数据科学导论 2022秋 第一次作业

第一题 - 采样

TA 最近在抖腿公司实习，他想用公司的数据做一些实验。但是公司的数据太多太大了，大约有5个PB。TA 希望对这些数据进行采样，得到一份较小的数据，这样他的笔记本才能跑得动实验。

抖腿公司是做短视频的，他们的数据是一个**图（graph）**。图上有若干类型的**点（node）**：用户，短视频，音乐、特效等等。它们彼此之间也有若干类型的**连边（edge）**，代表着不同的关系，比如观看、点赞、使用，创作。

TA希望采样得到的数据能**尽量**保持原始数据的**特性。**图的特性有很多，比如两点是否连通、边的数量和点的数量的比例、度的分布、点的邻居的分布等等。你能帮帮TA吗？

请设计一种对图进行采样的算法，**简述**你的算法，然后说说这种算法的**能保持哪些特性、有什么缺点。(算法如有需要，可作相应的假设)**

第二题 - 相关分析

TA 最近得到了抖腿公司的一部分用户数据(**user_data.csv**)。他发现用户的年龄似乎和用户的观看时长有联系。请你帮帮TA，看看这两者到底有什么关系？

请先做简单的数据清洗，然后**选择一种**相关系数进行计算。然后说说你做了哪些数据清洗、选择的依据，以及这两者之间到底有什么关系。(推荐使用Excel实现或者编程实现)

（可选题）逻辑回归

TA最近回顾经典，再次观看《泰坦尼克号》。他好奇轮船上的最终生存与否和乘客本身的信息是不是具有一定的关系，于是他找到历史上的部分数据，想分析出其中的关联关系，你能帮帮TA吗？

请观察titanic文件夹中的train.csv和readme.md，了解乘客都有什么基本信息，思考怎么做数据清洗，以及怎么抽象成为逻辑回归的问题。

阅读 https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python/notebook ，简述他所采用的数据清洗方法、分析方法和最终的结论。

作业要求

- **严禁抄袭**!
- 上传**单个pdf文件**至网络学堂。
- 如果想上传第二题的代码。请写在pdf文件中。
- 推荐使用 markdown 或 latex 写作业。