机器学习课题报告

—　聚类实验

实验名称　LAB03: Clustering

小组编号　第 11 小组

小组成员 1　1852305　包广垠（聚类模型部分）

小组成员 2　1853790　庄镇华（数据分析部分）

小组成员 3　1854205　郑昕瑶（聚类模型部分）

小组成员 4　1852329　赵孟石（数据分析部分）

注：以上排名不分先后

聚类模型报告

一、数据准备

“Facebook Live Sellers in Thailand”属于 UCI 机器学习数据集之一。该数据集包含 7050 个样本，每个数据包含 12 个属性。数据的每一项是一种类型（文本、视频、直播和图片）的 Facebook 帖子，该数据与 Facebook 平台上的实时销售功能有关。对于每个 Facebook 帖子，数据集记录了三个部分：销售的实时信息、发布到 Facebook 的时间、用户参与度指标。其中，参与度指标包括分享、评论和表情符号反应，同时将传统的“喜欢”与最近引入的表情符号反应区分开来，分成了“爱”、“哇”、“哈哈”、“悲伤”和“生气”五种。通过与其他形式的内容（文本、视频、直播和图片）的比较研究以及对 Facebook Live 的季节性统计分析，该数据集可以作为研究客户与 Facebook Live 新销售渠道互动的基础。

数据集可以通过以下链接获得。

https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand

二、数据清洗与初步分析

2.1 数据清洗

打开并查看数据集

```
data = pd.read_csv ('Live.csv', index_col = 0)
data.info()
```

可以看到数据集的特征如下

```
Int64Index: 7050 entries, 1 to 7050
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   status_type       7050 non-null   object
 1   status_published  7050 non-null   object
 2   num_reactions     7050 non-null   int64
 3   num_comments      7050 non-null   int64
 4   num_shares        7050 non-null   int64
 5   num_likes         7050 non-null   int64
 6   num_loves         7050 non-null   int64
 7   num_wows          7050 non-null   int64
 8   num_hahas         7050 non-null   int64
 9   num_sads          7050 non-null   int64
 10  num_angrys        7050 non-null   int64
 11  Column1           0 non-null      float64
 12  Column2           0 non-null      float64
 13  Column3           0 non-null      float64
 14  Column4           0 non-null      float64
```

status_type	status_published	num_reactions	num_comments	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
video	4/22/2018 6:00	529	512	262	432	92	3	1	1	0
photo	4/21/2018 22:45	150	0	0	150	0	0	0	0	0
video	4/21/2018 6:17	227	236	57	204	21	1	1	0	0
photo	4/21/2018 2:29	111	0	0	111	0	0	0	0	0
photo	4/18/2018 3:22	213	0	0	204	9	0	0	0	0
photo	4/18/2018 2:14	217	6	0	211	5	1	0	0	0
video	4/18/2018 0:24	503	614	72	418	70	10	2	0	3
video	4/17/2018 7:42	295	453	53	260	32	1	1	0	1
photo	4/17/2018 3:33	203	1	0	198	5	0	0	0	0
photo	4/11/2018 4:53	170	9	1	167	3	0	0	0	0
photo	4/10/2018 1:01	210	2	3	202	7	1	0	0	0
photo	4/9/2018 2:06	222	4	0	213	5	4	0	0	0

该数据集共有 7050 项，每项数据有 15 个特征。**原始数据存在以下问题：**

- 数据的第 1 列和第 2 列是非数值化的文本数据和时间数据；
- 数据的后四项是空数据，值均为 NAN；
- 数据的 num_loves, num_wows, num_hahas, num_sads, num_angrys 较为稀疏；
- 一些组合特征需要被挖掘；

针对以上问题，进行数据清洗工作，包括对非数值化数据的编码、删除空数据、增加新的数据特征。

status_type 一共有四种取值，故编码为 1~4；对于时间，我们调用 labelEncoder()进行编码；在组合特征构造上，我们构造 new_reaction_count(新反应类型数)去记录用户对于新表情的使用情况，构造了

postive_reactions 和 negative_reactions 去记录用户的两种反应态度，构造了'interesting'去记录用户对于此条动态的感兴趣程度。

```
data_clean = data.dropna(axis=1)
# 非数值数据的处理
labelEncoder = LabelEncoder()
labelEncoder.fit(data_clean['status_type'])
data_clean['status_type'] = labelEncoder.transform(data_clean['status_type'])
labelEncoder.fit(data_clean['status_published'])
data_clean['status_published'] =
labelEncoder.transform(data_clean['status_published'])
# 构造组合特征
data_clean['new_reaction_count'] = data_clean.iloc[:,6:11].sum(axis=1)
data_clean['postive_reactions'] = data_clean.iloc[:,5:8].sum(axis=1)
data_clean['negative_reactions'] = data_clean.iloc[:,8:11].sum(axis=1)
data_clean['interesting'] =
(data_clean.num_comments+data_clean.num_shares)/data_clean.num_reactions
data_clean.replace([np.inf, -np.inf, np.nan], 0.0, inplace=True)
# 保存清洗完的数据
data_clean.to_csv("data_clean.csv", index=True)
```

通过以上处理，完成了数据清洗和预处理操作。

2.2 初步分析

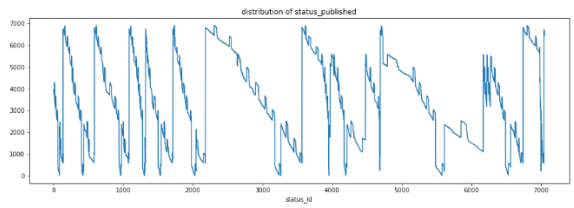
现对得到的预处理过后的数据进行展示，展示数据的组成：

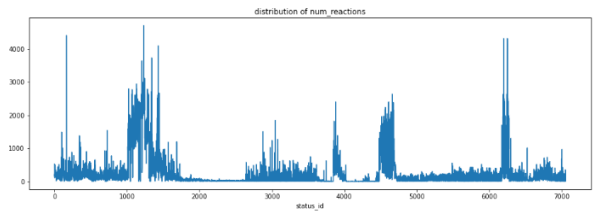
```
#   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   status_type          7050 non-null   int32
 1   status_published     7050 non-null   int32
 2   num_reactions        7050 non-null   int64
 3   num_comments         7050 non-null   int64
 4   num_shares           7050 non-null   int64
 5   num_likes            7050 non-null   int64
 6   num_loves            7050 non-null   int64
 7   num_wows             7050 non-null   int64
 8   num_hahas            7050 non-null   int64
 9   num_sads             7050 non-null   int64
 10  num_angrys           7050 non-null   int64
 11  new_reaction_count   7050 non-null   int64
 12  postive_reactions    7050 non-null   int64
 13  negative_reactions   7050 non-null   int64
 14  interesting          6934 non-null   float64
```

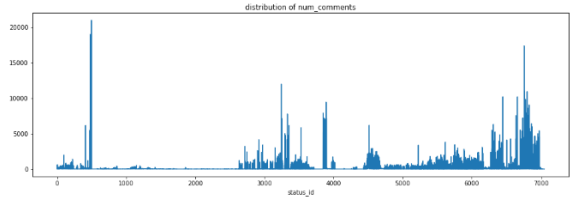
num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys	new_reaction_count	postive_reactions	negative_reactions	interesting
262	432	92	3	1	1	0	97	527	2	1.463138
0	150	0	0	0	0	0	0	150	0	0.000000
57	204	21	1	1	0	0	23	226	1	1.290749
0	111	0	0	0	0	0	0	111	0	0.000000
0	204	9	0	0	0	0	9	213	0	0.000000
0	211	5	1	0	0	0	6	217	0	0.027650
72	418	70	10	2	0	3	85	498	5	1.363817
53	260	32	1	1	0	1	35	293	2	1.715254
0	198	5	0	0	0	0	5	203	0	0.004926
1	167	3	0	0	0	0	3	170	0	0.058824
3	202	7	1	0	0	0	8	210	0	0.023810

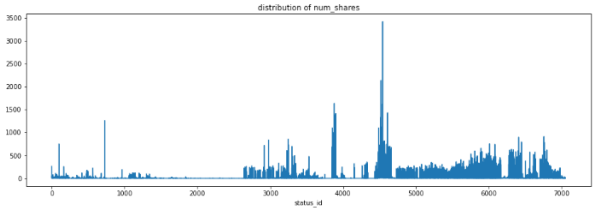
展示各个特征的取值分布：

```
data_clean.xxx(要展示的特征).plot(kind='line', figsize=(16,5))
```









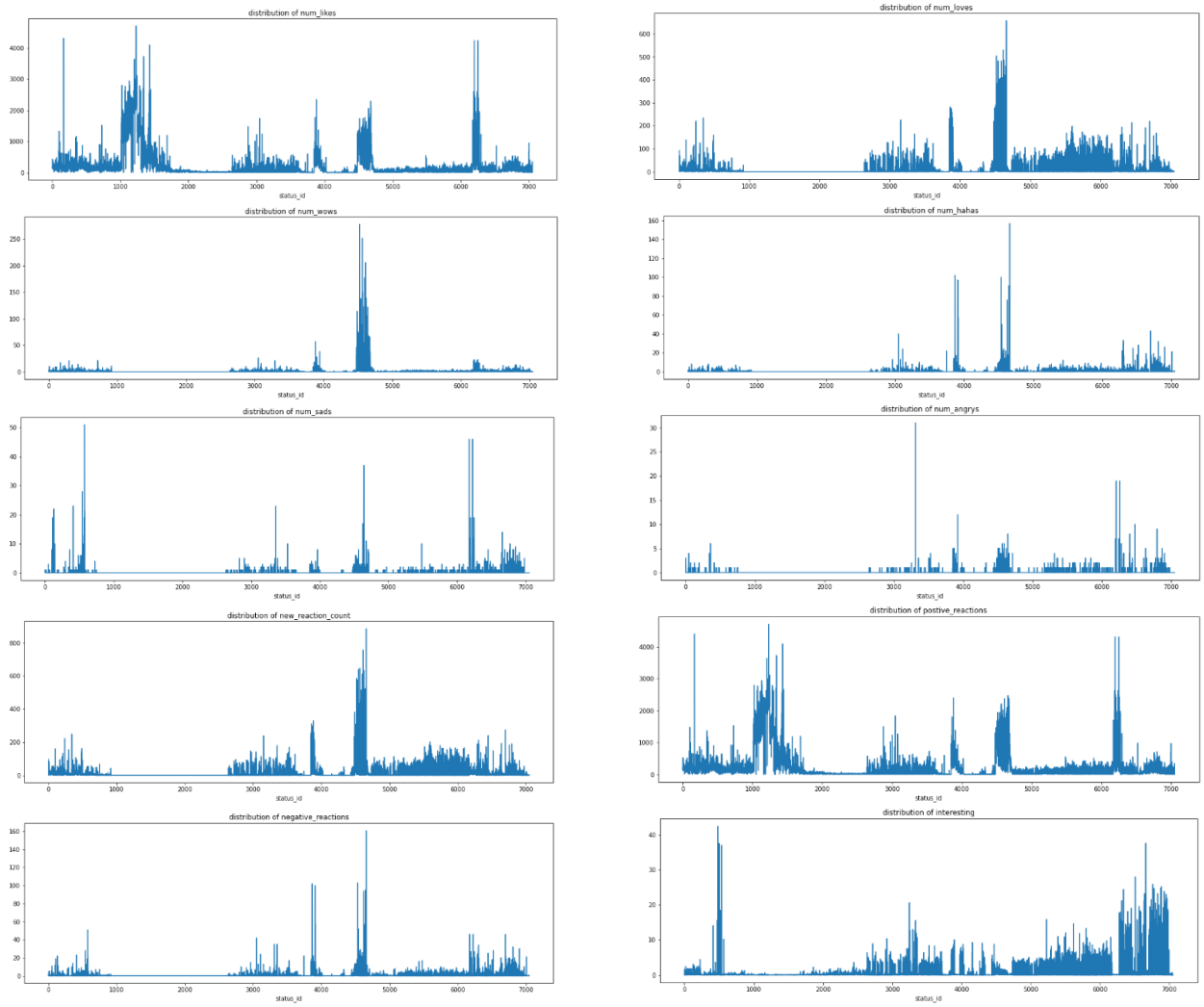
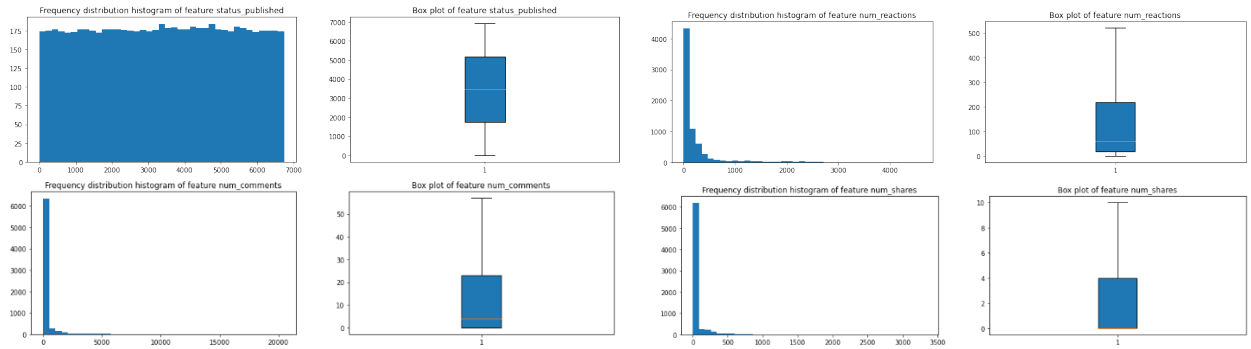


图 2-1 各个特征的取值分布

各个特征的取值的频率分布直方图与箱线图：

```
plt.hist(D_array[:,feature], np.arange(min(D_array[:,feature]),
max(D_array[:,feature]), (max(D_array[:,feature])-
min(D_array[:,feature]))/40).tolist())
plt.title('Frequency distribution histogram of feature ' + names[feature])
plt.boxplot(D_array[:,feature], showfliers = True, sym = '', vert=True,
patch_artist=True)
plt.title('Box plot of feature ' + names[feature])
```



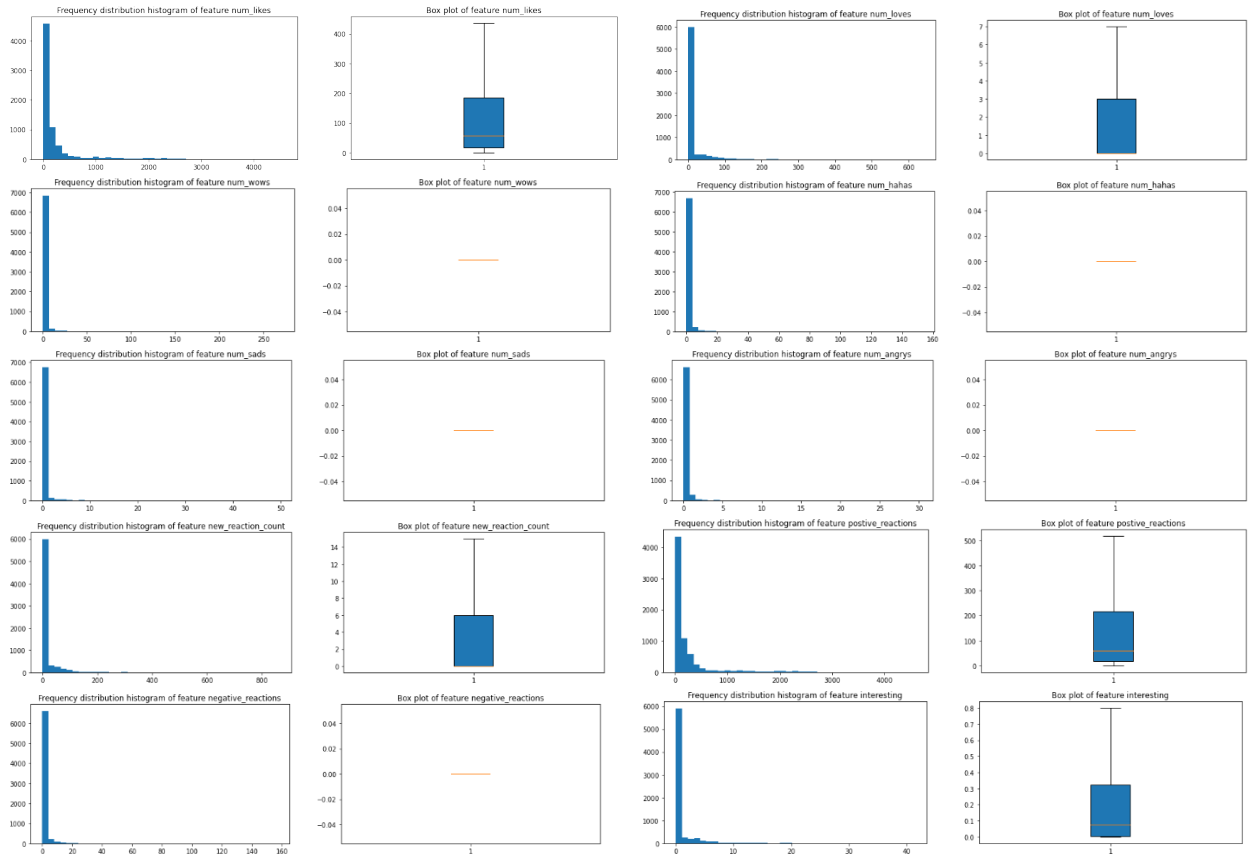


图 2-2 各个特征的取值的频率分布直方图与箱线图

由频率分布直方图和箱线图可以看出，数据在较小值(0~10)是出现的频率较高，在较大值时出现的频率很低，甚至会被当作异常值出现。

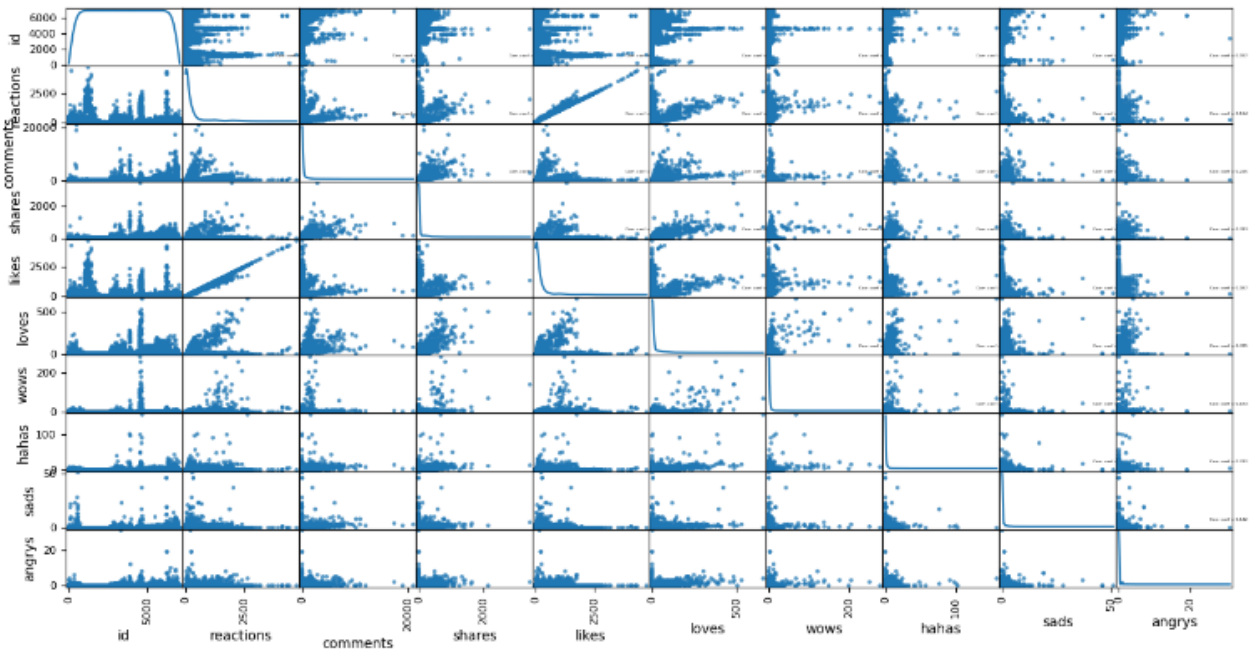


图 2-3 整体散点密度图

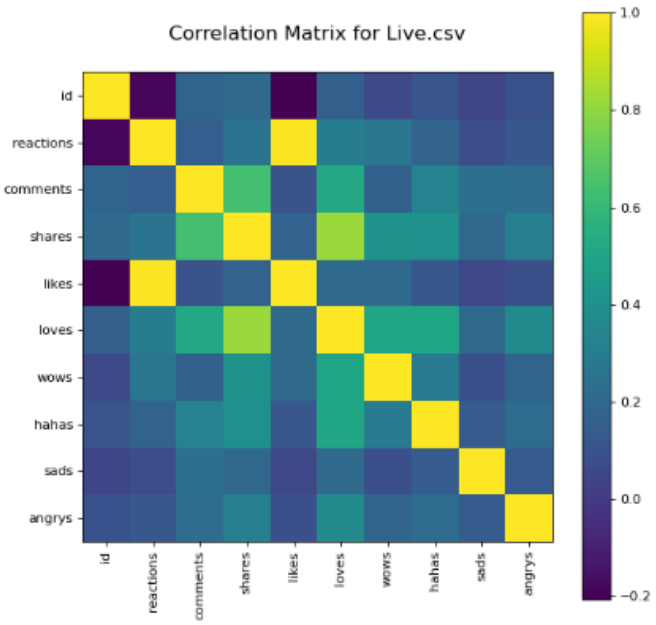
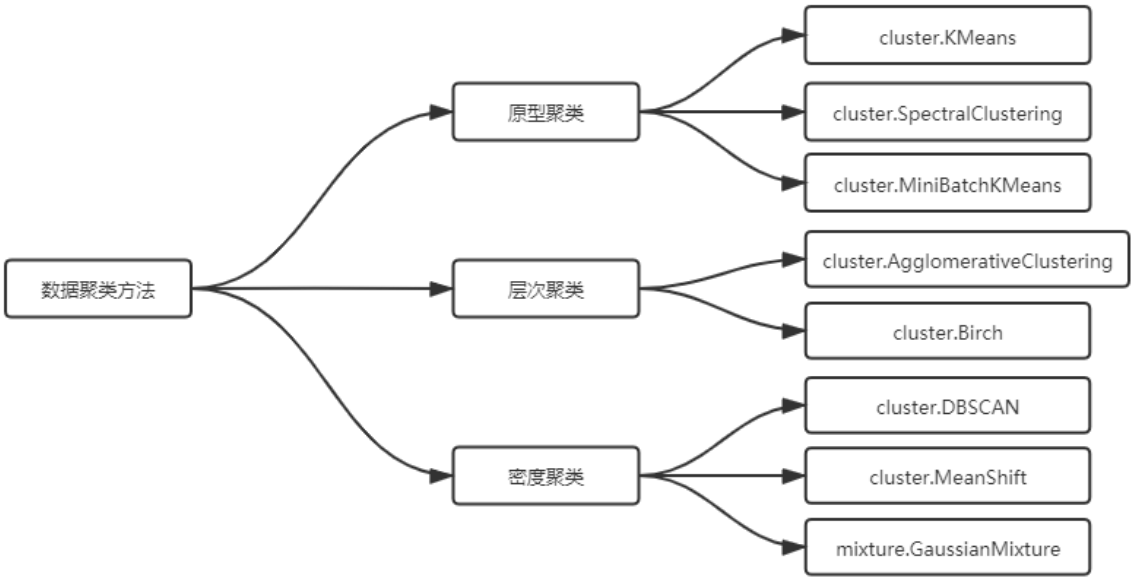


图 2-3 相关矩阵图

三、模型搭建

3.1 模型概要

在聚类模型的选择上，我们选取了 sklearn 中的多种类别的多种模型聚类模型。



3.2 模型详解

算法名称	算法概要	算法步骤

KMeans	KMeans 是我们最常用的基于欧式距离的聚类算法，其认为两个目标的距离越近，相似度越大。	(1)选择初始化的 k 个样本作为初始聚类中心 $a = a_1, a_2, \dots, a_k$； (2)针对数据集中每个样本 x_i 计算它到 k 个聚类中心的距离并将其分到距离最小的聚类中心所对应的类中； (3)针对每个类别 a_j，重新计算它的聚类中心 $a_i = \frac{1}{\lvert c_i \rvert}\sum_{x \in c_i} x$（即属于该类的所有样本的质心）； (4)重复上面(2)(3)两步操作，直到达到某个中止条件（迭代次数、最小误差变化等）。
SpectralClustering	谱聚类和一般的原型聚类方法相比有着不少优点，它只需要数据之间的相似度矩阵，对于不规则的误差数据不是那么敏感，计算复杂度比 KMeans 要小。	(1)根据数据构造一个 Graph，它的每一个节点对应一个数据点，将相似的点连接起来，并且边的权重用于表示数据之间的相似度。把这个 Graph 用邻接矩阵的形式表示出来，记为 W； (2)把 Graph 的每一列元素加起来得到 N 个数，把它们放在对角线上（其他地方都是零），组成一个 $N \times N$ 的矩阵，记为 D。并令 $L = D - W$； (3)求出 L 的前 k 个特征值以及对应的特征向量； (4)把这 k 个特征向量排列在一起组成一个 $N \times k$ 的矩阵，将其中每一行看作 k 维空间中的一个向量，并使用 KMeans 算法进行聚类。聚类的结果中每一行所属的类别就是原来 Graph 中的节点亦即最初的 N 个数据点分别所属的类别。
MiniBatchKMeans	MiniBatchKMeans 算法是 KMeans 算法的一种优化变种，采用小规模数据子集减少计算时间，同时优化目标函数，该算法可以加快收敛速度，但结果会略差于标准 KMeans 算法。	(1)抽取部分数据集，使用 KMeans 算法构建出 k 个聚簇点的模型； (2)继续抽取剩余训练数据，加到模型中，分配给最近的聚簇中心点； (3)更新聚簇的中心点； (4)循环(2)(3)操作，直到中心点稳定或者到达迭代次数上限。
AgglomerativeClustering	AgglomerativeClustering 是一种自下而上的层次聚类方法，聚类开始将每个数据点当作独立的一类，然后度量各个类别之间的距离(单链接、复链接、类平均距离) 进行聚类。	(1)将每个样本都作为一个簇 (2)计算聚类簇之间的距离，找出距离最近的两个簇，将这两个簇合并，直到迭代到指定类别数。
		(1)将所有的样本依次读入，在内存中建立一颗 CF Tree；

Birch	Birch 聚类是利用层次方法的平衡迭代规约和的聚类，它是用层次方法来聚类和规约数据。它运行速度很快，只需要单遍扫描数据集就能进行聚类。Birch 算法的主要过程，就是建立 CF Tree 的过程。	(2)第一步建立的 CF Tree 进行筛选，去除一些异常 CF 节点； (3)利用其它的一些聚类算法比如 K-Means 对所有的 CF 元组进行聚类，得到一颗比较好的 CF Tree； (4)利用(3)生成的 CF Tree 的所有 CF 节点的质心，作为初始质心点，对所有的样本点按距离远近进行聚类。
DBSCAN	DBSCAN 是一个比较有代表性的基于密度的聚类算法，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。另外，DBSCAN 不需要事先知道要形成的簇类的数量，对于数据库中样本的顺序不敏感。	(1)根据 eps 邻域和密度阈值 MinPts，判断一个点是核心点、边界点或者离群点，并将离群点删除； (2)如果核心点之间的距离小于 MinPts，就将两个核心点连接在一起，这样就形成了若干组簇； (3)将边界点分配到距离它最近的核心点范围内，形成最终的聚类结果。
MeanShift	Meanshift 算法是利用概率密度的梯度爬升来寻找局部最优的一种聚类算法。	
GaussianMixture	GaussianMixture 聚类算法与 KMeans 聚类类似，只不过 GaussianMixture 能发现椭圆形的聚簇。	

不同聚类算法在生成数据集上的聚类表现(算法聚类特性)比较：

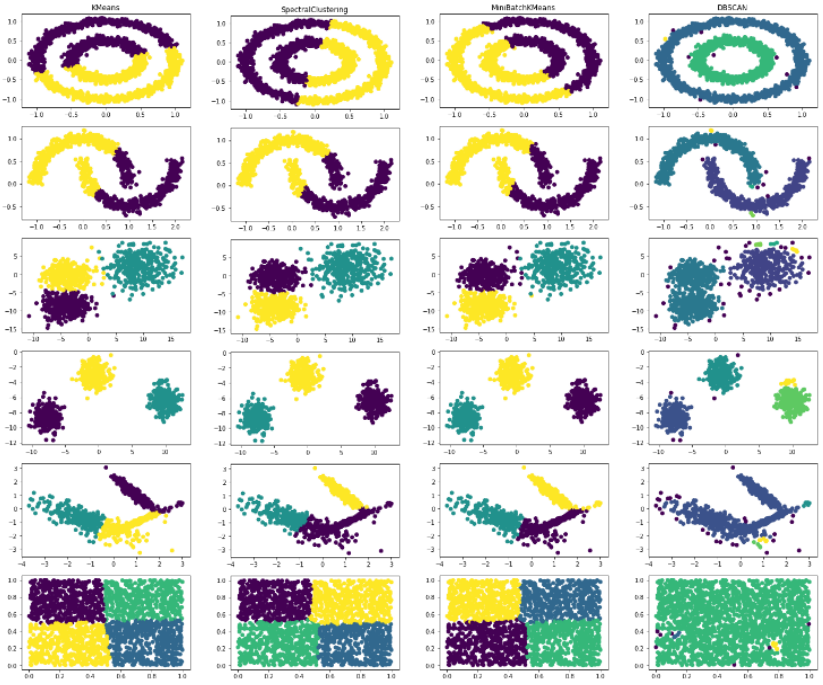


图 3-1 K-Means、Spectral Clustering、Mini Batch K-Means、DBSCAN 聚类效果

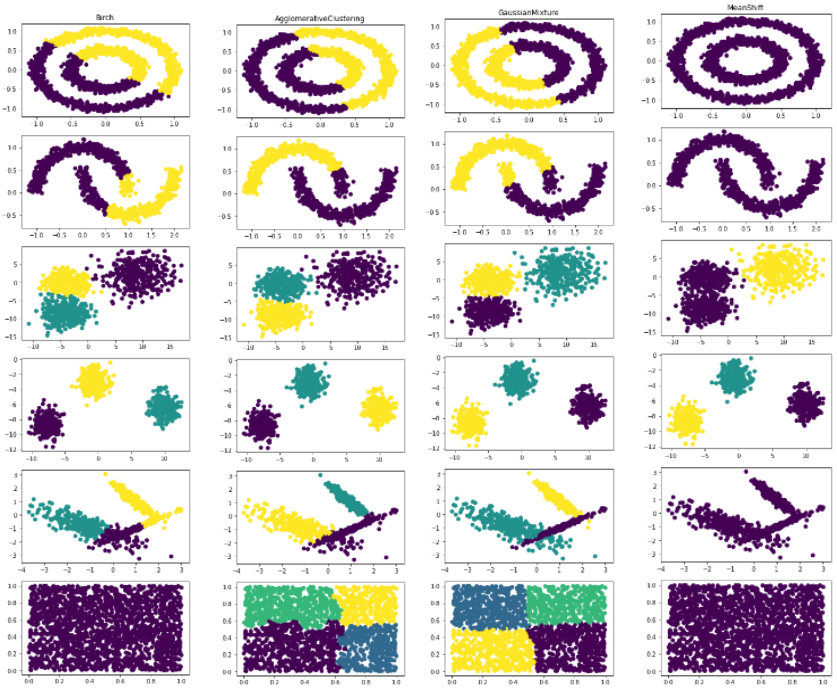


图 3-2 Birch、Agglomerative Clustering、Gaussian Mixture、Mean Shift 聚类效果

四、模型训练与结果可视化

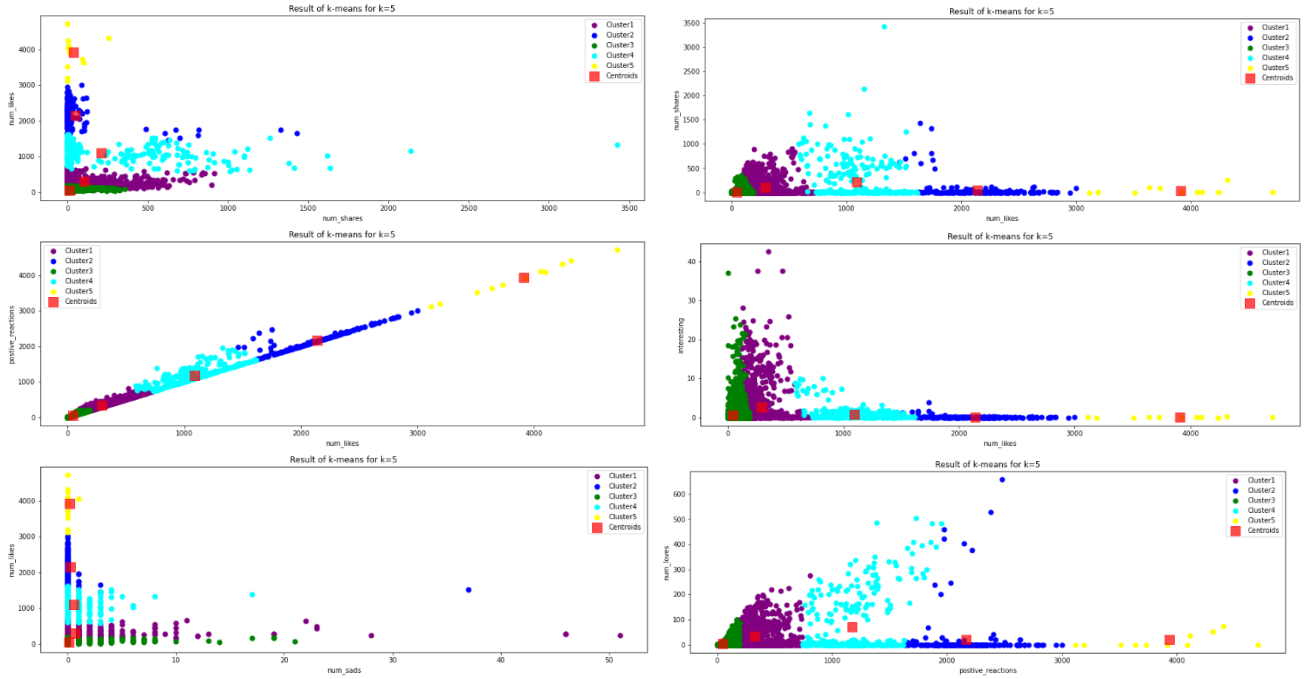
本小节内容聚类标准为原始分类（即 video、photo、link、status），自己设计的分类标准见第二部分——数据分析（超链接）。

4.1 K-Means 模型训练与可视化

选择模型的部分特征，使用 sklearn.cluster 中的模型进行训练，进行结果的可视化。（以 KMeans 为例）

```
# 使用 k-means 方法,k 选
km = KMeans(n_clusters=5, init='kmeans++', max_iter=300, n_init=10, rando
m_state=0)
result = km.fit_predict(X)
# 进行结果可视化
plt.figure(figsize=(16, 5))
plt.scatter(X[result==0,i], X[result==0,j], s=50, c=result, label='Cluste
r1')
plt.scatter(X[result==1,i], X[result==1,j], s=50, c=result, label='Cluste
r2')
...
plt.scatter(km4.cluster_centers_[:,i], km4.cluster_centers_[:,j], s=200,
marker='s', c='red', alpha=0.7, label='Centroids')
...
plt.show()
```

根据聚类的结果绘制在特定特征维度上的聚类结果图，如下：

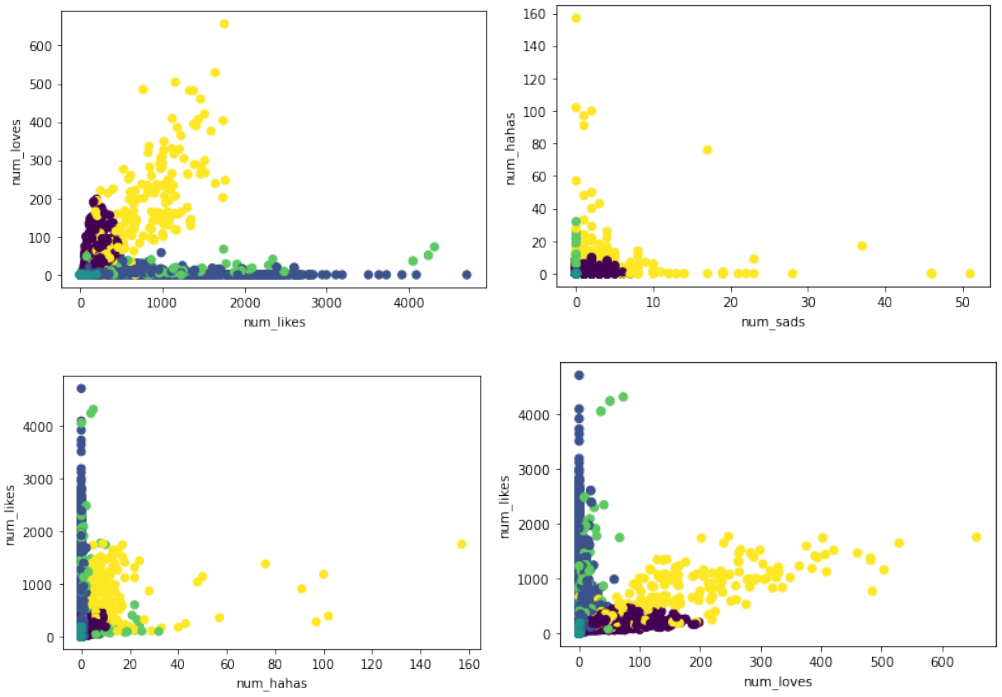


为了分析 KMeans 的聚类效果，将进一步在 5.3 小节（超链接）使用 Silhouette Score 评价指标来进行模型的评价和优化。

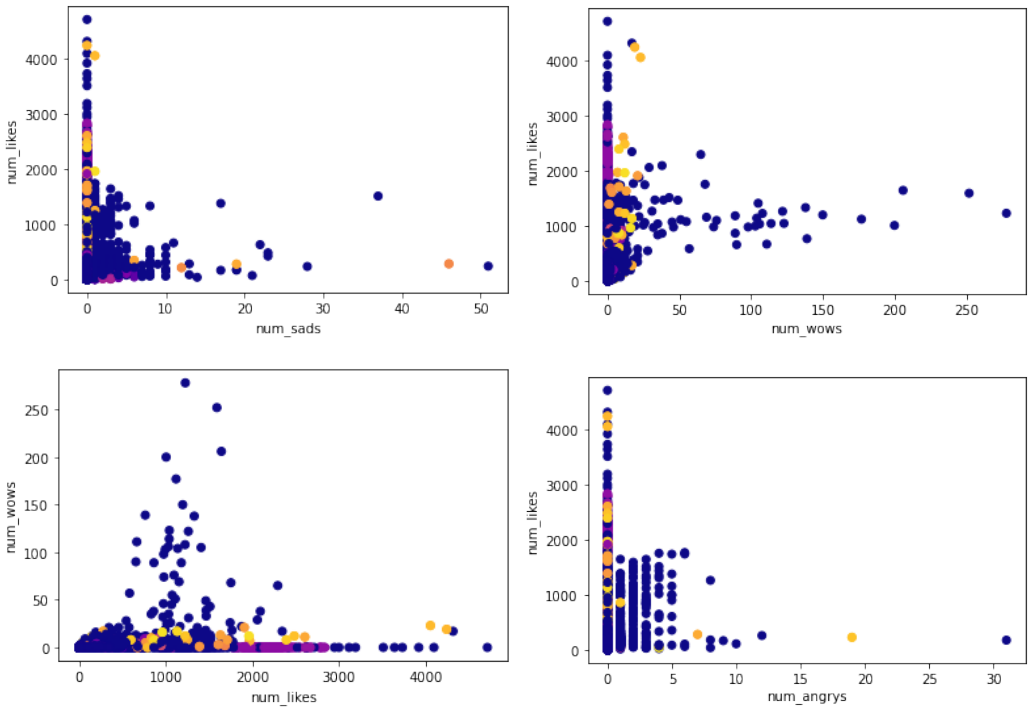
4.2 其他部分模型训练与可视化

由于篇幅限制原因，仅挑选以下几种聚类方法进行分析，图中不同颜色代表原始类别（video、photo 等），可以看出没优化之前聚类效果不是特别理想。

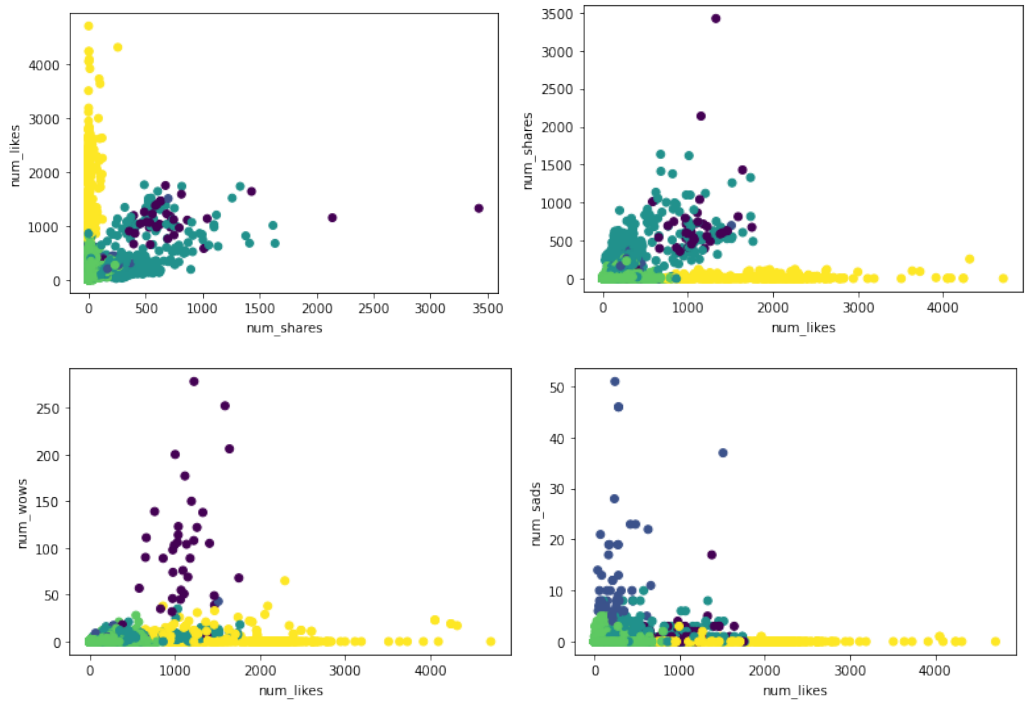
高斯混合模型聚类



DBSCAN 密度聚类



Agglomerative 层次聚类



五、模型优化

5.1 评价指标

聚类任务有多种评价指标，例如 Jaccard 系数、FM 系数、Rand 指数、DB 指数、Dunn 指数等。在这些聚类评价指标中，有的需要样本的类别标签，有的不需要样本的类别标签。

在我们本次的聚类任务中，数据集并没有提供可以直接使用的类别标签，因此需要使用“内部度量”标准去评价聚类结果的好坏。

我们采用三种内部度量指标：轮廓系数，CH 得分，DB 指数。

轮廓系数（Silhouette Coefficient）

轮廓系数（Silhouette Coefficient），是聚类效果好坏的一种评价方式，它结合内聚度和分离度两种因素，可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。

具体过程如下：

(1)计算样本i到同簇其他样本的平均距离a_i。a_i越小，说明样本i越应该被聚类到该簇。将a_i称为样本i的簇内不相似度。某一个簇C中所有样本的a_i均值称为簇C的簇不相似度。

(2)计算样本i到其他某簇C_j的所有样本的平均距离b_{ij}，称为样本i与簇C_j的不相似度。定义为样本i的簇间不相似度:$b_i = \min\{b_{i1}, b_{i2}, \dots, b_{ik}\}$，即某一个样本的簇间不相似度为该样本到所有其他簇的所有样本的平均距离中最小的那一个。b_i越大，说明样本i越不属于其他簇。

(3)根据样本i的簇内不相似度a_i和簇间不相似度b_i，定义某一个样本样本i的轮廓系数：

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

(4)判断：s_i接近 1，则说明样本i聚类合理；s_i接近-1，则说明样本i更应该分类到另外的簇；若s_i近似为 0，则说明样本i在两个簇的边界上。

(5)所有样本的s_i的均值称为聚类结果的轮廓系数，定义为 S。聚类结果的轮廓系数的取值在[-1,1]之间。观测值越大越好。

CH 分数（Calinski harabasz Score）

Calinski-Harabasz 分数值 ss 的数学计算公式是

$$ss = \frac{T_r(B_k)}{T_r(W_k)} \times \frac{N-k}{k-1}$$

其中，Bk 称之为簇间色散平均值，Wk 称之为群内色散之间，计算方式为

$$B_k = \sum_q n_q (c_q - c)(c_q - c)^T$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

ss 分数值越大越好。

DBI 指数（Davies-Bouldin Index）

DBI 又称为分类适确性指标，它度量每个簇类最大相似度的均值。

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{avg(C_i)+avg(C_j)}{d_{cen}(u_i,u_j)}\right)$$

DBI 指数越小越好。

5.2 特征提取

特征提取从初始的一组测量数据开始，并建立旨在提供信息和非冗余的派生值（特征），从而促进后续的学习和泛化步骤，并且在某些情况下带来更好的可解释性。特征提取与降维有关。特征的好坏对泛化能力有至关重要的影响。我们使用了两种特征提取方法：PCA 与 FA。

主成分分析（Principal Component Analysis）

主成分分析是一种统计方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。在用统计分析方法研究多变量的课题时，变量个数太多就会增加课题的复杂性。人们自然希望变量个数较少而得到的信息较多。在很多情形，变量之间是有一定的相关关系的，当两个变量之间有一定相关关系时，可以解释为这两个变量反映此课题的信息有一定的重叠。主成分分析是对于原先提出的所有变量，将重复的变量（关系紧密的变量）删去多余，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

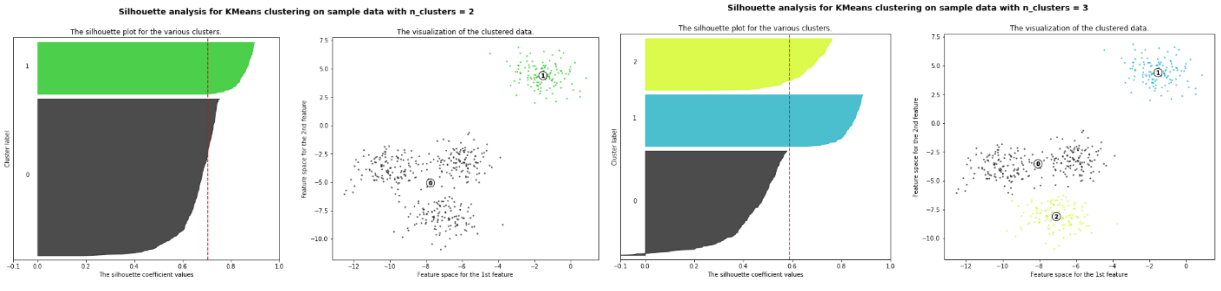
因子分析（Factor Analysis）

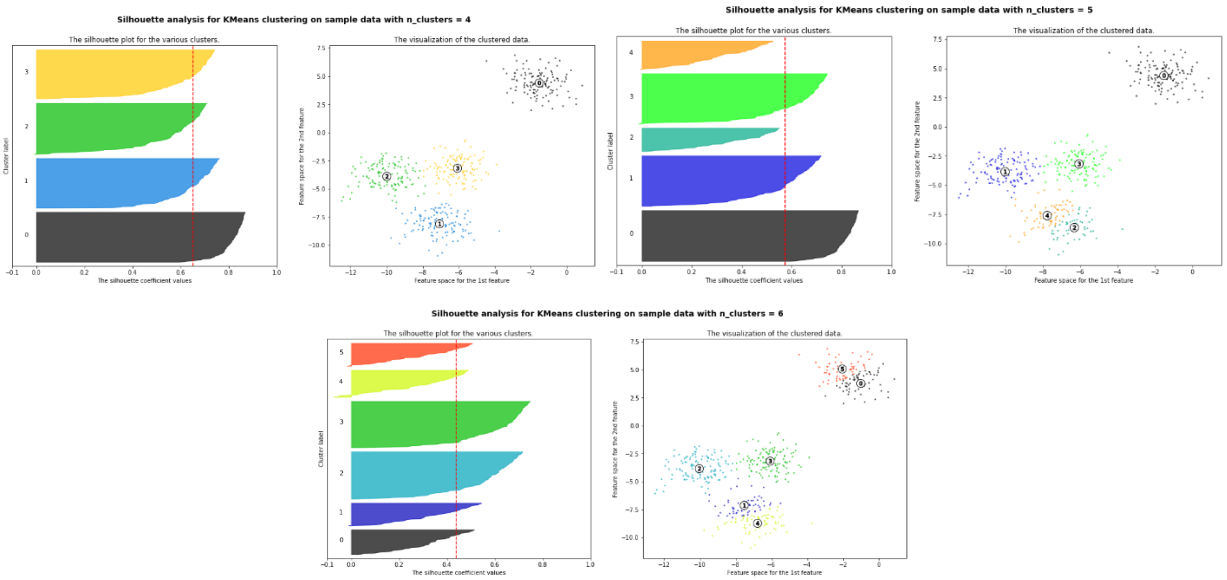
因子分析是指研究从变量群中提取共性因子的统计技术。因子分析的主要目的是用来描述隐藏在一组测量到的变量中的一些更基本的，但又无法直接测量到的隐性变量。这些变量无法直接测量。可以直接测量的可能只是它所反映的一个表征，或者是它的一部分。表征是由这个隐性变量直接决定的。隐性变量是因，而表征是果。因子分析就是从表征出发寻找原因的过程。

5.3 优化过程

以 KMeans 为例，以 PCA 为特提取方式，以 Silhouette Score 为评价指标，在自己生成的数据集上，进行模型的优化。

下图展示了在聚类类别数为 2 ~ 6 时，Silhouette 分析的结果，左边是 Silhouette 得分不断优化提升的过程，右边是生成的数据集的数据分布。





在逐步调整 KMeans 参数的过程中，可以看到各个聚类类别的 ss 得分总体上呈现上升趋势，这表明模型的聚类表现在不断的优化。

六、模型比较

这一部分将使用数据集中 num_comments, num_shares, num_likes, um_loves ,num_wows, num_hahas, num_sads, num_angrys 这 8 个特征进行聚类，目标类别为 5 类，意在对该帖子的热度和受欢迎程度进行区分，分为 5 个不同等级。

- 使用了三种聚类内部评价指标，分别是轮廓系数、CH 得分、DB 指数。
- 使用了两种特征提取方法，分别是 PCA 主成分分析、FA 因子分析。
- 使用了前述的八种聚类模型，比较了相同情况下不同模型的性能差异。

6.1 结果展示

- **轮廓系数**

Silhouette Score

算法	直接使用特征	PCA(n=2)	FA(n=2)	PCA(n=4)	FA(n=4)
K-means	**0.81**	0.71	0.74	0.68	0.76
Mini Batch K-Means	**0.74**	0.70	0.65	0.51	0.71
Agglomerative Clustering	0.68	0.71	**0.74**	0.65	0.69
Birch	**0.80**	0.70	0.84	-0.43	0.73
DBSCAN	-0.62	-0.60	**0.66**	-0.70	-0.22
Mean Shift	**0.70**	0.08	0.07	-0.61	0.43
Gaussian Mixture	**0.38**	0.17	0.01	0.30	-0.33

- **CH 得分**

Calinski Harabasz Score					
算法	直接使用特征	PCA(n=2)	FA(n=2)	PCA(n=4)	FA(n=4)
K-means	**10173.53**	2553.61	6613.68	1915.11	4001.59
Mini Batch K-Means	**7475.21**	2453.88	6378.83	2152.79	4806.7
Agglomerative Clustering	2892.67	2600.75	**6149.86**	1942.21	4026.79
Birch	**9037.45**	1432.02	3291.35	773.85	2548.71
DBSCAN	179.79	329.67	**650.81**	202.31	403.73
Mean Shift	2184.4	598.39	**3463.22**	443.08	749.32
Gaussian Mixture	**1292.3**	1024.54	1003.98	1172.31	732.62

- **DB 指数**

Davies Bouldin Score					
算法	直接使用特征	PCA(n=2)	FA(n=2)	PCA(n=4)	FA(n=4)
K-means	**0.60**	1.12	0.74	1.82	1.72
Mini Batch K-Means	**0.68**	1.39	0.69	1.40	0.96
Agglomerative Clustering	1.00	1.06	**0.75**	5.93	0.84
Birch	0.63	1.46	**0.34**	2.53	0.92
DBSCAN	6.43	4.80	1.84	3.20	**1.18**
Mean Shift	**0.56**	4.47	0.57	3.38	1.68
Gaussian Mixture	1.46	**1.31**	1.35	1.81	3.40

从三个评价指标下的观测值可以看出：

对于**特征提取**而言，**直接使用特征**和进行 **n=2 的因子分析**作为特征提取的聚类效果较好；

对于**聚类方法**而言，前两个指标都表明 **K-Means** 效果最好，第三个指标表明 **Mini Batch K-Means** 效果最好，该方法其实是 K-Means 的变形，故总的来说 K-Means 效果最好。

数据分析报告

一、行业分析

本次实验的数据是从 2012 年 3 月至 2018 年 6 月期间从 10 家泰国时装和化妆品零售商的 Facebook 页面中提取的。该数据集通过 Facebook API 收集，对于每个 Facebook 帖子，数据集记录了由此产生的参与

度指标，包括分享、评论和表情符号反应，并且将传统的“喜欢”与最近引入的表情符号反应区分开来。该数据集可以作为研究客户与 Facebook Live 新销售渠道互动的基础。

1.1 泰国直播销售概况

目前，泰国直播已经变得非常流行了，FackBook、YouTube、Tiktok 上已经陆续开始，泰国最大的两家电商巨头 Lazada 和 Shopee 也已经开设直播频道。以泰国 Lazada 直播频道为例：

- ✧ 其在 2018 年 11 月上线；
- ✧ 2019 年双 11，Lazada Super Show 购物狂欢夜直播晚会，观看人次 1300 万+；
- ✧ 2020 年 4 月期间美妆品牌 Shiseido 直播观看人数累计 9 万人，GMV 增长 40 倍，Lazada 独家特价爆款 30 分钟内售罄；60%用户为品牌商和其他商家，其余为个人用户；
- ✧ 总体而言，LazLive 活动类型多样（分销、游戏、电竞、音乐会），合作渠道广泛（MCN、B2B、政府等）内容生产和带货促销同步进行。

虽然泰国直播很火爆，但依然存在很多问题，例如：

- ✧ 网红机构模式未成熟，大多数网红带货都是一次性的商业活动；
- ✧ 电商直播集中在 Shopee 和 Lazada，其他小电商平台则只能借助 Facebook、YouTube 等渠道进行直播服务；
- ✧ 用户更广泛的社交媒体与电商渠道未完全打通、无法在社交平台直接购买、因此购物转化的能力一般。

1.2 直播销售商店特点

由上文分析可以得到泰国直播销售商店的一些信息，方便我们针对具体的数据进行聚类分析。

- ✧ 泰国直播销售十分火爆，观看参与人次较多，因此可以根据流量对商店进行聚类，评选最流行商店、较流行商店和一般流行商店；
- ✧ 数据集记录了观众的参与度指标，包括分享、评论和表情符号反应，并且将传统的“喜欢”与最近引入的表情符号反应区分开来，因此我们可以考虑对商店的销售手段的传统和现代程度进行聚类；
- ✧ 一般直播销售都有网红参与，但泰国的网红机构模式并未成熟，大多数网红带货都只是一次性的商业活动，因此，商店的信誉值得关注，而信誉可以通过分享数体现，因此可以考虑对商店信誉进行聚类分析；
- ✧ 数据集主要收集泰国时装和化妆品零售商的资料，考虑这些物品的受众一般为女性或者较为注意形象的男性，因此可以考虑对商店的主要受众性别进行聚类；
- ✧ 最后，上文提到，泰国直播销售市场由国际品牌、本土品牌共同占据，而两者也存在一定的差异，因此我们可以考虑对商店的国际性和本土性进行聚类。

二、数据分析

主要通过 5 种聚类标准对原始泰国 FaceBook 直播销售数据进行分析，详情如下：

2.1 聚类分析

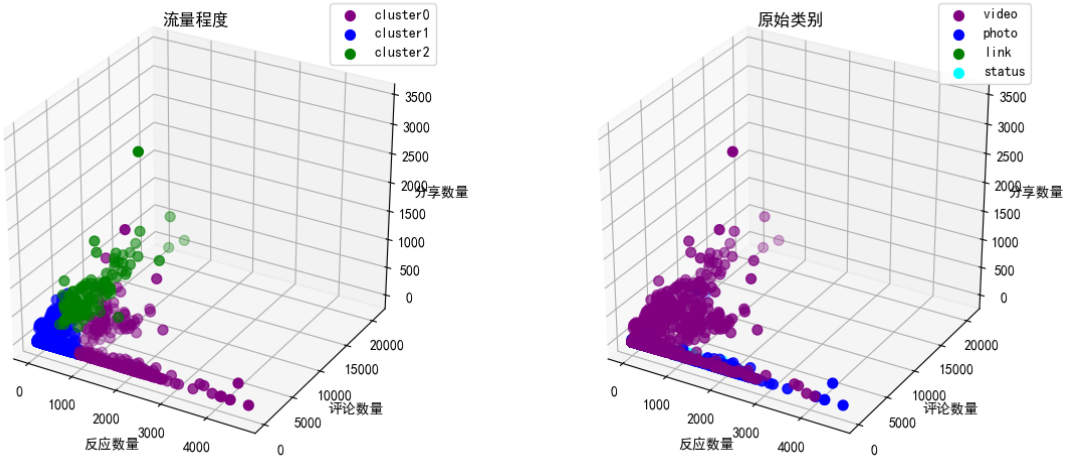
2.1.1 流行程度

直播销售，质量取胜还是流量取胜？这个问题较难回答，但不可否认的是，流量在直播销售中的巨大价值。不管商品质量怎么样，只要有流量，就有更大的可能被买家看到，就多了一些卖出去的可能。因此，流量这个指标对直播销售数据的分析有重要的意义。

这里所指的流量，不仅仅是指好评，也可以是差评，只要能火，无论好坏，因此不需要对评论种类区分，也不需要对表情代表的褒贬义进行区分。统计的仅仅是流量和人气。

纵观数据集中的 12 个属性，我们经过精心挑选，特地选出 3 个相关属性进行分析，它们分别是：num_comment 评论数量，num_sahre 分享数量，num_reaction 反应数量。针对这三个属性对原始数据进行三个类别的聚类分析，分别评选出最具人气商店、较具人气商店和一般人气商店。

聚类结果如下图所示：



其中，左图表示依据流量程度对原始商店进行聚类的结果，右图则是原始类别的分布图（即原始数据中视频、图片、链接的帖子分类）。

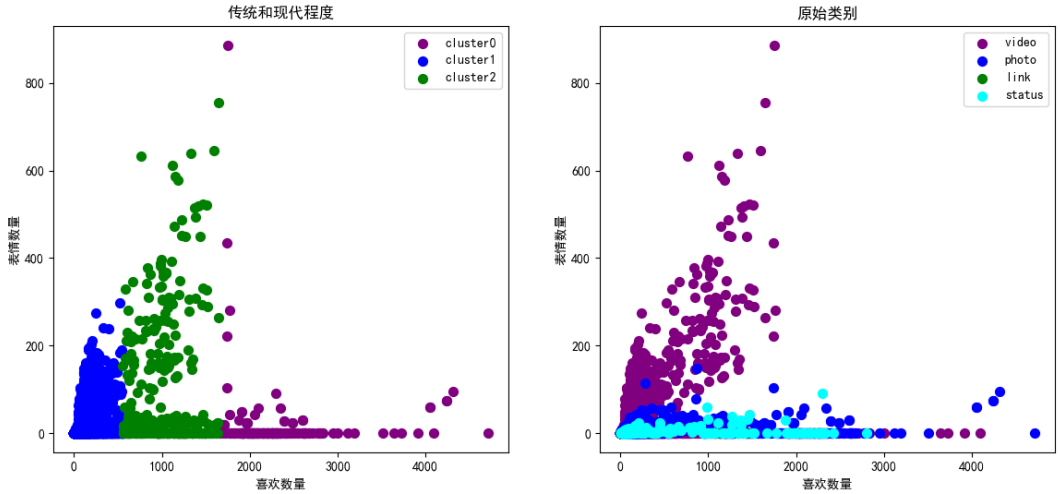
观察可以发现，左图蓝色 cluster1 代表一般流行商店，其反应数量、评论数量、分享数量都较低；绿色 cluster2 代表较为流行商店，其反应数量较少，但评论数量较多、分享数量尚可；紫色 cluster0 代表最流行商店，其反应数量、分享数量都较多，并且评论数量较可。与右图对比，还可以发现，依据人气评判和原始依据视频、图片等类别评判关联并不大，二者没什么必然联系。

2.1.2 销售手段的传统和现代程度

直播销售引入了新的交互手段——表情，这个设计原本是为了更加吸引受众，毫无疑问，这种革新式的方式都会吸人眼球，也会想让人研究研究到底原本的设想是否成立，即到底这种新交互手段能不能吸引更多的客户，给商店带来更多的利益。

因而，我们也根据销售手段的传统和现代程度对原始商店进行了聚类，主要分为三类，即传统方法一类、现代方法一类以及中庸一类，我们这里主要根据传统的“喜欢”与最近引入的表情符号这两种属性进行聚类分析，评选出最终的传统和现代程度销售手段商店。

聚类结果如下图所示：



其中，左图表示依据传统和现代程度对原始商店进行聚类的结果，右图则是原始类别的分布图（即原始数据中视频、图片、链接的帖子分类）。

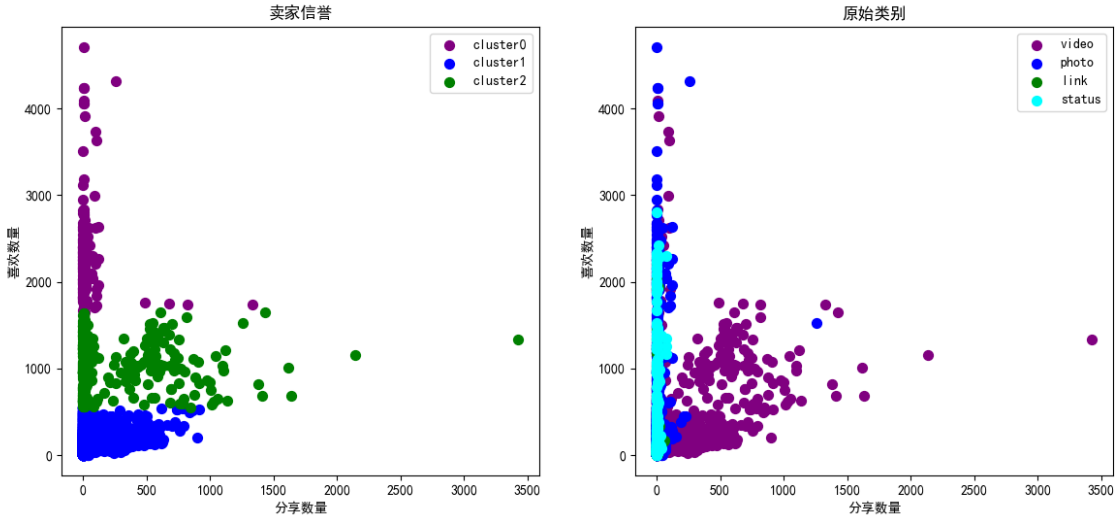
观察可以发现，左图绿色 cluster2 代表现代销售方法商店，其表情数量较高、传统“喜欢”适中；紫色 cluster0 代表传统销售方法商店，其传统“喜欢”数量较高，但表情数量较少；蓝色 cluster1 代表中庸销售方法商店。与右图对比，还可以发现，大部分 video 视频类型的帖子销售方法都属于现代一类，而 photo 图片帖子销售方法大部分都属于传统一类，这也是较为符合常识的。

2.1.3 信誉程度

在当前电子商务法律法规不完善和监管不到位的情况下，信誉成为维持网络交易秩序最重要的机制之一，而信誉价值的存在则是信誉机制发挥作用的基础。信誉价值是指，在同等条件下信誉好的企业能比信誉差的企业卖出更高的价格，即信誉溢价。信誉溢价能够激励企业在长期内维持信誉，而不是试图通过降低质量来获得短期内的收益，这种激励作用在信息不对称的网络交易环境下更为重要。

因此毫无疑问，信誉是该类数据集都应该分析的重要指标。我们精心选取了 num_share 分享数量和 num_likes 喜爱数量这两个属性对原始数据进行了分析，聚类除了三类商店：信誉最佳商店、信誉良好商店和信誉一般商店。

聚类结果如下图所示：



其中，左图表示依据信誉程度对原始商店进行聚类的结果，右图则是原始类别的分布图（即原始数据中视频、图片、链接的帖子分类）。

观察可以发现，左图紫色 cluster0 代表信誉最佳商店，其分享数量不多但喜欢数量很多，颇有酒香不怕巷子深之风度；绿色 cluster2 代表信誉良好商店，其分享数量较多，但喜欢数量不如 cluster0，可能只注重前期宣传，实际质量并不怎么样，所以喜欢数量不多，但考虑到其分享数量较多，定为信誉良好商店；蓝色 cluster1 喜欢数量和分享数量都不多，定位信誉一般商店。与右图对比，还可以发现，大部分 photo 图片类型的帖子都属于信誉最佳一类，可能 photo 属于传统一类，年份较久，较得客户信任，而大部分 video 视频帖子分享数量都很多，可能是新开的商店或商品，比较注重宣发，这也是较为符合常识的。

2.1.4 受众性别

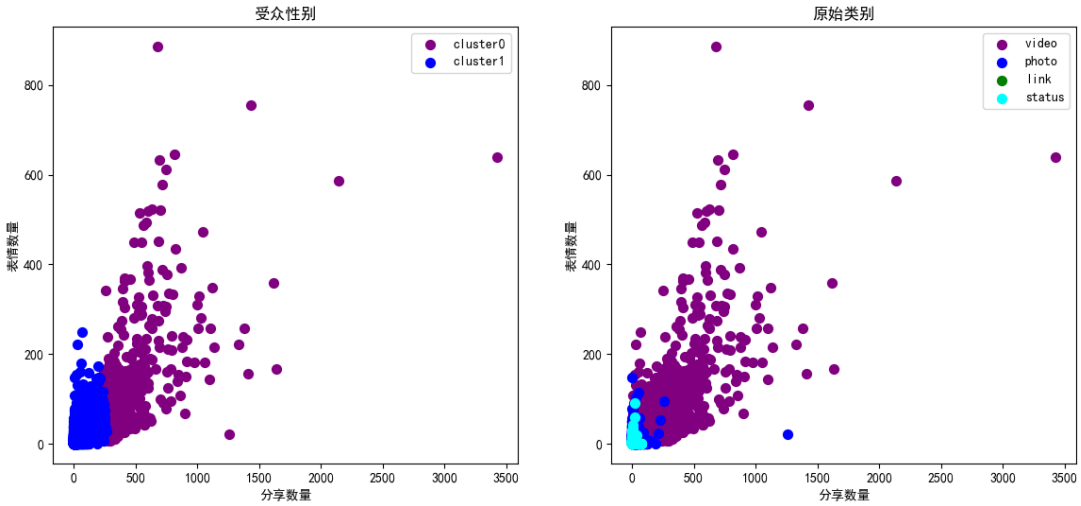
本次数据都是关于化妆品和时装的，并且在经历二十世纪九十年代的经济危机之后，泰国经济已复苏且正迅速崛起，消费者生活水平的提高，导致其需求也来越复杂。作为国际旅游城市之一的泰国，发达的旅游业为为泰国的国际和国内化妆品品牌带来了新的发展机遇。

与亚洲其他国家一样，泰国女性对肌肤美白的需求很大，有很多认为拥有美白肌肤不仅有助于她们寻觅到更好的丈夫，也是社会地位的象征，因此，越来越多的美白产品及多种功能于一身，如保湿、抗皱等。

此外，随着社会的进步，越来越多的泰国男士进入职场，他们开始注重自己的形象，所以泰国男性化妆品市场发展势头也很强劲。男性为泰国个人护理市场的增长带来了强大动力。

根据以上的背景信息，我们知道分析商店受众的性别也是一件有意义和有趣的事情，根据男女性心理特点，我们选取了两个属性，分别 num_share 分享数量和 num_loves + num_wows + num_hahas + num_sads + num_angrys 表情数量。这是因为一般男性很少会向朋友家人推荐化妆品和时装，并且他们也较少使用可爱的表情符号，相反，这些一般都是女性所喜爱的事情。

聚类结果如下图所示：



其中，左图表示依据受众性别对原始商店进行聚类的结果，右图则是原始类别的分布图（即原始数据中视频、图片、链接的帖子分类）。

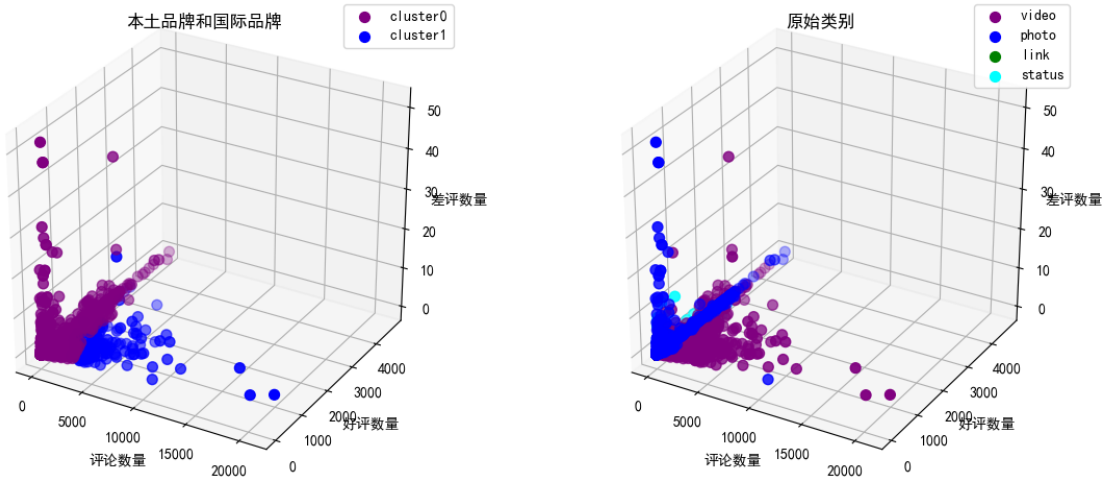
显而易见，左图蓝色 cluster1 代表受众为男性的商店，它们的分享数量和表情数量都很少，而紫色 cluster0 代表受众为女性的商店，它们的分享数量和表情数量都出奇的多。另外，值得欣喜的是，根据这两个属性可以将原始类别里 video 视频类别的大部分帖子分的成功分离。

2.1.5 国际品牌与本土品牌

由上文背景信息可知，在泰国时装和化妆品行业，国际品牌和本土品牌基本算是平分秋色，我们不妨分析一下各自的特点，本土品牌薄利多销，可能评论多但是毁誉参半，国际品牌较为昂贵，在普通民众之间分享少，评论少，但是好评多。

因此根据上述分析，我们选取了三个属性对原始数据进行了聚类，分别是代表评论水平的 num_comment 和代表好评数量的 num_likes + num_loves + num_wows + num_hahas 和代表差评数量的 num_sad + num_angrys。

聚类结果如下图所示：



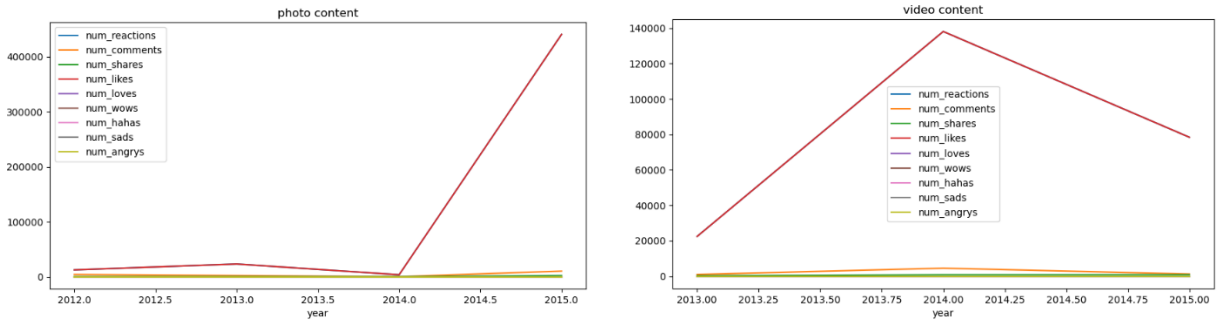
可以发现，聚类结果和初期设想的有所差距，这可能是因为评论少的商品可能本身就不怎么好，并不是因为受众小，由于这个因素的干预，所以和预期有一定差距，但是排除这类商品，其他的国际品牌的口碑也还不错。此外，利用这三个属性，对原始数据类别中 video 视频和 photo 图片的分类效果也还不错，这也是值得欣慰的。

2.2 时间序列分析

因为 FaceBook 直播模式开始于 2015 年 8 月，因此我们将数据一分为二，分别为 2016 年前和 2016 年后，并且针对视频、图片这两种类型分别在年、月、日、小时等时间程度进行数据分析。

2016 年前

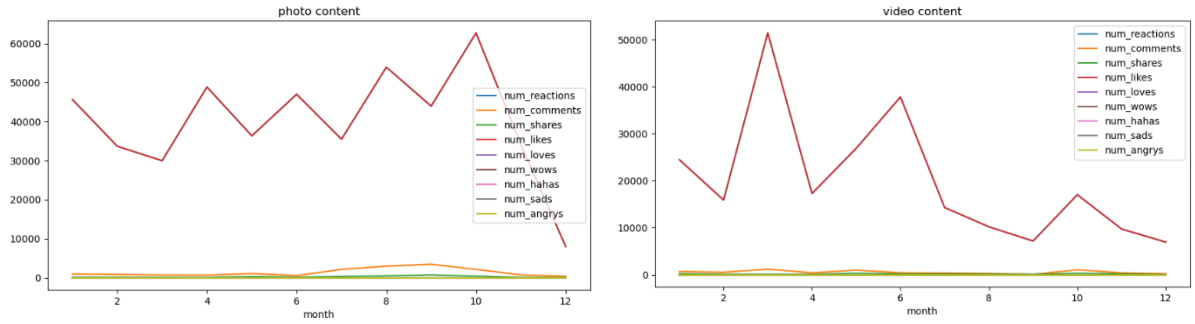
A．以**年**为单位：



对于图片内容，喜欢和反应数量从 2014 年到 2015 年增长显著；对于视频内容，2014 年时喜欢数量达

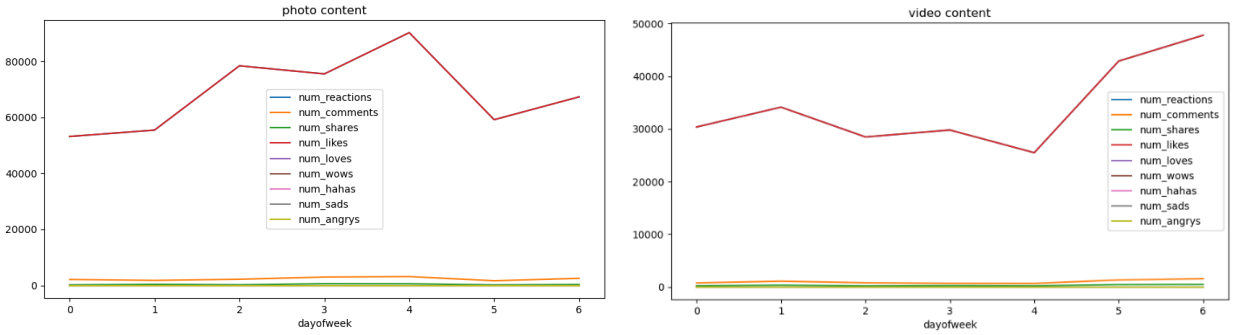
到最高点。

B．以**月**为单位：



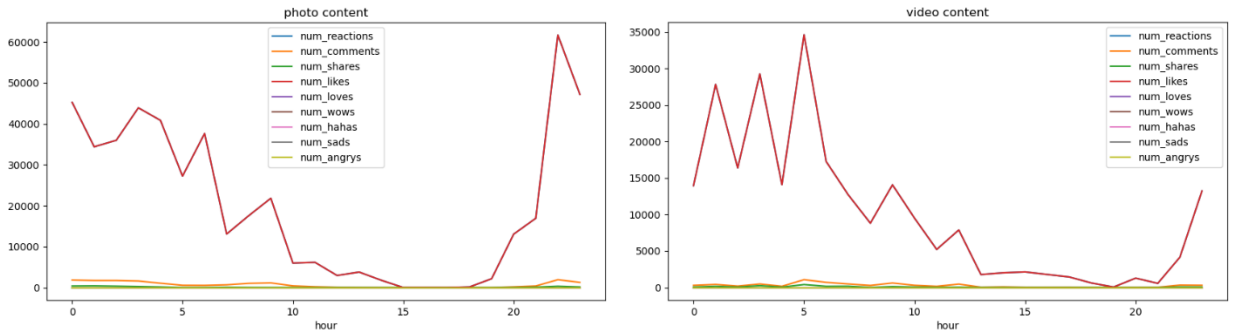
无论是视频内容还是图片内容，喜欢和反应数量都随着月份而震荡，但仍可以发现对于视频内容整体趋势是下降的，对于图片内容，发布在冬天似乎会收到更少的喜欢和反应数量。

C．以**天**为单位



对于视频内容，发布在周末可以收获更多的喜欢数量；对于图片内容，发布在周五可以收获更多的喜欢数量。这可能与周末大家一般都进行休闲娱乐有关系。

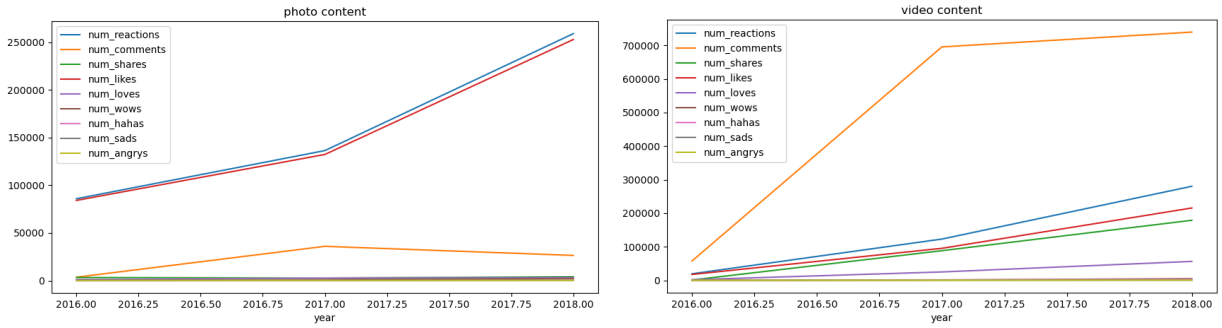
D．以**小时**为单位



对于图片内容，晚上 10 点喜欢数量达到顶峰，凌晨 1 点到 5 点喜欢数量也较多；对于视频内容，午夜 12 点到凌晨 5 点喜欢数量也处于较高的水平，这可能是晚上大家躺在床上玩手机的结果。

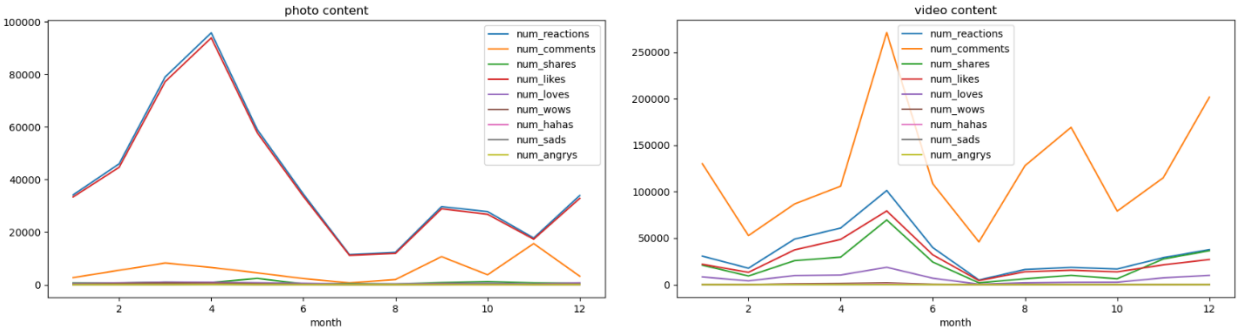
2016 年后

A．以**年**为单位：



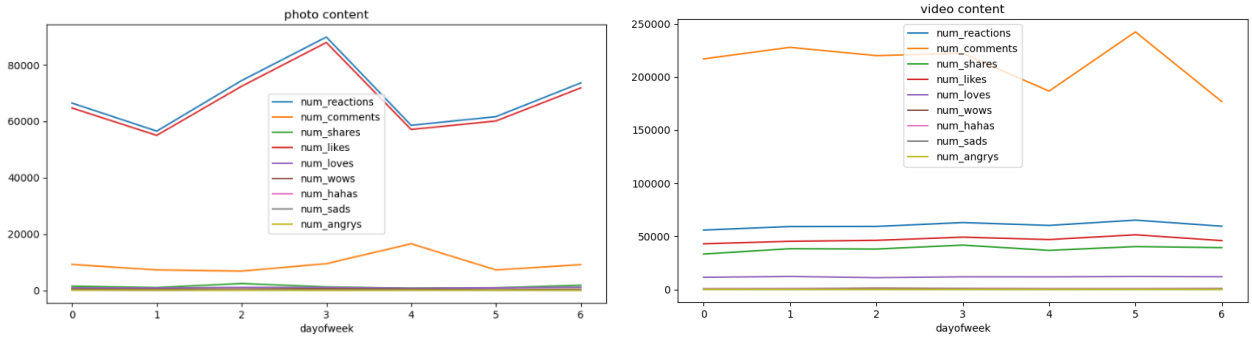
可以看出，图片内容的喜爱数量很多，而视频内容的评论和分享数量很多。

B．以**月**为单位：



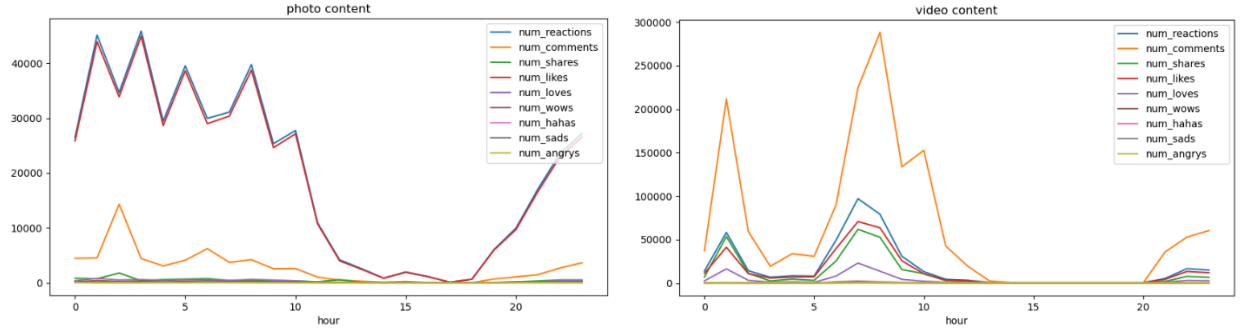
图片内容在夏天得到了最多的喜欢数量，在秋天喜欢数量较少；视频内容在 5 月份、9 月份、12 月份的喜欢数量都达到了顶峰。

C．以**天**为单位



对于图片内容，星期三和星期六都得到了最多的喜欢数量，但是整个喜欢数量分布都较为均匀；对于视频内容，在周六收获了最多的评论数量，与图片内容类似的是，整个分布也较为均匀。

D．以**小时**为单位



对于图片内容，晚上 8 点到次日凌晨 6 点喜欢数量达到较高的水平；对于视频内容，凌晨 1 点到早上 8 点分享和评论数量处于较高的水平，这可能是晚上直播开始的结果。

三、分析总结

3.1 分析总结

上述第二部分，我们选用不同的特征分别从流行程度、销售手段的现代和传统程度、信誉程度、受众性别、国际品牌和本土品牌等五个方面对原始数据集进行了聚类，并根据实际泰国化妆品和时装业 FaceBook 直播销售情况分析了这些聚类的合理性。

值得注意的是，在最后两种聚类标准中，即依据分享数量和表情数量的受众性别聚类和依据评论数量、赞美数量、反感数量的本土品牌和国际品牌聚类中，对原始 video、photo、link、status 分类的效果较为不错。这也说明，视频类的以及交互手段更多的帖子更受女性以及影响力大的商店的欢迎。

3.2 建议意见

从**时间序列分析**这一部分可以得出以下建议：

- 对于直播商家来说，发布图片类型的帖子可以获得更多的喜欢数量，发布视频类型的帖子可以获得更多的分享和评论数量。
- 关于季节性，无论是发布图片还是发布视频类型的帖子，最好都选择在夏季发布，并且避免在冬季发布，春季和秋季的效果差不多。
- 关于工作日，最好选择在周末发布，这样能获得更多的喜欢、分享和评论数量。
- 关于一天的时刻，最好选择在晚上 10 点以后到次日凌晨发布，这样也能获得更多的喜欢、分享和评论数量。

四、参考文献

[1]殷红.网络交易中信誉价值的影响因素研究——基于淘宝网的实证分析[J].商业经济与管理,2017(07):16-28.

[2]萧进才. 在线评论对搜索型商品销量的影响机制研究[D].暨南大学,2020.

[3]张艳辉,李宗伟,赵诣成.基于淘宝网评论数据的信息质量对在线评论有用性的影响[J].管理学报,2017,14(01):77-85.

[4]泰国化妆品市场有哪些特征. http://www.ouquan.cn/myzx/mydt/2015-09-25/5972.html