2021 年春季学期本科生《Python 数据分析》课程

小组作业描述

本文档最后修订于 2021 年 3 月 31 日。

作业描述

学生自由组队（每组不得少于 3 人，不得多于 5 人；需指定一名组长），每组自选数据，（至少）使用 Python 程序设计语言，对数据进行分析，对分析结果进行解读，并撰写报告。

组队时，2020 级本科生之间可以组队，高年级本科生之间也可以组队，但 2020 级本科生不可以和高年级本科生组队。本课程同时开设两个平行班，两班之间允许跨班组对。

为了完成本次作业，小组可以选择感兴趣的数据集。以下列举出了数据集的可能来源；当然，小组也可以从其他渠道获得喜欢的数据集。

1. 世界银行公开数据库（链接：https://data.worldbank.org/）。

世界银行公开数据库列出了世界银行数据库的 7000 多个指标，所有用户都可以免费使用和分享数据。可以按照国家、指标、专题和数据目录浏览数据。这些指标覆盖了国家层面经济、政治、公共卫生、教育等多个维度。你可以选择一部分你感兴趣的数据集，并自行下载这部分数据集再进行分析。当然，如果你发现有其他数据可以和本数据集有关，你也可以将它们结合起来分析。

2. Kaggle 平台上上百个数据集（链接：https://www.kaggle.com/datasets）。

Kaggle 是一个主要为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台。该平台已经吸引了 80 万名数据科学家的关注。该平台还开源了数百个数据集供用户免费使用。你可以选择一个或多个你感兴趣的数据集，并自行下载这部分数据集再进行分析；你也可以只使用一个数据的子集。当然，如果你发现有其他数据可以和本数据集有关，你也可以将它们结合起来分析。

3. Data.gov 网站上超过 22 万个数据集（链接：https://catalog.data.gov/dataset）

Data.gov 是美国提供的数据集平台。在它的官网上，用户可以根据数据内容、标签、格式、发布机构等进行筛选。你可以选择一个或多个你感兴趣的数据集，并自行下载这部分数据集再进行分析；你也可以只使用一个数据的子集。当然，如果你发现有其他数据可以和本数据集有关，你也可以将它们结合起来分析。

作业评分

本次作业在期末总评中共占 15 分，其中课堂展示占 5 分，报告 10 分。报告需要至少包括引言、数据集、方法、结果与讨论（针对结果进行的解读）、结论等部分。

提交步骤与时间

课程第 14 周（2021 年 6 月 7 日），小组**全部成员**在课堂上进行汇报，汇报的注意事项另行通知。

报告和 PPT 需要在 2021 年 6 月 28 日 23:59:59 前提交到教学网指定位置；每组由组长提交即可。请将全部内容包含在一个压缩包中，并命名为“[组序号].zip”（如“3.zip”）。

除遇不可抗力（不包括时间管理不善、课程冲突、数据或文档丢失等问题），如作业迟交在 24 小时以内，总分扣除 20%；迟交在 24 至 48 小时之间，总分扣除 40%；迟交在 48 至 72 小时之间，总分扣除 60%；迟交在 72 至 96 小时之间，总分扣除 80%；迟交 96 小时以上，该次作业不计入总分。严禁抄袭、套作。不得照搬或抄袭他人观点文字，需列出全部参考资料，必须遵照学术规范与诚信，否则本次作业记为 0 分。