LAB03：Clustering

小组编号：

小组成员 1：孙士晨

小组成员 2：李兵磊

小组成员 3：董祥虎

小组成员 4：聂　尧

姓名	学号	分工	占比
孙士晨	1853427	多角度分析、数据分析报告	25%
李兵磊	1852024	PCA、多角度分析	25%
董祥虎	1850718	k-means、模型训练	25%
聂　尧	1851909	数据清洗、PCA	25%
总　计			100%

目录

1. 数据准备

本课题使用的数据集为来自 UCI 数据库的“Facebook Live Sellers in Thailand Data Set”(https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand)，数据集共包含 7051 个样本，每个样本包括 12 个特征，即卖家编号，卖家类型，加入时间，反应数量，评论数量，分享数量，喜欢数量，喜爱数量，支持数量，欢笑数量，悲伤数量，生气数量。

我们直接将数据下载为本地 csv 文件，供分析使用。

2. 数据清洗

2.1. 缺失值处理

首先我们检查每个样本的 12 个特征，发现不含缺失值。

```
status_id               0
status_type             0
status_published        0
num_reactions           0
num_comments            0
num_shares              0
num_likes               0
num_loves               0
num_wows                0
num_hahas               0
num_sads                0
num_angrys              0
Column1              7050
Column2              7050
Column3              7050
Column4              7050
```

但是 UCI 数据集含有 4 个空特征列，所以我们删除了这 4 个空特征。

.2.2. 数据降维

12 个特征中，有些特征是对于卖家类型 status_type 的分类是没有影响的，比如卖家编号 status_id 和加入时间 status_published，属于无关变量，所以我们将这两个变量删除。

```
df.drop(['status_id', 'status_published'], axis=1, inplace=True)
```

我们将这样处理后的数据按照 status_type 分组绘制柱状图，如图表 1。

2.3. 数据编码

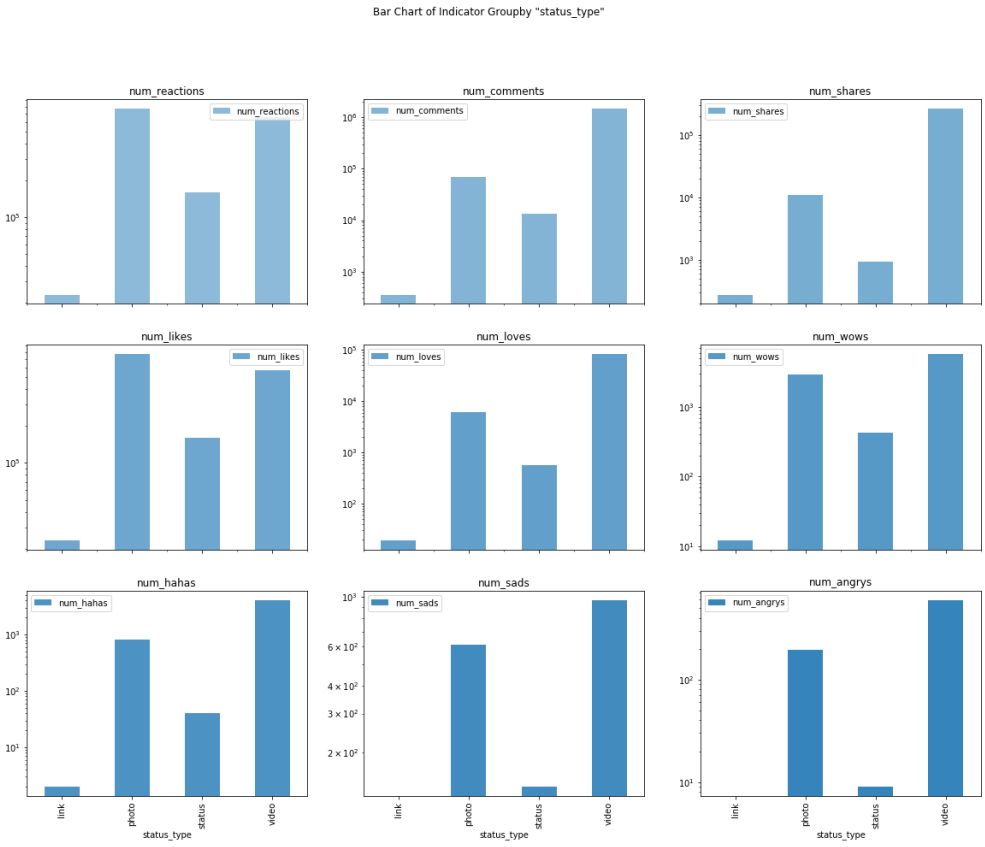
由于分类变量是 object 类型，所以需要对该变量进行编码。我们按照下表进行编码。

0	1	2	3
Link	Photo	Status	Video

2.4 正态标准化

之后我们对每个特征的数据正态标准化。即

图表 1



$$x' = \frac{x - \mu}{\sigma}$$

3. 模型搭建

我们主要采用 k-means 方法进行聚类。

2.1. 数学形式与基本原理

k-means 的基本思想是，通过迭代寻找 K 个簇（Cluster）的一种划分方案，使得聚类结果对应的损失函数最小。其中，损失函数可以定义为各个样本距离所属簇中心点的误差平方和：

$$J(c, \mu) = \sum_{i=1}^{M} |x_i - \mu_{c_i}|^2$$

其中x_i代表第i个样本，c_i是x_i所属的簇，μ_{c_i}代表簇对应的中心点，M是样本总数。

2.2. 特点

2.2.1 K-MEANS 的优点

- 计算复杂度为$O(NKt)$，接近于线性，收敛速度快；
- 可解释性强。

2.2.2 K-MEANS 的缺点

- 受初始值和异常点影响，聚类结果可能不是全局最优而是局部最优；
- k 是超参数，不好选择；
- 样本点只能划分到单一的类中。

2.3. 搭建过程

```
model = cluster.KMeans(n_clusters=4, random_state=2021)
```

4. 模型训练测试

4.1. 确定 K 值

由于 k-means 聚类需要首先确定超参数聚类数量 k，我们采用肘部法则来确定最佳的 k 值。

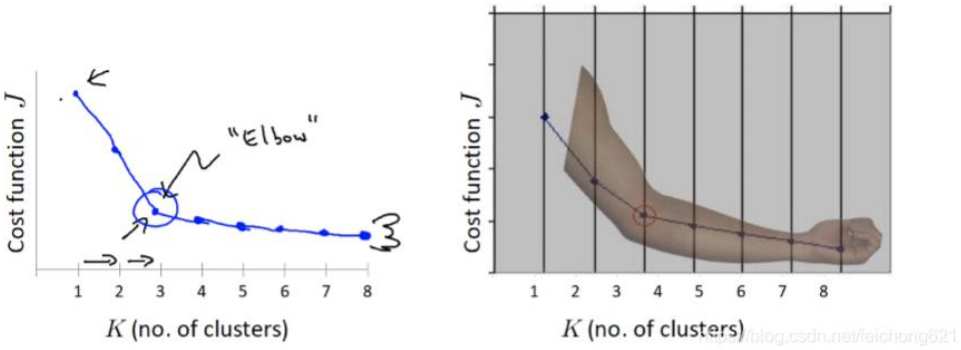
肘部法则简述如下：

（1）对于 n 个点的数据集，迭代计算 k 从 1 到 n，每次聚类完成后计算每个点到其所属的簇中心的距离的平方和；

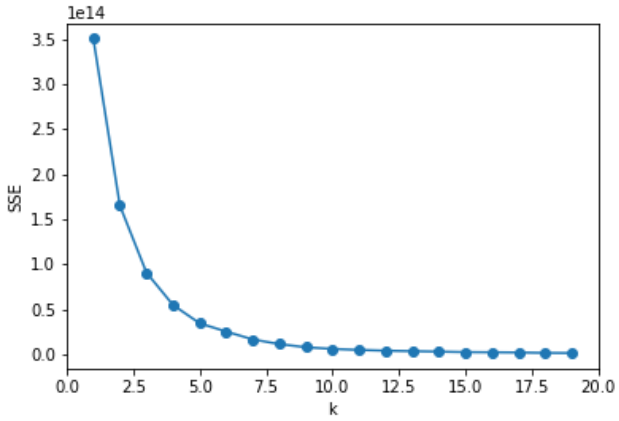
（2）平方和逐渐变小，直到 k=n 时平方和为 0；

（3）在平方和变化过程中，会出现一个拐点即“肘”点，“肘”点处下降率突然变缓。“肘”点对应的 k 值即佳最 k 值；

（4）在决定什么时候停止训练时，肘形判据同样有效。



本数据集的肘部图如下图。y 轴为 SSE（Sum of the Squared Errors），x 轴为 k 的取值，随着 x 的增加，SSE 会随之降低，当下降幅度明显趋向于缓慢的时候，取该值为 K 的值。显然，k=4 最为合适。



4.2. 聚类训练

用确定的超参数 k=4 进行聚类。

```
model.fit(df_std[zcols])
```

5. 结果可视化

下表是聚类结果的统计情况。

ClusterType	Link	Photo	Status	Video	num reactions	num comments	num shares	num likes	num loves
Cluster0	0	1	14	2074	168.38	379.28	74.77	143.01	23.42
Cluster1	14	212	78	71	1815.95	60.69	11.53	1810.22	3.08
Cluster2	49	4062	273	0	92.00	11.73	1.73	89.95	1.24
Cluster3	0	13	0	189	922.15	3540.53	564.56	713.44	169.47

该表显示了数据的不平衡性，即有照片和视频站大多数。

- 群组 0：video 为主，没有 link。
- 群组 1：photo 为主，没有 video。
- 群组 2：video 为主，但明显比群组 0 小，没有 link 或 status。status_type 对这个群组似乎不那么重要。
- 群组 3：photo 为主，但确实包括所有四种 status_type。和群组 2 一样，status_type 对这个群组显得不那么重要。

最终聚类结果为：

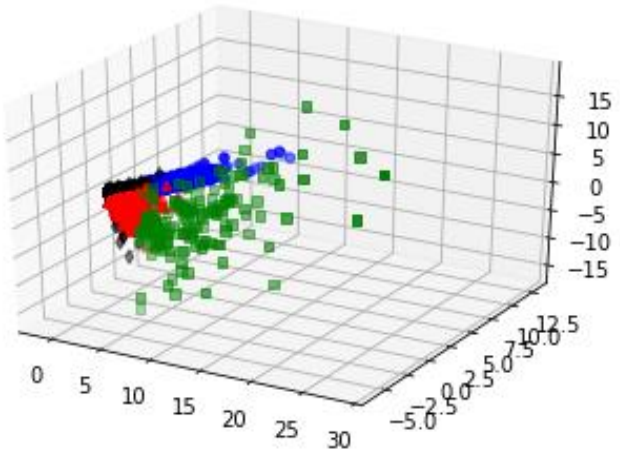
- 第 0 组：视频。
- 第 1 组：照片，没有分享、评论或喜爱，反应和喜欢很少。

- 第 2 组：与状态类型相关性不大，更多与帖子上的大量评论有关。
- 第 3 组：与状态类型相关性不大，更多与有类似平均喜欢和反应数量的帖子有关。

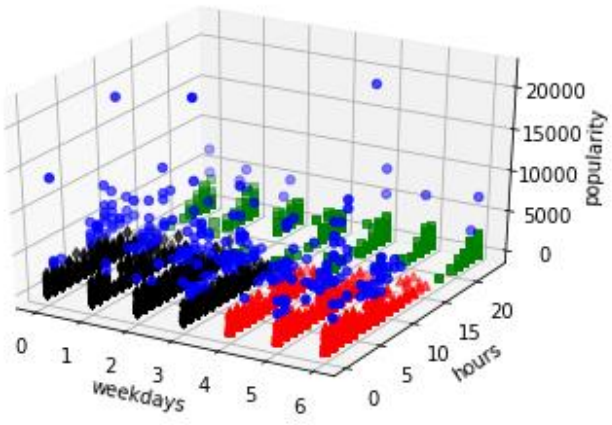
6. 模型优化

k-means 聚类模型的优点是可解释性强、速度快，但是他也有相应的缺点，即超参数不好确定、容易受初始值和异常点的影响等。针对超参数不好确定的问题，我们用肘部法则确定 k 的数量；针对异常点的问题，我们对数据进行了正太标准化，削弱了异常点的影响。

另外，原始数据的维度较多，可能有噪声，我们采用 PCA 的方式给数据降维，并按照降维后的数据分类。



我们首先将数据降维到 3 维。降维后的结果如上图所示。按照降维后的数据进行聚类，可以看到聚类效果提升明显，不同类在样本空间中分布较为分散。



7. 多角度分析（数据分析报告）

为了分析更具有说服力，本部分模拟 Facebook 官方的角度，对已有数据进行分析，希望能对卖家给出合理的建议，帮助卖家吸引更多的用户，让用户有更好的体验。

7.1. 添加新指标

为了对时间进行进一步分析，新增的两个变量

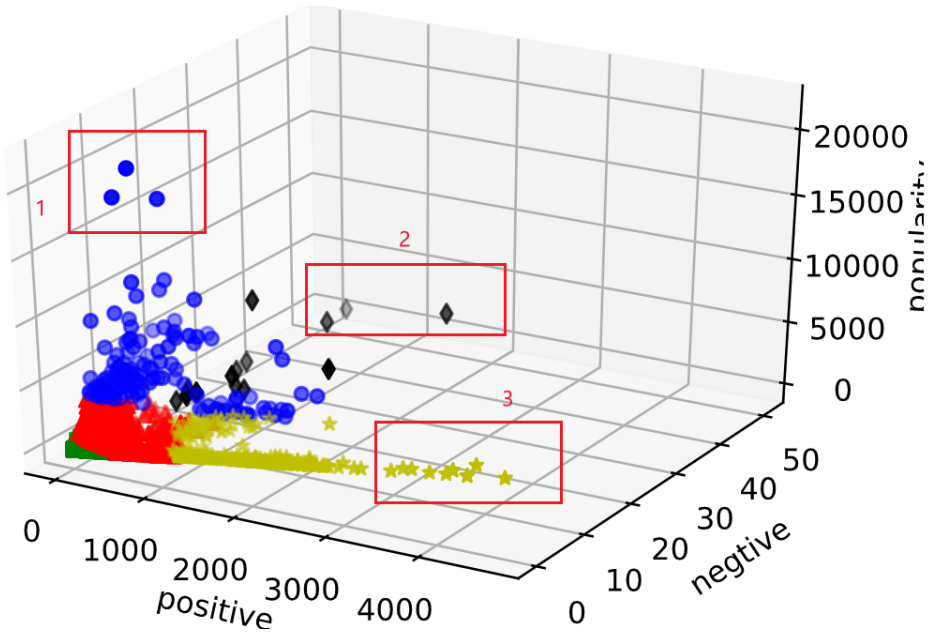
- weekdays 发布日期是星期几，范围从 1-7
- hours 发布的时间，在几点

根据已有的变量，进行组合，得到三个新的指标

- popularity：活跃数，是把 reactions，comments 和 shares 三个值相加，代表当前内容的活跃数
- positive：积极情感的个数，likes，loves，wows, hahas 的值相加，代表对当前内容有积极情感倾向的人数
- negtive：消极情感的个数，sads，angrys 的值相加，代表对当前内容有消极情感倾向的人数

7.2. 对情感的分析

通过对活跃数，积极情感，消极情感三个变量进行聚类，得到下面的结果。



对结果的分析：

1. 首先是离群点
   - 对于 1 位置，可以发现，这部分的内容有较大的活跃度，有更多的用户去评论，分享，回复。但是用户对这部分内容有较少的情感倾向。
   - 对于 2 位置，可以发现，这部分的内容的消极占比较多。

- 对于 3 位置，可以发现，该部分虽然活跃度较少，但是能获得很高的的积极态度，这部分内容可能更受大家喜欢。

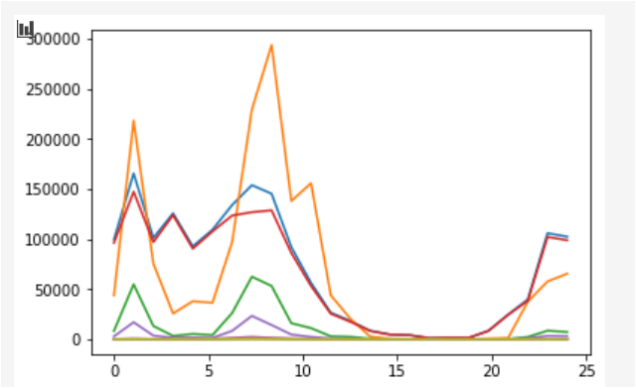
2. 聚类结果的分析：
   - 对绿色和红色部分，这部分内容的特点是活跃度少，积极和消极的情感态度也较少。说明这部分内容热度少，用户的关注也较少。
   - 对于蓝色部分，这部分内容的特点是活跃度提高，但是积极和消极的情感态度提高相对较少。
   - 对于黄色部分，这部分内容能够得到用户较多的积极情感，说明该部分内容比较优质。

给出的建议：

- 对于聚类中黄色部分，这部分卖家能够收获用户大量的积极态度，应该给予适当的支持，如适当的主页推广。尤其是上图方框 3 中的部分用户，应该给予更多的关注，这部分卖家能够显著得到用户的积极态度，未来或许能够有更好的发展。
- 对于聚类中的蓝色部分，这部分卖家有较高的用户活跃度，但是相对来说，收获较少的用户积极态度，可以对这部分卖家进行通知，让他们和用户有更多的互动。

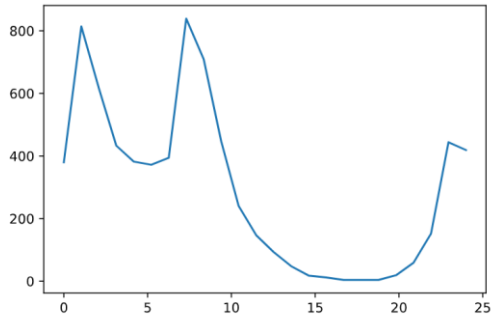
7.3. 对时间的分析

根据给出的信息，绘制每小时各个数据信息图，如下图。可以从图中看到各种数据的值在不同的时段有明显的不同，因此可以根据小时数据进行分析。



7.3.1. 小时信息的分析

统计每个小时的发布的卖家数量，根据小时数和卖家数（横坐标小时，纵坐标卖家数），得到下述的图像。



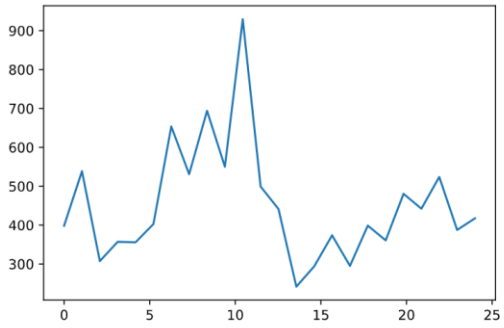
可以得到如下的结论

- 在 15.00-20.00 卖家发布较少
- 在 23.00-1.00　8.00-10.00 用户较多

绘制平均每个卖家的活跃度，公式为

$$\frac{\sum(shars + comments + reactions)}{n}$$

即，该时段的总活跃度除以该时段的卖家总数。得到下面的结果



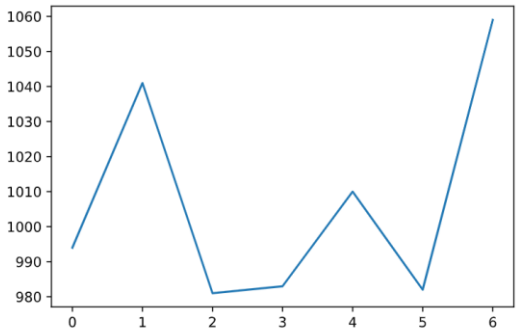
可以发现，在 11 点的时候最多，并且大致走势和上面的每小时的卖家数大致相同，说明卖家趋向于选择人数较多的时候发布信息。

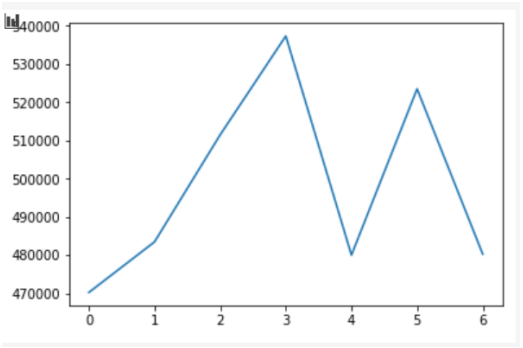
- 针对上面两个图片走势大致相同，有如下的两个解释
  - 卖家倾向于选择用户在线较多的时候，如晚上 22.00 之后，更多人空闲，所以会有更多卖家发布信息，
  - 同时，因为卖家的集中发布，会影响用户倾向于在卖家较多的时段参与直播，会导致该时段用户较多。
- 根据上述的两个图片
  - 可以给出卖家如下的建议：如果希望稳妥的话，希望卖家考虑上午 8.00-11.00，晚上 22.00-1.00 这段时间，这段时间里用户数较多，平均每个卖家能够得到更多的用户
  - 平台方可以尝试引导更多的卖家在下午 14.00-16.00 这段时间入驻，让这段时间能够活跃起来

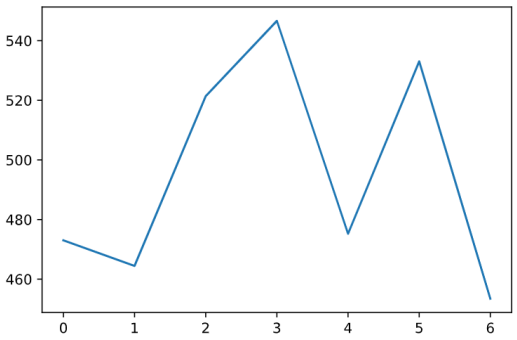
7.3.2. 周信息的分析

首先绘制星期几和卖家发布个数之间的关系（下左图）。可以看到最大的值大约为 1060，最小值大约为 980，相差其实不大。说明星期几其实对卖家发布数量影响较少。

然后用 shares+comments+reactions 的和代表当天的活跃度，绘制用户活跃度和星期几之间的关系，得到下面的关系（下右图）。可以发现，最高点在周三，最低点在周天，改变幅度超过 10%，说明每周每天的活跃人数其实有较大的差距。同时，可以发现一个和常识不太相同的事情：周末的活跃人数并不会显著增加。







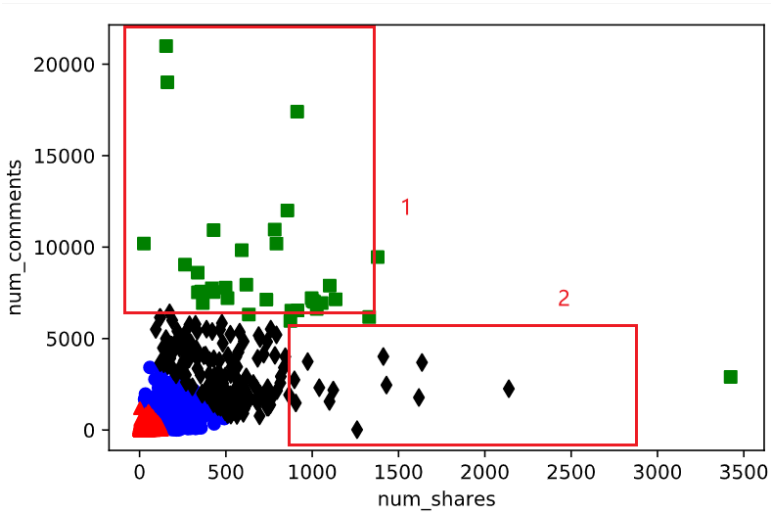
然后绘制平均的用户活跃度（上左图）可以看到，卖家发布数多，则平均得到的用户活跃度就会变小，如周六，有最多的卖家发布，但是平均每个卖家有的用户活跃数最少。

因此可以给出如下的建议：

- 对用户来说，每天的直播数量大致相同，不论周几都有大量的直播卖家
- 对于卖家来说，如果想要有更多的用户参与，可以尝试多在周三，周五发布，因为在周三，周五，平均用户数更多，平均的用户参与度也最高
- 对于平台方，通过上图可以发现，在周末用户的活跃度其实较低。用户在周末是休息，平台方可以增加对用户的吸引力，在周末吸引更多的用户

7.4. COMMENT 和 SHARE

对评论和分享聚类，得到下述的结果。可以看到，分布大致在对角线上，分享和评论成正相关，这也符合我们的预取想法。分享和评论都代表着用户的赞同，这两个应该成正相关。



我们侧重关注方框的部分，可以看到

方框 1 中的内容，有较多的评论，但是分享较少。较多的评论说明这部分卖家有较好的互动性，对这部分卖家可以给予适当的推广，引入更多的关注。

方框 2 中的内容，有较少的评论，但是分享较多。对这部分卖家可以基于建议，多多和用户互动。

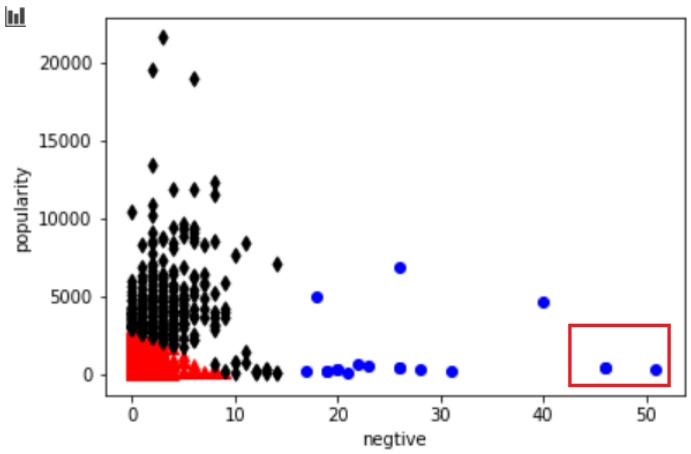
7.5. 找优秀卖家和不良卖家

作为平台方，应该对卖家的性质进行合理分类，帮助商家认识自己的不足并且发扬自己的长处。

找到一些不良的卖家，通过活跃度和消极的情况进行聚类，得到下面的结果

聚类主要分为三类，

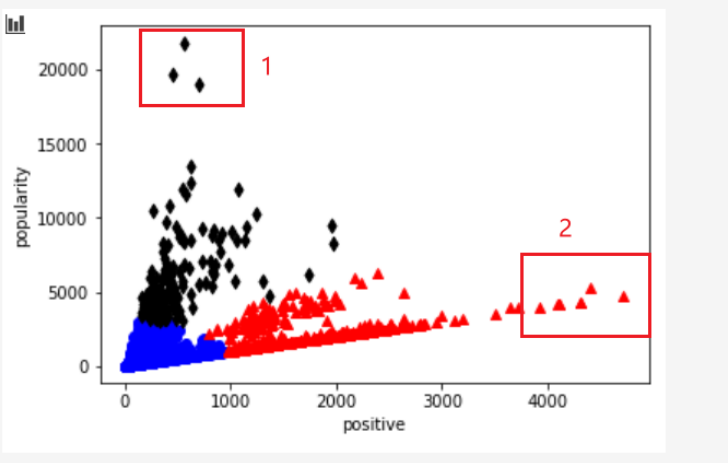
- 红色部分的买家用户活跃度较少
- 黑色部分的卖家占大多数，该部分的卖家有着合理的用户消极评价数
- 蓝色部分的卖家用户消极评价稍微超过正常标准，尤其是方框部分，卖家有较多的消极评价



给出的建议：

作为平台方，对方框中的离群点代表的卖家进行适当的通知，希望卖家能够进行一些改变，不会给用户带来不好的体验。

找到一些比较优秀的卖家，通过活跃度和积极的情况进行聚类，得到下面的结果



聚类主要分为三类，

聚类主要分为 3 类

- 蓝色部分这部分卖家属于活跃度和用户积极反馈都较少的
- 黑色部分的卖家，有较多用户活跃度，但是用户给予更少的积极评价
- 红色部分的卖家，虽然用户活跃度较少，但是用户给了更多的积极反馈

应该重点关注的是两个方框中的内容。

观察上述两个图片，发现方框 1 位置都有三个孤立的点，说明这三个卖家，有较高的用户活跃度，但是积极评价和消极评价都较少，因为现在的信息不够无法继续深入分析，可以后续使用更多资料，查看具体原因

方框 2 的位置，较少的活跃度，较大的积极反馈，说明这部分卖家能更好的吸引用户，但是受限于自身的情况，可能获得的关注不够多。

给出的具体建议：

- 作为平台方，对方框 1 中的卖家，进一步进行分析，查看具体的原因
- 对上述方框 2 中的用户，平台方可以给予适当的推广，帮助卖家得到更多的用户。

7.6. 结论

我们小组通过对给定数据集的分析，从情感，时间，行动，特殊的卖家等多个角度，通过机器学习的方法，对数据进行处理，然后进行分析，站在 facebook 的角度，希望能够给平台，卖家，用户都带来正向的影响。

结论总结:

- 从平台的角度：通过用户的倾向和活跃度之间的关系，对优秀用户予以推广，对相对获得更多消极评价的卖家，给予通知。对卖家的发布时间进行合理的引导，期望全天候的吸引跟多的用户。
- 从卖家的角度，可以选择合理的时间发布自己的内容，得到更多的用户关注。根据自己 在聚类中所处的类型，找到自己的不足，并向其他卖家学习。
- 从用户的角度，可以了解合适参与能够有更多的卖家，更多的选择。选择一些获得更积极的卖家，能够带来更多的预约。