

数据分析

2021年3月29日 13:28

Excel

数据透视表, vlookup, 常用函数的使用, 基础图表的制作

SQL

<https://zhuanlan.zhihu.com/p/38354000>

1. 简单查询

- %表示任意字符串
- 利用count()统计数量

2. 汇总分析

- 利用sum()计算总和
- DISTINCT 用于返回唯一不同的值。
- 分组group by
- 对分组结果指定条件 having
- 排序order by asc升序 desc降序

3. 复杂查询

- 子查询
- limit 从查询结果中取出指定行, 例如前n个, top(n), 排序后limit
- in 和 not in

4. 多表查询

联结

select * from kemu left join score on kemu.id = score.id

- left join 就是“左连接”, 表1左连接表2, 以左为主, 表示以表1为主, 关联上表2的数据, 查出来的结果显示左边的所有数据, 然后右边显示的是和左边有交集部分的数据。
- right join 同理
- join, 其实就是“inner join”, 为了简写才写成join, 两个是表示一个的, 内连接, 表示以两个表的交集为主, 查出来是两个表有交集的部分, 其余没有关联就不额外显示出来

case表达式

CASE WHEN <表达式> THEN <表达式>
ELSE <表达式>

END

5. 窗口函数

<窗口函数> over (partition by <用于分组的列名>
order by <用于排序的列名>)

partition by用来对表分组

order by子句的功能是对分组后的结果进行排序

group by分组汇总后改变了表的行数，一行只有一个类别。而partition by和rank函数不会减少原表中的行数

<窗口函数>的位置，可以放以下两种函数：

1) 专用窗口函数，rank (1114) , dense_rank (1112) , row_number (1234)
等专用窗口函数。

2) 聚合函数，如sum, avg, count, max, min等
窗口函数原则上只能写在select子句中

窗口函数解决3类问题：

1) 分组排名 2) topN问题 3) 每个组内比较

练习题：<https://zhuanlan.zhihu.com/p/152233908>

Python

- python基本语法
- 数据分析包 numpy pandas matplotlib
- 使用python操作结构化数据，进行数据清洗，数据抽取，数据可视化
- python操作数据库

Hive、spark

统计概率

- 描述性统计（平均值，标准差，中位数）
- 概率（独立事件，相关事件，期望，贝叶斯）
- 概率分布（离散概率分布，连续概率分布）
- 统计推断（抽样，执行区间，假设检验）

机器学习

- 分类算法：逻辑回归、贝叶斯、决策树、随机森林

- 回归算法：线性回归
- 聚类算法：K-means

- 特征工程
- 模型评价
- 交叉检验

业务问题

业务指标：漏斗思维、分类思维、平衡思维、A/Btest，金字塔原理

相关性和因果关系的区别

DAU下降（日活跃用户数量） https://www.sohu.com/a/449047677_114819