2021 年春季学期本科生《Python 数据分析》课程

期中练习描述

本文档最后修订于 2021 年 4 月 26 日。

作业描述

本次作业的数据来自于北京大学中国社会科学调查中心 2018 年所做的“中国家庭追踪调查”中的数据。做问卷时候的跳转逻辑和数据的详细说明见 PDF 文件（“CFPS2018（第 5 轮调查）CFPS 2018 Wave 5 Questionnaires（CHN）.pdf”）。

本数据已经在原始数据（包含约 1000 个字段）基础上进行了字段和样本的处理，目前只保留了已婚受访者的数据。

需要注意的是，调查过程中为了实现无遗漏的调查，在问卷中设置了大量的跳题，根据在之前问题中回答结果的不同跳转到不同的问题中，使得大量的问题没有被回答，即在各个记录中都有大量的字段为空值或为逻辑缺失值。

各个字段的含义如下：

字段名称	字段含义
PID	样本 ID
cyear	调查年份
IBIRTHY	出生年份
AGE	年龄
Sex	性别
Education	教育程度
SpouseAge	配偶年龄
SpouseEducation	配偶教育程度
MeetWay	认识方式
MarryLast	婚姻持续时间
Cohabitation	是否婚前同居
FirstMarry	是否为初婚
Marriage	是否结婚
Jiazhuang	嫁妆的价值
Marriage_stf	对当前的“婚姻/同居”生活的满意度
Economy_contrb_stf	对配偶经济贡献的满意度
Housework_contrb_stf	对配偶家务贡献的满意度
Job_code	行业编码
Income_month	个人月收入

Income	个人年度总收入
Life_stf	生活满意度
General_stf	婚姻满意度

以下为各字段赋值的含义，其中负数可视为缺失值：

性别：1 为男，0 为女；

教育程度：0 为文盲/半文盲，3 为小学，4 为初中，5 为高中、中专、技校或职高，6 为大专，7 为大学本科，8 为硕士，9 为博士；

认识方式：1.在学校自己认识 ，2. 在工作场所自己认识， 3. 在居住地自己认识，4.在其他地方自己认识， 5. 经亲戚介绍认识 ，6. 经朋友介绍认识，7.经婚介介绍认识， 8. 父母包办 ，9. 经过互联网认识的，77. 其他；

是否字段：1 为是，5 为否。

Marriage：这个字段在本数据集中没有意义。因为目前遴选的数据全都是已婚受访者。

行业编码：1 农、林、牧、渔业 ，2 采矿业， 3 制造业， 4 电力、燃气及水的生产和供应业 ，5 建筑业 ，6 交通运输、仓储和邮政业 ，7 信息传输、计算机服务和软件业 ，8 批发和零售业， 9 住宿和餐饮业 ，10 金融业， 11 房地产业 ，12 租赁和商务服务业 ，13 科学研究、技术服务和地质勘查业 ，14 水利、环境和公共设施管理业 ，15 居民服务和其他服务业 ，16 教育， 17 卫生、社会保障和社会福利业， 18 文化、体育和娱乐业， 19 公共管理与社会组织， 20 国际组织 ，21 其他行业，99 其他

各类满意度字段：数字越大，满意度越高。

请针对于本数据集，使用 Python 程序设计语言，首先进行（必要的）变量处理，然后结合可视化方法进行一定的探索式数据分析（EDA）。随后，自行提出 2-3 个研究问题（例如“教育程度和结婚年龄有何关系”）并通过数据解答。务必要对数据分析结果进行解读。

作业评分

本次作业在期末总评中共占 15 分。评分细则如下：

- 问题界定（15%）；
- 代码正确性、可读性和完整性（20%）；
- 数据处理与探索式数据分析（15%）；
- 结果可视化（25%）；
- 结果的解读（25%）。

本次作业最多会有 10%的额外加分。

提交步骤与时间

请将全部代码、可视化和解读包含在一个 Jupyter Notebook 内，并在 2021 年 5 月 10 日 15:09 之前将本次作业提交到教学网。Notebook 命名为“期中-[学号]-[姓名].ipynb”（如“期中-张三-2000016601.ipynb”）。

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。