电商行为数据分析

python+Mysql+Tableau 电商用户行为数据分析实战_使用python对mysql进行数据分析案例-CSDN博客

电商用户行为数据分析实战（MySQL +PowerBI）_powerbi和mysql-CSDN博客

数据集

阿里移动推荐算法数据集的全部Notebook信息_数据集-阿里云天池 (aliyun.com)
天池数据集_阿里系唯一对外开放数据分享平台-阿里云天池 (aliyun.com)
fork from 淘宝用户购物行为数据可视化分析_天池notebook-阿里云天池 (aliyun.com)

一、分析背景及数据说明

1. 背景及目的

中国电商行业经过初期的粗狂式发展，从有货就能卖的模式逐渐转变到精细化运营的模式，通过对大量数据进行深入分析，发现数据背后的用户需求逐渐伴随在电商运营的工作中。随着电商行业发展日趋成熟，加上对于数据的重视，数据基础平台以及数据库的完善，所收集到的数据更加完整，对于分析提供了强有力的支持，同时通过数据分析来为企业经营提供决策变得越来越重要，本文在这个背景下，基于电商用户数据开展分析

2. 数据说明

2.1 来源

数据集：
通过百度网盘分享的文件：基于100万真实电商用户的1亿条行为数据分析
链接：https://pan.baidu.com/s/13-9VFu-6Gu5g38RAuK0tmQ 提取码：dmwf
淘宝用户购物行为数据集_数据集-阿里云天池

2.2 数据相关信息

Behavior type	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

二、数据处理

3.1 修改列名、去空、去重、去异常

```sql
use taobao;
desc user_behavior;
select * from user_behavior limit 5;

-- 改变字段名
alter table user_behavior change timestamp timestamps int(14);
desc user_behavior;

-- 检查空值
select * from user_behavior where user_id is null;
select * from user_behavior where item_id is null;
select * from user_behavior where category_id is null;
select * from user_behavior where behavior_type is null;
select * from user_behavior where timestamps is null;

-- 检查重复值
select user_id,item_id,timestamps from user_behavior
group by user_id,item_id,timestamps
having count(*)>1;

-- 去重
alter table user_behavior add id int first;
select * from user_behavior limit 5;
alter table user_behavior modify id int primary key auto_increment;

delete user_behavior from
user_behavior,
(
select user_id,item_id,timestamps,min(id) id from user_behavior
group by user_id,item_id,timestamps
having count(*)>1
) t2
where user_behavior.user_id=t2.user_id
and user_behavior.item_id=t2.item_id
and user_behavior.timestamps=t2.timestamps
and user_behavior.id>t2.id;

-- 新增日期：date time hour
-- 更改buffer值
-- Buffer Pool(缓冲池)
show VARIABLES like '%_buffer%';  -- 查看当前缓冲池大小
set GLOBAL innodb_buffer_pool_size=1070 000 000;  -- 设缓冲池为10G  535000000
--设置缓冲值大小要看自己的电脑的内存大小的70%左右。
--像我电脑16G内存，但是可能15G能用，设置了缓冲5G，总内存就有12G在使用中了
```

```sql


-- datetime
alter table user_behavior add datetimes TIMESTAMP(0); -- 秒后位数为0，没有毫秒等


-- 将字段类型为int（11）的事件类型转换为年月日
update user_behavior set datetimes=FROM_UNIXTIME(timestamps);
select * from user_behavior limit 5;


-- date
alter table user_behavior add dates char(10);
alter table user_behavior add times char(8);
alter table user_behavior add hours char(2);


-- 一次性插入三个语段
-- update user_behavior set dates=substring(datetimes,1,10),times=substring(datetimes,12,8),hours=substring(datetimes,12,2);
-- 分次插入
update user_behavior set dates=substring(datetimes,1,10);
update user_behavior set times=substring(datetimes,12,8);
update user_behavior set hours=substring(datetimes,12,2);
select * from user_behavior limit 5;

-- 去异常
select max(datetimes),min(datetimes) from user_behavior;
delete from user_behavior
where datetimes < '2017-11-25 00:00:00'
or datetimes > '2017-12-03 23:59:59';

-- 数据概览
desc user_behavior; -- 查看特定表的详细设计信息
select * from user_behavior limit 5;
SELECT count(1) from user_behavior; -- 100095496条记录
```

3.5 一致化处理

```sql
-- datetime
alter table user_behavior add datetimes TIMESTAMP(0); -- 秒后位数为0，没有毫秒等


-- 将字段类型为int（11）的事件类型转换为年月日
update user_behavior set datetimes=FROM_UNIXTIME(timestamps);
select * from user_behavior limit 5;


-- date
alter table user_behavior add dates char(10);
alter table user_behavior add times char(8);
alter table user_behavior add hours char(2);


-- 一次性插入三个语段
-- update user_behavior set dates=substring(datetimes,1,10),times=substring(datetimes,12,8),hours=substring(datetimes,12,2);
-- 分次插入
update user_behavior set dates=substring(datetimes,1,10);
update user_behavior set times=substring(datetimes,12,8);
update user_behavior set hours=substring(datetimes,12,2);
select * from user_behavior limit 5;
```

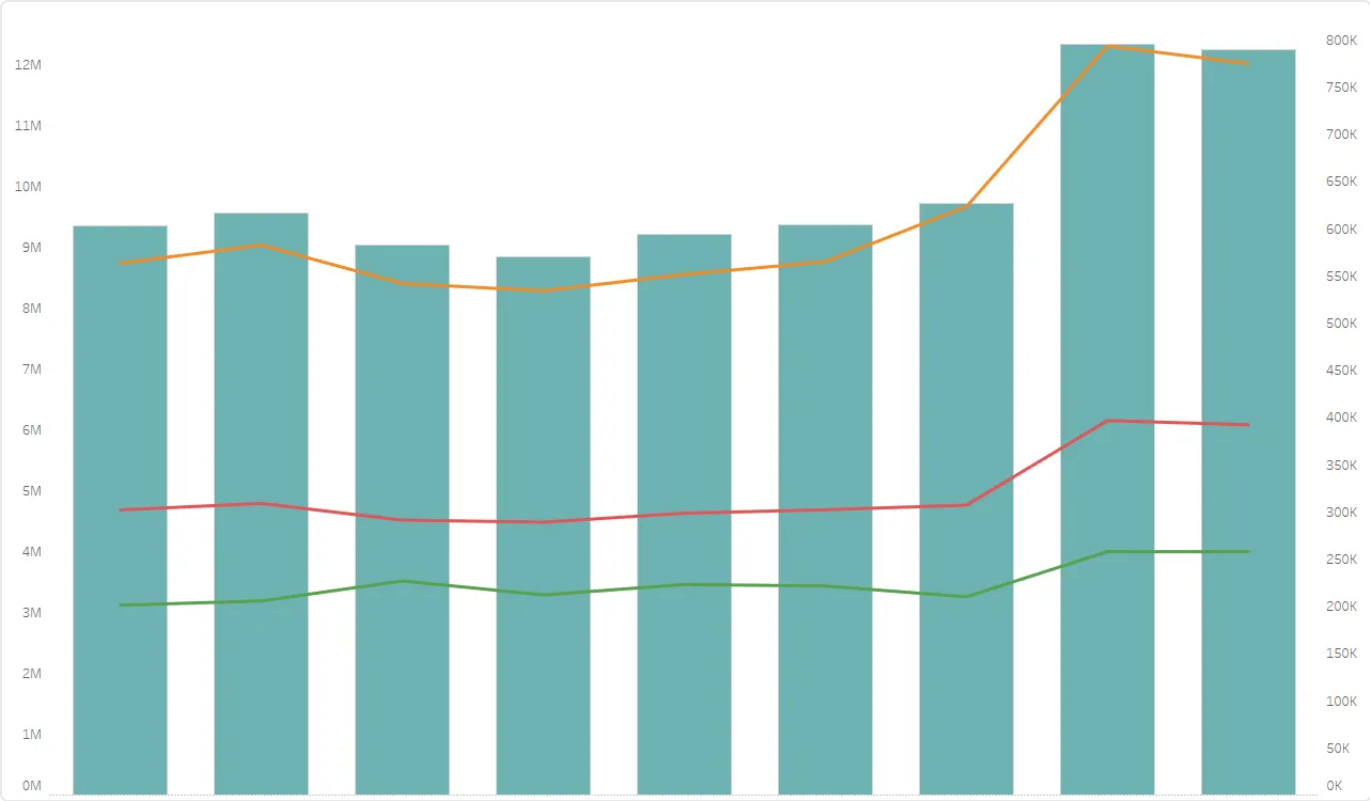
三、数据分析

4.1 流量分析

4.1.1 基于用户行为进行分析

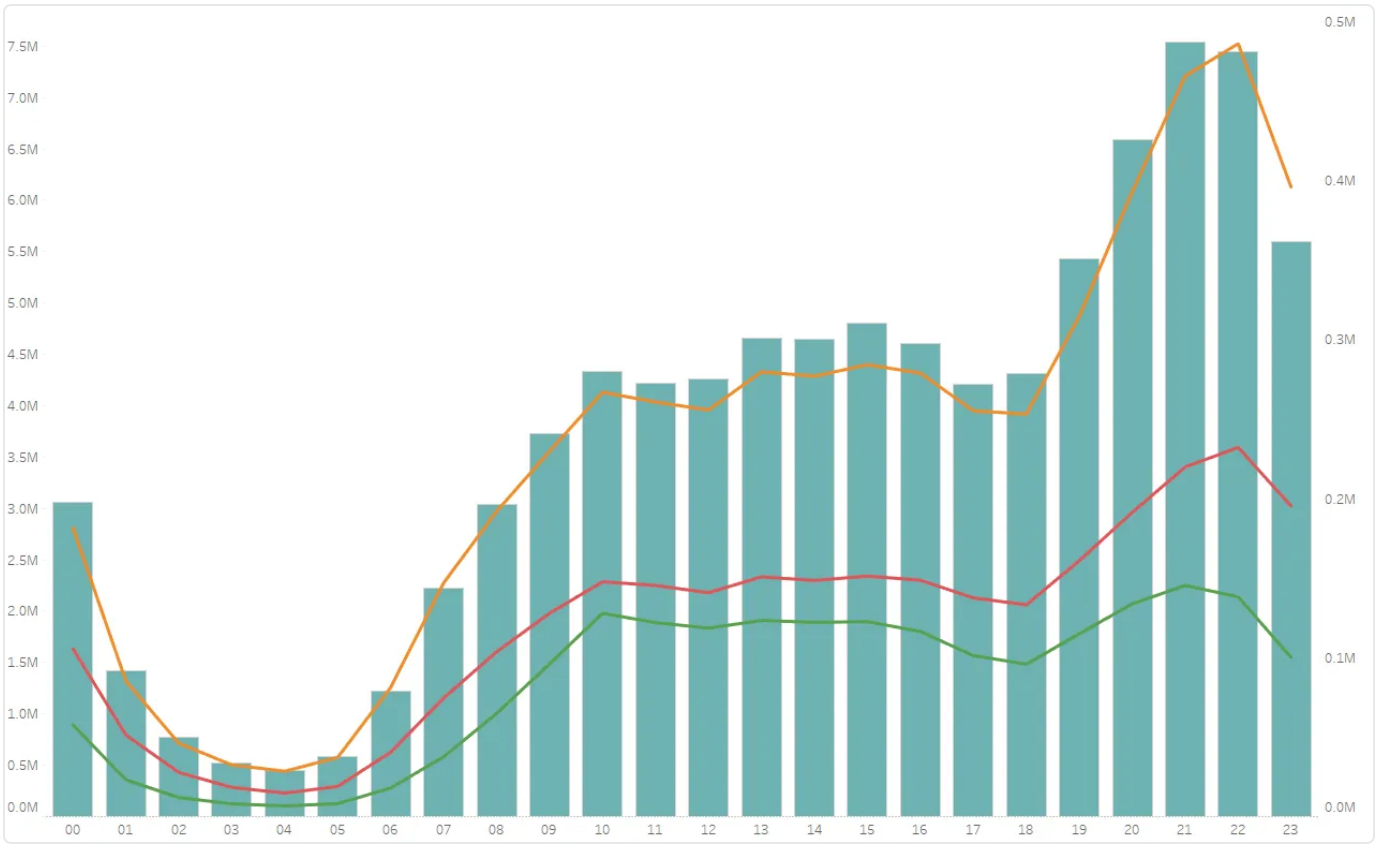
每天的用户行为分析

可以看出，每天四种行为（浏览、收藏、加购、购买）的变化趋势基本是一致的，浏览量在 12 月 2 日猛增，其原因可能是双十二到来，活动大促开始，吸引用户浏览各类商品，为双十二购物做准备



每小时的用户行为分析

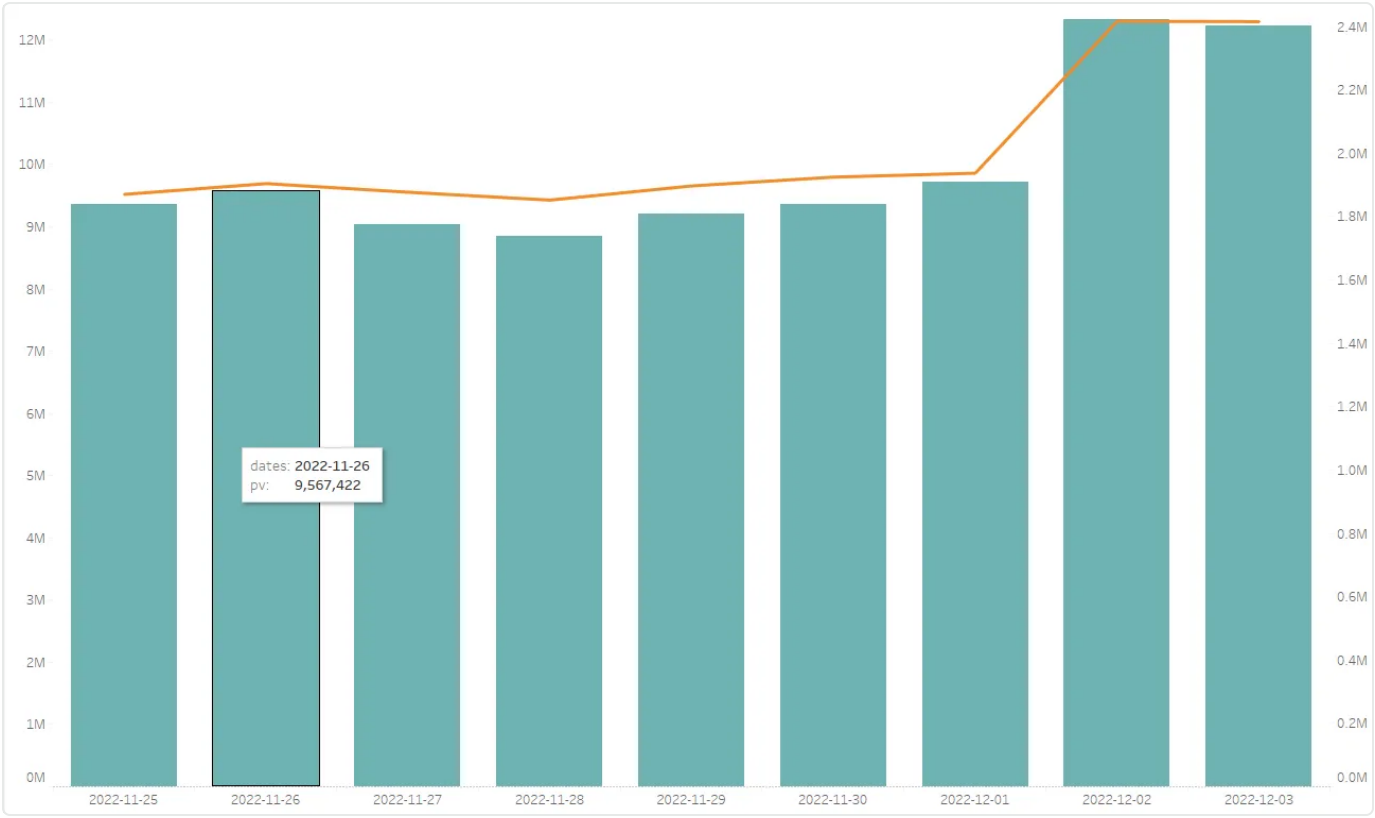
用户访问量在凌晨四点到达嘀咕，晚上 9–10 点到达顶峰，这可能是因为淘宝用户大多是工薪阶级，晚上 8 点之后，开始有时间购买商品，因此到达顶峰



4.1.2 基于用户数进行分析

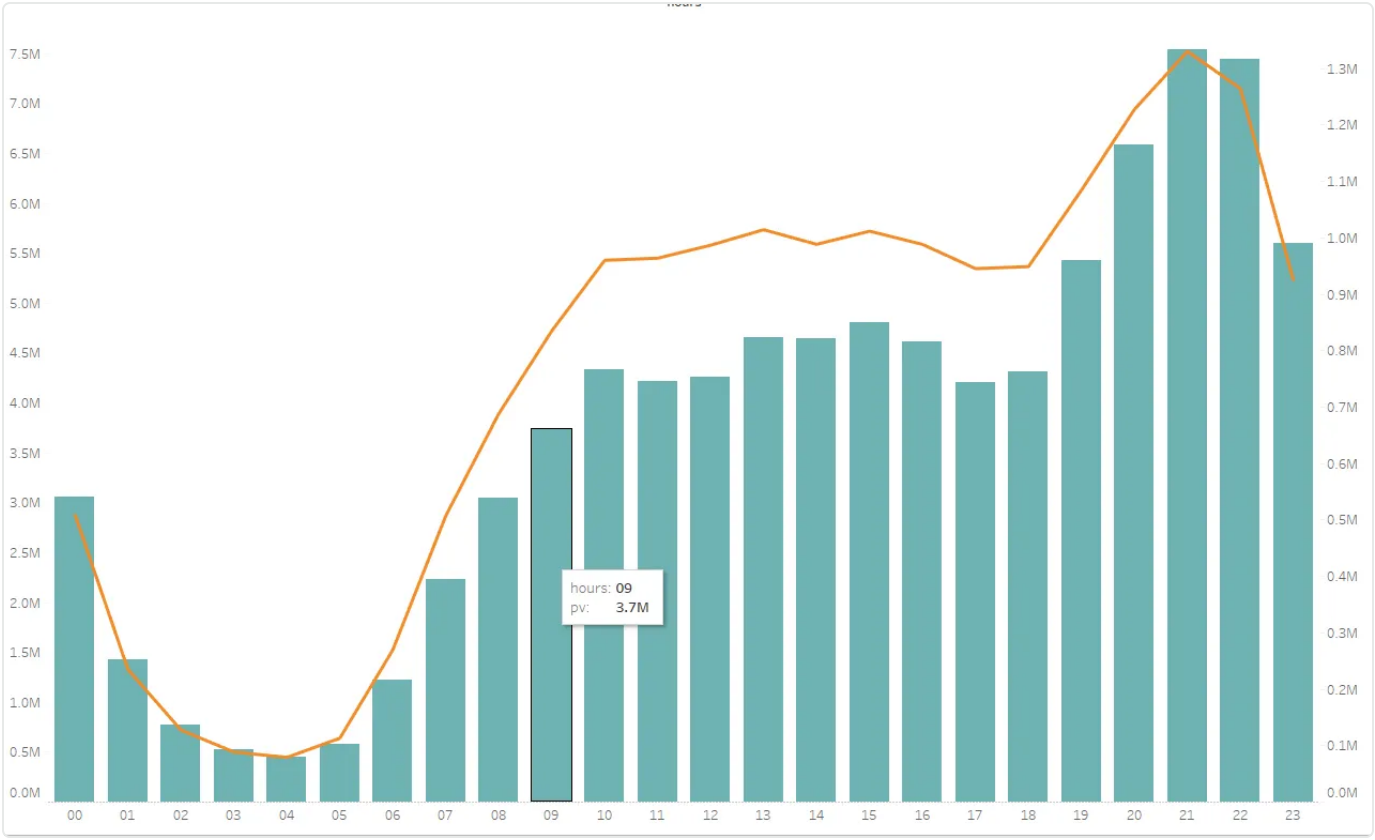
每天的 uv 数

用户数与浏览数基本呈正比



每小时的 uv 数

可以看出，用户数自 18：00 以后开始出现猛烈上升 ，这最大的原因可能是淘宝的主力消费人群是工薪阶级，18：00 是下班的时间，在下班后，他们开始使用淘宝，浏览商品；凌晨四点是 uv 最低的时间，这也符合人们的作息习惯；上午 10 点到下午 18 点期间，用户数基本稳定，不会产生很大的波动



4.2 用户流失分析

本次分析使用漏斗模型分析、假设验证分析。

在用户维度，使用漏斗模型分析顾客点击、加购、收藏、购买不同阶段的转化率，找出转化率在一天中不同时段的分布情况，针对地在转化率高的时段加大投放，争取最大化提高转化率，从而提高流量的利用和店铺销售额

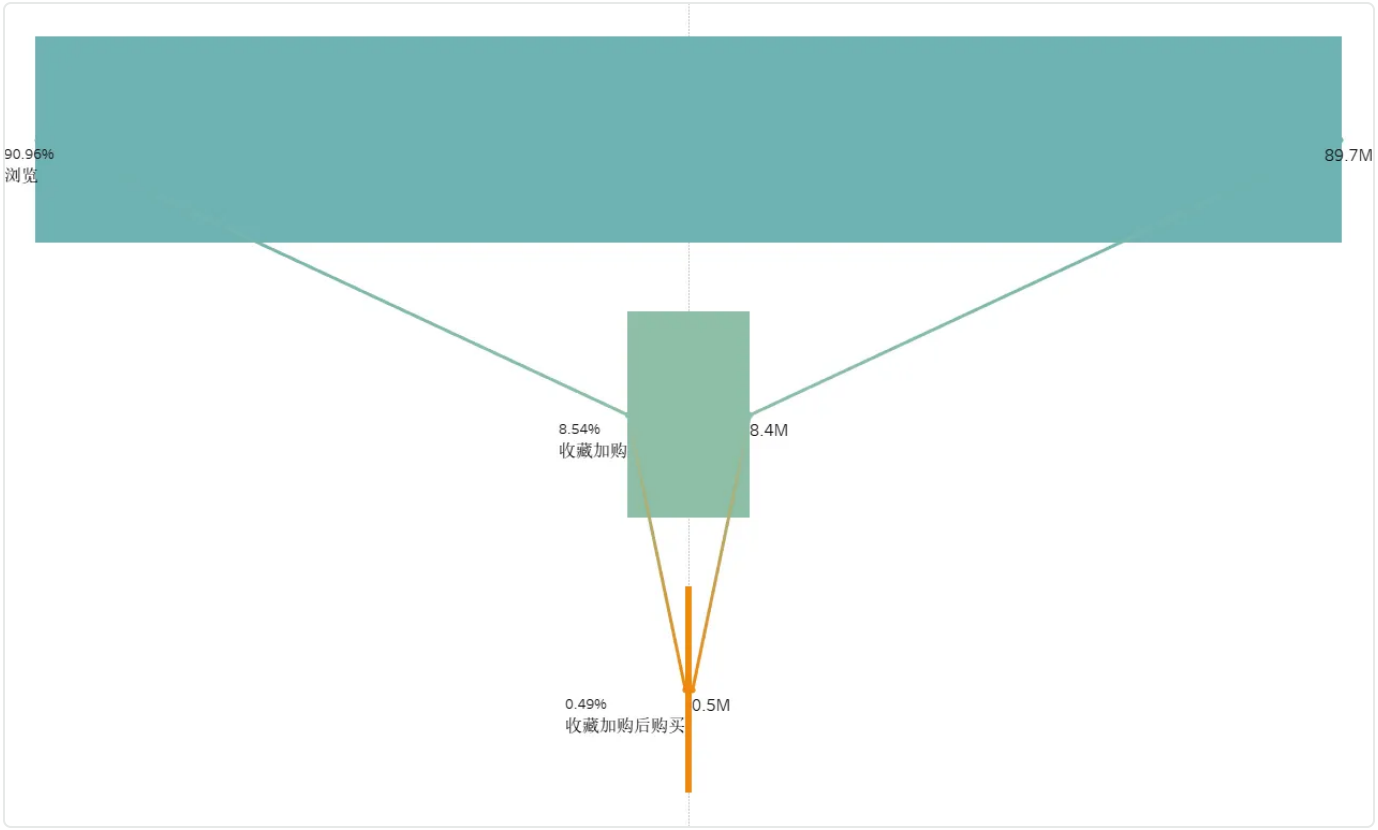
在商品维度，通过假设验证，分析流量商品与畅销产品的分布情况，优化新品引进、付费推广的投入、产品备货量和备货周期

1.1. 用户情况分析

1.1.1. 用户情况

商品类别	商品数	浏览数
9,437	4,161,140	89,660,986

加购数	收藏数	购买数
5,530,445	2,888,258	2,015,807



用户的行为包括点击、加入购物车、收藏以及购买，图 1 显示，浏览占总行为数的 90.96%，而收藏加购的只占 8.54%，购买占比只有不到 0.5%，从漏斗图，我们分析得到，用户流失主要在收藏加购这一环节。

于是，我们作出假设：出现这种情况的原因可能是，用户花了大量的时间浏览商品搜寻不到自己需要的产品，以至于放弃搜做，转而去其他平台购买或者放弃购买。

针对这个假设，我们从以下 3 个维度进行分析，验证此假设：（商品、用户、平台）

1. 用户想要找到什么样的商品
2. 用户习惯什么时间购买
3. 平台推送的商品是否可以满足用户的需求

1.1.2. 用户流失情况原因分析

（1）用户想要找到什么样的商品

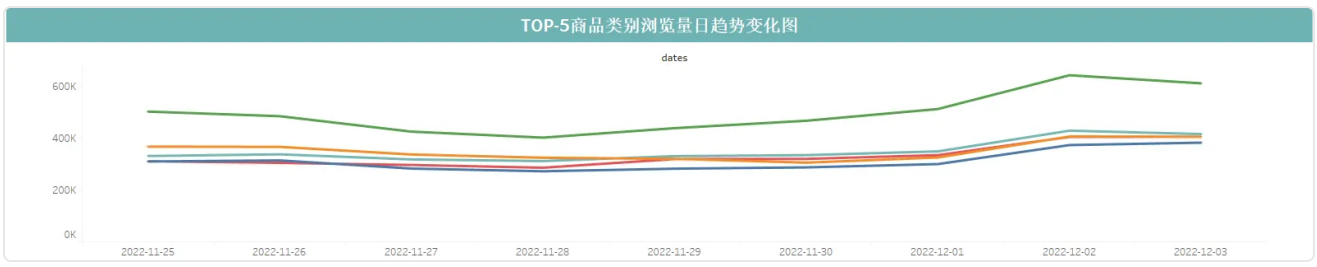
衡量用户最想要的商品，可以从商品、品类的点击率进行分析。通过这项指标，可以最大程度地分析那类商品的用户需求量比较大，哪类商品的需求量比较小

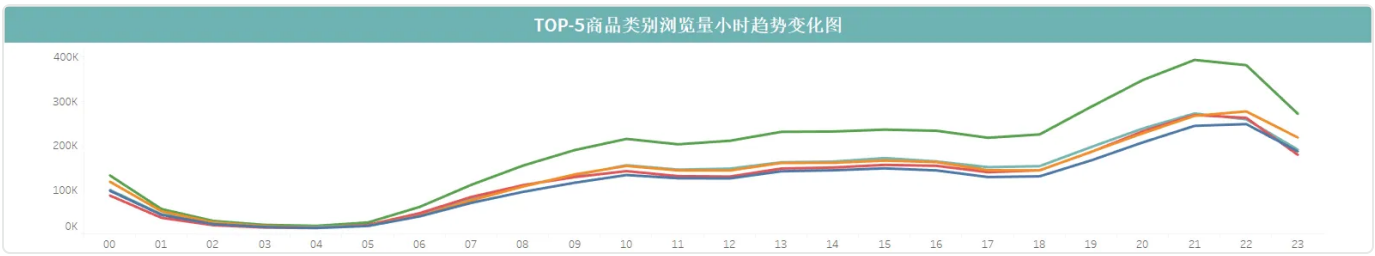
从图中可以看出，热门商品类别前五名是 4756105、2355072、4145813、3607361、982926 ，浏览量最高，说明用户对这五类商品的需求量的最高

前五类热门商品类别中，浏览量最高的五种商品如上图所示。销冠当属编号为 812879 的商品，浏览量为 30079，当前类别中的商品浏览量都在一万以上。

（2）用户习惯什么时候购买商品

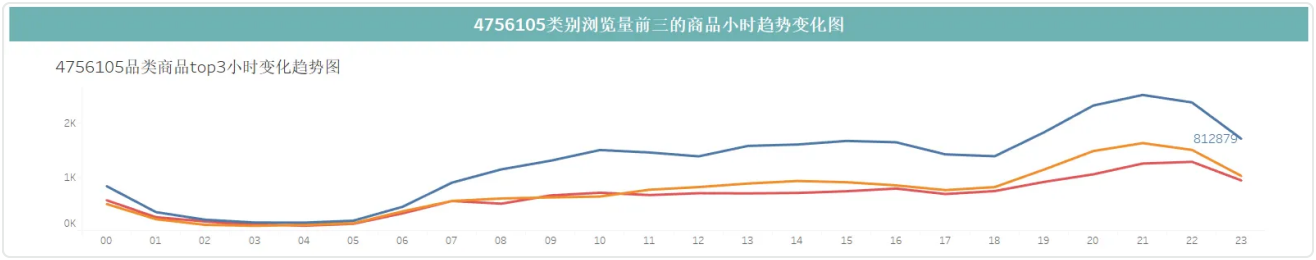
根据上面的结论，分析每天、每小时用户对物类商品的浏览量。





TOP-5商品类别浏览量日趋势变化图呈现了九天中用户前五名热门商品的浏览量变化趋势，从 12 月 2 日开始用户对各品类的浏览量逐渐增加。

从TOP-5商品类别浏览量小时趋势变化图中可以看出，用户的点击量主要集中在晚上 18 点-00 点，在这期间，用户的浏览量迅速上升，并在 **21 点，到达当日浏览量的最高值**。2 点之后点击量逐渐下降，4 点左右到达当日低谷，早晨 5 点到 10 点，浏览量缓慢上升，10 点-18 点期间浏览量比较平稳。



再看4756105 商品类别浏览量前三的商品（812879，4211339，4649427）日趋势变化图，浏览量趋势与商品类别趋势基本一致，图中可以看到，18-00 点，用户明显活跃起来。对此，有营销活动的商家可以在这段时间加大活动宣传力度，争取流量获得最大转化。

综上分析可知，用户在平台最想购买的商品类别是 4756105、2355072、4145813、3607361、982926 这五类商品，在这五类商品中，888 的需求量是最高的，用户主要在下午 18-00 点之间浏览这几类商品。

因为数据集有限，只能大致分析出用户想要寻找哪几类商品。若要更细致地分析出用户想要什么样的商品，还需知道用户使用的搜索高频词，利用该项数据建立用户搜索画像，并结合商品点击数数据，建立搜索点击率指标，分析总结出点击率高的搜索高频词和点击率低的搜索高频词。从而更精确地总结出用户在淘宝平台最想要寻找什么商品。

（3）平台推送的商品是否满足用户需求

从上一步的分析中，我们可以知道用户想要寻找什么样的商品，接下来需要知道平台推送的商品是否满足了用户需求。

首先，从商品数量占比上分析这五类商品是否有宠物的商品供用户选择。

Top-5商品占比

categories 总数

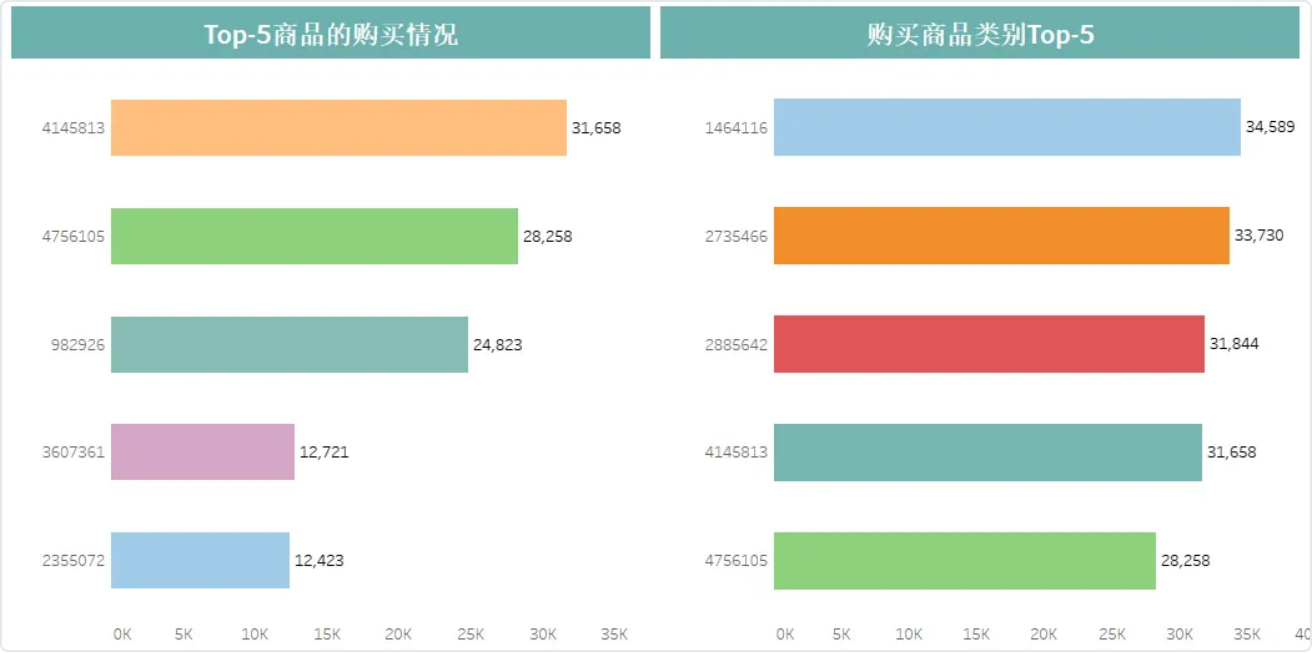
9,437

item_id 总数

4,161,140

category_id	num	ratio
4145813	68,328	1.64%
2355072	57,568	1.38%
982926	54,331	1.31%
4756105	47,851	1.15%
3607361	35,341	0.85%

从图上的结果可以看出，淘宝共有 9437 个大类商品，下属商品共 4161140 种商品，而从右图中看到热度最高的物类商品的商品数占比都比较低，说明淘宝平台提供的选择过少，这会使用户搜索很长时间，我发货的想要的商品，从而导致降低用户的购买欲，转战其他平台。



从上图可以看出，用户购买商品类别 Top-5 除 4145813、4756105 两类热门商品类别外，其他浏览量高的商品的购买率并不高，说明淘宝对浏览量高的五类商品的推荐机制并不合理，没有在用户进行搜索浏览时，满足用户的需求，**即用户在点击查看该类商品后，发现并不是自己想要的商品放弃加入购物车，造成转化率的降低。**

1.2. 建议

根据上述分析，证明了前面的假设，淘宝用户流失的原因是，用户在淘宝上花了大量的时间搜索浏览商品，但无法找到自己需要的产品，因此放弃购买。针对此，我有几点建议：

1. 推荐算法部门建议：建议淘宝算法部门优化推荐算法，针对需求最高的品类（ 4756105、2355072、4145813、3607361、982926）进行监控，分析这物类商品的购买书，将搜索数最高的商品优先推送给用户，并在搜索界面将浏览量最高的商品放在前面，使用户可以更快的找到热门商品，缩小用户寻找商品的时间，提高用户的转化率。
2. 市场部门建议：用户对 4756105、2355072、4145813、3607361、982926 品类的商品需求量比较高，因此建议市场部门可以增加这五类商品的种类，为用户提供更多的选择。
3. 运营部门：淘宝用户浏览商品的时间段主要集中在下午 6 点到晚上十二点，可以在在这个时间段多策划一些营销活动，刺激用户消费，提高转化率。并对在用户的主要搜寻时间对这五类商品中需求很高的商品进行促销活动。