数据分析分享：常用的术语和指标

互联网常用名词解释

PV（Page View）页面浏览量

指某段时间内访问网站或某一页面的用户的总数量。通常用来衡量一篇文章(比如阅读量为10万+的文章)或一次活动带来的流量效果，也是评价网站日常流量数据的重要指标。PV可重复累计，以用户访问网站作为统计依据，用户每刷新一次即重新计算一次。

UV（Unique Visitor）独立访客

指来到网站或页面的用户总数。这个用户是独立的，同一用户不同时段访问网站只算作一个独立访客，不会重复累计，通常以PC端的Cookie数量作为统计依据。

CTR （Click–Through–Rate）点击率

指某个广告、横幅、URL被点击的次数和被浏览的总次数的比值。一般用来考核广告投放的引流效果。CTR=点击数（click）/被用户看到的次数

Conversion rate 转化率

指用户完成设定的转化环节的次数和总会话人数的百分比，通常用来评价一个转化环节的好坏，如果转化率较低则急需优化该转化环节。转化率=转化会话数/总会话数

漏斗

通常指产生目标转化前的明确流程，比如在淘宝购物，从点击商品链接到查看详情页，再到查看顾客评价、领取商家优惠券，再到填写地址、付款，每个环节都有可能流失用户，这就要求商家必须做好每一个转化环节，漏斗是评价转化环节优劣的指标。

ROI（Return On Investment ）投资回报率

反映投入和产出的关系，衡量我这个投资值不值得，能给到我多少价值的东西（非单单的利润），这个是站在投资的角度或长远生意上看的。通常用于评估企业对于某项活动的价值，ROI高表示该项目价值高。投资回报率（ROI）= 年利润或年均利润/投资总额×100%。

重复购买率

指消费者在网站中的重复购买次数。

顾客的生命周期价值 (Lifetime Value, LTV)

顾客在他/她的一生中为一个公司产生的预期折算利润。

留存/顾客留存（Retention / Customer Retention)

指建立后能够长期维持的客户关系的百分比。

统计学名词解释

绝对数和相对数

绝对数：是反应客观现象总体在一定时间、一定地点下的总规模、总水平的综合性指标，也是数据分析中常用的指标。比如年GDP，总人口等等。

相对数：是指两个有联系的指标计算而得出的数值，它是反应客观现象之间的数量联系紧密程度的综合指标。相对数一般以倍数、百分数等表示。相对数的计算公式：

相对数=比较值（比数）/基础值（基数）

百分比和百分点

百分比：是相对数中的一种，他表示一个数是另一个数的百分之几，也成为百分率或百分数。百分比的分母是100，也就是用1%作为度量单位，因此便于比较。

百分点：是指不同时期以百分数的形式表示的相对指标的变动幅度，1%等于1个百分点。

频数和频率

频数：一个数据在整体中出现的次数。

频率：某一事件发生的次数与总的事件数之比。频率通常用比例或百分数表示。

比例与比率

比例：是指在总体中各数据占总体的比重，通常反映总体的构成和比例，即部分与整体之间的关系。

比率：是样本(或总体)中各不同类别数据之间的比值，由于比率不是部分与整体之间的对比关系，因而比值可能大于1。

倍数和番数

倍数：用一个数据除以另一个数据获得，倍数一般用来表示上升、增长幅度，一般不表示减少幅度。

番数：指原来数量的2的n次方。

同比和环比

同比：指的是与历史同时期的数据相比较而获得的比值，反应事物发展的相对性。

环比：指与上一个统计时期的值进行对比获得的值，主要反映事物的逐期发展的情况。

变量

变量来源于数学，是计算机语言中能储存计算结果或能表示值抽象概念。变量可以通过变量名访问。

连续变量

在统计学中，变量按变量值是否连续可分为连续变量与离散变量两种。在一定区间内可以任意取值的变量叫连续变量，其数值是连续不断的，相邻两个数值可作无限分割，即可取无限个数值。如:年龄、体重等变量。

离散变量

离散变量的各变量值之间都是以整数断开的，如人数、工厂数、机器台数等，都只能按整数计算。离散变量的数值只能用计数的方法取得。

定性变量

又名分类变量，观测的个体只能归属于几种互不相容类别中的一种时，一般是用非数字来表达其类别，这样的观测数据称为定性变量。可以理解成可以分类别的变量，如学历、性别、婚否等。

均值

即平均值，平均数是表示一组数据集中趋势的量数，是指在一组数据中所有数据之和再除以这组数据的个数。

中位数

对于有限的数集，可以通过把所有观察值高低排序后找出正中间的一个作为中位数。如果观察值有偶数个，通常取最中间的两个数值的平均数作为中位数。

缺失值

它指的是现有数据集中某个或某些属性的值是不完全的。

缺失率

某属性的缺失率=数据集中某属性的缺失值个数/数据集总行数。

异常值

指一组测定值中与平均值的偏差超过两倍标准差的测定值，与平均值的偏差超过三倍标准差的测定值，称为高度异常的异常值。

方差

是在概率论和统计方差衡量随机变量或一组数据时离散程度的度量。概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数。在许多实际问题中，研究方差即偏离程度有着重要意义。方差是衡量源数据和期望值相差的度量值。

标准差

中文环境中又常称均方差，是离均差平方的算术平均数的平方根，用σ表示。标准差是方差的算术平方根。标准差能反映一个数据集的离散程度。平均数相同的两组数据，标准差未必相同。

相关系数

相关系数是最早由统计学家卡尔·皮尔逊设计的统计指标，是研究变量之间线性相关程度的量，一般用字母r表示。由于研究对象的不同，相关系数有多种定义方式，较为常用的是皮尔森相关系数。

特征值

特征值是线性代数中的一个重要概念。在数学、物理学、化学、计算机等领域有着广泛的应用。设A是向量空间的一个线性变换，如果空间中某一非零向量通过A变换后所得到的向量和X仅差一个常数因子，即AX=kX，则称k为A的特征值，X称为A的属于特征值k的特征向量或特征矢量。
