文本描述

> 全部内容来自于2023年数据科学导论课程proj的数据集描述以及相关网站
>
> 内容均为**机翻加上部分个人理解**，个人并未仔细研究每一个数据集，因此可能存在**错误判断**
>
> **全文仅供参考，不代表任何官方意见，如果因为使用了出现纰漏，本人概不负责**
>
> 欢迎与我联系交流，QQ：1138352969
>
> 整理by：郑茗献

1. air quality

数据集信息：

> 本数据集包括 12 个国控空气质量监测点的每小时空气污染物数据。空气质量数据来自北京市环境监测中心。每个空气质量监测点的气象数据均与中国气象局最近的气象站相匹配。时间段为 2013 年 3 月 1 日至 2017 年 2 月 28 日。缺失数据以 NA 表示。

属性信息：

- No：行号
- 年份：该行数据的年份
- 月：本行数据的月份
- 日：本行数据的日
- hour：本行数据的小时数
- PM2.5：PM2.5 浓度（微克/立方米）
- PM10： PM10 浓度（微克/米^3）
- SO2： SO2： 二氧化硫浓度（微克/米^3）
- NO2： NO2：二氧化氮浓度（微克/米^3）
- CO： CO 浓度（微克/米^3）
- O3: O3：O3 浓度（微克/立方米）
- TEMP：温度（摄氏度）
- PRES：气压（hPa）
- DEWP：露点温度（摄氏度）
- RAIN：降水量（毫米）
- WD：风向
- WSPM：风速（米/秒）
- station：空气质量监测点名称

2.airport airline

airport

- 机场 ID 该机场的 OpenFlights 唯一标识符。
- 机场名称。可能包含也可能不包含城市名称。
- 城市 机场服务的主要城市。拼写可能与名称不同。
- 国家（Country） 机场所在的国家或地区。请参阅 "国家 "与 ISO 3166-1 代码对照。
- IATA 3 个字母的 IATA 代码。如果未指定/未知，则为空。
- ICAO 4 个字母的 ICAO 代码。
- 未指定时为空。
- 纬度 十进制度数，通常为六位有效数字。负数表示南纬，正数表示北纬。
- 经度 十进制度数，通常为六位有效数字。负数表示西，正数表示东。
- 高度 以英尺为单位。
- 时区 从UTC偏移的小时数。小数小时用小数表示，如印度为 5.5。
- DST 夏令时。E（欧洲）、A（美国/加拿大）、S（南美洲）、O（澳大利亚）、Z（新西兰）、N（无）或 U（未知）之一。另请参阅： 帮助： 时间
- Tz 数据库时区 以 "tz"（奥尔森）格式表示的时区，例如 "America/Los_Angeles"。
- 类型 机场类型。空港为 "airport"，火车站为 "station"，渡口为 "port"，未知为 "unknown"。在 airports.csv 中，只包括 type=airport。
- 数据来源 "OurAirports" 表示数据来源于 OurAirports，"Legacy" 表示与 OurAirports 不匹配的旧数据（主要是 DAFIF），"User" 表示未经验证的用户贡献。在 airports.csv 中，只包含 source=OurAirports。

注：夏令时规则每年和每个国家都有变化。当前数据为 2009 年的近似值，以国家为单位。在通常遵守夏令时的国家（如美国的 AL、HI，澳大利亚的 NT、QL，加拿大的部分地区）中，无夏令时地区的大多数机场标记不正确。

airline

- 航空公司 ID 该航空公司在 OpenFlights 中的唯一标识符。
- 名称 航空公司名称。
- 别名 航空公司的别名。例如，全日空航空公司通常被称为 "ANA"。
- IATA 2 个字母的 IATA 代码（如有）。
- ICAO 3 个字母的 ICAO 代码（如有）。
- 呼号 航空公司呼号。
- 国家（Country） 机场所在的国家或地区。请参阅 "国家 "与 ISO 3166-1 代码对照。
- 如果航空公司正在运营或直到最近一直在运营，则为 "Y"；如果已停业，则为 "N"。这个字段并不可靠：特别是那些早已停飞但尚未重新分配其 IATA 代码的主要航空公司（如安塞特/南航），会错误地显示为 "Y"。
- 数据采用 UTF-8 编码。NULL "使用特殊值 \N，表示没有可用值，如果导入，MySQL 会自动理解。

注：代码/呼号/国家为空的航空公司通常代表用户添加的航空公司。由于数据主要用于当前航班，因此一般不包括失效的 IATA 代码。例如，"Sabena "没有列出 SN IATA 代码，因为 "SN "目前由其后继者布鲁塞尔航空公司使用。

route

- 航空公司的 2 个字母（IATA）或 3 个字母（ICAO）的代码。
- 航空公司 ID 航空公司的唯一 OpenFlights 标识符（请参阅航空公司）。
- 来源机场 来源机场的 3 个字母 (IATA) 或 4 个字母 (ICAO) 代码。
- 来源机场 ID 来源机场的唯一 OpenFlights 标识符（请参阅 "机场"）。
- 目的地机场 目的地机场的 3 个字母 (IATA) 或 4 个字母 (ICAO) 的代码。
- 目的地机场 ID 目的地机场的唯一 OpenFlights 标识符（请参阅 "机场"）。

- 代码共享 如果此航班为代码共享航班（即不是由航空公司运营，而是由其他航空公司运营），则为 "Y"，否则为空。
- 停靠站点 此航班的停靠站点数量（"0 "表示直达）。
- 设备 此航班通常使用的飞机类型的 3 个字母代码，用空格分隔
- 数据采用 UTF-8 编码。NULL "使用特殊值 \N，表示没有可用值，如果导入，MySQL 会自动理解。

注意

- 航线是定向的：如果一家航空公司运营从 A 到 B 和从 B 到 A 的航线，则 A-B 和 B-A 均单独列出。
- 一家航空公司同时运营自己的航班和代码共享航班的航线只列出一次

plane

- 名称 飞机全名。
- IATA 代码 飞机唯一的三个字母的 IATA 识别符。
- ICAO 代码 飞机的唯一四字母 ICAO 识别符。
- 数据采用 UTF-8 编码。NULL "使用特殊值 \N，表示没有可用值，如果导入，MySQL 会自动理解。

备注：
有 IATA 但没有 ICAO 代码的飞机通常是飞机类别：例如，IATA "747 "可以是任何类型的波音 747，而 IATA "744"/ICAO "B744 "则专门指波音 747-400。

country

- name 国家或地区的全称。
- iso_code 该国家或地区唯一的双字母 ISO 3166-1 代码。
- dafif_code DAFIF 中使用的 FIPS 国家代码。
- 数据采用 UTF-8 编码。NULL "使用特殊值 \N 表示没有可用值，如果导入，MySQL 会自动理解。

注释
有些条目有 DAFIF 代码，但没有 ISO 代码。这些主要是没有机场的无人居住的岛屿，在大多数情况下可以忽略。

3.artists_directors

artist

- 每一行都是一个人的作品列表、
- 职业生涯中的作品用"|"分割
- 每件作品的第一个元素是落锤价，第二个元素是制作年份

director

- 每行是个人作品列表、
- 职业生涯中的作品用"|"分割
- 每部作品的第一个元素是 IMDB 评分，第二个元素是发行年份

4.bilibili

> 数据包含 Bilibili UP 主（UP主）的信息。包括ID、粉丝数量、性别、发布视频数量、代表作品数量、专辑数量、文章数量、频道数量、最近视频的平均长度、平均播放量、关联的第三方平台、自身标签等。

数据说明

变量类别		变量名	变量说明	取值范围
因变量		粉丝数	连续变量	13 ~ 6939197
自变量	个人信息	代表作个数	单位：个	0 ~ 4
		个人标签	文字	/
		是否关联新浪微博	定性变量（2水平）	是：1，否：0
		是否关联微信	定性变量（2水平）	是：1，否：0
		是否关联QQ	定性变量（2水平）	是：1，否：0
		是否关联淘宝	定性变量（2水平）	是：1，否：0
		是否关联邮箱	定性变量（2水平）	是：1，否：0
		是否关联小红书	定性变量（2水平）	是：1，否：0
		是否关联抖音	定性变量（2水平）	是：1，否：0
	主页信息	发布文章数	单位：个	0 ~ 2196
		发布相册数	单位：个	0 ~ 4128
		发布视频数	单位：个	51 ~ 24294
		发布频道数	单位：个	0 ~ 11
		主要发布的视频品类	定性变量（11水平）	动画、娱乐、影视、时尚、游戏、生活、科技、舞蹈、音乐、鬼畜、其他
		主品类视频占所有视频比例	连续变量	0 ~ 1
	近期视频信息	近期视频平均播放量	单位：次	3 ~ 6711586
		近期视频平均时长	单位：秒	12 ~ 146084.2

5.COVID-19

Confirmed cases

- total_cases COVID-19 确诊病例总数。如有报告，计数可包括疑似病例。
- new_cases COVID-19 新确诊病例。计数可包括报告的疑似病例。在极少数情况下，如果我们的数据源因数据更正而报告了负的每日变化，我们会将此指标设为 NA。
- new_cases_smoothed COVID-19 新确诊病例（7 天平滑）。如有报告，计数可包括疑似病例。
- total_cases_per_million 每 100 万人中 COVID-19 确诊病例总数。计数可包括报告的疑似病例。
- 每 100 万人中 COVID-19 新确诊病例数。计数可包括报告的疑似病例。
- new_cases_smoothed_per_million 每 100 万人中 COVID-19 的新确诊病例（7 天平滑）。计数可包括报告的疑似病例。

Confirmed deaths

- total_deaths COVID-19 导致的死亡总数。如有报告，计数可包括可能的死亡。
- new_deaths 归因于 COVID-19 的新死亡人数。在有报告的情况下，计数可包括可能的死亡。在极少数情况下，如果我们的数据源因数据校正而报告了负的每日变化，我们会将此指标设为 NA。
- new_deaths_smoothed COVID-19 的新死亡人数（7 天平滑）。如有报告，计数可包括可能死亡人数。
- total_deaths_per_million 每 100 万人中 COVID-19 导致的死亡总人数。计数可包括报告的可能死亡人数。
- new_deaths_per_million 每 100 万人中 COVID-19 导致的新死亡人数。计数可包括报告的可能死亡人数。
- new_deaths_smoothed_per_million 每 100 万人中 COVID-19 的新增死亡人数（7 天平滑）。计数可包括报告的可能死亡人数。

注：
由于在死因归因方面存在不同的规程和挑战，确诊死亡人数可能无法准确代表 COVID-19 导致的真实死亡人数。

Excess mortality

- excess_mortality 2020年至2021年每周或每月死亡人数与基于前几年的预测死亡人数之间的百分比差异。更多信息，请参阅 https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality
- excess_mortality_cumulative 自2020年1月1日以来死亡人数的累计差异与基于前几年的预测死亡人数之间的百分比差异。更多信息，请参阅 https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality
- excess_mortality_cumulative_absolute 自2020年1月1日以来死亡人数的累计差异与基于前几年的预测死亡人数之间的累计百分比差异。更多信息，请参阅 https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality
- excess_mortality_cumulative_per_million 自2020年1月1日以来死亡人数的累计差异与基于前几年的预测死亡人数之间的累计百分比差异，以每百万人口为单位。更多信息，请参阅 https://github.com/owid/covid-19-data/tree/master/public/data/excess_mortality

Hospital & ICU

- `icu_patients`：给定日期内重症监护病房（ICU）中的COVID-19患者数量
- `icu_patients_per_million`：给定日期内每100万人口中的重症监护病房（ICU）中的COVID-19患者数量
- `hosp_patients`：给定日期内医院中的COVID-19患者数量
- `hosp_patients_per_million`：给定日期内每100万人口中的医院中的COVID-19患者数量

- `weekly_icu_admissions`: 在给定周内（报告日期及前6天），新入住重症监护病房（ICU）的COVID-19患者数量
- `weekly_icu_admissions_per_million`: 在给定周内（报告日期及前6天），每100万人口中新入住重症监护病房（ICU）的COVID-19患者数量
- `weekly_hosp_admissions`: 在给定周内（报告日期及前6天），新入住医院的COVID-19患者数量
- `weekly_hosp_admissions_per_million`: 在给定周内（报告日期及前6天），每100万人口中新入住医院的COVID-19患者数量

Policy responses

- `stringency_index`: 这是一个综合指标，基于9个响应指标，包括学校关闭、工作场所关闭和旅行禁令等。该指数重新缩放为从0到100的值，其中100表示最严格的响应。

Reproduction rate

- `reproduction_rate`: COVID-19实时有效再生数（R值）的估算。详见 https://github.com/crondonm/TrackingR/tree/main/Estimates-Database。

Tests & positivity

- `total_tests`: COVID-19的总检测数
- `new_tests`: COVID-19的新检测数（仅计算连续天数）
- `total_tests_per_thousand`: COVID-19的每千人口总检测数
- `new_tests_per_thousand`: COVID-19的每千人口新检测数
- `new_tests_smoothed`: COVID-19的新检测数（7天平滑）。对于那些不以每日基础报告测试数据的国家，我们假设在没有报告数据的任何期间，测试以每日基础均等变化。这产生了一个完整的每日数据系列，然后在一个滚动的7天窗口上进行平均。
- `new_tests_smoothed_per_thousand`: COVID-19的新检测数（7天平滑）每千人口
- `positive_rate`: COVID-19检测中呈阳性的比例，以滚动的7天平均值表示（这是tests_per_case的倒数）
- `tests_per_case`: COVID-19每个新确诊病例的检测数，以滚动的7天平均值表示（这是positive_rate的倒数）
- `tests_units`: 地点用于报告其测试数据的单位。一个国家文件不能包含混合单位。有关测试数据的所有度量标准都使用指定的测试单位。有效单位包括'people tested'（测试人数）、'tests performed'（执行的测试次数，一个人在一天内可以进行多次测试）和'samples tested'（测试的样本数。在某些情况下，执行给定测试可能需要多个样本）。

Vaccinations

- `total_vaccinations`: COVID-19疫苗接种的总剂次数
- `people_vaccinated`: 收到至少一剂疫苗的人数总计
- `people_fully_vaccinated`: 接受了初始疫苗接种方案规定的所有剂次的人数总计
- `total_boosters`: COVID-19疫苗加强剂次的总数（超出疫苗接种方案规定的剂次数）
- `new_vaccinations`: 新的COVID-19疫苗接种剂次（仅计算连续的天数）
- `new_vaccinations_smoothed`: 新的COVID-19疫苗接种剂次（7天平滑）。对于那些不以每日基础报告疫苗接种数据的国家，我们假设在没有报告数据的任何期间，疫苗接种在每日基础上均等变化。这产生了一个完整的每日数据系列，然后在一个滚动的7天窗口上进行平均。
- `total_vaccinations_per_hundred`: 每100人口中的总COVID-19疫苗接种剂次数
- `people_vaccinated_per_hundred`: 每100人口中至少接种一剂疫苗的人数总计
- `people_fully_vaccinated_per_hundred`: 每100人口中接受了初始疫苗接种方案规定的所有剂次的人数总计
- `total_boosters_per_hundred`: 每100人口中的总COVID-19疫苗加强剂次数

- `new_vaccinations_smoothed_per_million`: 每100万人口中的新COVID-19疫苗接种剂次（7天平滑）
- `new_people_vaccinated_smoothed`: 每日接受第一剂疫苗接种的人数（7天平滑）
- `new_people_vaccinated_smoothed_per_hundred`: 每100人口中每日接受第一剂疫苗接种的人数（7天平滑）

Others

- `iso_code`: ISO 3166-1 alpha-3 – 三位字母的国家代码。请注意，OWID定义的地区（例如，'Europe'这样的大洲）包含前缀'OWID_'。
- `continent`: 地理位置所在的大洲。
- `location`: 地理位置。
- `date`: 观察日期。
- `population`: 人口（最新可用值）。详见 https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv 获取完整的数据源列表。
- `population_density`: 人口与土地面积的比值，以平方千米为单位，最近一年可用。
- `median_age`: 人口的中位年龄，根据2020年联合国预测。
- `aged_65_older`: 65岁及以上人口占比，最近一年可用。
- `aged_70_older`: 70岁及以上人口占比，截至2015年。
- `gdp_per_capita`: 按购买力平价计算的国内生产总值（恒定于2011国际美元），最近一年可用。
- `extreme_poverty`: 生活在极端贫困中的人口占比，自2010年以来的最近一年可用。
- `cardiovasc_death_rate`: 2017年心血管疾病的死亡率（每10万人口的年度死亡人数）。
- `diabetes_prevalence`: 2017年糖尿病患病率（20至79岁人口的百分比）。
- `female_smokers`: 抽烟女性的比例，最近一年可用。
- `male_smokers`: 抽烟男性的比例，最近一年可用。
- `handwashing_facilities`: 具备基本洗手设施的人口占比，最近一年可用。
- `hospital_beds_per_thousand`: 每千人口的医院床位数，自2010年以来的最近一年可用。
- `life_expectancy`: 2019年出生时的预期寿命。
- `human_development_index`: 人类发展指数（HDI），衡量人类发展的三个基本维度的平均成就：健康寿命、知识水平和体面生活水平。数值为2019年，从 http://hdr.undp.org/en/indicators/137506 导入。

6.diabetes

- `Diabetes_012`: 糖尿病相关信息
- `HighBP`: 高血压
- `HighChol`: 高胆固醇
- `CholCheck`: 胆固醇检查
- `BMI`: 体质指数
- `Smoker`: 是否吸烟
- `Stroke`: 中风
- `HeartDiseaseorAttack`: 心脏病或心脏病发作
- `PhysActivity`: 身体活动水平
- `Fruits`: 水果消费
- `Veggies`: 蔬菜消费
- `HvyAlcoholConsump`: 是否有大量饮酒
- `AnyHealthcare`: 是否接受任何医疗保健
- `NoDocbcCost`: 是否因为费用原因而没有医生
- `GenHlth`: 一般健康状况
- `MentHlth`: 心理健康状况
- `PhysHlth`: 身体健康状况
- `DiffWalk`: 行走困难程度
- `Sex`: 性别
- `Age`: 年龄
- `Education`: 受教育程度
- `Income`: 收入

7. hongloumeng

详见数据集

8.houseprice_BJ

- CATE 区
- bedrooms 卧室数
- halls 厅数量
- AREA 房屋面积
- floor 楼层位置（相对）
- subway 是否有地铁站
- school 是否有学校
- price 每平方米价格，单位：$万元/\mathrm{m}^2$(从网上查询北京房价对比数据大概得知)

9. houseprice_GD

> 广州、深圳和厦门的二手房价格

- `danjia`: 单价，单位面积单价，单位：$元/\mathrm{m}^2$
- `area`: 面积
- `floor`: 楼层位置
- `hall`: 厅数量
- `room`: 房间数
- `school`: 是否有学校
- `chaoxiang`: 朝向

- `year`: 年份
- `subway`: 是否有地铁
- `district`: 区
- `city`: 城市

10. input_output

> 《世界投入产出数据库》（WIOD）2016年11月发布版包括一系列数据库，涵盖了28个欧盟国家和世界上其他15个主要国家，时间跨度为2000年至2014年。

> 在投入产出表中，表的每一行和每一列代表一个产业或经济部门。表中某一行和某一列的交叉点（单元格）的数字表示这两个部门之间的交互关系，具体来说是一个部门对另一个部门的产出、投入或需求。
>
> - **行（r）**： 代表产出部门或行业。
> - **列（c）**： 代表投入部门或行业。
>
> 具体而言，表中某一行（r）的第 i 个元素表示产业 r 对自身的投入或需求，而该行的其他元素表示产业 r 对其他各行业的投入或需求。同样，表中某一列（c）的第 i 个元素表示行业 c 对各行业的投入或需求，而该列的其他元素表示行业 c 对自身的投入或需求。这些数字通常可以解释为货物和服务的交换，代表经济中不同产业之间的相互依赖关系。这种关系对于分析整个经济系统的运作和评估政策的影响非常有用。

11. Job

> 数据科学家的职位信息

- index：索引
- Job Title：职位标题
- Salary Estimate：薪资估算
- Job Description：工作描述
- Rating：评分
- Company Name：公司名称
- Location：位置
- Headquarters：总部
- Size：规模
- Founded：成立时间
- Type of ownership：所有权类型
- Industry：行业
- Sector：部门
- Revenue：收入
- Competitors：竞争对手
- Easy Apply：易申请与否

12. linkedIn_job_posting

- `Co_Nm`（公司名称）：数据类型为对象（Object）
- `Co_Pg_Lstd`（公司页面列出）：数据类型为布尔（Bool）
- `Emp_Cnt`（公司雇员人数）：数据类型为整数（int64）
- `Flw_Cnt`（公司关注者人数）：数据类型为整数（int64）
- `Job_Ttl`（职位标题）：数据类型为对象（Object）
- `Job_Desc`（职位描述）：数据类型为对象（Object）
- `Is_Supvsr`（是否是主管职位）：数据类型为布尔（Bool）
- `max_sal`（最高工资）：数据类型为浮点数（Float64）
- `med_sal`（中位数工资）：数据类型为浮点数（Float64）
- `min_sal`（最低工资）：数据类型为浮点数（Float64）
- `py_prd`（支付周期）：数据类型为分类（Category），包括 {Not Listed, YEARLY, HOURLY, MONTHLY, Unpaid, WEEKLY, ONCE}
- `py_lstd`（是否列出支付）：数据类型为布尔（Bool）
- `wrk_typ`（工作类型）：数据类型为分类（Category），包括 {Full-time, Contract, Part-time, Temporary, Internship, Other, Volunteer}
- `loc`（工作地点）：数据类型为对象（Object）
- `st_code`（工作州代码）：数据类型为对象（Object）
- `is_remote`（是否远程工作）：数据类型为布尔（Bool）
- `views`（发布浏览次数）：数据类型为整数（int64）
- `app_typ`（申请类型）：数据类型为分类（Category），包括 {Offsite Apply, SimpleOnSiteApply, ComplexOnSiteApply}
- `app_is_off`（是否是异地申请）：数据类型为布尔（Bool）
- `xp_lvl`（经验水平）：数据类型为分类（Category），包括 {Mid-Senior level, Not Listed, Entry level, Associate, Director, Internship, Executive}
- `domain`（发布域）：数据类型为对象（Object）
- `has_post_domain`（是否有发布域）：数据类型为布尔（Bool）
- `is_sponsored`（是否赞助）：数据类型为布尔（Bool）
- `base_comp`（是否有基本薪酬）：数据类型为布尔（Bool）

13. mathematician

> 一个包含来自维基百科的8500多位著名数学家信息的数据库。

- **mathematicians** 数学家
- **occupation** 职业
- **country of citizenship** 国籍
- **place of birth** 出生地
- **date of death** 逝世日期
- **educated at** 教育背景
- **employer** 雇主
- **place of death** 逝世地点
- **member of** 成员身份
- **employer** 雇主（列表中出现两次）
- **doctoral advisor** 博士导师
- **languages spoken, written or signed** 使用的语言
- **academic degree** 学位
- **doctoral student** 博士生
- **manner of death** 逝世方式
- **position held** 担任职务
- **field of work** 工作领域
- **award received** 获奖情况
- **Erdős number** Erdős 数
- **instance of** 实例，**Q5**在wiki标签中代表人类实体，**eunuch**表示为阉人
- **sex or gender** 性别
- **approx. date of birth** 是否为大致出生日期
- **day of birth** 出生日
- **month of birth** 出生月
- **year of birth** 出生年
- **approx. date of death** 是否为大致逝世日期
- **day of death** 逝世日
- **month of death** 逝世月
- **year of death** 逝世年

14. media_spreading

豆瓣

doubancom.xlsx

> 豆瓣平台短评数据

- **ID:** 评论发布者的用户昵称
- **Short:** 评论的文本内容
- **Votes:** 该短评的“点赞”数
- **Time:** 评论发布时间
- **Emotion:** 该评论的情感得分

doubanrp.xlsx

> 豆瓣平台用户被关注数据

- **ID:** 评论发布者的用户昵称
- **Time:** 评论发布时间
- **Content:** 评论的文本内容
- **Agreenum:** 该短评的“点赞”数
- **Follower:** 该评论发布者的被关注数

微博

wbcom.xlsx

> 微博平台博文数据

- **ID:** 博文发布者的用户昵称
- **Content:** 博文的文本内容
- **Time:** 博文发布时间
- **Agree:** 该博文的“点赞”数
- **Emotion:** 该博文的情感得分

wbrp.xlsx

> 微博平台用户被关注数据

- **ID:** 博文发布者的用户昵称
- **Agreenum:** 该博文的“点赞”数
- **Follower:** 该博文发布者的被关注数

知乎

zhihucom.xlsx

> 知乎平台回答数据

- **ID:** 回答者的用户昵称
- **Tag:** 回答者的个人标签
- **Agree:** 该回答的“点赞”数
- **Time:** 回答发布时间
- **Comment:** 回答的文本内容
- **Emotion:** 该回答的情感得分

zhrp.xlsx

> 知乎平台用户被关注数据

- **Question:** 问题内容
- **Agreenum:** 该回答的“点赞”数
- **Content:** 回答的文本内容
- **ID:** 回答者的用户昵称
- **Follower:** 该回答发布者的被关注数

15.netflix_movie

> Netflix 电影评分数据。

movie

- **电影ID (Movie_ID)：** 电影的唯一标识符
- **年份 (Year)：** 电影发布的年份
- **名称 (Name)：** 电影的名称或标题

rate

- **用户ID (User_ID):** 用户的唯一标识符
- **评分 (Rating):** 用户对电影的评分
- **电影ID (Movie_ID):** 电影的唯一标识符

16.Nobel

> 自 1901 年以来的诺贝尔奖信息。

- **awardYear** 获奖年份
- **category** 奖项类别
- **categoryFullName** 奖项全名
- **sortOrder** 排序顺序
- **portion** 奖项的一部分
- **prizeAmount** 奖金金额
- **prizeAmountAdjusted** 调整后的奖金金额
- **dateAwarded** 颁奖日期
- **prizeStatus** 奖项状态
- **motivation** 获奖动机
- **categoryTopMotivation** 奖项顶层动机
- **award_link** 奖项链接
- **id** 奖项ID
- **name** 姓名
- **knownName** 已知姓名
- **givenName** 名字
- **familyName** 姓氏
- **fullName** 全名
- **penName** 笔名
- **gender** 性别
- **laureate_link** 获奖者链接
- **birth_date** 出生日期
- **birth_city** 出生城市
- **birth_cityNow** 出生城市当前状态
- **birth_continent** 出生大陆
- **birth_country** 出生国家
- **birth_countryNow** 出生国家当前状态
- **birth_locationString** 出生地点字符串
- **death_date** 逝世日期
- **death_city** 逝世城市
- **death_cityNow** 逝世城市当前状态
- **death_continent** 逝世大陆
- **death_country** 逝世国家
- **death_countryNow** 逝世国家当前状态
- **death_locationString** 逝世地点字符串
- **orgName** 组织名称
- **nativeName** 本地名称
- **acronym** 缩写
- **org_founded_date** 组织成立日期
- **org_founded_city** 组织成立城市
- **org_founded_cityNow** 组织成立城市当前状态
- **org_founded_continent** 组织成立大陆
- **org_founded_country** 组织成立国家
- **org_founded_countryNow** 组织成立国家当前状态
- **org_founded_locationString** 组织成立地点字符串
- **ind_or_org** 个人或组织
- **residence_1** 居住地1
- **residence_2** 居住地2

- **affiliation_1** 从属机构1
- **affiliation_2** 从属机构2
- **affiliation_3** 从属机构3
- **affiliation_4** 从属机构4

17.shared_room

> 2019年10月从蛋壳公寓解析的合租租金数据

以下是关于房屋的一些属性的中文翻译，并按照无序列表的格式输出：

- **rent** 月租金
- **bedroom** 卧室数
- **livingroom** 厅数
- **bathroom** 卫生间数
- **area** 租赁面积
- **room** 租赁房间
- **floor_grp** 所在楼层分组
- **subway** 是否邻近地铁
- **region** 城区
- **heating** 供暖方式

18. Stack Overflow 2023 survey dataset

> 2023 年 5 月，9 万多名开发人员参与了 Stack Overflow 年度调查，了解他们如何学习和提高水平、正在使用哪些工具以及想要哪些工具。
>
> 调查时间为 2023 年 5 月 8 日至 2023 年 5 月 19 日。合格受访者的调查时间中位数为 17 分钟。
>
> 受访者主要通过 Stack Overflow 拥有的渠道招募。前五大受访者来源分别是现场消息、博客文章、电子邮件列表、meta.stackoverflow 文章、横幅广告和社交媒体文章。由于受访者是通过这种方式招募的，Stack Overflow 上的高参与度用户更有可能注意到调查链接并点击开始调查。

schema

> 包含 78 行和 6 列，涵盖了调查的基本模式。

- **qid** 问题ID
- **qname** 问题名称
- **question** 问题描述
- **force_resp** 是否强制回答
- **type** 问题类型
- **selector** 选择器

public

> 包含 89184 行和 84 列，涉及与用户和他们想要探索的工具相关的不同问题。
>
> 具体问题参考文件夹中pdf问卷

- **ResponseId** 响应标识
- **Q120** Q120（问题编号）
- **MainBranch** 主要领域
- **Age** 年龄
- **Employment** 就业状况
- **RemoteWork** 远程工作
- **CodingActivities** 编码活动
- **EdLevel** 教育水平
- **LearnCode** 学习编码
- **LearnCodeOnline** 在线学习编码
- **LearnCodeCoursesCert** 通过课程或认证学习编码
- **YearsCode** 编码年数
- **YearsCodePro** 专业编码年数
- **DevType** 开发者类型
- **OrgSize** 组织规模
- **PurchaseInfluence** 购买影响力
- **TechList** 技术列表
- **BuyNewTool** 购买新工具
- **Country** 国家
- **Currency** 货币
- **CompTotal** 总薪资
- **LanguageHaveWorkedWith** 曾使用的编程语言
- **LanguageWantToWorkWith** 想要使用的编程语言
- **DatabaseHaveWorkedWith** 曾使用的数据库
- **DatabaseWantToWorkWith** 想要使用的数据库
- **PlatformHaveWorkedWith** 曾使用的平台
- **PlatformWantToWorkWith** 想要使用的平台
- **WebframeHaveWorkedWith** 曾使用的Web框架
- **WebframeWantToWorkWith** 想要使用的Web框架
- **MiscTechHaveWorkedWith** 曾使用的其他技术
- **MiscTechWantToWorkWith** 想要使用的其他技术
- **ToolsTechHaveWorkedWith** 曾使用的工具和技术
- **ToolsTechWantToWorkWith** 想要使用的工具和技术
- **NEWCollabToolsHaveWorkedWith** 曾使用的协作工具
- **NEWCollabToolsWantToWorkWith** 想要使用的协作工具
- **OpSysPersonal use** 个人使用的操作系统
- **OpSysProfessional use** 专业使用的操作系统
- **OfficeStackAsyncHaveWorkedWith** 异步办公软件栈曾使用的工具
- **OfficeStackAsyncWantToWorkWith** 异步办公软件栈想要使用的工具
- **OfficeStackSyncHaveWorkedWith** 同步办公软件栈曾使用的工具
- **OfficeStackSyncWantToWorkWith** 同步办公软件栈想要使用的工具
- **AISearchHaveWorkedWith** 曾使用的AI搜索工具
- **AISearchWantToWorkWith** 想要使用的AI搜索工具
- **AIDevHaveWorkedWith** 曾使用的AI开发工具
- **AIDevWantToWorkWith** 想要使用的AI开发工具
- **NEWSOSites** 新的社交网站

- **SOVisitFreq** 访问Stack Overflow的频率
- **SOAccount** Stack Overflow账户
- **SOPartFreq** Stack Overflow参与频率
- **SOComm** Stack Overflow社区参与
- **SOAI** Stack Overflow上的人工智能相关问题
- **AISelect** 选择人工智能工具的重要性
- **AISent** 人工智能伦理的重要性
- **AIAcc** 人工智能的账户
- **AIBen** 人工智能的好处
- **AIToolInterested in Using** 感兴趣使用的人工智能工具
- **AIToolCurrently Using** 当前使用的人工智能工具
- **AIToolNot interested in Using** 不感兴趣使用的人工智能工具
- **AINextVery different** 未来人工智能工具期望的相似度
- **AINextNeither different nor similar** 未来人工智能工具期望的相似度
- **AINextSomewhat similar** 未来人工智能工具期望的相似度
- **AINextVery similar** 未来人工智能工具期望的相似度
- **AINextSomewhat different** 未来人工智能工具期望的相似度
- **TBranch** 目标领域
- **ICorPM** 是否是个体负责人或项目负责人
- **WorkExp** 工作经验
- **Knowledge_1** 知识1
- **Knowledge_2** 知识2
- **Knowledge_3** 知识3
- **Knowledge_4** 知识4
- **Knowledge_5** 知识5
- **Knowledge_6** 知识6
- **Knowledge_7** 知识7
- **Knowledge_8** 知识8
- **Frequency_1** 频率1
- **Frequency_2** 频率2
- **Frequency_3** 频率3
- **TimeSearching** 搜索时间
- **TimeAnswering** 回答时间
- **ProfessionalTech** 专业技术
- **Industry** 行业
- **SurveyLength** 调查长度
- **SurveyEase** 调查难度
- **ConvertedCompYearly** 年度转换薪资

19.statistics_big4_citation_network

> 在《JASA》、《Biometrika》、《J. Roy. Stat. Soc. Series B》和《Annals of Statistics》上发表的论文及其引用网络。

information

> 5,746 篇论文的完整列表，包括四种期刊每篇论文的标题、出版商、doi、摘要、关键词、参考文献、论文编号（详细列表如下）。
>
> [1] JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
> [2] BIOMETRIKA
> [3] JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY
> [4] ANNALS OF STATISTICS

- **paper_id** 论文ID
- **title** 论文标题
- **publisher** 出版商
- **doi** DOI（数字对象标识符）
- **abstract** 摘要
- **references** 参考文献
- **keywords** 关键词

citation

> 节点 "paper_id "的引用网络中的边。共有 23737 条边，每条边代表 "来源 "引用 "目标 "一次。论文之间的边是有向的。

- **source** 源数据
- **target** 目标数据

20. statistics_papers

> 1951 年至 1960 年统计领域发表的 10 年论文。数据包括论文的子领域、发表年份、每篇论文的参考文献列表（即引用网络）。

field

- **paper** 论文ID
- **field** 子领域
- **year** 发表年份

network

- **paper** 论文ID
- **reference** 引用

21.TMDB 6000

> 本数据集由电影数据库 API 生成。本产品使用 TMDb API，但未经 TMDb 认可或认证。本数据集包含截至 2023 年 10 月的最新电影数据。

credits

- **id** 编号
- **cast** 演员表
- **crew** 制作组

movie

- **budget** 预算
- **genres** 类型
- **homepage** 官方网站
- **id** 编号
- **keywords** 关键词
- **original_language** 原始语言
- **original_title** 原始标题
- **overview** 简介
- **popularity** 受欢迎程度
- **production_companies** 制作公司
- **production_countries** 制作国家

- **release_date** 上映日期
- **revenue** 收入
- **runtime** 播放时间
- **spoken_languages** 口语语言
- **status** 状态
- **tagline** 宣传词
- **title** 标题
- **vote_average** 平均评分
- **vote_count** 评分人数

22.top200000scientists

> 数据来源于https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/1，每列的描述在Excel表中。
>
> **描述**：引文指标被广泛使用和滥用。我们建立了一个包含 10 万名顶尖科学家的公开数据库，提供有关引用、h 指数、共同作者调整后的 hm 指数、不同作者位置的论文引用以及综合指标的标准化信息。职业生涯影响和单年影响的数据分开显示。此外，还给出了自引和未自引的指标，以及引文与引用论文的比率。科学家分为 22 个科学领域和 176 个子领域。此外，还提供了发表过至少 5 篇论文的所有科学家的特定领域和子领域百分位数。职业生涯数据更新至 2017 年底和 2018 年底，以便进行比较。

VARIABLES	BASIS	DESCRIPTION
authfull		作者的全名
inst_name		机构名称（仅限大型机构）
cntry		与最近机构相关联的国家
np6022		1960-2022年间发表的论文数量
firstyr		首次发表的年份
lastyr		最近发表的年份
rank (ns)	排除自引用	基于复合分数c的排名
nc9622 (ns)	排除自引用	1996-2022年间的总引用数
h22 (ns)	排除自引用	截至2022年底的h指数
hm22 (ns)	排除自引用	截至2022年底的hm指数
nps (ns)	排除自引用	单一作者发表的论文数量
ncs (ns)	排除自引用	单一作者发表的论文总引用数
cpsf (ns)	排除自引用	单一+第一作者发表的论文数量
ncsf (ns)	排除自引用	单一+第一作者发表的论文总引用数
npsfl (ns)	排除自引用	单一+第一+最后作者发表的论文数量
ncsfl (ns)	排除自引用	单一+第一+最后作者发表的论文总引用数
c (ns)	排除自引用	复合分数c（排除自引用）
npciting (ns)	排除自引用	不同引用论文的数量（排除自引用）
cprat (ns)	排除自引用	总引用与不同引用的比率（排除自引用）
np6022 cited9622 (ns)	排除自引用	1960-2022年间至少被引用一次的论文数量（排除自引用）
self%	自引用百分比	自引用百分比
rank	所有引用	基于复合分数c的排名
nc9622	所有引用	1996-2022年间的总引用数
h22	所有引用	截至2022年底的h指数
hm22	所有引用	截至2022年底的hm指数
nps	所有引用	单一作者发表的论文数量
ncs	所有引用	单一作者发表的论文总引用数
cpsf	所有引用	单一+第一作者发表的论文数量

VARIABLES	BASIS	DESCRIPTION
ncsf	所有引用	单一+第一作者发表的论文总引用数
npsfl	所有引用	单一+第一+最后作者发表的论文数量
ncsfl	所有引用	单一+第一+最后作者发表的论文总引用数
c	所有引用	复合分数c
npciting	所有引用	不同引用论文的数量
cprat	所有引用	总引用与不同引用的比率
np6022 cited9622	所有引用	1960-2022年间至少被引用一次的论文数量
np6022_d	1960-2022年间标题已在Scopus中停用的论文数量	1960-2022年间在Scopus中停用的标题的论文数量
nc9622_d	由于在Scopus中停用的标题而产生的1996-2022引用总数	由于在Scopus中停用的标题而产生的1996-2022引用总数
sm-subfield-1	所有引用	作者的排名第一的Science-Metrix类别（子领域）
sm-subfield-1-frac	所有引用	关联类别的分数
sm-subfield-2	所有引用	作者的排名第二的Science-Metrix类别（子领域）
sm-subfield-2-frac	所有引用	关联类别的分数
sm-field	所有引用	作者的排名第一的更高级别Science-Metrix类别（领域）
sm-field-frac	所有引用	关联类别的分数
rank sm-subfield-1	所有引用	类别sm-subfield-1中c的排名
rank sm-subfield-1 (ns)	排除自引用	排除自引用的类别sm-subfield-1中c(ns)的排名
sm-subfield-1 count	类别sm-subfield-1中的总作者数	类别sm-subfield-1中的总作者数

注：数据在第二个sheet中

23.xiyouji

> 吴承恩所著《西游记》的人物-事件出场矩阵。数据以 R 数据帧的形式存储，您需要使用 Pyreadr 等软件包在 Python 中读取数据。

列名

出场章数，角色名称

24. yelp

> Yelp 应用程序数据。完整的评论文本数据，包括撰写评论的用户 ID 和撰写评论的企业 ID。

- `review_id`：字符串，22 个字符的唯一评论 ID
- `user_id`：字符串，22 个字符的唯一用户 ID，映射到 user.json 中的用户
- `bussiness_id`：字符串，22 个字符的企业 ID，映射到 business.json 中的企业
- `stars`：整数，星级
- `date`：字符串，日期格式 YYYY-MM-DD
- `text`：字符串，评论本身
- `useful`：整数，获得的有用票数
- `funny`：整数，收到的有趣投票数
- `cool`：整数，收到的 "酷" 票数