第二次作业要求

作业内容

运用课程中学习的编程方法与数据分析方法，进行数据的生成、保存、载入、预处理、数据分析、预测、数据可视化等。

数据集

- 可以是模拟深圳市个人财务情况和综合信息，进行年龄，学历，工作经验，职级，薪资，个人存款，个人贷款，个人所得税，子女数量，信用情况等进行分析、统计或预测
- 基础字段：id/身份证号，姓名，性别，年龄，学历，职位，月工资，存款，贷款，月供
- 可以添加自己认为有意义的任意字段，可以添加公司、家庭等数据集进行综合分析
- 可以自由探索数据集，例如可以从包括但不限于东方财富，雪球等网站，或者任何可以合法使用的数据集，或者自己生成其他自己感兴趣的数据集
- 数据集样本数量不少于5000个(自己生成数据集的要求）

补充说明

- 本作业属于半命题作业，给大家充分的发挥空间。可以只满足基础和优秀档，也可以去挑战满分档。
- 如果担心自己生成数据集有困难或者不好进行数据分析，可以采用网上或者别的同学的数据集，要求和金融相关，要求分析内容和结果不能雷同
- 自己生成数据集会有加分，好的数据分析也有加分；自己生成好分析的数据会有一定挑战性；其他数据集可能更好分析，但是过于复杂或只适用于机器学习；同学们自行选择自己的策略
- 无论是自己生成还是其他数据集，都需要包含有一定自己设计的算法逻辑 例如自己写的函数等，不能只是pandas已有函数的调用
- 数据分析指的是课堂讲过的例如平均值，分组，频率分布等分析方法。数据挖掘机器学习等属于想拿满分同学的可选策略，如果有同学恰好学过这部分或者想去自行调研，是不错的选择。当然也可以选别的拿满分的策略，只要满足满分要求中的3条即可拿满分。

基础档要求

作业需要满足以下基本要求：

1. 以*.ipynb格式提交作业
2. 使用pandas和numpy来处理分析数据
3. 使用markdown来组织作业，作业具有逻辑性，具体格式样例见作业模板
4. 包含能利用工资计算个税的函数（如果数据不相关则不需要）
5. 能够分析数据的基本信息，例如总数，平均值，分组平均值，求和等
6. 至少得到一张聚合后的series，或dataframe，形成新知识
7. 至少得到一个条形图和一个饼图
8. 至少研究两个具体问题，例如工资和工作年限的关系，工资和行业的关系等

优秀档要求

做到以下列表中任意2条，作业即可达到优秀：

1. 使用了自己生成的和工作业务相关的数据集
2. 有分组对比的图表
3. 表的种类多于3种，研究的具体问题多于5种
4. 逻辑清晰、算法清晰，分析的问题有价值，研究的结果有意义

满分档要求

在优秀档次的基础上，如果能做到以下列表中的任意3条，作业即可获得满分：

1. 使用多于一张表，例如citizens表，company表，family表等进行聚合分析
2. 运用学过的函数和方法在互联网上抓取并处理好新的数据集
3. 生成的数据分布很符合逻辑，或符合事实
4. 能够利用工具包建立回归等模型，进行收入等的预测
5. 用除了matplotlib的其他可视化工具进行数据呈现（plotly等）
6. 运用机器学习等高级工具进行了数据分析或预测
7. 数据集不少于100,000个样本(自己生成数据集的要求）