Welcome to the CoGrammar K-means Clustering

The session will start shortly...

Questions? Drop them in the chat. We'll have dedicated moderators answering questions.

CoGrammar

Data Science Session Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. Moderators are going to be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: **Questions**

Data Science Session Housekeeping cont.

- For all **non-academic questions**, please submit a query: **www.hyperiondev.com/support**

- Report a **safeguarding** incident: **www.hyperiondev.com/safeguardreporting**

- We would love your **feedback** on lectures: **Feedback on Lectures**

Skills Bootcamp
8-Week Progression Overview

Fulfil 4 Criteria to Graduation

✅ Criterion 1: Initial Requirements

Timeframe: First 2 Weeks
Guided Learning Hours (GLH): Minimum of 15 hours
Task Completion: First four tasks

Due Date: 24 March 2024

✅ Criterion 2: Mid-Course Progress

60 Guided Learning Hours

Data Science - **13 tasks**
Software Engineering - **13 tasks**
Web Development - **13 tasks**

Due Date: 28 April 2024

CoGrammar

Skills Bootcamp Progression Overview

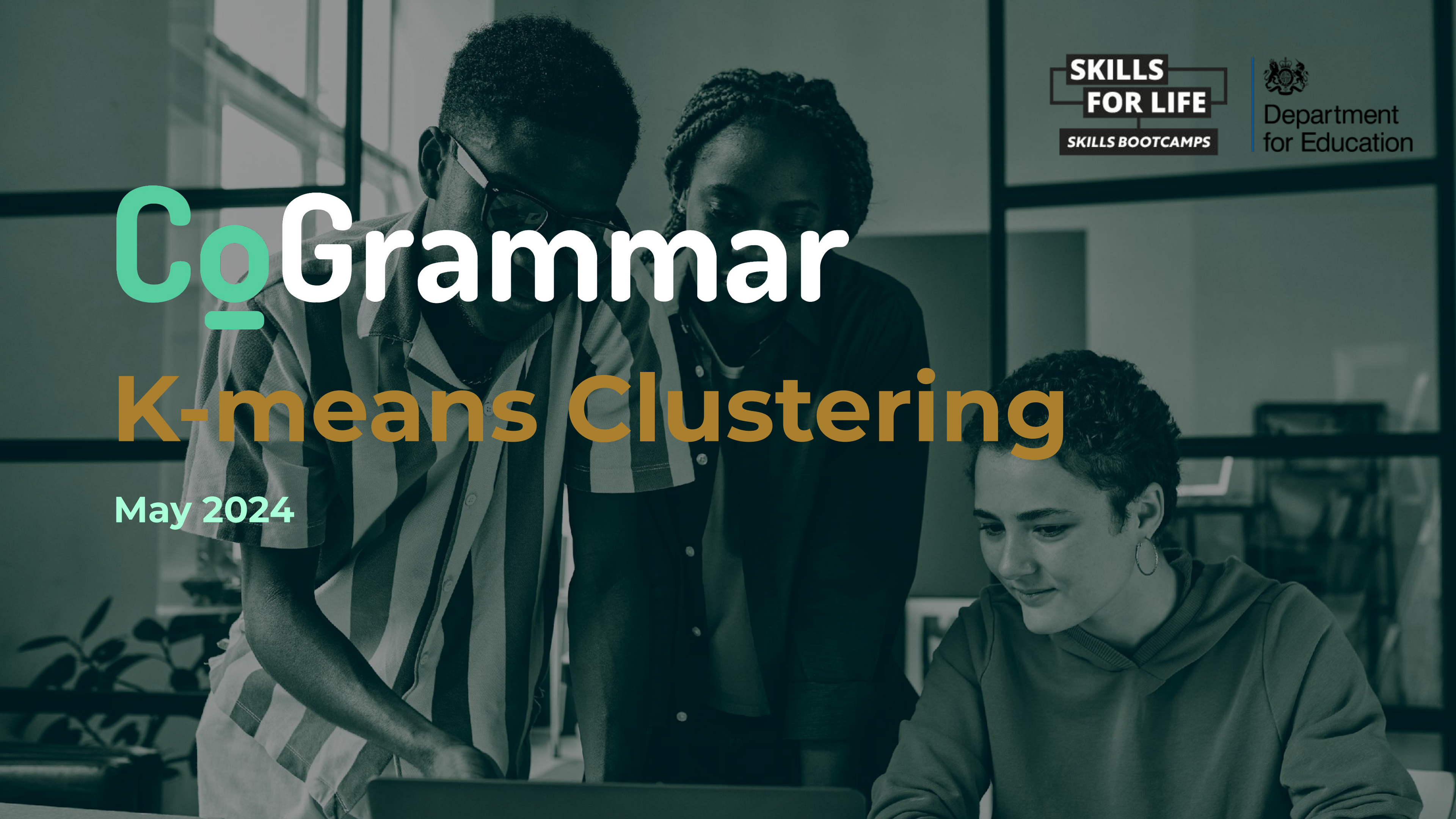
✅ Criterion 3: Course Progress

Completion: All mandatory tasks, including Build Your Brand and resubmissions by study period end
Interview Invitation: Within 4 weeks post-course
Guided Learning Hours: Minimum of 112 hours by support end date (10.5 hours average, each week)

✅ Criterion 4: Demonstrating Employability

Final Job or Apprenticeship Outcome: Document within 12 weeks post-graduation
Relevance: Progression to employment or related opportunity

CoGrammar



SKILLS FOR LIFE
SKILLS BOOTCAMPS
Department for Education

CoGrammar

K-means Clustering

May 2024

Learning Objectives

- **Understand** the concept and goal of K-means clustering
- Identify common **applications** of K-means clustering
- Describe the steps of the **K-means clustering algorithm**
- Explain techniques for **choosing the optimal number of clusters (K)**

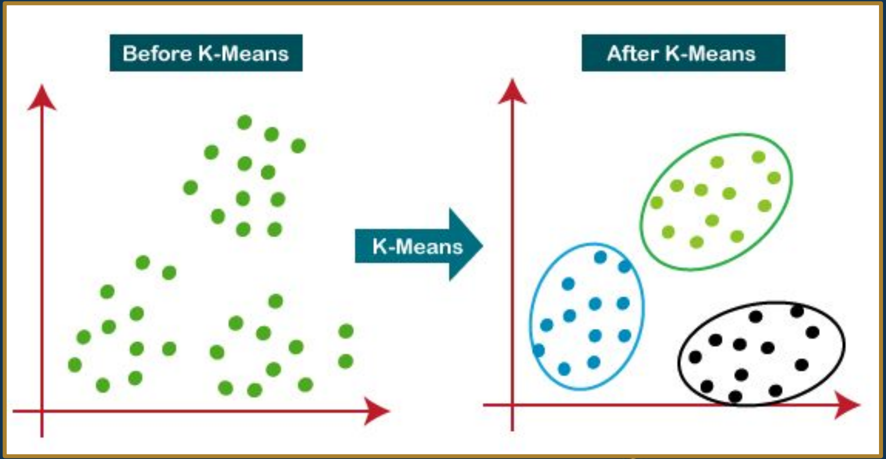
Learning Objectives

- **Implement** K-means clustering in Python using scikit-learn
- **Visualise and interpret the results** of K-means clustering

CoGrammar

Introduction

- ❖ K-means clustering is an **unsupervised learning algorithm used to partition a dataset into K distinct clusters.**
- ❖ The goal is to **minimise the within-cluster variation** and **maximise the separation between clusters**.
- ❖ It is a popular technique for **exploratory data analysis** and **pattern discovery**.



Source: Javapoint

CoGrammar

Applications of K-means Clustering

CoGrammar

Some Applications

- **Customer Segmentation:** Grouping customers based on their purchasing behaviour, demographics, or preferences.
- **Image Compression:** Reducing the colour palette of an image by clustering similar colours.
- **Anomaly Detection:** Identifying unusual or anomalous data points that do not belong to any cluster.
- **Document Clustering:** Grouping similar documents based on their content or topics.

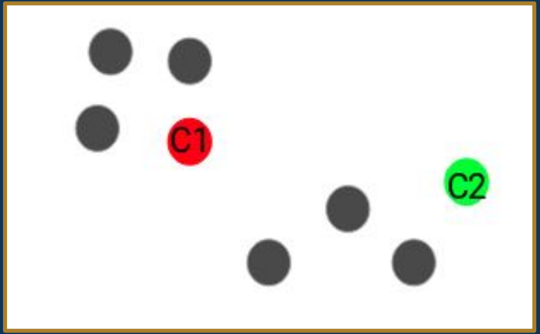
CoGrammar

K-means Clustering Algorithm

CoGrammar

Step 1

- Initialise **K centroids randomly** or using a specific strategy (e.g., K-means++).

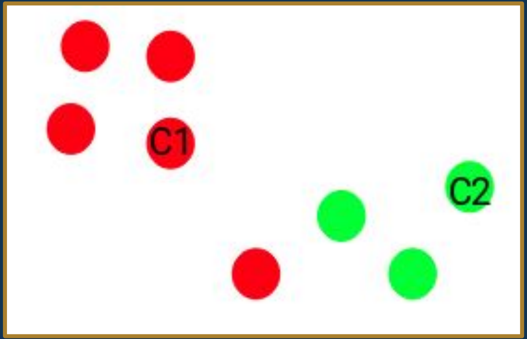


Source: Analytics Vidhya

CoGrammar

Step 2

- Assign each data point to the **nearest centroid** based on a distance metric (e.g., Euclidean distance).

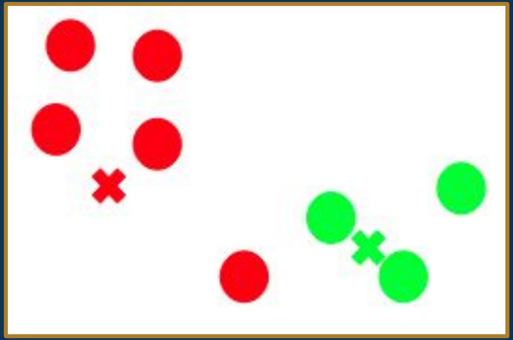


Source: Analytics Vidhya

CoGrammar

Step 3

- Update the centroids by calculating the **mean of the data points assigned to each cluster**.

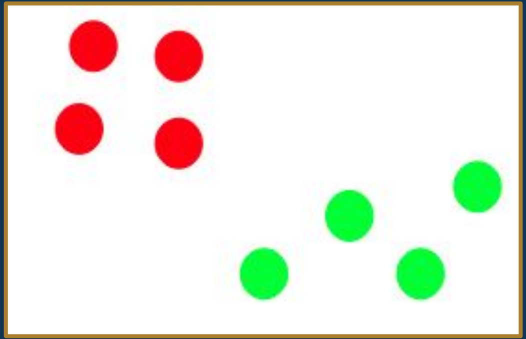


Source: Analytics Vidhya

CoGrammar

Step 4

- Repeat steps 2 and 3 **until convergence (centroids no longer change)** or a **maximum number of iterations** is reached.



Source: Analytics Vidhya

CoGrammar

Goals

- The algorithm aims to minimise the **sum of squared distances** between data points and their assigned centroids.

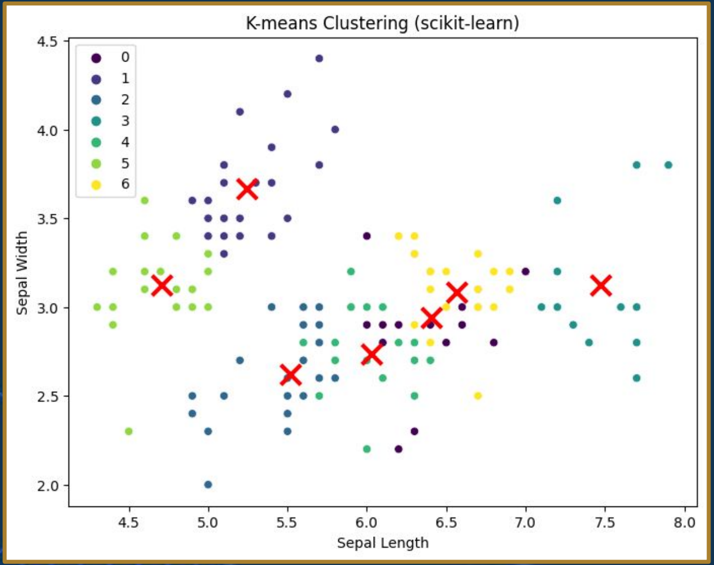
CoGrammar

Choosing the Number of Clusters (K)

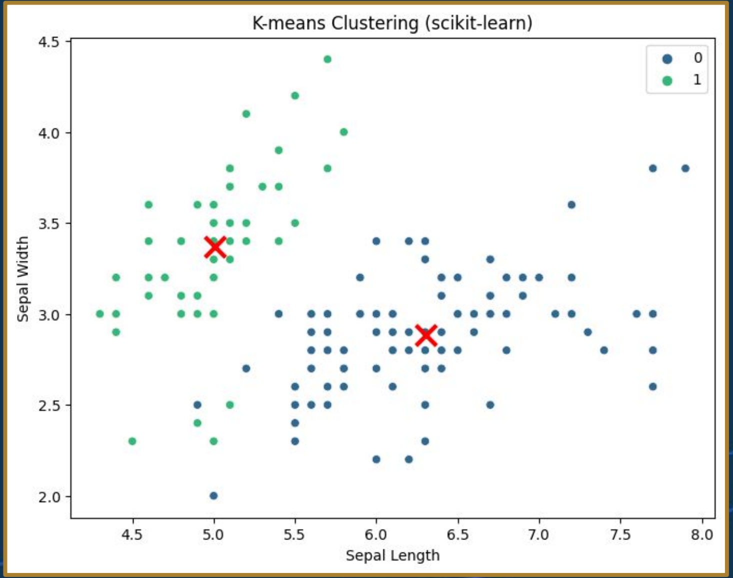
CoGrammar

Goals

- Selecting an appropriate number of clusters (K) is crucial for effective clustering.

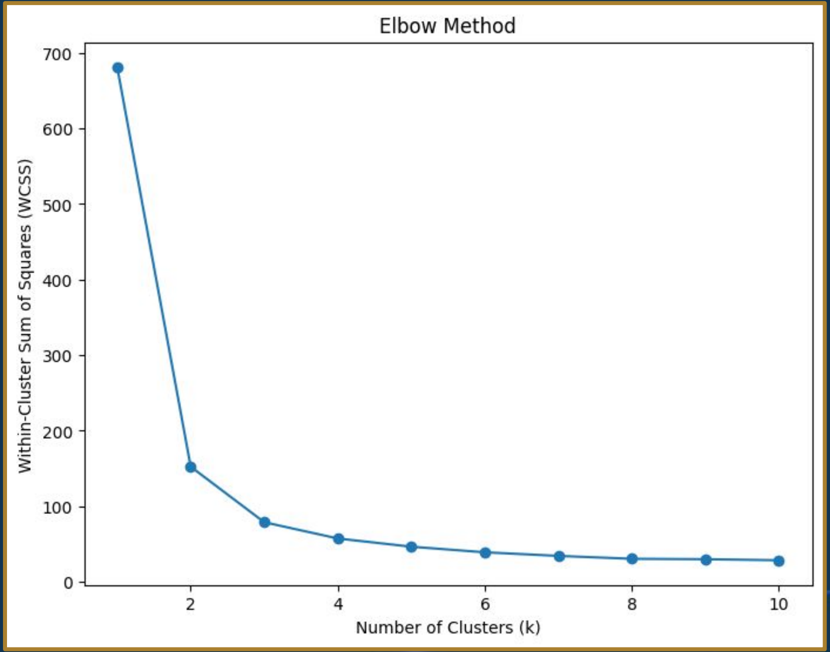


VS



Techniques

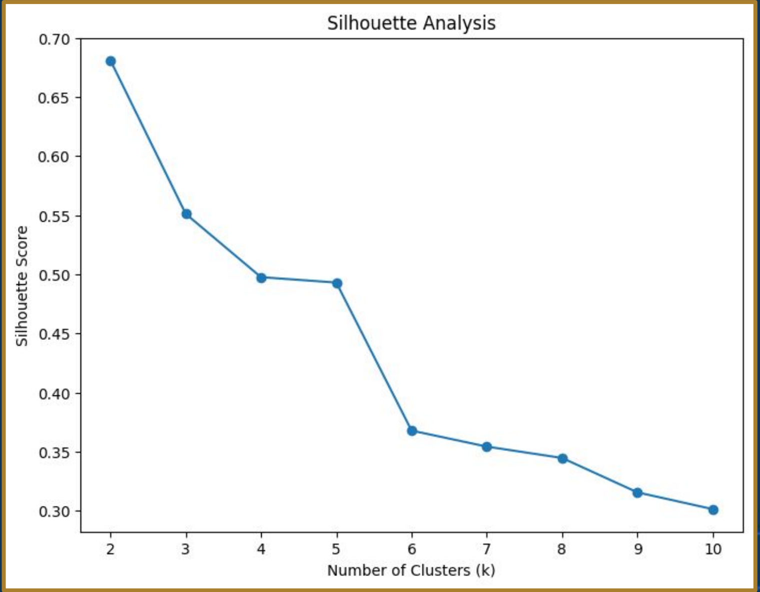
- Elbow Method:
  - Plot the within-cluster sum of squares (WCSS) against different values of K.
  - Look for the **"elbow point"** where the rate of decrease in WCSS slows down significantly.



CoGrammar

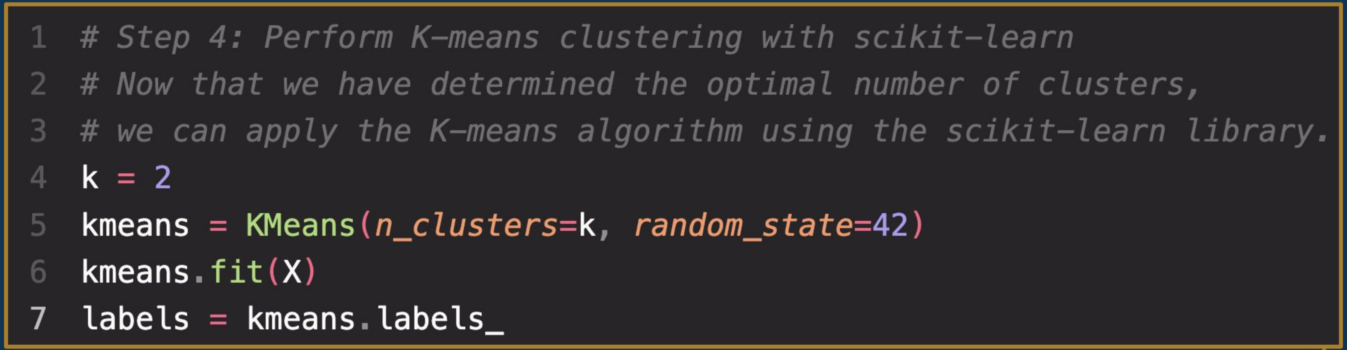
Techniques

- Silhouette Analysis:
  - Measure the quality of clustering based on the compactness and separation of clusters.
  - Calculate the silhouette score for each data point and average them for different values of K.
  - Choose the K value that **maximises the average silhouette score.**



CoGrammar

Implementation

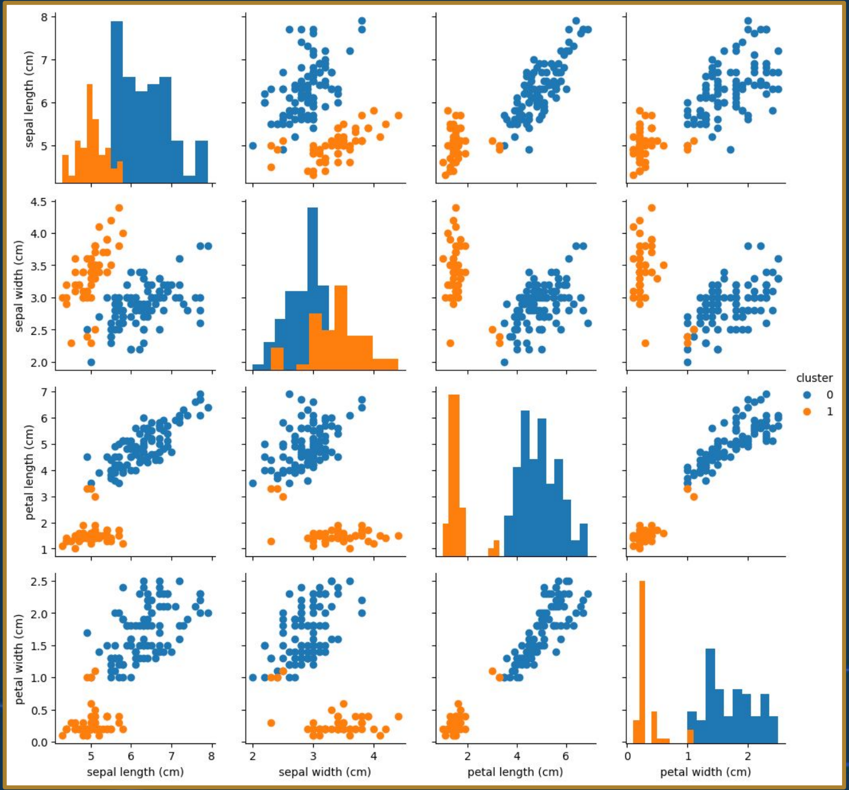


```python
# Step 4: Perform K-means clustering with scikit-learn
# Now that we have determined the optimal number of clusters,
# we can apply the K-means algorithm using the scikit-learn library.
k = 2
kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(X)
labels = kmeans.labels_
```

Visualising Cluster Results

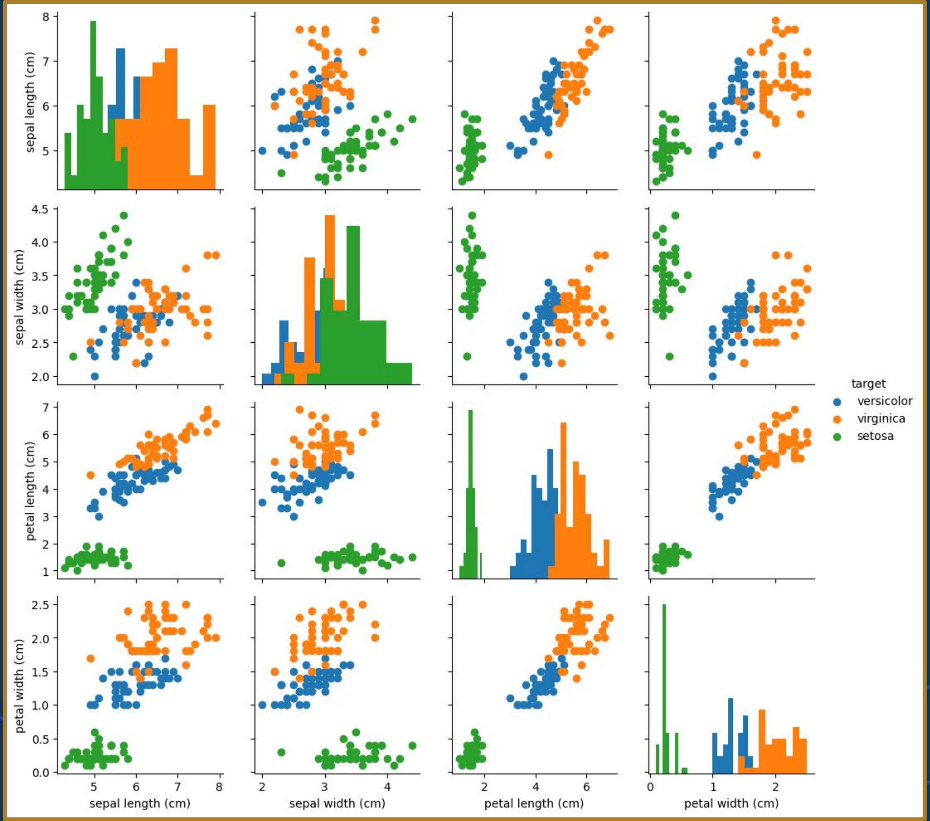
CoGrammar

Calculated Clusters



CoGrammar

Actual Species



CoGrammar

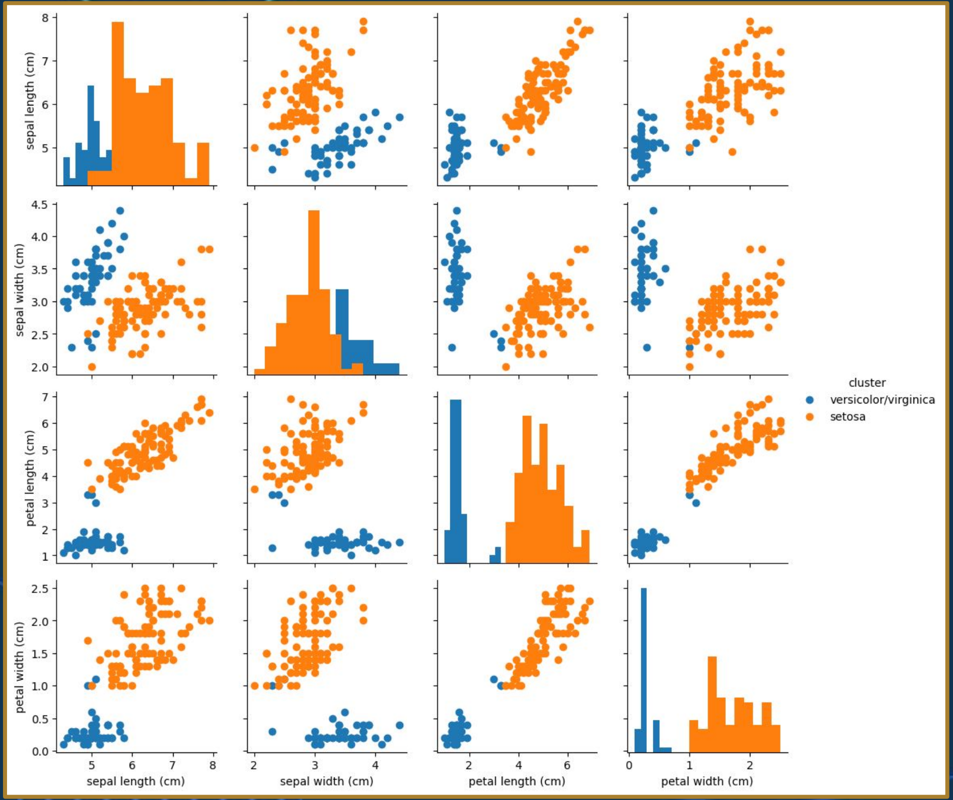
Interpreting Clusters

CoGrammar

Importance

- After obtaining the clustering results, it's important to **interpret and analyse the clusters.**
- Examine the characteristics and centroids of each cluster to **understand their distinguishing features.**
- **Assign meaningful labels or descriptions to the clusters based on domain knowledge and the patterns observed in the data.**

Interpretation



It is easy to distinguish between setosa and the other two species, which could guide the type of model you build later on. It also indicates that differentiating between versicolor and virginica would be more difficult, although could be made easier when choosing petal metrics instead of sepal metrics given the separation is more apparent.

Limitations and Considerations

CoGrammar

Limitations

- The algorithm is **sensitive to the initial positions of the centroids**, which can lead to different results in each run.
- It assumes that **clusters are spherical and have equal sizes, which may not always be the case in real-world data**.
- **Outliers and noise** in the data can affect the clustering results.
- **Scaling and normalisation of features** may be necessary to ensure equal contribution to the distance calculations.
- Handling **high-dimensional data** can be challenging due to the curse of dimensionality.

CoGrammar

Questions and Answers

CoGrammar

Thank you for attending

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar