Group Members

1.Suthasinee Pojam 6220422065

2.Siraprapa Chunloy

3.Teremate Tangpatong

Load Dependencies

```
!pip install pycaret
```

```
  Downloading pydantic-1.8.2-cp37-cp37m-manylinux2014_x86_64.whl (10.1 MB)
     |████████████████████████████████| 10.1 MB 33.9 MB/s
Requirement already satisfied: tqdm>=4.48.2 in /usr/local/lib/python3.7/dist-packages (from pandas-profi
Collecting visions[type_image_path]==0.7.4
  Downloading visions-0.7.4-py3-none-any.whl (102 kB)
     |████████████████████████████████| 102 kB 8.3 MB/s
Requirement already satisfied: attrs>=19.3.0 in /usr/local/lib/python3.7/dist-packages (from visions[type_
Requirement already satisfied: networkx>=2.4 in /usr/local/lib/python3.7/dist-packages (from visions[type
Requirement already satisfied: Pillow in /usr/local/lib/python3.7/dist-packages (from visions[type_image_p
Collecting imagehash
  Downloading ImageHash-4.2.1.tar.gz (812 kB)
     |████████████████████████████████| 812 kB 42.9 MB/s
Collecting scipy<=1.5.4
  Downloading scipy-1.5.4-cp37-cp37m-manylinux1_x86_64.whl (25.9 MB)
     |████████████████████████████████| 25.9 MB 1.6 MB/s
Requirement already satisfied: retrying>=1.3.3 in /usr/local/lib/python3.7/dist-packages (from plotly>=4.
Requirement already satisfied: wcwidth in /usr/local/lib/python3.7/dist-packages (from prompt-toolkit<2.0
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /usr/local/lib/python3.7/dist-packages (from requ
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.7/dist-packages (from requests>=2
Requirement already satisfied: charset-normalizer~=2.0.0 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.7/dist-packages (from requests
Requirement already satisfied: wasabi<1.1.0,>=0.4.0 in /usr/local/lib/python3.7/dist-packages (from spac
Requirement already satisfied: srsly<1.1.0,>=1.0.2 in /usr/local/lib/python3.7/dist-packages (from spacy<
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from spac
Requirement already satisfied: blis<0.5.0,>=0.4.0 in /usr/local/lib/python3.7/dist-packages (from spacy<2
Requirement already satisfied: plac<1.2.0,>=0.9.6 in /usr/local/lib/python3.7/dist-packages (from spacy<
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.7/dist-packages (fro
Requirement already satisfied: thinc==7.4.0 in /usr/local/lib/python3.7/dist-packages (from spacy<2.4.0->
Requirement already satisfied: catalogue<1.1.0,>=0.0.7 in /usr/local/lib/python3.7/dist-packages (from sp
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from spa
Requirement already satisfied: importlib-metadata>=0.20 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-packages (from importlib-metada
Requirement already satisfied: notebook>=4.4.1 in /usr/local/lib/python3.7/dist-packages (from widgetsnb
Requirement already satisfied: nbconvert in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1
Requirement already satisfied: terminado>=0.8.1 in /usr/local/lib/python3.7/dist-packages (from noteboo
Requirement already satisfied: Send2Trash in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4
Requirement already satisfied: pyzmq>=13 in /usr/local/lib/python3.7/dist-packages (from jupyter-client->
Requirement already satisfied: ptyprocess in /usr/local/lib/python3.7/dist-packages (from terminado>=0.8
Requirement already satisfied: PyWavelets in /usr/local/lib/python3.7/dist-packages (from imagehash->vis
```

```
Collecting databricks-cli>=0.8.7
  Downloading databricks-cli-0.16.2.tar.gz (58 kB)
     |████████████████████████████████| 58 kB 5.8 MB/s
Requirement already satisfied: click>=7.0 in /usr/local/lib/python3.7/dist-packages (from mlflow->pycaret
Requirement already satisfied: sqlparse>=0.3.1 in /usr/local/lib/python3.7/dist-packages (from mlflow->p
Collecting docker>=4.0.0
  Downloading docker-5.0.3-py2.py3-none-any.whl (146 kB)
     |████████████████████████████████| 146 kB 49.0 MB/s
Collecting alembic<=1.4.1
  Downloading alembic-1.4.1.tar.gz (1.1 MB)
     |████████████████████████████████| 1.1 MB 45.6 MB/s
Collecting prometheus-flask-exporter
  Downloading prometheus_flask_exporter-0.18.7-py3-none-any.whl (17 kB)
Requirement already satisfied: packaging in /usr/local/lib/python3.7/dist-packages (from mlflow->pycaret)
Requirement already satisfied: entrypoints in /usr/local/lib/python3.7/dist-packages (from mlflow->pycaret
Requirement already satisfied: cloudpickle in /usr/local/lib/python3.7/dist-packages (from mlflow->pycaret
Collecting gitpython>=2.1.0
  Downloading GitPython-3.1.24-py3-none-any.whl (180 kB)
```

```
import pandas as pd
from pycaret.clustering import *
```

Load Data

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
df = pd.read_csv('/content/drive/MyDrive/SupermarketData.csv')
```

```
df.shape
```

```
(956574, 22)
```

```
df.head()
```

	SHOP_WEEK	SHOP_DATE	SHOP_WEEKDAY	SHOP_HOUR	QUANTITY	SPEND	PROD_

```
df['SHOP_DATE'] = df['SHOP_DATE'].apply(lambda x: pd.to_datetime(str(x), format='%Y%m%d'))
```

1	200733	20071010	4	20	3	6.75	PRD0!

```
df2=df
```

```
df.describe()
```

	SHOP_WEEK	SHOP_WEEKDAY	SHOP_HOUR	QUANTITY	SPEND	BAS
count	956574.000000	956574.000000	956574.000000	956574.000000	956574.000000	9.565
mean	200702.251671	3.996021	14.950665	1.514577	1.871697	9.941
std	65.857803	1.997058	3.636119	1.621021	2.767820	3.332
min	200607.000000	1.000000	8.000000	1.000000	0.010000	9.941
25%	200637.000000	2.000000	12.000000	1.000000	0.750000	9.941
50%	200713.000000	4.000000	15.000000	1.000000	1.200000	9.941
75%	200742.000000	6.000000	18.000000	1.000000	2.060000	9.941
max	200819.000000	7.000000	21.000000	129.000000	476.160000	9.941

```
df.info
```

```
...          ...        ...          ...     ...    ...   ...
956569    200617 2006-06-22            5      12      3   3.96
956570    200633 2006-10-13            6      20      3   3.96
956571    200617 2006-06-22            5      18      3   3.96
956572    200619 2006-07-06            5      19      3   3.96
956573    200635 2006-10-23            2      21      3   3.96

         PROD_CODE PROD_CODE_10 PROD_CODE_20 PROD_CODE_30 PROD_CODE_40  \
0       PRD0900001      CL00072     DEP00021       G00007       D00002
1       PRD0900001      CL00072     DEP00021       G00007       D00002
2       PRD0900001      CL00072     DEP00021       G00007       D00002
3       PRD0900001      CL00072     DEP00021       G00007       D00002
4       PRD0900001      CL00072     DEP00021       G00007       D00002
...            ...          ...          ...          ...          ...
956569  PRD0904997      CL00074     DEP00021       G00007       D00002
956570  PRD0904997      CL00074     DEP00021       G00007       D00002
956571  PRD0904997      CL00074     DEP00021       G00007       D00002
956572  PRD0904997      CL00074     DEP00021       G00007       D00002
956573  PRD0904997      CL00074     DEP00021       G00007       D00002

             CUST_CODE CUST_PRICE_SENSITIVITY CUST_LIFESTAGE        BASKET_ID  \
0       CUST0000583261                     UM             YF  994107800547472
1       CUST0000537317                     MM             OF  994107900512001
2       CUST0000472158                     MM             YF  994108700468327
3       CUST0000099658                     LA             OF  994107700237811
```

```
4              NaN                NaN          NaN  994108300002212
...            ...                ...          ...  ...
956569             NaN                NaN          NaN  994101100088778
956570             NaN                NaN          NaN  994102700099738
956571  CUST0000544241               LA           YA  994101100506174
956572  CUST0000423155               LA           YF  994101300433650
956573             NaN                NaN          NaN  994102900104676

       BASKET_SIZE BASKET_PRICE_SENSITIVITY BASKET_TYPE  \
0               L                       MM      Top Up
1               L                       MM   Full Shop
2               L                       MM   Full Shop
3               L                       LA   Full Shop
4               L                       MM   Full Shop
...           ...                      ...         ...
956569          M                       MM      Top Up
956570          L                       LA      Top Up
956571          L                       LA      Top Up
956572          L                       LA   Full Shop
956573          L                       MM      Top Up

       BASKET_DOMINANT_MISSION  STORE_CODE STORE_FORMAT STORE_REGION
0                      Grocery  STORE00001           LS          E02
1                        Fresh  STORE00001           LS          E02
2                      Grocery  STORE00001           LS          E02
3                        Mixed  STORE00001           LS          E02
4                        Fresh  STORE00001           LS          E02
...                        ...         ...          ...          ...
956569                   Fresh  STORE00002           LS          W01
956570                   Fresh  STORE00002           LS          W01
956571                   Fresh  STORE00002           LS          W01
956572                   Fresh  STORE00002           LS          W01
```

Prepare customer single view

Define features

Total visits = COUNT(DISTINCT BASKET ID)

Ticket size = SUM(SPEND)/COUNT(DISTINCT BASKET ID)

Total no. of SKUs

FirstDate min SHOP_Date

LastDate max SHOP_Date

Calculate features

```
##prepare customer single view
df_csv = df_groupby = df[df['CUST_CODE'].notnull()].groupby(by=['CUST_CODE']).agg(TotalSpend=('SPEND', 'sum
                                                                                  TotalVisits=('BASKET_ID', 'nunique'),
                                                                                  TotalSKUs=('PROD_CODE', 'nunique'),
                                                                                  TotalSKUs_10=('PROD_CODE_10', 'nunique'),
                                                                                  TotalSKUs_20=('PROD_CODE_20', 'nunique'),
                                                                                  TotalSKUs_30=('PROD_CODE_30', 'nunique'),
                                                                                  TotalSKUs_40=('PROD_CODE_40', 'nunique'),
                                                                                  FirstDate=('SHOP_DATE', 'min'),
                                                                                  LastDate=('SHOP_DATE', 'max'),

                                                                                  ).reset_index()


##calculate ticket size
df_csv['TicketSize'] = df_csv['TotalSpend']/df_csv['TotalVisits']


##find max date in the dataset
max_date = df_csv['LastDate'].max()


##calculate total days of the relationship
df_csv['total_days'] = (df_csv['LastDate'] - df_csv['FirstDate']).dt.days + 1


##calculate recency days
df_csv['recency'] = (max_date - df_csv['LastDate']).dt.days
```

```
df_csv.head(5)
```

	CUST_CODE	TotalSpend	TotalVisits	TotalSKUs	TotalSKUs_10	TotalSKUs_20	Tot
0	CUST0000000181	2.44	1	1	1	1	
1	CUST0000000369	959.33	220	189	81	36	
2	CUST0000000689	328.57	16	116	73	41	

```
df_csv.shape
```

```
(6100, 13)
```

```
df2['attend']=1
```

```
df2.head()
```

	SHOP_WEEK	SHOP_DATE	SHOP_WEEKDAY	SHOP_HOUR	QUANTITY	SPEND	PROD_
0	200732	2007-10-05	6	17	3	6.75	PRD09
1	200733	2007-10-10	4	20	3	6.75	PRD09
2	200741	2007-12-09	1	11	1	2.25	PRD09
3	200731	2007-09-29	7	17	1	2.25	PRD09
4	200737	2007-11-10	7	14	3	6.75	PRD09

```
##prepare customer single view
df_csv2 = df2[df2['CUST_CODE'].notnull()].groupby(by=['CUST_CODE','SHOP_WEEK']).agg(TotalAtt=('attend', 'sum
```

```
df_csv3 = df_csv2[df_csv2['CUST_CODE'].notnull()].groupby(by=['CUST_CODE','SHOP_WEEK']).agg(TotalAttMin=('
```

```
df_csv3.head()
```

	CUST_CODE	SHOP_WEEK	TotalAttMin	TotalAttMax
0	CUST0000000181	200645	1	1
1	CUST0000000369	200607	4	4
2	CUST0000000369	200608	4	4
3	CUST0000000369	200609	3	3
4	CUST0000000369	200610	12	12

```
df_csv_final = pd.concat([df_csv3, df_csv], ignore_index=True)
```

```
df_csv.head()
```

	CUST_CODE	TotalSpend	TotalVisits	TotalSKUs	TotalSKUs_10	TotalSKUs_20	Tot
0	CUST0000000181	2.44	1	1	1	1	
1	CUST0000000369	959.33	220	189	81	36	
2	CUST0000000689	328.57	16	116	73	41	

```
df_csv.dtypes
```

```
CUST_CODE              object
TotalSpend            float64
TotalVisits             int64
TotalSKUs               int64
TotalSKUs_10            int64
TotalSKUs_20            int64
TotalSKUs_30            int64
TotalSKUs_40            int64
FirstDate      datetime64[ns]
LastDate       datetime64[ns]
TicketSize            float64
total_days              int64
recency                 int64
dtype: object
```

```
#df_final=df_csv.join(df_csv3,how='left',on='CUST_CODE',c)
```

```
#result = pd.concat([df_csv, df_csv3], axis=1, join="left",on='CUST_CODE')
```

```
merged = pd.merge(df_csv,df_csv3, on=['CUST_CODE'])
```

```
df_csv.shape
```

```
(6100, 13)
```

```
merged.shape
```

```
(78137, 16)
```

```
#df_csv=merged
```

```
df_csv.head()
```

	CUST_CODE	TotalSpend	TotalVisits	TotalSKUs	TotalSKUs_10	TotalSKUs_20	Tot
0	CUST0000000181	2.44	1	1	1	1	
1	CUST0000000369	959.33	220	189	81	36	
2	CUST0000000689	328.57	16	116	73	41	

Cluster customers

```
exp_clu = setup(data=df_csv, ignore_features=['CUST_CODE','FirstDate', 'LastDate'], normalize=True)
```

	Description	Value
0	session_id	3728
1	Original Data	(6100, 13)
2	Missing Values	False
3	Numeric Features	9
4	Categorical Features	1
5	Ordinal Features	False
6	High Cardinality Features	False
7	High Cardinality Method	None
8	Transformed Data	(6100, 18)
9	CPU Jobs	-1
10	Use GPU	False
11	Log Experiment	False
12	Experiment Name	cluster-default-name
13	USI	8620
14	Imputation Type	simple
15	Iterative Imputation Iteration	None
16	Numeric Imputer	mean
17	Iterative Imputation Numeric Model	None
18	Categorical Imputer	mode
19	Iterative Imputation Categorical Model	None
20	Unknown Categoricals Handling	least_frequent
21	Normalize	True
22	Normalize Method	zscore

```
models()
```

	Name	Reference
ID		
kmeans	K-Means Clustering	sklearn.cluster._kmeans.KMeans
ap	Affinity Propagation	sklearn.cluster._affinity_propagation.Affinity...
meanshift	Mean Shift Clustering	sklearn.cluster._mean_shift.MeanShift
sc	Spectral Clustering	sklearn.cluster._spectral.SpectralClustering
hclust	Agglomerative Clustering	sklearn.cluster._agglomerative.AgglomerativeCl

```
get_metrics()
```

	Name	Display Name	Score Function	Sco
ID				
silhouette	Silhouette	Silhouette	<function silhouette_score at 0x7fa001689d40>	make_scorer(silhouette_sco
chs	Calinski-Harabasz	Calinski-Harabasz	<function calinski_harabasz_score at 0x7fa0016...	make_scorer(calinski_harabasz_sco
db	Davies-	Davies-	<function davies_bouldin_score	make_scorer(davies_bouldin_sco

Compare model performance

```
metrics = []
for model in models().index:
    if model in ['meanshift', 'optics']:
        continue
    create_model(model)
    metric_result = pull()
    metric_result['model'] = model
    metrics.append(metric_result)
```

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	-0.0878	290.8768	2.9148	0	0	0

```
cluster_metrics = pd.concat(metrics)
cluster_metrics.set_index("model", inplace=True)
```

```
cluster_metrics.sort_values(by='Silhouette', ascending=False, inplace=True)
cluster_metrics.style.highlight_max(subset=['Silhouette', 'Calinski-Harabasz'], color = 'green', axis = 0).highlight_mi
```

model	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
sc	0.672900	40.611800	0.247600	0	0	0
birch	0.457400	1768.474900	0.855000	0	0	0
kmeans	0.292500	3714.085100	1.172000	0	0	0
hclust	0.281800	3289.183500	1.256000	0	0	0
ap	0.000000	0.000000	0.000000	0	0	0
dbscan	-0.035400	219.369900	1.584800	0	0	0

Spectral Clustering Clustering

```
sc = create_model('sc')
```

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.6729	40.6118	0.2476	0	0	0

```
plot_model(sc)
```

KMeans Clustering

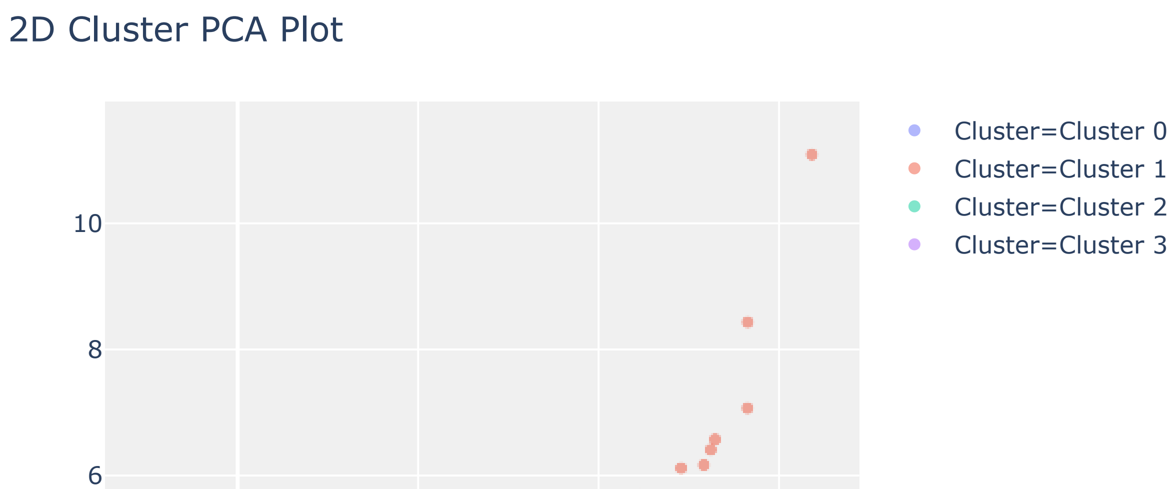
```
kmeans = create_model('kmeans')
```

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.2925	3714.0851	1.172	0	0	0

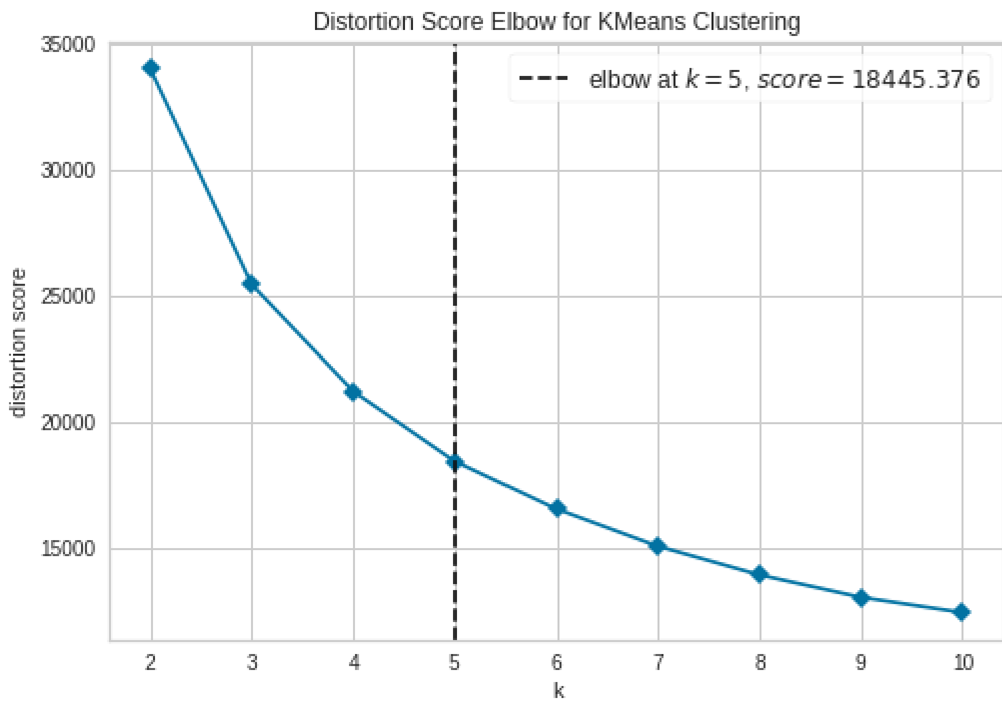
```
print(kmeans)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=4, n_init=10, n_jobs=-1, precompute_distances='deprecated',
       random_state=3728, tol=0.0001, verbose=0)
```

```
plot_model(kmeans)
```

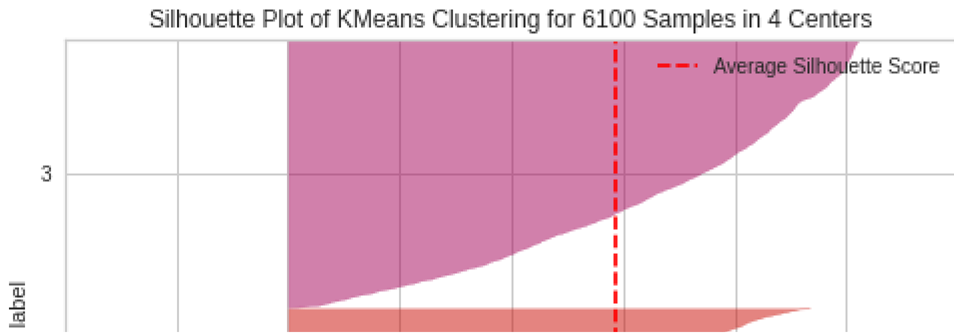


plot_model(kmeans, plot = 'elbow')



plot_model(kmeans, plot = 'silhouette')

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html



Interpret results and plan for actions

```
kmeans_df = assign_model(kmeans)
kmeans_df
```

	CUST_CODE	TotalSpend	TotalVisits	TotalSKUs	TotalSKUs_10	TotalSKUs_20
0	CUST0000000181	2.44	1	1	1	1
1	CUST0000000369	959.33	220	189	81	36
2	CUST0000000689	328.57	16	116	73	41
3	CUST0000000998	5.95	3	4	4	4
4	CUST0000001163	39.74	4	24	21	15
...	...	...	...	...	...	...
6095	CUST0000999593	453.58	30	206	91	50
6096	CUST0000999645	105.11	11	46	36	27

```
final_df= kmeans_df.drop(columns=['CUST_CODE','FirstDate','LastDate'])
member_df = final_df[['Cluster']]
member_df['member_count'] = 1
member_df = member_df.groupby(by=['Cluster']).agg('sum').reset_index()
final_df = final_df.groupby(by=['Cluster']).agg('mean').reset_index()
final_df = final_df.merge(member_df,how='left',on='Cluster')
import seaborn as sns
pink = sns.light_palette('pink', as_cmap = True)
s = final_df.style.background_gradient(cmap=pink)
s
```

	Cluster	TotalSpend	TotalVisits	TotalSKUs	TotalSKUs_10	TotalSKUs_20	TotalSKUs_
0	Cluster 0	39.472479	7.125345	14.867477	11.889564	9.276091	6.4483
1	Cluster 1	2500.713525	173.713115	369.357923	120.882514	53.551913	22.3032
2	Cluster 2	412.896452	35.329372	107.172326	59.922750	34.348896	17.3446

Cluster	Character	Name	Action
Cluster 0	ซื้อสินค้าน้อย มีการเข้ามาดูสินค้าน้อย ไม่ค่อยมีความสนใจกับสินค้าของเรามาก	เพื่อนบ้านที่ห่างไกล	ยิง ads ให้ลูกค้ารู้จักสินค้าเรามากขึ้น , เพิ่ม promotion
Cluster 1	ซื้อสินค้าปริมาณค่อนข้างสูง มีการตอบสนองต่อ promotion ที่ดี	คนสนิทแต่ยังไม่ใช่แฟน	เพิ่ม promotion เพื่อให้การซื้อสูงขึ้น
Cluster 2	ซื้อสินค้าปานกลาง เข้าชมสินค้าปานกลาง	เพื่อนบ้านในหมู่บ้านเดียวกัน	ยิง ads ให้ลูกค้ารู้จักสินค้าเรามากขึ้น , เพิ่ม promotion
Cluster 3	ซื้อสินค้าในปริมาณที่สูง ตอบสนองต่อ promotion ดีมาก	คนที่รู้ใจ	พยายามเสนอ promotion ที่ถูกใจโดยพิจารณา ความชอบของแต่ละบุคคล พยายามรักษาลูกค้า