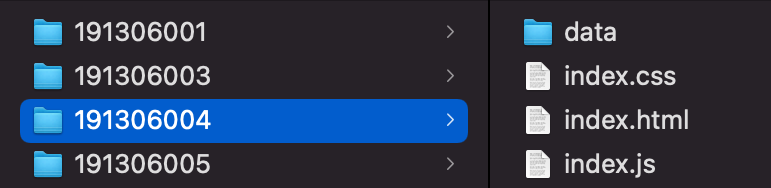
《数据可视化技术》课程报告

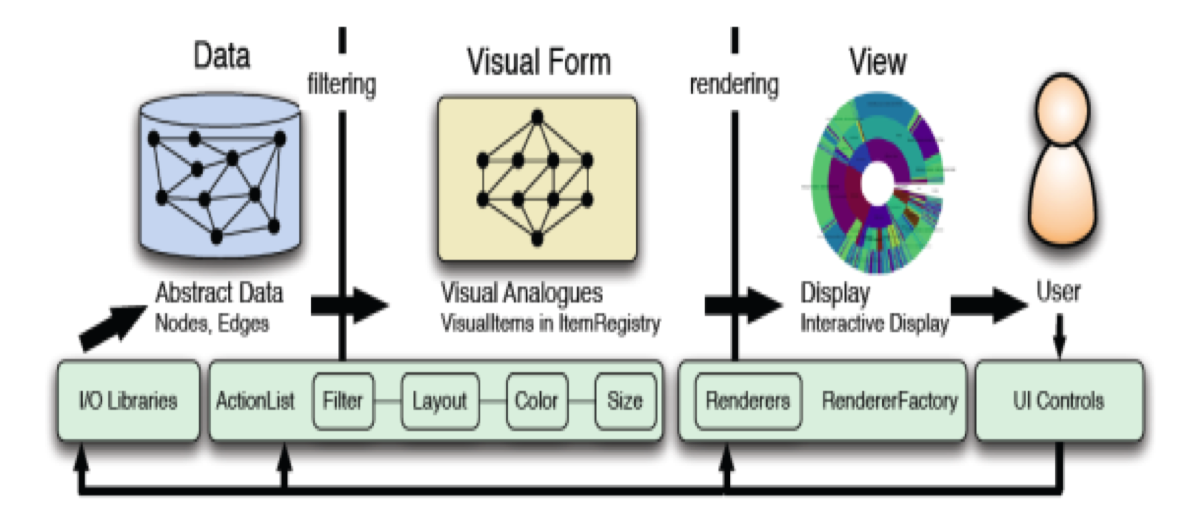
|  |  |  |  |
| --- | --- | --- | --- |
| 数 计 学院 | 软件工程 专 业 | | 21 级 班 |
| 学 生 姓 名 | 蒋天宇 | 学 号 | 211308121 |
| 指 导 教 师 | 王诗莹 | 职 称 | 硕 士 |
| 完 成 日 期 | 2024 年06月25日 | | |

**教务处 制**

**《数据可视化技术》课程设计实施与评价规则**

1. 本次课程设计由每人独立完成设计，任课老师将考量设计与他人或线上作品的相似度，**涉嫌抄袭的作品将以0分处理**，考量标准任课教师保留解释权。
2. 本次课程设计必须使用 D3.js 框架作为项目的技术手段，**不采用D3.js 框架的作品将以0分处理**，**如有特殊情况请向任课教师申请，**任课教师保留最终解释权。
3. 本次课程设计建议使用WebStorm作为项目的编辑器。
4. 本次课程设计包括9个课程实验报告，完成时间为1周。请在教室中完成资料查找与编码，务必按时上交单次报告，**逾期不接受补交**。
5. 本次课程设计最终提交内容包括如下清单：
6. 1. 项目作品打包（**源代码**）命名规则：学号.rar 如 201122001.rar
7. 2. 本实验报告（完整填写）命名规则：学号.docx如 201122001. docx
8. 本次课程设计项目（**源代码**）必须包含4个文件：index.html、index.js、index.css（任课教师提供模版）、所用data文件、其他相关项目文件，项目结构可以参考如下：
9. 
10. 本次课程设计的**评分比例**如下：
11. 1. 项目评分 50 %
12. 2. 实验报告评分 40 %
13. 3. 课堂表现 10 %
14. **以上评分若有一项为0分，课程设计总分将直接划为0分。**
15. 本实验报告的模板原内容仅作为提醒，并非硬性要求，根据自身项目完成即可。报告内容需逻辑清晰、注意格式、作品正确且美观。
16. 本次课程设计最终评价阶段将通过课程答辩方式，个人**制作PPT**进行演讲答辩、展示项目，限时**3分钟**以内。务必参加答辩，评判标准任课教师保留解释权。

**数据可视化流程：**



**数据可视化设计规范：**

[数据可视化设计规范指南](https://m2.material.io/design/communication/data-visualization.html)

课程实验报告 1 ： 数据采集

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 | 蒋天宇 | 学 号 | 211308121 |
| 1. **数据来源：Kaggle**    1. **world-economic-data.csv**   世界1960 - 2022年GDP数据，包括GDP增长率、个人GDP、失业率共9个字段，这些字段提供了对一个国家经济方面的洞察，包括个人汇款流入、失业率以及GDP及其增长的不同衡量标准：  地址：<https://www.kaggle.com/datasets/mittvin/world-economic-indicators-1960-2022-dataset>   * 1. **world-terrorist-data.csv**   世界1970 - 2023年恐袭数据，包括恐袭次数、恐袭类型、死亡人数等21个字段：  地址：<https://www.kaggle.com/datasets/rafsunahmad/terrorist-attacks-data-since-1970-2023>   * 1. **world-unemployment-data.csv**   世界2014 - 2024年失业数据，包括性别分组、年龄分组、失业率等16个字段，这一数据集从2014年到2024年，提供了各种人口参数下失业的综合情况。重点在于基于国家、年龄段和性别的数据剖析，提供对影响全球劳动力市场的各种因素的洞察：  地址：<https://www.kaggle.com/datasets/sazidthe1/global-unemployment-data>   * 1. **world-population-data.csv**   世界1970、1980、1990、2000、2010、2015、2020、2021、2022、2023年的人口数据，该数据集提供了全球国家和地区的全面历史人口数据，提供了对各种参数的洞察，如面积大小、大陆、人口增长率、排名和世界人口百分比等17个字段。从1970年到2023年，它包括不同年份的人口数字，从而能够详细检查人口趋势和随着时间的变化：  地址：<https://www.kaggle.com/datasets/sazidthe1/world-population-data>   1. **原始数据读取代码截图：**   使用await在异步环境下等待读取所有数据集，并保存在不同的变量中。其中world-110m.json是geo json格式，用来绘制世界地图，而world-110m-country-codes.json是保存所有国家名与国家id，可以一一对应出总共173个国家。     1. **原始数据各项属性分析：**   **3.1. world\_economic\_data.csv**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **属性名** | **字段类型** | **取值范围** | **是否使用** | **备注** | | Country Name | string | 国家英文名称 | 是 | 国家名称 | | Country Code | string | 唯一的三个字母 | 否 | 国家代码 | | Year | int | 1960-2022 | 是 | 年份 | | Personal remittances, received (% of GDP) | double | 0-235.9279913 | 是 | 个人GDP | | Unemployment, total  (% of total labor force) | double | 0.095-38.8 | 是 | 失业率 | | GDP (current US$)\_x | double | 8824746.238-100562000000000 | 是 | 国家GDP | | GDP growth (annual %)\_x | double | -64.04710697-149.9729635 | 是 | 国家GDP增长率 | | GDP (current US$)\_y | double | 8824746.238-100562000000000 | 否 | 国家GDP（方案2） | | GDP growth (annual %)\_y | double | -64.04710697-149.9729635 | 否 | 国家GDP增长率（方案2） |   **3.2. world\_unemployment\_data.csv**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **属性名** | **字段类型** | **取值范围** | **是否使用** | **备注** | | country\_name | string | 国家英文名称 | 是 | 国家名称 | | indicator\_name | string | 指标英文名称 | 是 | 指标名称 | | sex | string | Female / Male | 是 | 性别 | | age\_group | string | 无 | 是 | 年龄组 | | age\_categories | string | Youth / Adults / Children | 是 | 年龄分类 | | 2014 | double | 0.027-74.485 | 是 | 2014年失业率 | | 2015 | double | 0.034-74.655 | 是 | 2015年失业率 | | 2016 | double | 0.038-74.72 | 是 | 2016年失业率 | | 2017 | double | 0.035-75.416 | 是 | 2017年失业率 | | 2018 | double | 0.044-76.395 | 是 | 2018年失业率 | | 2019 | double | 0.036-77.173 | 是 | 2019年失业率 | | 2020 | double | 0.056-83.99 | 是 | 2020年失业率 | | 2021 | double | 0.064-82.135 | 是 | 2021年失业率 | | 2022 | double | 0.067-78.776 | 是 | 2022年失业率 | | 2023 | double | 0.063-78.541 | 是 | 2023年失业率 | | 2024 | double | 0.06-78.644 | 是 | 2024年失业率 |   **3.3. world\_terrorist-data.csv**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **属性名** | **字段类型** | **取值范围** | **是否使用** | **备注** | | Entity | string | 国家英文名称 | 是 | 国家英文名称 | | Code | string | 唯一的三个字母 | 否 | 国家代码 | | Year | int | 1970-2021 | 是 | 年份 | | Terrorist attacks | int | 0-16820 | 是 | 空袭数 | | Terrorism deaths | int | 0-44576 | 是 | 空袭死亡数 | | Attack method: Hijacking | int | 0-58 | 是 | 恐袭方式 | | Attack method: Hostage Taking (Barricade Incident) | int | 0-85 | 是 | 恐袭方式 | | Attack method: Unarmed Assault | int | 0-106 | 是 | 恐袭方式 | | Attack method: Facility/Infrastructure Attack | int | 0-776 | 是 | 恐袭方式 | | Attack method: Hostage Taking (Kidnapping) | int | 0-1409 | 是 | 恐袭方式 | | Attack method: Assassination | int | 0-1007 | 是 | 恐袭方式 | | Attack method: Armed Assault | int | 0-4037 | 是 | 恐袭方式 | | Attack method: Bombing/Explosion | int | 0-8739 | 是 | 恐袭方式 | | Death Age 100+ | int | 0-29 | 是 | 各段年龄死亡数 | | Death Age: 51-99 | int | 0-52 | 是 | 各段年龄死亡数 | | Death Age : 21-50 | int | 0-275 | 是 | 各段年龄死亡数 | | Death Age : 11-20 | int | 0-522 | 是 | 各段年龄死亡数 | | Death Age : 6-10 | int | 0-853 | 是 | 各段年龄死亡数 | | Death Age : 1-5 | int | 0-6612 | 是 | 各段年龄死亡数 | | Terrorist Death Type : Suicide | int | 0-10337 | 是 | 死亡类型 | | Terrorist Death Type : Killed | int | 0-38627 | 是 | 死亡类型 |   **3.4. world\_population\_data.csv**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **属性名** | **字段类型** | **取值范围** | **是否使用** | **备注** | | rank | int | 1-234 | 否 | 国家队列 | | cca3 | string | 唯一的三个字母 | 否 | 国家代码 | | country | string | 国家英文名称 | 是 | 国家英文名称 | | continent | string | 大洲英文名称 | 否 | 大洲英文名称 | | 2023 population | int | 518-1428627663 | 是 | 各国家人口数 | | 2022 population | int | 510-1425887337 | 是 | 各国家人口数 | | 2020 population | int | 520-1424929781 | 是 | 各国家人口数 | | 2015 population | int | 564-1393715448 | 是 | 各国家人口数 | | 2010 population | int | 596-1348191368 | 是 | 各国家人口数 | | 2000 population | int | 651-1264099069 | 是 | 各国家人口数 | | 1990 population | int | 700-1153704252 | 是 | 各国家人口数 | | 1980 population | int | 733-982372466 | 是 | 各国家人口数 | | 1970 population | int | 752-822534450 | 是 | 各国家人口数 | | area (km虏) | double | 0.44-17098242 | 是 | 地区面积 | | density (km虏) | int | 0-21403 | 是 | 人口密度 | | growth rate | double | -7.45%-4.98% | 是 | 人口生长率 | | world percentage | double | 0.00%-17.85% | 是 | 世界人口占比 |  1. **可视化数据筛选相关代码截图：**   使用delete函数删除表中所不需要的列： | | | |

课程实验报告 2 ： 可视化方案构建

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 | 蒋天宇 | 学 号 | 211308121 |
| 1. **设计草图：**    1. **主页面**      * 1. **数据统计页面**      * 1. **数据分析页面**      1. **设计方案描述**    1. **主页面**   此页面用于存放整个项目的主页面，上面可以通过日期选择框选择需要生成数据的年份，点击生成数据按钮就可以在左图中显示出世界地图，并使用颜色的深浅来显示各个地区数据的大小，数据统计中可以按顺序显示前十名的国家以及其GDP总值。在第一次进入页面时，左图默认显示2022年全球存在数据的各个国家GDP总值，也就是数据中的最后一个年份的数据。图中地图通过国家分割，并且每个国家都可以点击跳转至数据统计页面。   * 1. **数据统计页面**   此页面用于存放所点击的国家的国家人口大小、国家人口失业率、国家遭遇恐袭次数等数据，以条形图或折线图的方式存在。点击数据统计页面中的数据分析按钮，即可跳转至数据分析页面。   * 1. **数据分析页面**   此页面用于存放一种可能原因的数据与国家GDP总值的对比，将两者放在同一张图上，并用不同的颜色描绘出来。通过对两条线在图上的分布显示，可以观察出是否存在线性回归等特性，这样就可以很明显地看出所选择的数据是否与国家GDP总值呈现相关性。 | | | |

课程实验报告 3 ： 数据清理

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 | 蒋天宇 | 学 号 | 211308121 |
| 1. **数据错误查询：**   在数据集中，有时不可避免会出现错误的数据，比如出现负数，因此我们对数据进行了检测，代码如下：      并且得到结果如下：    这个结果说明两个数据集均未出现负数异常。   1. **数据缺失处理：**   由于在首页中需要默认显示2022年全球173个地区的GDP总值，我们需要获取的数据有：   1. 国家名称 2. 年份 3. GDP总值   根据整个数据集来看，我们需要手动来切割数据集，在里面挑出我们需要的数据，也就是2022年的世界各个国家的GDP总值，代码如下：    在控制台中打印出我们的数据，可以得到以下结果：    其中不仅包括了国家的GDP总值，还包括了一些地区的GDP总值，包括世界、中亚、欧洲等一些未被world-110m所收录的地区，所以这些是多余的数据，我们需要剔除。  由于数据量太大，我们使用国家对应GDP的办法，先筛出2022年的所有数据，再根据173个地区名对应数据，得到一份国家名-GDP总值-id的数据：    但是我们发现数据只有148条，在173个地区中是有未统计的数据，包括北极地等一些得不到数据的地区。这些数据无法通过原有数据的基础上推测或分析出来，因为原有数据并不存在一个一定的趋势，所以我们承认这个数据缺失的情况出现，并通过图例以不同颜色的形式在地图上表示出来。  其他数据集也同以上方式处理，由于这些数据不存在趋势，因此我们只能承认数据存在确实的情况，并在图上用图例的方式说明，这一点我们在下面进行讨论。  **3. 数据缺失处理**  以GDP总值为例，我们对数据进行了统计分析，并求了以下数据：   1. 最值 2. 平均值 3. 中位数 4. 标准差   我们实现了求这些值的方法，并打包成一个math-utils.js文件，代码如下所示：    结果如下所示：    可以得出结论：   1. 标准差巨大，各个国家的GDP总值差异非常大，波动也非常大。 2. 中位数远低于平均数，说明GDP总值有的非常高，但是数量非常少 3. 总体来看，大部分国家GDP总值较低，只有极少数国家GDP总值非常高   同样，在人口数据集上也有类似的分布：  代码如下所示：    结果如下所示：    通过这两种数据可以观察到，世界GDP总值分布和世界人口分布非常散乱，差异也非常大，因此在数据缺失时，我们无法通过常规情况下根据趋势推测出数据，因此我们检测离群点也毫无意义。因为无法恢复数据，所以我们承认数据的丢失，并将会在之后的图中以图例或其他的方式展现出来。 | | | |

课程实验报告 4 ： 数据过滤和筛选

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 | 蒋天宇 | 学 号 | 211308121 |
| 1. **数据归一化**   在数据处理过程中，我们猜测人口大小与国家GDP总值呈现一定的关系，那么我们需要将国家人口规模大小与国家GDP放在同一个折线图中进行展示，表现在时间的流逝中，人口的变化和GDP的变化之间是否呈现相关性；或者我们将国家人口规模大小与国家GDP增长率放在同一个图中，来表现随着时间的流逝，人口的变化与GDP的变化幅度之间是否呈现相关性。  那么我们就遇到了一件问题，两种数据的范围是完全不相关的，也就是说，两幅图画出来不会有任何的相交，甚至差距非常远，这非常不利于我们观察图像。于是，我们使用了对数据进行归一化的操作：    线性函数将原始数据线性化的方法转换到[0 1]的范围，这样我们就可以把横坐标定为年份，纵坐标定位相同的0-1范围，代码和效果如图：  代码（xValue是x轴数据，yValue是y轴数据，我们对y轴进行归一化操作）：    图 1：数据归一化代码  效果（左边是归一化前，右边是归一化后）：    图 2：数据归一化前 图 3：数据归一化后  这边以中国人口数据为例子，可见两幅图画的变化基本一致，因此简单判断归一化可行，并且将大数据范围变为0-1的小数据范围，图像呈现也会更好。   1. **数据筛选**   数据集中的数据总是那么的多，我们不可能直接就这样显示在图中，因此我们需要对数据进行筛选，筛选出我们想要呈现的。  在这个中国的例子中，我们只需要用到人口数量的归一化数据与年份就可以了，我们可以通过JavaScript中简单的函数对数据进行处理即可，具体过程我们在下文中阐述。   * 1. **原数据分析**   首先，我们先打印出原数据，并对原数据进行分析，确定我们需要的数据或者需要添加的数据。原数据应该是从world-population-data.csv这个文件中读取出来的，打印结果如下图所示：    图 4：原数据  由图可见，这是一个很多对象的数组，一个对象就代表一条数据。这个分布显然无法提供我们使用，我们决定将数据集改变为以 {年份，人口规模，国家} 的对象形式呈现。  代码：    图 5：重写数据结构  效果：    图 6：重写后数据结构  可以看到，这个已经是我们想要的效果了。  对数据结构的处理，改变了数据的排列方式，这步操作非常有意义，对我们接下来的处理迈出了至关重要的一步。   * 1. **数据处理**   根据2.1得到的数据，我们需要对其进行数据处理，因为有些数据并不是描述一个国家，而是一个地区，比如说亚洲、欧洲等，我们需要通过world-110m-country-codes.json这个数据集过滤出存在的国家，对于两个不同的数据集，我们只取交集部分。  代码如下：    图 7：处理world-country-codes数据集    图 8：添加id数据段  效果如下：    图 9：添加后数据结构  这个结构差不多就是我们需要的结构，由多个对象构成，每个对象包含国家、国家id、人口大小、年份四个字段。同时可以很明显地看到，数据从原来的2106条变成了1350条，说明在处理过程中已经自动排除了一些我们不需要或者错误的数据。   * 1. **数据提取**   接下来我们可以开始处理业务逻辑了，在一幅折线图中，我们只需要根据国家来呈现在不同年份下人口的变化。  代码（以中国为例）：    图 10：根据国家名筛选数据  效果：    图 11：中国人口数据  到目前为止，我们已经成功地提取到我们想要的数据，这是一个通过国家名进行查询的方法，我们也可以通过别的字段进行提取，只需要修改部分代码即可。下一步，我们将要把数据放在我们需要的svg图中进行显示即可。 | | | |

课程实验报告 5 ： 视觉编码

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 |  | 学 号 |  |
| 视觉编码的设计是指如何使用位置、尺寸、灰度值、纹理、色彩、方向、形状等视觉通道，以映射我们要展示的每个数据维度。背景知识 – 视觉编码：<https://zhuanlan.zhihu.com/p/24835448>  视觉编码总体描述：（文字+代码截图） | | | |

课程实验报告 6 ： 图例&标签

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 |  | 学 号 |  |
| 简介：在简单图表中，可以使用直接标签。在密集的图表（或更大的图表组的一部分）中，可以用图例。    项目中使用图例&标签场景描述：（文字+实现截图） | | | |

课程实验报告 7 ： 布局

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 |  | 学 号 |  |
| 数据可视化中，好的布局方式，往往能带来舒服的视觉效果。项目中，通过改变可视化部件的位置、大小、角度来获得我们需要的效果  背景知识：<https://blog.csdn.net/github_36904248/article/details/73342373>  布局的设计描述：（文字+实现截图） | | | |

课程实验报告 8 ： 交互

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 |  | 学 号 |  |
| 常见的交互方式包括：  1.滚动和缩放  2.颜色映射的控制  3.数据映射方式的控制：这个是指用户对数据可视化映射元素的选择，一般一个数据集，是具有多组特征的，方便用户按照自己感兴趣的维度去探索数据背后的信息。  4.数据细节层次控制：比如隐藏数据细节，hover或点击才出现。  背景知识：<https://zhuanlan.zhihu.com/p/49528045>  <http://www.qiutianaimeili.com/html/page/2020/03/8rqxzs4fnl.html>  <https://www.jianshu.com/p/c49dfb38fc0d>  项目中使用交互场景描述：（文字+代码截图+实现截图） | | | |

课程实验报告 9 ： 项目评价和改进

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 |  | 学 号 |  |
| 项目最终效果呈现：（文字+实现截图）  项目优化描述： | | | |