作业说明

——完成人：曾凡霞（Fanxia Zeng）

一． 解题思路

运行方式：

1） **安装 python 包：**打包文件中包含了 poetry 相关文件，也将程序所需包导出为 requirements 了，通过 poetry 或 pip install requirements 进行包安装均可；

2） **运行命令：streamlit run test_excel.py，即出来网页显示。**
在 vscode 终端或者，linux 终端均可访问。（*ps:a. 前提 VPN 能行保证 LLM 的 key endpoint 可访问，b. key 和 endpoint 的访问方式，在 test_azure.py 里和 main.py 的注释里，均有所提供*）

解决思路：

1）沙盒环境:

建立 1 个 docker 镜像（包括程序所需安装包），可建容器，实现沙盒环境；

2）自然语言方式:

a. 通过调用 Azure OpenAI 的模型，实现 LLM 对话解析，根据用户对话内

容，理解其每次 promt 的需实现的功能，来决定是上传 Excel，清洗数据，特征工程，因子分析或画图功能;

b. 调 key 或 endpoint，linux 或 Windows 均需设置 VPN 才可访问外网。

3）langchain:

a.可实现多次连续对话：由于 LLM 的 endpoint 和 key 过期，并未完全实现 NLP 对话;

b.可实现数据分析功能：因为本身 LLM 只能进行自然语言分析，不能进行数据分析，通过代理和工具，自定义函数，或 python 工具，或 pandas 工具来实现数据分析功能。

4） streamlit:

实现 demo，可供用户会话，获取用户意愿，与程序进行交互，并对数据分析情况进行展示。

二、作业说明

1. 总体说明

本人是第一次实现 langchain、streamlit 等相关作业，基于自己的 python 基础和数据分析理论功底，时间有限因此仅实现了功能和 demo，在不考虑时间情况下可从以下角度更漂亮和全面一些。整个说明，结合功能和自己理论功底，进行陈述。

可将 promt 作为 query，判断该 query 对应的功能，如上传文件或特征清洗等。但这里仅将 promt 作为模板进行管理，选择功能。

可相对现在版本实现更多的交互：因时间内仓促，该作业仅简单采用 docker 实现沙盒目的（因未 model 的 key 和 endpoint 过期，未全部进行调试），并结合 streamlit 制作 demo 作为用户和程序之间的交互，langchain 则是调用 llm，若全调试完毕，可重新根据 promt 的反馈，设置更多选择供用户使用、可更倾向于 NLP 处理，实现对应的数据分析和可视化功能。

2. 数据:

仅 1 个表格：来自于 machine learning 公 1 开 UCI 数据库，数据自己随机

设置少部分为 null，仅为了配合数据清洗功能。

3. 数据清洗功能：

Key 和 endpoint 过期了，无法调用 model，因此对四项功能进行多选，以**实现用户定义性。**

1）**清洗仅考虑具有缺失值行的删除：**

因时间有限，未考虑数据重复等其它数据清洗功能，原因是因为 demo 和代码结构两者相似，只是代码调用 Python 函数不同。若不考虑时间，可以根据架构，进行重复数据处理、缺失值填充、清洗格式等多种功能扩充。

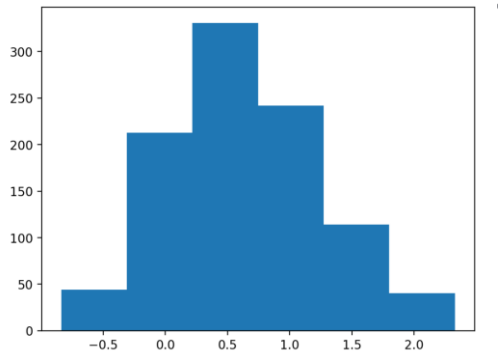
2）**虽未数据填充，但以前接触过，多说几句：**

数据清洗的方式有很多，仅从机器学习或数据科学角度考虑，比如对缺失值可以采用中值、均值、众数、随机填充、K 近邻填充、相似性填充等等进行填充，因时间有限未进行考虑。未来可通过本作业搭建的 langchain 和 streamlit 的 Python 代码，可在设置默认方式下，增加选择框可供用户选择的均值填充方式。（以前接触过所以知道方式比较多，这方面的工作也有很多 paper 包括 survey 可供相关研究查阅，但方法都比较直接明了。时间充裕可考虑 demo 应提供多种方式便于用户实际操作）【ps：这方面的工作，其实广西师范大学的 shichao zhangn 老师多年前在这方面做了不少工作】

3）**特征工程：**

仅实现数据 attribute 6 离散化化及可视化： 特征工程的范围其实很大，很广，包括了数据清洗、数据归一化、离散化、特征选择、降维、回归、甚至假设检验等等方式，这些计算本身并不是难事，可是自己写具体的计算函数代码，也可以调用 Python 函数（当然虽然以前做这些工作都是自己写代码，但平心而论直接调 Python 函数遇到 error 的处理机制更优良）。【这里根据题目文字中功能的并列关系，本人自动将特征工程单独建为一个功能，主要去实现 demo】

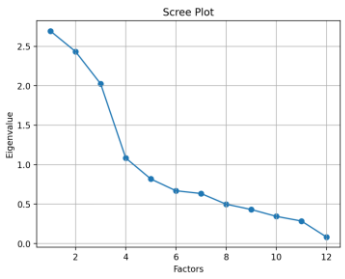
	attribute 6
0	5
1	5
2	5
3	4
4	3
5	3
6	3
8	5
9	5
11	5



4）**因子分析：**

先进行充分性检验，再计算因子旋转矩阵。

这里也单纯按题目语句仅对因子分析这一功能进行实现。若不考虑时间限制，其实可添加主成分分析(pca)、线性判别分析（LDA）、典型相关分析（cca）等等功能。 这些方法，都是基于不同的**投影目标**角度出发，提取数据的主要成分，可按有监督或无监督进行区分。



5）**画图功能：**

仅实现了因子分析的饼状图：若不考虑时间，其实应提供用户交互选项，可展示更多的图，包括扇形图（饼形图），直方图（柱状图）、火柴杆图、散点图、折线图。这里仅展示关于因子分析的折线图。