Assignment 6 Documentation

Pavan Kalyan Rao Yelamati
700729168

2. CC General Dataset---PCA---Hierarchical Clustering

- Read cc general dataset using pandas.
- Check head to see columns and type of data.
- Drop CUST_ID column.
- Check for null values using isnull().sum().
- Fill missing values with column means using fillna() method.
- Apply standard scaling and normalization on this data.
- Do PCA on this transformed data with 2 components.
- Use this pca data to fit different variations(linkage=ward, single, complete, average) of AgglomerativeClustering with k=2.
- Note the silhouette score for each variant.
- Now repeat this process for k=3, 4, 5 and save silhouette scores.
- Plot the clusters for different combinations using matplotlib.
- Also use silhouette scores to barplot and understand for how many clusters and which linkage the clustering apt for.

Data:

```
cc.head()
#cc.describe()
```

]:

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FRE
0	C10001	40.900749	0.818182	95.40	0.00	95.4	0.000000	
1	C10002	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	
2	C10003	2495.148862	1.000000	773.17	773.17	0.0	0.000000	
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.0	205.788017	
4	C10005	817.714335	1.000000	16.00	16.00	0.0	0.000000	

CUST_ID categorical column not required.

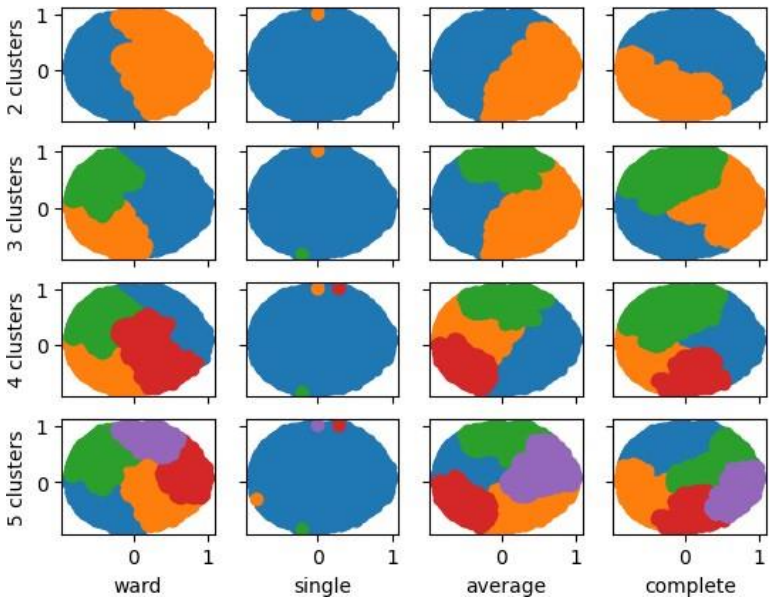
Missing values:

```
print(cc.isnull().sum()) # check for null values
```

```
BALANCE                                0
BALANCE_FREQUENCY                      0
PURCHASES                              0
ONEOFF_PURCHASES                       0
INSTALLMENTS_PURCHASES                 0
CASH_ADVANCE                           0
PURCHASES_FREQUENCY                    0
ONEOFF_PURCHASES_FREQUENCY             0
PURCHASES_INSTALLMENTS_FREQUENCY       0
CASH_ADVANCE_FREQUENCY                 0
CASH_ADVANCE_TRX                       0
PURCHASES_TRX                          0
CREDIT_LIMIT                           1
PAYMENTS                               0
MINIMUM_PAYMENTS                     313
PRC_FULL_PAYMENT                       0
TENURE                                 0
dtype: int64
```

These missing values should be filled with mean of the columns.

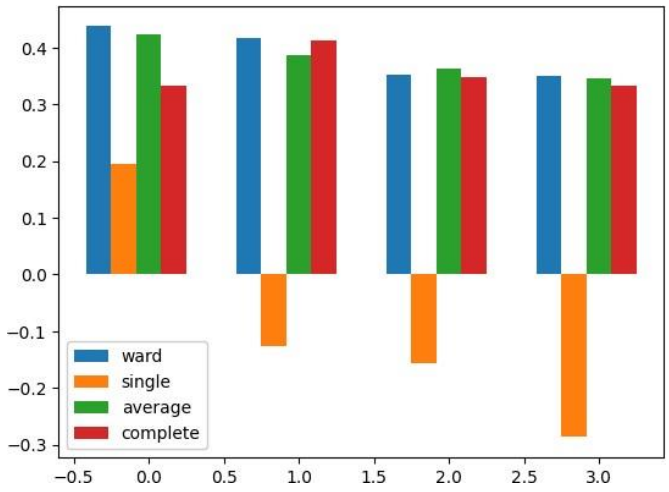
Clusters:



Silhouette Scores:

```
Silhouetter Score for 2 cluster ward linkage : 0.437324287290409
Silhouetter Score for 2 cluster single linkage : 0.19450588814856126
Silhouetter Score for 2 cluster average linkage : 0.4224295086990182
Silhouetter Score for 2 cluster complete linkage : 0.33218800469234344
Silhouetter Score for 3 cluster ward linkage : 0.416722988721946
Silhouetter Score for 3 cluster single linkage : -0.12691534093830453
Silhouetter Score for 3 cluster average linkage : 0.38721918047525616
Silhouetter Score for 3 cluster complete linkage : 0.4125839066592452
Silhouetter Score for 4 cluster ward linkage : 0.35310619113979114
Silhouetter Score for 4 cluster single linkage : -0.1565509511680269
Silhouetter Score for 4 cluster average linkage : 0.36203366257683434
Silhouetter Score for 4 cluster complete linkage : 0.3481845744776147
Silhouetter Score for 5 cluster ward linkage : 0.3505507966819357
Silhouetter Score for 5 cluster single linkage : -0.2868555209728266
Silhouetter Score for 5 cluster average linkage : 0.344783541461381
Silhouetter Score for 5 cluster complete linkage : 0.3328569302298077
```

Silhouette Score bar plot:



Maximum silhouette score is 0.437 from forming 2 clusters with ward linkage clustering. We also observe that except for single linkage all other types have more or less the same score.