自强studio数分2019分流考核（一）

创建时间： 2019/9/29 17:12 **更新时间：** 2019/9/30 15:56

自强studio数分2019分流考核（一）

> 本次的学习重点：python--pandas库的使用，体会数据分析的思想和方法
> 考核周期：整个10月？

引入

咣当！自强studio掌上武大每年都会产生大量的用户成绩信息，利用这些数据进行脱敏然后做一些统计是很有意思的事情。

去年数分刚成立干了很多次这样的事23333，现在作为数分组的新成员，选择故事模式开始游戏，你回到了历史上的今天！

人物介绍：

- 你：悲催的数据处理员，不知道自己接了什么黑锅。
- 哨加胃，掌武负责人，没有经过你的同意就把你拉到一个不知道是什么的群里，告诉你要做一个学期成绩的分析。
- 松蛾，你的顶头上司，监视你的完成进度，没有完成任务时会鞭笞你。
- 洋芋豪，后台组腹黑小哥哥，你要取的数据全部是他管理，所以要跟他搞好关系。
- 大橙子：仁慈和蔼亲善的小哥，任务过程中出现了什么问题都可以咨询他~~~

关卡一：你挠了挠头，觉得头有点大

你有点迷，不知道如何下手，觉得数据分析数据分析应该先把数据搞到，然后你就做伸手觉找洋芋豪了。

洋芋豪：**SQL**还不会嘛！？还要麻烦我给你取数据，好烦人！喏，这是你要的这学期的数据。（丢给你两张表）下一次一定要自己把**SQL**学了啊！（发现下一次任务的痕迹！）

你：？？？解释一下这都是啥。。。

洋芋豪：这里有两张表：course和grade，course记录了这学期所有课程的信息，grade记录了同学的成绩信息，它们的字段都很简单，自己研究去吧

你：好好好！谢谢哥！

> 请你观察检视这两张表的信息，观察他们的字段和他们之间的关系，为下一阶段的任务打基础

关卡二：有点意思奥

你还是有点迷，还是不知道要如何着手，于是你找到了哨加胃，想询问一下业务需求

哨加胃：这个嘛，我觉得你要先给我统计出来：

- 我们的开课情况，比如说各种类型的课的占比
- 哪个学院开课最多
- 哪个老师开课最多
- 每个人的平均成绩
- 修了最多学分的是哪个bt
- 修了最多门课的是哪个bt

这个只是初步，鉴于你是个新手，就先让你做这些简单的练手，以后会有更难的嘿嘿嘿

你：是的哥是的哥！

> 请利用手头数据，结合加胃哥给你提出的要求，利用pandas库来进行统计分析，得到这些问题的结果后交给大橙子

大橙子：什么？你还不会pandas？哈哈哈哈，我这里有教程，你拿去看吧！

> **中文参考教程**：
> https://www.kesci.com/home/project/59e77a636d213335f38daec2
> **英文参考教程(我更推荐)**：https://www.kaggle.com/learn/pandas
> **参考书籍**：《利用python进行数据分析》中关于pandas的部分
> **推荐工具**：ipython，jupyternotebook，pycharm

大橙子温馨提示：不要试图把所有东西完整的学完背过，再去做任务，而是先学习最小化门槛知识，然后在做中学，不会的查和记录，不断完善，迭代学习，尤其是你一直读那本书而不做的话会自闭的

关卡三：哈哈哈太轻松了！冲冲冲（技术）

你学了学pandas，觉得似乎也不是那么难啊。于是自信心爆棚，觉得数据分析不过如此，于是你自信地去找哨加胃，讨要下一阶段的任务了

你：好简单啊！我要冲冲冲

哨加胃：nbnb，之后可能就不会那么轻松了，那么下一阶段的需求是：

- 匹配每一位同学的学院
- 统计每一位同学本学期的GPA（规则百度）
- 计算平均GPA最高的前五个学院
- 清除冗余课程分类（只留下四种：专业必修，公共必修，公共选修，专业选修）
- 统计出上课人数最多的前十门课的平均成绩

（提示：利用好所给的id_college_map.json，使用pandas的apply函数很重要）

> 请利用手头数据，结合加胃哥给你提出的要求，利用pandas库来进行统计分析，得到这些问题的结果后交给大橙子

关卡四：不冲了不冲了（业务）

经过上一轮的摧残，你真的觉得好心累，还是不冲了不冲了。

松蛾：完成的很不错！我很满意23333！诶，但是我对于哨加胃提出的这些数据指标不满意，这些信息统计出来了好像也没有什么意思hhh，所以我觉得你应该提出更多的数据统计目标23333

你：收到！

> 请你提出几个可以从两张表成绩数据的字段中能挖掘出来的新指标，并谈一谈这些数据的业务价值。

大橙子：哈哈哈，缺少业务知识？我可以给你推荐一些资源

- 《深入浅出数据分析》
- 掌上武大部长邵家伟同学，跟我们合作了很多项目，可以向他取经

> 阅读《深入浅出数据分析》，并回答如下问题：
>
> - 数据分析过程中最重要的一件事？
> - 数据分析的主要过程？
> - 分解问题的意义是什么？
> - 评估的意义是什么？
> - 一个简单的数据分析报告应该包括什么？
> - 什么是混杂因素，如何管理混杂因素？
> - 图形化的原则是什么？
> - 假设检验的核心？
> - 你如何看待回归方法？有什么优点和缺点？
> - 你如何看待数据清洗工作在数据分析工作中的地位？
> - 谈谈你们读完此书的感想？

注：重点阅读章节：

- 数据分析引言
- 检验你的理论
- 数据图形化

- 假设检验
- 启发法
- 直方图
- 回归
- 误差
- 整理数据