**Part 1: PCA:**

**Problem Statement:** The [‘Hair Salon.csv’View in a new window](https://olympus.mygreatlearning.com/courses/80501/files/6794274/download?wrap=1) dataset contains various variables used for the context of Market Segmentation. This particular case study is based on various parameters of a salon chain of hair products. You are expected to do Principal Component Analysis for this case study according to the instructions given in the rubric. **Kindly refer to the**[**PCA\_Data\_Dictionary.jpg**](https://olympus.mygreatlearning.com/courses/80501/files/6794275/download?wrap=1)**[View in a new window](https://olympus.mygreatlearning.com/courses/80501/files/6794275/download?wrap=1) file for the Data Dictionary of the Dataset.**  
**Note: This particular dataset contains the target variable satisfaction as well. Please do drop this variable before doing Principal Component Analysis.**

|  |
| --- |
|  |

|  |
| --- |
| Part 1: PCA: Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented. (5marks) |
| Part 1: PCA: Scale the variables and write the inference for using the type of scaling function for this case study. (3 marks) |
| Part 1: PCA: Comment on the comparison between covariance and the correlation matrix after scaling. (2 marks) |
| Part 1: PCA: Check the dataset for outliers before and after scaling. Draw your inferences from this exercise. (3 marks) |
| Part 1: PCA: Build the covariance matrix, eigenvalues and eigenvector. (4 marks) |
| Part 1: PCA: Write the explicit form of the first PC (in terms of Eigen Vectors). (5 marks) |
| Part 1: PCA: Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame. (8 marks) |
| Part 1: PCA: Mention the business implication of using the Principal Component Analysis for this case study. (5 marks) |

**Part 2: Clustering:**

[The View in a new windowState\_wise\_Health\_income.csvView in a new window](https://olympus.mygreatlearning.com/courses/80501/files/6794277/download?wrap=1) dataset given is about the Health and economic conditions in different States of a country. The Group States based on how similar their situation is, so as to provide these groups to the government so that appropriate measures can be taken to escalate their Health and Economic conditions.  
2.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc, etc)  
2.2. Do you think scaling is necessary for clustering in this case? Justify  
2.3. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.  
2.4. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score.  
2.5. Describe cluster profiles for the clusters defined. Recommend different priority based actions that need to be taken for different clusters on the bases of their vulnerability situations according to their Economic and Health Conditions.  
**Data Dictionary for State\_wise\_Health\_income Dataset:**  
1. States: names of States  
2. Health\_indeces1: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in the State.  
3. Health\_indeces2: A composite index rolls several related measures (indicators) into a single score that provides a summary of how the health system is performing in certain areas of the States.  
4. Per\_capita\_income-Per capita income (PCI) measures the average income earned per person in a given area (city, region, country, etc.) in a specified year. It is calculated by dividing the area's total income by its total population.  
5. GDP: GDP provides an economic snapshot of a country/state, used to estimate the size of an economy and growth rate.

**Dataset for Part 1: PCA:**[**Hair Salon.csv**](https://olympus.mygreatlearning.com/courses/80501/files/6794274/download?wrap=1)**[View in a new window](https://olympus.mygreatlearning.com/courses/80501/files/6794274/download?wrap=1)**

**Dataset for Part 2: Clustering:**[**State\_wise\_Health\_income.csv**](https://olympus.mygreatlearning.com/courses/80501/files/6794277/download?wrap=1)

|  |
| --- |
| Part 2: Clustering: Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, etc) |
| Part 2: Clustering: Do you think scaling is necessary for clustering in this case? Justify. |
| Part 2: Clustering: Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them. |
| Part 2: Clustering: Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and find the silhouette score. |
| Part 2: Clustering: Describe cluster profiles for the clusters defined. Recommend different priority based actions that need to be taken for different clusters on the bases of their vulnerability situations according to their Economic and Health Conditions. |