# 质量调查与评价

## 1. 引言

在数字化时代背景下，开放数据的质量直接影响政策制定、商业决策及科研分析的有效性。通过对全国除港澳台地区以外的31个省级行政区公共数据平台建设情况的考察（见附录：我国除港澳台外，各省级行政区公共数据平台建设汇总表），我发现山东、四川、辽宁、广西、宁夏五省的公共数据平台如出一辙。实际测试后发现，只有四川公共数据开放网的建设情况较为可观，同时支持使用爬虫进行数据集的自动化抓取。山东公共数据开放网的问题是，数据集详情页中的主要内容使用动态加载，大大增加了数据抓取的难度；辽宁省公共数据开放平台的问题是，数据集数量过少，其中无条件开放且提供csv格式的数据集仅86个（截至2024年5月6日）；而我的故乡，宁夏公共数据开放平台，十分遗憾已经停止更新近两年，所有数据集下载和预览等功能均无法使用。故本次实验选取四川公共数据开放网作为数据来源。四川省作为中国重要的省份之一，其公开的数据集包含了丰富的社会经济、文化及环境信息，这些数据的准确性和完整性对于推动地方发展具有重要意义，为本项目提供了丰富的数据基础。

本项目使用Scrapy爬虫框架来实现数据的自动化抓取，依据数据科学的方法来评价数据的结构完整性、唯一性、数据完整性和及时性这四个维度，旨在全面了解数据质量，并通过可视化技术使分析结果直观呈现。

## 2. 问题调查与分析

### 2.1数据来源

本项目数据源为四川公共数据开放网，以下简称**开放网**。为便于下载和处理，选取了开放类型为**无条件开放**且提供**CSV**格式文件的数据集。数据集搜集于2024年4月30日进行，共搜集到**4022**个数据集的相关文件。

### 2.2数据特征

开放网上的数据集具有多样性和广泛的应用潜力。根据数据目录页领域筛选的选项，这些数据集包括教育文化、生活服务、地理空间、社保就业等21个重点领域。每个数据集都附带有详细的元数据，包括数据集名称、发布和更新时间、来源部门等信息，这些元数据对于后续理解和分析数据至关重要。

### 2.3面临的挑战

在数据抓取和质量评估的过程中，我面临了多项挑战：

* **数据集规模**: 一些数据集规模庞大，包含数十万条记录，开放网限制了可直接下载的数据文件最多包含前1000条数据，全量数据需要通过API接口方式获取，因此本项目中使用到的数据集有许多不是完整的。
* **数据规范性问题**: 部分数据集存在缺失数据或格式错误的问题；部分数据集并非原始数据，即csv格式通常的一行一记录模式，而是经过人为编辑修改后的报告表格；更有甚者完全是以html的格式呈现，导致数据集无法被正常读取并处理。我选择专门设立一项度量指标：结构完整性，在后续实验过程中首先评估，以作为能否进一步进行数据质量检测的必要条件。
* **动态内容加载问题**: 如引言部分提到的，一些数据开放平台在不同的地方使用JavaScript动态加载技术，这增加了爬虫的复杂度，使得数据爬取被卡在某一处无法继续，最终我选择暂时放弃考察这些平台。

综上所述，虽然四川公共数据开放网为本项目提供了宝贵的数据资源，但在数据抓取和处理过程中所遇到的技术和方法论挑战也促使我不断优化和改进抓取与分析方法，以确保数据质量和项目的成功实施。

刚拿到4022份数据集文件时我束手无策，对代码复杂性的预测让我没有选择直接挨个读取的走一步看一步式处理。通过手动浏览部分数据集内容，观察脏数据的出现类型和方式，我选择回到开放网的数据集详情页，设计爬虫以生成一份数据集目录以储存数据集的元数据。经过测试和修改，我开始收集每个数据集的：名称、开放网ID、URL、来源部门、重点领域、发布时间、最后更新时间、更新频率、一条示例数据，以上共9条信息作为一个数据集的元数据，于2024年4月30日收集并整理得到包含**4011**条元数据记录的**数据集目录**。

由于数据集爬取和元数据记录收集的分别执行，导致数据集文件和元数据记录出现冲突。为同步数据集文件和数据集目录内记录，从**全部数据集**中去除了**234**个数据目录未记录的数据集文件，从**数据集目录中**去除**211**条文件未下载的数据集记录。同步后，此时共有**3788**个数据集和**3800**条元数据记录，进一步排查发现元数据记录中存在**23**条数据集为重名：名称重复出现两次的数据集共**10**个，三次的共**1**个。访问数据集对应URL发现确有此事，四川公共数据开放网**存在**重名的数据集，这些数据集的内容和结构彼此略有出入。解决这一冲突，本次实验共剩下**3777**条数据集记录及对应的**3777**个csv数据集文件，并以这些记录和文件开展进一步分析。

通过初步调查，发现开放网上数据集存在的脏数据问题较多，包括但不限于：数据值缺失、数据格式（或表达）不一致、未知值表达不一致、数据脱敏程度不一致、时间值不合理、无数据值单位、数据重复、不合理值或错误值、数据过于笼统、数据过于精准等等。

表1展示了眉山市\_市生态环境局\_建设项目国家报送审批数据（省市县三级）中的部分脏数据问题：占地面积、项目总投资、项目环保投资3列均存在数据格式（或表达）不一致和无数据值单位的脏数据问题，占地面积列数据值大量缺失，电话列脱敏程度不一致，环境保护联系人列出现不合理值或错误值。

表..1眉山市\_市生态环境局\_建设项目国家报送审批数据（省市县三级）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 项目名称 | 占地面积 | 项目总投资 | 项目环保投资 | 电话 | 环境保护联系人 |
| 平昌县邱家红红建材项目 | 1320m2 | 150 | 19 | 139\*\*\*\*\*\*\*2 | 苏小红 |
| 内嵌式中低速磁悬浮交通系统研发及产业化（二期） |  | 47277 | 104 |  |  |
| 德发农业气调库建设项目 | 60000 | 8000 | 40 | 138\*\*\*\*\*\*\*2 | 杨明海 |
| 南充市顺庆区共兴污水处理站及配套管网建设工程 | 2599.81平方米 | 1525.93 | 118 | 176\*\*\*\*\*\*\*7 | 党宁 |
| 资源循环利用项目 |  | 11500 | 158 | 139\*\*\*\*\*\*\*2 | 13980506862 |
| 蓝光天娇城二期建设项目 | 99995㎡ | 89000 | 1219.5 | 185\*\*\*\*\*\*\*9 | 史琨 |
| 红原县牦牛、藏绵羊毛技工创富增收建设项目 |  | 2800 | 79.5 | 1390904724 | 阿云峰 |
| 成都市青白江区姚渡镇卫生院 |  | 69 | 17.5 | 136080\*\*\*\*\*5 | 陈秋霖 |

表2展示了攀枝花市\_市水利局\_攀枝花\_市水利局\_巡河记录表中的部分脏数据问题，两列日期的格式不一致且有明显错误，县级是否代填一列数据格式不一致（是和0），河长类型列值意义不明，以后的几列取值大量重复且不合理（会议名称与实际河段冲突）。在没展示出的后几列关于会议内容、问题类型和问题备注的取值，1000行数据中大量重复，实际只有6到7种唯一的值。

表 2攀枝花市\_市水利局\_攀枝花\_市水利局\_巡河记录表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 开始日期 | 结束日期 | 县级是否代填 | 河段 | 河长类型 | 主要巡河人员 | 巡河人员数 | 巡河会议名称 |
| 2022/8/1018:10 | 2022/8/1018:09 | 0 | 安宁河 | 否 | 肖济今 | 2 | 巡河 |
| 2022/8/1018:07 | 2022/8/1018:03 | 是 | 白沙沟 | 否 | 肖济今 | 2 | 金沙江河段巡河情况 |
| 2022/8/1018:00 | 2022/8/1018:00 | 0 | 小河 | 否 | 肖济今 | 2 | 巡河 |
| 2022/8/1017:53 | 2022/8/1018:00 | 是 | 安宁河 | 否 | 肖济今 | 2 | 金沙江河段巡河情况 |
| 2022/8/1017:47 | 44760 | 0 | 安宁河 | 否 | 肖济今 | 2 | 巡河 |
| 2022/8/1017:47 | 2022/8/1017:53 | 是 | 长江（金沙江） | 否 | 肖济今 | 2 | 金沙江河段巡河情况 |
| 2022/8/1017:39 | 44760 | 0 | 小河 | 否 | 肖济今 | 2 | 巡河 |
| 2022/8/1017:32 | 2022/8/1017:46 | 是 | 灰老河 | 否 | 肖济今 | 2 | 金沙江河段巡河情况 |
| 2022/8/1017:32 | 44760 | 0 | 安宁河 | 否 | 肖济今 | 2 | 巡河 |
| 2022/7/29 | 2022/7/29 | 是 | 长江（金沙江）(攀枝花市)银江镇段 | 否 | 肖济今 | 2 | 金沙江河段巡河情况 |

## 3. 质量维度与度量指标

### 3.1数据质量维度

依据四个关键的数据质量维度来评估数据集的质量，每个维度针对数据质量的不同方面进行了深入分析：

1. **结构完整性 (Integrity)**
   * **定义**: 结构完整性是通过评估数据文件是否符合标准CSV格式来衡量的，确保每个数据集的每一行都应正确分隔，无缺失字段。
   * **测量方法**: 此指标通过解析每个CSV文件，验证字段数与表头一致性来评估。数据集若完全符合预期格式，则此值为1；若存在格式错误，如字段错位或缺失，则逐渐接近于0。
2. **唯一性 (Uniqueness)**
   * **定义**: 唯一性衡量数据集中记录的重复性。完全没有重复记录的数据集视为具有最高的唯一性。
   * **测量方法**: 通过检查数据集中每条记录的唯一性来评定。如果所有记录均为唯一，则此指标为1；如果数据集中存在重复记录，则此值为0。
3. **数据完整性 (Completeness)**
   * **定义**: 数据完整性指数据集中所需字段的完整填充程度。
   * **测量方法**: 该指标通过计算数据集中非空字段的比例来评定。完全没有缺失值的数据集完整性为1，缺失值越多，指标值越低。
4. **及时性 (Timeliness)**
   * **定义**: 及时性衡量数据集的更新频率是否符合预定的更新计划。
   * **测量方法**: 该指标基于数据集的最后更新时间与预定更新频率的对比。如数据定期更新且最近的更新符合预定频率，则此值为1；如果更新滞后或不符合预期频率，则此值为0。

### 3.2度量指标定义

以下是每个维度的具体定义与解读：

* **结构完整性**:
  + *值范围 [0, 1]*: 1表示数据完全按照CSV格式要求构建，0则表示数据文件存在严重的结构问题，如嵌入的HTML或其他格式错误。
  + *NaN（空数据集）*: 如果CSV文件为空，则此维度不适用，记录为空值（以下用NaN表示）。
* **唯一性**:
  + *1（唯一）*: 表示数据集中无重复记录。
  + *0（不唯一）*: 表示数据集中存在重复记录。
  + *NaN*: 如果由于结构完整性问题无法正确解析数据，则无法判断唯一性，记录为NaN。
* **数据完整性**:
  + *值范围 [0, 1]*: 表示数据字段的填充程度，1表示无任何缺失字段，0表示所有字段均为空。
  + *NaN*: 如果数据集无法被解析（如结构问题严重），则完整性无法评定，记录为NaN。
* **及时性**:
  + *1（及时）*: 数据更新符合预定频率。
  + *0（不及时）*: 数据更新不符合预定频率。
  + *NaN（无法判断）*: 如果缺乏关于预定更新频率的信息，或更新频率为不定期，此维度记录为NaN。

以上度量标准为实际操作提供了明确的评估基准，确保数据质量评估的客观性和准确性。在后续的综合分析部分，我将根据这些度量指标的具体数值，对数据集的整体质量进行深入的分析与讨论。

## 4. 综合分析

### 4.1分析方法论

采用系统化的数据分析流程来评估每一份数据集的质量。首先，我们将数据集目录导入python，检测并移除有冲突的数据集文件和数据集记录，发现并移除重名的数据集记录和对应的数据集文件，通过分析数据集记录的示例数据将空数据集转移到单独一个文件夹，检测并转移不具备结构完整性的数据集到单独一个文件夹，至此准备工作结束。

接下来对于每个数据质量维度，运用以下方法进行深入分析：

* **结构完整性**: 编写自动化脚本逐行检查CSV文件的格式规范性，主要通过验证每行的字段数。
* **唯一性**: 利用数据查重，识别数据集中重复记录。
* **数据完整性**: 通过统计各字段非空值的比例，评估数据填充的完整性。
* **及时性**: 对比数据的最后更新时间与预定更新频率，评估数据是否按计划更新。

得到并记录评估结果后，采用数据可视化技术以帮助直观地展示分析结果，使得数据质量问题一目了然。

### 4.2结果总结

通过对3777份数据集的综合分析，我发现以下几个典型的数据质量问题：

* **结构完整性问题**: 39个数据集存在一定的结构完整性问题，包括错误的字段分隔、嵌入的html等，其中11个数据集存在十分严重的结构完整性问题。
* **数据重复问题**: 705个数据集存在一定程度的记录重复，约占全部数据集的19%。
* **不完整的记录**: 2392个数据集的完整性小于1，约占全部数据集的64%，主要是由于缺失值较多，3777个数据集中只有1385个数据集是完全完整的，有2666个数据集的完整性在0.75以上。
* **更新不及时**: 963个数据集的实际更新频率未能达到其预定的更新周期，仅有996个数据集做到更新及时。1818个数据集无法判断其及时性。

### 4.3可视化解读

通过绘制的可视化图表，我们可以清晰地识别出数据质量的不同维度中存在的问题：

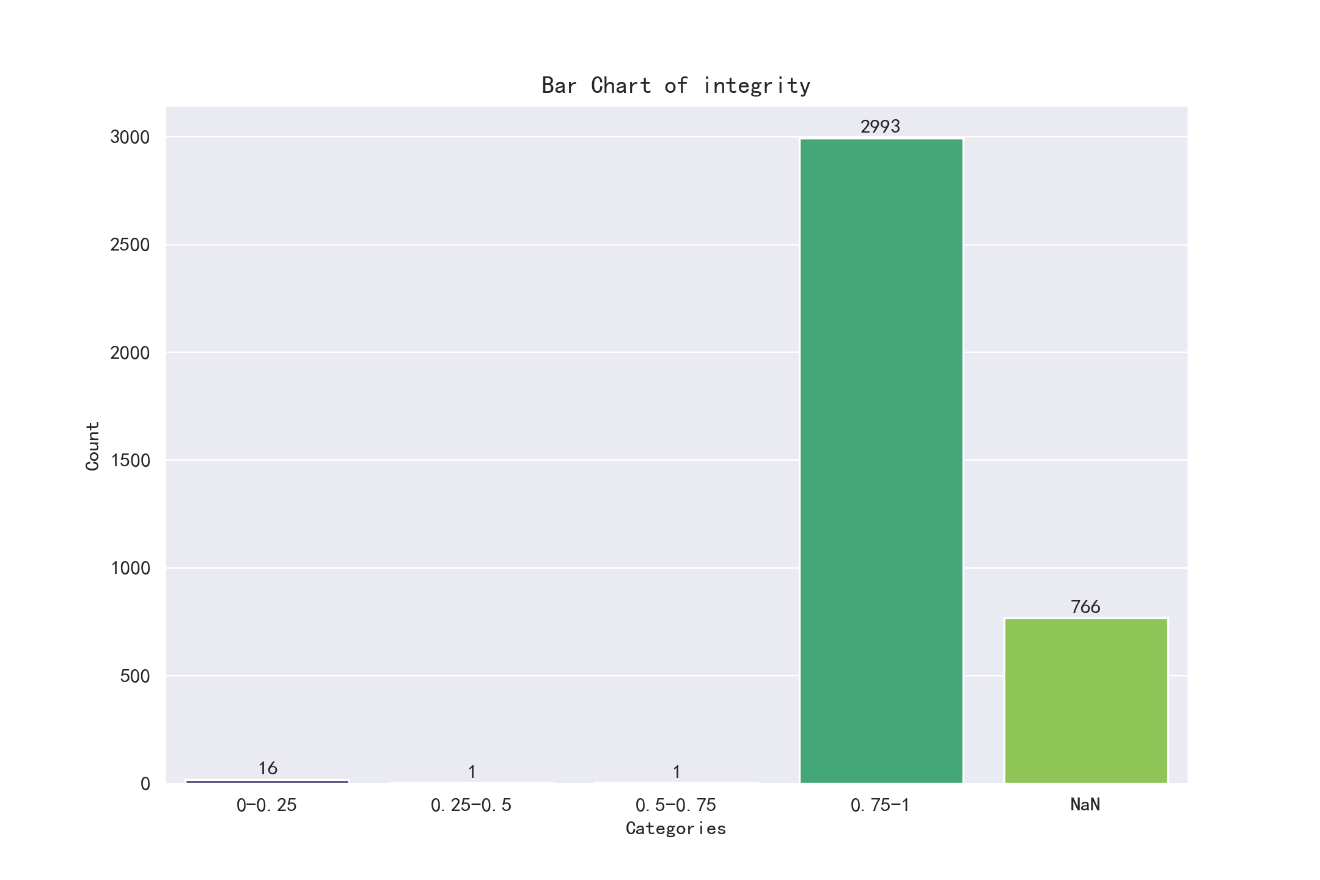


图1 数据集结构完整性统计图

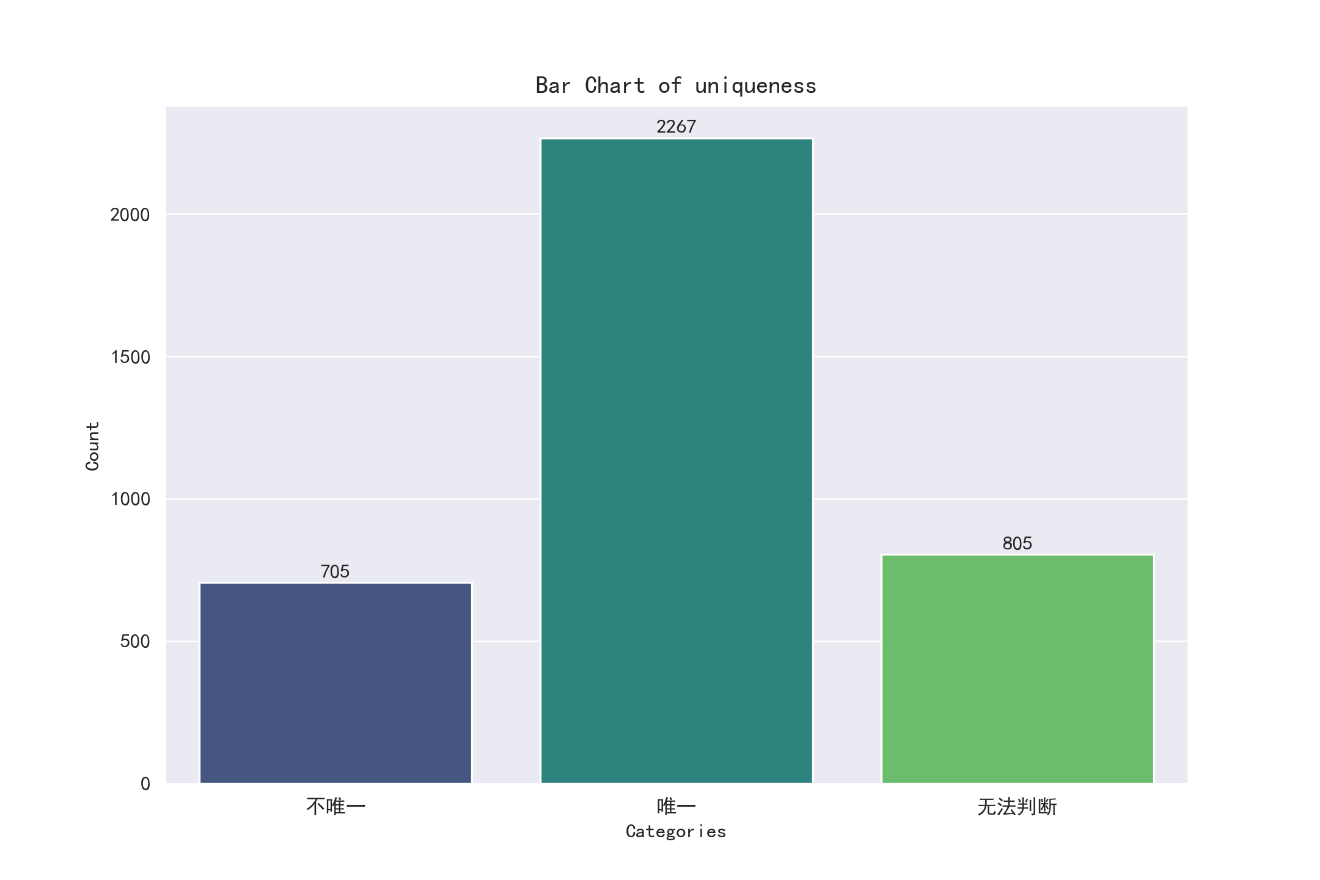


图 2数据集记录唯一性统计图

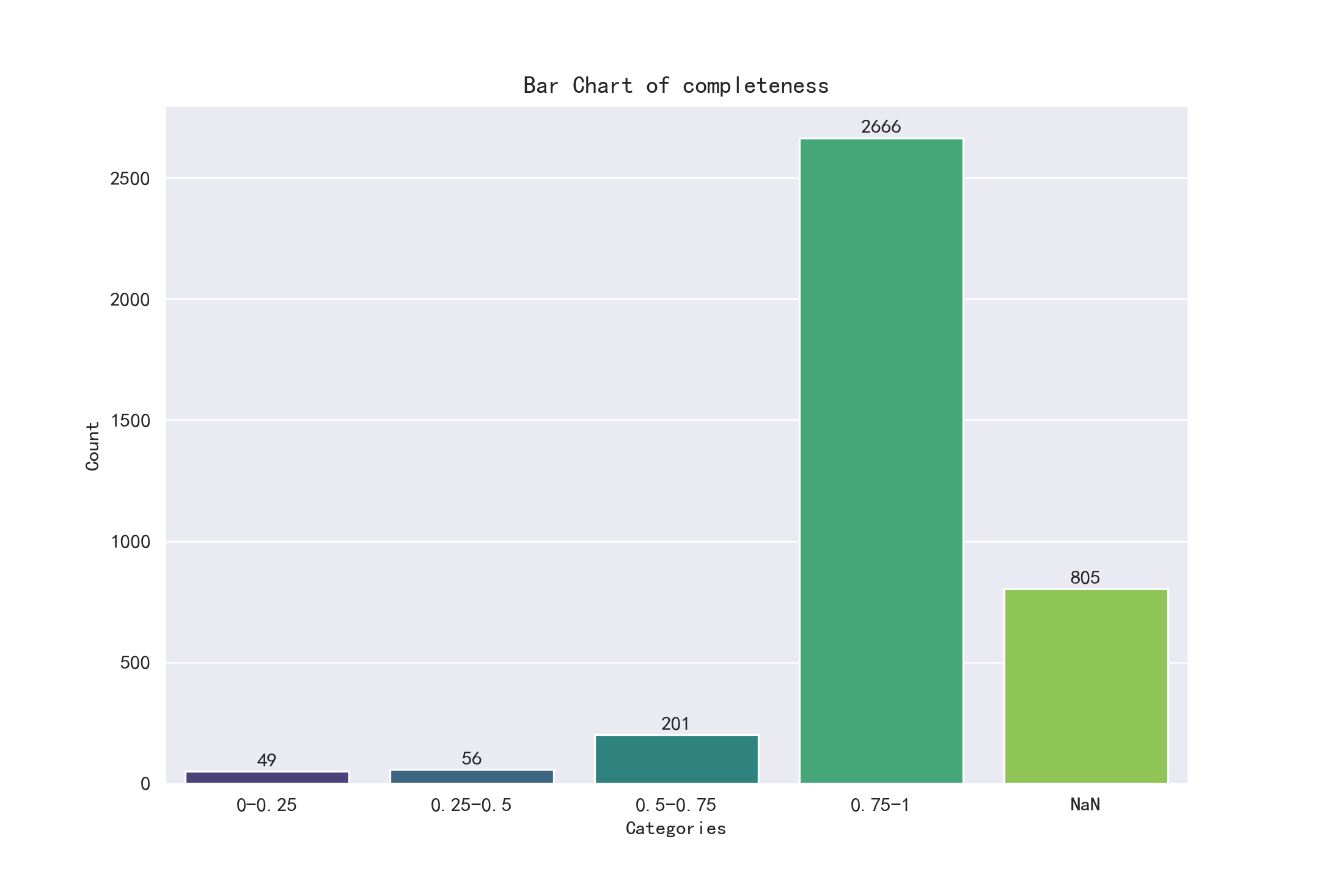


图3 数据集完整性统计图

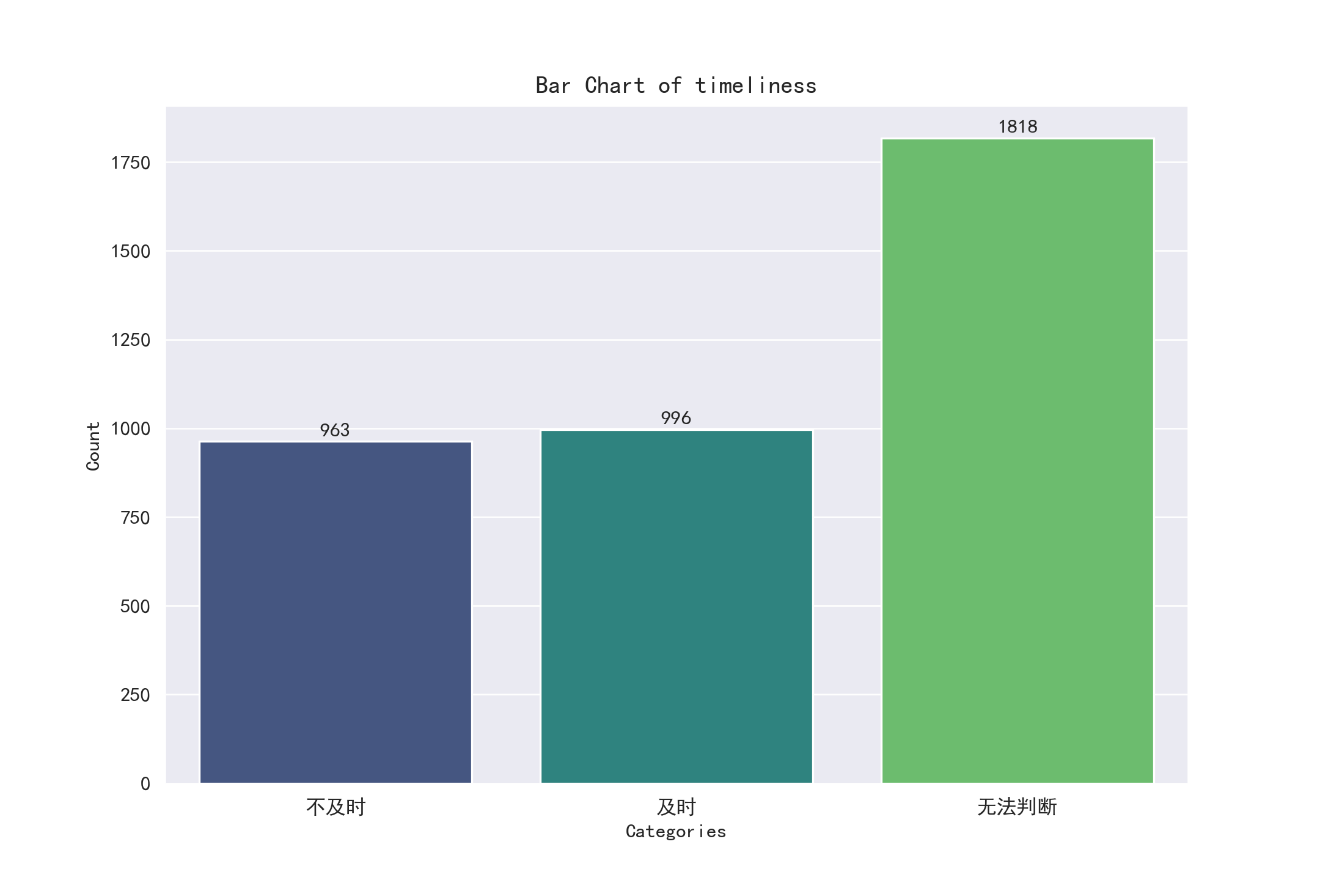


图 4数据集更新及时性统计图

|  |  |  |
| --- | --- | --- |
| 值 | 数量 | 占比 |
| 1 | 2972 | 78.7% |
| NaN | 766 | 20.3% |
| < 1 | 39 | 1.0% |

首先，观察各数据集在结构完整性上的得分（表格1），可以发现在3777个数据集中，有766个空数据集，即数据集文件中除了一行列标题以外什么都没有，占全部数据集的20%，另有39个，约占1%的数据集具有不同程度的结构完整性问题。这反映出数据管理过程中存在的严重问题，其中包括数据录入错误、数据保管不善或数据集更新和维护的疏忽。这种情况不仅降低了数据的可用性，也可能导致对数据的误解，进而影响基于这些数据做出的决策的有效性。

结构完整性不佳的数据集难以被标准的数据处理工具读取和分析。对于数据分析师和数据科学家而言，这意味着在使用这些数据之前，需要投入额外的时间和资源进行数据清洗和预处理工作，这不仅增加了工作量，也可能因为处理不当而引起数据信息的损失。

表格 1结构完整性统计表

|  |  |  |
| --- | --- | --- |
| 值 | 数量 | 占比 |
| 1 | 2267 | 60.0% |
| NaN | 805 | 21.3% |
| 0 | 705 | 18.7% |

表格 2唯一性统计表

|  |  |  |
| --- | --- | --- |
| 值 | 数量 | 占比 |
| 1 | 1385 | 36.7% |
| NaN | 805 | 21.3% |
| < 1 | 1587 | 42.0% |

接着讨论数据的唯一性（表格2）和完整性（表格3）维度，这两个维度对于保证数据集的分析价值尤为重要。在结构完整性得分为1的2972个数据集中，有2267个数据集显示无重复记录，占比达到60.0%，显示良好的唯一性。然而，还有705个数据集存在重复记录问题，占比约18.7%，这些数据集中的重复记录需要通过适当的数据清洗措施进行处理，以确保数据分析的准确性。这说明在这些数据集中，大多数（60%）已经达到了可供数据分析正常使用的标准，但仍有一定比例的数据集需要进一步处理来去除重复记录，以提高数据的准确性和可靠性。

表格 3数据完整性统计表

|  |  |  |
| --- | --- | --- |
| 值 | 数量 | 占比 |
| 1 | 996 | 26.4% |
| NaN | 1818 | 48.1% |
| 0 | 963 | 25.5% |

在数据完整性方面，1385个数据集的记录完全无缺失，占比为36.7%。这表明这部分数据集的数据填充度非常高，可以直接用于高质量的数据分析。然而，还有1587个数据集显示出不同程度的缺失问题，占比42.0%，这些数据集需要进行数据填充或其他预处理工作，以确保分析的完整性。虽然大部分数据集（约78.7%）在结构完整性上表现良好，但在唯一性和数据完整性上仍有改进的空间。特别是数据完整性方面，近一半的数据集存在缺失值问题，这可能对依赖精确数据输入的分析造成影响。

表格 4及时性统计表

最后，关于数据集的及时性，分析结果显示（表格4），这一指标表现不尽如人意。在3777个数据集中，有将近一半的数据集（1818个，占48.1%）无法判断及时性，因为这些数据集的更新频率标记为“不定期”。此外，963个数据集（占25.5%）未能按照预定更新频率进行更新，而仅有996个数据集（占26.4%）能够满足及时性标准。

此数据表明，更新频率较高的数据集（如每日更新和实时更新）较难维持最新状态，可能由于这些数据集需要持续的维护和监控，从而增加了数据管理的复杂性和成本。反观更新频率较低的数据集（如每年更新），由于更新周期较长，通常有足够的时间进行数据处理和审核，从而更易保持数据的及时更新。

在进一步的调查中，对满足和不满足及时性的数据集进行了更新频率的分布分析。结果显示（见图5），即使在符合及时性的数据集中，绝大多数（98%）都采用每年更新。而在更新不及时的数据集中，最常见的更新频率是每日（34%）、实时（25%）、和每月（22%）。这种分布明显地揭示了不同更新需求对数据管理策略的影响。

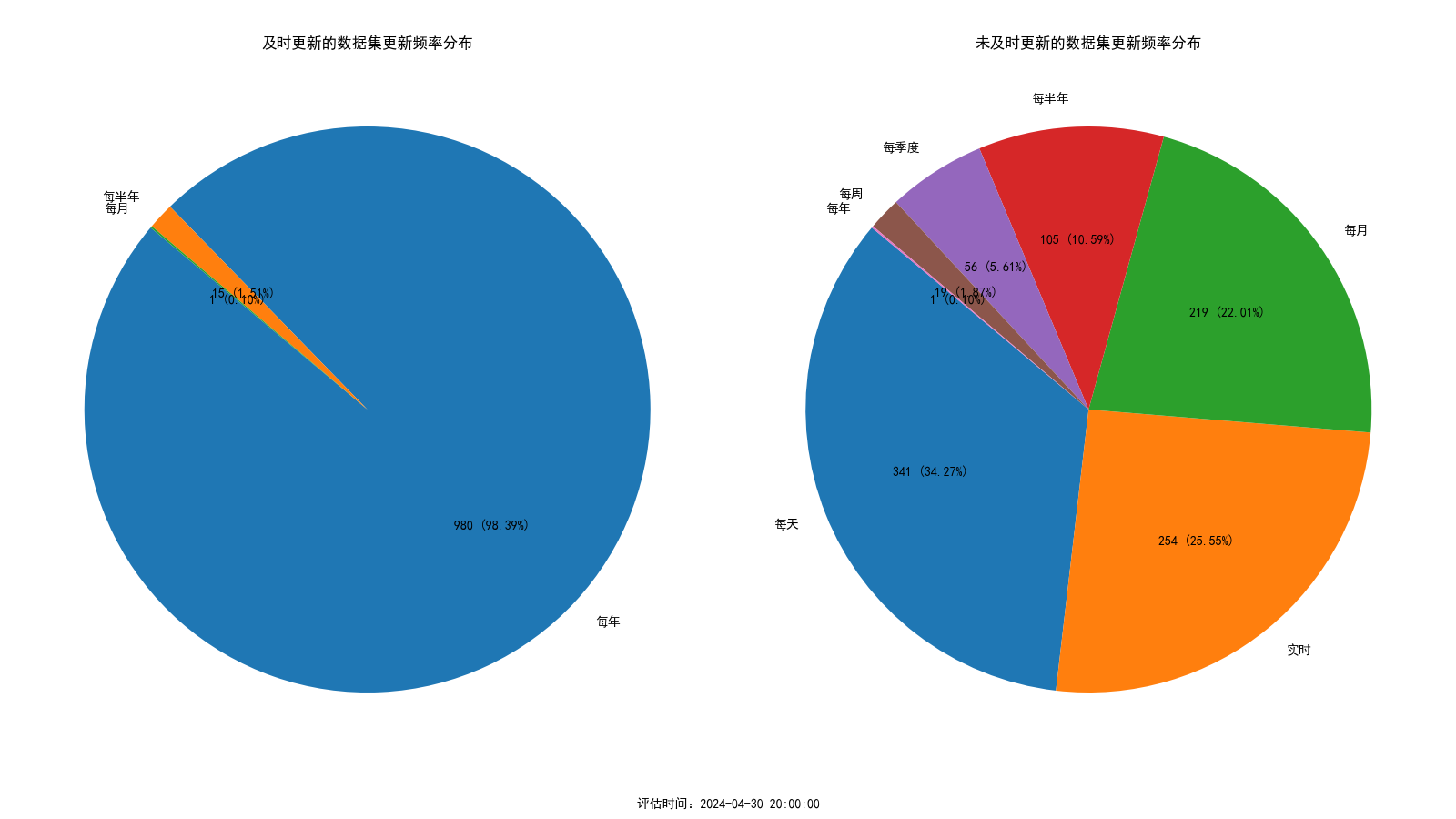


图5数据集及时性下的更新频率分布饼图

## 5. 对策建议

**结构完整性**

* **建立严格的数据验证规则**：实施自动化的数据验证系统，在数据上传至平台前自动检查其结构完整性。通过软件工具检查文件格式、字段完整性和数据一致性，确保所有数据集严格符合组织的数据标准。
* **定期的数据质量审查**：设立定期审查机制，由数据质量团队定期对数据库进行审核，识别并纠正数据质量问题。这包括对数据格式、数据完整性和历史数据的回溯验证。
* **提高数据录入质量**：对数据录入人员进行定期培训，强调数据质量的重要性，并确保他们熟悉并能遵守数据录入标准。实施奖励机制以鼓励高质量的数据输入。

**唯一性和完整性**

* **数据去重**：使用先进的去重技术（如基于哈希表的算法），自动识别并处理数据重复问题，保证数据库中每条记录的唯一性。
* **缺失值处理**：针对数据完整性问题，采用多种技术处理缺失数据，例如利用统计方法、机器学习预测或最近邻插值等方法填补缺失值，或在必要时剔除含有大量缺失值的记录。

**及时性**

* **提高更新监控的自动化**：为频繁需要更新的数据集建立自动化的监控系统，实时监控数据更新状态，确保所有数据集按照既定频率及时更新，减少人工监控负担。
* **优化数据更新流程**：重新评估和优化数据更新流程，简化数据提交和审批流程，增加数据处理资源，并引入高效的技术工具以缩短数据更新周期，提升数据更新效率。

通过实施这些对策，不仅可以解决当前面临的数据质量问题，还可以建立起一种持续关注并改进数据质量的文化和机制。这将为数据驱动的决策提供更坚实的基础，提高决策的效率和准确性，最终推动开放数据建设的整体业务发展。

## 6. 结语

通过对3777个数据集的综合分析，初步了解了数据的当前质量状态，还识别出了结构完整性、唯一性、完整性和及时性等关键质量维度中的普遍问题，积累了对数据质量检测的经验。

接下来的工作还可以往以下方向扩展：

* **数据源**：本次只选取了四川公共数据开放网内支持无条件开放并提供csv文件的数据集，未来还可以拓展到其它开放条件和格式，以及其它数据开放平台内的数据源。
* **数据质量度量指标**：本次实验使用到的指标有结构完整性、唯一性、数据完整性和及时性，希望未来能够实现更多种类的脏数据自动检测。
* **结果分析**：本次实验仅对四个度量指标分别进行了简单统计分析，对及时性进一步调查了更新频率的分布，除此之外仍应有大量信息可供探索发现。

## 附录

表 我国除港澳台外，各省级行政区公共数据平台建设汇总表（最后测试时间2024-05-07）

| 序号 | 省份 | 平台名称 | 平台链接 | 备注 | 最近更新时间 |
| --- | --- | --- | --- | --- | --- |
| 1 | 北京市 | 北京市公共数据开放平台 | <https://data.beijing.gov.cn/> | 无开放类型，下载均需登录，按主题、机构 | 2024-05-07 |
| 2 | 天津市 | 天津市信息资源统一开放平台 | <https://data.tj.gov.cn/> | 无条件开放下载需登录，按主题、市级、区级、综合得分 | 2024-04-30 |
| 3 | 河北省 | 河北省公共数据开放网 | <http://hebdata.hebyun.gov.cn/> | （资源目录）大概率无法下载（提示注册），按主题、按部门 | 2023-11-29 |
| 4 | 山西省 | 山西省公共数据开放网 | <http://data.shanxi.gov.cn/> | **无法下载**，按省直部门、领域筛选、开放类型 | 2023-12-29 |
| 5 | 内蒙古自治区 | 内蒙古自治区公共信息资源开放平台 | <http://open.nmgdata.org.cn/> | 下载需登录，按数据提供单位、主题、开放方式 | 2024-04-12 |
| 6 | 辽宁省 | 辽宁省公共数据开放平台 | <https://data.ln.gov.cn/> | 无条件开放直接下载，有条件开放申请，按领域、格式、开放类型、省直部门、地市、企事业单位，无障碍浏览 | 2024-03-22 |
| 7 | 吉林省 | 吉林·数据开放 | <https://data.jl.gov.cn/> | 登陆下载，按数据提供部门、主题分类、数据格式 | 2024-04-25 |
| 8 | 黑龙江省 | null | null | null | null |
| 9 | 上海市 | 上海市公共数据开放平台 | <https://data.sh.gov.cn/> | 有条件开放申请，无条件开放直接下载，按数据领域、资源类型数据、提供单位、综合得分、其他数据 | 2024-05-07 |
| 10 | 江苏省 | 江苏省公共数据开放平台 | <http://data.jszwfw.gov.cn:8118/extranet/> | 下载需登录，按省级部门、设区市、所属主题、所属行业、资源类型、开放类型；无障碍浏览，老年人专区 | 2024-05-07 |
| 11 | 浙江省 | 浙江·数据开放 | <https://data.zjzwfw.gov.cn/> | 下载需登录，按数据来源单位、领域、标签、格式；网站无障碍，进入老年模式 | 2024-05-07 |
| 12 | 安徽省 | 安徽省数据开放平台 | <http://data.ahzwfw.gov.cn:8000/dataopen-web/index.html> | 下载需登陆，按开放属性、所属领域、所属区域、提供部门 | 2024-01-09 |
| 13 | 福建省 | 福建省公共数据资源统一开放平台 | <https://data.fujian.gov.cn/>‍ | 下载需登录，按领域、格式、开放类型；无障碍浏览，长者模式 | 2024-01-31 |
| 14 | 江西省 | 江西省公共数据开放平台 | <https://data.jiangxi.gov.cn/> | 下载需登录，按部门、领域、开放类型、责任清单 | 2024-04-28 |
| 15 | 山东省 | 山东公共数据开放网 | <https://data.sd.gov.cn/> | 无条件开放直接下载，有条件开放申请，按领域、格式、开放类型、省直部门、地市、企事业单位、综合得分；无障碍通道 | 2024-05-07 |
| 16 | 河南省 | 河南省公共数据运营服务平台 | <http://hndataops.com/> | 只有几十个数据接口，按数据来源、行业领域、服务类型 | 2024-03-08 |
| 17 | 湖北省 | null | null | null | null |
| 18 | 湖南省 | 湖南政务大数据公众门户 | <https://data.hunan.gov.cn/etongframework-web/> | 建设中 | 2024-03-28 |
| 19 | 广东省 | “开放广东”全省政府数据统一开放平台 | <https://gddata.gd.gov.cn/> | 下载需登录，按部门、主题分类、格式、开放模式，无障碍浏览 | 2024-05-06 |
| 20 | 广西壮族自治区 | 广西壮族自治区公共数据开放平台 | <https://data.gxzf.gov.cn/> | 下载需登录，按区直部门、地市、企事业单位、领域、格式、数据格式、开放类型；长者专区 | 2024-04-23 |
| 21 | 海南省 | 海南省政府数据统一开放平台 | <https://data.hainan.gov.cn/> | 下载需登录，按主题、部门、综合得分；网站无障碍，适老版 | 2024-05-07 |
| 22 | 重庆市 | 重庆市公共数据开放系统 | <http://data.cq.gov.cn/> | 下载需登录申请，按数据部门、区县、行业、主题、格式、开放属性 | 2024-05-07 |
| 23 | 四川省 | 四川公共数据开放网 | <https://www.scdata.net.cn/> | 无条件开放直接下载，有条件开放申请，按领域、格式、开放类型、省直部门、地市 | 2024-04-11 |
| 24 | 贵州省 | 贵州省政府数据开放平台 | <http://data.guizhou.gov.cn/> | 下载需登录并订阅，按主题、场景、部门、类型、格式 | 2024-05-06 |
| 25 | 云南省 | 云南省公共数据平台 | <http://106.227.94.96:8894/> | 正在建设中 | null |
| 26 | 西藏自治区 | null | null | null | null |
| 27 | 陕西省 | null | null | null | null |
| 28 | 甘肃省 | null | null | null | null |
| 29 | 青海省 | null | null | null | null |
| 30 | 宁夏回族自治区 | 宁夏公共数据开放平台 | <https://opendata.nx.gov.cn/portal/index> | 名存实亡，没有一个能下载的 | 2022-06-23 |
| 31 | 新疆维吾尔自治区 | null | null | null | null |