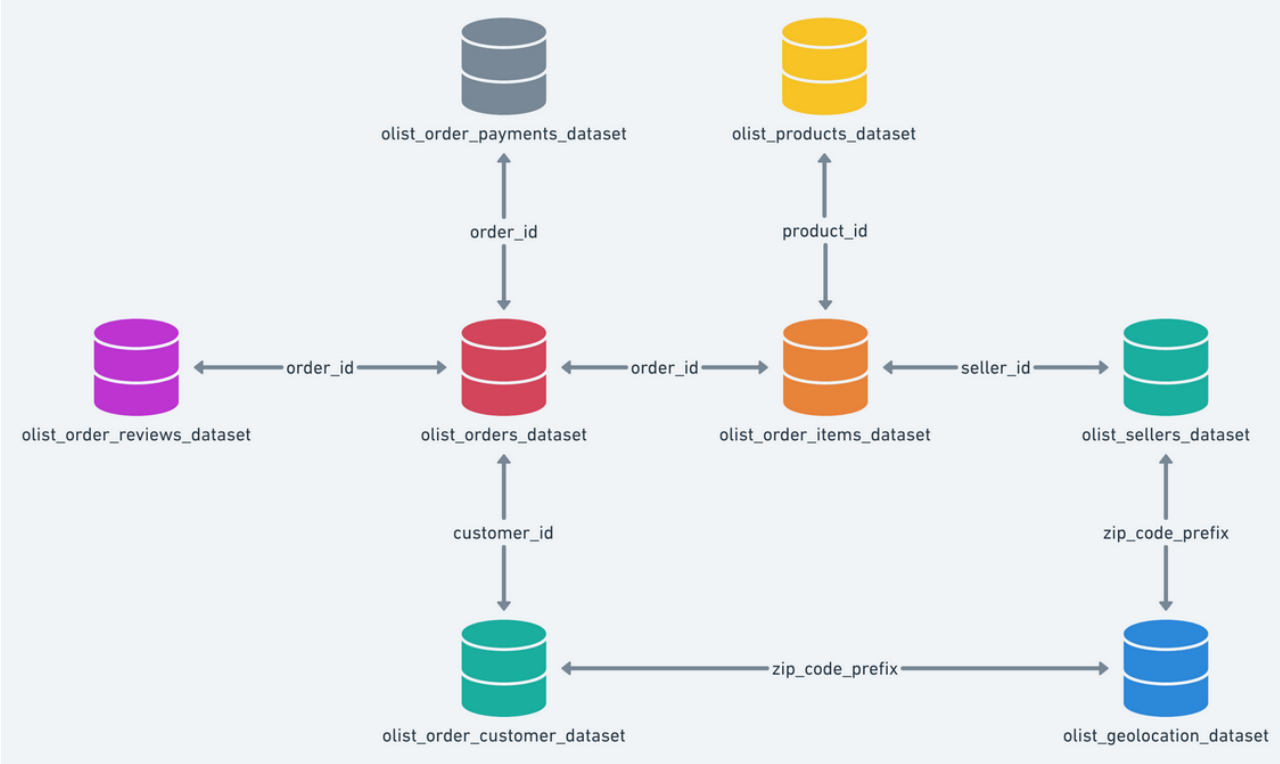
表关系



csv文件【9】

表名称	说明
olist_products_dataset.csv	**产品**：产品id 产品名称长度 类别 描述 照片数量 长宽高 重量
olist_sellers_dataset.csv	**卖家**：卖家id 邮编 城市 国家
olist_geolocation_dataset.csv	**地理位置**：邮编 经纬度 城市 国家
olist_order_items_dataset.csv	**订单商品项**：订单id 商品项id 产品id 卖家id 揽收开始时间限制 单价 运费（需叠加计算）
olist_order_payments_dataset.csv	**支付方式**：订单id 支付顺序（多种支付方式情况） 支付类型 分期数 交易价值（金额）
olist_customers_dataset.csv	**客户**：客户id（连接订单表） 唯一id 邮编 城市 国家
olist_order_reviews_dataset.csv	**评价**：评价id 订单id 评分 评价标题 评价内容 调查发送时间（卖家） 调查回答时间（客户）
olist_orders_dataset.csv	**订单**：订单id 客户id 订单状态 下单时间 审批时间 揽收时间 实际送达时间 预计送达时间
product_category_name_translation	**产品类别翻译**：类别名称 翻译为英文

数据清洗

- olist_products_dataset.csv：610行/32951缺失，只有id其他信息为空。**缺失填充--name**：BLANK 其余数值字段：0。`补充：导入tableau中即不需要填充`
- olist_order_items_dataset.csv：一个订单可能有n件同一个商品，仅order_item_id数值不同（1、2、3、4递增）。订单金额和运费为同一个order_id下的所有项求和
- olist_order_payments_dataset.csv：一个订单可由多种支付方式组成，按支付顺序排，金额为同一order_id的所有金额合计
- olist_order_reviews_dataset.csv：`review_id` 和`order_id`都有**重复值**，是因为 ***评价和订单之间不是严格的一对一关系***。order_id重复说明一个订单可能会收到多次评价，客户评价后平台可能再次发送提醒，修改/再评，于是同一个 `order_id` 会出现多个 review 记录。`review_id` 应该是唯一的，可以认为是 **数据质量问题**。重复删除：每个review_id保留1条记录，选时间最新的

  ```
  =COUNTIF(B:B,B2)，往下填充:当前行在所有行中，共出现的次数。
  =COUNTIF($B$2:B2,B2),往下填充，结果是该行在整个区域中从上到下第几次出现，我们只要筛选把>=2的数值删除，即可删除重复行
  注：数据太多，按公式太卡。筛选出review_id重复项后，按回答时间降序，再使用公式。
  ```

- olist_customers_dataset.csv：同一个 `customer_unique_id` 会对应多个 `customer_id`，原因是customer_id相当于一次交易会话ID，每次下单系统都会重新生成。`customer_unique_id` → 跨订单识别用户，`customer_id` → 识别某个订单里的用户。防止信息泄露：数据集是公开的，Olist 在脱敏时避免直接用一个用户 ID 贯穿所有订单。
- olist_geolocation_dataset.csv：同一个 zip code 前缀会对应多行数据，经纬度差异往往只有几百米到几公里，原因：地理数据来自客户的送货地址，有的人在同一个 zip code 下不同街道。

  ```
  处理方式
  1. 取均值 / 中位数（适合做聚合分析）√
      对同一 zip_code_prefix，只保留一个中心点
  2.聚类（KMeans 或 DBSCAN）（适合更精细化分析）
      如果某个 zip code 覆盖面积较大，可以用聚类把经纬度点分成几个簇，取簇中心作为代表
  3.选择代表点（适合地图可视化）
      可以直接取频次最高的 (lat, lng) 作为该 zip code 的代表，或者取靠近中位数位置的点
  4. 保持所有点（适合研究"地理分布噪声"）
      如果要做精细物流分析的话
  ```

经纬度的数值精度
一般是 小数点后 6 位左右
·实际应用中的选择
城市级 / zip code 级分析 → 3 位小数就够（百米精度）
路线、商圈、门店选址 → 建议保留 5-6 位小数（米级精度）
·推荐做法:
存储时保留6位小数（国际通用标准，Google Maps 也用这个）
如果要做 zip code 聚合，可以先计算均值/中位数，再保留到 3-4 位，避免伪精度。

- orders.csv：根据订单状态，补全数据。为delivered需要补全，其他保持为空。
order_approved_at 14个空值，用order_purchase_timestamp填充；
order_delivered_carrier_date 2个空值，用平均揽收时间计算后填充

=DATEDIF(order_approved_at,order_delivered_carrier_date,"d"),
小数格式，填充=order_approved_at+avg（datedif），设置时间格式

- product_category_name_translation：添加仅product表中，用vlookup匹配。匹配不到的手动补充翻译。

逻辑异常值

数据加工

抽取：字段合并、分列、匹配

计算：数值、函数

分组

转换：转置

数据分析方法

电商订单销售数据，数据表包括订单-商品-评价-支付方式、买家、卖家、产品明细，属于B to C模式。

数据指标体系：**人-货-场**

分析維度	分析指標	分析方法	圖表
產品維度	銷售額	帕累託分析	帕累託圖
	銷量、訂單	對比分析	組合圖
	利潤、客單價	波士頓矩陣	四象限圖
客戶維度	銷售額佔比(性別)	對比分析	餅圖
	客戶數量(性別時間)	對比分析	多系列柱形圖
	客戶數量(性別渠道)	對比分析	多系列柱形圖
	購買數量(客戶年齡)	對比分析	組合圖:面積
	客戶數量(客戶年齡)	對比分析	組合圖:柱形
區域維度	銷售額	對比分析	柱形圖
	銷量	對比分析	詞雲圖
	利潤	對比分析	條形圖
	銷量(城市渠道)	對比分析	多系列柱形圖
時間維度	系列指標	對比分析	分組表

在 **B2C 电商（Business-to-Consumer）** 数据分析里，核心指标可以分为 **销售指标、用户指标、订单履约指标、产品指标** 四大类。以下是比较全面的整理：

1️⃣ 销售指标（Revenue / GMV）

指标	说明	计算方式
GMV (Gross Merchandise Value)	商品总交易额	`sum(order_item_price * order_item_quantity)`
订单总数	一段时间内的订单数量	`count(order_id)`

指标	说明	计算方式
平均订单价值 (AOV)	每笔订单平均金额	`GMV / 订单总数`
支付方式分布	不同支付方式占比	`count_by(payment_type) / total_orders`
退款率 / 退货率	被退回或取消订单占比	`canceled_or_refunded_orders / total_orders`

2️⃣ 用户指标（Customer）

指标	说明	计算方式
活跃用户数 (Active Users)	一段时间内下单的独立用户数	`count(distinct customer_unique_id)`
新用户数	第一次下单用户数	通过 `first_order_date` 计算
复购率 (Repeat Purchase Rate)	多次下单用户占比	`users_with_2+_orders / total_users`
用户生命周期价值 (LTV)	用户在生命周期内产生的总收入	累计订单金额
平均购买频次	用户平均下单次数	`总订单数 / 活跃用户数`

3️⃣ 订单履约指标（Order / Delivery）

指标	说明	计算方式
订单准时率	订单按预估时间送达占比	`on_time_delivered_orders / delivered_orders`
平均发货时间	下单到卖家发货时间	`avg(order_delivered_carrier_date - order_purchase_timestamp)`
平均配送时间	发货到客户收到时间	`avg(order_delivered_customer_date - order_delivered_carrier_date)`
订单取消率	被取消订单占比	`canceled_orders / total_orders`
订单完整率	成功完成交付的订单占比	`delivered_orders / total_orders`

4️⃣ 产品指标（Product）

指标	说明	计算方式
销量前 N 产品	最畅销商品	`sum(order_item_quantity)` 按产品排序
品类贡献	不同类别对 GMV 的占比	`sum(订单金额)` 按 `product_category_name` 分组
退货率 / 评价分布	不同商品退货率及用户评分	`refunded_orders / total_orders`，平均评分
库存周转（如果有库存）	产品销售速度	`库存量 / 日均销量`

5️⃣ 客户体验 / 评价指标

指标	说明
平均评分	用户评价的平均星级
评论数量	产品或卖家收到的评价数
差评率	低于某一星级（如 3 星）的占比

6️⃣ 衍生指标（高级分析）

- **城市/区域 GMV 分布** → `zip code / state` 聚合
- **复购率分布** → 不同时间段的新老用户分析
- **渠道 ROI** → 不同支付方式或营销渠道的 GMV 对比
- **用户留存率** → N 天留存、复购周期分析

🔑 核心思路

1. **销售** → 看 GMV、订单、平均价值
2. **用户** → 看活跃、复购、LTV
3. **履约** → 看发货、配送、取消率
4. **产品** → 看销量、类别、评价

问题

1. 2016-2018，销售情况如何？是否有明显的时间性变化趋势？

2016年订单量少、销售额少。订单量从2017年开始呈现大幅增长，在2018年保持平稳，在2017/11月订单量较上月增长63%，属于特殊时间点增张（11·11/黑五）。销售额在2017年多数保持增长，在2018年保持平稳。平均订单价格较平稳，但较2017年初有小幅下降（10%以内）

2. 顾客消费水平及地域分布情况？复购率？顾客价值分层分布情况?
3. 各类商品的结构及销售额如何？商品描述质量其与销量有什么关系？商品单价及运费水平在什么范围？退货商品有哪些?
4. 订单评分情况？交付性能?

分析内容

订单量月度图:

包含全部时期（月）的月订单数量和月环比，查看订单时间分布柱状图，可知：2016年订单量很少，主要订单分布在2017年和2018年1-3季度，故使用的样本是**下单时间在2017/01-2018/08 的数据**。且2017/01环比出现异常大值（outlier），`故所有的同比环比计算均不包括2016/12及之前的数据`。

1销售分析

GMV：用payment表的value计算GMV

（1）时间维度

- 日 / 周 / **月 GMV 趋势**
- GMV 季节性（节假日、双十一、黑五）
- GMV 环比、**同比增长**

（2）用户维度

- 新用户 vs 老用户 GMV 占比
- 人均 GMV（= GMV / 活跃用户数）
- 复购用户贡献的 GMV 占比

（3）订单维度

- 平均客单价 (AOV = GMV / 订单数)
- 不同订单来源渠道的 GMV

（4）产品维度

- 各品类 GMV 占比
- Top N 热销商品 GMV
- 长尾商品 GMV 占比

（5）卖家维度

- 单个卖家 GMV 分布（集中度分析）
- 平均每卖家 GMV

（6）支付维度

- 支付方式 GMV 占比（信用卡 / boleto / 转账等）
- 分期付款 GMV 占比

（7）地域维度

- 不同州 / 城市 GMV 分布
- 城市等级 / 区域差异

订单量：月订单数，月环比

平均订单价格：每月，环比

支付方式分布

哪种支付方式最常用？（分布）信用卡

信用卡用户更喜欢分多少期？（分期）1期

不同支付方式的客单价差异大吗？存在差异，`支付方式客单价/整体客单价`

哪些支付方式更容易导致订单未完成？（风险）

```
理解顺序（典型生命周期）

created → 用户下单

approved → 支付成功，订单确认
```

```
processing → 商家准备货物

invoices → 出具发票

shipped → 发货

delivered → 完成(成功)

canceled / unavailable → 异常终止（失败）
```

信用卡支付失败排第一

支付方式的偏好随时间/地区/用户群体有变化吗?

2买家分析

分析类型	计算字段	备注
GMV	SUM([payment_value])	含运费、手续费
商品销售额	SUM([price])	只看商品价值
订单总额	SUM([price] + [freight_value])	接近支付金额
RFM / LTV	`{FIXED [customer_unique_id] : SUM([payment_value])}`	衡量客户贡献
毛利分析	SUM([price] - 成本 - freight_value)	需结合成本数据

✅ **结论总结：**

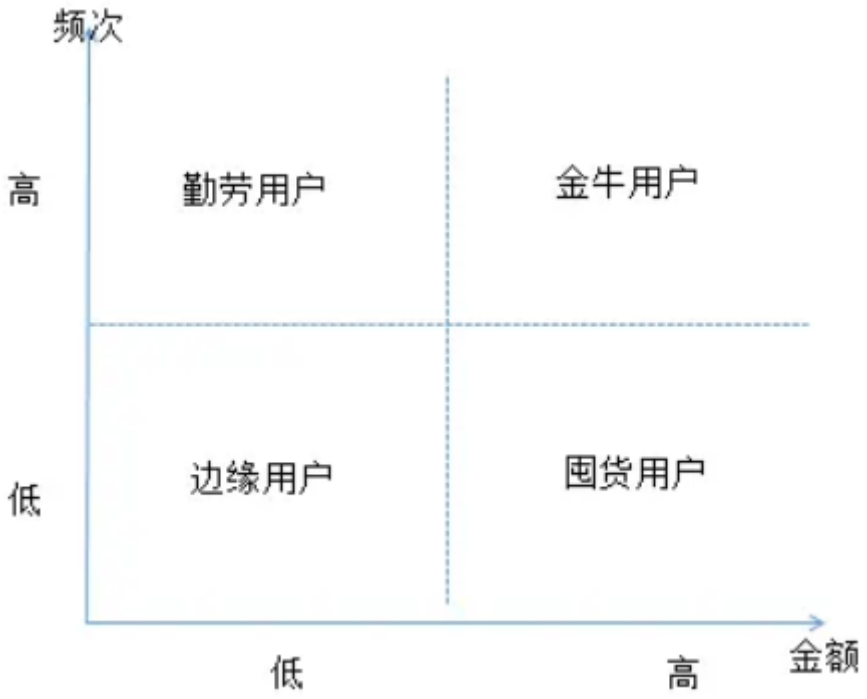
> - “如果你关心“客户花了多少钱” → 用 `payment_value`。”
> - “如果你关心“商品卖了多少钱” → 用 `price`。”
> - “如果你关心“商家赚了多少钱” → 还要引入成本、运费。”

RFM模型(用payment计算)

· Recency最近一次消费时间，最近一次（max_time）到当前时间的间隔，如7天、30天、90天未到店消费（用户唤醒机制）

· Frequency一定时间段内的消费频率(order_count)，如一年内几个月有消费、一个月内几天到店。频率越高用户忠诚度越高（用户激励机制）

· Monetary一定时间段内的累计消费金额（只用商品金额），如一年内多少消费金额，买的越多价值越大（VIP机制）

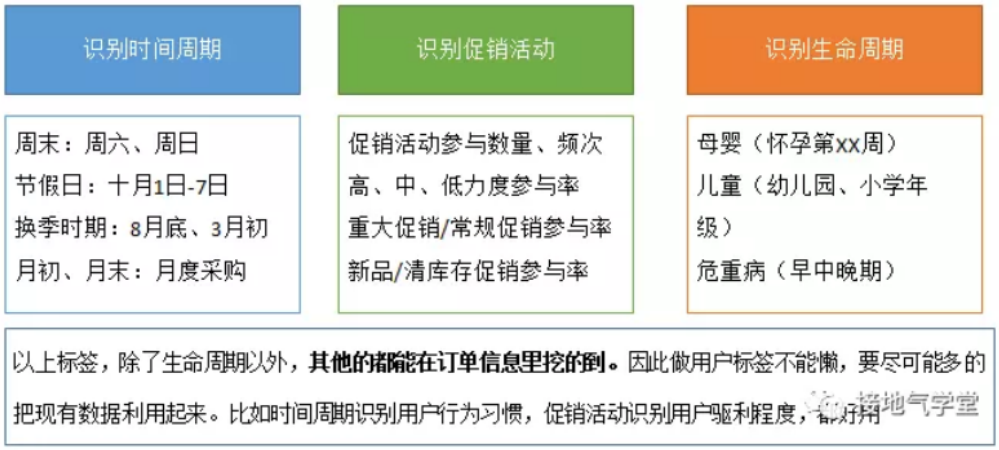


Q：取数的时间怎么分？
消费频次本身越高的业务，取的时间应该越短。更多的做法是按月取，比如R按月取，F、M算最近一年内的数值。

RFM本身并没有错，在数据匮乏（特别是缺少埋点数据）的情况下，用RFM比不用RFM好太多了。

综合RFM失效的场景，可以看出：季节性、商品特征、促销活动、节假日事件、用户生命周期，这五大要素，都会影响到用户的行为

促销活动也是同理，促销活动可以直接从订单识别出来，因此也很容易给用户贴上——促销敏感型的标签。



用户生命周期，需要数据采集，而且是采集一个最关键的数据即可。

将客户分为“高价值”“忠诚用户”“沉睡用户”“潜力用户”等。
方法：打分（1–5），再聚类或分层。

```
Q：数据选取事件范围为2017/01 – 2018/08，Recency 基准时间应该选哪里？
有3种选择
·用数据集的最大日期（推荐），适合静态分析（一次性算 RFM）
    如果筛选到 2018/08，就取
    基准日期=2018-08-31+1
·√用一个固定观察点，适合做动态趋势或队列分析（Cohort）
    如每月 RFM 分析，基准就是当月月末
    2017/01 分析 → 基准 = 2017-01-31
·用今天（NOW()）作为基准
    在实际业务系统里，通常取 当前日期
```

活跃用户数：R-F-countd(customer_uid)热力图，

近30天有购买的客户数6153，占所有客户数比例为6%

近30-90天有购买的客户数12250，占所有客户数比例为12.8

90-180天19950，占20%

一年内有购买的占76%

购买频率：1年内1次最多

复购率：频次Frequency，购买仅1次占比96.9%，2次占比2.8%，3次及以上占比0.3%

新用户数：从2017-01至2017-11每月新用户数呈现递增状态，2018年1-8月每月新增用户数保持平稳约6.5k

生命周期价值LifeTimeValue：客户在其整个生命周期内（从第一次下单到流失之前）为企业带来的平均收益。即一个客户从第一次下单到不再回来的这段时间，总共贡献了多少钱?

LVT分布在0-14k之间，总体分布区间[60%,80%]为[171, 228]

地区分布（购买力）：客户分群

基于统计：按地区、性别、年龄、支付方式等切分。

基于行为：购买品类、购买周期、支付偏好。

转化漏斗（无）

路径分析：浏览 → 下单 → 支付 → 收货。

漏斗分析：计算每一步转化率，找出流失点

3订单履约分析

准时率：准时送达/已送达，92%。on_time_delivered_orders / delivered_orders

平均发货时间：3.5天。avg(order_delivered_carrier_date - order_purchase_timestamp)

平均配送时间：9.2天。avg(order_delivered_customer_date - order_delivered_carrier_date)

取消率：1%，被取消订单占比。canceled_orders / total_orders

交付完整率：93%，成功完成交付的订单占比。delivered_orders / total_orders

4产品分析

销量前N:TopN

品类贡献:health_beauty

退货率:0.2%

评价分布:4.1

5客户体验

平均评分：4.1

评论数：40414

评论回答时间（间隔）：3.3天

差评率：15%

6卖家分析

谁在卖？（概况）卖家数量、**地区分布**、增长趋势

卖的怎么样？（销售表现）销量、GMV、客单价、转化率

卖的好不好？（服务质量）发货效率、退货率、评价、留存

7衍生指标分析

城市GMV分布

地图:

客户、卖家：订单密度、GMV、平均客单价、配送时长

复购率分布

支付方式渠道ROI

用户留存率

结论与建议

1.数据源Orders

【订单量月度图】，**订单量条形图与环比折线图**双轴。看出订单数主要分布在2017/1-2018/8月。

【月度GMV_订单量表】，月份为行，`月度GMV、GMV环比、GMV同比`、`月度订单量、订单量月环比、订单量同比`、`月平均订单价格、订单价格环比、订单价格同比`。2017年11月gmv最高，且之后的月度GMV都超过平均值797.2k，GMV呈现递增-平稳状态。订单量趋势相同，在2017年11月达到顶峰。平均订单价格呈现略微下降趋势，平均价格为162，2018年1月同比下降11%。

【订单量与GMV关系图】，列为订单量，行为月度GMV，由**散点图与趋势线**看出，呈正相关`月度GMV = 157.901*月订单量 +14890.2`

【支付方式分布表】，行为支付方式，列为`订单数量、订单数占比、支付金额、金额占比、客单价`。其中信用卡订单数量占比最多，其次是boleto。信用卡支付金额也最多，占所有的78.4%。

【支付方式分期期数分布图】，期数0-24期，信用卡1期的订单数最多，其次是boleto1期，90%一张的订单泛起数在10期及以内。

【支付方式_未完成表】，信用卡取消和不可用的订单数最多，合计858单；其次是boleto 242单。但占所有订单数的比例很小，可以忽略不计。

【订单履约分析表】，行为时间，列为`总订单数、已送达订单数、准时送达订单数、取消订单数`，以及`订单完整率、准时率、取消率`，`平均发货时间、平均配送时间`。订单完整率呈递增状态，由93%上升到98%；准时率在2017年10月及以前保持在95%以上，11月开始随订单数量的增多趋于不稳定，2018年3月最低为79%，2018年8月最高99%；取消率维持在平均线0.6%左右，但2017年2月、3月与2018年2月取消率在1%以上，可能与季节性有关。平均发货时间维持在平均值3.2天左右，2017年11月最长为4.1天，与购物节订单量剧增有关。配送时间均值为9.3天，2017年4月、2017年11月-2018年3月都达到10天以上。

【订单取消率与发货时间、配送时间的关系】发货时间与配送时间趋势同步，最长都出现在2018年2月。

在tableau用双坐标轴折线图分析相关性，通过这按图的趋势看，取消率和配送时间的相关性更高，取消率存在滞后性，需要向右移动一个单元。

【客户体验表】，行为时间，列为`评价订单数、评论数、差评数、平均评分、差评率、平均评分、平均评价时间`，平均评分4.1分，差评率均值14%，2018年3月最高23%。

【差评与发货/配送时间关系】，行为时间，列为差评率与货/配送时间。双轴折线图看出，差评与配送时间相关性更强，配送时间长差评率高。改善差评率可以从缩短配送时间下手。

2.数据源Products

【产品】，行为产品ID和产品类别，`列为产品销量、销售额、单价、订单数、退货订单数、客户评分、评分人数、产品下单评分率。`根据产品销量排序，选择TopN。Top100中，最高销量：527件家居装饰类，最高销售额：63.9k 健康美容类，最高单价：350.8 健康美容类，最多订单数：467 床桌家具类，退货订单0-3单，退货率均值为0.2%，客户评分分布在3.8-4.6之间、均值4.1，评分率均值为99%。

【品类贡献】，行为产品类目西语/英文，列为`销量、销售额、类目下产品数、GMV贡献率`。共72个类目，销量、销售额、贡献率最高的类目是健康美容类，销量9670件、销售额1258.7k、贡献率9.3%；贡献率前5是健康美容、手表、桌浴室家具、运动器具、电脑配件，占总体的39.9%。

【客户RFM】，`行为R_score、F_score，列为M_score，详细信息为Customer_uid计数`。根据Recency、Frequency、Monetary按5分制打分分类，按平均值区分得分高低。根据颜色深度，显示绝大部分用户是存在1分的项。高价值客户：173，重点保持：71，重点发展：34840，重点挽留：22371，一般价值：0，一般保持：3，一般发展：22527，潜在客户：15788，合计95773。

R_SCORE	F_SCORE	M_SCORE	客户类型
高	高	高	高价值客户
低	高	高	重点保持
高	低	高	重点发展
低	低	高	重点挽留
高	高	低	一般价值
低	高	低	一般保持
高	低	低	一般发展
低	低	低	潜在客户

金额高的都是重点客户。

【RFM热力图】，行为R_score，列为F_score，详细信息为M，颜色深浅表示消费金额均值高低。随着R_score和F_score的分值增大，颜色越来越深，说明客户最近一次交易时间越近、交易次数越多，其平均交易金额越高。

【RFM分布图】【RFM散点图】

【每月新用户数】，呈递增状态，2017年11月新用户数最多达7.3k。说明平台仍在吸引新客，有扩大消费者的潜力。

【LTV首单月份】：在该月首次购买客户的平均生命周期价值，最高186，最低154，呈现波动状态。

【LTV客户分布】，主要分布在57-228之间

【每月新增卖家数】，最低76，最高228，均值150。平台对卖家入驻仍有吸引力，但需要增强在商家中的推广

【卖家活跃状态】，90天内活跃占59%，不活跃占41%，需要为不活跃卖家推流。

【卖家销售分析】，按销售额TopN筛选前100，最高销售额253k，最多订单数1.8k，平均客单价309，最多上架产品数399，平均每单商品数1。

【卖家服务质量】，发货时长平均7天，配送时长平均9.8天，服务评分平均4.1，差频率平均13%

【卖家GMV与评分气泡图】，评分集中在3-5分，，销售额集中在50k以内。评分平均值4，GMV平均值5k，存在GMV高评分低的商家。

3.数据源Geolocatiom

【地图_各州订单数】，颜色深度表示订单数多少，最高41606，密集分布在沿海地区。

【地图_城市GMV分布】，颜色深度表示，最高是belo horizonte为419k

【客户数-销售额双层地图】，SP州颜色最深，客户数40k，销售额5976k

【各州月订单量动态变化】，page是月份，从2017年1月至2018年8月

【卖家/客户地区分布】，散点，卖家分布在沿海集中，客户分布较广，多数在靠近沿海的南方地区。

注释

环比与数据范围

有效数据的日期筛选范围是2017/01-2018/08，数据文件中的范围是2016/09-2018/10.

在计算环比时，如2017/01上个月的数据不在显示范围内，如何计算?

解决思路：要做环比，**必须保证上月数据也在计算范围内**。

方法一：使用 **表计算（LOOKUP）**

时间维度放到行/列（比如 `订单日期` → 月）

```
[环比增长] = (SUM([GMV]) - LOOKUP(SUM([GMV]), -1)) /
LOOKUP(SUM([GMV]), -1)
LOOKUP取上一个月的数据，即使你没在筛选器里选它，只要它在表格里就能计算。
```

这里 `SUM([GMV])` 必须是聚合字段，否则会报错

方法二：双重筛选，「计算字段 + 过滤器」

步骤 1：确保数据可以环比计算

```
([月度GMV] - LOOKUP([月度GMV], -1)) / LOOKUP([月度GMV], -1)
```

注意：这里 `[月度GMV]` 必须是聚合字段，否则会报错。

√步骤 2：创建「用于显示的日期」字段

创建一个计算字段，用于在图表上只显示 2017/01 之后的数据：

1. 在左侧数据栏 → 右键 → **创建计算字段**
2. 命名为 `显示日期`（名字可自定义）

```
IF [日期] >= DATE("2017-01-01") THEN [日期] END
这个字段会在 2017/01 之后保留日期，2016/12 会返回 Null。
```

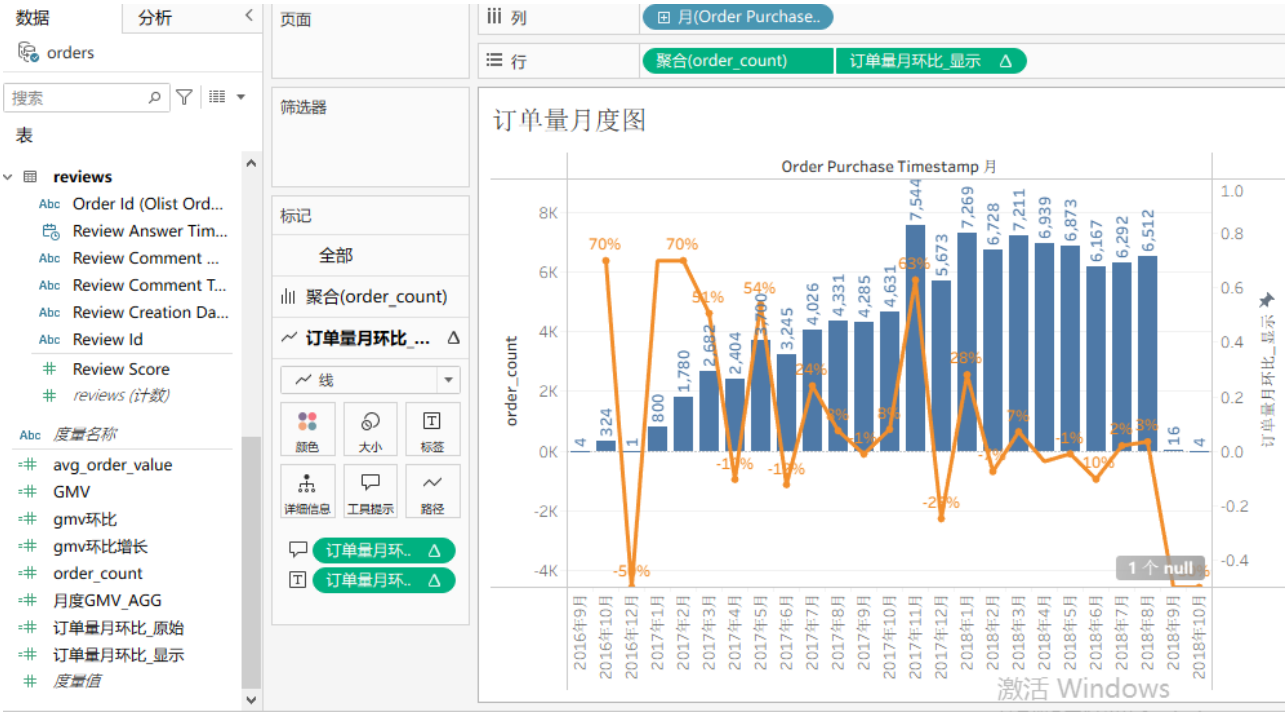
步骤 3：将「显示日期」放到列或行,Tableau 会自动忽略 Null

方法三：用 **LOD 表达式**

```
[上月GMV] = { FIXED DATETRUNC('month', DATEADD('month', -1, [订单日期])) : SUM([GMV]) }
[环比增长] = (SUM([GMV]) - [上月GMV]) / [上月GMV]
这样即使上月被筛掉了，LOD 也会强制取到
```

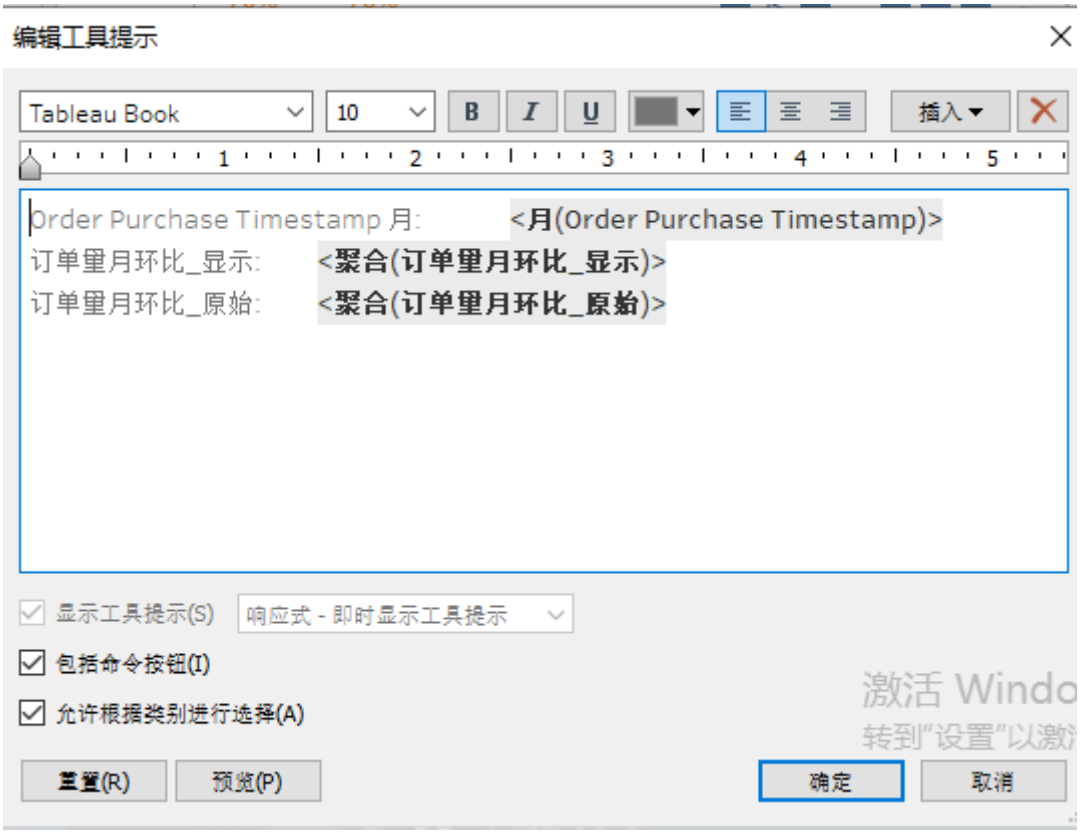
折线图异常大/小值

在计算同比、环比时，会出现异常大值，导致其他正常值显示在底部，折线可视化差，需要调整。采用数据截断，借助标记-工具提示(Tooltip)显示真实值。




```
折线图异常值处理.png月度订单量=DATETRUNC('month', [Order Purchase Timestamp]):COUNT([Order Id]) }
订单量月环比_原始=ZN(SUM([月度订单量]))/LOOKUP(sum([月度订单量]), -1) - 1
订单量月环比_显示={
IF [订单量月环比_原始] > 0.7 then 0.7
ELSEIF [订单量月环比_原始] < -0.5 then -0.5
ELSE [订单量月环比_原始]
END}
```

工具提示-插入：设置显示格式



csv文件中的逗号问题

CSV文件以逗号作为分隔符，正常情况下，**CSV 遇到双引号括起来的字段，里面的逗号应该被当作普通字符，不作为分隔符**。在reviews表的评价内容里，有包含逗号的信息，如："Amei a mesinha, que pena que me confundir na cor, comprei achando que seria branca, mas é nude"。在tableau中即使用了引号包围，仍按照逗号分割成多个列，导致原数据中后面的字段内容错误，需要处理逗号。

·方法1:在 Tableau 中用 Data Interpreter（数据解释器）
在 Tableau 导入数据时，勾选 Use Data Interpreter。Tableau 会尝试重新解析引号和分隔符
√·方法 2：用 Python 清洗（最稳妥）
如果文件太大，建议用 Python 的 pandas 重新导出

```
import pandas as pd

df = pd.read_csv("olist_order_reviews_dataset.csv", sep=",",
quotechar='"', encoding="utf-8")
df.to_csv("olist_order_reviews_clean.csv", index=False,
encoding="utf-8")
```

数据埋点 Data buried point

也称事件追踪，在产品或应用中添加代码，以捕获用户行为的细节/数据，将信息发送到数据服务器。

数据构成：事件（如：按钮点击 搜索 滑动）+属性（如：用户性别、来自渠道）

作用：了解用户的行为轨迹、偏好和需求，优化产品、提高用户留存率、转化率和满意度；评估产品效果和用户需求反馈

技术手段：

· 代码埋点（也称自定义埋点，适用于网站或应用开发过程中，需要开发人员配合）

· 可视化埋点（通过可视化工具，在页面上选择需要埋点的元素，即可自动生成代码。如Google Analytics、Mixpanel、Amplitude）

· 无/全埋点（通过前端技术，自动收集用户的行为数据，无需手动添加埋点代码。适用于简单的数据采集需求，在实际业务上用的不多）

埋点机制：（与捕鱼很像）

· 事件检测（上报时机）

· 参数采集（采集策略分成静态和动态）

静态的采集策略是全部采集还是部分采集。按需采集

动态的采集策略是在交互动作前后，参数的采集策略。动态采集策略的不同常常在关键漏斗转化上数据对不上，上游点击并不等于下游曝光

· 上报传输：指将采集到的公参和私参传输入库再清洗

埋点机制	自定义埋点	可视化埋点	〔全〕无埋点
事件检测	检测时机灵活可调	检测时机相对固定	检测时机相对固定
参数采集	采集部分参数〔开发代码操作〕	采集部分参数〔可视化界面操作〕	采集全部参数〔无需操作〕
上报传输	技术选型可以一致	技术选型可以一致	技术选型可以一致

埋点方案

· 理解产品：从业务目标入手，拆解北极星指标

· 翻译产品：数据化过程，流程图 框架设计

· 表达产品：即设计埋点方案，如事件驱动型、层次结构驱动型。最主要的目标是设计有效的统计参数，让埋点开发在执行过程中有重心

库存指标

一、库存周转率（Inventory Turnover Ratio）

📘 **定义：**衡量库存被售出的速度，反映库存资金的利用效率。

📊 **计算公式：**库存周转率=销售成本（或销售额）/平均库存成本（或金额）

📈 **举例：**

- 一季度销售成本为 300 万，平均库存为 100 万，
  → 库存周转率 = 3 次。
  意思是一季度内库存周转了 3 次，平均每月转一次。

📌 **业务解读：**

- 周转率越高 → 库存利用率越好，积压少；
- 周转率过低 → 可能存在滞销、库存积压、占用资金问题；
- 但太高也可能缺货或供应不稳定。

💡 **补充指标：**可反推**库存周转天数**

库存周转天数=365/库存周转率，表示平均库存卖完所需天数。

二、滞销率（Slow-Moving Inventory Rate）

📘 **定义：**反映滞销库存占总库存的比例，用于识别库存结构是否健康。

*📊 *计算公式：*滞销率=滞销商品库存金额/总库存金额×100%

📈 **举例：**

- 仓库中 50 万库存商品中，有 10 万超过 60 天未出库；
  → 滞销率 = 10 ÷ 50 = **20%**

📌 **业务解读：**

- 滞销率高说明选品或采购计划不合理；
- 电商常按SKU维度看滞销SKU数量、滞销金额、滞销周期；
- 通常结合动销率（动销SKU / 总SKU）一起分析。

三、采购准确率（Procurement Accuracy）

📘 **定义：**衡量采购计划的执行精度与实际需求的匹配程度。

📊 **常见计算方式（两种视角）：**

1️⃣ **数量维度：**

采购准确率=1−∣实际采购量−需求预测量∣/需求预测量

或

采购准确率=按时且数量准确到货的采购订单数/总采购订单数×100%

2️⃣ **时间维度：**

按时到货率=按时交货的采购批次/总采购批次×100%📈 **举例：**

- 预测要进 1000 件，实际到货 950 件 → 准确率 = 95%；
- 若总 50 笔订单中 45 笔按时到 → 按时到货率 = 90%。

📌 **业务解读：**

- 准确率低：说明预测不准、供应商不稳定或采购计划不合理；
- 电商常结合销售预测模型调整采购策略。

三轴图画法

度量名称：度量名称放在筛选器，筛选2个。度量值放在行与标记；第3个度量值单独放入行，选择双轴。

分析相关性的五种常用方法

1.图表相关分析（折线图及散点图）

双坐标轴折线图，数据的变化和趋势大致相同。散点图更直观，去除了时间维度的影响。

优点：对相关关系的展现清晰。

缺点：无法对相关关系进行准确的度量，缺乏说服力；当数据超过两组时，无法完成各组数据间的相关分析。

2.协方差及协方差矩阵

协方差用来衡量两个变量的总体误差。如果两个变量的变化趋势一致，协方差就是正值，说明两个变量正相关。如果两个变量的变化趋势相反，协方差就是负值，说明两个变量负相关。如果两个变量相互独立，那么协方差就是0

$$cov(X, Y) = \sum(Xi - X')(Yi - Y')/(n - 1)$$

协方差只能对两组数据进行相关性分析，当有两组以上数据时就需要使用协方差矩阵

3.相关系数

相关系数(Correlation coefficient)是反应变量之间关系密切程度的统计指标，相关系数的取值区间在1到-1之间。1表示两个变量完全线性相关，-1表示两个变量完全负相关，0表示两个变量不相关。数据越趋近于0表示相关关系越弱。

优点：通过数字对变量的关系进行度量，并且带有方向性

缺点：无法利用这种关系对数据进行预测，简单的说就是没有对变量间的关系进行提炼和固化，形成模型

4.一元回归及多元回归

回归分析（regression analysis)是确定两组或两组以上变量间关系的统计方法。按照变量的数量分为一元回归和多元回归。两个变量使用一元回归，两个以上变量使用多元回归。进行回归分析之前有两个准备工作，第一确定变量的数量。第二确定自变量和因变量。

$$Y = b0 + b1X1 + b2X2 + ... + bnXn$$

5.信息熵及互信息

度量这些文本特征值之间相关关系的方法就是互信息。通过这种方法我们可以发现哪一类特征与最终的结果关系密切。《决策树分类和预测算法的原理及实现》

小结

图表方法最为直观，相关系数方法可以看到变量间两两的相关性，回归方程可以对相关关系进行提炼，并生成模型用于预测，互信息可以对文本类特征间的相关关系进行度量。

数据异动归因分析

北极星指标：也叫唯一关键指标(one metric that matters)，产品现阶段最关键的指标。 梁宁《增长思维30讲》《精益数据分析》《硅谷增长黑客实战笔记》。目标制定最好是能符合SMART原则。目标制定最好是能符合SMART原则。目标制定最好是能符合SMART原则：具体(Specific)、可以衡量的(Measurable)、可实现(Attainable)、相关性(Relevant）、截止期限(Time-bound)。

做业务分析时最常用到的方法：指标拆解

当一个指标波动时，我们首先需要从业务视角判断其波动是否异常，即异动检测，其次判断异常背后的原因是什么，即异动归因。

1.发现指标异常

量化度量：**均值+标准差**。选取近30日指标数据，计算指标均值和标准差，选取正常范围（+-一倍标准差），当波动大于阈值时，可重点关注。

2.定位异动原因

针对指标进行**下钻**，通过相对熵挖掘异常维度，结合指标贡献度定位到可能的问题维度。相对通用且高效的排查思路**宏观→微观→宏观**，涵盖六步：

步骤1 宏观-多维钻取：为了达到全面、高效的目的，需要做两点--1维度累积（日常工作中我们对于业务的理解） 2工具沉淀（*多维分析工具*），产出内容包括但不限于：维度波动分数（例如：JSD分数）、当天pv、同期pv、pv变化绝对diff、pv变化相对diff等。

步骤2 宏观-异常聚焦：完成多维钻取，产出各维度涨降幅数据后，需要将细分维度对大盘的影响进行量化度量，即我们常说的：*贡献度度量，核心在于量化*。

经过这2步可以处理80%左右的实际波动问题，剩余20%需要考虑下钻到微观层面探索问题本质。

步骤3 微观-Case分析：主要是通过用户详细行为数据或者轻度汇总数据进行挖掘，发现可能的异常问题。由于是单用户行为，会存在诸多干扰的情况，很容易被带偏。核心在于*发现问题，快速试错*。

步骤4 微观-假设及场景模拟：到此步，还没能找出核心的问题点，则说明此问题相对隐蔽。需要我们根据业务场景，假设问题出现的原因，并尝试找到可度量的维度进行验证。

到这一步，基本可以解决95%左右的实际波动问题。

步骤5 宏观-维度上升：进行完步骤三和步骤四之后，我们已经对可能的问题有了一定的累积，可以将此维度上升到宏观层面，利用多维钻取工具进行挖掘，判断是否此问题导致的指标大幅波动。

步骤6 宏观-输出结论：对问题下结论，结论输出需要涵盖“*数据发现问题*”和“*业务解释问题*“，在与产运方沟通后，得出“*业务化*”*的结论*。另外一个需要注意的，则是*大盘指标变动往往受到多方因素的影响，无法100%精确解释，或多或少会存在预估的成分，这个属于正常情况*。

A/B Test(实验)

概念

源于生物医学的「双盲测试」，测试中病人被随机分为AB两组，在不知情的情况下分别注射安慰剂和测试药剂，通过用户的表现来评估药物是否有作用。

日常工作如：针对某APP的产品改动，将用户群体随机划分AB两组，A赋予实验策略，B保持原有策略，通过用户的行为表现数据，评判实验对产品改进是否有增益。如有增益，则用新策略A替代原有策略B。

为什么要进行AB实验

AB实验可以最大程度避免「***时间差异***」「***用户差异***」等混淆因子所带来的影响，纯粹度量策略对用户的影响程度，是度量策略「因果效应」最直接的方式。

最佳AB实验流程（七个步骤）

步骤1 实验设计阶段

·**建立假设**【产品运营】，涵盖明确原则（满足AB实验的原则）、明确背景（设计实验的前提）、探索改进点（根据历史数据及产品经验判断）、筛选实验对象（全量用户or部分标签用户，实验单元是用户or会话or页面）、预估实验影响（可能对哪些核心指标有影响、程度多大、与业务账期目标是否一致）

·**明确实验类型**【数分】，根据流量关系划分为「正交实验」以及「互斥实验」。正交实验，实验之间互不干扰，可以复用相同的流量，只要保证各实验用户是随机分布即可，即「常规AB实验」。互斥实验，实验之间会相互干扰，X实验和Y实验之间会出现冲突的情况。「层域AB实验」、「Interleaving实验」、「包含网络效应实验」以及「Multi-Armed Bandit，MAB实验」等。

·**筛选评估指标**【产品运营/数分】，考虑2个方面，1根据实验类型筛选（如页面停留时长、内容点击率），2根据度量方向筛选（北极星指标、期望提升指标、预警指标、观察指标）

·**设定实验周期及流量**【数分】，实验周期和流量的设定是个权衡的问题，一方面，我们希望有足够多的流量，保证实验策略可以充分体现出来；另一方面，也希望缩短实验周期、提升迭代效率、降低实验风险。在评估流量的时候，我们需要计算，在满足评估的前提下，获取最小的样本量。根据统计功效、显著性水平、相对差异、最小可检测效果，利用公式反推实验的最小样本量。*注意：周期效应、新奇效应(等实验指标趋于平稳之后再进行评估)*。

步骤2 实验研发阶段【研发】

主要是对「策略的生效」以及对「生效人群的打标」。

步骤3 实验运行阶段

·**实验数据质量检测：AA实验**【数分】，AA空跑实验是AB实验的一种特殊形态，同AB实验一样，将线上用户随机进行分组，将各分组用户采用相同的无策略空跑实验。根据空跑实验指标显著性情况，探查是否存在SDK(软件开发工具包)接入、指标接入等问题。判断无问题后，才会挑出其中一个分组附上实验策略，进入常规AB实验。

·**验证样本均衡性**【数分】，开始上线实验前，还需要验证AA桶的流量是否均衡，是否会出现SRM（sample ratio mismatch）的问题。如果出现实际样本比例与理论样本比例不同的情况，就要排查其中的原因。可通过「*卡方检验*」的方式，度量两组流量的一致性。

出现SRM问题排查：实验配置阶段、实验上报阶段

·**验证策略是否生效**【产品运营】，在上线实验前，选部分测试用户，将其直接进入实验策略，验证用户端是否生效。在上线阶段，这部分用户不会记录到真实的实验分桶中，防止对上线实验的干扰。

步骤4 实验评估阶段【数分/产品运营】

·**评估指标敏感度及显著性**，敏感度：实验累计样本是否满足最小样本量要求、当达到实验预期的周期后，评估各观测指标的MDE（minimun detectable effect，最小检测变化）是否小于实验前设定的最小变化阈值。显著性：通过假设检验方式进行判断。

·**拆解维度评估**，当遇到实验效果「超预期」或者「不符合预期」的时候，需要对维度进行下钻评估。必要时，还可以挖掘一些badcase/goodcase，从而探索实验可能的问题点。

·**评估对大盘的影响程度**，实验是否可以上线，主要还是要看对大盘的影响程度，这里指标可以划分为「绝对指标」和「相对指标」。相对指标可直接通过指标的DIFF判断对大盘的影响；而绝对指标需要先将指标折合到大盘比例，再进行计算。

步骤5 实验报告输出阶段【数分】

在申请放量之前，一方面用于实验的总结和实验放量的依据，另一方面用于经验的沉淀及后续流程的复用。报告主要涵盖：实验背景、实验设计、实验数据、实验结论。

步骤6 实验放量阶段【数分/产品运营】

综合考虑三个因素：效率、质量、风险。

·**第一阶段 小流量**，整体放量比例控制在5%以下，评估实验是否对产品北极星指标有负向影响。同时验证策略的触发，以及排查是否存在潜在风险。在无风险的前提下，建议持续3-5日左右，进入下一个阶段。

·**第二阶段 放量**，随着样本量的逐渐放开，实验的结果也会更加精准。伴随而来可能会出现流量压力等问题的发生，因此在此阶段需要跟进放量，观察是否有出现问题。逐级放量建议持续至少一周，以观测周中和周末的影响。

·**第三阶段 长期存放**，针对部分实验，如果希望长期观测实验效果，可以保留5%以下的原始策略，作为「*反转桶*」。

步骤7 实验报告归档阶段【数分】

将实验的经验沉淀到平台或者文档中，为未来规范化实验评估提供理论依据。包括内容：

·**实验分类**，按照类型分，尽量保障同类型实验的评估方式规范化。例如：页面改版实验、功能更改实验、福利策略实验

·**实验评估经验**，评估流量（最优流量大小）、评估指标（哪些指标能更清晰地度量）、波动阈值（范围 微显著、显著、非常显著）

·实验成功/失败的经验

Excel查找和引用函数

1、Lookup：当您需要查询一行或一列并查找另一行或列中的相同位置的值时。向量形式:

```
LOOKUP(lookup_value, lookup_vector, [result_vector])
```

Lookup_value 在第一个向量中搜索的值，可以是数字、文本、逻辑值、名称或对值的引用

lookup_vector 只包含一行或一列的区域，可以是文本、数字或逻辑值

*result_vector*只包含一行或一列的区域，参数必须与 *lookup_vector* 参数大小相同

注：如果 **LOOKUP** 函数找不到 ***lookup_value***，则该函数会与 ***lookup_vector*** 中小于或等于 ***lookup_value*** 的最大值进行匹配。如果 ***lookup_value*** 小于 ***lookup_vector*** 中的最小值，则 **LOOKUP** 会返回 #N/A 错误值。

2、VLookup：在表格或区域中按行查找内容时，如根据员工 ID 查找员工姓名。

```
=VLOOKUP(查找值，包含查找值的范围，包含要返回的值的范围内的列号[，返回表示为 1/TRUE 或 0/FALSE 的近似或精确匹配项)
```

3、HLookup：在表格的首行或数值数组中搜索值，然后返回表格或数组中指定行的所在列中的值。 H 代表“行”。

```
=HLOOKUP(lookup_value, table_array, row_index_num, [range_lookup])
```

lookup_value 要在表格的第一行中**查找的值**，数值、引用或文本字符串。

table_array 查找数据的**信息表**，第一行的数值可以为文本、数字或逻辑值。**如果 range_lookup 为 TRUE，则 table_array 的第一行的数值必须从左到右按升序排列**。否则，HLOOKUP 将不能给出正确的数值。 如果 range_lookup 为 FALSE，则 table_array 不必进行排序。

row_index_num 返回匹配值的table_array中的**行号**。如果row_index_num小于 1，HLOOKUP 将返回 #VALUE！ error 值;如果row_index_num大于table_array上的行数，HLOOKUP 将返回 #REF！ 错误值。

[range_lookup] **匹配模式，TRUE 或省略，则返回近似匹配值**。 如果为 **False**， 查找精确匹配值，找不到则返回错误值 #N/A。

例：在首行查找车轴，并返回同列（列 A）中第 2 行的值。

=HLOOKUP("车轴", A1:C4, 2, TRUE)

4、XLookup：按行查找表或区域中的项，如根据员工 ID 查找员工姓名。XLOOKUP 函数搜索区域或数组，然后返回与它找到的第一个匹配项对应的项。 如果不存在匹配项，则 XLOOKUP可以返回最接近的 (近似) 匹配项。 XLookup是VLookup和XLookup的改进版本。

```
=XLOOKUP(lookup_value, lookup_array, return_array, [if_not_found],
[match_mode], [search_mode])
```

*lookup_value*要搜索的值，如果省略，XLOOKUP将返回在 lookup_array 中找到的空白单元格。

*lookup_array*要搜索的数组或区域

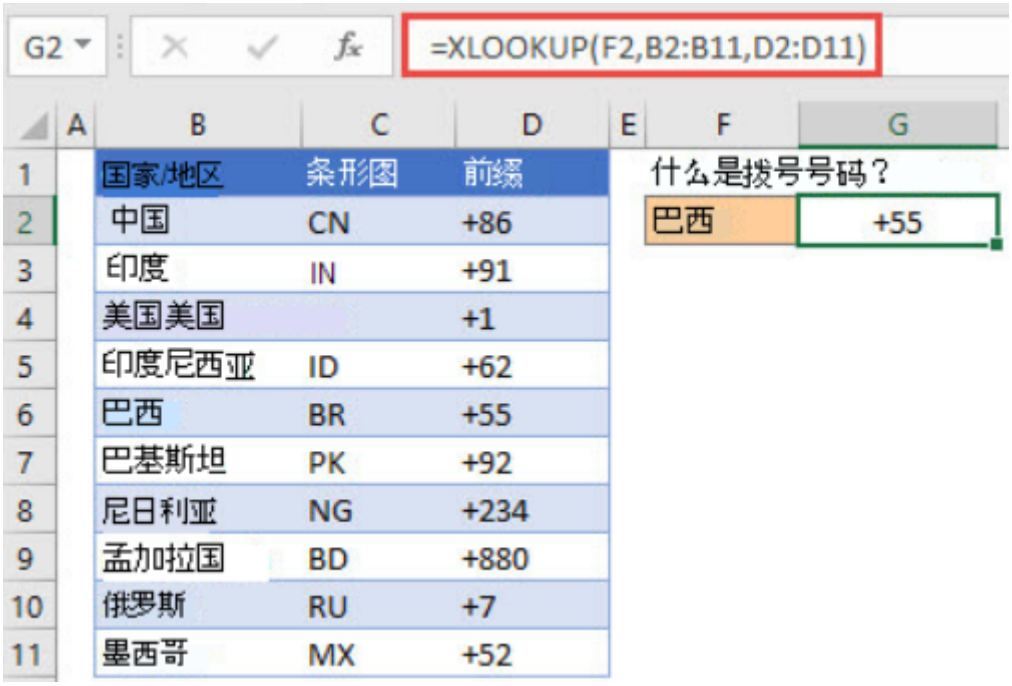
*return_array*要返回的数组或区域

[if_not_found] 如果未找到有效的匹配项，则返回提供的文本。

[match_mode] 指定匹配类型。**0完全匹配**， 如果未找到，则返回 #N/A，这是默认选项。-1完全匹配，如果没有找到，则返回下一个较小的项。1完全匹配， 如果没有找到，则返回下一个较大的项。2通配符匹配。

[search_mode] 指定要使用的搜索模式。**1从第一项开始执行搜索**，这是默认选项。-1从最后一项开始执行反向搜索。2执行依赖于 lookup_array 按升序排序的二进制搜索。 如果未排序，将返回无效结果。

例1：根据F2的国家/地区查找电话前缀



=xlookup(F2, B2:B11, D2:D11)

XLOOKUP使用**查找数组和返回数组**，而 VLOOKUP 使用**后跟列索引号的单个表数组**。 在这种情况下，等效的 VLOOKUP 公式为： **=VLOOKUP (F2，B2：D11,3，FALSE)**

例2：根据员工ID查找姓名与部门

=xlokup(B2, B5:B14, C5:D14, "ID not found")

例3：使用嵌套XLOOKUP函数执行垂直和水平匹配。

首先查找 B 列中的**“毛利润”**，然后在表 (范围 C5：F5) 的上一行中查找 **Qtr1**，最后返回两者交集处的值。

=xlookup(D2, $B$6:$B$17, xlookup($C3, $C5:$G5, $C6:$G17)) 行->列 确定值的位置。可使用HLookup、index+match替换

=HLOOKUP(C3,B5:G17,4,FALSE)

=INDEX($B$5:$G$17,MATCH(D2,$B$5:$B$17,0),MATCH($C3,$B$5:$G$5,0))

5、index+match：类似于XLookup执行垂直和水平匹配。

index(array, match(), mach())

·index：返回表格或区域中的值或值的引用

--数组形式：`INDEX(array, row_num, [column_num])`

array 单元格区域或数组常量

row_num 选择数组中的某行，函数从该行返回数值。

[column_num] 选择数组中的某列，函数从该列返回数值。 如果省略 column_num，则需使用 row_num。

如果同时使用 row_num 和 column_num 参数，则 INDEX 返回 row_num 和 column_num 交叉处的单元格中的值。

例：位于区域 A2:B3 中第二行和第二列交叉处的数值

=INDEX(A2:B3,2,2)

--引用格式：

```
INDEX(reference, row_num, [column_num], [area_num])
```

reference 对一个或多个单元格区域的引用。如果为引用输入一个不连续的区域，必须将其用括号括起来。

row_num 引用中某行的行号，函数从该行返回一个引用。

[column_num] 引用中某列的列标，函数从该列返回一个引用

[area_num] 选择要返回 row_num 和 column_num 的交叉点的引用区域

例：第一个区域 A1:C6 中第二行和第二列的交叉处，即单元格 B2 的内容。

=INDEX((A1:C6, A8:C11), 2, 2, 1)

·match：在 范围 单元格中搜索特定的项，然后返回该项在此区域中的**相对位置**。

```
MATCH(lookup_value, lookup_array, [match_type])
```

lookup_value 查找值，可以为值（数字、文本或逻辑值）或对数字、文本或逻辑值的单元格引用。

lookup_array 要搜索的单元格区域。

[match_type] 匹配模式。1 或省略，查找小于或等于 lookup_value 的最大值，默认值为 1。0查找完全等于 lookup_value 的第一个值。-1查找大于或等于 lookup_value的最小值。

例：

=match(39,B2:B5,1) 返回单元格区域 B2:B5 中最接近的下个最小值 (38) 的位置：2

=match(41,B2:B5,0) 返回单元格区域 B2:B5 中值 41 的位置：4