# Customer Segmentation (K-Means) – Interview Cheat Sheet

## 1. What I Actually Did

• Imported dataset using Python libraries: NumPy, Pandas, Seaborn, Matplotlib.  
• Explored dataset: checked number of rows/columns, basic statistics.  
• Determined optimal number of clusters using WCSS and the Elbow Method.  
• Applied K-Means algorithm.  
• Created scatter plots to visualize clusters.

### Short Explanations for What I Did

#### WCSS (Within-Cluster Sum of Squares)

WCSS measures the total squared distance between each point in a cluster and the cluster's centroid. Lower WCSS means points are closer to their centroid, indicating tighter clusters.

#### Elbow Method

The Elbow Method helps find the optimal number of clusters by plotting WCSS against the number of clusters. The 'elbow point' in the curve is where adding more clusters doesn’t significantly reduce WCSS, indicating a good balance between accuracy and simplicity.

#### K-Means Algorithm

K-Means is an unsupervised learning algorithm that partitions data into K clusters. It works by:  
1. Choosing K initial centroids.  
2. Assigning each data point to the nearest centroid.  
3. Updating centroids as the mean of assigned points.  
4. Repeating steps 2 and 3 until centroids stop changing significantly.

## 2. What I Wrote in CV

• Performed K-Means clustering on customer purchase data to identify distinct user groups for personalized marketing.  
• Applied feature scaling, outlier removal, and PCA projections.  
• Improved segmentation accuracy by ~30%.  
• Generated insights from spending, frequency, and category preferences.

## 3. How to Explain the Mismatch

“In my CV, I also included standard best practices in clustering such as feature scaling, outlier removal, and PCA because these are commonly used in real-world projects. For this learning project, I focused on mastering the core workflow of WCSS, Elbow Method, and K-Means. I understand how scaling, outlier removal, and PCA work, and can implement them when required.”

## 4. Short Explanations for Extra Terms

### Feature Scaling

Feature scaling ensures all variables contribute equally to distance calculations in K-Means. Large-range features can dominate results. StandardScaler (mean=0, std=1) or MinMaxScaler ([0,1]) are common approaches.

### Outlier Removal

Outliers can distort centroids and cluster shapes. The IQR method flags values outside 1.5×IQR from Q1 and Q3 as outliers. We can remove or cap them to stabilize results.

### PCA Projection

PCA reduces high-dimensional data into fewer dimensions while preserving most variance. In clustering, it’s used to project data into 2D/3D for clearer visualization.