**单变量分析**

变量值有效性检验

1. 每个变量为空的记录数
2. 每个变量不合法的记录数（比如：超过合理范围，不是给定的选项等）

场景：

1. 业务想要分析的属性，与实际数据质量之间的差距

数值变量 numeric/continuous variable，

1. histogram
2. density
3. boxplot：只有一个 box，如果增加一个 categorical 变量，则可以有多个 box

场景：

1. 采购金额，采购量分布；

Categorical 变量

1. 不同 levels 的计数Barchart

场景：

1. 不同选项的数量；每种类型对应的数量；每个城市对应的数量；
2. 每个供应商的采购订单数量；
3. 每颗料的采购订单数量；

**多变量分析**

两个 numerical

1. Scatter plot
2. 回归分析

场景

1. 两个变量是否存某种关系；

多个 numerical

1. Pairs 的相关系数，热度图；
2. 协方差矩阵

场景

1. 两两变量间的相关性

两个变量：Categorical + numerical

1. 相当于对 这一个 numerical 变量做了分组，也即，相当于多组 numerical；
2. 所有组的Boxplot；
3. 每组的均值是否相等 t.test，组与组之间是否有明显差异 ANAVO；
4. 每组中位数是否相等 wilcox.test；

场景：

1. 某一类是否比另一类更好；
2. 某个物资，供应商供应价格差异（供应商：A/B/C；单价）
3. 某项措施实施前后，对于某个指标是否有明显的效果（而不是随机导致）

两个变量：categorical + categorical

1. 其实还需要一个 numerical 变量，或者一个计数变量；
2. 不同 levels 下，做成 contingency table，判断两 categorical变量 是否独立；chi-square

场景：

1. 吸烟是否导致更容易得肺癌？（吸烟：Y/N；肺癌: Y/N）
2. 研发中，哪家供应商配合更容易成功? ( 供应商: A/B/C；研发成功: Y/N )

集中度分析：

1. 一颗物料有多少供应商；供应来源太多，或是太单一；
2. 一个供应商，供应多少物料；对供应商过于依赖，或供应商只是中间商；

波动分析：

1. 一颗料，不同时间，价格波动，与采购量波动关系；
2. 一颗料，不同供应商，价格波动，与采购量关系；供应商配额与采购价格管理；
3. 一个 numerical，在同一个 categorical 下不同 levels 之间的均值；

业务节拍/合作模式

1. 供应商送货频率，固定时间送货(每周二、五送货)，还是随时送货；供应商配送成本，仓储成本；寄售模式；

**PCA分析：数据降维**

1. 每行都是一个观测对象；
2. 每列都是一个维度；categorical 类型的，也用数字表示；
3. 每个维度做标准化（不同维度间可比性）；
4. 做 PCA 分解，获取主成分；
5. 观察 scree 图，一般情况下，前两个成分占比超过 70%；
6. 用前两个成分做轴，对原来的数据进行散点图；
7. 新旧维度之间的转换关系，雷达图；
8. PCA 使用 SVD，SVD 又是基于 eigenvector;

**回归分析**

线性回归

1. 单变量：numerical ~ numerical
2. 双变量：numerical ~ numerical + categorical
3. 多变量：numerical ~ 多个 numerical
4. 多变量：numerical ~ 多个 numerical + 多个categorical

每个 categorical，就相当于多条回归线(levels 的数量)；

总结

1. 线性回归，不一定都是直线，多项式也是线性回归；
2. 不一定只是一条线，categorical 的不同 levels 就对应不同的线；
3. 线性指的是形式： y = xt \* beta + alpha，其中，
   * x,beta,alpha 都是列向量；
   * x1, x2, xn 就是上面的变量，可以有很多个，也可以是不同的次方(1, x, x2, x3);
   * beta是模型参数（未知常数），但是不一定是一个数，而是一个列向量；
   * alpha 是随机误差；这里和简单的 ax + b 中的 b 不是一个概念，因为 b 对应的 x1 = 1 时的 beta1 的值；
   * alpha 在 GLM 中，会被抵消掉，结果就是：GLM 中 log Rate 之和 beta 和 x 有关，和 alpha 无关；
   * GLM 中不同之处只是左边的y，右边形式是不变的；
   * GLM 的 y 都是 log 形式，所以，两个 log 相减，右边 alpha 就抵消了，beta 作为公因数可以提取，剩下(x1 – x2)t \* beta；同时log 相减等于里面数值的相除，也即rate 的含义；

游戏分析

1. Metascore 和 userscore 的相关性如何，差异大的是哪些游戏；
2. 作品最多的10 家开发商，各自所有游戏的得分情况 boxplot， 最低得分的 10 部游戏是什么；
3. 按时间来看，评分有什么趋势没；
4. 游戏类型，年龄层次，有什么关系，肯定不相同，但是哪些组合最常见 (PCA中的维度散点，观测对象散点)；

GLM

逻辑/指数归回：

1. 不是最小二乘，而是极大似然；
2. 是分类
3. 两步映射：
   1. 把 0~1 映射到 0到正无穷；
   2. 映射后，1 是分界点，也即0~1 和 1~正无穷 在逻辑上是对称的，但是在数轴上却不是对称的，所以，再取 log，映射到 负无穷到正无穷，保证的对称性；
4. 转换函数
   1. P： 0~1
   2. Odds：0~正无穷
   3. Log Odds：负无穷 到 正无穷
5. 关注的是比例，比例对应条件概率：
6. 无后效性；

泊松回归

Cox 回归