**基于Github部分Top300项目的仓库指标的**

**可视化分析报告**

**一、引言**

在开源蓬勃发展的当下，开源项目产生海量数据，这些数据对了解项目动态、社区活跃度及技术走向意义重大。本大作业围绕开源的Top300项目的仓库指标数据，通过系统的数据处理与可视化，将复杂数据转化为直观信息，为开发者、项目管理者等提供决策辅助，推动开源生态发展。

**二、数据选择**

**1. 原始数据概述**

原始数据取自上一届 OpenSODA 挑战赛开放数据集，该数据集为比赛整理，背景与开源发展紧密相关。它聚焦 GitHub Top 300 项目， 即使用 2022 全年全域 OpenRank 全球 Top300 项目，这些项目通常拥有大量的贡献者和用户。这些项目涵盖了各种语言、各种应用场景，是当前最受欢迎的开源项目。

数据集含 2 份日志数据与 1 份指标数据助力探索。本次选用了 Top 300 仓库 OpenDigger 指标数据，下载后为大文件夹，内含 300 个项目文件夹，每个项目有 1 - 2 个子文件夹对应仓库，子文件夹里是指标的 JSON 源文件，涵盖项目活跃度（日、月）、关注度、巴士系数、PR 相关（数量、生命时长、响应时间等）、代码变更（添加、删除、总变更行数）、贡献者相关（邮箱后缀、新老贡献者数量与名单）、issue 相关（数量、评论数、生命时长、响应时间）、OpenRank 值、参与者人数、收藏量、fork 数量等指标，为后续分析提供多维度数据基础。

原始数据网址：**https://xlab2017.yuque.com/staff-kbz9wp/olpzth/tq36xvyzg9b880hy?singleDoc#**

1. **本次特有的数据选择**

鉴于每个项目可能关联多个仓库，且部分指标值存在缺失风险，为确保最终可视化成果具备高度一致性与可靠性，对项目实施了严格筛选流程。最终筛选出了204个项目，详情可见数据处理。

基于研究与应用需求，从原始数据选关键指标数据，包括 stars（收藏量）、technical\_fork（fork数）、participants（参与者数量）、openrank（项目综合排名相关值）、activity（活跃度）、attention（关注度）、change\_requests（变更请求数，即 PR 总数）、code\_change\_lines\_sum（代码改动总量）等。

选这些数据因它们从不同角度反映开源项目状况：stars 直观体现项目受喜爱程度；fork 数量表明项目可拓展性与认可度；参与者数量反映社区规模；openrank 助评估项目相对地位；activity 和 attention 追踪项目动态变化；PR 总数及code\_change\_lines\_sum 关联项目开发推进情况。综合分析这些数据，能洞察开源项目发展态势。

**三、可视化工具选择**

选可视化工具时，调研、试用多款主流工具后，选定 Dataease 桌面版。其优势明显：一是数据连接便捷，能对接 MySQL 数据库，在配置界面输入主机、用户、密码、数据库名等参数，快速将处理好的数据表转为可视化数据集，提升效率；二是可视化组件多样，有条形图、折线图、饼图、面积图、双线图等，满足不同数据指标与分析需求，充分展现数据内涵；三是操作方便，采用可视化流程，新手易上手，通过简单拖拽、配置操作，就能完成复杂图表创建与布局调整，便于推进项目。

**四、数据处理、分析**

**1. 数据处理思想**

数据处理按模块化、结构化原则，将原始数据清洗、转换、整合为可用素材。针对项目，在前期数据准备阶段，为应对数据来源的复杂性，即每个项目可能对应多个仓库且部分指标值存在缺漏的问题，编写专门的筛选程序。针对不同类型指标数据设计专属流程。对时间序列数据，统一时间格式、聚合不同时段数据；对 JSON 格式复杂数据，解析、扁平化处理，提取嵌套数据为独立字段以便统计。同时，以项目名称等关键字段关联不同表数据，避免错配、重复统计，构建逻辑严密的数据体系。   
 详情的数据处理请见数据处理代码。

**2. 数据分析的意义与选择缘由**

选上述关键指标处理，因其各有独特信息价值，构成反映开源项目全貌的信息网。从项目生命周期看，stars、fork、participants 等在项目起步、成长阶段重要，高收藏量吸引开发者加入，fork 行为推动项目传播，参与者增多助力迭代。项目稳定期，openrank、activity、attention 等衡量健康状况与竞争力，openrank 反映项目行业地位及变化，activity 与 attention 揭示项目热度与关注变化。在项目迭代微观层面，PR 总数、code\_change\_lines\_sum 把控开发节奏与质量，PR 总数反映开发者参与积极性，code\_change\_lines\_sum 量化代码改动规模。综合处理分析这些数据，能全面了解开源项目成长轨迹、发现问题、预测走向。

**五、可视化设计**

可视化设计依据视觉传达原理，打造高效信息传递空间。整体布局合理：上方正中间标题点明大屏主题“部分 Top300 项目仓库指标洞察大屏 2018 - 2023”及时间跨度，助用户聚焦。下方 4 张指标卡对应 stars、fork、PR、参与者等核心指标总量，卡片设计简洁，用鲜明色彩（如蓝色示收藏量、黄色示 fork 数等）、合理字体搭配突出关键数据，方便用户识别。左侧两张条形图展示收藏量、参与者排行前十项目，条形图遵循视觉比较原则，统一宽度、精确高度比例，标注项目名与坐标轴刻度，方便用户直观对比项目差异。右中折线图呈现 openrank 值随时间变化，折线颜色清新醒目，坐标轴刻度均匀、标注简洁，便于用户追踪 openrank 升降轨迹。最右侧面积图展示每月代码修改总量，利用色彩填充与层次感强化数据累积效应，配合下方时间轴，展现代码开发进程。最下方双线图结合柱状图呈现每月活跃度与关注度，图例可交互，用户点击可隐藏或显示线条，坐标轴动态调整，满足个性化解读需求。大屏背景用稍暗蓝色营造科技氛围，各图边框依内容特色配色，强化视觉区分，引导用户视线，提升信息传递效率。

1. **可视化的实际实现**

在 Dataease 桌面版构建可视化：先连接 MySQL 数据库，在配置界面准确输入主机、用户、密码、数据库名等参数，引入前期数据表，转为可视化数据集。接着按布局创建图表。对指标卡，从组件库拖曳数据卡片组件至指定位置，通过数据绑定关联对应总量数据表字段，如将 stars\_total\_metrics 表中 total\_stars 字段绑至收藏量指标卡数据区域，再在样式编辑面板调字体、颜色、背景等参数优化外观。创建条形图时，深入条形图组件设置，从收藏量、参与者总数相关表选项目名与对应指标数据，分别设为横轴、纵轴数据源，调整条形颜色并优化轴标签字体、角度，确保清晰可读。折线图、面积图、双线图构建精细，以折线图为例，从处理后的 openrank 时间序列数据表提取项目名、时间、openrank 值等关键列，将时间与 openrank 值分别绑至横轴、纵轴，精心设折线颜色、粗细、标记样式，精细打磨坐标轴范围、刻度、标签，依据数据最值设刻度间隔，加简洁标注，确保用户能清晰追踪 openrank 变化。面积图与双线图类似处理。最后，利用 Dataease 交互功能，为图表加鼠标悬停提示，关联数据点与显示逻辑，鼠标滑过图表点时弹出含横轴、纵轴详细数据的提示框；在左上方查询组件区域绑项目列表数据源，设关联关系，用户选不同项目时，除左侧两张排行图，其余图表实时更新数据，呈现所选项目全方位信息，让可视化大屏具交互性与动态性。

**七、实际应用**

**一、开源化促进知识共享**

本大作业成果全面开源，原始数据、数据处理代码、Dataease 中数据集、最终展示大屏模板均传至 github 和 Dataease。在 github 上，用户可借版本管理与社区协作功能，回溯数据历史版本、看代码迭代轨迹、参与讨论优化方案；在 Dataease 平台，用户能在线体验可视化大屏，下载数据集个性化分析。凭详细注释与文档，新手和专业人士都能依原始资料复现项目，实现知识无障碍传播与再创造，为开源生态添活力。

**二、助力项目管理精细化**

对项目管理者，可视化大屏是精准监控工具。通过观察收藏量、参与者数量，可快速评估项目在开源社区人气与人力汇聚情况，指标不佳时，及时反思推广、运营策略；从代码改动总量、PR 总数动态变化，精准把控开发节奏，若异常，排查技术难题或协调团队协作；依据 openrank 及其趋势图，看清项目行业竞争地位与发展走向，排名下滑时，对标竞品、制定赶超策略，确保项目稳健前行。

**三、助力开发者技能提升与项目选择**

开发者可借助本大作业的可视化成果，深入了解 Top300 项目的各项关键指标。通过分析 stars、fork、participants 等数据，发现热门项目背后的用户需求与技术趋势，从而有针对性地选择参与那些与自身技能提升方向相契合的开源项目。例如，若某一项目在 stars 和 fork 数量上持续增长，且参与者数量众多，说明该项目具有较高的技术价值和社区认可度，开发者参与其中能够接触到前沿技术，与优秀开发者交流学习，快速提升自身技术水平。同时，开发者还可以参考本大作业中的数据处理和可视化方法，将其应用到自己的项目中，优化项目的数据管理和展示效果，增强项目的吸引力和竞争力。

**四、推动开源项目研究与优化**

行业研究人员可以利用本大作业提供的 Top300 项目仓库指标可视化大屏，深入研究开源项目的运行规律和发展模式。通过对不同项目在 stars、fork、openrank、activity 等指标上的对比分析，挖掘出影响项目成功的关键因素，为开源项目的优化提供理论依据。例如，研究发现某些项目在特定时期内 stars 和 fork 数量大幅增长，而活动度和关注度也相应提高，研究人员可以进一步探究这一现象背后的原因，是项目发布了重大更新、解决了关键问题，还是受到了媒体关注等。这些研究成果不仅可以帮助项目管理者更好地制定项目发展战略，还可以为开源社区的建设和发展提供参考，促进开源项目的整体质量和影响力提升。