Project Title:
Customer Segmentation using data science

Phase 2: Innovation

1.Data Preprocessing:

- Before applying dimensionality reduction, it's essential to preprocess your data. This typically includes handling missing values, scaling features, and encoding categorical variables.

2.Standardization:

- Standardize the features if they are on different scales. PCA is sensitive to the scale of the data, so it's crucial to have all features with a similar range.

3.PCA (Principal Component Analysis):

- PCA is a linear dimensionality reduction technique that identifies the principal components (linear combinations of the original features) that explain the most variance in the data. Follow these steps for PCA: a. Compute the covariance matrix of your standardized data. b. Calculate the eigenvalues and eigenvectors of the covariance matrix. c. Sort the eigenvalues in descending order and select the top k eigenvectors corresponding to the highest eigenvalues, where k is the desired number of dimensions. d. Project your data onto the k-dimensional subspace defined by the selected eigenvectors.

4.t-SNE (t-Distributed Stochastic Neighbor Embedding):

- t-SNE is a non-linear dimensionality reduction technique that focuses on preserving the pairwise similarities between data points. It is often used for visualization purposes. Here's how to use t-SNE: a. Compute the pairwise similarities (typically using a Gaussian kernel) between data points in the high-dimensional space. b. Create a probability distribution over pairwise similarities. c. Create a similar probability distribution in a lower-dimensional space. d. Use gradient descent to minimize the Kullback-Leibler divergence between the two probability distributions, optimizing the embedding in the lower-dimensional space.

5.Visualization:

- After applying PCA or t-SNE, you'll have a reduced-dimensional representation of your data. You can now visualize this lower-dimensional data to discover patterns and insights.
- Use scatter plots, heatmaps, or other visualization techniques to explore and interpret the data in its reduced form.

6.Interpretation:

- Pay attention to the clustering or grouping of data points in the visualization. Clusters may indicate customer segments or patterns that were not apparent in the original high-dimensional data.
- Examine the loadings of the original features on the principal components (for PCA) or the pairwise similarities (for t-SNE) to understand which features are contributing the most to the observed patterns.