EXPLAINATION OF STEP-4:

In data analysis, exploratory data analysis (EDA) is a crucial step that aids in understanding the underlying structure of the data and locating patterns, linkages, and potential discoveries. When used with K-means clustering, EDA aids in data preparation for clustering, selecting the right K number of clusters, and gaining understanding of the generated clusters. I'll go over the main procedures and methods used in EDA for K-means clustering in this essay.

1: Data Collection and Pre-processing

Before diving into EDA for K-means clustering, you first need to collect your data and pre-process it. Data pre-processing involves tasks like handling missing values, scaling features, and encoding categorical variables. The quality of pre-processing can significantly impact the success of the clustering analysis.

2: Univariate Analysis

EDA often begins with univariate analysis, where you examine individual features one by one.

1. Summary Statistics: Calculate basic statistics such as mean, median, standard deviation, minimum, and maximum for each feature. This helps you understand the central tendency and spread of the data.
2. Histograms: Plot histograms to visualize the distribution of each feature. Histograms provide insights into the data's skewness and multimodality.
3. Box Plots: Box plots display the distribution's quartiles, helping identify outliers and the presence of skewness.
4. Kernel Density Estimation (KDE): KDE plots provide smoothed representations of feature distributions and can help identify underlying patterns.

By examining these statistics and plots, you can identify potential outliers, assess feature scaling requirements, and detect any data anomalies.

3: Bivariate Analysis

After analysing individual features, it's essential to explore relationships between pairs of features. Bivariate analysis can uncover correlations and dependencies that might influence the clustering results. Bivariate analysis can uncover correlations and dependencies that might influence the clustering results.

1. Scatter Plots: Create scatter plots to visualize the relationship between pairs of continuous features. Patterns, clusters, or trends may emerge.
2. Correlation Matrix: Calculate the correlation coefficients between all pairs of continuous features. A correlation matrix helps identify highly correlated or collinear features.
3. Categorical Feature Analysis: For categorical features, create contingency tables or bar plots to explore relationships and dependencies.

Bivariate analysis allows you to make informed decisions about feature selection, identify potentially redundant features, and assess which features might contribute most to clustering.

4: Principal Component Analysis

In cases where you have a high-dimensional dataset, dimensionality reduction techniques like Principal Component Analysis (PCA) can be valuable. These techniques help reduce the number of features while retaining essential information. Visualizing the data in lower dimensions can reveal cluster structures more effectively.

5: Elbow/ Scree plot

Choosing the right number of clusters (K) is a critical step in K-means clustering. The Elbow Method is a common EDA technique for determining K. It involves running K-means clustering for a range of K values and plotting the within-cluster sum of squares (WCSS) or variance explained. The point where the WCSS starts to level off (resembling an "elbow" in the plot) is often considered the optimal K. However, the choice of K can be somewhat subjective and may require domain knowledge.

6: Visualizing clusters

Once you've chosen K and performed K-means clustering on your data, it's time to visualize the clusters.

1. Scatter Plots: Create scatter plots where each data point is coloured or marked according to its cluster assignment. This provides an intuitive view of cluster separation.
2. Centroid Plots: If your features are continuous, you can plot the cluster centroids to understand their location in feature space.
3. Cluster Profiles: Analyse cluster profiles by calculating the mean or median values of features within each cluster. This helps interpret the characteristics of each cluster.
4. Dimensionality Reduction: Apply dimensionality reduction techniques like PCA or t-SNE to visualize the data and clusters in lower dimensions.

7: Cluster Validation Metrics

To quantitatively assess the quality of your clusters, you can use various cluster validation metrics, such as:

Inertia: measures the sum of squared distances from each point to its assigned cluster's centroid. Lower inertia indicates tighter clusters.
These metrics provide objective measures of clustering quality and can help you fine-tune your clustering approach.