**PHASE 2:**

**INNOVATION**

Performing clustering analysis to identify patterns among different industrial categories and age groups can provide valuable insights for various purposes, such as marketing strategies, customer segmentation, or policy planning. Here are the steps you can follow to conduct this analysis:

**1. Data Collection:**

Gather data that includes information about individuals or companies, their industrial categories, and age groups. Ensure that your dataset is clean and properly formatted.

**2. Data Preprocessing:**

- Handle missing data: Fill in missing values or remove incomplete records.

- Normalize or standardize the data: Ensure that all variables are on the same scale if they have different units or ranges.

- Encode categorical variables: Convert categorical variables (such as industrial categories) into numerical format using techniques like one-hot encoding.

**3. Feature Selection:**

- Decide which features (variables) you want to use for clustering. In your case, the industrial category and age group are relevant, but you may have other variables as well.

**4. Choose Clustering Algorithm:**

- Select an appropriate clustering algorithm based on your dataset and the nature of the problem. Common clustering algorithms include K-Means, Hierarchical Clustering, and DBSCAN.

**5. Determine the Number of Clusters:**

- Decide how many clusters you want to create. You can use methods like the elbow method or silhouette analysis to find the optimal number of clusters.

**6. Perform Clustering:**

- Apply the chosen clustering algorithm to your preprocessed data.

**7. Visualization:**

- Visualize the clusters to understand the patterns and relationships. You can use techniques like scatter plots, heatmaps, or dimensionality reduction methods (e.g., PCA) to visualize the data.

**8. Interpretation:**

- Analyze the clusters to gain insights into the relationships between industrial categories and age groups. Are there any significant patterns or trends? What do these clusters represent?

**9. Validation:**

- Evaluate the quality of your clusters using internal validation metrics (e.g., silhouette score) or external validation measures if you have ground-truth labels.

**10. Actionable Insights:**

- Based on your findings, derive actionable insights. For example, you might identify that certain age groups are more prevalent in specific industrial categories, which could inform targeted marketing strategies or policy decisions.

**11. Repeat and Refine (if necessary):**

- If the results are not satisfactory or you want to explore different aspects of the data, you can refine your clustering analysis by adjusting parameters, trying different algorithms, or including additional features.

**12. Reporting and Communication:**

- Present your findings in a clear and understandable way to stakeholders, whether through reports, visualizations, or presentations.Remember that the success of clustering analysis depends on the quality of your data, the choice of algorithm, and the interpretation of the results. It's essential to approach the process systematically and iteratively to gain meaningful insights from your data.

**Analyzing the Clusters**

We now know that our data can be divided into 5 clusters, and we can start to interpret them and draw insights. We can sort them like this:

* Green corresponds with an average annual income and average spending score
* Blue corresponds with a low annual income and high spending score
* Red corresponds with a high annual income and low spending score
* Purple corresponds with a high annual income and high spending score

**Python program:**

Import pandas as pd

Import numpy as np

Import matplotlib. pyplot as plt

Import seaborn as snssns.barplot(x='cluster\_pred',y='Age',palette="plasma",data=avg\_data)  
sns.barplot(x='cluster\_pred',y='Annual Income (k$)',palette="plasma",data=avg\_data)  
sns.barplot(x='cluster\_pred',y='Spending Score (1-100)',palette="plasma",data=avg\_data)

plt.figure(figsize=(10,10))  
plt.scatter(clusters\_new['Annual Income(k$)'],clusters\_new['Spending Score (1-100)'],c=clusters\_new['cluster\_pred'],cmap='rainbow')  
plt.title("Clustering customers based on Annual Income and Spending score", fontsize=15,fontweight="bold")  
plt.xlabel("Annual Income")  
plt.ylabel("Spending Score")

import pandas as pd

import numpy as np

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

data = pd.read\_csv('data.csv')

selected\_features = data[['IndustrialCategory', 'Age']]

selected\_features = selected\_features.dropna()

selected\_features = pd.get\_dummies(selected\_features, columns=['IndustrialCategory'])

scaler = StandardScaler()

scaled\_data = scaler.fit\_transform(selected\_features)

for i in range(1, 11):

kmeans = KMeans(n\_clusters=i, init='k-means++', max\_iter=300, n\_init=10, random\_state=0)

kmeans.fit(scaled\_data)

wcss.append(kmeans.inertia\_)

plt.figure(figsize=(8, 6))

plt.plot(range(1, 11), wcss, marker='o', linestyle='--')

plt.title('Elbow Method')

plt.xlabel('Number of clusters')

plt.ylabel('WCSS')

plt.show()

# Based on the elbow method, choose the optimal number of clusters (e.g., 3)

num\_clusters = 3

# Perform K-Means clustering with the chosen number of clusters

kmeans = KMeans(n\_clusters=num\_clusters, init='k-means++', max\_iter=300, n\_init=10, random\_state=0)

cluster\_labels = kmeans.fit\_predict(scaled\_data)

data['Cluster'] = cluster\_labels

plt.figure(figsize=(10, 8))

for cluster\_num in range(num\_clusters):

plt.scatter(data[data['Cluster'] == cluster\_num]['Age'],

data[data['Cluster'] == cluster\_num]['IndustrialCategory\_YourCategory'],

label=f'Cluster {cluster\_num}')

plt.title('Clustering Analysis')

plt.xlabel('Age')

plt.ylabel('Industrial Category (Your Category)')

plt.legend()

plt.show()

**OUTPUT:**

cluster\_pred Age Annual Income (k$) Spending Score (1–100)  
0 24.960000 28.040000 77.000000  
1 43.727273 55.480519 49.324675  
2 32.692308 86.538462 82.1282387  
4 45.217391 26.304348 20.913043

