SCALER - Clustering in Learner Profiling Business Case

❖ Topic: Clustering for Learner Profiling & Company Profiling
❖ Duration: 1 week

Why this case study?

From the company's perspective:

- Scaler, as an emerging tech-versity, endeavors to provide world-class education in computer science & data science domains.
- A significant challenge for Scaler is understanding the diverse backgrounds of its learners, especially in terms of their current roles, companies, and experience. Clustering similar learners helps in customizing the learning experience, thereby increasing retention and satisfaction.
- Analyzing the vast data of learners can uncover patterns in their professional backgrounds and preferences. This allows Scaler to make tailored content recommendations and provide specialized mentorship.
- By leveraging data science and unsupervised learning, particularly clustering techniques, Scaler can group learners with similar profiles, aiding in delivering a more personalized learning journey.

From the learner's perspective:

- This case presents a valuable opportunity to delve deep into how online tech-versities like Scaler operate and the challenges they encounter in learner profiling.

- Clustering, especially methods like K-means and Hierarchical Clustering, can reveal hidden patterns and structures within data that are not immediately obvious.

- Through this exercise, participants refine their expertise in exploratory data analysis, feature engineering, data pre-processing, and unsupervised learning.

- Furthermore, learners get a chance to navigate real-world business scenarios, translating raw datasets into valuable insights that can steer organizational strategies.

Dataset Explanation: scaler_kmeans.csv

1. Unnamed 0: The index of the dataset.
2. Email_hash: An anonymized identifier representing the email of the learner.
3. Company_hash: An anonymized identifier indicating the current employer of the learner.
4. orgyear: Represents the year the learner began employment at the current company.
5. CTC: Current Compensation to the Company (CTC) of the learner.
6. Job_position: Represents the job profile or role of the learner within their company.
7. CTC_updated_year: The year in which the learner's CTC was most recently updated. This could be due to yearly increments, promotions, or other factors.

What is Expected?

Assuming you're a data scientist at Scaler, you're tasked with the responsibility of analyzing the dataset to profile the best companies and job positions from Scaler's database. Your primary goal is to execute clustering techniques, evaluate the

coherence of your clusters, and provide actionable insights for enhanced learner profiling and course tailoring.

Submission Process:

- Upon wrapping up the case study...
- Document your insights, methodologies, and outcomes in a Jupyter Notebook.
- Within your notebook, make sure you:
    - Demonstrate the Python code for every data processing, feature engineering, clustering method, and cluster analysis.
    - Incorporate visualizations, such as bar charts, dendrograms, Elbow plots, silhouette plots, and more, to reinforce your analysis.
    - Culminate with profound insights extracted from the dataset and propose actionable recommendations for Scaler to refine its course offerings and learner support.
- Convert your Jupyter Notebook into a PDF (Leverage the Chrome browser's Print function for this).
- Stick to the submission procedures and upload your PDF on the assigned platform.
- Be mindful that post your submission, there's no facility to revisit or revise your work.

General Guidelines:

This exercise mirrors genuine challenges, capturing the essence of tasks that data scientists often undertake. Seize this moment to delve deep and mimic an actual professional scenario.

During your exploration, you might encounter difficulties or even feel overwhelmed:

- Periodically review the problem statement to ensure you're on track with the objectives.
- Break down complex tasks into more straightforward, digestible segments.
- When confronted with coding dilemmas or challenges, resort to online communities or official documentation. Developing a problem-solving mindset is crucial for data scientists.
- Engage with your peers. Participating in discussion forums might provide varied viewpoints, helping you navigate challenges or sparking innovative approaches.
- Re-engage with lessons or seek external sources for concepts that seem elusive.
- For any overarching concerns or if any part of the problem statement seems

unclear, reach out to your Instructor without hesitation.

Remember, every obstacle is an avenue for personal growth. Tackle this case with enthusiasm, commitment, and an adaptable mindset.

What does 'good' look like?

1. Define Problem Statement and perform Exploratory Data Analysis

	Hint	Approach
a. Definition of problem	Start by honing in on the main goal. What does Scaler aim to achieve? Why is understanding a learner's preferences or career ambitions vital?	The primary objective is to profile learners effectively to tailor course content, ensuring a more personalized and impactful learning experience.
b. Observations on Data	Having a grip on the dataset's layout is fundamental. Examine the data's dimensions, the data types of every variable, transformation of categorical variables to 'category' (if necessary), spotting missing values, and a brief statistical overview.	a. Utilize commands such as data.info(), data.describe(), and data.shape in Python. b. Distinguish between numerical and categorical variables. If required, modify categorical data types employing astype('category').
c. Univariate Analysis	Embark with single variables. For numerical attributes, lean towards distribution graphs, and for categorical variables, consider bar or frequency plots.	For numerical attributes, histograms or density charts are apt. In the case of categorical attributes, frequency plots are suitable. Libraries like Seaborn simplify the creation of such visualizations. This is instrumental in deciphering the dispersion of individual variables.
d. Bivariate Analysis	Dive into relationships between two variables. (Relationships between important variable)	Employ scatter plots for continuous-continuous relationships, boxplots for categorical-continuous

		correlations, and crosstab or stacked bar plots for categorical-categorical correlations. For instance, understanding how a driver's ratings correlate with the frequency of rides can offer key insights.
e. Illustrate the insights based on EDA	Every graph and table should deliver an insight	Take notes on surprising distributions, high correlations, or peculiar behaviors seen in the bivariate analysis.
f. Comments on range of attributes, outliers of various attributes	Outliers and ranges of attributes can significantly influence the subsequent clustering process.	Box plots or IQR calculations can help detect outliers. Understand if these outliers represent genuine high or low earners or if they're data errors. A comment on the spread of CTC or years of experience can offer context on the overall dataset's composition.
g. Identify normal vs skewed distributions and understand why.	For numerical attributes, comment on the skewness or symmetry of the distribution..	When analyzing relationships, offer insights into observed correlations or patterns, such as if individuals in specific job roles consistently earn more irrespective of their company.
h. Comments for each univariate and bivariate plots	Just plotting isn't enough, explain them.	Mere visualization isn't sufficient; accompanying narratives are essential. Each chart or plot should be paired with a brief commentary. For example, "The histogram reveals that a majority of the learners fall within the 6-10 lakh CTC range. However, a scatter plot of CTC against years of experience demonstrates a clear trend of increasing salary with experience."

2. Data Preprocessing

	Hint	**Approach**
a. Duplicate value check	Given the nature of the dataset, there might be duplicate entries	Start by examining the dataset for any repeated patterns based on attributes like

	for learners.	'Email_hash' or 'Company_hash'. Consider removing the duplicates based on a combination of features to ensure the uniqueness of each learner's data.
b. Missing value treatment	Crucial step since incomplete data can introduce bias.	a. Begin by detecting columns that have missing values. b. Strategies can include imputation using central tendencies, imputation with advanced techniques like KNN, or even deletion based on the nature of the data. c. Pay special attention to columns like 'CTC', 'Job_position', and 'orgyear' since they play vital roles in clustering.
c. Outlier treatment	Given that salary (CTC) might have extreme values, detecting and handling outliers is essential.	a. Visualize 'CTC' and other continuous variables to identify outliers. b. Techniques can be used such as capping based on domain knowledge, transformations, or even removal if justified.
d. Feature engineering	With the current set of features, you have the opportunity to extract more insights..	a. 'orgyear': Use it to compute 'Years of Experience' by subtracting from the current year. This can provide better insights than just the starting year of employment. b. Extract specific features or flags from 'Job_position' to highlight prominent roles. c. From 'CTC_updated_year', create a flag indicating if a person got an increment or promotion this year. d. Consider categorizing 'CTC' into bins like 'Low', 'Average', 'High' for a more generalized view. e. Feature Mining from Aggregated Data: Derive new features from aggregated

		data, like average CTC per company, role, or years of experience.
e. Data preparation for modeling	Once the dataset is clean and enhanced, the next step is to prepare it for clustering.	a. Depending on the model's requirements, consider scaling the features. The choice of scaling technique would vary based on data distribution and model sensitivity. b. Different encoding techniques are suitable for different types of categorical variables: I. Label Encoding: Use for ordinal categories with a natural order (e.g., Low, Medium, High). II. One Hot Encoding: Best for nominal categories without inherent order. c. After feature engineering, evaluate the distributions of newly created features. Examine skewness, kurtosis, and other statistical measures. Any transformations or treatments should be justified with appropriate logic.

3. Model building

	Hint	**Approach**
a. Data Splitting	Given that clustering is an unsupervised learning technique, there's no 'target' variable to split on.	However, if you have a notion of 'good' clusters based on domain knowledge, you can reserve part of the dataset to evaluate the clustering's effectiveness. A typical 80-20 split can still be employed to have an 'unseen' dataset to evaluate clustering robustness.
b. Checking Clustering Tendency	Before performing clustering, it's useful to determine if the data has a natural tendency to form clusters.	Use the Hopkins Statistic: It gives a value between 0 and 1, where a value around 0.5 suggests no significant clustering structure, and values deviating from 0.5 suggest a clustering tendency.

c. Selecting Optimal Number of Clusters	Before applying clustering, decide on the optimal number of clusters.	a. Use the Elbow Method: Plot inertia (within-cluster sum of squares) against a range of cluster numbers. The 'elbow' of the curve indicates an optimal number .
d. K-means Clustering	K-means is a widely-used clustering technique.	a. Use KMeans from sklearn.cluster. b. Set the number of clusters based on prior analysis. c. Train the model on the training set and label the validation set.
e. Hierarchical Clustering	A method that builds a hierarchy or tree of clusters.	a. Use 'AgglomerativeClustering' from 'sklearn.cluster'. b. Visualize using a dendrogram to help decide the optimal number of clusters. c. Train on the dataset.
f. Evaluating K-means Clustering	While many standard metrics like silhouette score exist, let's look at more general methods:	a. Within-Cluster Sum of Squares (WCSS): It's the inertia value, and it should be relatively low, indicating tight clusters. b. Between-Cluster Sum of Squares (BCSS): This should be high, indicating well-separated clusters. c. Visual Inspection: Depending on the data dimensionality, visualizing clusters using scatter plots or similar can be useful. Dimensionality reduction techniques like PCA can be employed for higher-dimensional data. d. Cluster Sizes: Ensure that there's a reasonable distribution of data points among the clusters. Highly imbalanced clusters might suggest that the data hasn't been segmented meaningfully. e. Domain Validation: If available, use domain-specific knowledge to verify if the clusters make intuitive sense.

4. Results Interpretation & Stakeholder Presentation

	Hint	Approach
a. Understand the Business Context	While clustering doesn't provide explicit "performance" metrics like supervised learning, it reveals data structures and segmentations that might be invaluable.	a. Determine the main goals behind clustering: Is the aim to segment customers for targeted marketing? Understand user behavior? Find anomalies? b. Understand the value of each cluster: Does each cluster represent a unique user group with distinct behaviors or needs?
b. Insights from Unsupervised Clustering	Analyze each cluster's characteristics to derive insights.	a. Profile each cluster: What commonalities exist within members of a cluster? Are they high-value customers, frequent shoppers, or occasional visitors? b. Examine the central tendencies (mean/median) of features within each cluster. This helps in understanding the dominant characteristics of each group.
c. Visual Representations	Visuals are essential in making cluster insights comprehensible.	a. Use bar plots or pie charts to showcase the size and distribution of each cluster. b. Consider scatter plots (possibly using PCA or t-SNE for dimensionality reduction) to visually demonstrate how clusters are separated from each other.

d. Trade-off Analysis	There might be trade-offs associated with catering to or targeting specific clusters.	a. Evaluate the cost of targeting a high-value cluster against the potential ROI. b. Weigh the benefits of creating tailored experiences for a niche cluster versus a generalized approach for all.
e. Actionable Insights & Recommendations	Insights are valuable only when they lead to actionable steps.	a. If a cluster represents high-value but infrequent shoppers, recommend strategies to increase their purchase frequency. b. For a cluster indicating potential churn, suggest retention strategies or loyalty programs. c. If a particular cluster seems to favor specific products or services, advocate for targeted marketing or offers for those preferences.
f. Feedback Loop	Clustering should be a periodic activity, especially with evolving user behavior and business dynamics.	a. Suggest a regular re-run of the clustering process to ensure that the segments are still valid. b. Recommend channels to continuously collect data on customer feedback, preferences, and behaviors to fine-tune future clustering exercises.

Questionnaire (Answers should present in the text editor along with insights):

1. What percentage of users fall into the largest cluster?
2. Comment on the characteristics that differentiate the primary clusters from each other.
3. Is it always true that with an increase in years of experience, the CTC increases? Provide a case where this isn't true.
4. Name a job position that is commonly considered entry-level but has a few learners with unusually high CTCs in the dataset.
5. What is the average CTC of learners across different job positions?
6. For a given company, how does the average CTC of a Data Scientist compare with other roles?
7. Discuss the distribution of learners based on the Tier flag:
    1. Which companies dominate in Tier 1 and why might this be the case?
    2. Are there any notable patterns or insights when comparing learners from Tier 3 across different companies?
8. After performing unsupervised clustering:
    1. How many clusters have been identified using the Elbow method?
    2. Do the clusters formed align or differ significantly from the manual clustering efforts? If so, in what way?
9. From the Hierarchical Clustering results:
    1. Are there any clear hierarchies or patterns formed that could suggest the different levels of seniority or roles within a company?
    2. How does the dendrogram representation correlate with the 'Years of Experience' feature?