

Machine Learning - Clustering and Dimensionality Reduction

Assignment 3

Rhichard Koh

- Flatten the data in 4070 columns (each corresponding to a stock-code), and 4373 rows (each corresponding to a customer id). Your numbers may differ depending on the way you do flattening.

```
import pandas as pd
from sklearn.cluster import KMeans, DBSCAN
from sklearn.preprocessing import StandardScaler, MinMaxScaler
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
from sklearn.decomposition import PCA
```

```
df = pd.read_excel('Online Retail.xlsx')
```

df

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680.0	France

541909 rows × 8 columns

```
# Checking Unique Items
len(df["StockCode"].unique())
```

4070

```
# Pivoting the CustomerID and StockCode with the values as the Quantity
pivot_df = df.pivot_table(index='CustomerID', columns='StockCode', values='Quantity', aggfunc='sum', fill_value=0)
pivot_df
```

StockCode	10002	10080	10120	10125	10133	10135	11001	15030	15034	15036	...	90214Y	90214Z	BANK CHARGES	C2	CRUK	D	DOT	M	PADS	POST
CustomerID																					
12346.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	0
12347.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	0
12348.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	9
12349.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	1
12350.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	1
...
18280.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	0
18281.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	0
18282.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	0
18283.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	2	0	0
18287.0	0	0	0	0	0	0	0	0	0	0	...	0	0		0	0	0	0	0	0	0

4372 rows × 3684 columns

```
# Checking which item is bought the most.
pivot_df.max().sort_values()
```

```
StockCode
21412      0
79320      0
85023C      0
35832      0
85098B      0
...
21915     8120
17003    10077
84077    10080
22197    11692
84826    12540
Length: 3684, dtype: int64
```

```
# Converting the columns to be all string.
pivot_df.columns = pivot_df.columns.astype(str)
```

```
# Checking if the conversion happened.
pivot_df.columns[1]
```

'10080'

```
# Scaling the Quantity Values.
scaler = StandardScaler()
scaled = scaler.fit_transform(pivot_df)

scaled

array([[ -0.04089274, -0.04348365, -0.04752299, ..., -0.00995053,
        -0.03026138, -0.15831688],
       [ -0.04089274, -0.04348365, -0.04752299, ..., -0.00995053,
        -0.03026138, -0.15831688],
       [ -0.04089274, -0.04348365, -0.04752299, ..., -0.00995053,
        -0.03026138,  1.91678387],
       ...,
       [ -0.04089274, -0.04348365, -0.04752299, ..., -0.00995053,
        -0.03026138, -0.15831688],
       [ -0.04089274, -0.04348365, -0.04752299, ...,  0.01737593,
        -0.03026138, -0.15831688],
       [ -0.04089274, -0.04348365, -0.04752299, ..., -0.00995053,
        -0.03026138, -0.15831688]])
```

✓ Apply PCA and pick top 3 PCs to transform your 4070x4373 matrix into 3x4373 one.

```
# Applying PCA so we are only left with 3 columns.
pca = PCA(n_components=3)
scaled_pca = pca.fit_transform(scaled)

scaled_pca

array([[ -2.40439108, -1.35824465, -1.12467836],
       [  4.53401998,  4.77331158,  5.52102463],
       [  0.43233602, -0.89957571, -0.58101546],
       ...,
       [ -2.09091299, -0.80098264, -0.41156032],
       [  0.59991994, -1.23800721, -0.47412906],
       [  1.99955835,  7.31157919, -1.31877326]])

# Converting the scaled data into a dataframe for easier use.
scaled_df = pd.DataFrame(data=scaled, columns=pivot_df.columns, index=pivot_df.index)
scaled_df
```

StockCode	10002	10080	10120	10125	10133	10135	11001	15030	15034	15036	...	90214Y	90214Z	BANK CHARGES	C2	CRUK	D	
CustomerID																		
12346.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
12347.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
12348.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
12349.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
12350.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
...
18280.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
18281.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
18282.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
18283.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015
18287.0	-0.040893	-0.043484	-0.047523	-0.038833	-0.102894	-0.076524	-0.051014	-0.032103	-0.047351	-0.077025	...	-0.018645	-0.015125	-0.045418	-0.023847	0.015125	0.023437	-0.015

4372 rows × 3684 columns

```
# Converting the scaled and PCA transformed data into a dataframe for easier use.
scaled_pca_df = pd.DataFrame(data=scaled_pca, columns=['Principal Component 1', 'Principal Component 2', 'Principal Component 3'], index=pivot_df.index)
scaled_pca_df
```

	Principal Component 1	Principal Component 2	Principal Component 3
CustomerID			
12346.0	-2.404391	-1.358245	-1.124678
12347.0	4.534020	4.773312	5.521025
12348.0	0.432336	-0.899576	-0.581015
12349.0	-0.266184	-0.597366	0.383031
12350.0	-1.918710	-0.589165	-0.547577
...
18280.0	-2.315659	-1.261694	-0.943462
18281.0	-2.089671	-1.432836	-1.041452
18282.0	-2.090913	-0.800983	-0.411560
18283.0	0.599920	-1.238007	-0.474129
18287.0	1.999558	7.311579	-1.318773

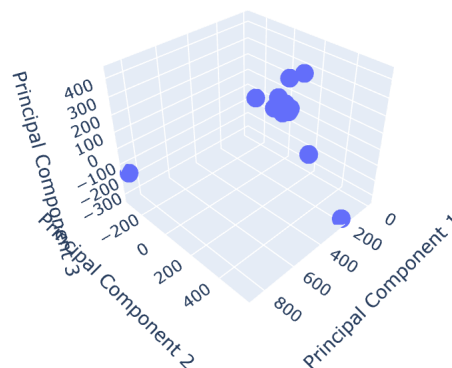
4372 rows × 3 columns

Next steps:

[Generate code with scaled_pca_df](#)[View recommended plots](#)

```
# Viewing a 3D visual of the scaled and PCA transformed data.
import plotly.express as px
fig = px.scatter_3d(scaled_pca_df, x='Principal Component 1', y='Principal Component 2', z='Principal Component 3', title='Scaled PCA DataFrame')
fig.show()
```

Scaled PCA DataFrame



✓ Get cluster labels from each algorithm.

✓ K-Means

```
# Finding the optimal number of clusters for K-Means.
wcss = []
```

```
for each in range(1, 100):
    kmeans = KMeans(n_clusters=each)
    kmeans.fit(scaled_df)
    wcss.append(kmeans.inertia_)
```

```
plt.plot(range(1, 100), wcss)
plt.title('Elbow plot')
plt.xlabel("Number of k value (each)")
plt.ylabel("WCSS")
plt.show()
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]