dataset, customer\_data.csv, includes the following variables:

* + customer\_id: Unique identifier for each customer
  + age: Customer's age
  + income: Customer's annual income
  + purchase\_frequency: Number of purchases made in the past year
  + average\_order\_value: Average value of customer orders
  + website\_visits: Number of visits to the website in the past year
  + time\_spent\_on\_site: Average time spent on the website per visit

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | age | income | purchase\_frequency | average\_order\_value | website\_visits | time\_spent\_on\_site |
| 1 | 35 | 50000 | 12 | 100 | 50 | 30 |
| 2 | 28 | 60000 | 8 | 150 | 30 | 25 |
| 3 | 42 | 45000 | 10 | 80 | 40 | 35 |
| 4 | 30 | 70000 | 15 | 200 | 60 | 40 |
| 5 | 25 | 55000 | 11 | 90 | 45 | 32 |
| 6 | 38 | 65000 | 9 | 130 | 35 | 28 |
| 7 | 40 | 40000 | 8 | 70 | 30 | 30 |
| 8 | 27 | 75000 | 16 | 220 | 70 | 45 |
| 9 | 32 | 50000 | 10 | 100 | 40 | 35 |
| 10 | 35 | 60000 | 12 | 150 | 50 | 30 |
| 11 | 45 | 55000 | 13 | 120 | 45 | 38 |
| 12 | 32 | 65000 | 7 | 100 | 30 | 25 |
| 13 | 28 | 40000 | 9 | 80 | 35 | 30 |
| 14 | 35 | 70000 | 14 | 180 | 55 | 40 |
| 15 | 27 | 50000 | 10 | 90 | 40 | 32 |

Use R or Python to run principal component analysis on that data, show and explain the R or Python code you are using, and interpret the findings. Suggest some policies from your analysis.

Question 3:

You have been provided with a dataset containing information about 15 individuals, with three variables recorded for each person: Age, Income, and Education Level.

*Table 5*

|  |  |  |  |
| --- | --- | --- | --- |
| **Person** | **Age** | **Income** | **Education Level** |
| 1 | 30 | 50000 | 12 |
| 2 | 35 | 60000 | 14 |
| 3 | 25 | 40000 | 10 |
| 4 | 40 | 70000 | 16 |
| 5 | 45 | 80000 | 18 |
| 6 | 28 | 45000 | 11 |
| 7 | 38 | 65000 | 15 |
| 8 | 50 | 90000 | 20 |
| 9 | 32 | 55000 | 13 |
| 10 | 42 | 75000 | 17 |
| 11 | 33 | 58000 | 14 |
| 12 | 48 | 85000 | 19 |
| 13 | 27 | 48000 | 12 |
| 14 | 36 | 62000 | 14 |
| 15 | 44 | 78000 | 18 |

1. Discuss the differences between Principal Component Analysis (PCA) and Clustering.
2. Choose whether to apply PCA or Clustering to the given dataset and provide justification.
3. Explain the results of the chosen method.