数据分析

2021年3月29日 13:28

Excel （略）

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partiition 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（概率论与数理统计）

- 见计算机笔记中的《概率论》和《数理统计》

机器学习

分类算法：见计算机笔记中的《分类算法》

- 回归算法：线性回归
  【算法】七种常用的回归算法
  https://cloud.tencent.com/developer/article/1102103
- 聚类算法：K-means

- **特征工程**
  深度了解特征工程 https://zhuanlan.zhihu.com/p/111296130
- **模型评价**
  模型的评价方法 https://zhuanlan.zhihu.com/p/53774460
- **交叉检验**
  Cross-Validation（交叉验证）详解 https://zhuanlan.zhihu.com/p/24825503

业务问题

- 数据分析思维
  数据驱动决策的13种思维方式 https://www.jianshu.com/p/5a8f01fe7f2a
  数据分析师必备的20种分析思维 https://zhuanlan.zhihu.com/p/83138160
  业务指标：漏斗思维、分类思维、平衡思维、A/Btest，金字塔原理

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

- **漏斗思维**
  漏斗思维 https://www.jianshu.com/p/32414ca5895f
  五个经典漏斗模型，看漏斗思维穿透流程化的本质
  http://www.woshipm.com/operate/3269415.html

- **A/B test**
  (分割测试或桶测试）是一种将网页或应用程序的两个版本相互比较以确定哪个版本的

性能更好的方法 https://cloud.tencent.com/developer/article/1496302
分离式组间试验，也叫对照试验，科研领域（药物测试）中已广泛应用
https://blog.csdn.net/m0_37773338/article/details/108561193

- **金字塔原理**
  金字塔原理（简析） https://zhuanlan.zhihu.com/p/44423303
  什么是金字塔原理，如何运用金字塔原理？ https://www.sohu.com/a/402879711_404038

- **相关性和因果关系的区别**