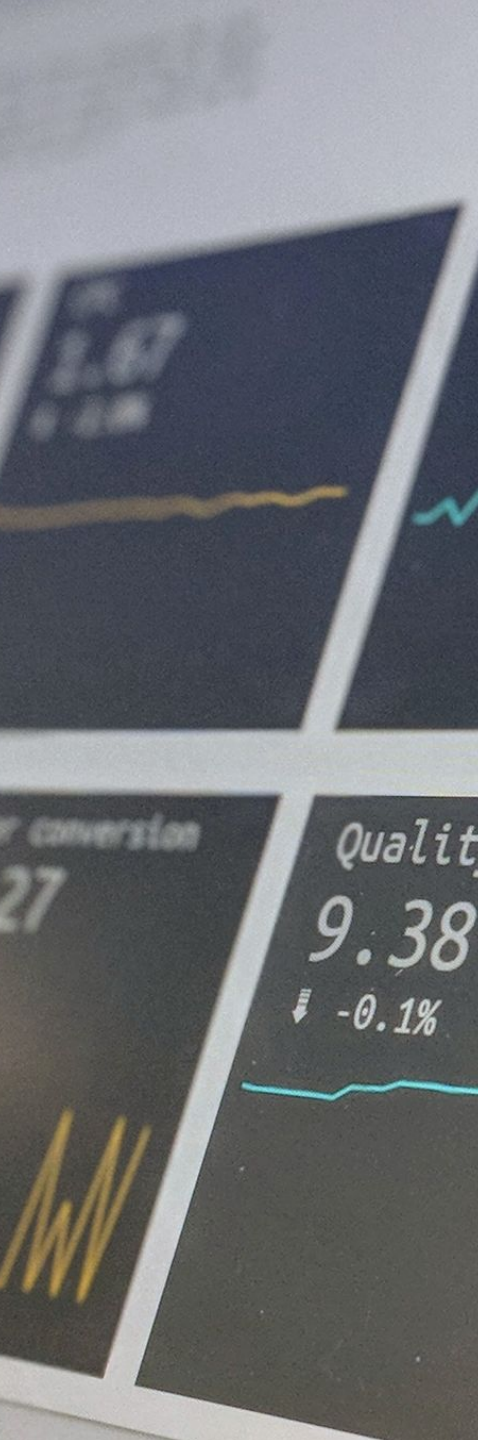
“Explicable unsupervised clustering with mixed data and small data“

Inria
UNIVERSITÉ CÔTE D'AZUR | 3iA Côte d'Azur Interdisciplinary Institute for Artificial Intelligence

*Pierpaolo **Goffredo**, Nicholas **Halliwell**, Alexandra **Würth**, Xuchun **Zhang***



Overall Tasks

- Construct homogeneous cluster
- Find best algorithm
- Implement replicable and automatic process
- Explain cluster(s) behavior

Pipeline

Data Overview

- **Correlation** matrix
- **Histograms**

➡ **Redundant** features

Data Clustering

- **K-Means** Clustering
- **Silhouette-Score**

➡ Optimal number of clusters

Data Normalization

➡ Normalize data between *0* and *1*

Cluster Analysis

- **PCA** transformation
- **Visualization** of data

➡ Find **representative** features

K-Means Clustering

- Choose **number** of clusters **K**
- **Randomly** initialize K **centroids**
- **Pseudo code**

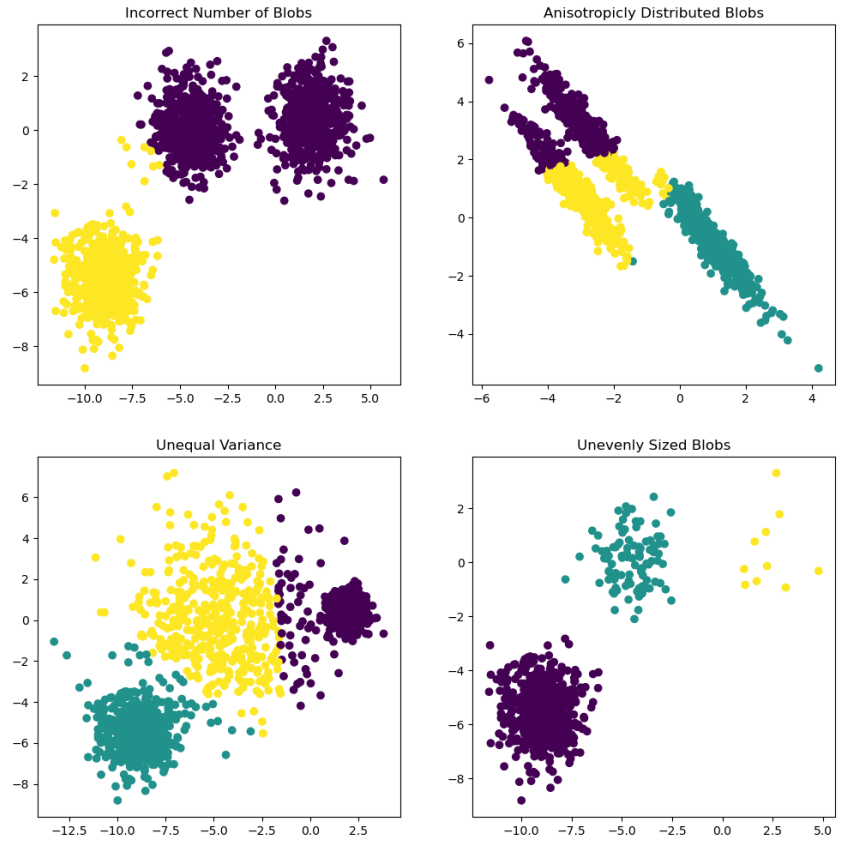
  While current_iter != **max_iter**:

  Assign each point to the nearest centroid

  Recompute centroids

  If labels are **unchanged**

    Terminate



UNSUPERVSIED LEARNING ALGORITHM

Principal Component Analysis (PCA)

1. Reduces **dimensionality** of data while preserving as much information as possible
2. Compute **covariance** matrix of input data
3. Compute **eigenvectors** and **eigenvalues** of covariance matrix
4. Select **p** highest eigenvalues
5. Use p corresponding eigenvectors of covariance matrix to project data to **p dimensions** capturing the most variance

Silhouette Score

- **Measures cluster density for unsupervised data**
- **Value between -1 and 1**
  - **1**: well separated clusters
  - **0**: overlapping clusters
  - **-1**: wrong clusters assignment
- **S = (B-A) / max(A,B)**

  A = mean distance between the observation and points in same cluster

  B = mean distance between the observation and all points in the next nearest cluster

Dataset Overview

	# Observations	# Features
Banking Dataset	1,912	18
Marketing Dataset	71,141	36
Supply Chain Dataset	740	11

Banking

Dataset

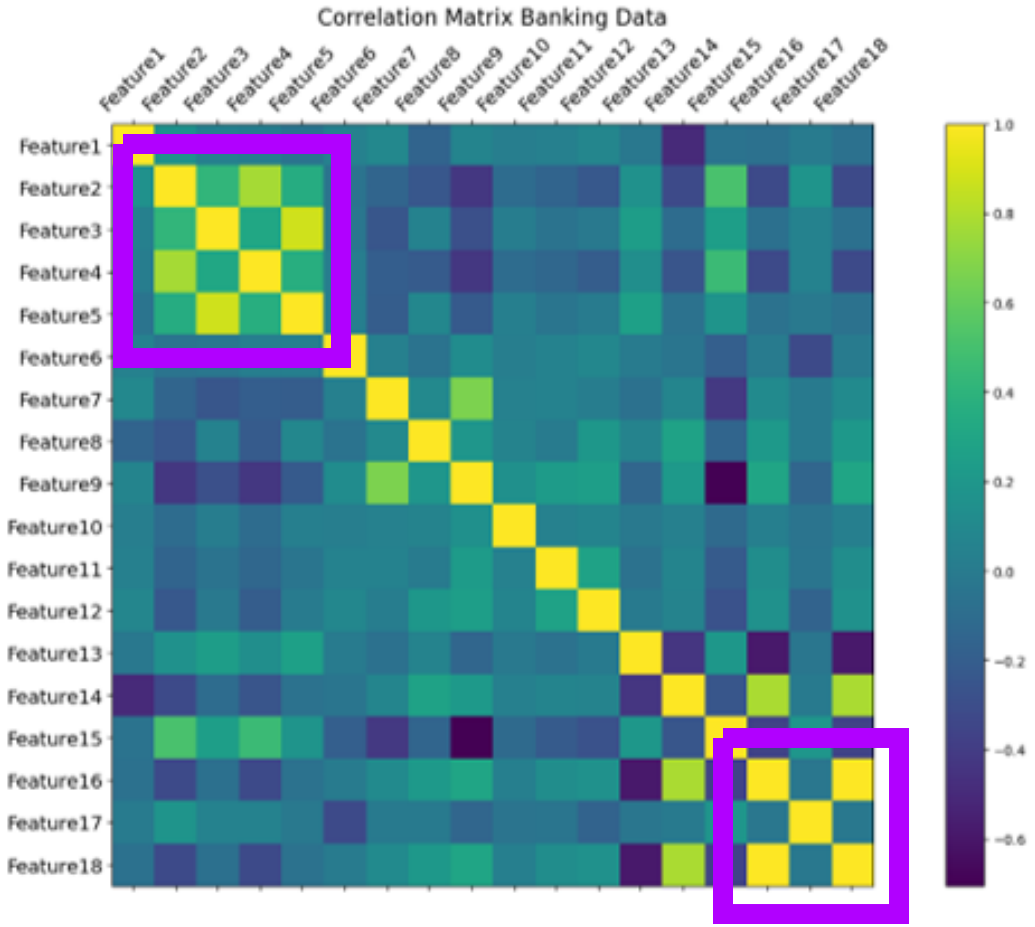
Data Overview

Banking Dataset

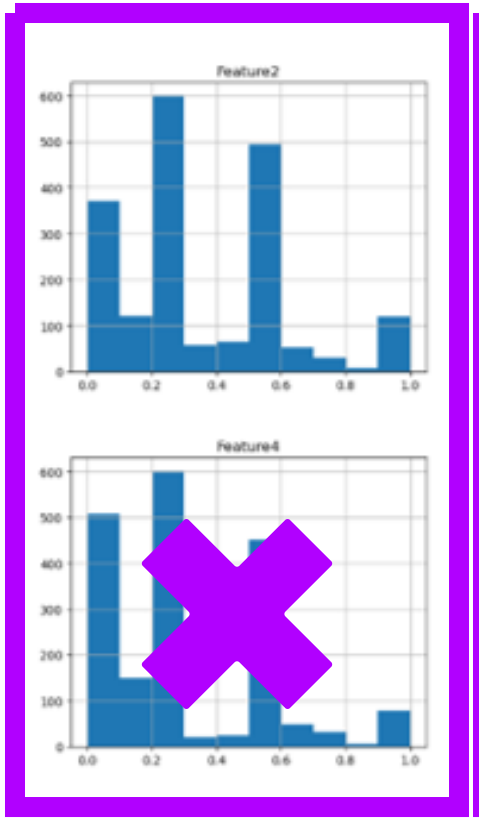
Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	Feature11	Feature12	Feature13	Feature14	Feature15	Feature16	Feature17	Feature18
39590	3	2	3	2	4	0	15011.75	1	0	0	0	2	4665	2	1	261	1
32696	6	8	6	8	5	0	0	0	0	0	0	3	11559	2	1	230	1
40124	6	8	6	8	4	0	31.25	0	0	0	0	9	0	2	0	188	0
36631	8	2	8	2	3	1	63.09	1	455.5	0	0	9	0	1	0	285	0
41206	5	8	3	0	1	0	0.29	0	0	0	0	3	3049	2	1	286	1
40695	4	8	2	4	1	0	2083.77	0	0	0	0	9	0	2	0	286	0
41188	4	8	0	7	1	0	1462.49	0	0	0	0	3	0	2	0	286	0
39170	10	8	10	2	3	0	402.57	0	0	0	0	9	0	2	0	286	0
43810	8	2	8	2	3	1	0.47	1	0	0	0	2	0	1	0	207	0
34943	1	2	1	2	4	0	0	1	1295.52	0	0	2	9312	0	1	76	1
40799	1	0	1	0	4	1	32.53	1	4305.14	0	0	1	3456	0	1	244	1

Data Overview

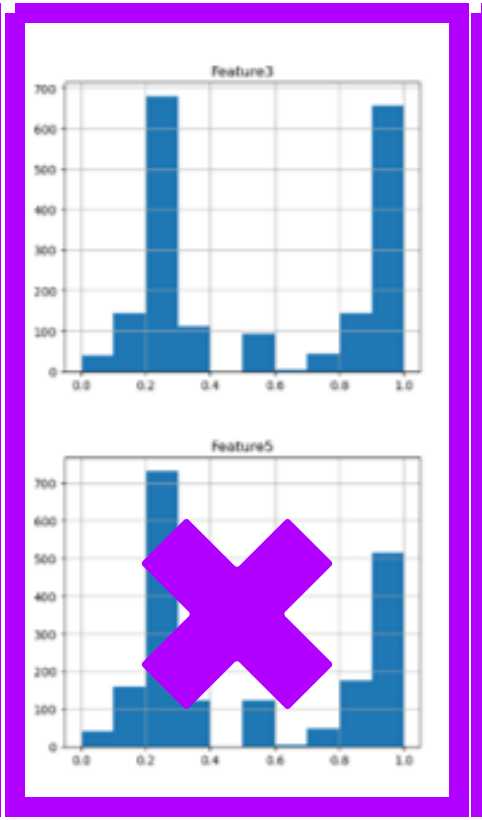
Banking Dataset



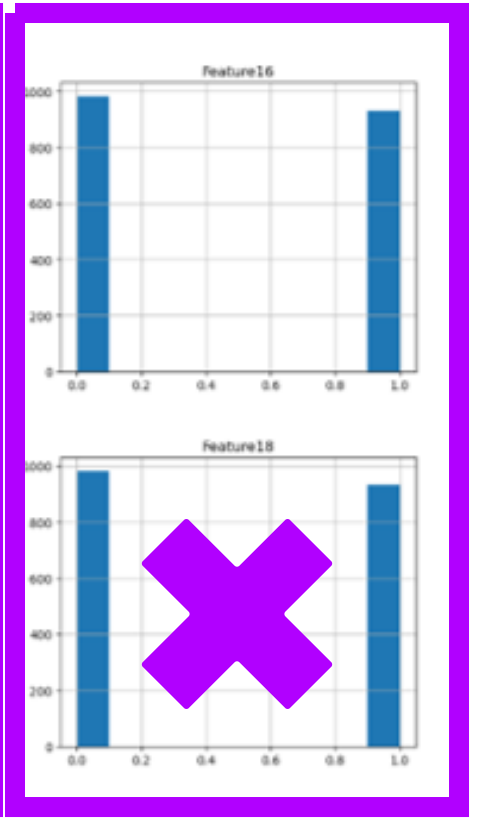
F2 & F4



F3 & F5



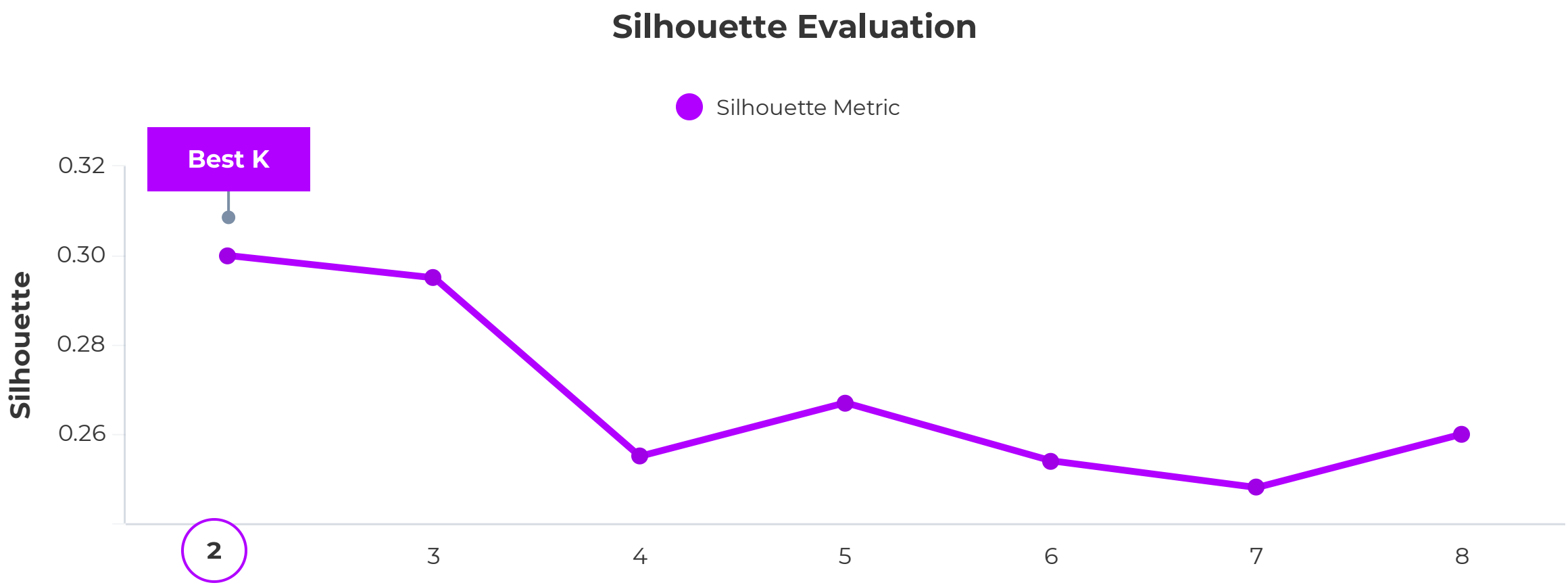
F16 & F18



MYDATAMODELS

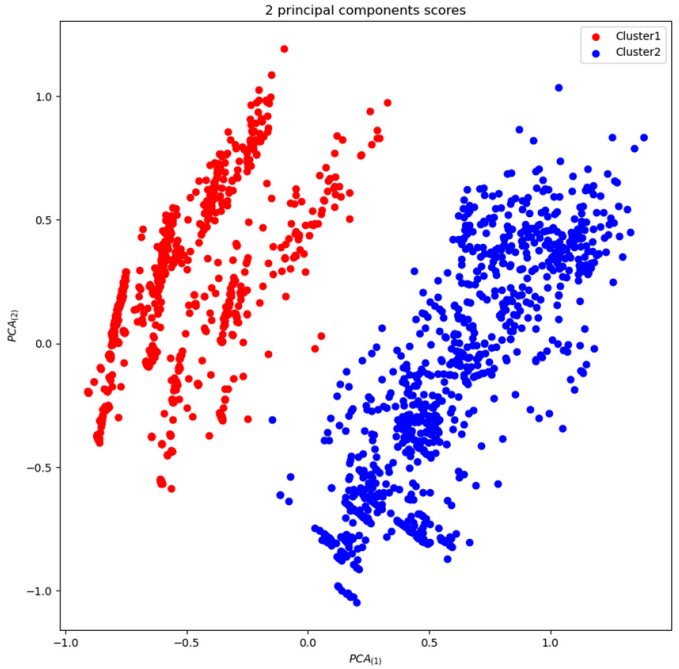
Data Clustering

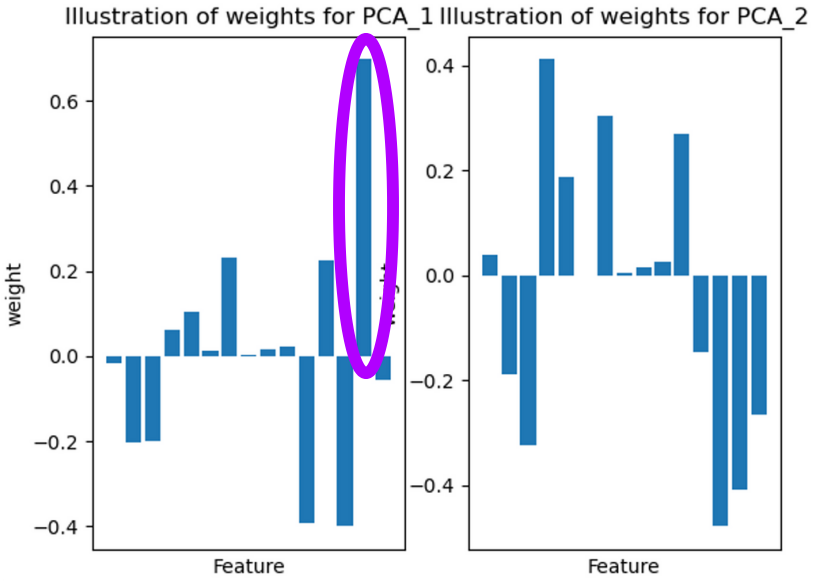
Banking Dataset



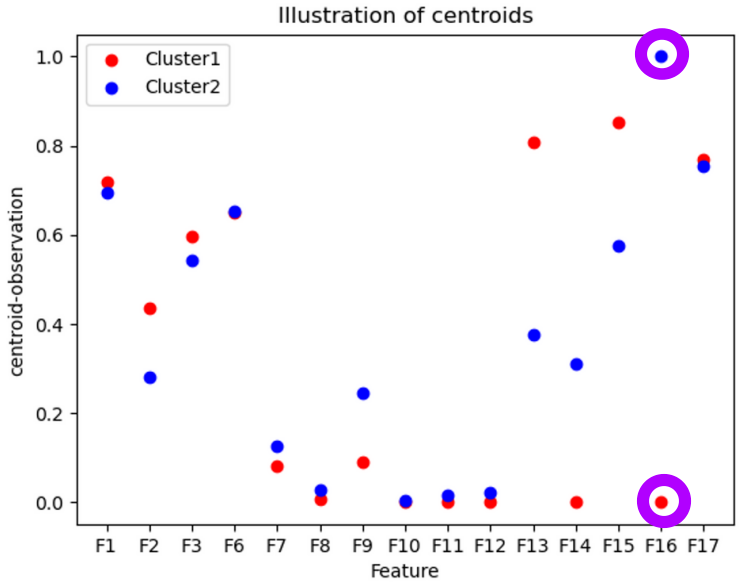
Cluster Analysis

Banking Dataset





F16



Marketing

Dataset

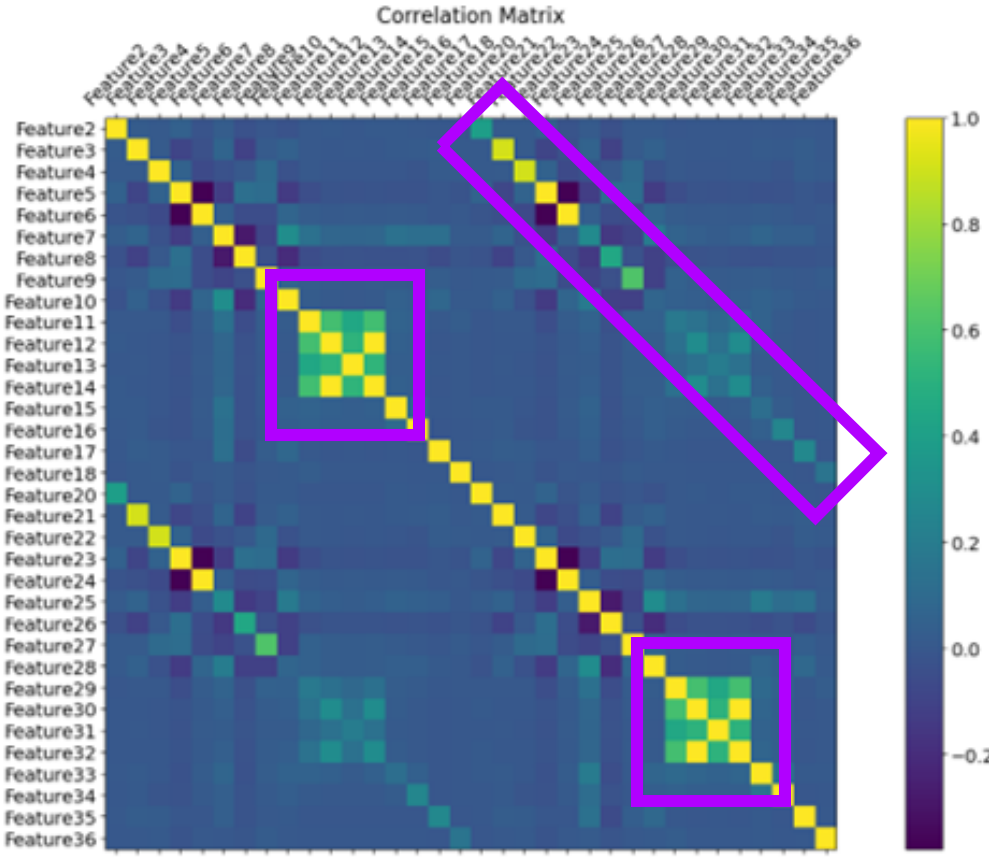
Data Overview

Marketing Dataset

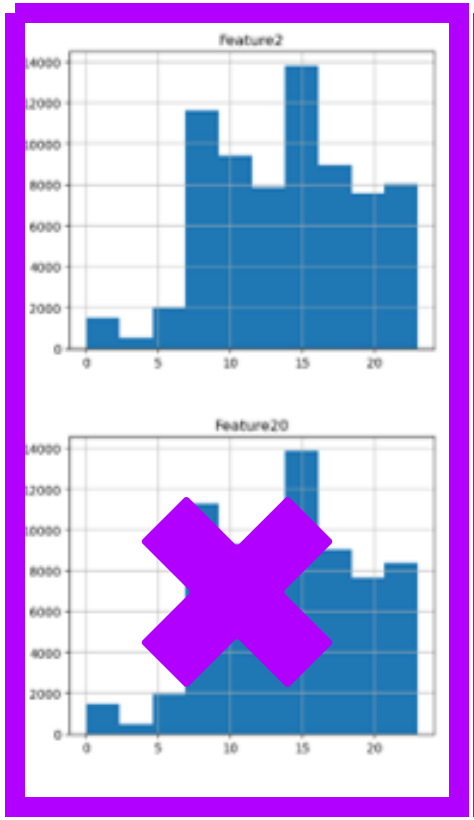
Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	Feature11	Feature12	Feature13	Feature14	Feature15	Feature16	Feature17	Feature18
6/1/2019	8	6	3	0	3	1	1	0	0	0	0	0	0	0	0	0	0
6/1/2019	1	4	3	0	4	3	0	0	4	0	0	0	0	0	0	0	0
6/2/2019	0	6	3	0	4	5	0	0	34	0	0	0	0	0	0	0	0
6/2/2019	23	1	3	1	6	1	1	0	0	0	0	0	0	0	0	0	0
6/2/2019	22	1	3	1	6	1	1	0	0	0	0	0	0	0	0	0	0
6/3/2019	16	1	42	0	3	14	0	0	3831	0	0	0	0	0	0	0	0
6/3/2019	15	1	42	0	3	3	0	0	88	0	0	0	0	0	0	0	0
6/3/2019	14	4	3	0	4	16	0	0	573	0	0	0	0	0	0	0	0
6/3/2019	11	1	42	0	3	33	0	0	2047	0	0	0	0	0	0	0	0
6/3/2019	16	1	42	0	4	4	0	0	249	0	0	0	0	0	0	0	0

Data Overview

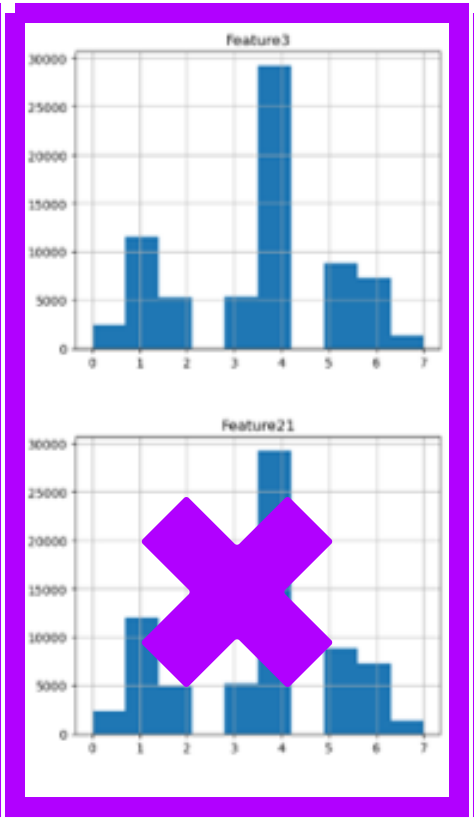
Marketing Dataset



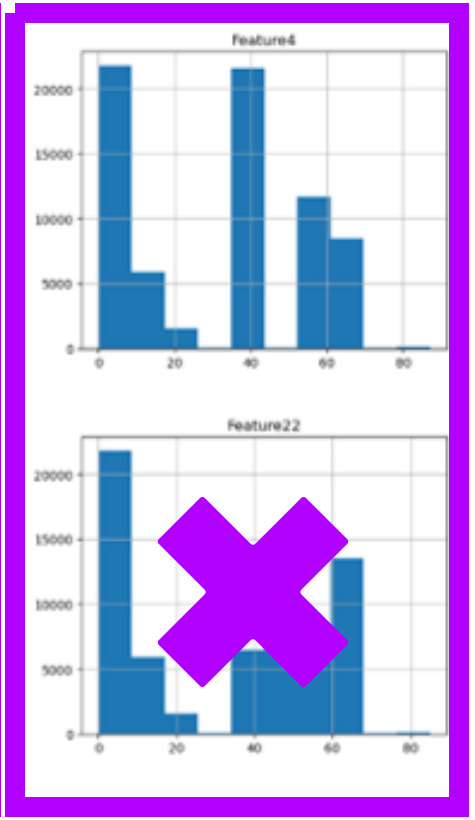
F2 & F20



F3 & F21

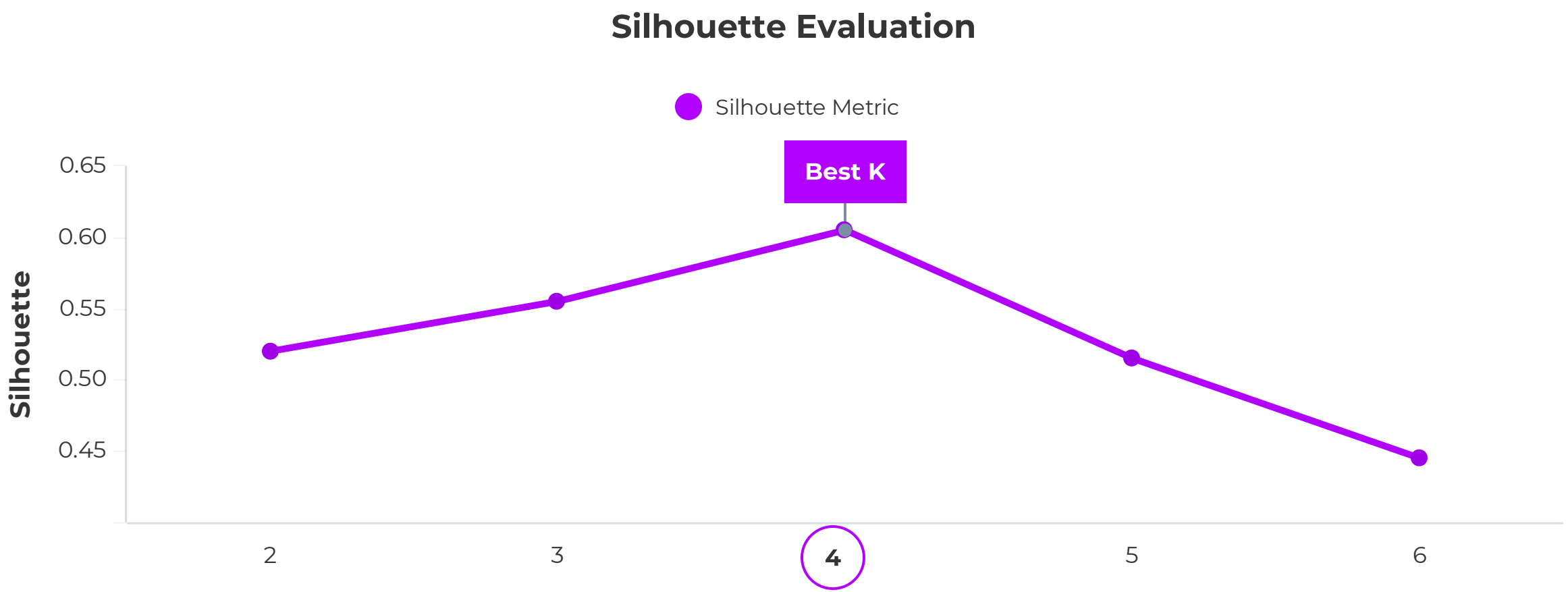


F4 & F22



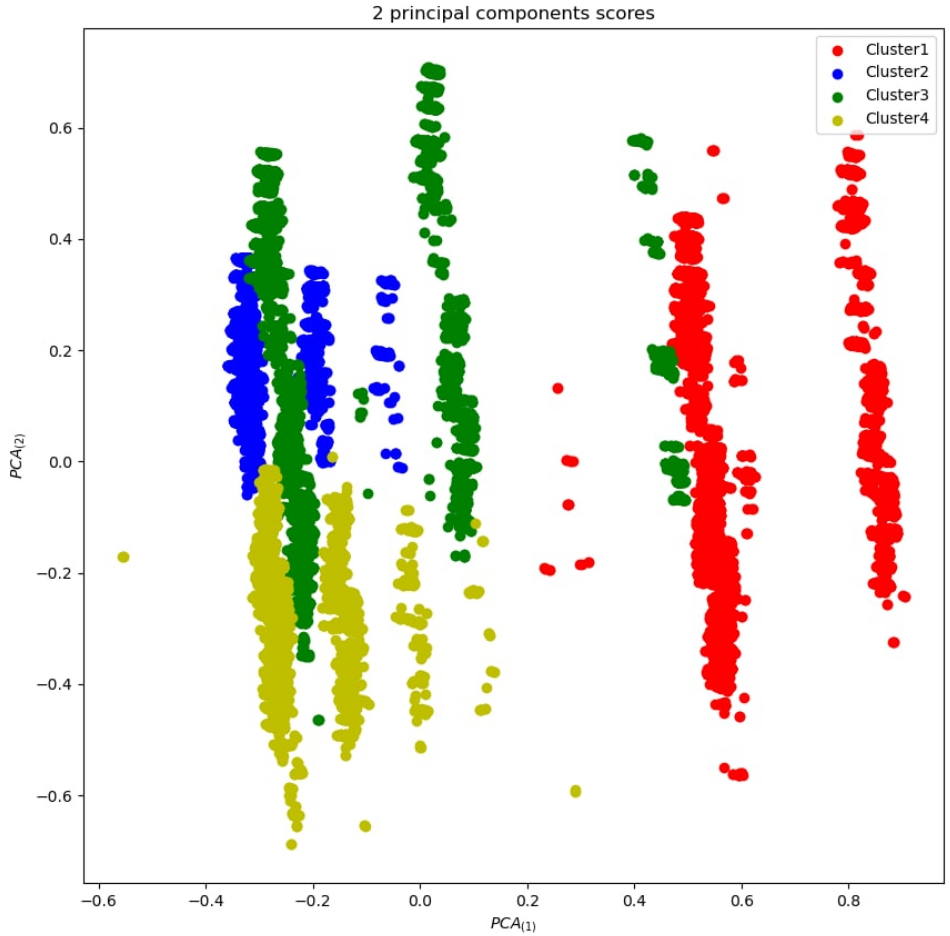
Data Clustering

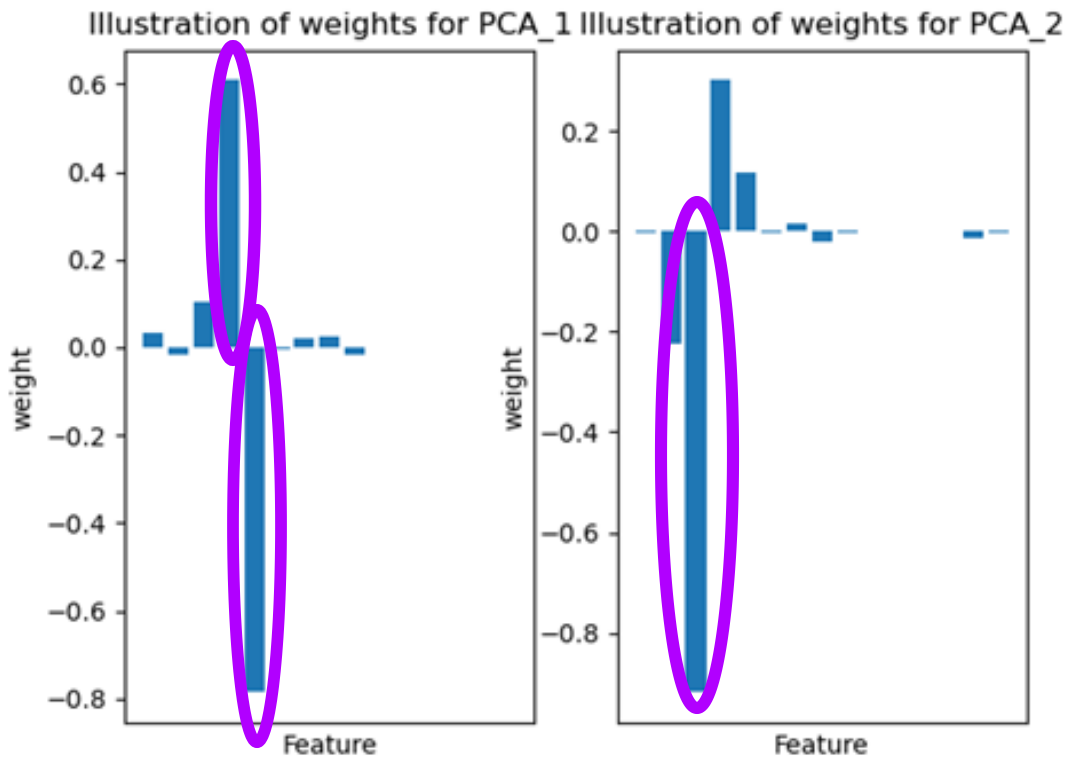
Marketing Dataset



Cluster Analysis

Marketing Dataset





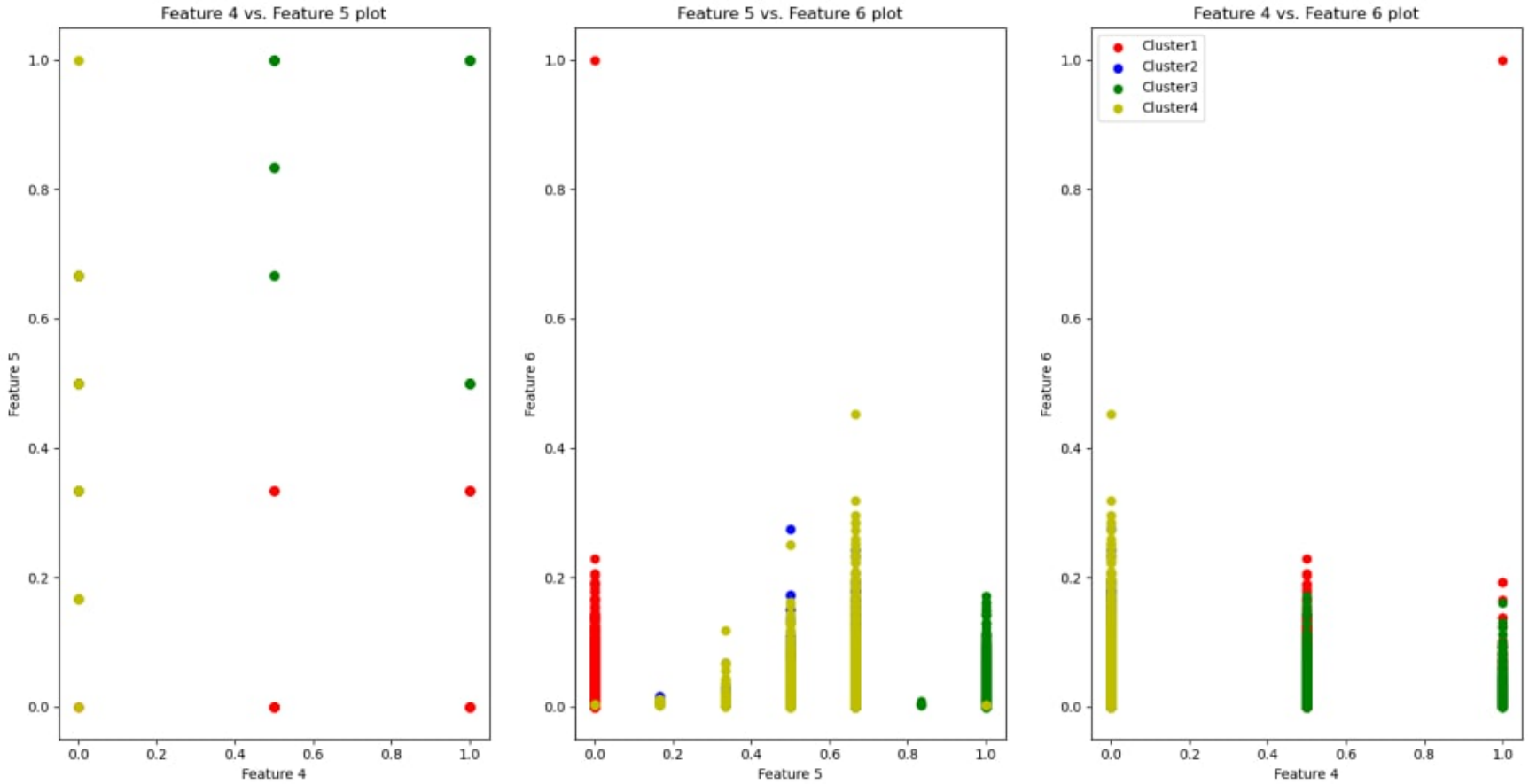
F5 & F6

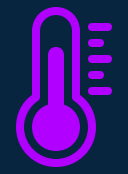
F4

Cluster Analysis

Marketing Dataset

The most relevant features are **F4**, **F5**, **F6**





Supply Chain

Dataset

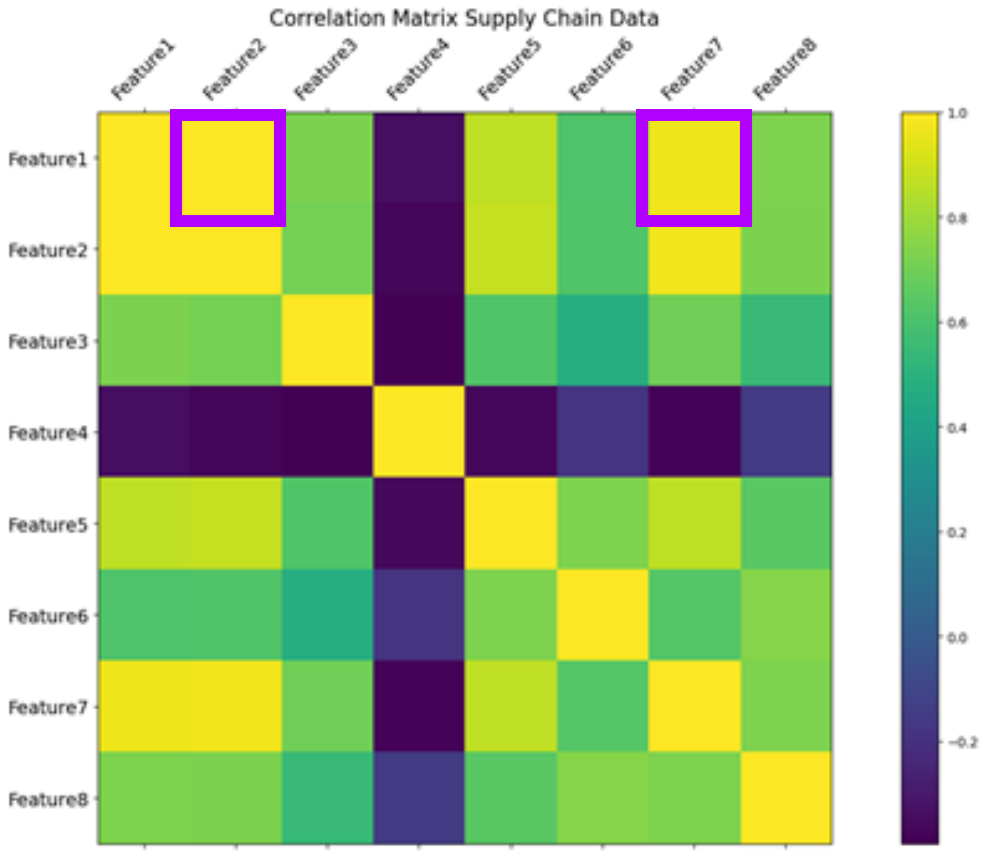
Data Overview

Supply Chain Dataset

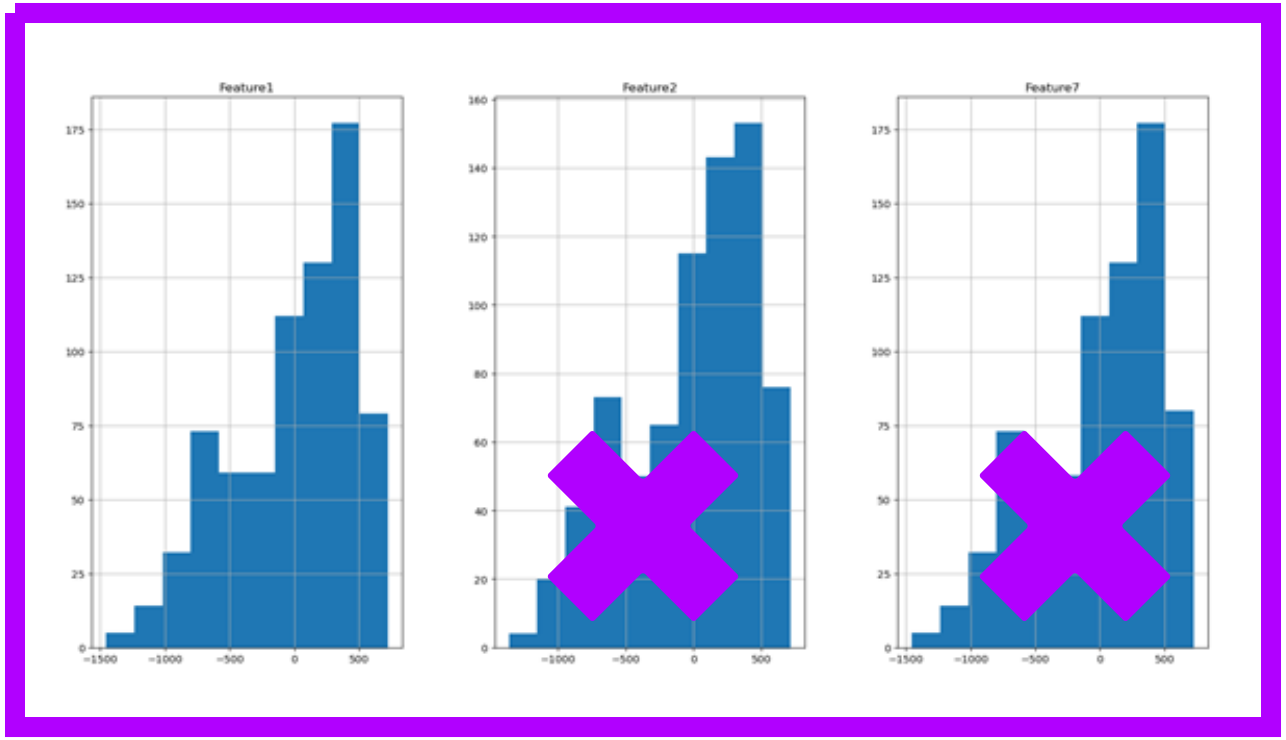
Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature10	Feature11
361.639	372.1742	46.85512	-149.491	69.5343	8.8	556.3684	14.4	Tuesday	April	4/1/2019
349.3086	415.9496	-28.3256	-221.788	-19.3297	9.2	361.639	15.7	Wednesda	April	4/2/2019
227.4342	292.1828	9.003453	-284.948	130.8434	10	349.3086	10.4	Thursday	April	4/3/2019
191.1352	243.2636	23.05316	-210.158	88.46188	10.5	227.4342	6.6	Friday	April	4/4/2019
267.4178	256.1257	11.37405	-150.983	145.4092	12.9	191.1352	7.4	Saturday	April	4/5/2019
437.2506	419.0309	16.01892	-116.733	478.7565	13.8	267.4178	9.9	Sunday	April	4/6/2019
500.1424	449.4386	39.74919	-147.253	556.3684	14.4	437.2506	12.6	Monday	April	4/7/2019

Data Overview

Supply Chain Dataset

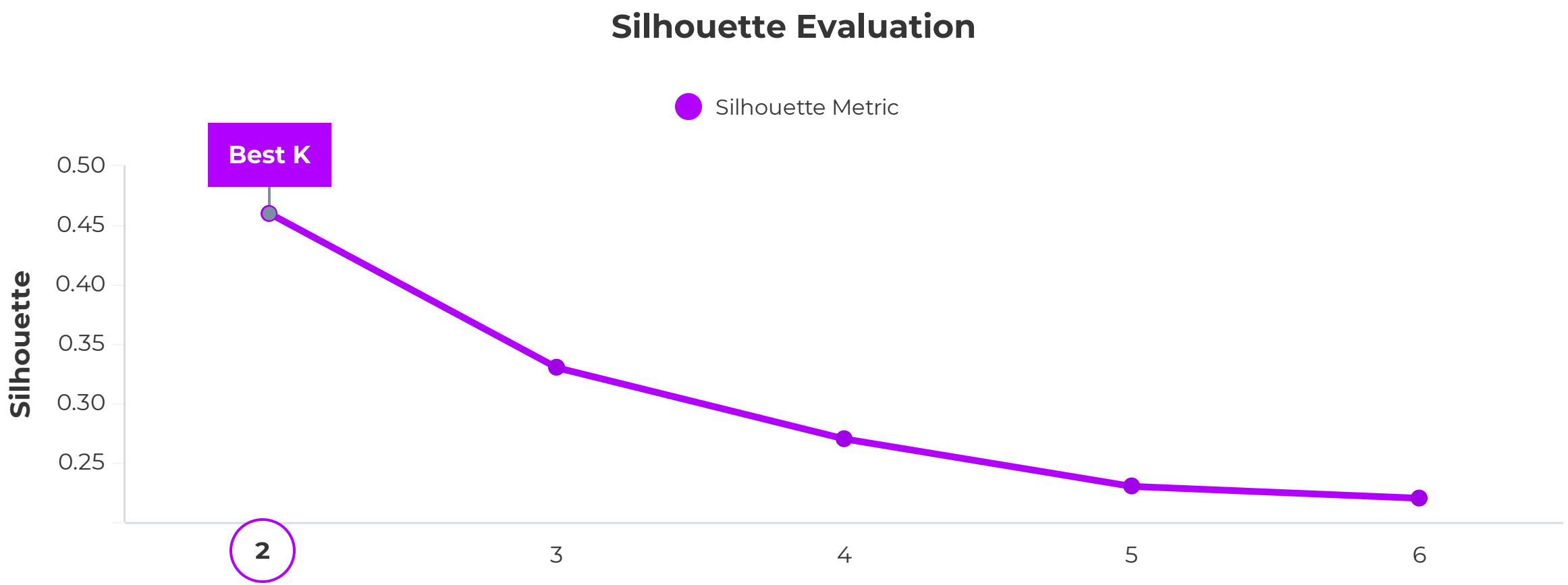


F1, F2 & F7



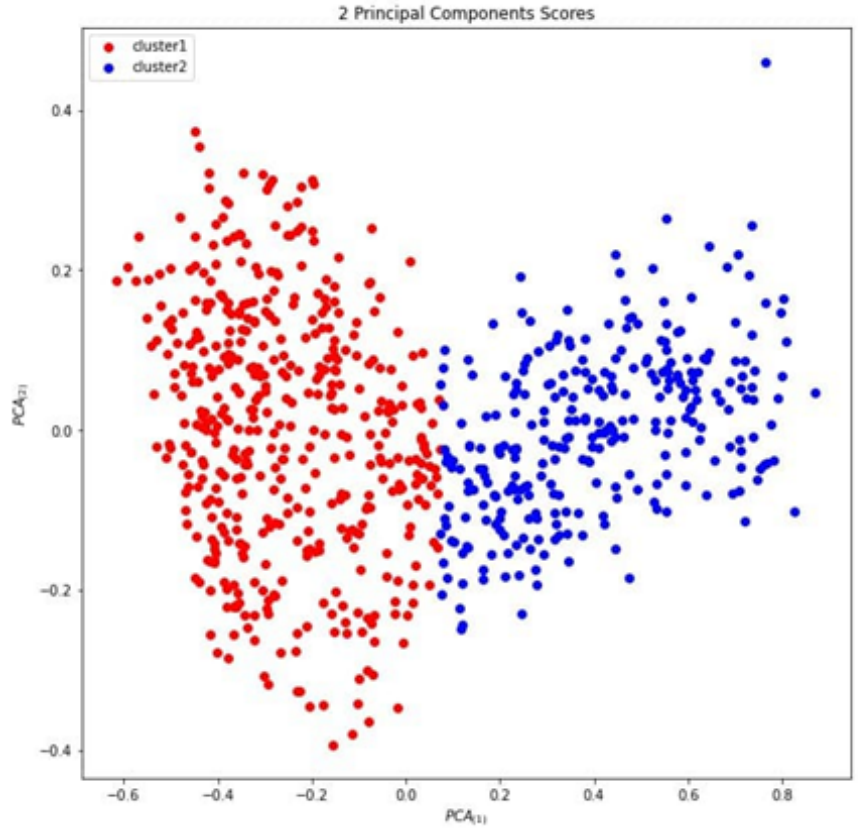
Data Clustering

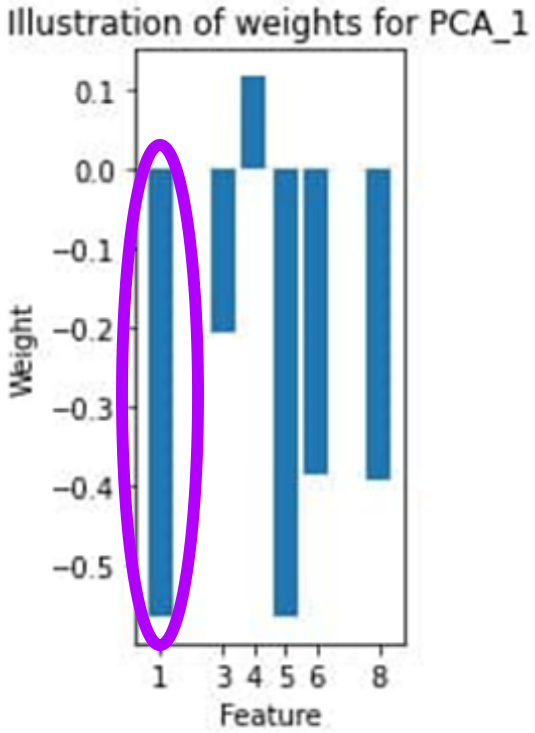
Supply Chain Dataset

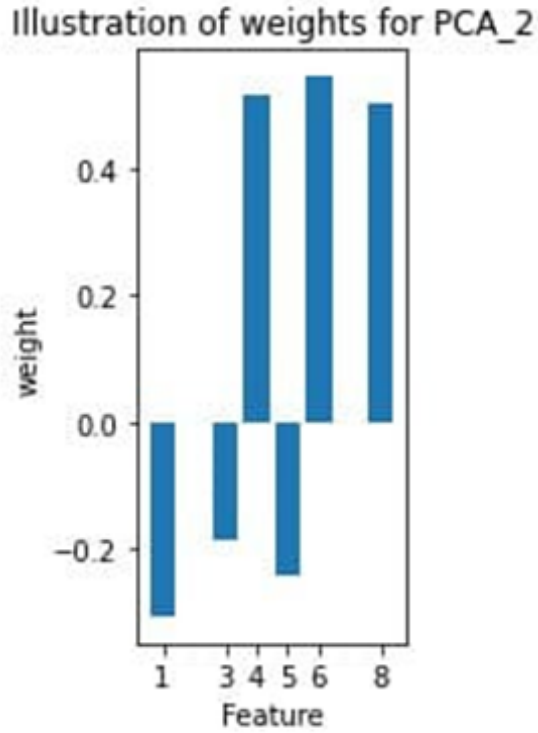


Cluster Analysis

Supply Chain Dataset



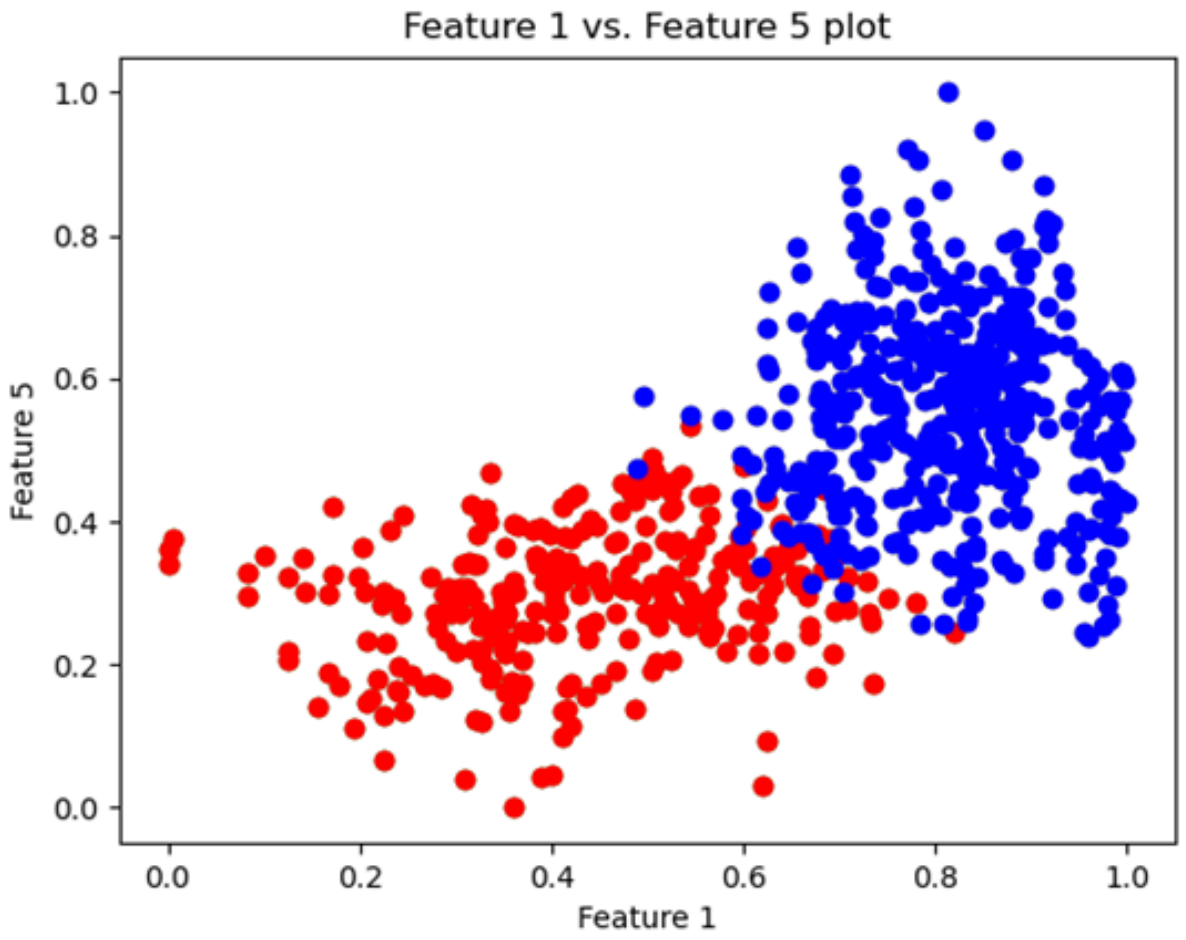




F1 (F5)

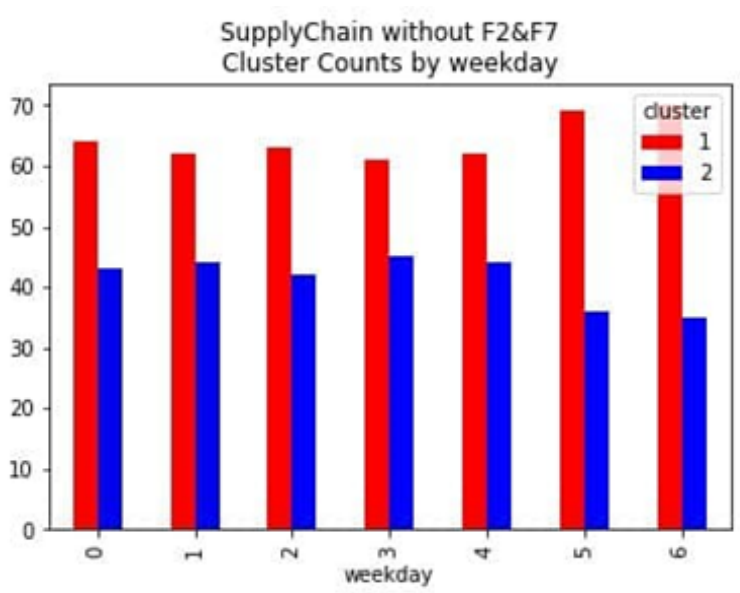
Cluster Analysis

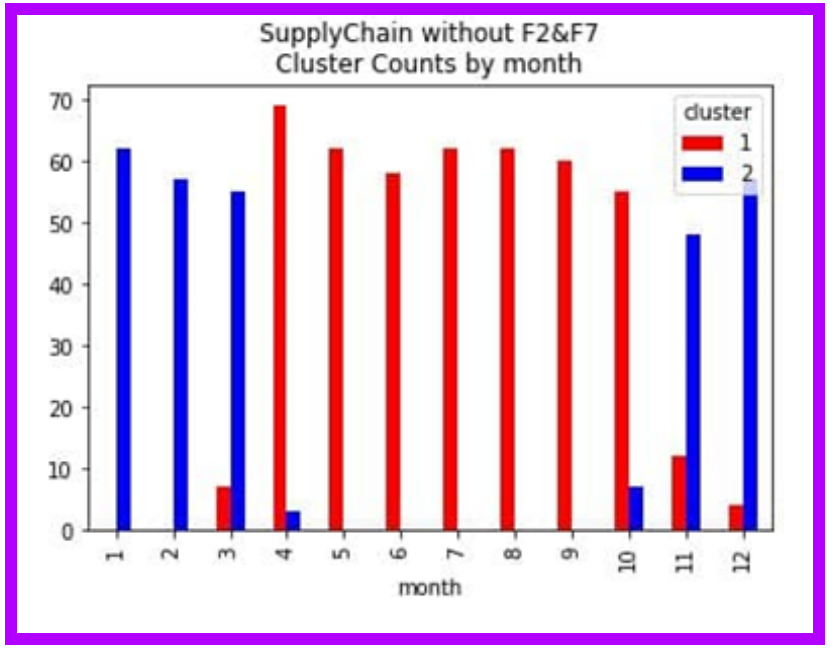
Supply Chain Dataset

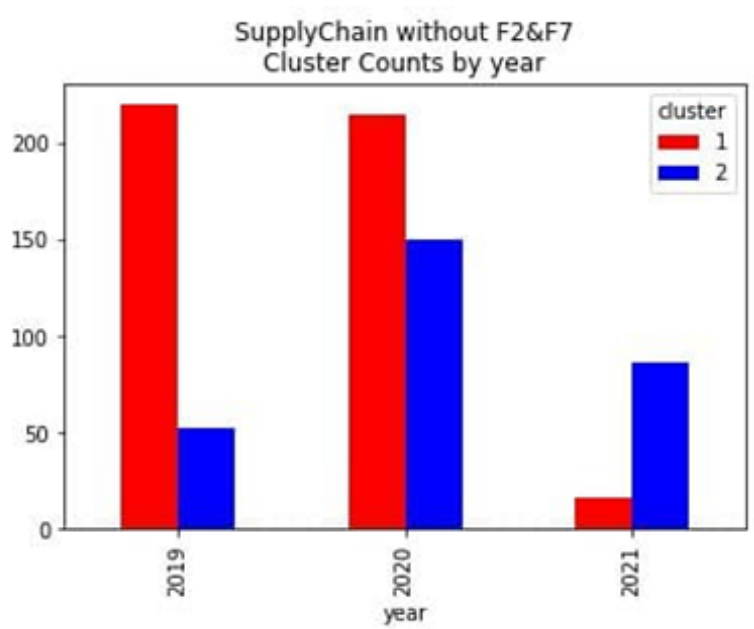


Cluster Analysis

Supply Chain Dataset







Cluster 1: "summer" season

Cluster 2: "winter" season

Thank you for your attention!

Any questions?

Google Colab references

- **Banking/Marketing Dataset**

  https://bit.ly/3l95bOC

- **Supply Chain Dataset**

  https://bit.ly/3CYqEA0