项目背景：随着移动互联网多年的快速发展，移动互联网已进入下半场 ，不再依靠用户红利来经营，发展业务，告别粗糙的/高成本企业发展的方式，开始转而精细化管理，结合市场、渠道、用户行为等数据分析，对用户展开有针对性的运营活动，提供个性化、差异化的运营策略，以实现运营目的行为。本文利用SQL 对淘宝用户行为数据进行分析，通过用户行为分析业务问题，提供针对性的运营策略。

分析步骤：

1. 提出问题
2. 数据理解
3. 数据清洗
4. 构建模型
5. 总结与建议

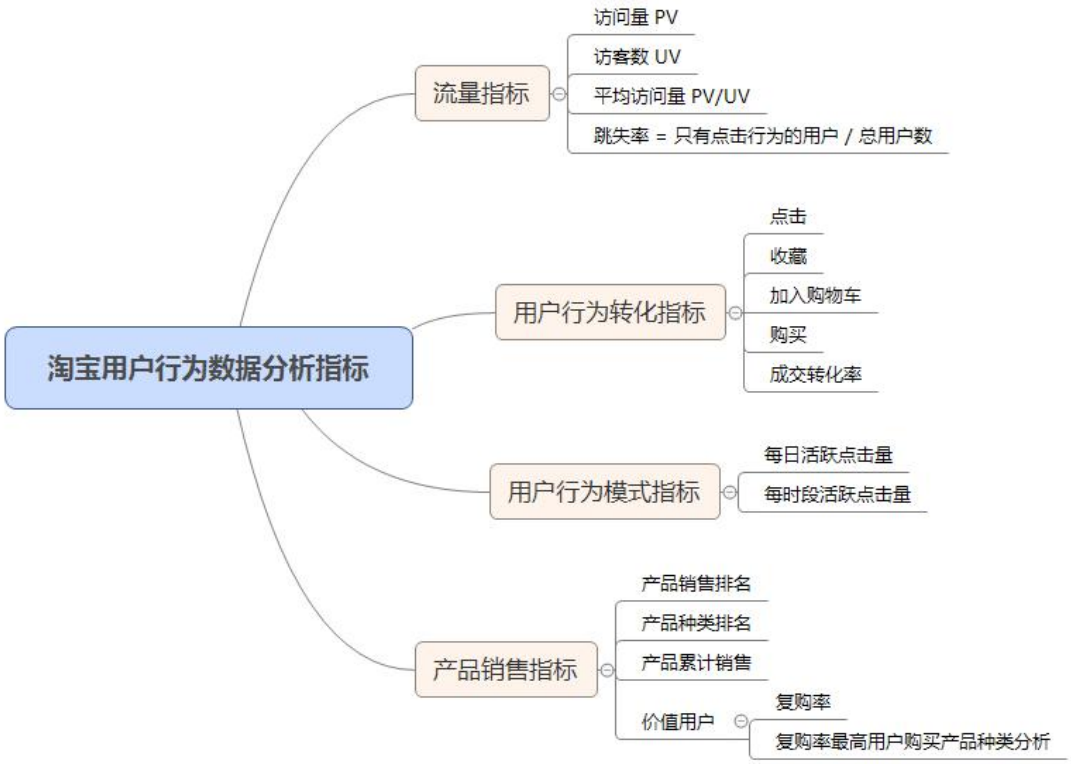
一、提出问题

1. 本次分析的业务问题及适用指标

本次分析的目的是想通过对淘宝用户行为数据分析，为以下问题提供解释和改进建议：

1. 用户从浏览到最终购买的整个过程的流失情况，确定夹点位置，提出改善转化率的意见。
2. 在研究的时间段里找出用户最活跃的日期以及每天活跃的时间段，了解用户的行为时间模式。
3. 什么产品以及产品类目的购买率最高，找出最受欢迎的产品，优化产品销售。
4. 哪些用户购买次数最多，找出最核心的付费用户群，并且统计出这些用户购买的产品以及类目，针对这些用户的购买偏好推送个性化的产品销售方案。

针对上面的业务问题，下面是适用的业务指标：



2. 基于 AARRR 漏斗模型分析用户行为

本项目通过常用的电商数据分析业务指标，采用 AARRR 漏斗模型拆解用户进入 APP 后的每一步行为。AARRR 模型是根据用户使用产品全流程的不同阶段进行划分的，针对每一环节的用户流失情况分析出不同环节的优化优先级，主要通过

以下个各阶段来进行分析：

二、数据理解

本项目数据来源于阿里云天池，可登陆阿里云天池下载数据，地址如下：[User Behavior Data from Taobao for Recommendation]

本数据集包含了 2017 年 11 月 25 日至 2017 年 12 月 3 日之间，有行为的约一百万随机用户的所有行为（行为包括点击、购买、加购、喜欢）。数据集的组织形式和 MovieLens-20M 类似，即数据集的每一行表示一条用户行为，由用户 ID、商品 ID、商品类目 ID、行为类型和时间戳组成，并以逗号分隔。关于数据集中每一列的详细描述如下：

列名称	说明
用户ID	整数类型，序列化后的用户ID
商品ID	整数类型，序列化后的商品ID
商品类目ID	整数类型，序列化后的商品所属类目ID
行为类型	字符串，枚举类型，包括('pv', 'buy', 'cart', 'fav')
时间戳	行为发生的时间戳

注意到，用户行为类型共有四种，它们分别是：

行为类型	说明
pv	商品详情页pv，等价于点击
buy	商品购买
cart	将商品加入购物车
fav	收藏商品

关于数据集大小的一些说明如下：

维度	数量
用户数量	987,994
商品数量	4,162,024
商品类目数量	9,439
所有行为数量	100,150,807

三、数据清洗

1. 观察记录

原数据集数据记录达到 1 亿条，数据量庞大，为了方便分析与效率，本项目将选取了从 500 万行至 800 万的 300 万条记录进行分析。

2. 一致化处理

原数据时间戳使用的是 epoch&unix timestamp 格式，需要转换为标准可读的日期时间形式。在原数据表增加 3 个新字段 datetime、dates、hours，把转换好的日期时间放进去。

```
ALTER TABLE userbehavior ADD COLUMN datetime TIMESTAMP(0) NULL;
UPDATE userbehavior SET datetime=FROM_UNIXTIME(timestamps);

ALTER TABLE userbehavior ADD COLUMN date CHAR(10) NULL;
UPDATE userbehavior SET date=SUBSTRING(datetime FROM 1 FOR 10);

ALTER TABLE userbehavior ADD COLUMN hour CHAR(2) NULL;
UPDATE userbehavior SET hour=SUBSTRING(datetime FROM 12 FOR 2);
```

user_id	item_id	category_id	behavior	timestamps	datetime	date	hour
309818	1461532	3102419	cart	1511959462	2017-11-29 20:44:22	2017-11-29	20
309818	4710383	1792277	pv	1511959603	2017-11-29 20:46:43	2017-11-29	20
309818	1421743	4069500	pv	1511959759	2017-11-29 20:49:19	2017-11-29	20
309818	800137	1216617	pv	1511959828	2017-11-29 20:50:28	2017-11-29	20
309818	2493122	1216617	pv	1511959953	2017-11-29 20:52:33	2017-11-29	20
309818	1461532	3102419	pv	1511998449	2017-11-30 07:34:09	2017-11-30	07
309818	3648099	4220654	buy	1512003819	2017-11-30 09:03:39	2017-11-30	09
309818	149503	58836	pv	1512003932	2017-11-30 09:05:32	2017-11-30	09

3. 异常值处理

检查日期是否在规定范围内（2017 年 11 月 25 日至 2017 年 12 月 3 日），将不符合规定的数据删除。

```
SELECT MAX(timestamps),
       MIN(timestamps),
       MAX(datetime),
       MIN(datetime)
FROM userbehavior;
```

MAX(timestamps)	MIN(timestamps)	MAX(datetime)	MIN(datetime)
1634651606	-1586903608	2021-10-19 21:53:26	2017-04-11 11:40:48

```
DELETE FROM userbehavior
WHERE datetime<'2017-11-25 00:00:00' OR datetime>='2017-12-04 00:00:00';
```

一共删除了 1689 行数据，再次验证日期时间的准确性，下面结果满足要求：

MAX(timestamps)	MIN(timestamps)	MAX(datetime)	MIN(datetime)
1512316798	-1586903608	2017-12-03 23:59:58	2017-11-25 00:00:00

四、构建模型

1. 流量与用户行为转化分析

解决问题：用户从浏览到最终购买的整个过程的流失情况，确定夹点位置，提出改善转化率的意见。

1）访客数 UV、访问量 PV、平均访问量 PV/UV：

```
SELECT
    COUNT(DISTINCT user_id) AS 'UV',
```

```
    (SELECT COUNT(*) FROM userbehavior WHERE behavior='pv') AS
'PV',
    (SELECT COUNT(*) FROM userbehavior WHERE
behavior='pv')/(COUNT(DISTINCT user_id)) AS 'PV/UV'
FROM userbehavior;
```

UV	PV	PV/UV
29233	2683658	91.8023

2）跳失率(只有点击行为的用户/总用户数)：

```
SELECT COUNT(DISTINCT user_id)
FROM userbehavior
WHERE user_id NOT IN(SELECT DISTINCT user_id FROM userbehavior
WHERE behavior = 'fav')
    AND user_id NOT IN(SELECT DISTINCT user_id FROM userbehavior
WHERE behavior = 'cart')
    AND user_id NOT IN(SELECT DISTINCT user_id FROM userbehavior
WHERE behavior = 'buy');
```

结果显示只有点击行为却没有收藏、加入购物车以及购买行为的用户数是 1628，

除以总用户数 29233，则跳失率为为 5.57%。

3）用户总行为漏斗：

```
SELECT behavior,COUNT(*)
FROM userbehavior
GROUP BY behavior;
```

behavior	COUNT(*)
cart	168771
pv	2683658
buy	59329
fav	86553

由于收藏和加入购物车都为浏览和购买阶段之间确定购买意向的用户行为，且不分先后顺序，因此将其算作同一阶段。可以看到从浏览到有购买意向只有**9.50%**的转化率，当然也有部分用户是直接购买而未通过收藏和加入购物车，但是这仍说明大多数用户浏览页面次数较多，而使用加入购物车和收藏功能较少。另外，购买次数占加入购物车和收藏功能的**23.53%**左右，说明从浏览到收藏和加入购物车的阶段是指标提升的重点环节。

4）独立访客行为漏斗：

```
SELECT behavior,
       COUNT(DISTINCT user_id) AS DIS_user
FROM userbehavior
GROUP BY behavior;
```

behavior	DIS_user
buy	19865
cart	22131
fav	11638
pv	29116

上图展示的是每一步用户行为的独立访客数的分布情况，可以看出使用 APP 的用户中 PUR 约为 68.2%，用户付费成交转化率相当高，说明用户的购买欲望还是挺大的。

2. 用户行为模式分析

解决问题：在研究的时间段里找出用户最活跃的日期以及每天活跃的时间段，了解用户的行为时间模式。

1)每日活跃点击量：

```
SELECT date,COUNT(*) as pv
FROM userbehavior
WHERE behavior='pv'
GROUP BY date
ORDER BY date;
```

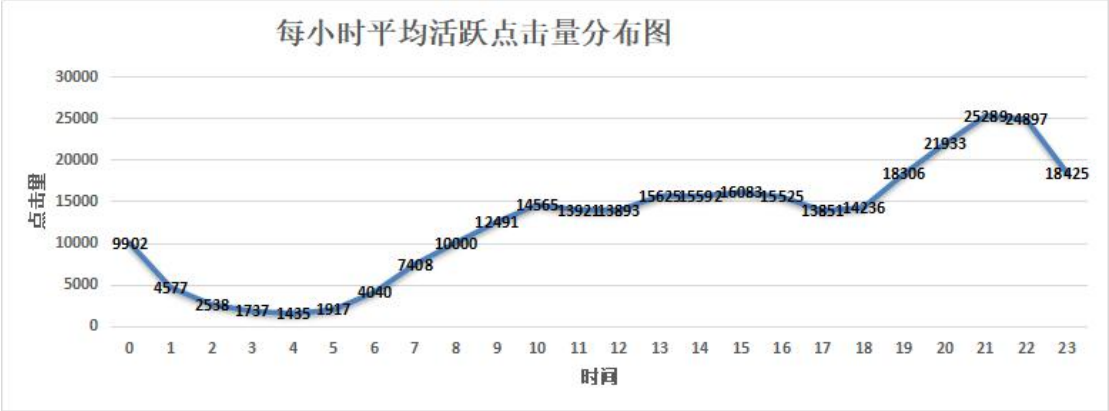
date	pv
2017-11-25	283830
2017-11-26	287519
2017-11-27	271512
2017-11-28	267496
2017-11-29	272311
2017-11-30	278325
2017-12-01	291565
2017-12-02	368515
2017-12-03	362584

从上图可以看出 **11** 月 **25** 日**-12** 月 **1** 日保持稳定的水平，**12/2** 开始出现较为明显的增长，点击量陡增，增长率约为 **26.4%**。推测是上班族因工作逛淘宝的时间少，而周末**(12** 月 **2** 日**-12** 月 **3** 日**)**有充足的精力和有较多空闲时间访问淘宝。因此平日运营可以将活动集中在周末进行。

2)每时段的活跃点击量：

```
SELECT `hour`,COUNT(*)/9
FROM userbehavior
WHERE behavior = 'pv'
GROUP BY `hour`
ORDER BY `hour`;
```

hour	hour_pv
00	9901.6667
01	4576.8889
02	2537.5556
03	1736.8889
04	1435.3333
05	1917.1111
06	4039.5556
07	7408.4444
08	10000.0000
09	12490.5556
10	14565.0000
11	13920.6667
12	13892.7778
13	15624.8889



在数据集观察的 9 天里，从 18 点开始点击量稳步上升，到 21 点到达顶峰，22 点稍有回落，到 23 点明显下降，说明大部分用户会在晚上 18 点到 22 点时段频繁点击浏览网页，符合大部分人的作息时间。

3. 产品销售分析

解决问题 1：什么产品以及产品类目的购买率最高，找出最受欢迎的产品，优化产品销售。

解决问题 2：哪些用户购买次数最多，找出最核心的付费用户群，并且统计出这些用户购买的产品以及类目，针对这些用户的购买偏好推送个性化的产品销售方案。

1）浏览次数、收藏次数、加入购物车次数以及购买次数最多的商品：

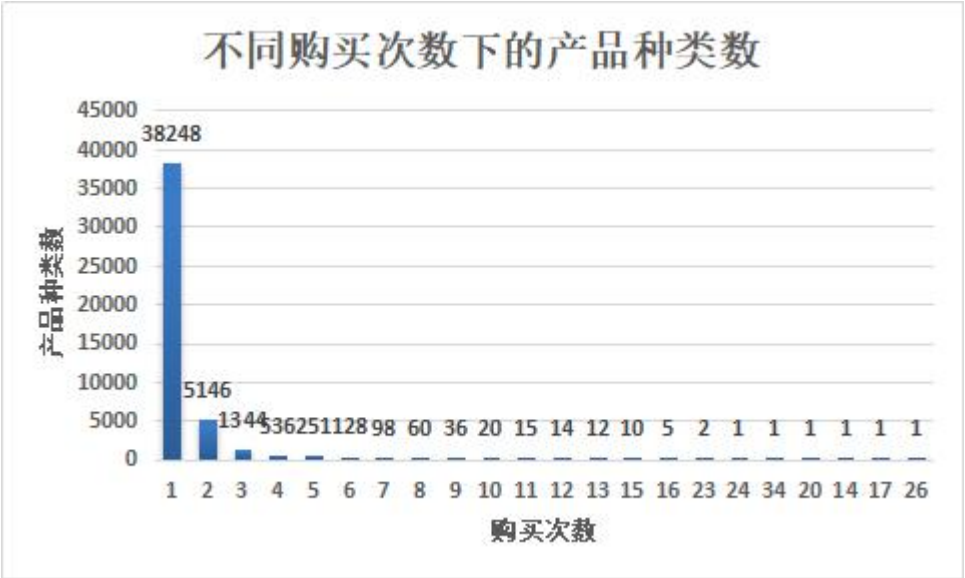
```sql
SELECT
    item_id,
    count(user_id) AS times_pv
FROM
    userbehavior
WHERE
    behavior='pv'
GROUP BY
    item_id
ORDER BY
    times_pv DESC;
```

在销量榜单中并没有看到浏览量第一第二的商品，说明这些吸引用户更多注意力的商品并没有很好的转化为实际销量，仅更多的加入收藏中（浏览量前排的商品均能在收藏量前列中，说明浏览量与收藏的关系更为直接）。

2）产品销售排名：

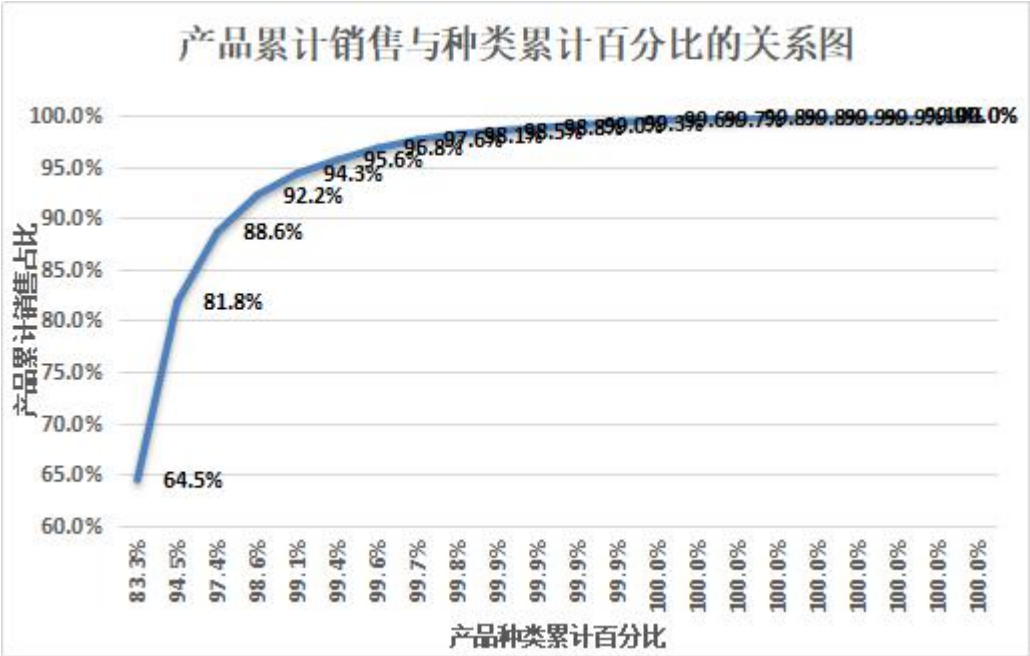
```
-- 计算不同购买次数下的产品种类数
SELECT
    a.buy_num AS buy_count,
    COUNT(a.item_id) AS item_num
FROM
    ( SELECT item_id, COUNT(user_id) AS buy_num FROM userbehavior
WHERE behavior='buy' GROUP BY item_id ) AS a
GROUP BY
    a.buy_num
ORDER BY
```

```
    item_num DESC;
```



从上图可以看出只被购买一次的产品有 **38248** 种，被购买两次的产品有 **5146** 种，本次分析的产品（**item_id**）有 **45931** 种，只被购买一次的产品占到 **83.3%**，意味着并没有销售非常集中的产品。为了看清楚这一部分，我们来看看产品种类的累计销售情况。

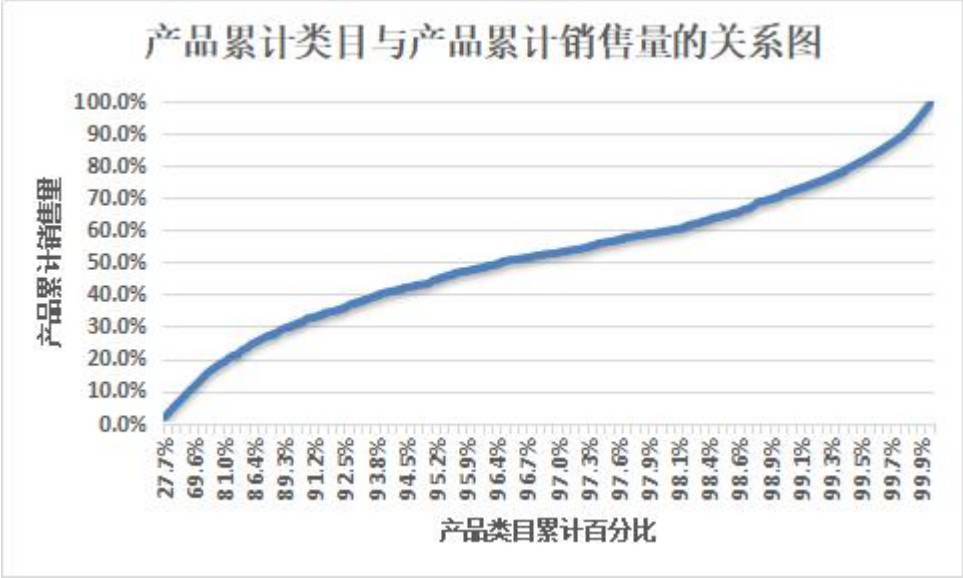
3）产品种类的累计销售情况：



从上图可以看出 **83.3%**的产品贡献了 **64.5%**的销售量，不符合传统零售业的二八法则，说明电商靠长尾理论累计销售，而不是制造爆款商品带动销量。

4）产品类目的累计销售情况：

```
-- 计算不同购买次数下的商品类目数量
SELECT
    a.cat_buytimes,
    COUNT(category_id) AS cat_type_count
FROM
  -- 每种商品类目的购买次数
    ( SELECT category_id,COUNT(user_id) AS cat_buytimes FROM
UserBehavior WHERE behavior='buy' GROUP BY category_id ) AS a
GROUP BY
    a.cat_buytimes
ORDER BY
    a.cat_buytimes;
```

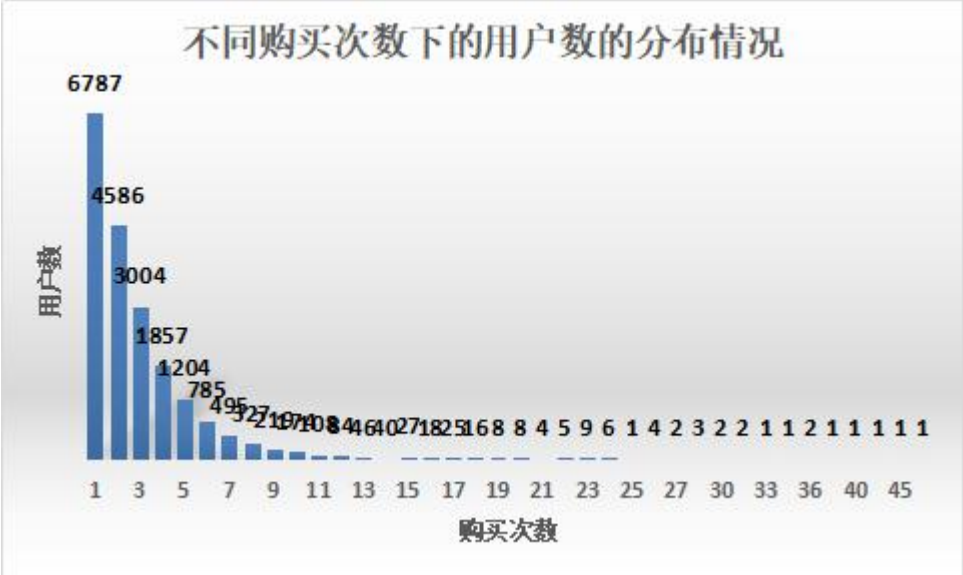


accu_cat_buytimes	accu_cat_type_count
1.6%	27.7%
3.2%	40.7%
4.7%	49.2%
6.1%	55.2%
7.4%	59.6%
8.9%	63.7%
10.2%	66.9%
11.5%	69.6%
12.9%	72.1%

从上表可以更清楚的看出 **27.7%**的产品类目贡献了 **1.6%**的销售量，**69.6%**的产品类目贡献了 **11.5%**的销售量，不符合传统零售业的二八法则，同样可以说明其依靠长尾理论累计销售。

5）复购率：

```
-- 统计不同购买次数下的用户数
SELECT
    a.buy_times,
    COUNT(user_id) AS '人数'
FROM
  -- 有购买行为的用户各自的购买次数
    ( SELECT user_id,COUNT(behavior) AS buy_times FROM
userbehavior WHERE behavior='buy' GROUP BY user_id ) AS a
GROUP BY
    a.buy_times
ORDER BY
  a.buy_times;
```



从上图可以得知整体复购率为(59329-6787)/59329=88.6%，即有购买行为的用户中大概有 88.6%的用户会重复购买。上面是复购情况的可视图，可以看出大部分买家还是只购买一次。

6）找出复购率最高的用户以及他们购买的产品：

```
SELECT
  user_id,
    COUNT(behavior) AS buy_times
FROM
  userbehavior
WHERE
  behavior='buy'
```

```
GROUP BY
  user_id
ORDER BY
  buy_times DESC;
```

user_id	buy_times
337305	93
381745	45
368575	43
347596	40
42193	39
398534	36
400569	36
343823	35

从上面 **SQL** 语句的执行结果可以看到用户 **user_id=337305** 购买次数最多，高达 **93** 次。下面以复购率最高的用户 **user_id=337305** 为例研究说明。

```
SELECT
    category_id,
    COUNT(*)
FROM
    UserBehavior
WHERE
    behavior='buy'
    AND user_id=337305
GROUP BY
    category_id
ORDER BY
    COUNT(*) DESC;
```

category_id	COUNT(*)
2465336	14
4690421	8
3002561	5
4801426	4
4756105	4
4148053	4
4217906	3
4643350	3

可以看出复购率最高用户 user_id=337305 购买的商品类目主要集中在上面表格中的前 3 大类，可以参考这些商品类目的 id 来确定产品种类。

这种针对某些用户做的分析可以更好地了解和发现价值用户，如果数据集提供产品价格信息，就可以通过上面的数据分析很容易地找到高价值用户。了解高价值用户的购买行为，比如购买时间、购买产品以及品类等等以推出有针对性的产品推荐，通过个性化的推荐提高产品销售情况。

五.总结与建议

本次分析利用 MySQL 语句执行，数据集大约有 300 万条淘宝用户行为数据，针对用户行为问题我们使用 AARRR 漏斗模型进行业务分析，结合上述分析的业务指标，下面提出修改建议：

1. 获取客户(Acquisition)：关键点是语言市场匹配和渠道产品匹配。

每天晚上 16 点到 22 点是用户频繁访问的时间，也是获取更多潜在客户的黄金时间，平台开展活动获取客户应首选这个时间段进行。
淘宝是电商第一平台，用户基数大，可以利用用户转发的方式获取新客户，比如在晚间时段做促销活动，邀请朋友拼团享受优惠来增加用户数，适合利用口碑渠道获取新客户。也可以进行小游戏邀请、KOL 推广、热门社交或小视频平台合作推广、淘宝 app 卖家推送等。

2. 激活用户(Activation)：摸清楚产品的“啊哈”时刻，用户从浏览到最终购买整个过程的流失情况，确定夹点位置，提出改善转化率的建议。

用户行为包括点击、加入购物车、收藏以及购买，点击量占总行为的 89.5%，而加入购物车和收藏只占 6%，最后实际购买跌至 2%，夹点位置在收藏和加入购物车环节上，可能出现的原因是用户花了大量时间寻找合适的产品。根据数据分析结果改善转化率的建议有：

（1）优化电商平台的筛选功能，增加关键词的准确率，让用户可以更容易找到合适产品；

（2）给客户提供同类产品比较的功能，让用户不需要多次返回搜索结果反复查看，便于用户确定心怡产品；

（3）精简下单步骤，提供一键下单服务，比如只包含点击-购买-支付三个环节，缩短购买流程，提高用户体验。

3. 第三个环节提高留存（Retention）：让用户养成使用习惯。

让用户保持使用淘宝电商平台的习惯是提高留存率的关键，可采用的方案可能有：

（1）按照使用频率和购买次数积攒积分，每天上线点击量达到某个数值即可自动领取积分，到月末换取购物礼券；

（2）对于年购买次数和金额达到规定量的客户推出 VIP 服务，享受全场不限时 9.5 折优惠，购买次数同比上升之后相应福利也上升，利用这种方法提高高价值用户的留存率和对平台的忠诚度。

4. 第四个环节增加收入（Revenue）：提高成交转化率、复购率及产品和类目的购买率情况。

独立用户从点击到最后购买的转化率约为 68.2%，用户购买诚意还是很足的，所以通过合理优化电商平台的筛选功能可以提高最终购买的转化率。

有购买行为的用户中，大概有 88.6%的用户会重复购买。在独立用户中，最高的复购次数是 93 次，我们可以通过复购率、购买金额（本次数据集没有提供）等来确定价值用户，通过分析找出价值用户的购买偏好，产品和类目等，给价值用户制定个性化的产品推荐，从而提高用户体验和电商平台销售情况。

83.3%的产品贡献了 64.5%的销售量，27.7%的产品类目贡献了 1.6%的销售量，69.6%的产品类目贡献了 11.5%的销售量，不符合传统零售业的二八法则，电商靠长尾理论累计销售。

以上数据显示淘宝平台的最大优势是产品种类和类目丰富，用户可选择的范围非常广，吸引不同类型的客户群，所以应该继续保持这个优势。可能合适的提高方案有：

（1）内容营销：使用“没有找不到的产品，只有想不到的产品”来宣传平台购物种类丰富，让用户形成“只要买东西上淘宝一定有”的思维习惯；

（2）针对前面确定的价值用户提供个性化产品推荐，比如最关心的产品类目和种类，上新之后定时推送给用户；

（3）针对复购率，可以推出 3 个月内复购优惠活动，让客户保持购买频率。

5. 第五个环节推荐（Refer）:用户推荐给其他人，关注转发率、转化率和 K 因子。

针对淘宝平台，让用户推荐给其他人的方案有：

（1）产品在购买的时候提供拼团服务，让用户主动推荐给其他人；

（2）每当推出新功能，比如前面提到的一键下单，让体验过的用户转发和分享领取优惠券，快速实现新功能推广；

（3）当用户使用优惠券购物或者通过某种行为积分购物之后提供朋友圈打卡功能，分享给好友，实现传播功能。

在实行以上方案之后需要关注转发率、转化率、通过用户分享链接点击购买的用户比例以及 K 因子来检测提出方案的有效性。

本文来源：https://segmentfault.com/a/1190000023242953