描述统计分析
相关性分析

讲师：萨缪尔 Samuel

萨缪尔老师

网易数据分析教研负责人；

前盛大游戏战略规划总监、前腾讯游戏商业智能中心Leader

上海交通大学高金硕士 、《哈佛管理导师》外部导师

- **知乎大V：「萨缪尔」**主要聚集与商业分析、行业与战略分析、Python数据分析等
- **BAT**互联网巨头商业洞察分析 & 咨询公司战略咨询背景
- 擅长行业趋势研究和战略管理咨询工具，**为20多家上市公司提供战略发展决策建议**

课程亮点

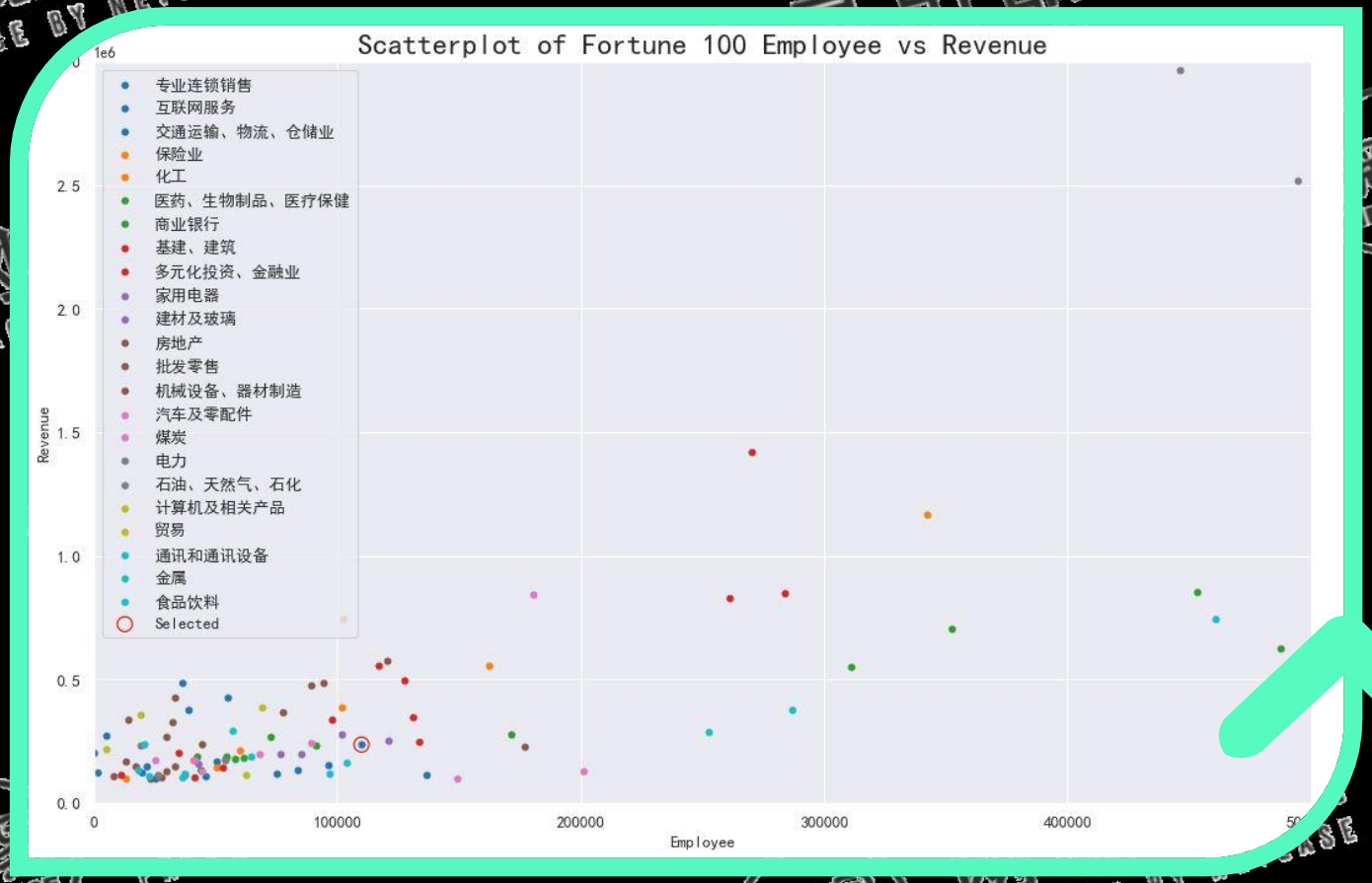
1 相关性的定义与类型

2 探索相关性

3 解释相关性

什么是相关性

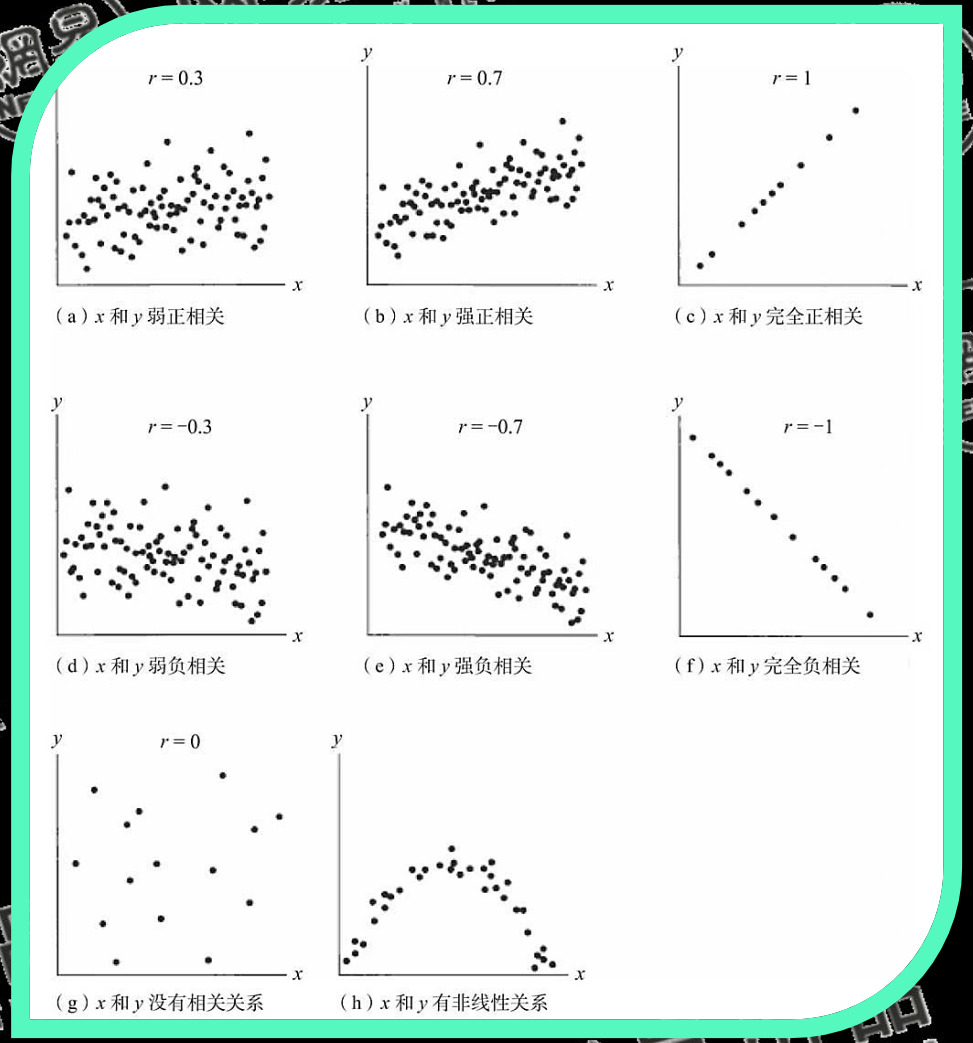
案例：世界500强企业营业收入跟员工规模之间的关系



相关性（Correlation，或称相关系数或关联系数）：显示两相关变量之间线性关系的强度和方向。

相关的类型

- **正相关**：两个变量同时增加（或减小）
- **负相关**：两个变量变化的趋势相反，一个变量增加而另一个变量减小。
- **不相关**：两个变量间没有明显的（线性）关系。
- **非线性关系**：两个变量有关联，但是以散点图呈现的相关关系不是直线形状。



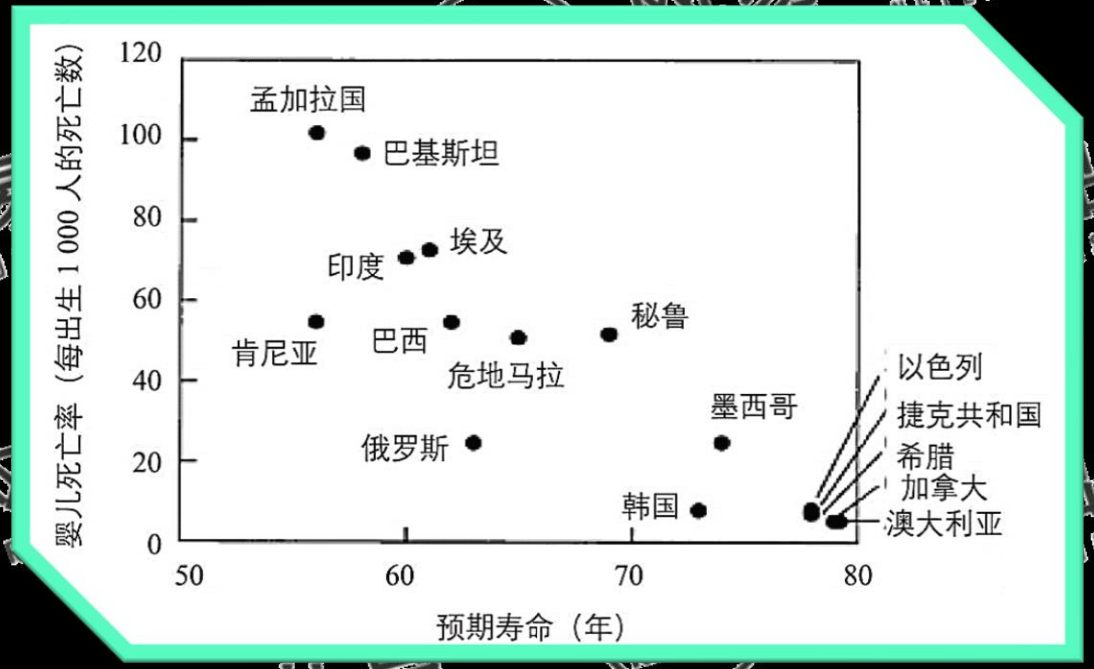
（a）x 和 y 弱正相关　（b）x 和 y 强正相关　（c）x 和 y 完全正相关

（d）x 和 y 弱负相关　（e）x 和 y 强负相关　（f）x 和 y 完全负相关

（g）x 和 y 没有相关关系　（h）x 和 y 有非线性关系

探索相关性

预期寿命与婴儿死亡率之间是什么关系?



预期寿命和婴儿死亡率数据的散点图

左图为16个国家的“预期寿命”和“婴儿死亡率”这两个变量间关系的散点图。

它属于哪种相关关系?

这种相关关系有意义吗?

其中有因果关系吗?

探索相关性：相关系数的计算

1 **Pearso相关系数：**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，最常用

$$\rho_{X,Y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

2 Spearman相关系数：又称秩相关系数，根据随机变量的等级而不是其原始值衡量相关性的一种方法。对原始变量的分布不作要求，适用范围更广些。不服从正态分布的变量、分类或等级变量之间的关联性可采用Spearman秩相关系数

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} (R_i - Q_i)^2}{n(n^2 - 1)}$$

对两个变量成对的取值分别按照从小到大（或者从大到小）顺序编秩，R_i 代表 x_i 的秩次，Q_i 代表 y_i 的秩次，$R_i - Q_i$ 为 x_i、y_i 的秩次之差，n为样本量。

3 其他：kendall相关系数、卡方检验、Fisher检验等

探索相关性：相关系数的计算

指标类型		相关性算法	应用示例
连续型指标	连续型指标	Pearson	商品曝光量和购买转化率
有序离散型指标	有序离散型指标	Spearman Kendall	用户等级和活跃程度
有序离散型指标	无序离散型指标	卡方检验	满意度和手机品牌
无序离散型指标	无序离散型指标	卡方检验 Fisher检验	手机品牌和年龄段
二分型指标	连续型指标	Point-biserial	性别和阅读率
二分型指标	有序离散型指标	Biserial	性别和商品评分
有序离散型指标	连续型指标	无直接算法，但可将连续型指标离散化后进行处理	商品评分和购买转化率

解释相关性

1. 发现异常值，并处理
2. 注意不恰当的分组
3. 相关并不蕴含因果关系
4. 寻找业务跟相关性相关的解释

课程总结

1. 相关系数的定义:显示两相关变量之间线性关系的强度和方向。
2. 相关的类型有正相关、负相关、不相关和非线性关系
3. 相关系数计算的方法： Pearso相关系数、 Spearman相关系数、kendall相关系数、卡方检验、Fisher检验等

网易云课堂 | 涨薪计划

谢谢观看