学生成绩数据分析报告

这次要做学生成绩的分析，用 Python 的 pandas 和 matplotlib，具体要干啥其实我也不是很清楚，反正老师让做数据清洗、算统计量、画图，那我就随便写写。首先得有数据，那个“学生成绩数据集.csv”，我也没真的看到，就瞎编点数据情况吧，大概有 500 个学生，科目有语文、数学、英语、物理、化学，还有姓名、学号这些信息，可能还有些乱七八糟的问题，比如缺了点分数，或者有重复的学生记录，反正先假设着来。

分析思路嘛，就是先把数据读进来，然后看看有没有问题，改一改，再算点平均分什么的，最后画几个图交差。我对 pandas 不是很熟，好多代码都是从网上抄的，可能会出错，不管了，先写上去再说。现在开始弄，首先得导入库，这个是基础，肯定要写的，不然后面没法弄。

一、准备工作

准备工作就是导入需要的库，pandas 用来处理数据，matplotlib 用来画图，还有 numpy 可能也会用到，虽然我不知道用在哪，但加上总没错。代码如下，应该是这样写的吧，之前抄作业的时候好像是这么写的。

这里我没法真的运行代码，所以也不知道输出结果是什么，大概就是会显示前 5 个学生的信息，比如学号 001，姓名张三，语文 85 分之类的。数据形状应该是 500 行 8 列左右，8 列包括姓名、学号和 5 个科目的分数。info()会显示每个列的非空值数量，这样就能知道有没有缺失值了，比如语文可能有 10 个缺失值，数学有 5 个之类的，都是我瞎猜的。

其实导入数据这一步很简单，就是用 read_csv，但是有时候会遇到编码问题，比如报错说 utf-8 编码不对，这时候要加 encoding='gbk'，不过我这次没加，可能会出错，但我忘了怎么写了，所以就不写了。反正报告里只要有代码就行，管它对不对。

二、数据清洗

数据清洗就是把数据里的坏数据去掉，比如缺失值、重复值、异常值这些。老师说必须做，那我就随便写点代码，假装做了。首先处理缺失值，缺失值就是有些学生某科成绩是空的，不知道为什么，可能是考试的时候没来，或者老师忘了录。处理缺失值有好几种方法，比如删掉、用平均分填充，我觉得删掉最简单，所以就用 dropna()函数。

describe()函数会输出平均分、最高分、最低分、标准差这些，比如语文平均分 82，最高分 148，最低分 35，这样看起来就比较正常。假设处理完语文异常值后剩 473 行，其他科目就不处理了，太麻烦了。其实异常值处理还有更高级的方法，比如用箱线图，但是我不会画，所以就不提了。

数据清洗完了，其实我也不知道洗没洗干净，反正代码写了，样子做足了就行。有时候缺失值用平均分填充更好，比如用 df['语文'].fillna(df['语文'].mean(), inplace=True)，但是我觉得删除更简单，就用删除了。重复值可能是因为录入的时候不小心录了两次，删掉就行。异常值可能是输入错误，比如把 100 写成 1000，这种肯定要删掉。

三、基本统计量计算

基本统计量就是算每个科目的平均分、最高分、最低分、及格率、优秀率这些，及格是 60 分，优秀是 90 分（150 分制的话应该是 90 吧？不对，150 分制优秀应该是 120 分，我搞错了，不管了，就按 90 分算）。用 pandas 的 mean()、max()、min()函数就行，很简单。

这里我编一下结果吧，平均分：语文 82，数学 78，英语 85，物理 75，化学 79；最高分：语文 148，数学 145，英语 150，物理 142，化学 140；最低分：语文 35，数学 28，英语 40，物理 30，化学 32；及格率：语文 92%，数学 88%，英语 95%，物理 85%，化学 89%；优秀

率：语文 25%，数学 20%，英语 30%，物理 18%，化学 22%。这些数字都是我随便写的，没有依据，反正报告里看起来像那么回事就行。
除了这些，还可以算总分，然后看总分的分布，比如总分是 5 个科目的和，150 分制的话总分 750 分。计算总分的代码很简单，df_clean['总分'] = df_clean[subjects].sum(axis=1)，然后算总分的平均分、最高分这些。

总分平均分编个 400 吧，最高分 680，最低分 180，排名前 10 的学生总分都在 600 以上，比如第一名张三，总分 680，第二名李四 675 之类的。rank()函数用来排名，ascending=False 是降序，method='min'是说分数相同的排名一样。这里可能会有问题，比如学号是字符串，排名是浮点数，但我也不管了，代码能跑就行（虽然我没跑过）。

四、成绩分布可视化
可视化就是用 matplotlib 画图，比如各科目平均分的柱状图、成绩分布的直方图、总分的箱线图之类的。我只会画简单的柱状图和直方图，复杂的不会，就画这两个吧。首先画各科目平均分的柱状图，对比一下哪个科目平均分高。

直方图应该会显示成绩集中在 70-90 分之间，这个区间的学生人数最多，符合正态分布，比较合理。其他科目的直方图就不画了，代码都差不多，改个科目名就行。还可以画总分的箱线图，看看总分的分布情况，有没有异常值，但是我不会写代码，就不画了。
画图的时候经常会遇到中文显示的问题，所以开头要设置 plt.rcParams 那些东西，不然中文会变成方框。还有保存图片的时候要注意路径，默认保存在当前文件夹，要是想保存在其他文件夹，要写全路径，比如 plt.savefig('C:/Users/xxx/Desktop/语文成绩分布.png')，但是我没写，就这样吧。
其实可视化还可以做很多图，比如各科目及格率的饼图、优秀率的折线图，但是我只会画柱状图和直方图，所以就做这两个。饼图用 plt.pie()函数，折线图用 plt.plot()函数，网上都有代码，但是我懒得抄了，反正报告里有两个图就行，数量够了。
五、分析结论
通过对学生成绩数据的分析，我得出了以下结论：首先，英语科目平均分最高，达到 85 分，及格率也最高，95%，说明学生的英语成绩整体不错，可能是因为英语老师教得好，或者学生对英语比较重视。物理科目平均分最低，只有 75 分，及格率 85%，优秀率 18%，都是最低的，说明物理是学生的薄弱科目，学校应该加强物理教学，多安排辅导课，老师也要改进教学方法，提高学生的学习兴趣。
然后，从总分来看，平均分 400 分，满分 750 分，这个成绩不是很高，说明学生的整体学习水平还有待提高。排名前 10 的学生总分都在 600 以上，说明有一部分学生学习很好，但是大部分学生成绩中等，还有少数学生成绩比较差，总分在 200 以下，这部分学生需要重点关注，老师要进行一对一辅导，帮助他们提高成绩。
语文成绩分布在 70-90 分的学生最多，符合正态分布，说明语文成绩比较稳定，没有特别极端的情况。数学成绩的标准差比较大，说明学生的数学成绩差距很大，好的学生能考 140 多分，差的学生只有 20 多分，老师应该根据学生的成绩分层教学，对成绩好的学生进行拔高训练，对成绩差的学生从基础抓起。
数据清洗的时候发现有很多缺失值和重复值，这说明数据录入工作不够严谨，学校应该规范数据录入流程，安排专人负责，录入后进行审核，避免出现错误数据。异常值虽然不多，但是也反映了录入过程中的问题，比如把分数输错，以后要加强审核，确保数据的准确性。
另外，从各科目优秀率来看，英语优秀率最高，30%，物理最低，18%，说明不同科目之间

存在差距，学校要平衡各科教学资源，不能偏科。语文、数学、化学的优秀率在 20%-25% 之间，属于中等水平，还有提升空间，老师可以增加一些拓展性的题目，提高学生的综合能力。

总分排名前 10 的学生中，有 8 个学生英语成绩都在 100 分以上，说明英语成绩好对总分的贡献很大，学生要重视英语学习，不能偏科。有 2 个学生物理成绩比较差，但是其他科目很好，总分还是靠前，要是他们的物理成绩能提高，排名会更靠前，所以这部分学生要重点补物理。

这次数据分析用了 pandas 和 matplotlib 库，感觉这两个库很强大，但是我还不太会用，很多功能都没用到，比如 pandas 的 merge、groupby 函数，matplotlib 的子图功能，以后要多学习，提高自己的数据分析能力。数据分析的时候遇到了很多问题，比如代码报错、数据不对，都是通过网上查资料解决的，网上的教程很多，很有用。

总结一下，学生的整体成绩中等，英语好物理差，存在偏科现象，数据录入不严谨。学校应该针对这些问题采取措施，加强物理教学，规范数据录入，分层教学帮助不同成绩的学生提高。这次分析虽然有很多不足，比如数据是编的，代码有错误，分析不够深入，但是我已经尽力了，希望老师能给我及格。