目录

[1. 数据集说明 2](#_Toc32013)

[2. 数据处理 2](#_Toc26853)

[2.1 数据导入 2](#_Toc18947)

[2.2 数据清洗 3](#_Toc19362)

[3.数据分析可视化 3](#_Toc24335)

[3.1 用户流量及购物情况 4](#_Toc1632)

[3.2 用户行为转换率 7](#_Toc11345)

[3.3 用户行为习惯 11](#_Toc25787)

[3.4 基于 RFM 模型找出有价值的用户 15](#_Toc764)

[3.5 商品维度的分析 18](#_Toc16178)

[3.6 时间维度的分析 18](#_Toc15696)

[3.7 用户行为路径分析 19](#_Toc929)

[3.8 结论与建议 19](#_Toc9180)

[3.10 竞品分析 20](#_Toc2579)

[3.11 总结与展望 21](#_Toc24934)

# 1. 数据集说明

这是一份来自淘宝的用户行为数据，时间区间为 2017-11-25 到 2017-12-03，总计 100,150,807 条记录，大小为 3.5 G，包含 5 个字段。

# 2. 数据处理

## 2.1 数据导入

将数据加载到 hive， 然后通过 hive 对数据进行数据处理。

-- 建表drop table if exists user\_behavior;create table user\_behavior (`user\_id` string comment '用户ID',`item\_id` string comment '商品ID',`category\_id` string comment '商品类目ID',`behavior\_type` string comment '行为类型，枚举类型，包括(pv, buy, cart, fav)',`timestamp` int comment '行为时间戳',`datetime` string comment '行为时间')

row format delimited

fields terminated by ','

lines terminated by '\n';

-- 加载数据

LOAD DATA LOCAL INPATH '/home/getway/UserBehavior.csv'

OVERWRITE INTO TABLE user\_behavior ;

## 2.2 数据清洗

数据处理主要包括：删除重复值，时间戳格式化，删除异常值。

--数据清洗，去掉完全重复的数据

insert overwrite table user\_behaviorselect user\_id, item\_id, category\_id, behavior\_type, timestamp, datetimefrom user\_behaviorgroup by user\_id, item\_id, category\_id, behavior\_type, timestamp, datetime;

--数据清洗，时间戳格式化成 datetime

insert overwrite table user\_behaviorselect user\_id, item\_id, category\_id, behavior\_type, timestamp, from\_unixtime(timestamp, 'yyyy-MM-dd HH:mm:ss')from user\_behavior;

--查看时间是否有异常值select date(datetime) as day from user\_behavior group by date(datetime) order by day;

--数据清洗，去掉时间异常的数据

insert overwrite table user\_behaviorselect user\_id, item\_id, category\_id, behavior\_type, timestamp, datetimefrom user\_behaviorwhere cast(datetime as date) between '2017-11-25' and '2017-12-03';

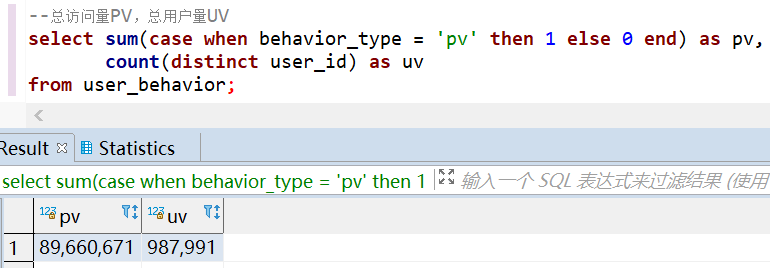
--查看 behavior\_type 是否有异常值select behavior\_type from user\_behavior group by behavior\_type;

# 3.数据分析可视化

## 3.1 用户流量及购物情况

--总访问量PV，总用户量UVselect sum(case when behavior\_type = 'pv' then 1 else 0 end) as pv,

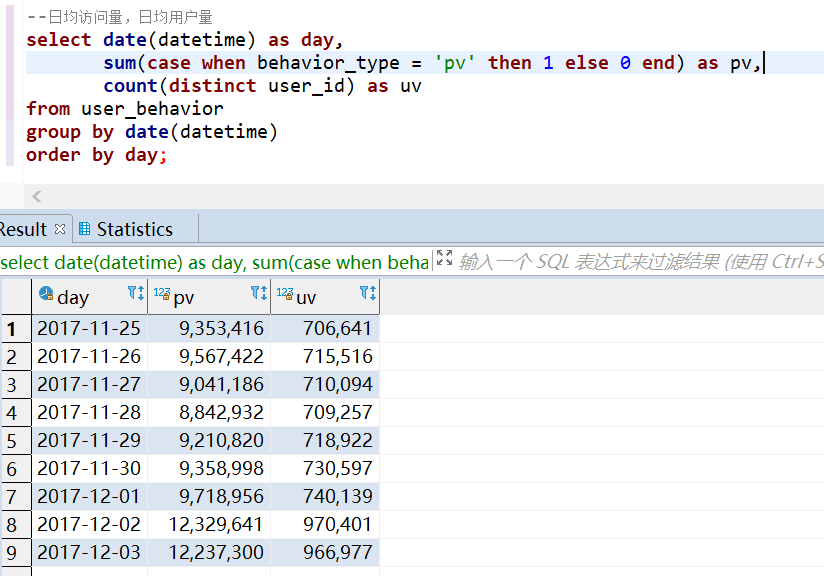
count(distinct user\_id) as uvfrom user\_behavior;

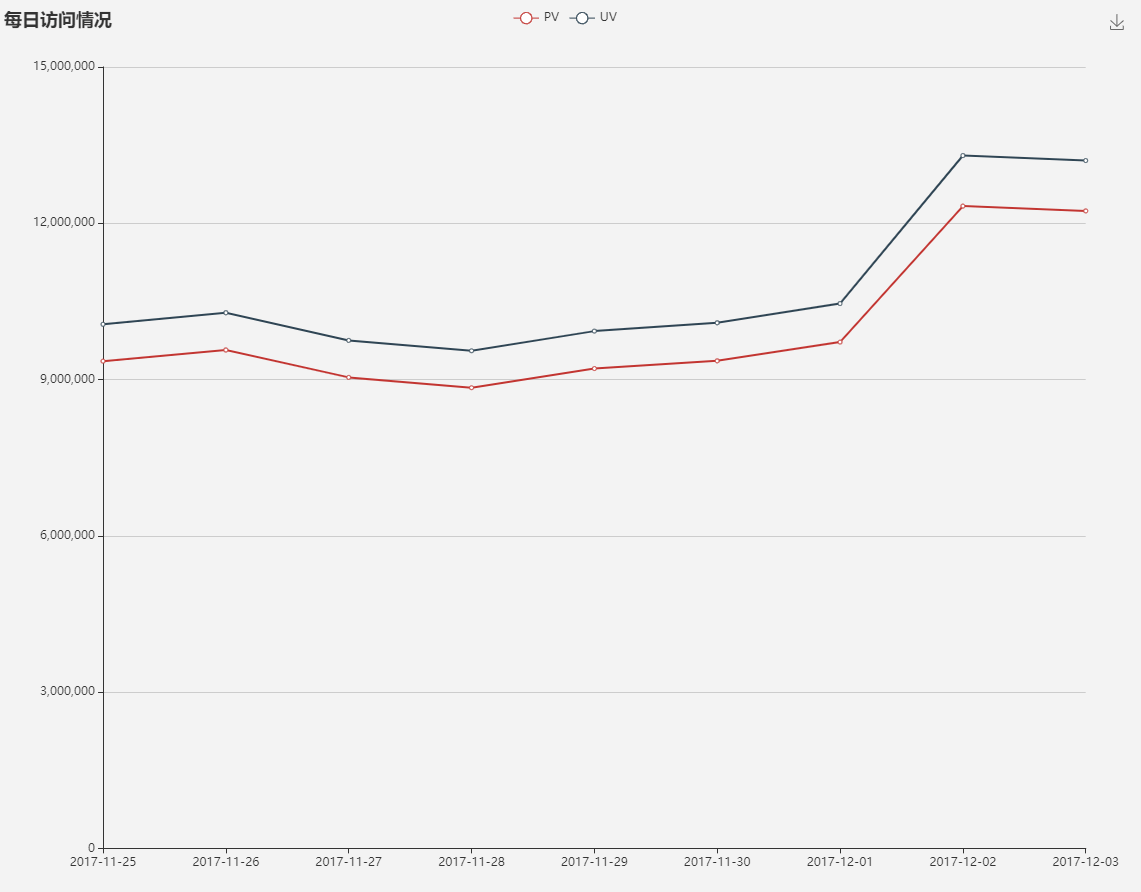
[](https://camo.githubusercontent.com/be6bbe2ff10e6c70951bb716251b27770a97ee25c599e8a40924bca251f8a290/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383134353433363833382e706e67)

--日均访问量，日均用户量select cast(datetime as date) as day,

sum(case when behavior\_type = 'pv' then 1 else 0 end) as pv,

count(distinct user\_id) as uvfrom user\_behaviorgroup by cast(datetime as date)order by day;

[](https://camo.githubusercontent.com/aea1d14f073b024f8934c6196ba580924ce48bfd43ff048c51993105e5f1ed24/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383135313035383237392e706e67)

[](https://camo.githubusercontent.com/64112f0e52b884ffa9891183b6cdc14897c2d32e977a12f2c109c489f73c1191/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383135313533353339332e706e67)

--每个用户的购物情况，加工到 user\_behavior\_countcreate table user\_behavior\_count asselect user\_id,

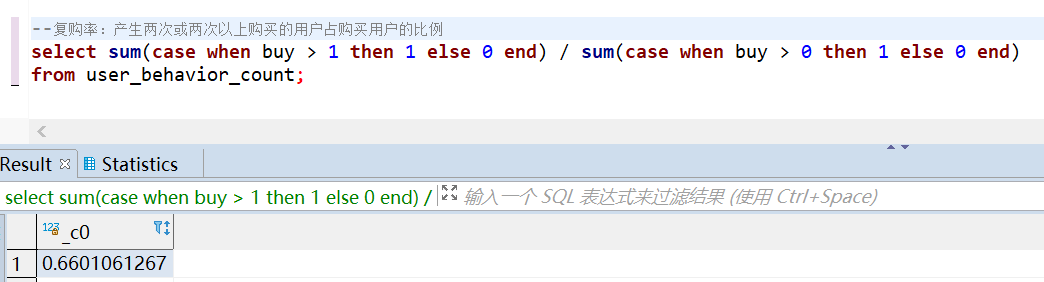
sum(case when behavior\_type = 'pv' then 1 else 0 end) as pv, --点击数

sum(case when behavior\_type = 'fav' then 1 else 0 end) as fav, --收藏数

sum(case when behavior\_type = 'cart' then 1 else 0 end) as cart, --加购物车数

sum(case when behavior\_type = 'buy' then 1 else 0 end) as buy --购买数from user\_behaviorgroup by user\_id;

--复购率：产生两次或两次以上购买的用户占购买用户的比例select sum(case when buy > 1 then 1 else 0 end) / sum(case when buy > 0 then 1 else 0 end)from user\_behavior\_count;

[](https://camo.githubusercontent.com/5d5140b16713ce55603a800f49005e4015c1dc837cd2f6cff325c9a1716c6a85/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383135323030343433322e706e67)

* 小结：2017-11-25 到 2017-12-03 这段时间，PV 总数为 89,660,671 ，UV 总数为 987,991。从日均访问量趋势来看，进入 12 月份之后有一个比较明显的增长，猜测可能是因为临近双 12 ，电商活动引流产生，另外，2017-12-02 和 2017-12-03 刚好是周末，也可能是周末的用户活跃度本来就比平常高。总体的复购率为 66.01%，说明用户的忠诚度比较高。

## 3.2 用户行为转换率

--点击/(加购物车+收藏)/购买 , 各环节转化率select a.pv,

a.fav,

a.cart,

a.fav + a.cart as `fav+cart`,

a.buy,

round((a.fav + a.cart) / a.pv, 4) as pv2favcart,

round(a.buy / (a.fav + a.cart), 4) as favcart2buy,

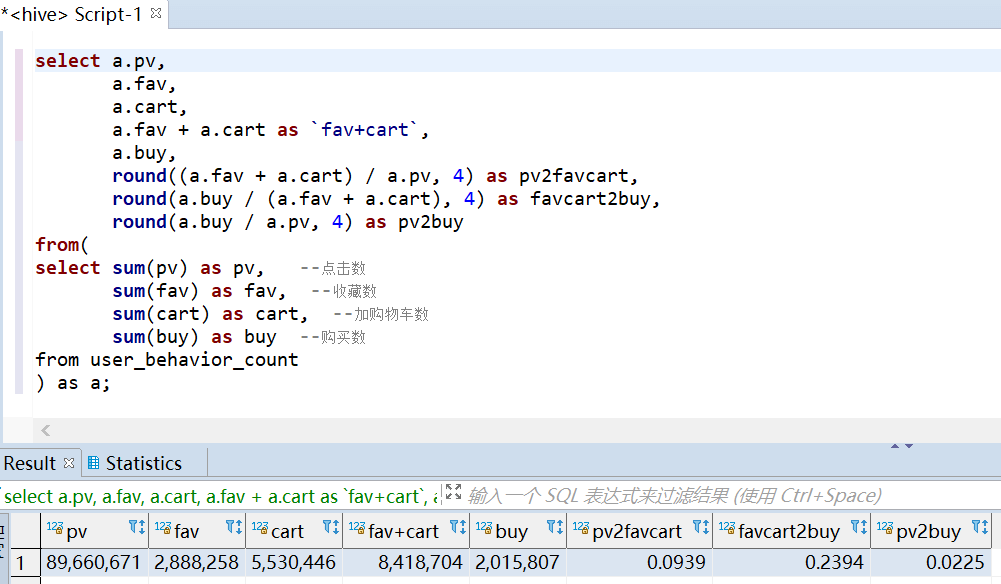
round(a.buy / a.pv, 4) as pv2buyfrom(select sum(pv) as pv, --点击数

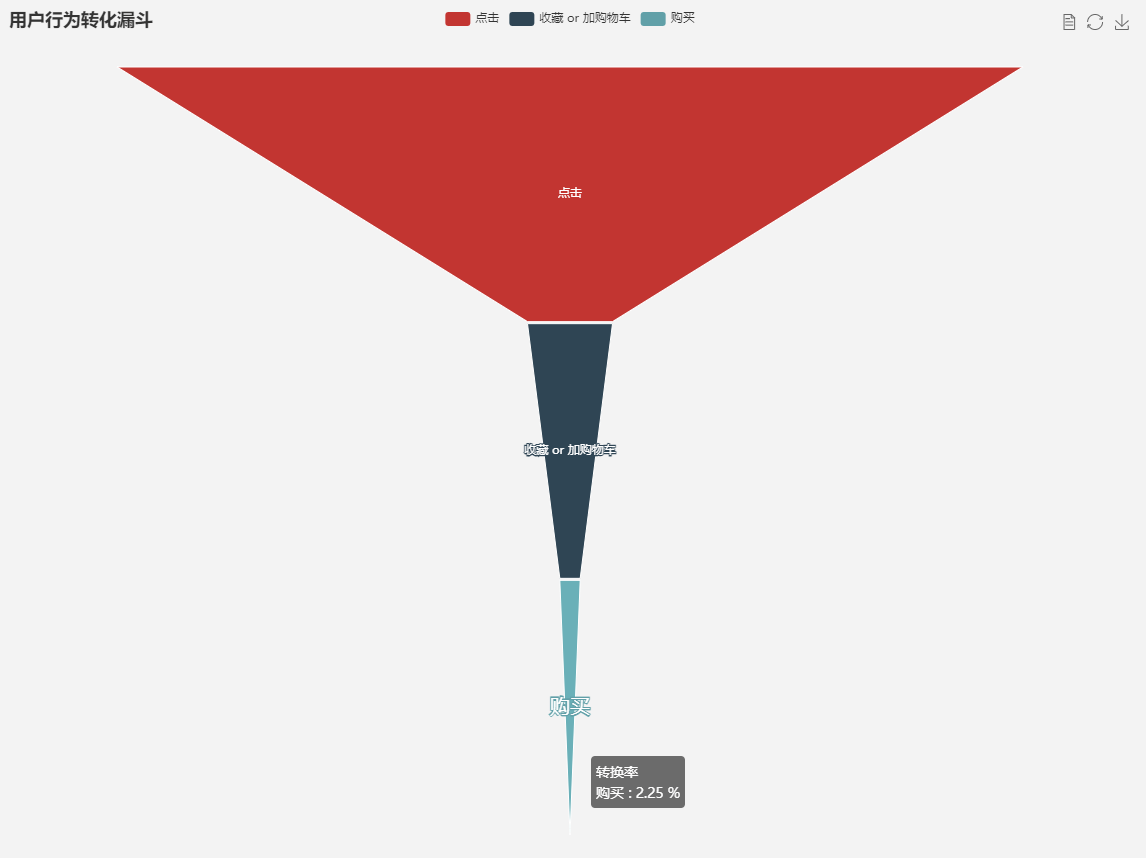
sum(fav) as fav, --收藏数

sum(cart) as cart, --加购物车数

sum(buy) as buy --购买数from user\_behavior\_count

) as a;

[](https://camo.githubusercontent.com/70de1d34d9747ba6e2ed7a758c2f24ae906cf4719264f8e39e6f3b724d84c26d/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383134343935383735372e706e67)

[](https://camo.githubusercontent.com/655193469878152969919075383bdf9f254fdc74ef0412a087d74067a0eacb41/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383134343831343737332e706e67)

* 小结：2017-11-25 到 2017-12-03 这段时间，点击数为 89,660,671 ，收藏数为 2,888,258，加购物车数为5,530,446，购买数为 2,015,807。总体的转化率为 2.25%，这个值可能是比较低的，从加到购物车数来看，有可能部分用户是准备等到电商节日活动才进行购买。所以合理推断：一般电商节前一段时间的转化率会比平常低。

## 3.3 用户行为习惯

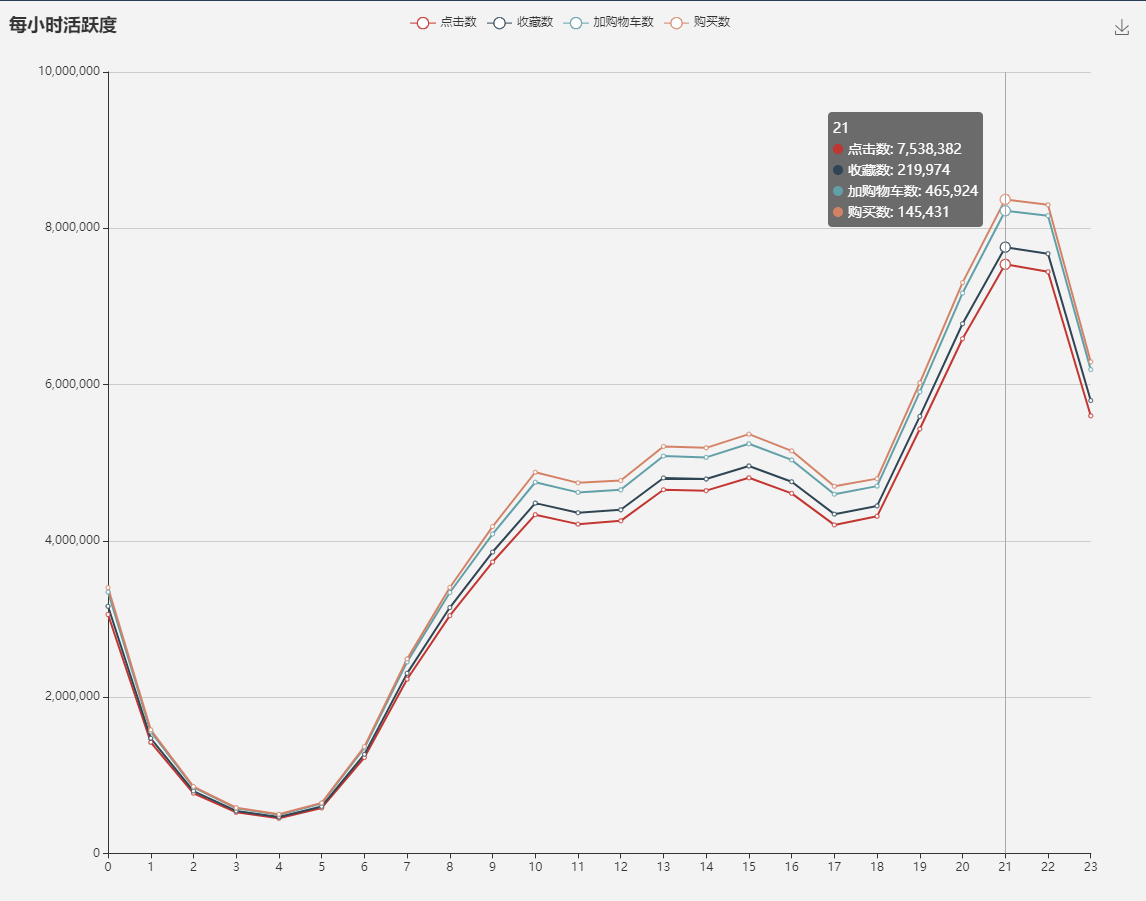
-- 一天的活跃时段分布select hour(datetime) as hour,

sum(case when behavior\_type = 'pv' then 1 else 0 end) as pv, --点击数

sum(case when behavior\_type = 'fav' then 1 else 0 end) as fav, --收藏数

sum(case when behavior\_type = 'cart' then 1 else 0 end) as cart, --加购物车数

sum(case when behavior\_type = 'buy' then 1 else 0 end) as buy --购买数from user\_behaviorgroup by hour(datetime)order by hour;

[](https://camo.githubusercontent.com/4a02d31d9faad9a11e379c1c0345c7d15646ea82d239431add7ff9829103ca2f/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383135333230363934372e706e67)

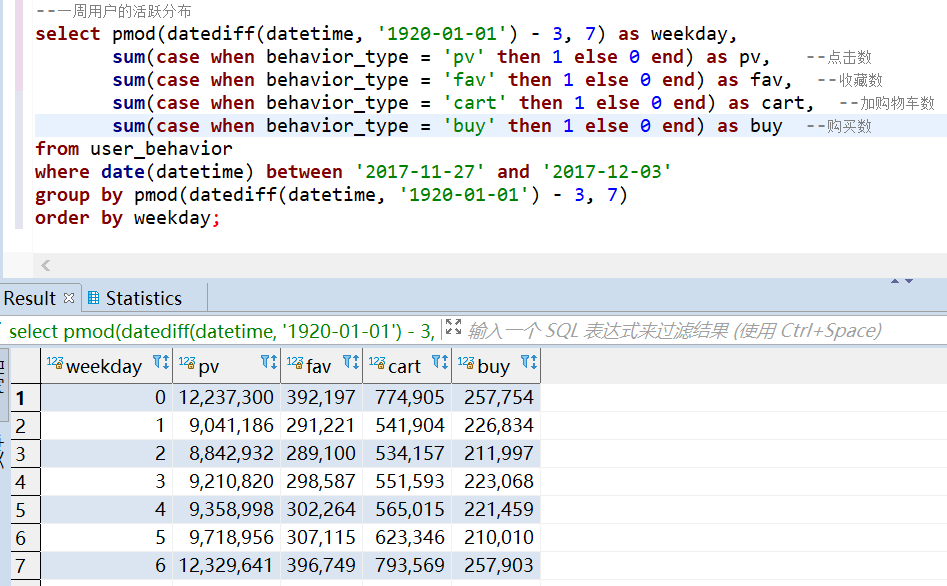
--一周用户的活跃分布select pmod(datediff(datetime, '1920-01-01') - 3, 7) as weekday,

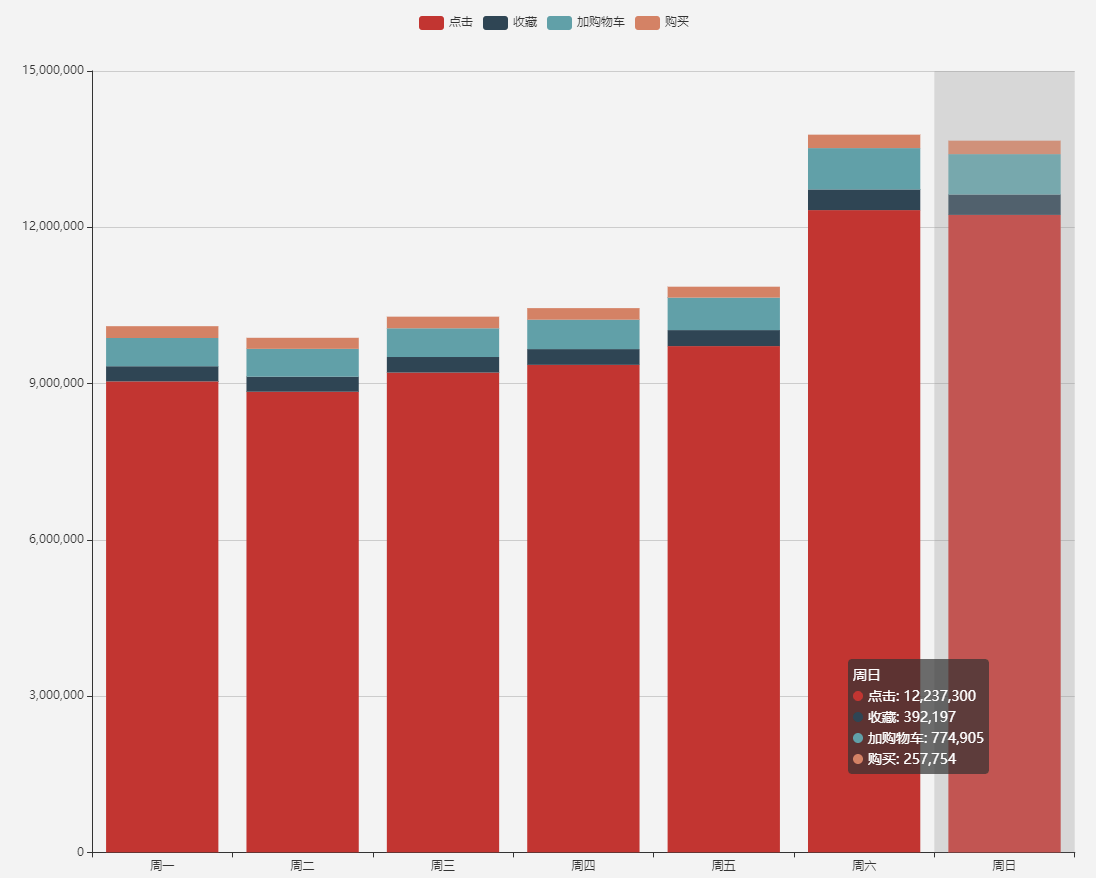
sum(case when behavior\_type = 'pv' then 1 else 0 end) as pv, --点击数

sum(case when behavior\_type = 'fav' then 1 else 0 end) as fav, --收藏数

sum(case when behavior\_type = 'cart' then 1 else 0 end) as cart, --加购物车数

sum(case when behavior\_type = 'buy' then 1 else 0 end) as buy --购买数from user\_behaviorwhere date(datetime) between '2017-11-27' and '2017-12-03'group by pmod(datediff(datetime, '1920-01-01') - 3, 7)order by weekday;

[](https://camo.githubusercontent.com/d8951b820928b92ca3de7efc483d0932d516b5b0a8f11aa0ea353c7ade96b358/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383135333735313934332e706e67)

[](https://camo.githubusercontent.com/8a1da491012682e4c863c4f72c616f21f30604093d4bbf22050dc85061308793/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383135343533333936382e706e67)

* 小结：晚上21点-22点之间是用户一天中最活跃的时候，凌晨 4 点，则是活跃度最低的时候。一周中，工作日活跃度都差不多，到了周末活跃度有明显提高。

## 3.4 基于 RFM 模型找出有价值的用户

RFM 模型是衡量客户价值和客户创利能力的重要工具和手段，其中由3个要素构成了数据分析最好的指标，分别是：

* R-Recency（最近一次购买时间）
* F-Frequency（消费频率）
* M-Money（消费金额）

--R-Recency（最近一次购买时间）, R值越高，一般说明用户比较活跃select user\_id,

datediff('2017-12-04', max(datetime)) as R,

dense\_rank() over(order by datediff('2017-12-04', max(datetime))) as R\_rankfrom user\_behaviorwhere behavior\_type = 'buy'group by user\_idlimit 10;

--F-Frequency（消费频率）, F值越高，说明用户越忠诚select user\_id,

count(1) as F,

dense\_rank() over(order by count(1) desc) as F\_rankfrom user\_behaviorwhere behavior\_type = 'buy'group by user\_idlimit 10;

--M-Money（消费金额），数据集无金额，所以就不分析这一项

对有购买行为的用户按照排名进行分组，共划分为5组， 前 - 1/5 的用户打5分 前 1/5 - 2/5 的用户打4分 前 2/5 - 3/5 的用户打3分 前 3/5 - 4/5 的用户打2分 前 4/5 - 的用户打1分 按照这个规则分别对用户时间间隔排名打分和购买频率排名打分，最后把两个分数合并在一起作为该名用户的最终评分

with cte as(select user\_id,

datediff('2017-12-04', max(datetime)) as R,

dense\_rank() over(order by datediff('2017-12-04', max(datetime))) as R\_rank,

count(1) as F,

dense\_rank() over(order by count(1) desc) as F\_rankfrom user\_behaviorwhere behavior\_type = 'buy'group by user\_id)

select user\_id, R, R\_rank, R\_score, F, F\_rank, F\_score, R\_score + F\_score AS scorefrom(select \*,

case ntile(5) over(order by R\_rank) when 1 then 5

when 2 then 4

when 3 then 3

when 4 then 2

when 5 then 1

end as R\_score,

case ntile(5) over(order by F\_rank) when 1 then 5

when 2 then 4

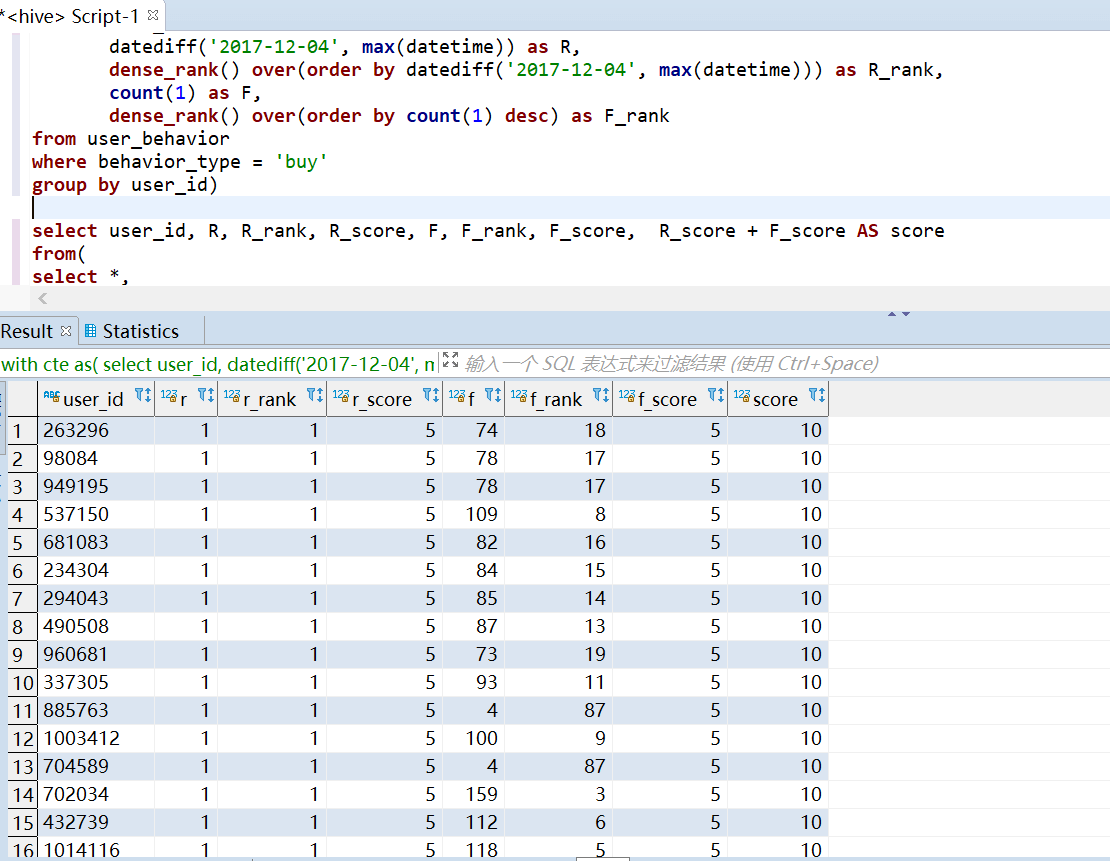
when 3 then 3

when 4 then 2

when 5 then 1

end as F\_scorefrom cte

) as aorder by score desclimit 20;

[](https://camo.githubusercontent.com/483218c0702338ead8214e9f656ef2daf98979f02bc018cbab060c332d00095f/687474703a2f2f7069632e747572626f7761792e746f702f626c6f67696d672f696d6167652d32303230313232383135353730303634362e706e67)

* 小结：可以根据用户的价值得分，进行个性化的营销推荐。

## 3.5 商品维度的分析

--销量最高的商品select item\_id ,

sum(case when behavior\_type = 'pv' then 1 else 0 end) as pv, --点击数

sum(case when behavior\_type = 'fav' then 1 else 0 end) as fav, --收藏数

sum(case when behavior\_type = 'cart' then 1 else 0 end) as cart, --加购物车数

sum(case when behavior\_type = 'buy' then 1 else 0 end) as buy --购买数from user\_behaviorgroup by item\_idorder by buy desclimit 10;

--销量最高的商品大类select category\_id ,

sum(case when behavior\_type = 'pv' then 1 else 0 end) as pv, --点击数

sum(case when behavior\_type = 'fav' then 1 else 0 end) as fav, --收藏数

sum(case when behavior\_type = 'cart' then 1 else 0 end) as cart, --加购物车数

sum(case when behavior\_type = 'buy' then 1 else 0 end) as buy --购买数from user\_behaviorgroup by category\_idorder by buy desclimit 10;

* 小结：缺失商品维表，所以没有太多分析价值。假如有商品维表，可以再展开，以商品纬度进行分析，比如不同行业、不同产品的转化率，还有竞品分析等等。

## 3.6 时间维度的分析

--按日统计销量select date(create\_time) as date,

sum(case when behavior\_type = 'buy' then 1 else 0 end) as sales\_countfrom user\_behaviorgroup by dateorder by date asc;

--按小时统计点击量select hour(create\_time) as hour,

sum(case when behavior\_type = 'pv' then 1 else 0 end) as pv\_countfrom user\_behaviorgroup by hourorder by hour asc;

小结：时间维度的分析可以帮助我们了解用户的行为模式，例如哪个时间段的用户活跃度最高，哪个时间段的转化率最高等。同时，对于运营人员来说，也可以根据这些数据来制定更精准的活动策略，比如在高峰期加大促销力度，或者在低峰期进行新品的推广等。

## 3.7 用户行为路径分析

用户行为路径分析是通过观察和分析用户在系统内的行为序列，来揭示用户的行为模式和偏好。对于电商平台来说，用户行为路径的分析至关重要，因为它可以帮助我们了解用户的购物习惯，优化用户体验，提高转化率。

我们可以通过构建用户行为序列的模型，来分析用户的购物路径。比如，我们可以按照用户的访问顺序，将用户的点击、浏览、加入购物车、购买等行为进行串联，形成用户的行为路径。然后，我们可以对这些路径进行统计和分析，找出用户最常见的购物路径，以及不同路径下的转化率等关键指标。

然而，由于本次提供的数据只包含了用户的单次行为记录，并没有包含用户的ID信息，因此我们无法进行用户行为路径的分析。如果我们能够获取到用户的ID信息，那么我们就可以进行更深入的用户行为路径分析了。

## 3.8 结论与建议

综合以上分析，我们可以得出以下结论：

1. 用户画像方面，我们可以根据用户的属性和行为数据，构建出不同维度的用户画像，为后续的精准营销和个性化推荐提供数据支持。

2. 转化率方面，虽然本次数据中没有直接提供转化率相关的指标，但我们可以根据用户的行为数据，如点击、收藏、加购物车、购买等，来间接推算出转化率，并对其进行优化。

3. 商品维度方面，由于缺失商品维表，我们无法进行深入的商品维度分析。如果后续能够补充商品维表，那么我们可以进一步分析不同商品、不同行业的转化率、点击率等指标，为商品推荐和选品提供数据支持。

4. 时间维度方面，我们可以通过分析不同时间段的用户活跃度和行为数据，来揭示用户的行为模式和偏好，为运营策略的制定提供数据支持。

5. 用户行为路径方面，由于本次数据中没有包含用户的ID信息，我们无法进行用户行为路径的分析。如果后续能够获取到用户的ID信息，那么我们可以进行更深入的用户行为路径分析，以揭示用户的购物习惯和偏好。

基于以上结论，我们提出以下建议：

1. 尽快补充商品维表和用户ID信息，以便进行更深入的商品维度分析和用户行为路径分析。

2. 根据用户画像和转化率的分析结果，制定更精准的营销策略和个性化推荐方案，提高用户满意度和转化率。

3. 关注时间维度的分析结果，结合用户的活跃度和行为模式，制定更合适的运营策略，比如在高峰期加大促销力度，或者在低峰期进行新品的推广等。

4. 在进行数据分析时，注重数据的准确性和完整性，避免因为数据问题导致的分析结果偏差。同时，也要不断学习和探索新的数据分析方法和工具，以提高数据分析的效率和准确性。3.9 用户留存与流失分析

用户留存与流失分析是电商数据分析中不可或缺的一部分。通过了解用户的留存情况，我们可以判断平台或产品的用户黏性，而分析用户流失的原因则有助于我们制定有效的用户召回策略。

首先，我们可以计算用户的留存率。以新用户为例，我们可以统计首次访问后，第二天、第七天、第三十天等时间节点的回访用户数量，并计算相应的留存率。留存率的高低可以反映出平台对用户的吸引力以及用户的忠诚度。

其次，对于流失用户的分析，我们可以通过分析流失用户在流失前的行为数据，如访问频率、购买频次、浏览的商品类别等，来识别可能导致用户流失的关键因素。同时，我们还可以结合用户画像，了解流失用户的特征和偏好，为后续的召回策略提供参考。

为了更精准地定位流失用户，我们还可以利用机器学习算法构建用户流失预测模型。通过模型预测，我们可以提前识别出潜在流失用户，并主动采取措施进行干预，降低用户流失率。

## 3.10 竞品分析

竞品分析是电商数据分析中的另一个重要环节。通过对竞品的深入了解，我们可以发现自身的优势和不足，为后续的改进和优化提供依据。

在竞品分析中，我们可以关注以下几个方面：

1. 竞品的产品定位、功能特点以及用户体验等方面的差异和优劣；

2. 竞品的用户规模、活跃度以及留存率等关键指标；

3. 竞品的营销策略和推广渠道，以及其在市场中的表现；

4. 竞品的发展趋势和创新方向，以及对行业的影响和启示。

通过竞品分析，我们可以不断学习和借鉴竞品的优点，同时针对自身的不足进行改进和优化，以提升自身的竞争力。

## 3.11 总结与展望

本次电商数据分析主要围绕用户画像、转化率、商品维度、时间维度、用户行为路径、用户留存与流失以及竞品分析等方面进行了深入探讨。通过对这些方面的分析，我们得以对电商平台或产品的运营状况有了更为全面和深入的了解。

然而，数据分析并非一劳永逸的工作。随着市场环境的变化、用户需求的升级以及技术的不断进步，我们需要不断更新和优化数据分析的方法和手段，以应对新的挑战和机遇。

未来，我们可以在以下几个方面进行进一步的探索和实践：

1. 引入更多的数据源和维度，以丰富数据分析的内容和深度；

2. 利用机器学习和人工智能等技术手段，提升数据分析的效率和准确性；

3. 加强与其他部门的沟通和协作，将数据分析的成果更好地应用于实际业务中；

4. 持续关注行业动态和竞品情况，以便及时调整和优化数据分析的策略和方向。

总之，电商数据分析是一个持续不断的过程，需要我们不断学习和探索，以不断提升数据分析的能力和水平，为电商平台的稳健发展提供有力的数据支持。