数据集的名称及对应含义——中文理解版本

## user字段

|  |  |  |  |
| --- | --- | --- | --- |
| user\_id | string | 22 character unique user id, maps to the user in user.json | 22 个字符的唯一用户 ID，映射到 user.json 中的用户（包括商家和消费者吗） |
| name | string | the user's first name | 昵称？jack？（违规昵称？） |
| review\_count | integer | the number of reviews they've written | 【Review】user-》reviews |
| yelping\_since | string | when the user joined Yelp, formatted like YYYY-MM-DD | 【时间】user——》注册时间 |
| friends | array | array of strings, an array of the user's friend as user\_ids | 【用户】user——》friends【user\_ids】 |
| useful | integer | number of useful votes sent by the user | 【用户】user——》useful votes —— ALL商家？ 非商家维度，而是其他用户对该用户**所有评论**的 “有用” 点赞数。例如：用户 A 的某条评论被 10 人点击 “有用”，则 useful 计数 + 10。 |
| funny | integer | number of funny votes sent by the user | 【用户】user——》funny votes —— ALL商家？ |
| cool | integer | number of cool votes sent by the user | 【用户】user——》cool votes —— ALL商家？ |
| fans | integer | number of fans the user has | 【用户】粉丝 |
| elite | array | the years the user was elite | 【时间】优质用户 |
| average\_stars | float | average rating of all reviews | 【review】平均评分 |
| compliment\_hot | integer | number of hot compliments received by the user | 用户收到的热门赞美数量【用户】点赞   Yelp 用户可对其他用户发送不同类型的赞美，“hot” 可能指内容优质、受欢迎的赞美（如评论被大量推荐）。 |
| compliment\_more | integer | number of more compliments received by the user | ？？用户收到的更多表扬数量——?回复评论的人想要知道更多关于商家的内容的评论  实际指其他用户认为该用户的内容（如评论、资料）提供了**额外价值**，给予 “more” 类型的赞美（如 “想了解更多该商家信息，感谢分享”）。 |
| compliment\_profile | integer | number of profile compliments received by the user | 用户收到的个人资料赞美数量——？个人资料包括哪些？能看到usr到处游玩的经历，看到这些所以夸usr的评论 指用户个人主页信息（如头像、简介、打卡记录等）被其他用户赞美的次数，例如：资料填写详细、排版美观等。 |
| compliment\_cute | integer | number of cute compliments received by the user | 用户收到的可爱赞美数量，被夸可爱 对用户内容（如评论语气、头像风格）的 “可爱” 属性赞美，常见于轻松、有趣的互动场景。 |
| compliment\_list | integer | number of list compliments received by the user | ？？用户收到的列表赞美数，区别于profile？  **区别于 profile？** Yelp 用户可创建商家列表（如 “必吃汉堡店”），该字段记录列表被其他用户赞美的次数，与个人资料无关。 |
| compliment\_note | integer | number of note compliments received by the user | 用户收到的 note 赞美数 可能与 Yelp 的 “笔记” 功能相关（用户可在商家页面添加私人笔记），其他用户对笔记内容的赞美计数。 |
| compliment\_plain | integer | number of plain compliments received by the user | 用户收到的普通赞美数量 无特定分类的基础赞美，可能指简单的 “赞” 或 “不错” 等非定制化评价。 |
| compliment\_cool | integer | number of cool compliments received by the user | 用户收到的 COOL 赞美数量 对用户行为或内容的 “酷”“新潮” 属性赞美，如探店体验独特、评论视角新颖。 |
| compliment\_funny | integer | number of funny compliments received by the user | 用户收到的有趣赞美数量， 区别于评论投票的 “funny”，此处指其他用户直接对用户发送的 “有趣” 类赞美（非投票，而是主动留言赞美）。 |
| compliment\_writer | integer | number of writer compliments received by the user | 用户收到的作家赞美数量——夸用户长得好看？ 赞美用户的**写作能力**，如评论语言流畅、文采出众。 |
| compliment\_photos | integer | number of photo compliments received by the user | 用户收到的照片赞美数量——夸用户拍的好； 其他用户对该用户上传的商家照片（如菜品图、环境图）的质量赞美，如 “照片拍得专业”。 |

## business数据

### Yelp 商户数据集 business.json 字段含义解析表

| 字段名 | 数据类型 | 官方定义简述 | 详细解释（结合平台特性补充） |
| --- | --- | --- | --- |
| business\_id | string | 22 字符唯一商户 ID | 用于唯一标识每个商户，可与用户评论、用户互动数据关联（如通过 business\_id 匹配某商户的所有评论）。 |
| name | string | 商户名称 | 注册时填写的正式名称（如 “Starbucks Coffee”），可能包含分店名（如 “Starbucks Coffee - Times Square”）。 |
| address | string | 完整地址 | 【location】包含门牌号、街道、街区等详细地址信息（如 “123 Main St, Suite 45”），用于用户导航或地理位置分析。 |
| city | string | 所在城市 | 【location】如 “San Francisco”“New York”，可结合 state 和 postal code 定位具体区域。 |
| state | string | 州代码（2 字符） | 【location】美国各州缩写（如 “CA” 加州、“NY” 纽约州），非美国商户可能为空或特殊标识。 |
| postal\_code | string | 邮政编码 | 【location】用于细分地理位置，如美国 zip code（“94103”），可结合经纬度进行区域商业分析（如不同邮编区的商户分布）。 |
| latitude | float | 纬度 | 【location】地理坐标纬度值（如 37.7749°N），与 longitude 配合可精确定位商户位置，用于地图展示或空间分析。 |
| longitude | float | 经度 | 【location】地理坐标经度值（如 - 122.4194°W），与 longitude 配合可精确定位商户位置，用于地图展示或空间分析。 |
| stars | float | 星级评分（四舍五入到半星） | 【商家】**平均评分**：基于所有用户评论的星级汇总，如 4.5 星表示该商户整体评价较高，半星机制（如 3.5 星）体现更细致的评分差异。 |
| review\_count | integer | 评论数量 | 【商家】该商户收到的用户评论总数，可反映商户的关注度（如 review\_count 高的商户可能更热门）。 |
| is\_open | integer | 营业状态（0 = 关闭，1 = 营业） | 【当前状态、时间？】实时记录商户是否营业，0 表示已永久关闭或临时停业，1 表示正常营业（注意：Yelp 可能根据用户反馈更新该状态）。 |
| attributes | object | 商户属性键值对（部分值可能为对象） | 【商家info】**属性与设施**： 包含商户特色信息，例如： - WiFi: {"free": true}（免费 WiFi） - Outdoor Seating: true（有户外座位） - Price Range: 3（价格区间 3 星，即中高端） 不同行业属性不同（如餐厅含 “Accepts Credit Cards”，酒店含 “Parking”）。 |
| categories | array | 商户类别数组 | 【商家diff】所属行业分类（如["Restaurants", "Pizza", "Italian"]），用于搜索分类（用户可按类别筛选商户），类别层级从大到小（先主类后子类）。 |
| hours | object | 营业时间（键为星期，值为 24 小时制时间） | 【时间】例如：{"Monday": "10:00-22:00", "Sunday": "11:00-21:00"}，部分商户可能标注 “24 hours”（24 小时营业），用于用户查询商户营业时段。 |

### 补充说明：Yelp 商户数据的业务逻辑

1. **地理位置数据的应用**：
   * latitude和longitude可用于绘制商户分布热力图，分析区域商业密度（如市中心餐饮商户集中）；
   * postal\_code配合city和state可用于细分市场分析（如不同邮编区的消费偏好差异）。
2. **营业状态与属性的动态性**：
   * is\_open字段可能随时间变化（如商户倒闭后更新为 0），需结合数据更新时间判断时效性；
   * attributes中的部分属性（如 “WiFi 免费”“接受预约”）可能随商户政策调整而改变，字段值记录的是数据采集时的状态。
3. **类别体系的层级逻辑**：  
   Yelp 的categories采用 “主类别 + 子类别” 结构（如["Food", "Burgers", "Fast Food"]），越靠后的类别越具体，可用于商户精准分类（如筛选 “意大利披萨餐厅”）。
4. **评分与评论的关联性**：
   * stars是review\_count条评论的星级平均值，反映商户的整体口碑；
   * 可通过分析高评分商户的attributes（如是否有 “Outdoor Seating”）挖掘用户偏好（如有户外座位的餐厅评分更高）。

通过以上解析，可全面理解 Yelp 商户数据各字段的业务含义及潜在分析价值，为后续商业智能分析（如选址推荐、竞品对比、用户偏好建模）提供基础。

## review表

### Yelp 评论数据集 review.json 字段含义解析表

| 字段名 | 数据类型 | 官方定义简述 | 详细解释（结合平台互动逻辑补充） |
| --- | --- | --- | --- |
| business\_id | string | 22 字符商户 ID，映射到 business.json | 用于关联评论对应的商户，可结合商户数据（如类别、评分）分析不同类型商户的评论特征（如高端餐厅评论更关注服务细节）。 |
| cool | integer | 评论获得的 “酷” 投票数 | Yelp 用户对评论的互动反馈，“cool” 表示评论内容被认为 “新潮”“有个性”（如独特的探店视角），投票数越高说明该评论在 “酷” 维度越受认可。 |
| date | string | 评论发布日期（YYYY-MM-DD 格式） | 如 “2023-05-15”，可用于时间序列分析（如节假日前后评论量变化）或评论时效性判断（旧评论可能不反映当前商户状态）。 |
| funny | integer | 评论获得的 “有趣” 投票数 | 评论因幽默、生动的表达获得的投票，例如用段子描述用餐体验的评论可能积累较多 funny 投票，反映评论的娱乐性价值。 |
| review\_id | string | 22 字符唯一评论 ID | 用于唯一标识每条评论，可追踪评论的互动历史（如后续是否有回复或修改），或与其他数据集（如用户点赞记录）关联。 |
| stars | integer | 评论星级评分（1-5 星） | 用户对商户的直接评分，1 星为差评，5 星为好评，半星通过四舍五入转换（如 4.5 星在评论中可能记录为 5 星？需结合数据实际存储规则）。 |
| text | string | 评论正文内容 | 包含用户对商户的具体评价（如 “牛排煎得恰到好处”“服务态度恶劣”），是自然语言处理的核心数据，可提取关键词分析用户关注点（如 “排队”“价格” 高频出现可能反映商户痛点）。 |
| useful | integer | 评论获得的 “有用” 投票数 | 最常见的互动类型，用户认为评论对消费决策有帮助（如 “详细描述了菜品分量”）时点击 “有用”，是衡量评论信息价值的重要指标。 |
| user\_id | string | 22 字符用户 ID，映射到 user.json | 关联评论作者，可结合用户数据（如 review\_count、elite 身份）分析不同类型用户的评论特征（如精英用户的评论更详细）。 |

### 补充说明：Yelp 评论数据的业务场景与分析价值

1. **用户互动逻辑拆解**：
   * useful/funny/cool是 Yelp 预设的三种投票类型，分别对应评论的**信息价值**、**娱乐性**和**独特性**，三者可组合分析评论的综合影响力（如一条评论同时获得高 useful 和 funny 投票，说明既实用又有趣）。
   * 投票数由所有看到评论的用户产生，非商家或平台干预，因此能客观反映评论的受欢迎程度。
2. **评分与文本的关联性**：
   * stars是量化评分，text是定性描述，两者需结合分析：例如 5 星评论中高频出现 “服务周到”“环境优美”，可提炼商户优势；1 星评论中 “卫生差”“态度差” 高频出现，可定位问题点。
   * 部分评论可能存在 “评分与文本矛盾”（如 4 星评论正文抱怨排队久），需通过情感分析算法挖掘真实态度。
3. **时间维度的应用**：
   * date可用于分析评论量随时间的波动（如餐厅在新店开业期评论量激增），或商户口碑变化（如某商户 2022 年评分下降，可结合当时的评论文本查找原因）。
   * 结合business.json的is\_open字段，可分析闭店商户的历史评论，总结失败原因（如闭店前半年差评集中爆发）。
4. **数据关联分析场景**：
   * **用户 - 评论关联**：通过user\_id分析高频评论用户的偏好（如某用户常给日料店 5 星评论，可能是日料爱好者）；
   * **商户 - 评论关联**：通过business\_id分析某商户的评论情感倾向（如提取所有评论的 “positive”“negative” 关键词频率）；
   * **跨数据集关联**：将review.json与user.json、business.json结合，构建 “用户 - 商户 - 评论” 三角关系，用于推荐系统（如给 A 用户推荐与其评分偏好相似的 B 用户曾好评的商户）。

## checkin表

### Yelp 签到数据集 checkin.json 字段含义解析表

| 字段名 | 数据类型 | 官方定义简述 | 详细解释（结合平台功能与业务场景补充） |
| --- | --- | --- | --- |
| business\_id | string | 22 字符商户 ID，映射到 business.json | 用于关联签到对应的商户，可结合商户数据（如类别、营业时间）分析不同类型商户的签到特征（如快餐店签到高峰在午餐时段）。 |
| date | string | 签到时间字符串（多个时间点用逗号分隔） | 记录用户在商户页面主动签到的时间戳，格式为"YYYY-MM-DD HH:MM:SS, ..."（如示例中的 6 个时间点）。 **核心逻辑**： - 每个时间点代表一次用户签到行为（可能是同一用户多次签到，或不同用户的签到汇总）； - Yelp 允许用户在访问商户时 “签到” 以记录足迹，类似 “打卡” 功能，可用于社交分享或积累积分。 |

### 补充说明：Yelp 签到数据的业务价值与分析场景

1. **签到数据的本质与来源**：
   * 签到是用户主动操作的行为，不同于评论的 “深度反馈”，更偏向 “浅度互动”，主要用于记录 “我来过这里” 的足迹；
   * date字段存储的是**该商户所有签到事件的时间集合**，而非单个用户的签到记录（即未关联 user\_id，无法区分具体是谁签到）。
2. **时间维度的深度分析**：
   * **时段分布**：提取所有签到时间的小时数（如 12:00-14:00），可分析商户的客流高峰时段（如餐厅午市、晚市签到量高）；
   * **日期规律**：按周 / 月统计签到量（如周五晚签到量显著高于周一），可辅助商户制定排班或营销计划；
   * **特殊时间点**：如签到时间在凌晨 2 点（如示例中的 “2016-10-15 02:45:18”），可能反映商户 24 小时营业或夜间消费场景（如酒吧）。
3. **与其他数据集的关联应用**：
   * **结合商户属性**：若business.json中该商户标注 “Outdoor Seating: true”，可分析天气晴朗日的签到量是否更高；
   * **结合评论数据**：某商户在签到高峰时段后出现大量好评，可能说明客流高峰时服务质量稳定；
   * **地理位置关联**：通过business\_id获取商户经纬度，结合签到时间分析用户动线（如某用户在下午 18:00 签到 A 餐厅，19:30 签到附近 B 电影院，反映消费路径）。
4. **潜在业务场景**：
   * **商户运营**：根据签到高峰调整员工排班（如签到量集中在 12:00-13:00，增加该时段服务员）；
   * **用户行为研究**：分析不同城市用户的签到习惯（如纽约用户更爱在工作日晚上签到，洛杉矶用户周末签到更多）；
   * **推荐系统**：基于用户常签到的商户类型，推荐同类别其他商户（如频繁签到意大利餐厅的用户，推荐新开业的意大利餐厅）。

### 注意事项

* 签到数据未包含用户身份信息（无 user\_id），无法直接关联具体用户画像，需通过其他数据集（如评论的 user\_id）间接推断签到用户特征；
* 部分商户可能因人气低而签到记录少，或因数据采集限制未完整记录所有签到事件，分析时需注意样本量；
* 签到时间可能存在 “用户事后补签” 的情况（如隔天签到昨天的消费），需结合业务场景判断数据时效性。

通过以上解析，可明确 Yelp 签到数据的字段含义及应用方向，为客流分析、时间序列建模、地理位置商业洞察等提供数据支撑。

# #