

1. 数据分析基础

一、数据特性

计量层次对数据进行分类

计量层次低：字段信息无法进行计算比较 如 书本 黄瓜 番茄 树木 牛

计量层次高：数据可以进行计量比较 如 1 2 3 4 5 6

定类数据：

- (1) 按照类别属性进行分类，各类别之间是平等并列关系
- (2) 这种数据不带数量信息，并且不能在各类别间进行排序
- (3) 主要数值运算，计算每一类别中的项目的频数和频率

如 颜色：红黄蓝 性别：男女

定序数据

- 1) 各数据之间可以排序和比较优劣；
- 2) 通过对文字编码进行排序，可表示彼此的高低差异

如 教育程度：小学 初中 高中 大学 季度：春夏秋冬 等级：合格 良好 优秀

定距数据

- 1) 具有一定单位的实际测量值；
- 2) 精确性比定类数据和定序数据更高；
- 3) 可计算各变量之间的实际差距（加、减）

如 温度：33 21 9 年龄：30 93 24 成绩：50 70 80

定比数据

- 1) 可以比较大大小，进行加减乘除运算；
- 2) 定距尺度中，0表示数值，定比尺度中，0表示“没有”；如（定距数据）温度
- 3) 定比数据中存在绝对零点，定距数据中不存在。

如 利润：10万 30万 薪酬：1000 3000 4000 用户数：230 3500 3000

定性数据（定类数据、定序数据）是一组表示事物性质、规定事物类别的文字表述型数据

定量数据（定距数据、定比数据）指以数量形式存在着的属性，并因此可以对其进行测量

数据矩阵：属性 记录

统计指标：体现总体数量特征的概念和数值，根据数据分析的目的不同，统计指标也会变化

1. 总量指标：（GDP，总人口，销售总额）特定条件下的总规模、总水平或工作总量。是一种最基本的统计指标
2. 平均指标：用一个数字显示其一般水平，集中趋势指标
3. 相对指标：两个有联系的现象数值相比得到的比率，描述相对关系而不是总体
4. 比例=各数据/总比例=数据项：数据项倍数 突出上升、增长幅度
5. 环比增长率= $(\text{本期数} - \text{上期数}) / \text{上期数} * 100\%$
同比增长率= $(\text{本期数} - \text{同期数}) / \text{同期数} * 100\%$

1. 平均数

2. 中位数 排序后中间的数值 区分奇数偶数2种情况

3. 众数 各个数值出现的次数 一般按照区间来划分

离散趋势

极差 $\min - \max$ 的绝对值

平均差=|每个数据项的值-均值|的总和/数据项个数，数值小-离散小，但是异常值对其影响比较大（尤其是在样本量小的情况下）

标准差，放大离散程度 体现内部的离散程度 一般标准差是最常用的离散指标（股票、风控等）

分布形态：数据图表化后呈现出来的形态

平均值、中位数、众数----体现平均水平

极差、平均差、标准差-----体现一组数据样本

高度-----一般水平：均值

宽度-----离散程度

正态分布：左右对称：身高、体重、天气、降雨量（随着数据越多）

左偏分布（左高右低）：eg 考试成绩，死亡年龄、资产变化情况

右偏分布（右高左低）：eg 药物有效性、人类运动能力、财富分布

判断异常值：

值/平均数 = 倍数 根据倍数来观察是否存在异常值

异常值：与平均值偏差极大和极小的值，也叫做离群点；

处理异常值：

错误记录---修改正确

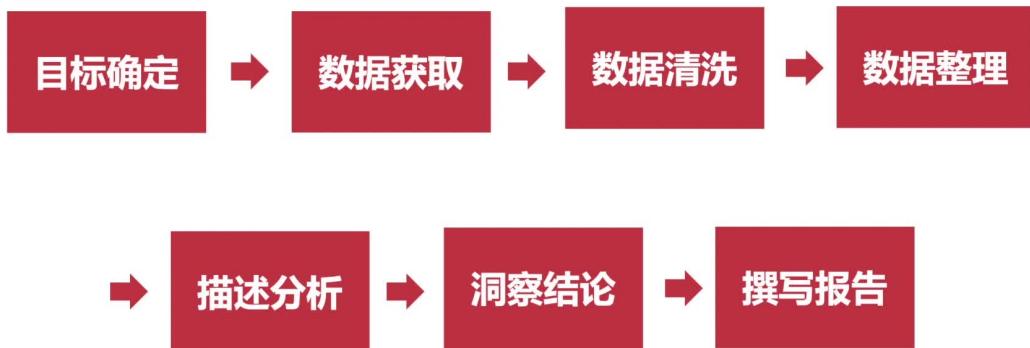
错误添加---删除

正确、真实---是否反映特殊事件---修改、调整或不做处理

错误数据---填充空值或填充样本平均值

正确、真实---根据实际情况调整（数值*需调整比率）

数据分析的流程



数据分析的流程：

目标确定：描述性分析，预测性分析

数据获取：字段设计（平均销售额，销售总额，增减幅度），数据提取（销售管理软件，导入导出），互联网企业使用SQL从数据库提取

数据清洗：异常值的识别判定处理，空白值，无效值，重复值

数据整理：格式化（日期的处理，行列的格式化），指标计算（基础的计算，如平均值、总额）

描述分析：数据描述（数据的基本情况，数据总数，时间跨度，数据来源等），指标统计（分析实际情况的数据指标）。（变化、分别、对比、预测）

洞察结论：数据报告的核心，体现数据分析能力撰写报告：

报告背景，报告目的，数据基本情况，可视化图表，策略选择

3. 互联网数据分析框架

数据分析 vs 商业分析

硬技能

数据分析工具

找数据、抓数据

统计分析

设计报表

可视化图表

软技能

行业经验

产品设计、优化

营销推广资源

基于业务场景，拆解指标

互联网商业模式

互联网简介：

- 基本特征：连接，技术，价值
- 普及率：逐年稳步上升
- 行业格局：企业，服务类型

产品分类：

- 服务对象：ToB (企业，社团，政府：理性，明确的指标，效率>体验)， ToC (个人用户：感性，功能设计突出，突出人性)
- 运营平台：移动，PC，智能设备 (iwatch, ipad)
- 用户需求：交易 (Taobao, JD)，社交 (Weibo)，工具 (jetbrains)，平台 (阿里巴巴，腾讯)，游戏，内容 (抖音)

行业分析：

- 以行业为分析对象：具有高度相似性，竞争性
- 企业相互作用关系：竞争，合作，供应商，服务商
- 产业本身发展：市场规模，需求，增速，未来潜力
- 联系，区域分布：产业链上下游，国家和城市分布

行业分析流程和方法：

确立目标 -> 收集资料 -> **结构化分析 (总量，细分，预测，竞争关系 (波特五力法))** -> 内容呈现

Z.B.: 直播电商：

- 发展背景 (过去): (更多详细行业报告可以Google, Baidu)
 - 宏观：时间线，背景 (为什么兴起, 满足了什么需求) [参照 10.4-5]
 - 微观：整体成交额，用户规模，对比其他方式的占比 [参照 10.4-5]
- 竞争分析：
 - 波特五力法：

- 潜在新进入者：新平台
 - 供应方议价能力 (品牌, 经销商, 工厂) [和KOL, 主播成反比。小主播就只能给什么条件接什么条件]: **主播, 平台**
 - 同业竞争: [内容平台和电商平台存在合作博弈关系]
 - 电商+直播 (淘宝, 京东)
 - 内容+电商 (小红书, 抖音, 快手)
 - 买方议价能力: 消费者 (选择直播平台, 需求, 价格, 信息不透明)
 - 潜在替代产品 [alternative]: 线下零售, 实体店, 智能化服务, 图文形式带货, 粉丝经济
- 生态分析:
 - 产业图谱:
 - (上下游公司和产品, 整体规模趋势, 市场数据表现, 上游:广告商利润, 中游:主播销售额, 下游:人均消费额)
 - 如何梳理产业链: **供应商 -> (MCN, 主播) -> 渠道 [JD, PDD] 电商 内容[Bilibili, Tiktok] 社交 [Wechat] -> 用户**
 - 产业链平台 (win.d), 数据服务平台 (**新榜newrank.cn** [平台数据], **itrustdata.cn** [数据服务提供商])
 - 趋势预测 (未来):
 - PEST分析法:
 - Politics: 政策, 监管, 处罚
 - Economy: 宏观经济, 资金, 经济状况 -> 消费行为, 口红效应 [基础消费品]
 - Society: 人口因素, 消费心理, 生活方式, 文化传统 -> 年龄段, 性价比, 品质, 消费升级, 国货 [战狼]
 - Technology: 5G, AR, 个性化推荐, 智能手机普及 -> 沉浸式用户体验

数据指标体系

[数据指标模型]

用户维度的分析模型:

生命周期模型：用户接触到离开产品的过程

- 导入期: 引导用户完成注册
- 成长期: 体验行为
- 成熟期: 依赖行为
- 休眠期: 一段时间未登录
- 流失期: 长时间未登录

AARRR模型 [自上而下]: 从产品营销角度, 实现用户管理

- Acquisition: 获取用户
- Activation: 提高用户活跃度
- Retention: 维持用户留存率
- Revenue: 获取收入
- Referral: 自传播

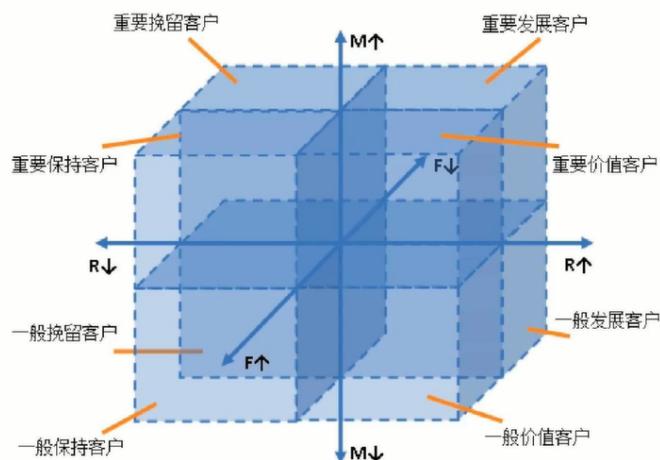
RARRA模型: (由AARRR转变而来)

- Retention: 维持用户留存率 [最重要, 有回头客, 建立用户群]
- Activation: 提高用户活跃度
- Referral: 自传播
- Revenue: 获取收入
- Acquisition: 获取用户

RFM模型 (CRM系统): 用户为分析维度, 从消费行为的角度采取差异化营销策略

更加准确的衡量客户价值和客户创造利润的能力, 并执行对应的营销手段.

RFM模型



用户分层模型

- 最近一次消费 Recency
- 消费频率 Frequency
- 消费金额 Monetary

5W2H: 提问方式, 快速掌握事件本质

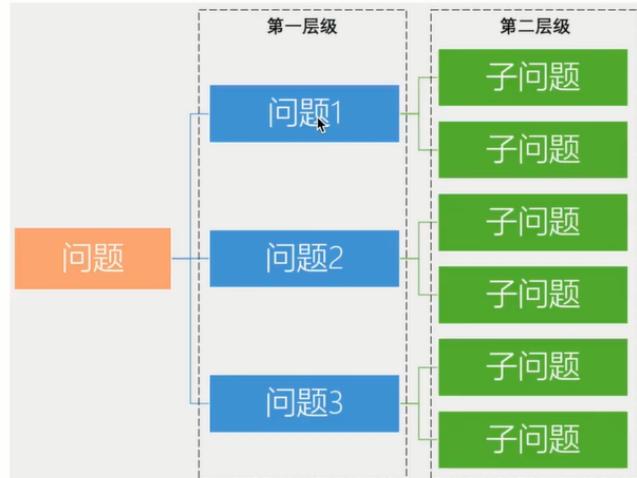
Z.B.: 创建电商用户画像

- why: 运营活动, 营销方案
- what: 用户画像是什么
- who: 年龄, 性别, 工作, 婚姻
- when: 时间, 频次, 金额
- where: 分布地点, 聚集性
- How: 行为轨迹, 支付方式
- How much: 消费水平, 价位

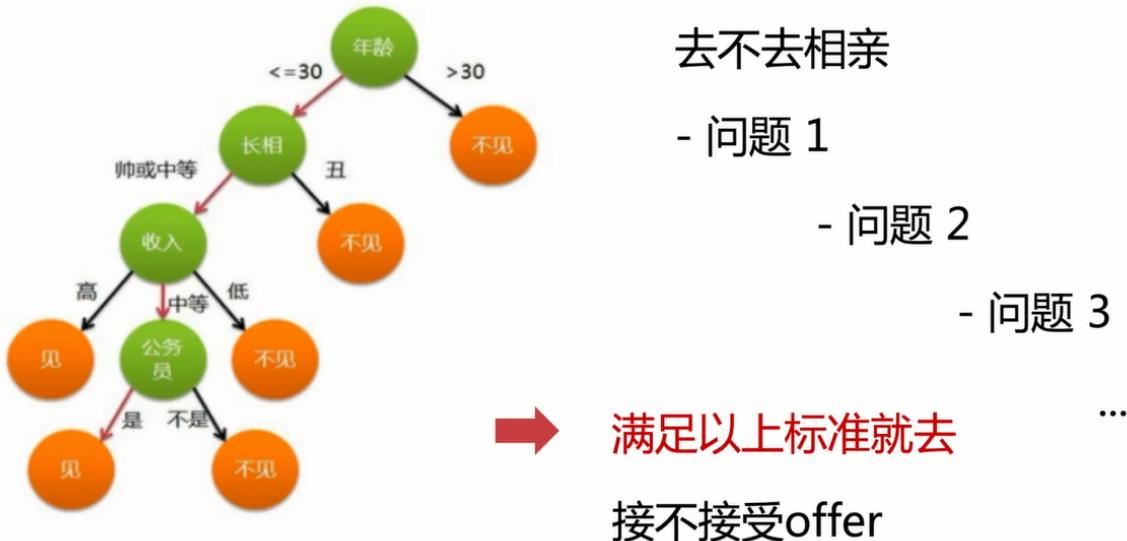
逻辑树: (决策树) 一步步拆解问题, 得出解决方案

- 确保考虑的完整性
- 避免重复思考
- 识别关键问题
- 大问题分解小问题

逻辑树模型

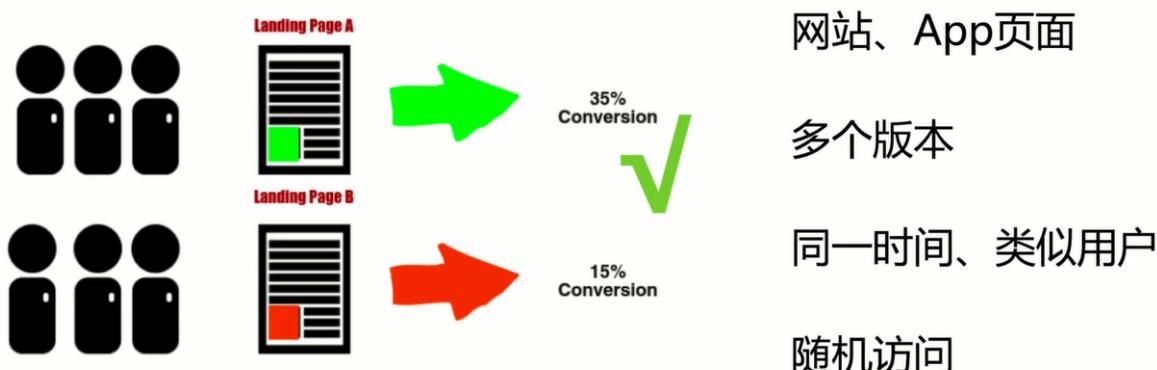


逻辑树模型



A/B 测试模型：(多个方案时，使用分组测试来筛选和确定最终方案)

A/B 测试模型

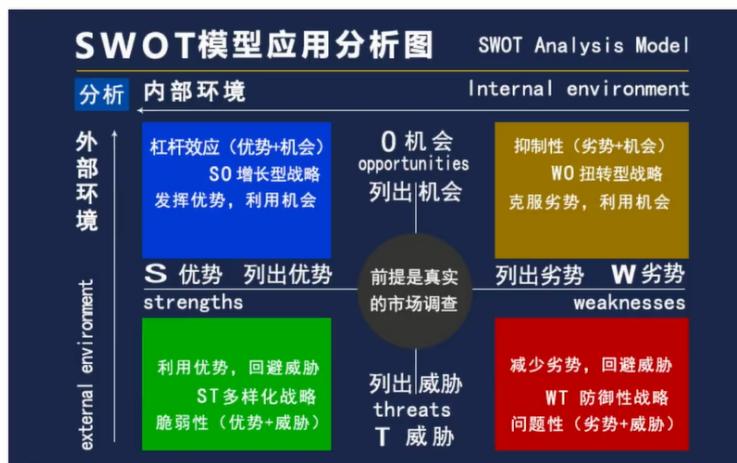


企业战略分析模型:

SWOT: 帮助企业在商业环境中找准自身定位，并在此基础上制定决策

- Strengths [内部有利因素]: SO增长型战略, ST多样化战略
- Weakness [内部不利因素]: WT防御性战略, WO扭转型战略
- Opportunity [外部有利因素]:
- Threat [外部不利因素]:

SWOT



1. 分析环境因素、列表
2. 构造矩阵、轻重缓急
3. 制定行动计划

SWOT

抖音SWOT分析

		机会 Opportunities	威胁 Threats
优势 Strengths	1-庞大的一二线城市用户量，超强的用户粘性 2-原创内容丰富，风格明显 3-个性化算法推荐 4-互动的社区氛围 5-有今日头条母公司强大的现金流支撑 6-明星入驻使抖音影响力更强 7-积极的出海政策使得抖音在全球市场中占据了优势	1-垂直内容维度细化增多 2-明星IP和粉丝经济 3-内容推荐机制智能化(PEST-T) 4-短视频营销价值高(PEST-E) 5-用户的短视频社交趋势 (PEST-S) 6-短视频的市场空间仍然巨大 (PEST-E) 7-海外市场潜力巨大	1-市场竞争激烈 (5-force) 2-BAT短视频业务布局市场 (5-force) 3-短视频竞品替代多(5-force) 4-更加严格的视频内容审核和版权问题 (PEST-P)
	S O策略	1-大力发展垂直领域内容 2-软件开发科学性，多多沉淀用户使用数据，推送个性化、制作精良的短视频，发挥流量价值；使用户在使用中获得满意的服，实现用户满意与短视频APP发展共赢； 3-继续拓展和巩固海外市场 4-除了明星，还可以借助热门综艺的影响力增长用户 5-增加社交功能	ST策略 1. 注重与其他类短视频APP、电商、资讯网站的合作。完善短视频服务产业链，提升用户黏性，用更科学、高明的手法抢占用户流量、提升变现能力 2. 通过不断完善短视频制作技术提升短视频质量，增加用户黏度；
劣势 Weaknesses	1. 泛娱乐内容较多 2. 精准的推荐算法也容易带来信息茧房的问题 3. 盈利模式比较单一	WO策略 1. 拓展多元盈利模式，加强造血 2. 增加推荐范围，拓展用户接收内容范围，避免用户信息疲劳 3. 增加短视频类型，除了娱乐，还可增加其他领域的短视频	WT策略 1-完善短视频审查制度

PEST:

PEST

基于公司战略的眼光，分析企业外部宏观环境



波特五力法：[更多可参考直播电商行业分析报告]

波特五力



④ 数据分析指标体系：[11.4 - 11.9.xlsx] ④

- 有机组成的统计指标
- 依据行业，业务属性而定
- 以AARRR模型为线索：拉新指标，活跃指标，留存指标，转化指标，传播指标 [参考11.4数据指标体系]
 - 拉新指标 [参与用户数，新增用户数，获客成本]: ·B.S. 拼多多红包裂变，转发好友机制实现促活和拉新
 - 线上，线下广告，应用商店优化，电子邮件推送，社交媒体传播，朋友推荐计划
 - 活跃指标：
 - 潜在用户真正使用产品
 - 流失率高达90%
 - 激活时刻：**行为 = (动力-阻力) *助推 + 奖励**
 - 行为：引导用户完成行为
 - 动力：需求强度
 - 阻力：完成行为需要的成本
 - 助推：提示用户完成行为
 - 奖励：完成行为后，用户得到的反馈
 - B.S.: 抖音: 直接刷视频，不需要注册登录，简化操作成本
 - 留存指标：
 - 提高留存率手段：
 - 产品核心价值
 - 满足用户需求：
 - 核心，延伸，需求触发，被动
 - B.S.: 支付宝
 - 核心：收付款
 - 延伸：余额宝，理财
 - 需求触发：app推送，短信提示
 - 被动需求：年度账单，蚂蚁森林
 - 转化指标 [变现]：
 - 免费用户付费
 - 变现模式：广告变现，增值服务 [VIP会员]，电商变现 [平台使用费，营销推广费]，直播变现，数据变现 [可视化，数据支持，解决方案]，游戏变现，金融变现 [金融产品，支付通道]，
 - 商业模式：低成本规模化获客，高效率持续发现
 - 引导用户付费
 - 传播指标：
 - 自传播：
 - **社交货币**: [评价对方的要素 (他穿A)，用户彰显自我的需求]
 - **诱因**: 一谈到什么就能想到什么，形象的绑定
 - **情绪**: 高唤醒，大量传播 [疫情消息]
 - **公共性**: 销量，评价，好评 [从中心理]
 - **实用性**: 奖励机制，拉新返现，送存储空间 [百度网盘]
 - **故事性**: 好奇，沉浸式，意外
 - 病毒/口碑传播

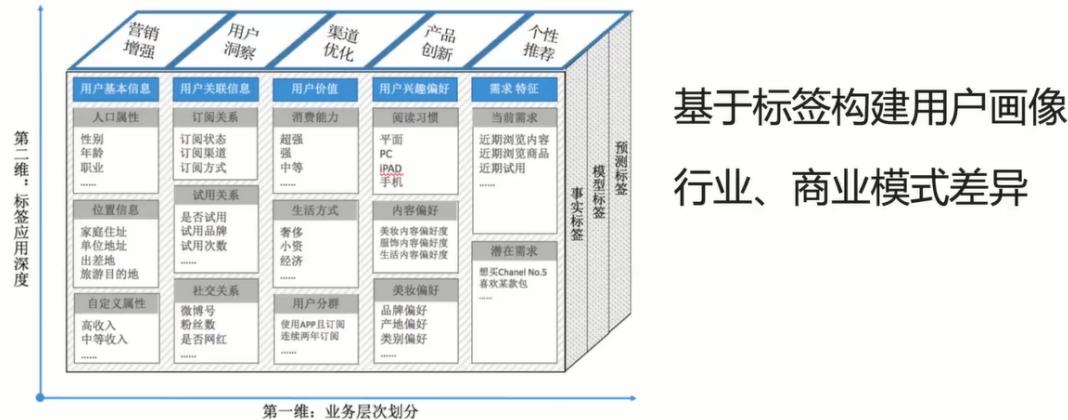
构建用户画像

用户画像：

- is what: 用一组标签 [用户属性, 兴趣偏好] 来形容用户特征, 各类描述用户数据的变量合集
- 用户分群: 品质生活优享, 经济实用, 数码狂热, 某粉狂热
-

用户画像 (标签) 体系

用户画像体系



- 作用：
 - 精细化运营：
 - 对不同用户角色采用不同的营销手段
 - 新产品, 中高端, 定价高
 - 营销手段：
 - 个性化推荐: 针对性推送
 - 广告投发系统: 微信朋友圈, 用户群
 - 活动营销: 微信社群
 - 内容推荐: 头条推荐算法, 兴趣偏好

数据标签系统: [12.2.xlsx 基于用户数据标签和用户画像对用户进行分组]

- 个性化短信推送: 1. 提升销售额 2. 老用户找回 3. 短信推送 4. 标签定位 5. 推送规则 6. 提取用户
7. 完成推送 8. 优化迭代

个性化短信推送			通过watcher查链接埋点情况		以手机号为验证元素导出的订单情况							
推送用户数	推送用户标签	优惠券力度	点击数UV	点击率	提交人数	提交率	提交金额	成交人数	成交金额	提交率	推送用户数/提交人数	提交-成交转化率
2290	领券/提交/未成交	10%	32	1.40%	12	37.50%	30,815	3	3665	0.52%	+	25.00%
92020	领券/提交/成交		1730	1.88%	1417	81.91%	2,385,230	949	1,390,271	1.54%		66.97%
108195	未领券/提交/未成交	6%	553	0.51%	415	75.05%	658,960	116	169,405	0.38%	+	27.95%
124692	领券+未领券/提交/未成交		381	0.31%	212	55.64%	316,457	74	103,554	0.17%		34.91%
90037	未领券/未提交/未成交用户		277	0.31%	58	20.94%	94,769	32	49,569	0.06%		55.17%

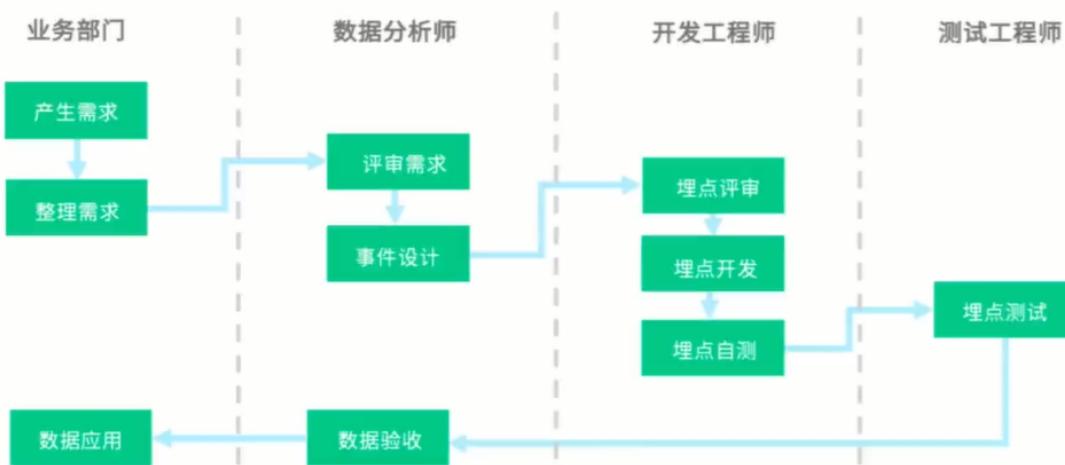
基于数据标签的技术架构



数据埋点：一种数据采集的方式 => 管理，统计带参数的URL

- UTM (Urchin Traffic Monitor)，可以理解为流量监控器，用于帮助监控流量的来源，系列参数：
 - source: 请求来源，类似referrer
 - medium: 用来标记Banner、CPC等广告形式
 - term: 用来标记广告关键词，主要用于SEM [Search Engine Marketing] 投放
 - campaign: 用来标记广告或运营活动的整体的名称
 - content: 用于A/B测试，标记同一广告间细微差别
-

埋点流程



构建用户画像:

- 如何构建用户画像:
 - 用户价值模型: 最近消费时间, 金额, 频数
 - 内容偏好模型: 分发机制, 评论, 点赞, 收藏, 交互行为
- 基础用户画像 [**12.4用户画像标签系统.xlsx**]: 人口属性, 用户行为, 消费习惯, 特征偏好
- 可视化工具: 神策, GrowingIO

构建商品画像:

- 根据消费对象, 构建商品标签
- 基础商品画像 [**12.6商品画像标签系统.xlsx**]: 价格, 库存, 销售, 交互行为

用户画像应用分析:

RFM模型:

- R, F, M 指标将用户分成8类.

RFM模型有什么用

客户类型	最近交易距离当前天数	累计单数	累计交易金额	用户精细化管理
重要价值客户	↑	↑	↑	RFM都很大, 优质客户, 需要保持
重要挽回客户	↓	↑	↑	交易金额和交易次数大, 但最近无交易。需要挽回
重要深耕客户	↑	↓	↑	交易金额大贡献度高, 且最近有交易。需要重点识别
重要挽留客户	↓	↓	↑	交易金额大, 潜在的有价值客户。需要挽留
潜力客户	↑	↑	↓	交易次数大, 且最近有交易。需要挖掘
新客户	↑	↓	↓	最近有交易, 接触的新客户, 有推广价值
一般维持客户	↓	↑	↓ ⁰⁰⁰	交易次数多, 但是贡献不大, 一般维持
流失客户	↓	↓	↓	FM值均低过平均值, 最近也没再发货相当于流失

- RFM模型具体应用实例: [**12.7RFM模型实例应用(K-Means).xlsx**]
 - KMeans: 计算分段阈值
 - 设计客户类型: 如上图
- 输出分析报告: [**12.15案例5: 基于RFM的用户精细化管理.ppt**]

4. 销售, 市场与运营数据分析

网站流量：

- 如何获取流量：
 - 别人的渠道：付费渠道，合作渠道，搜索引擎
 - 自己的渠道：Bilibili, TikTok, 快手，视频，图文，活动
 - 病毒渠道：分享，转发，邀请好友
- 流量数据指标：
 - 站外营销推广指标：衡量不同渠道的效果
 - 网站流量：数量指标 ⊲ 站外到站内
 - 网站流量：质量指标 ⊲ 留存指标
 - 流量指标：[13.2流量指标.xlsx]
- 分析模型：
 - 流量波动检测模型：异常数据告警
 - 渠道特征聚类模型：对投放渠道归类和分析
 - 流量预测模型：基于历史数据预测 多少投放量才能完成KPI

广告引流聚类分析：

- 数据分析：[Ad_Performance.py]
- 数据预处理：
 - 计算，合并相关性 -> 避免聚类算法重复计算 夸大特征表现
 - data.corr()
 - 热力图呈现
 - data.drop(['column'], axis=1)
 - 两个相关性高的变量只保留一个
 - 标准化：[标准化比较的维度，固定的值，统一的分析标准，规范不同规模和量纲的数据]
 - Z-score: $x' = (x - \text{mean}) / \text{std}$
 - 中心化，正太分布 [均值为0，方差为1]，会改变原有数据的分布形态
 - Min-Max: $x' = (x - \text{min}) / (\text{max} - \text{min})$
 - 数据进行线性变换，数据落入0-1的区间
 - MinMaxScaler()
 - Fit_transform(matrix)
 - 特征数字化 (One-Hot编码)：定类数据 [性别，职位，学历] 转为数值型数据，参与运算
 - 0和1 -> False, True
 - OneHotEncoder(matrix)
 - 平均轮廓系数：确定最佳K值 $s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$
 - KMeans.fit_predict()
 - Metrics.silhouette_score()
 - 聚类结果分析：[Ad_Performance.py]
- 可视化：雷达图 [Ad_Performance.py]

漏斗分析模型：

- 描述序列的环节与环节, 以及**关键节点**的转化程度, 有上下关系, 有步骤关系的模型
- 价值: 通过对转化的检测与优化提升转化效率
- 应用场景:
 - 运营过程, 效率: 找到薄弱环节, 针对性提升



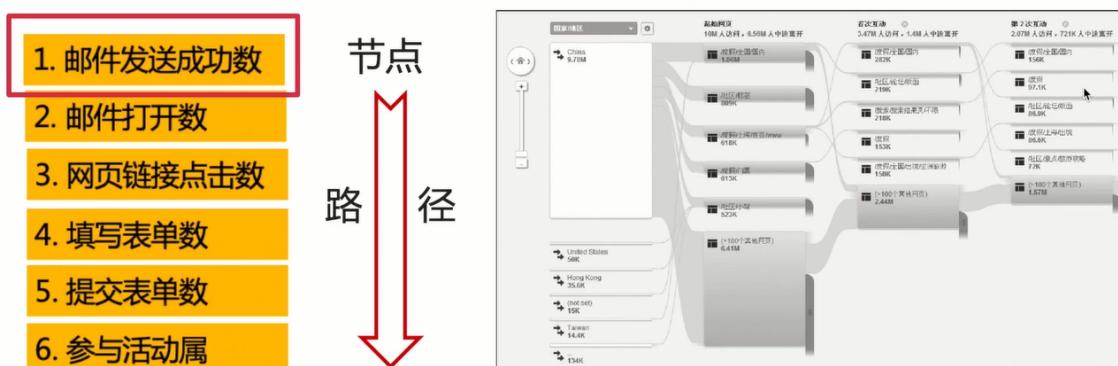
运营过程监控 运营效率分析

用户关键 路径分析

产品优化

- 识别用户行为特征

用户关键路径分析



Google Analytics 用户路径分析图

- 分析关键节点转化效率 VS 用户体验

•



- B.S.: 活动链接的邮件:

- 邮件发送成功数
- 邮件打开数
- 网页链接点击数

- 填写表单数
- 提交表单数
- 参与活动数

- 实例：用户下单流程

- 点击率，流失率判断流量来源的质量是否过关

落地页



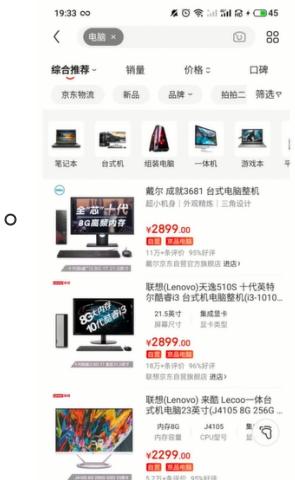
落地页



中间页



中间页：搜索页



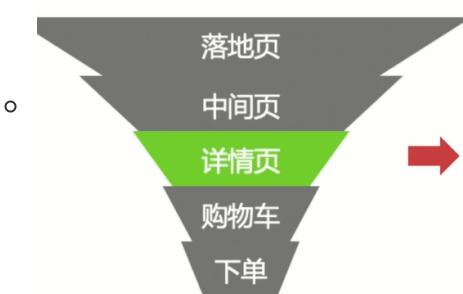
- 搜索点击率=点击次数/搜索次数
- UV到详情页转化率
- 搜索无结果次数
- 搜索次数：搜索词产生的搜索次数
- 搜索人数：搜索词被多少人搜索的数量
- 高级筛选项点击次数

中间页：活动页、频道页



- 点击率 = 页面点击数 / 页面UV数
- 到达详情页转化率
- 成交转化率 = 成交件数 / 详情页UV
- 图文不够吸引、调整优化
- 展示样式不合理、选品失误
- 展示区域太偏、调整位置

详情页



质量 ↑

转化率 ↑

- 平均页面停留时间：总时间/访问UV
- 加入购物车数：意向购买用户（详情页、客服服务、评价）

购物车



挑选多个商品时，节省付款时间

利用满减凑单等促销手段，提高客单价



App、短信、邮件催付

订单页



- 用EXCEL绘制漏斗模型图：[13.18构建漏斗模型.xlsx]

分析消费行为：

- 用户画像：



- 消费动机 (important) 满足客户某种需求，生理或者心理：



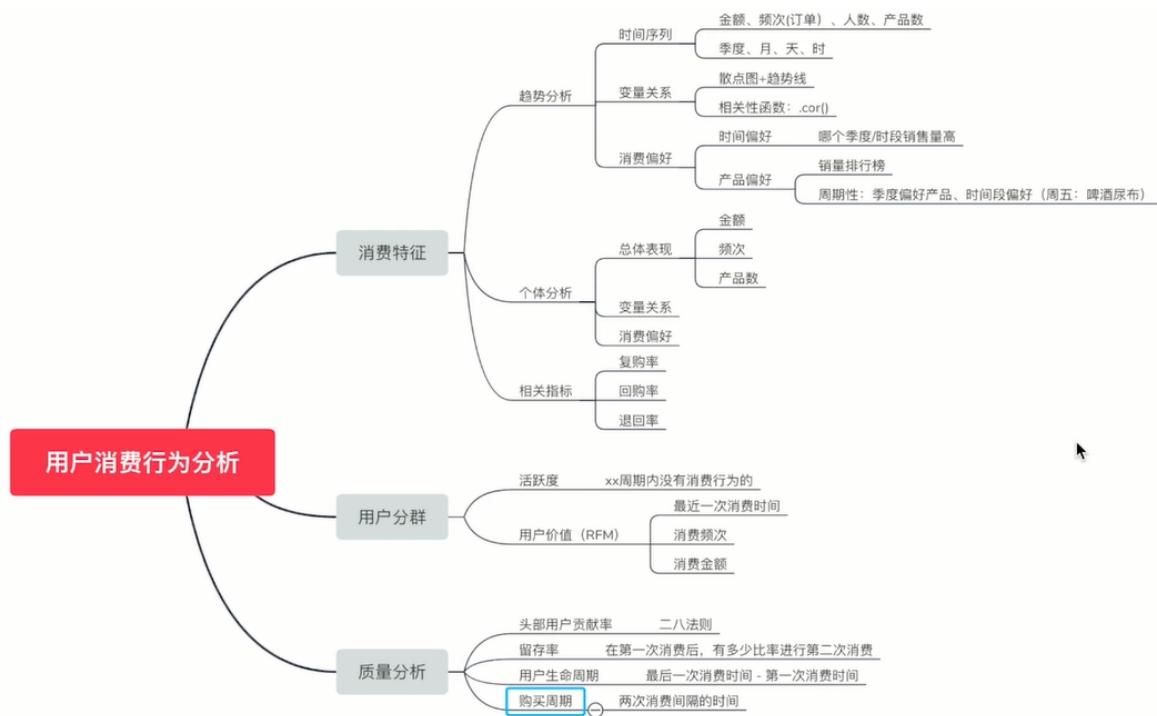
- 消费行为：

- 购买次数，购买总金额，客单价，Recency etc.



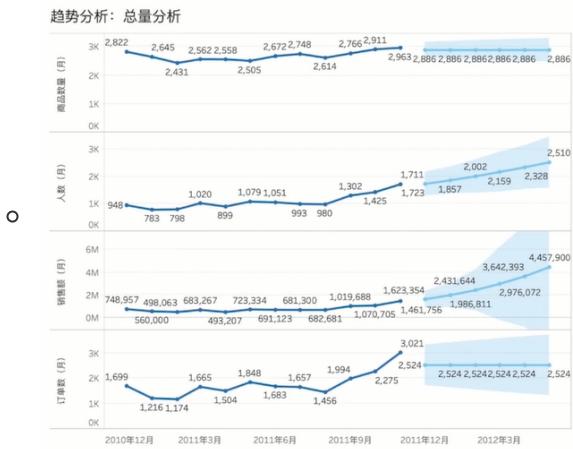
- 消费者行为模式随互联网的发展而变化：
 - 从被动接受，到主动出击 (搜索引擎)
 - interaction, share 用户之间相互传播

用户消费行为分析：

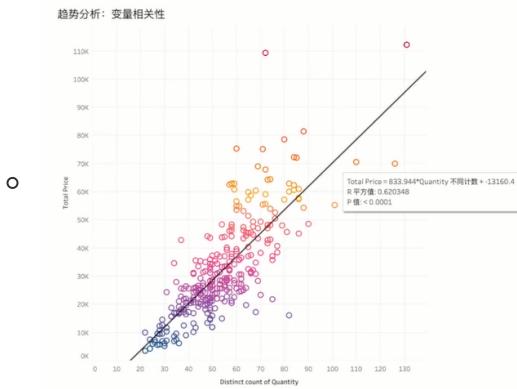


消费特征分析：

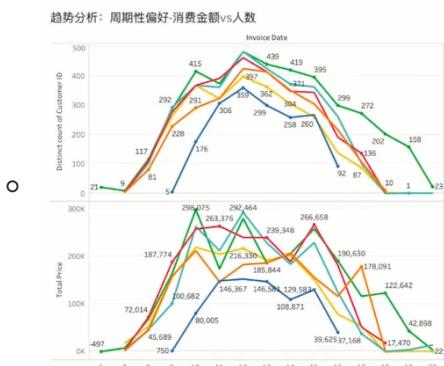
- 趋势分析：[14.4趋势：时间序列, 14.5趋势：变量关系, 14.6趋势：时间偏好]



- 月度消费次数、消费人数、消费产品数与月总销售额均呈上升趋势
- 8月开始数据陡升：是否存在营销推广活动（无法判断）
- 12月用户数据异常分析：数据范围限制、预测趋势两个号
- 11月销量最高：黑色星期五（11月购物节）促进了消费情况

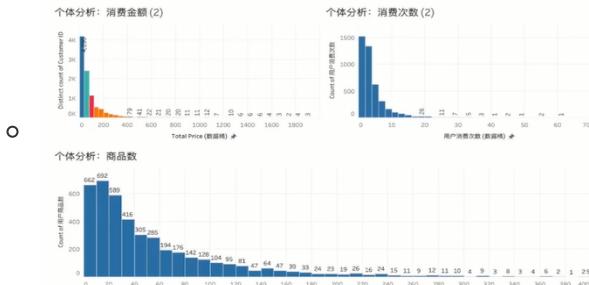


- 销售额与订单数相关系数 (R方) 为 0.62
- 销售额与商品数相关性为 0.52
- 销售额与用户数相关性为 0.51
- 以上指标整体趋势向好，其中销售额与订单数的相关性最高，可以通过促销组合购买等方式，提升用户下单率，进一步提升销售额



- 每月日均消费，以3~5天为一个周期，相较于月初、月末，月中消费金额较多
- 消费时间时段主要集中在10~15点，且习惯于周中消费，整体而言第4季度消费较高
- 周期性是否与会员活动等促销行为相关，且月初经济情况较好，可进一步采取促销措施（降价、优惠券、精细化营销）

• 个体分析：[14.7-14.8.ppt]



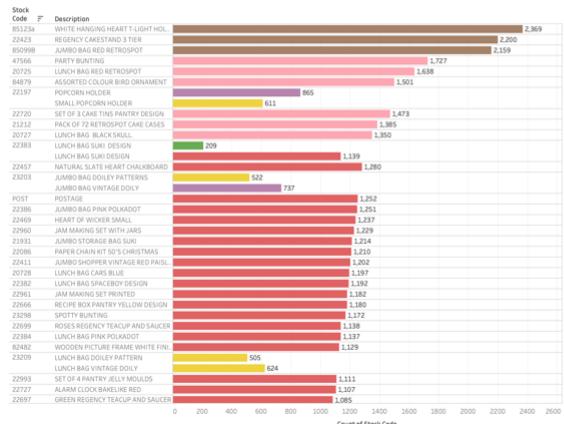
- 人均消费金额：£ 1,898
- 人均消费次数：5次
- 人均消费商品数：61类
- 整体呈现长尾分布，需重点关注头部用户消费偏好

• 相关指标：

- 人，货，场 [营销推广类]：

-

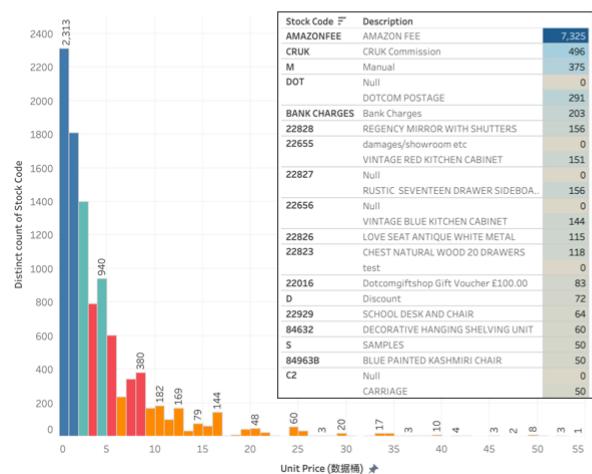
商品分析：销售表现较好的以家具装饰类为主



① 商品数总数3,958种，其中销量最高的是灯台（Light Holder），总销售次数达2,369次，销售表现较好的是家具、装饰类

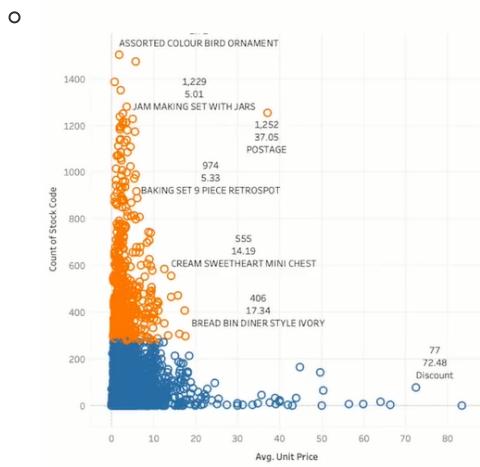
② 其中，商品销售次数主要集中在500次以内，平均每商品销售次数达60次

商品分析：90%商品单价低于20英镑



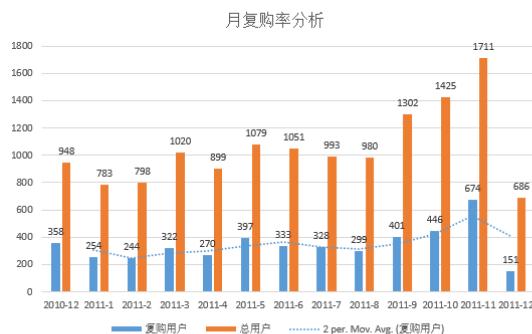
① 商品价格区间为 £ 0~7,325，其中价格超100的商品11种，其中价格最高的为AmazonFee

② 商品单价主要集中在 £ 20 以内，普遍价格比较优惠，可进一步观察销售量与价格的相关性，推广低价热销商品，提升消费次数



- 商品单价主要集中在20英镑以内
- 商品销售量主要在1000件以内，随着价格的上升，销量相应下降
- 对于销量较高的商品，可设计为爆款商品，进一步提升购买转化率

复购率：超过30%用户消费超过2次

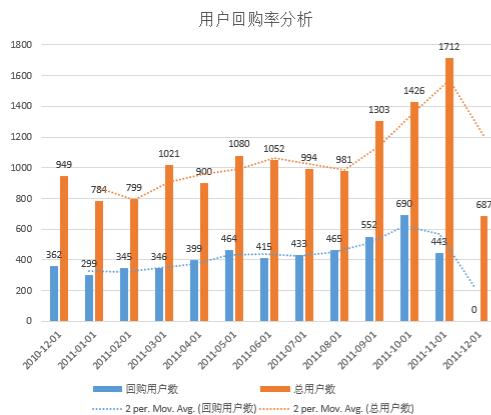


- ① 2010年12月、2011年5月、11月复购率超35%以上，其他月份在30%左右，平均10人里有3人会复购
- ② 新老用户发展情况呈正比，原因可能是商品类型以家居、日常用品为主，每月复购率都差不多
- ③ 5月、11月、12月年中和年末的促销活动较多因此促进了消费，增长了约5个百分比。

```
o
SELECT
    yr,
    mt,
    COUNT(
        IF(
            (
                t1.orders > 1,
                t1.orders,
                NULL
            )
        ) AS a,
        COUNT( t1.orders ) AS b, COUNT(
            IF(
                (
                    t1.orders > 1,
                    t1.orders,
                    NULL
                )
            )
        ) / COUNT( t1.orders )
    )
FROM
    (
        SELECT YEAR
            ( InvoiceDate ) AS yr,
            MONTH ( InvoiceDate ) AS mt,
            CustomerId,
            COUNT( DISTINCT InvoiceNo ) AS orders
        FROM
            OnlineRetail
        GROUP BY
            YEAR ( InvoiceDate ),
            MONTH ( InvoiceDate ),
            CustomerId
        ) AS t1
    GROUP BY
        yr,
        mt;
    (TIME, '%Y-%m')
```

- o

回购率：每月回购用户数超40%

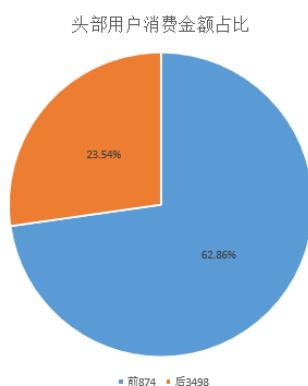


- ① 用户的回购率大于复购率，约40%左右浮动，回购人数趋近稳定，波动主要基于消费总人数的变动，存在营销与淡旺季的因素
- ② 部分回购用户消费行为稳定，与之前每月复购用户有重合，属于优质用户
- ③ 综合分析，可以得出，新客的整体质量低于老客，老客的忠诚度表现较好，消费频次稍次

```
• SELECT DATE_FORMAT(m1, '%Y-%m') , COUNT(m1) , COUNT(m2) , COUNT(m2)/COUNT(m1)
  FROM
  (SELECT CUSTOMERID,DATE_FORMAT(INVOICEDATE, '%Y-%m-01') as m1 FROM
  OnlineRetail
  -- WHERE DATE_FORMAT(INVOICEDATE, '%Y-%m') = '2011-01'
  GROUP BY DATE_FORMAT(INVOICEDATE, '%Y-%m-01'),customerid) A
  LEFT JOIN
  (SELECT CUSTOMERID,DATE_FORMAT(INVOICEDATE, '%Y-%m-01') as m2 FROM
  OnlineRetail
  -- WHERE DATE_FORMAT(INVOICEDATE, '%Y-%m') = '2011-02'
  GROUP BY DATE_FORMAT(INVOICEDATE, '%Y-%m-01'),customerid) B
  ON A.CUSTOMERID = B.CUSTOMERID
  AND m1 = DATE_SUB(m2, INTERVAL 1 MONTH)
  GROUP BY m1;
```

用户分层与质量分析：

- 消费金额前20%用户，贡献率达62%



- ① 总用户人数为4372人，总消费金额 £ 9,747,747
- ② 头部用户300人贡献了45%的消费金额，人均是后3500名客户的23倍
- ③ 消费金额靠后的3500名用户贡献额约23%左右，整体满足二八法则

```
• SELECT
  SUM(SALES)/9769872 AS '消费占比',
  COUNT(us)/4372 AS '用户占比',
  SUM(SALES)/COUNT(us) AS '客单价' FROM
```

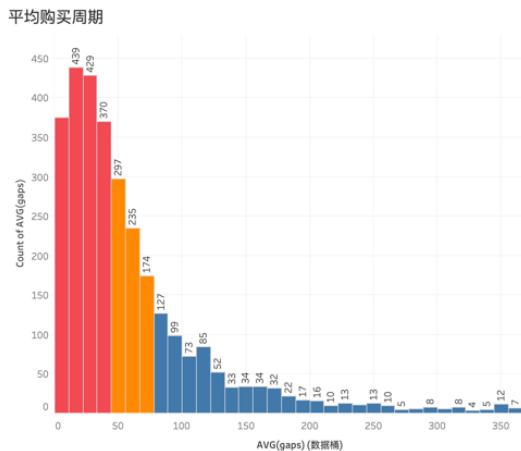
```

(SELECT CUSTOMERID AS us,ROUND(SUM(UNITPRICE * QUANTITY),2) AS SALES
FROM OnlineRetail WHERE CustomerID IS NOT NULL
GROUP BY CUSTOMERID
ORDER BY SUM(UNITPRICE * QUANTITY) DESC LIMIT 874) T1;
-- 判断质量较低的xxx个用户他们的总消费金额/总消费金额(80%的用户贡献了20%销售额)SELECT

SELECT SUM(SALES)/9769872 AS '消费占比',
COUNT(us)/4372 AS '用户占比',
SUM(SALES)/COUNT(us) AS '客单价' FROM
(SELECT CUSTOMERID AS us,ROUND(SUM(UNITPRICE * QUANTITY),2) AS SALES
FROM OnlineRetail WHERE CustomerID IS NOT NULL
AND QUANTITY>0
GROUP BY CUSTOMERID
ORDER BY SUM(UNITPRICE * QUANTITY) LIMIT 3000) T1;

```

- 平均购买周期约32天，价值仍有待挖掘



- ① 大部分用户消费间隔较短，在0~50天左右，平均消费间隔为32天
- ② 可通过消费后立即赠送优惠券，10天后询问用户体验，20天后提醒优惠到期，30天后短信推送等方式，缩短用户平均购买周期，提升销售额

- ```

SELECT CustomerID, AVG(gap) as 平均购买周期 FROM(
select CustomerID,time1,time2 ,datediff(time1,time2) AS gap
from(
select
CustomerID,
InvoiceNo,
InvoiceDate as time1,
-- ROW_NUMBER() OVER(PARTITION BY CustomerID ORDER BY InvoiceDate) AS rank1,
LAG(InvoiceDate,1) OVER(PARTITION BY CustomerID) AS time2 --获取时间窗口，上下N行
from OnlineRetail
WHERE CustomerID is not NULL
GROUP BY InvoiceNo,CustomerID,InvoiceDate)a)b
GROUP BY CustomerID HAVING AVG(gap) >0
ORDER BY AVG(gap);

```

## 预测销售额、调整运营策略

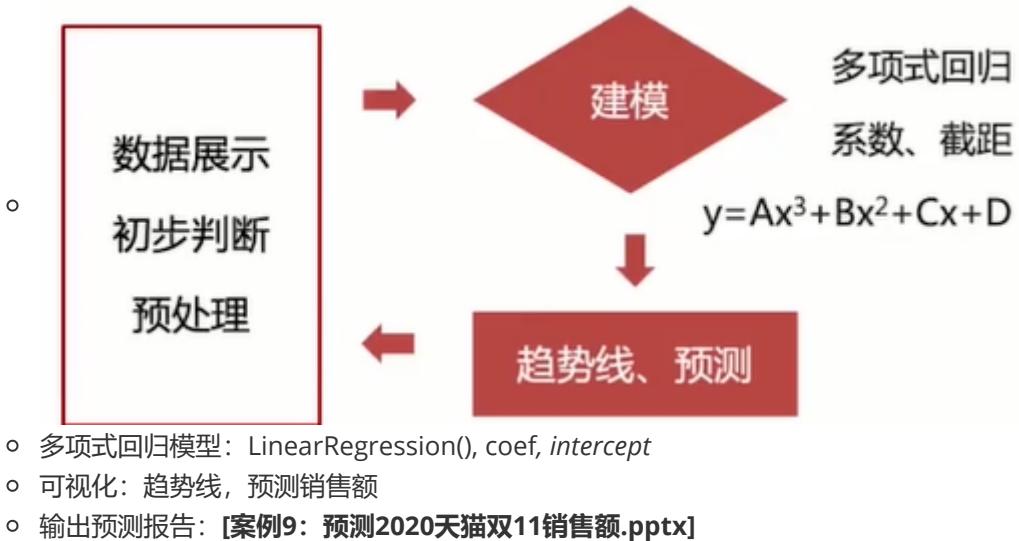
## GMV(销售额):

- 设定企业与各部门KPI指标，并以此为目的分配资源



## GMV预测模型:

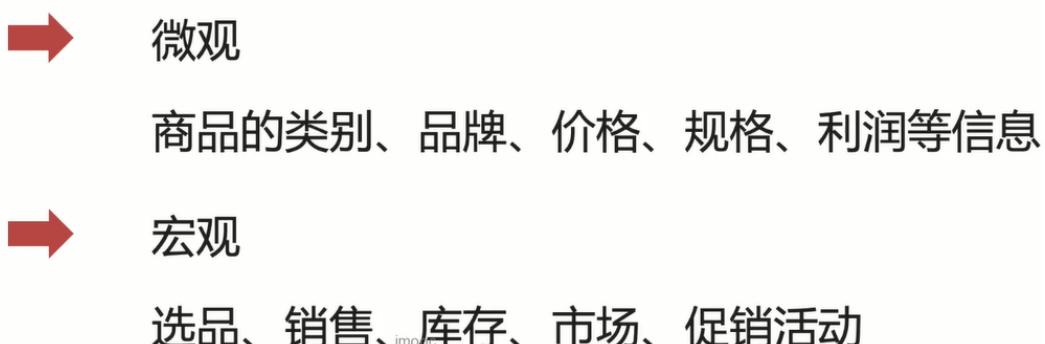
- Excel: 数据建立图标，插入多项式回归；
- Python回归分析: [predictive\_model.py]



## 商品分析:

- 基于商品基础数据, 销售数据, 选品, 销售, 库存, 市场, 促销活动
- 分析商品基础数据、销售数据

@300241



- 商品常用指标:

### 销售指标

- ✓ 订单转化率 : 订单量/访问量
- ✓ 支付转化率 : 完成支付的用户数/需要支付的用户数
- ✓ 订单有效率/废单率 : (有效订单量/订单量)/(1-订单有效率)
- ✓ 毛利 : 利润指标, 商品销售额-商品进货成本
- ✓ 毛利率 : 毛利/商品销售额

## 促销活动指标

- ✓ 每订单成本：费用/订单量（广告cps、运营优惠券）
- ✓ 每有效订单成本：费用/有效订单量（更真实的评估指标）
- ✓ 每优惠券收益：优惠券带来的订单成交金额/优惠券数量
- ✓ 每积分兑换收益：使用积分兑换的订单成交额/积分兑换量
- ✓ 活动直接/间接收入：活动引入，订单商品是/不是活动商品
- ✓ 活动拉升比例：活动期间收入/非活动期间收入-1

## 供应链指标

- ✓ 库存可用天数：库存商品数量/期内每日商品销售数量
  - ✓ 库存量：一定周期内全部库存商品的数量
  - ✓ 库龄：出库时间-入库时间，过长则滞销，果断则热销
  - ✓ 缺货率：缺货数量/用户订货数量，缺货量=订货量-库存量
- 层次分析法AHP：
    - 应用场景：
      - 新增了几个商品资源位，应该挑选哪些商品
      - 在商品资源位的决策中，哪些商品应该放在最佳位置
      - 在商品流量导航入口，哪些品类应该放在最明显的位置
      - 要选一个商品作为爆款商品推给用户，选哪个商品最合适
    - 构建层次结构：

决策目标

选择最优商品

准则层

引流  
能力

购买  
转化

毛利  
价值

品牌  
效应

方案层

商品A

商品B

商品C

- 构建对比矩阵：[15.12成对比较矩阵.xlsx]
- 方案判断矩阵：[15.13方案判断矩阵.xlsx]
- 计算权重得分：[15.14总得分.xlsx]

## 运营策略分析：

- **运营**：以用户为中心，通过各种方式拉新，促活，提升留存与价值，以满足转化目标的一切行为活动
  - 红包车：骑车发红包，达成用户增长，转化以及单车调配的目的
  - 基于业务需要设计采集的指标
- 如何策划一场运营活动：

### 策划期

- → ✓ 确定活动目标：拉新、转化  
✓ 量化活动目标和时间：需要完成的指标  
✓ 活动形式：玩法、活动页面、用户路径

### 筹备期

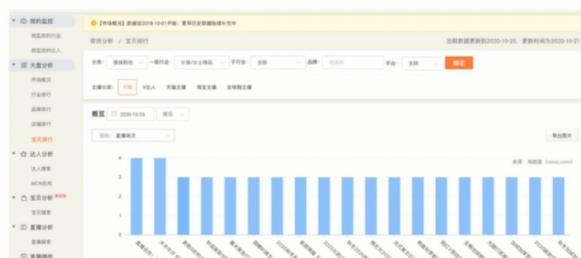
- ✓ 部门协作：分工安排、反馈时间  
✓ 活动物料与推广渠道的准备  
✓ 活动预热：双十一活动介绍页



| APP 项目计划表                          |  | 时间/天    | 一 | 二 | 三 | 四 | 五 | 六 | 日 | 一 | 二 | 三 | 四 | 五 | 六 | 日 | 一 | 二 | 三 | 四 | 五 |
|------------------------------------|--|---------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. 此项目计划表中的单色块代表每一个工作日。            |  |         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 2. 客户方所提供的物料须按照约定的物料清单提供。此项目的所有生效。 |  |         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 负责人                                |  | 工作需求描述  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 整个项目                               |  | 共20个工作日 | 一 | 二 | 三 | 四 | 五 | 六 | 日 | 一 | 二 | 三 | 四 | 五 | 六 | 日 | 一 | 二 | 三 | 四 | 五 |
| 分配工作、整理物料                          |  | 1天      | 黄 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 编辑分类物料                             |  | 3天      |   | 蓝 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 设计制作海报素材                           |  | 3天      |   |   | 红 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 设计制作APP海报                          |  | 2天      |   |   |   | 绿 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 编辑整理图片文字物料并分类                      |  | 6天      |   |   |   |   | 黄 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 技术架构APP环境、构建APP框架                  |  | 8天      |   |   |   |   |   | 绿 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 技术手册整理归档的物料                        |  | 2天      |   |   |   |   |   |   | 绿 |   |   |   |   |   |   |   |   |   |   |   |   |
| 编写白皮书或深入浅出                         |  | 2天      |   |   |   |   |   |   |   | 绿 |   |   |   |   |   |   |   |   |   |   |   |
| 项目测试和修改                            |  | 3天      |   |   |   |   |   |   |   | 绿 |   |   |   |   |   |   |   |   |   |   |   |
| 提交App Store审核（苹果审核日期另计）            |  | 7天左右    |   |   |   |   |   |   |   |   | 黄 |   |   |   |   |   |   |   |   |   |   |

### 发生期

- ✓ 关注进展：数据波动、用户反馈  
✓ 公示活动结果：双十一价格优惠、GMV

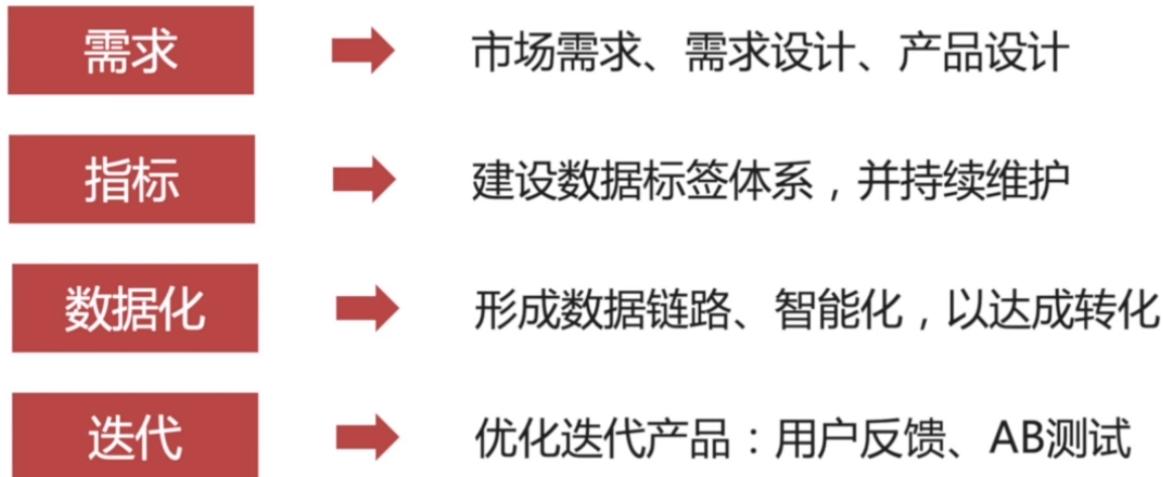


### 复盘

- → ✓ 超预期的部分，予以肯定  
✓ 未达预期的部分：分析问题、明确责任

# 基于数据驱动迭代产品设计

## 数据产品经理职能：



## 促活，提升用户留存：

- B.S.: 大众点评：
  - 普通会员，特殊会员：提供等级成长机制，等级越高，能获得更多优惠，更多实惠.
- 用户活跃度分析模型 RFE：
  - Recency: 最近一次
  - Frequency: 频率
  - Engagements (深度): 浏览时间，商品数，播放，点赞，转发
- RFE实例分析：[16.3练习：RFE模型.xlsx]
- 用户留存，价值分析：
  - Aha Moment: 让用户第一时间发现产品的核心价值，提升用户留存率
  - B.S.: Weibo，新用户自动关注up，自动加载资讯让用户进行交互
  - 步骤：
    - 提出假设：列举新用户可能出现的行为

| 提出假设                 | 创作者      | 消费者         |
|----------------------|----------|-------------|
|                      | 首日发视频个数  | 首日看视频个数     |
| ▪                    | 首个短视频浏览数 | 首日关注作者人数    |
|                      | 首个短视频评论数 | 次日接受push通知数 |
|                      | 首个短视频获赞数 | 首日赞视频个数     |
| ▪ 分组验证：验证每个行为对留存率的影响 |          |             |

## 分组验证

| 看视频个数 | 留存  | 用户占比 |
|-------|-----|------|
| 0     | 30% | 35%  |
| 1     | 40% | 20%  |
| 2     | 41% | 15%  |
| 3     | 42% | ...  |
| 4     | 42% | ...  |
| 5     | 49% | ...  |
| 6     | 50% | ...  |
| 7     | 52% | ...  |
| 8     | 49% | ...  |
| 9     | 53% | ...  |
| 10    | 48% | ...  |

新用户看视频个数



看视频个数 vs 用户留存



描述统计  
相关性分析

聚类分析

- 设计优化：对关键因素进行优化设计

## 设计优化

影响因素：至少看过一个视频

首个内容直接播放

观看后获取成长值（升等级）

增加新手引导

优化内容呈现方式

- 因果测试：持续监测，明确关系

## 因果测试

上线不同的优化策略



查看对应的留存变化



明确最佳优化策略

- 用户留存率计算实例：3天，7天，10天，30天，60天，看消费日期 - 首次消费日期的结果落在哪一个区间 [16.5练习：计算留存率.xlsx]

@300241

```

■
select InvoiceNo as '订单号',
CustomerID as '用户ID',time1 as '消费日期',time2 AS '首次消费日期',
datediff(time1,time2)gap from(
select
CustomerID,
InvoiceNo,
InvoiceDate as time1,
ROW_NUMBER() OVER(PARTITION BY CustomerID ORDER BY InvoiceDate) AS
rank1,
FIRST_VALUE(InvoiceDate) OVER(PARTITION BY CustomerID ORDER BY
InvoiceDate) AS time2
from OnlineRetail
WHERE CustomerID is not NULL
GROUP BY InvoiceNo,CustomerID,InvoiceDate)a;

```

#### ■ 通过时间间隔，计算留存率：

- 用户生命周期计算实例：最后一次消费日期 - 第一次消费日期 [16.6练习：计算用户生命周期.xlsx]
- 案例分析PPT：[案例8：基于电商的用户消费行为分析.xlsx]

## AB测试与功能迭代：

- AB测试基本流程：
  - 分析现状，建立假设：分析业务，确定优先级最高的地方，作出假设，提出优化建议
  - 设定观测指标：
  - 设计与开发：
  - 确定测试时长：
  - 确定分流方案：
  - 采集分析数据：
  - 确定方案，迭代优化：
- 假设检验：
  - 假设：H0原假设
  - 验证：H1备择假设
  - P值：判定假设检验结果的一个参数，具有一定主观性，阈值为0.05
  - 实例：[ABtest\_action.py] [案例13：网站主页改版.pptx]
    - CTR(点击率)差异分布，正态分布 => H1-H0 >1, p值= 0.006

## 异常数据检测：

- 基于孤立森林算法 (IsolationForrest)的异常检测：
  - 利用随机性和特征表现进行划分
  - 实例：[Exception\_Detection.py]
    - IsolationForrest(), decision\_function()[获取数据点的得分]
    - 随机森林 (Random Forrest) 是有监督学习，孤立森林是无监督学习

## 撰写数据报告：

| 报告框架 | 目的     | 内容                                     | 渠道、工具         |
|------|--------|----------------------------------------|---------------|
| 分析背景 | 明确分析目的 | 市场、产品、业务、数据                            |               |
| 分析思路 | 拆解数据维度 | 获取什么数据、从哪儿获取、怎么获取<br>采用什么分析方法（统计描述、建模） | 内部（数据库、爬取）、外部 |
| 分析主体 | 实现数据分析 | 从哪些维度分析、怎么分析、得出什么结论                    | 数据分析工具、思维模型   |
| 结论建议 | 提出解决方案 | 优化方向、如何解决、怎么提升表现                       |               |
| 附录   | 罗列参考资料 |                                        |               |

### 01. 市场是投机的吗

双十一那阵儿，股价跌得厉害，挺多文章从内需，以及新颁布的《关于平台经济领域的反垄断指南征求意见稿》等角度来分析，看着挺有意思。

- 当然了，我相信市场是有效的。投资者会基于自己的认知与经验，尝试理性地判断某件事或者政策的影响，正面还是负面，并最终通过价格的形式体现在股价上。

但同时，它也是投机的。投资者会因市场整体情绪产生过激反应，如2月3日上证暴跌7.72%，疫情引发的避险情绪飙升，市场恐慌抛售。而有人看到风险，就有人看到机会。

所以，在那个节点的买入行为，既是有效的也是投机的。

### 02. 双十一有周期性吗

受以上启发，我尝试思考一个问题，就是双十一有周期性吗？

也就是，是否存在一种规律，比如提前2~3周买入互联网公司的股票，等当天或者某个时间点卖出，就能稳稳的赚钱？

- 我尝试了一下，发现，还真有。并且同样的策略，每年执行，10日收益率能达到5.3%，换算成年化收益率就是193.4%。

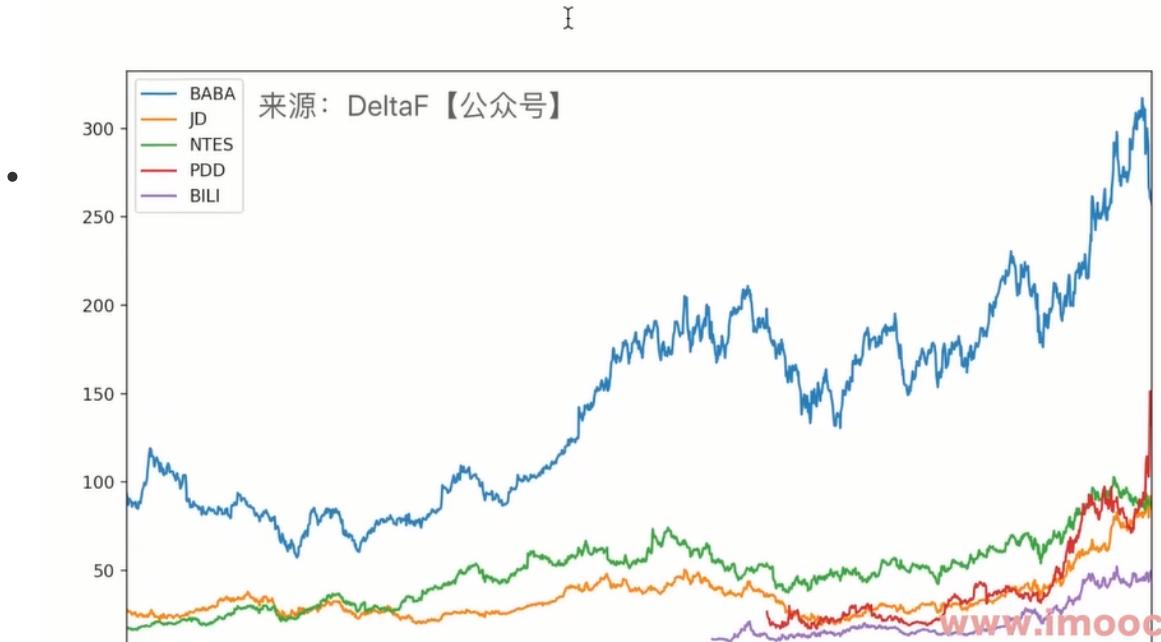
| DeltaF 交易策略 | 自然日 | 绝对收益率 | 年化收益率   |
|-------------|-----|-------|---------|
| 双十一         | 10  | 5.30% | 193.42% |

是不是还挺激动人心的？

接着，我就来给你解密一下。

其中，阿里巴巴和京东 2014 年上市，网易 2000 年，拼多多和哔哩哔哩都是 2018 年。而“双十一”这个概念是由淘宝 2009 年首创，对阿里巴巴股价的影响得从 2014 年算起。所以，就以阿里巴巴上市那天作为起始点。

从【图 1】可以看到，2014~2020 年这 5 家公司的价格涨幅，其中，股价最高的阿里巴巴，上市发行价每股 68 美元。一眼看上去，体量和股价波动都很大。



然而，当我将收盘价归一化，即将初始值设定为 1 美元后【图 2】，我们可以发现，原以为股价波动最剧烈的阿里巴巴，实际上表现相对稳定。而令我比较意外的是网易，历年累计涨幅居然是最高的。

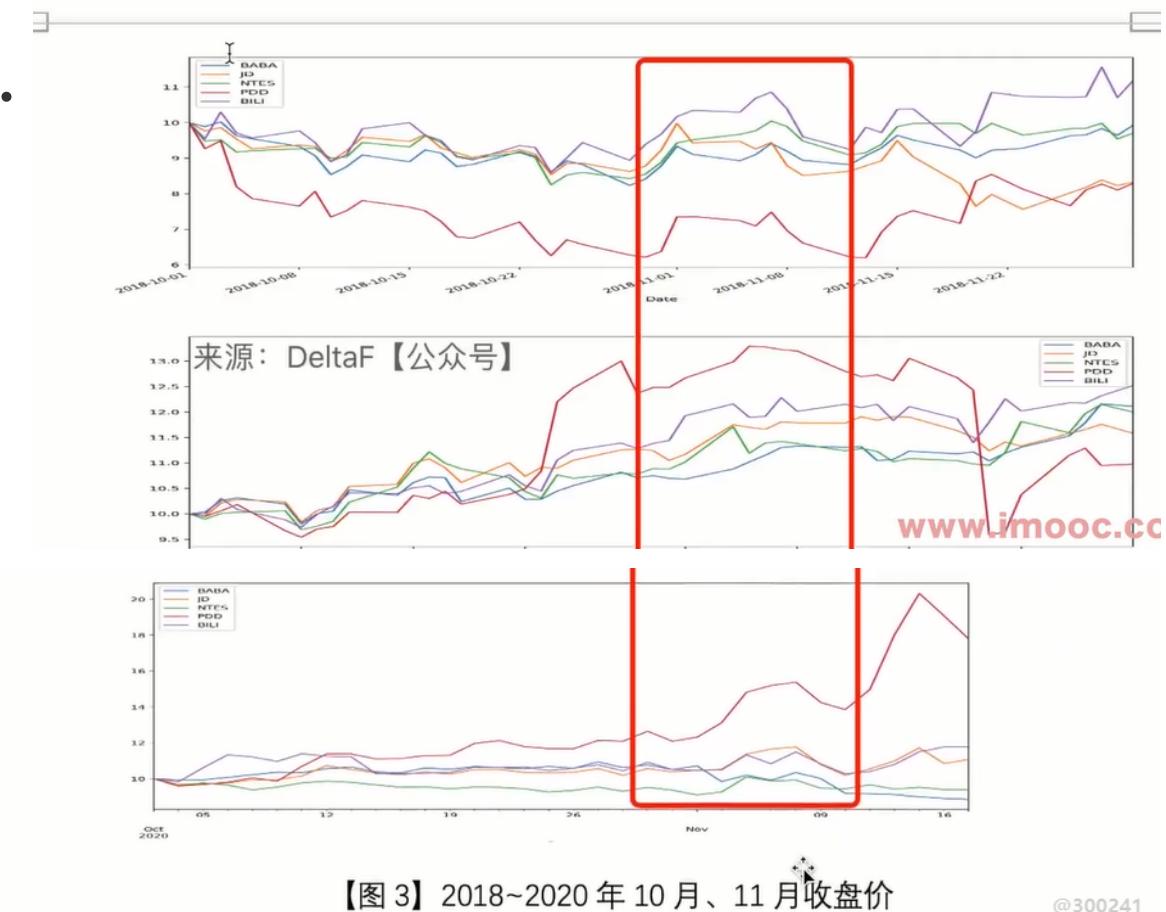
此外，深挖下沉市场与社交电商的拼多多，以及主打 PUGV 的内容平台 B 站，增速都非常迅猛，且自 11 月下旬三季报发布后，两家均暴涨 20%+，后市看好。



我将时间窗口缩小到 2018 年 10~11 月、2019 年 10~11 月以及 2020 年 10~11 月，也就是抽取 5 家互联网双十一前 6 周至后 3 周的数据，来进一步观测。

大家注意【图 3】红框内，也就是 11 月 1~14 日区间，在这 13 天个自然日内，每一年的趋势都有所雷同，即：以 10 月底为开始，以 11 月中旬为结束，形成了类似顶峰的形态。

那么，我们要怎么利用这个规律呢？大家可以思考一下。



我们可以尝试在最低点，也就是 10 月底买入，以及最高点，也就是 11 月的第 1 周卖出。

根据历史数据，2018~2020 年的阶段最低点，分别是 10 月 29、29、28 日，则暂定 10 月 29 日为卖出日。最高点，对应的分别是 11 月 7、8、6 日，我暂定 11 月 7 日为卖出点，时间跨度为 10 个自然日。

此外，如果当天不是交易日，那么就挑天数最近的交易日卖出。如果前后相邻的两个交易日间隔天数一致，那么买入点挑前置的交易日，卖出点挑后置的交易日。

我以收盘价为买卖点，通过计算逐年的买卖收益比，即：(卖出价格 - 买入价格) / 买入价格，得到了 2014~2020 年总共 7 年的交易数据。

如【图 4】所示，正收益率为黄、红色，负收益率为绿色，收益率越高颜色越深。其中，负收益率 8 次，正收益率 19 次，概率比 1 : 2.4，即赢率超过 1 倍。

| %    | BABA   | JD     | NTES   | PDD            | BILI   | AVG.   |
|------|--------|--------|--------|----------------|--------|--------|
| 2014 | 16.53% | 2.66%  | -0.23% | 来源：DeltaF【公众号】 |        | 6.32%  |
| 2015 | -1.02% | 3.86%  | 5.45%  |                |        | 2.77%  |
| 2016 | -2.96% | -1.62% | -5.94% |                |        | -3.51% |
| 2017 | 3.82%  | 6.33%  | 10.88% |                |        | 7.01%  |
| 2018 | 14.33% | 9.51%  | 19.25% | 19.29%         | 21.49% | 16.77% |
| 2019 | 5.52%  | 4.72%  | 5.94%  | 6.94%          | 8.86%  | 6.40%  |
| 2020 | -7.05% | 2.19%  | -0.42% | 12.77%         | -0.80% | 1.34%  |
| AVG. | 4.17%  | 3.95%  | 4.99%  | 13.00%         | 9.85%  | 5.30%  |

图 4：2014~2020 年收益率（10 月 29 日买入、11 月 7 日卖出）

我折合了一下收益率，10 天平均 5.30% 的收益，换算成年化收益大约为 193%，这投资回报比可真香。可比你听消息瞎蒙买入高得多。

## 06. 规则也会失效

这个策略，目前看是有效的，能让你大概率赚到钱。但这之中，也牵扯到资金、费率、突发事件、市场表现、投资者心理素质等各因素。

于是，免不了，我也得说一句，投资有风险，亏了请反省。[doge]

## 演讲技巧：

- **听众是谁 =>** 他们想听什么
- **怎么讲 =>** 通俗易懂，实操性强，注重互动
- **准备讲稿 =>** 大纲，逐字稿，测试演练