项目报告

一、项目背景描述

本项任务旨在探讨大五人格个性心理特征，与自尊、生活满意度、主观幸福感、身体健康等生存发展适应性指标之间的历史-社会生态因素动态关联性。任务主要基于心理学（大五人格理论）与社会学（宏观社会生态因素，如经济、政治、文化、人口等宏观社会结构）领域，并探讨宏观社会生态环境因素对大五人格特质的影响问题。

其中，大五人格理论是当今心理学中最广为人知和应用的个性特质理论之一。该理论认为，人格可以被归纳为五个基本维度：

```
外向性(Extraversion)
    反映个体在社交、活跃度、积极情绪等方面的特点
    高外向性者通常更社交、乐观、主动
    低外向性者则更内向、冷淡、安静

宜人性(Agreeableness)
    反映个体在友善、同情、合作等方面的特点
    高宜人性者更友善、信任、合作
    低宜人性者则更自我中心、对抗、独立

尽责性(Conscientiousness)
    反映个体在有条理、自律、负责等方面的特点
    高尽责性者更勤勉、有条理、可靠
    低尽责性者则更随意、无序、缺乏自律

神经质(Neuroticism)
    反映个体在情绪稳定、压力应对等方面的特点
    高神经质者更易焦虑、易怒、情绪化
    低神经质者则更平稳、冷静、自信

开放性(Openness to Experience)
    反映个体在创造力、想象力、好奇心等方面的特点
    高开放性者更富有创造力、喜欢新事物
    低开放性者则更保守、常规
```

其中除了神经质是负面性质外，其他均为正面性质

二、问题定义

问题定义为，宏观的社会生态因素对大五人格特质是否有影响，以及会产生怎样的影响。

需要完成的目标有：

- 提供基于传统心理测量法或大数据分析所得关键研究指标的信效度信息。
- 收集和整理典型的历史-社会生态宏观指标数据，并对其进行必要的降维简化处理。
- 对特定的亲和-匹配适应性模式进行较深入分析，并做出归纳总结。
- 对关键的分析过程或结果的模式，进行直观可视化展示。

三、技术方案

我们将解决方案分为以下几个步骤：

1. 寻找数据集

由于宏观社会生态因素涉及方面过多，本次任务将这方面聚焦于包括宏观/微观的经济（收入、就业）、环境（社区、教育、生态环境、安全）、健康（生活满意度、健康状况）三个方面。

对于人格特质方面，我们将选择五个国家（美国、加拿大、澳大利亚、印度、英国）的大五人格数据集，同时寻找相应国家的社会生态因素数据集。

2. 初步分析

找到数据集后，进行数据清洗（传统的去空，去重，异常值检测）和降维处理，将数据集中的数据进行标准化处理。

同时需要将人格数据集与社会因素数据集进行两表连接，进行数据的初步分析，包括关键指标的信效度信息分析，聚类分析。

3. 可视化分析

将经过处理的数据作为数据源制作可视化图表大屏，分析寻找社会生态因素对大五人格特质的影响关系并进行直观展示。

四、实验设置

以下是详细的实验过程：

1. 数据集

详细的数据集描述信息参见`dataset.pdf`，包括寻找到的两个原始数据集（人格数据集、社会因素数据集），以及实验过程中的产出数据集。

2. 初步分析

数据处理，代码详见src/country_select.py,new_output_example.py

原有人格数据集有很多单个的国家数据，没有具体分析意义。country_select.py筛选出人格数据集中数量最多的五个国家进行人格数据分析。
对两个人格与地域数据集进行分割连接，具体以Canada为例见new_output_example.py。由于国家数据集和人格数据集的国家表示格式不同只能用手动连接

pca降维处理，代码详见src/pca.py

将处理连接后的数据集进行pca降维处理后可视化展示

因子分析，代码详见src/relief.py

对数据进行因子分析聚类，得出经度（也就相当于国家）在总体人格数据集上起到最大的影响作用

克隆巴赫系数，代码详见src/relief.py

克隆巴赫系数体现了数据内部信度一致性。

聚类分析，代码详见src/k-means.ipynb文件。

对原始人格数据集进行数据清洗后挑选出数据集中50列和人格相关的数据集，在该数据基础上构建k-means模型将所有数据分为5类，其中kmeans模型为自行实现。该模型同时能够输入个人的人格数据对人格进行预测。
之后对五个人格维度分别聚合显示五类人格类型分布并可视化每个cluster中五个人格维度数据的表现。
使用人格相关的数据构建二维PCA模型并可视化聚类结果。

人格-全球类数据集，代码详见src/connect.py文件。

对以'output_'前缀的五个国家数据集进行数据清洗和整理，去除实验无关指标，其他指标以整体平均值为准。

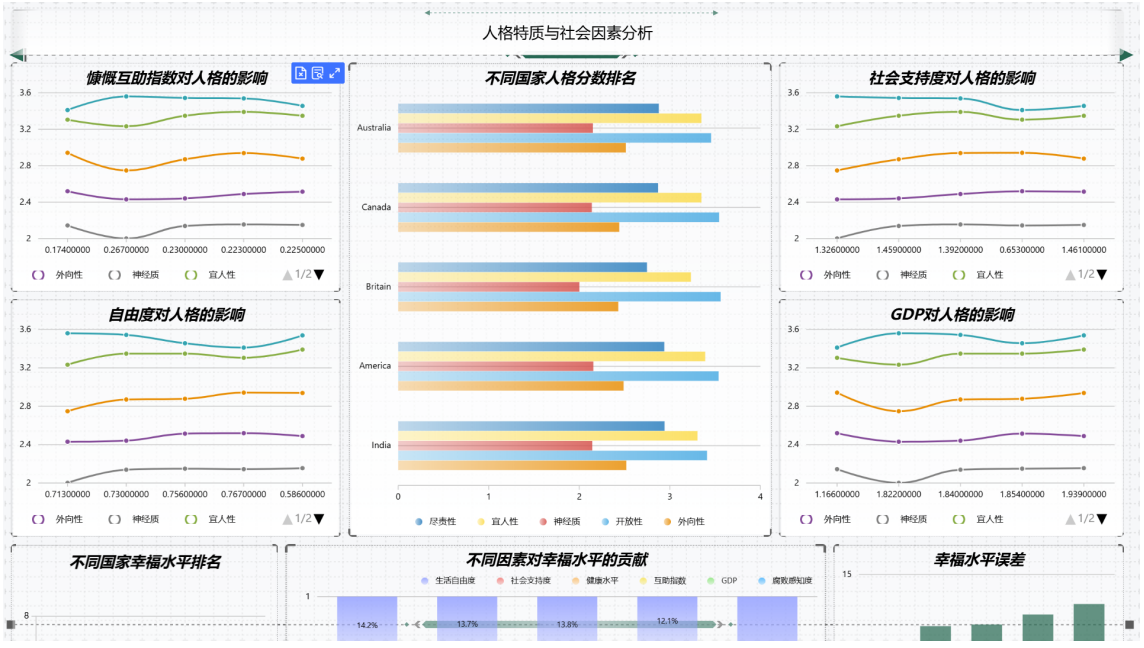
3. 可视化分析

在可视化分析中，我们选用[DataEase](DataEase)作为可视化分析工具。首先租用一台云主机（ucloud），然后按照官网示例将其部署为一个网页应用，之后便可通过网页来进行数据图表制作。

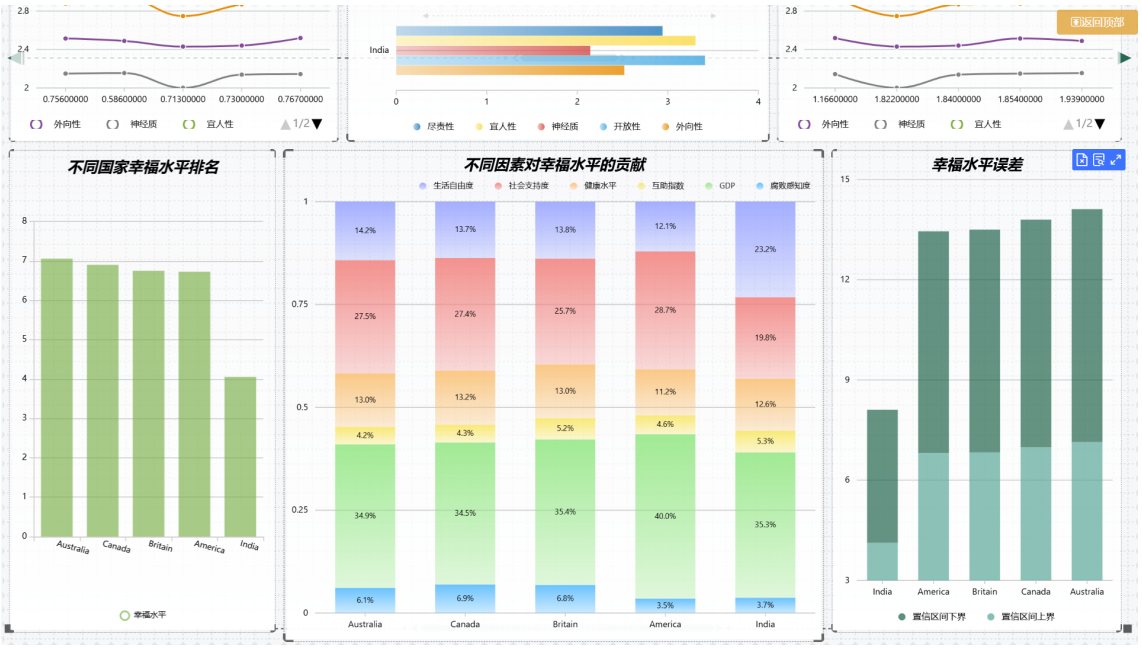
得到最终可视化大屏，可以在线互动，链接为：http://student.dataease.fit2cloud.com/link/vSYTHsBe

最终完成的可视化大屏如下：

- 中间条形图为不同国家不同人格分数的对比图，四个折线图分别为慷慨互助指数、自由度、社会支持度和GDP对人格分数的影响曲线。



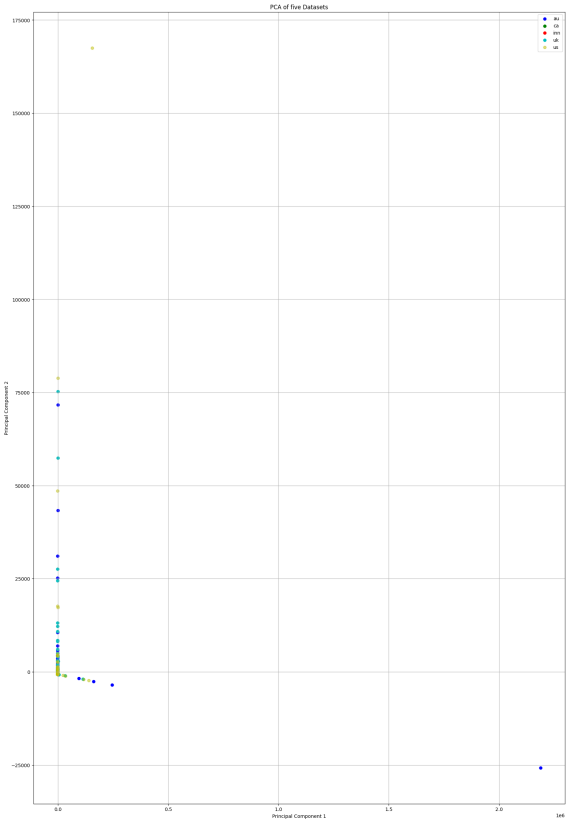
- 百分比柱状图显示了不同因素对幸福水平的影响，左图为5个国家幸福水平的排名，右图为幸福水平的置信区间。



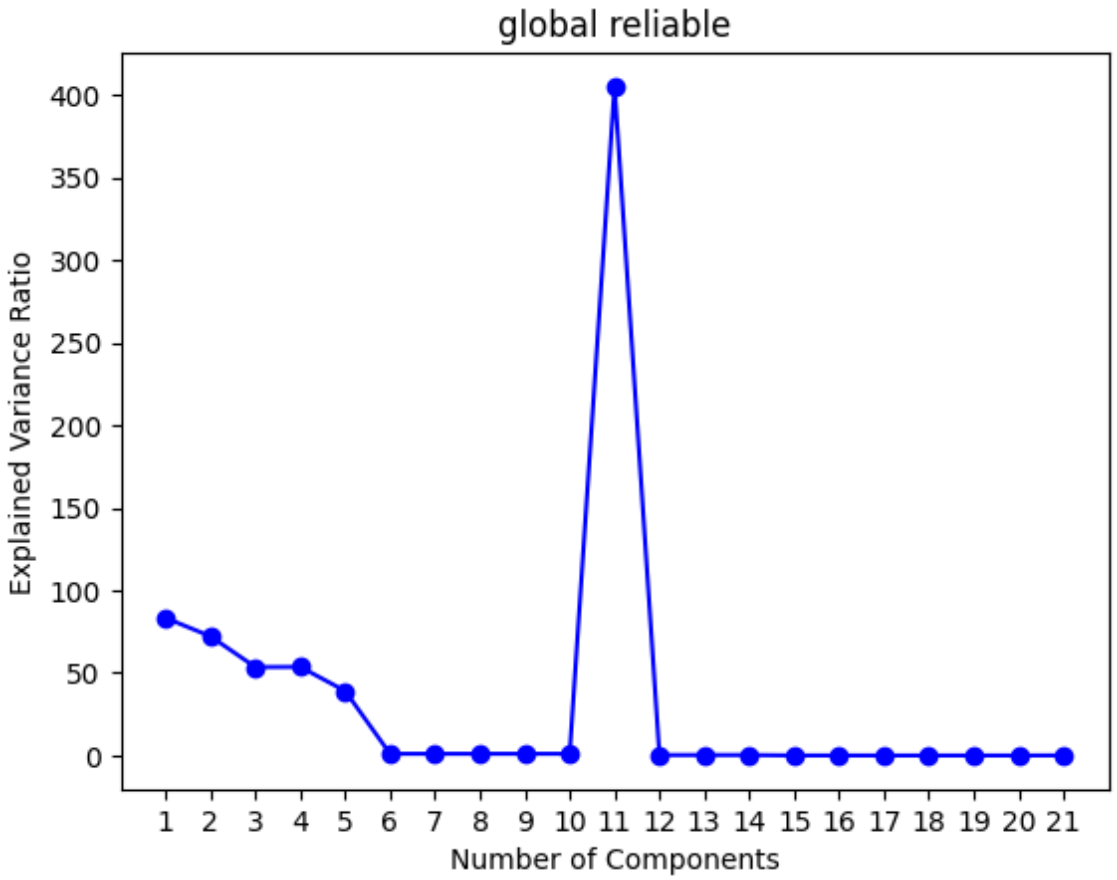
五、结果分析

1. 初步分析结果

- 通过pca降维分类画图之后可以看出不同国家总体人格倾向的聚集性，其中点的距离越近代表人格特质越相似：



- 通过因子分析进行绘图之后，通过碎石图我们可以得出不同特征对最终人格倾向的影响程度，其中影响最大的为经纬度特性，实际为各个国家间的社会因素不同造成的影响



- 在数据一致性检验方面，通过克隆巴赫系数可以看出总体人格数据的可信度是在接受范围内：

  Cronbach's α: (0.6943535047934039, array([0.694, 0.695]))

- 通过k-means聚合出的cluster分布如下，可以看出五类人格之间数量相差，以及分数相关指标，分数越低的cluster数量越少：

```
Clusters
0    191793
1    181019
3    179650
4    170459
2    143363
Name: count, dtype: int64
Clusters
0    164.043886
1    156.336578
2    148.665492
3    151.801547
4    162.266076
Name: score_sum, dtype: float64
```

- 每个cluster中五个人格维度的分数分布如下，可以从中看出cluster类别偏向指标，如cluster1类型明显看出CSN尽责性表现较低。

- 五个cluster的集群可视化结果：

2. 可视化分析结果

- 文化因素对人格分数影响是非常大的，但是文化因素非常难以量化，因此这里主要使用生活自由度,社会支持度和慷慨互助指数来分析对人格的影响。其中，随着生活自由度的上升，人格的尽责性和宜人性也在小幅度提升，但是同时神经质的分数也在上升。慷慨互助指数和社会支持度对人格也有一定的影响，但是影响的波动并不明显。
- 经济因素选择了GDP来分析对人格的影响，随着GDP的上升，不同人格的分数趋势都为先降低后小幅度上升，只有对开放性的人格影响主要是上升，其中的小幅波动可能是由于时间对不同流行趋势的影响导致。
- 在选取的五个国家中,开放性人格的分数都是最高的，神经质人格则分数最低，这样的数据也能看出即使不同国家的社会经济等条件都非常不同，但是主流的人格情况相似。
- 根据不同社会因素计算出的五个国家幸福水平中，澳大利亚的幸福水平最高，印度最低。可能由于选取的国家中只有印度不是发达国家，印度的幸福水平与其他四个国家的差距最大，刚刚超过4，其他四个国家幸福水平都在7左右，由此可见经济水平对幸福度的影响非常大。
- 在不同因素对幸福水平的影响中，普遍影响较大的是GDP、生活自由度和社会支持度，其中令人惊讶的是美国和印度的腐败感知度竟然对幸福水平影响都很小，美国为3.5%，印度为3.7%。

六、团队成员贡献

- 刘露莹：人格数据集聚类和可视化，五个国家人格和相关指标数据合并。
- 朱施颐：人格数据降维，特征分析，数据处理/分割/去重/，数据一致性检验。
- 朱文韬：文档组织与编写，可视化图表制作与结论分析，PPT制作与答辩。
- 徐婧： 社会因素数据收集，可视化面板制作，可视化分析，报告编写。