河海大學
HOHAI UNIVERSITY

大数据管理与应用专业

商务数据分析实验报告

高考志愿填报数据收集分析指导

姓名：李宗霖

学号：2063110124

时间：2022 年 6 月 28 日

指导老师：徐绪堪

目录

一、 背景与实验目的介绍

高考刚刚结束，又到了一年一度报考志愿的时候，俗话说的好：“志愿填报好，胜过好高考”。如何根据学生个人的分数，填报可以触及的最好大学，或者某些大学的最好专业，成为了 6、7 月份困扰大部分高三学子和家长的问题。在填报志愿时，我们往往需要考虑很多问题，首先是多少分可以报考什么学校，其次是学校提供哪些专业，该专业的全国排名的优劣以及专业的就业情况等等。通过对这些数据的横向对比和时间上的纵向对比可以较好地选择填报志愿。为了妥善地处理解决家长的这些问题，特此以此为专题，通过手收集数据分析的方法，进行科学志愿填报推荐。

二、 数据收集

在本部分，将具体介绍上文提及的报考时需要参考的数据，并详细讲解收集方法，具体数据展示均以江苏省考生及河海大学为例（仅为展示数据，总数据超过 10000 条）。

2.1 数据源与数据介绍

介绍数据来源，并对具体数据进行简单描述。

2.1.1 普通高校招生普通类本科批次平行志愿投档线

针对什么分数可以报考什么大学的问题，我们完全可以参考江苏省教育考试院每年发布的《普通高校招生普通类本科批次平行志愿投档线》。

在该文件中，很好地介绍了各个学校的最低录取分数，可以有效地帮助填报决策。（以 2021 年物理类为例，详见附件 1）

江苏省2021年普通高校招生普通类本科批次
平行志愿投档线
（物理等科目类）

院校代号	院校、专业组（再选科目要求）	投档最低分	投档最低分同分考生排序项					
			（一）	（二）	（三）	（四）	（五）	（六）
			语数成绩	语数最高成绩	外语成绩	首选科目成绩	再选科目最高成绩	志愿号
0105	陆军工程大学03专业组(不限)	537	194	100	108	59	89	9
0131	海军军医大学03专业组(不限)	581	218	115	121	78	89	9
0131	海军军医大学04专业组(化学)	579	218	122	132	59	88	16
0131	海军军医大学05专业组(化学或生物)	595	228	122	127	66	88	7
0201	中国人民公安大学10专业组(不限)	585	215	115	138	68	83	2

2.1.2 高等学校招生计划

为详细了解欲报考大学的具体专业，我们可以登入学校的招生信息网进行具体查询。（以河海大学为例）

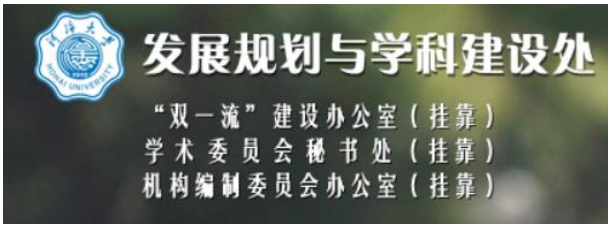
在该网页提供的数据中，我们可以知道该院校对报考省份提供哪些专业选择，往届招生详情，具体专业报考分数，对报考专业选择有很强的参考性。（本部分使用八爪鱼对数据进行爬取，依旧以河海大学为例，数据详见附件 2、3）

	A	B	C	D	E	F	G	H	I
1	省份	类型	招生专业	校区	学费	字段	科类	字段1	专业组序号
2	江苏	普通类	水利类(大	南京校区	6380	物理+不限	1	20	大二在常州校区学习一年
3	江苏	普通类	计算机类	南京校区	6380	物理+不限	1	10	大二在常州校区学习一年
4	江苏	普通类	水文与水	南京校区	6380	物理+不限	2	28	大二在常州校区学习一年
5	江苏	普通类	水务工程	南京校区	6380	物理+不限	2	20	大二在常州校区学习一年
6	江苏	普通类	水利水电	南京校区	6380	物理+不限	2	38	大二在常州校区学习一年
7	江苏	普通类	智慧水利	南京校区	5800	物理+不限	2	5	大二在常州校区学习一年
8	江苏	普通类	港口航道	南京校区	6380	物理+不限	2	36	大二在常州校区学习一年
9	江苏	普通类	海洋资源	南京校区	6380	物理+不限	2	2	大二在常州校区学习一年
10	江苏	普通类	船舶与海	南京校区	5800	物理+不限	2	5	大二在常州校区学习一年
11	江苏	普通类	土木工程	南京校区	6380	物理+不限	2	40	大二在常州校区学习一年
12	江苏	普通类	电气工程	南京校区	6380	物理+不限	2	25	大二在常州校区学习一年

	A	B	C	D	E	F	G	H	I
1	省份	类型	招生专业	校区	学费	字段	科类	字段1	专业组序号
2	江苏	普通类	水利类(大禹强化班)	南京校区	6380	物理+不限	1	20	大二在常州校区学习一年
3	江苏	普通类	计算机类(大禹强化班)	南京校区	6380	物理+不限	1	10	大二在常州校区学习一年
4	江苏	普通类	水文与水资源工程	南京校区	6380	物理+不限	2	28	大二在常州校区学习一年
5	江苏	普通类	水务工程	南京校区	6380	物理+不限	2	20	大二在常州校区学习一年
6	江苏	普通类	水利水电工程	南京校区	6380	物理+不限	2	38	大二在常州校区学习一年
7	江苏	普通类	智慧水利	南京校区	5800	物理+不限	2	5	大二在常州校区学习一年
8	江苏	普通类	港口航道与海岸工程	南京校区	6380	物理+不限	2	36	大二在常州校区学习一年
9	江苏	普通类	海洋资源开发技术	南京校区	6380	物理+不限	2	2	大二在常州校区学习一年
10	江苏	普通类	船舶与海洋工程	南京校区	5800	物理+不限	2	5	大二在常州校区学习一年
11	江苏	普通类	土木工程	南京校区	6380	物理+不限	2	40	大二在常州校区学习一年
12	江苏	普通类	电气工程及其自动化	南京校区	6380	物理+不限	2	25	大二在常州校区学习一年

2.1.3 专业排名

在选择专业时，还有一个很重要的参考项便是该学校专业在全国的评定层次，是否为双一流学科，是否为 A 类学科，都是报考时的重要参考选项。（以河海大学为例）



很遗憾，在学校官网给出的链接中无法进行数据查询，最后使用教育机构提供的参考数据，对专业评估进行整理。（以河海大学为例）。

序号	学校名称	一级学科名称	评估结果
1	河海大学	水利工程	A+
2	河海大学	土木工程	A-
3	河海大学	环境科学与工程	A-
4	河海大学	马克思主义理论	B+
5	河海大学	管理科学与工程	B+
6	河海大学	工商管理	B+
7	河海大学	社会学	B
8	河海大学	力学	B
9	河海大学	计算机科学与技术	B

2.1.4 毕业生就业质量年度报告

在本部分，通过使用高等学校就业创业信息网提供的毕业生就业质量报告来参考每届毕业生的就业情况，从而以未来职业规划为基础，更好地选择院校和专业。（以河海大学 2019 年毕业生就业质量年度报告为例，详见附件 4）

数据来源（上图）

2019 届本科毕业生分院系专业就业率

学　　院	毕业生数	就业率	升学和出国率
大禹学院	101	97.03%	84.16%
工程力学	24	100.00%	95.83%
土木工程	19	100.00%	78.95%
水利水电工程	20	95.00%	85.00%
水文与水资源工程	19	100.00%	78.95%
港口航道与海岸工程	19	89.47%	78.95%
水文水资源学院	278	97.12%	54.32%
自然地理与资源环境	31	96.77%	29.03%
水文与水资源工程	116	99.14%	64.66%
水务工程	131	95.42%	51.15%

部分就业率信息（上图）

2.1.5 一分一段表

（本部分为后续补充）

在高考成绩出分前后，各大平台陆续提出了高考填报指导的服务，这也说明我的选题方向符合当下热门。我随即向提供类似服务的夸克平台询问相关问题，得出以下回复：


« 返回　回复　回复全部　转发　删除　彻底删除　举报　拒收　标记为...　移动到...　上一封 下一封

夸克☆

发件人：夸克客服服务 <kksupport@service.alibaba.com>

时　间：2022年6月24日（星期五）上午10：41

收件人：木子宀示雨林 <2648330774@qq.com>

纯文本 |

关于您反馈的志愿填报问题【工单编号：20025004220771 38】，经核实，夸克高考智能选志愿给出的冲稳保数据，基于考生分数和位次计算，主要参考该分数和位次，在相同省份、相同选科情况下，在过去三年的录取情况，根据可报考学校的最低录取分数和位次进行计算，可冲击意味着在当年可录取概率相对较小，可保底意味着过去三年至少2年可录取，录取概率相对大，夸克高考相关数据和工具查询结果仅供参考，最终报考须学生和家长参考多方信息、谨慎决策。高考志愿是人生重要决定，是每个考生的个性化选择，夸克高考提供查询工具和智能志愿工具帮助考生了解自己想要什么，处在什么阶段，有哪些选择的可能性，最终报考须学生和家长参考多方信息，自主决策，愿你志如所愿~感谢您的支持，夸克~

在数据源方面，夸克提供的服务中，还提及了参考到分数的排名情况，经过思考，发现，在填报志愿时，除了参考成绩是否达到投档线外，在省内的排名也具有很大的参考价值，因此特查询江苏省公布的一分一段表，作为数据补充。（以江苏 2021 年为例，具体表见附件 5）

	A	B	C	D	E	F
1	年份	省份	科目	分数	人数	累计
2	2021	江苏	历史类	632	101	101
3	2021	江苏	历史类	631	14	115
4	2021	江苏	历史类	630	8	123
5	2021	江苏	历史类	629	18	141
6	2021	江苏	历史类	628	18	159
7	2021	江苏	历史类	627	13	172
8	2021	江苏	历史类	626	25	197
9	2021	江苏	历史类	625	21	218
10	2021	江苏	历史类	624	25	243
11	2021	江苏	历史类	623	26	269

2.2 存在问题

2.2.1 数据分散

因为是从各大学校的不同官网进行数据收集和爬取，所以想要完整的将全部高校的所有数据全部收集存在着工作量大的问题，应该适当考虑优化数据采集的方式来减小工作量。

2.2.2 数据缺失

在官方给出的数据中，存在着数据缺失的情况，部分专业信息没有在公告中给出，而且部分高校缺少最新数据，以河海大学为例，就业数据只给到 2019 年，没有最新数据。

2.2.3 数据爬取

“天下没有免费的午餐”，在使用了八爪鱼、浏览器插件等方式对数据进行爬取后，综合使用体验发现八爪鱼提供的数据抓取服务最方便简介，但是在使用“云收集”等高级服务时，便需要支付高额的会员费用（我没有^_^）。在爬取某些教育机构提供的网站数据时，发现不可爬取，受到限制，经过联系咨询之后，发现如果需要具体数据需要购买。

2.3 数据评估

2.3.1 来源与获取方法

数据来源为学校官网，江苏省教育厅官网，教育机构提供的学校招生数据等。

获取方法主要包括：官网下载、使用八爪鱼爬取、与机构沟通获取。

2.3.2 采准率

由于数据集的来源，在数据准确率方面得到了保证。

2.3.3 采全率

在采全率方面，因为要保证数据的采准率，没有采纳非官方提供的最新数据，因此在学校官网没有提供最新数据时，部分年份数据是缺失的。

三、 数据清洗与汇总

在数据预处理方面，我主要做了数据去重、空值处理以及数据汇总三个面，其中，空值处理又分为直接删除和回归补充，下面将进行具体介绍。

3.1 数据去重

为了保证数据的真实性和准确性，我将从教育机构的数据和从官网找到的同

类型数据进行了 Excel 的合并。只需将重复的数据进行去重处理，就可以得到不重复的最完整的数据。但在数据集合并之后，必然会出现大量的重复数据。或者，因为数据源的不准确，本来就存在重复的数据。这就需要我们采用高效地数据去重方法。

在本实验研究中，主要是使用了 Python 脚本对 Excel 进行对应的数据去重操作。（代码如下）

```python
# -*- coding = utf-8 -*-
# @Time : 28/6/2022 下午7:52
# @Author : 木子宀示雨林
# @File : quchong.py
# @Software : PyCharm
import pandas as pd
import numpy as np

df = pd.read_csv(r'C:\Users\商务数据分析\Desktop\data\附件1.csv')
df.sort_values('grade',ascending=False)#对文件按指定关键字进行排序
print(len(df))
a=df.drop_duplicates(["grade"])#对文件按指定列retro_templates去重
print(len(a))
a.to_csv(r'C:\Users\商务数据分析\Desktop\data\附件1.csv')#去重后文件重新保存到新文件
```

3.2 空值处理

由于专业名称的改变还有数据录入的异常等原因，导致了空值出现，空值在后续运算数据的过程中会造成报错，因此与要对空值进行处理，本实验主要采取了以下两种处理方式。

3.2.1 直接删除

对于时间特别久远的数据，比如在 2022 年进行报考，那么 2017 年的给出的各项数据的参考价值就偏低，因此对于 2017 年采集到的数据中的空值情况，采用了直接删除处理。因为每年的数据是单独的 Excel 进行存储的，所以可以直接对 2017 年的数据进行整个 Excel 文件的删除空值，而不必担心对之后的数据造成影响。本实验主要采用了 Python 脚本对空值进行处理。（代码如下）

```python
import pandas as pd
import numpy as np

df = pd.read_csv(r'C:\Users\商务数据分析\Desktop\data\附件1.csv')
# 准备工作
df.isnull().any()  #查看哪一列有空值，发现是<订单付款时间>列
print(df[df['rank'].isna().values==True])  #输出<订单付款时间>列存在空值的行
#清洗空值
df2 = df.dropna(axis=0,how='any',thresh=None,subset=None,inplace=False)  #删除含有空值的行或列
df2['rank'].isna().any()  #查看是否还存在空值
#再次查看
df2.shape

df.replace(to_replace=r'^\s*$',value=np.nan,regex=True,inplace=True)
df['rank'].dropna()
```

代码的前半部分主要用于空值的所在行的直接去除，而代码的后半部分则进行了去空后是否还有“空值”的存在（上半部分代码仅去除空值，不能排除空字符串的影响），然后进行再次去除，这样可以保证将空值和空字符串全部去除，避免对后续算法造成的影响。

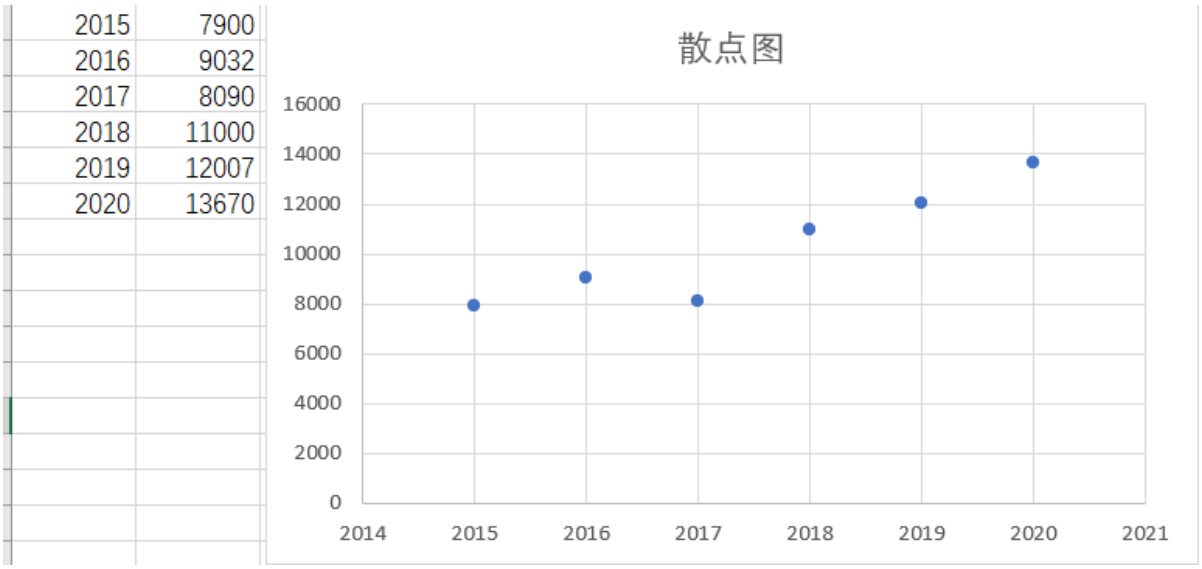
注：文件保存等代码未给出，仅展示了核心去空代码。

3.2.2 空值补全

对于最新的数据，如 2021 年的数据，则对 2022 年的高考填报有着重要的参考价值，因此，对于此类数据的缺失，需要进行人工补齐，本实验采用的方法主要是回归分析法，对近年来的其他数据进行回归分析，推测出缺失的数据。（以下为已知 2015~2020 年数据，2021 数据为空值的情况）

本实验主要采用了 Excel 内置的回归分析函数对数据进行回归分析。

首先，收集同类型数据的前 6 年数据，并建立新 Excel 表格进行存放，同时绘制散点图进行趋势预测。（本文展示的是某大学 2015~2020 年的录取分数排名，对 2021 排名进行预测）。



随后，进行回归分析。

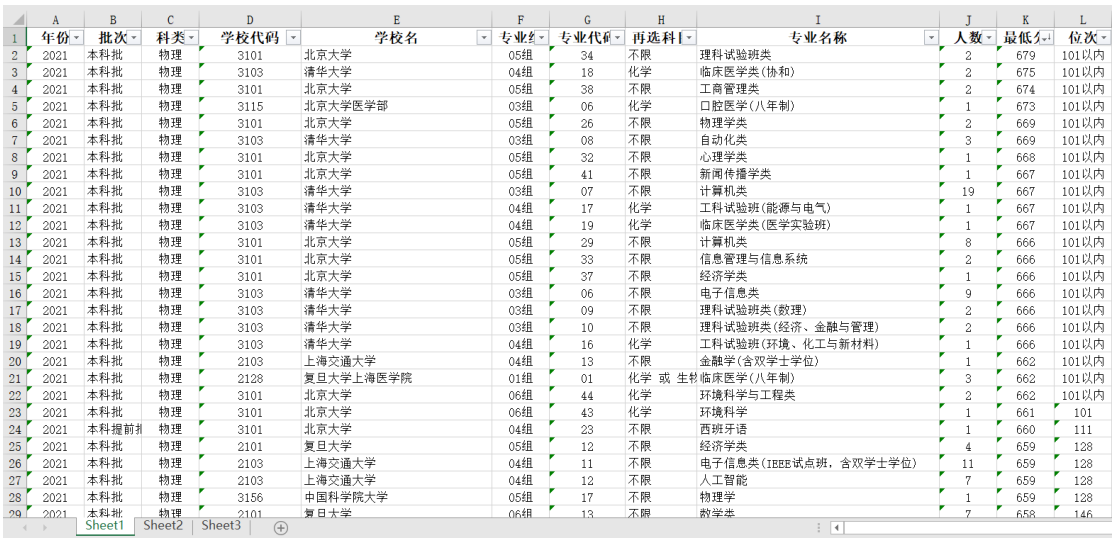
（回归分析结果如下图：）

SUMMARY OUTPUT								
回归统计								
Multiple R	0.935654							
R Square	0.875448							
Adjusted F	0.84431							
标准误差	917.0973							
观测值	6							
方差分析								
	df	SS	MS	F	gnificance F			
回归分析	1	23646703	23646703	28.11511	0.006077			
残差	4	3364270	841067.4					
总计	5	27010973						
	Coefficient	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	-2334916	442293	-5.27912	0.006174	-3562919	-1106914	-3562919	-1106914
X Variable	1162.429	219.2282	5.302368	0.006077	553.7536	1771.104	553.7536	1771.104

经过几次测试，发现江苏省内的录取分数排名十分稳定，基本上方差很小，因此，对于个别有变化趋势的数据（如上述内容），进行回归分析预测，用预测结果代替空值位置。同时，对于变化趋势不明显的数据（比如，仅 6 年的数据没有明显变化趋势，且波动小），则进行取前几年数据的均值的形式对空值进行补充。来最大程度上模拟实际可能出现的结果。

3.3 数据汇总

因为在最后志愿报考参考数据时，不可能想知道某项数据时，去单独的找某一张表进行比对，为了最大程度上的简化和方便用户的使用，我将之前收集到的，参考各大论坛评论，决定的对报考志愿最具参考性的几项数据（主要包括：年份、批次、科类、学校代码、学校名、专业组、专业代码、再选科目、专业名称、人数、最低分、位次）进行了数据表格的汇总处理，此部分主要采用人工汇总的方式，可以最大程度上的避免数据错误的情况，但也存在着工作量大的缺点，这个缺点可以后续使用 Excel 自带的相关功能以及 Python 脚本进行简化。整理好的数据表格如下（具体数据见附件 6）



	A	B	C	D	E	F	G	H	I	J	K	L
1	年份	批次	科类	学校代码	学校名	专业组	专业代码	再选科目	专业名称	人数	最低分	位次
2	2021	本科批	物理	3101	北京大学	05组	34	不限	理科试验班类	2	679	101以内
3	2021	本科批	物理	3103	清华大学	04组	18	化学	临床医学类(协和)	2	675	101以内
4	2021	本科批	物理	3101	北京大学	05组	38	不限	工商管理类	2	674	101以内
5	2021	本科批	物理	3115	北京大学医学部	03组	06	化学	口腔医学(八年制)	1	673	101以内
6	2021	本科批	物理	3101	北京大学	05组	26	不限	物理学类	2	669	101以内
7	2021	本科批	物理	3103	清华大学	03组	08	不限	自动化类	3	669	101以内
8	2021	本科批	物理	3101	北京大学	05组	32	不限	心理学类	1	668	101以内
9	2021	本科批	物理	3101	北京大学	05组	41	不限	新闻传播学类	1	667	101以内
10	2021	本科批	物理	3103	清华大学	03组	07	不限	计算机类	19	667	101以内
11	2021	本科批	物理	3103	清华大学	04组	17	化学	工科试验班(能源与电气)	1	667	101以内
12	2021	本科批	物理	3103	清华大学	04组	19	化学	临床医学类(医学实验班)	1	667	101以内
13	2021	本科批	物理	3101	北京大学	05组	29	不限	计算机类	8	666	101以内
14	2021	本科批	物理	3101	北京大学	05组	33	不限	信息管理与信息系统	2	666	101以内
15	2021	本科批	物理	3101	北京大学	05组	37	不限	经济学类	1	666	101以内
16	2021	本科批	物理	3103	清华大学	03组	06	不限	电子信息类	9	666	101以内
17	2021	本科批	物理	3103	清华大学	03组	09	不限	理科试验班类(数理)	2	666	101以内
18	2021	本科批	物理	3103	清华大学	03组	10	不限	理科试验班类(经济、金融与管理)	2	666	101以内
19	2021	本科批	物理	3103	清华大学	04组	16	化学	工科试验班(环境、化工与新材料)	1	666	101以内
20	2021	本科批	物理	2103	上海交通大学	04组	13	不限	金融学(含双学士学位)	1	662	101以内
21	2021	本科批	物理	2128	复旦大学上海医学院	01组	01	化学 或 生物	临床医学(八年制)	3	662	101以内
22	2021	本科批	物理	3101	北京大学	06组	44	化学	环境科学与工程类	2	662	101以内
23	2021	本科批	物理	3101	北京大学	06组	43	化学	环境科学	1	661	101
24	2021	本科提前批	物理	3101	北京大学	04组	23	不限	西班牙语	1	660	111
25	2021	本科批	物理	2101	复旦大学	05组	12	不限	经济学类	4	659	128
26	2021	本科批	物理	2103	上海交通大学	04组	11	不限	电子信息类(IEEE试点班，含双学士学位)	11	659	128
27	2021	本科批	物理	2103	上海交通大学	04组	12	不限	人工智能	7	659	128
28	2021	本科批	物理	3156	中国科学院大学	05组	17	不限	物理学	1	659	128
29	2021	本科批	物理	2101	复旦大学	06组	13	不限	数学类	7	658	146

Sheet1 Sheet2 Sheet3

至此，数据预处理工作得以全部完成。

四、 建模与问题解决

在本实验中，主要拟解决三个问题，分别是分数预测、学校波动情况与模拟报考。

4.1 考试分数预测

根据考生的高考分数，预测出考生在去年高考的大概分数，在高考填报志愿数据库中筛选掉分数相差特别大的学校及专业，对剩下报考的学校及专业进行处理。

首先建立高考成绩预测模型。高考分数和各省本科最低控制分数线发布后，考生可以根据高考的分数和所在省市分数线，对比去年的分数线，计算得到自己在去年的大概分数.预测分数公式为

$$d = c + (a - b), \tag{1}$$

其中：a 为去年的本省分数线；b 为今年的本省分数线；c 为考生的分数；d 为考生的高考成绩在去年的预测分数。

其次进行筛选数据。先在数据库中筛选出文科、理科。其次，将式（1）得到的分数 d 作为参照分数，在数据库的院校最低分数中选择的分数区间为[d-30,d+10]；在最高分数中选择的分数区间为[d,d+30]；使报考院校范围缩小.最后，根据省市和相关专业、录取人数进行条件筛选，再次缩小报考院校范围，得到部分适合报考的院校。

4.2 学校波动情况与报考概率

4.2.1 差分法确定学校波动类型

本文根据各院校不同专业每年最低录取分数的上下波动情况，采取差分法得到录取分数最大波动公式为：

$$\Delta M_j = \max(M_j) - \min(M_j), \tag{2}$$

其中：max(Mj)为近几年最低录取分中的最高分；min(Mj)为近几年最低录取分中的最低分。

某高校相邻两年同专业高考最低录取分数差公式为

$$\Delta M_i = M_i - M_{i-1}. \tag{3}$$

为预测录取分数并提高录取概率，由差分法判定各院校波动类型，分为平稳型、上升型和下降型，判定如下：

平稳型院校：ΔMj≤ΔMi；上升型院校：ΔMj>ΔMi，ΔMi>m(m 为定值)；下降型院校：ΔMj>ΔMi，ΔMi≤m(m 为定值)。

对高校波动类型进行分类后，为使考生能够了解各个高校不同专业今年的录取分数，以及所考分数被录取的概率，再次建立模型进行简单的分析计算。

4.2.2 预测录取成绩和录取概率模型

将各高校去年的专业最高分、最低分加上今年的本省分数线与去年分数线的差值，得出大概录取分数。对大概录取分数进行处理：若该专业为上升型，则在此成绩基础上减 5 分；若该专业为平稳型，则在此成绩不变；若该专业为下降型，则在此成绩基础上加 5 分。用上述方法可以大致推测出预测的录取成绩。

随后，根据预测的相关分数得到的录取概率模型为：

$$P = \frac{F - M_{\min}}{\Delta M}, \tag{4}$$

其中：F 为目标分数；Mmin 为预测分数最低分；ΔM 为预测的最高分与最低分的差值。当分数的录取风险概率 P 为 0～39%为高风险，40%～59%为中风险，60%～94%为低风险，95%～100%为保险型。

根据式（4）的结果，再逐个了解学校详细信息，选出最想考的几个院校专业进行下一步计算预测，计算报考各个学校的录取概率。最终综合考虑，可以得出最优的报考方案。

4.2.3 实例分析

问题：江苏省一理科考生 2022 年高考成绩总分为 580 分考生希望报考的大学离家近一些并且对师范类专业非常感兴趣，怎样填报高考志愿能够考上理想中的大学（第一志愿大学）？

(1) 初步筛选

根据式（1）可以推断出，本考生在 2021 年的分数大致为 580+（417-429）=568。根据上文，可以得出可选的低分数区间为[538,578]，高分数区间为[568,598]。同时，根据该学生给出的相应条件，可以在总表中得到以下信息：

A	B	C	D	E	F	G	H	I	J	K	L
年份	批次	科类	学校代码	学校名	专业组	专业代码	再选科	专业名称	人数	最低	位次
2021	本科批	物理	1381	扬州大学	19组	E9	不限	电气工程及其自动化	99	539	47564
2021	本科批	物理	1112	南京医科大学	07组	38	化学 或 生物	临床医学	449	587	13458
2021	本科批	物理	1106	南京信息工程大学	10组	A1	不限	计算机类(计算机科学与技术、软件工程、网络	443	561	29196
2021	本科批	物理	1106	南京信息工程大学	10组	A0	不限	电子信息类(电子信息工程、电子科学与技术、	420	559	30736
2021	本科批	物理	1381	扬州大学	24组	I0	化学 或 生物	临床医学	394	546	41356
2021	本科批	物理	1223	徐州医科大学	08组	84	化学 或 生物	临床医学	294	565	26308
2021	本科批	物理	1116	南京财经大学	03组	28	不限	金融学类(金融学、金融工程、保险学、投资学	233	549	38724
2021	本科批	物理	1222	江苏师范大学	14组	E1	化学 或 生物	生物科学(师范)	232	539	47564
2021	本科批	物理	1111	南京邮电大学	04组	36	不限	通信工程	211	576	19304
2021	本科批	物理	1116	南京财经大学	03组	46	不限	工商管理类(会计学、财务管理、审计学、资产	210	554	34573
2021	本科批	物理	1261	苏州大学	15组	B5	不限	电子信息类	177	588	13029
2021	本科批	物理	1223	徐州医科大学	08组	83	化学 或 生物	麻醉学	171	575	19888
2021	本科批	物理	1222	江苏师范大学	11组	B8	不限	数学与应用数学(师范)	168	556	32995
2021	本科批	物理	1119	南京审计大学	04组	30	不限	金融学	152	555	33778
2021	本科批	物理	1222	江苏师范大学	11组	C8	不限	物理学(师范)	152	549	38724
2021	本科批	物理	1116	南京财经大学	03组	26	不限	经济学类(经济学、经济统计学)	147	549	38724
2021	本科批	物理	1110	南京工业大学	06组	64	不限	土木类(土木工程、城市地下空间工程、铁道工	143	548	39585
2021	本科批	物理	1111	南京邮电大学	04组	38	不限	光电信息科学与工程	135	565	26308
2021	本科批	物理	1119	南京审计大学	04组	54	不限	审计学	130	566	25631
2021	本科批	物理	1301	南通大学	15组	E0	化学 或 生物	临床医学	130	546	41356
2021	本科批	物理	1111	南京邮电大学	04组	34	不限	电子信息工程	128	571	22337
2021	本科提前批	物理	1122	江苏警官学院	07组	28	化学 或 生物	刑事科学技术(浦口校区)	127	569	23578
2021	本科批	物理	1114	南京工程学院	03组	27	不限	电气工程及其自动化(电力系统及其自动化)	126	558	31480
2021	本科批	物理	1401	江苏大学	11组	66	不限	电气工程及其自动化	124	545	42191
2021	本科批	物理	1110	南京工业大学	06组	41	不限	机械类(机械工程、过程装备与控制工程、车辆	123	548	39585
2021	本科批	物理	1223	徐州医科大学	08组	86	化学 或 生物	口腔医学	120	564	27048
2021	本科批	物理	1109	中国药科大学	02组	04	化学	药学类(药学、药物制剂、药物分析、药物化学	106	576	19304
2021	本科批	物理	1223	徐州医科大学	08组	85	化学 或 生物	医学影像学	104	547	40477
2021	本科批	物理	1110	南京工业大学	06组	50	不限	计算机科学与技术	102	556	32995
2021	本科批	物理	1108	南京师范大学	22组	A2	化学 或 生物	生物科学类(生物科学(含师范、国家理科基地班	101	589	12546
2021	本科批	物理	1261	苏州大学	15组	A6	不限	数学类	100	589	12546
2021	本科批	物理	1111	南京邮电大学	04组	45	不限	计算机科学与技术	100	575	19888
2021	本科批	物理	1115	南京林业大学	06组	39	不限	机械类(机械设计制造及其自动化、机械电子工	100	547	40477

这些选出的学校都满足分数要求，且拟录取的人数偏多，竞争压力小，同时，在标黄的三所学校与专业中，更是满足了该同学的“师范类学校”，“离家近（江苏省内）”的要求。

(2) 概率分析

经过初步筛选，已经大大缩小了可选学校的范围，接下来，只需按照之前的步骤进行学校类型的波动性变化，经计算，上图所示的学校中，第 6、11、18、22、24 所学校为平稳型，第 1、3、4、7、12、15、16、23 为上升型，其余为下降型学校。

再经过概率计算，可以得到以下结果：

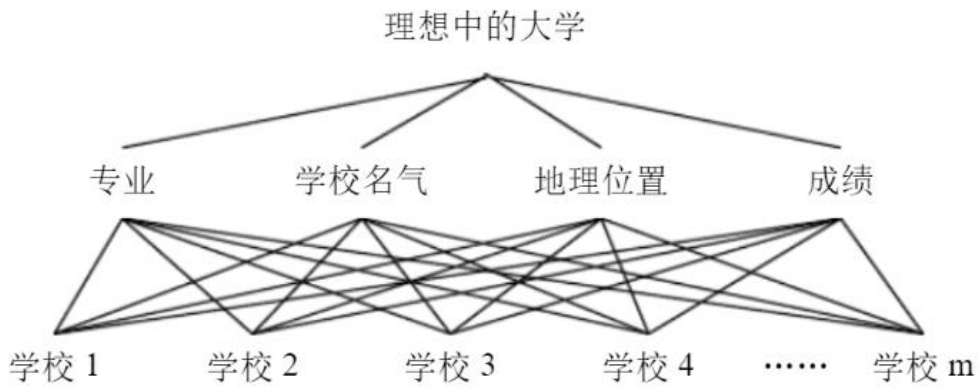
E	F	G	H	I	J	K	L	M
学校名	专业组	专业代码	再选科目	专业名称	人数	最低分	位次	概率
扬州大学	19组	E9	不限	电气工程及其自动化	99	539	47564	0.023766
南京医科大学	07组	38	化学 或 生物	临床医学	449	587	13458	0.371017
南京信息工程大学	10组	A1	不限	计算机类(计算机科学与技术、软件工程、网络	443	561	29196	0.55976
南京信息工程大学	10组	A0	不限	电子信息类(电子信息工程、电子科学与技术、	420	559	30736	0.993522
扬州大学	24组	I0	化学 或 生物	临床医学	394	546	41356	0.895271
徐州医科大学	08组	84	化学 或 生物	临床医学	294	565	26308	0.739864
南京财经大学	03组	28	不限	金融学类(金融学、金融工程、保险学、投资学	233	549	38724	0.801612
江苏师范大学	14组	E1	化学 或 生物	生物科学(师范)	232	539	47564	0.034671
南京邮电大学	04组	36	不限	通信工程	211	576	19304	0.583841
南京财经大学	03组	46	不限	工商管理类(会计学、财务管理、审计学、资产	210	554	34573	0.017859
苏州大学	15组	B5	不限	电子信息类	177	588	13029	0.214144
徐州医科大学	08组	83	化学 或 生物	麻醉学	171	575	19888	0.671669
江苏师范大学	11组	B8	不限	数学与应用数学(师范)	168	556	32995	0.911956
南京审计大学	04组	30	不限	金融学	152	555	33778	0.772546
江苏师范大学	11组	C8	不限	物理学(师范)	152	549	38724	0.901289
南京财经大学	03组	26	不限	经济学类(经济学、经济统计学)	147	549	38724	0.416003
南京工业大学	06组	64	不限	土木类(土木工程、城市地下空间工程、铁道工	143	548	39585	0.576207
南京邮电大学	04组	38	不限	光电信息科学与工程	135	565	26308	0.265983
南京审计大学	04组	54	不限	审计学	130	566	25631	0.574621
南通大学	15组	E0	化学 或 生物	临床医学	130	546	41356	0.765552
南京邮电大学	04组	34	不限	电子信息工程	128	571	22337	0.897046
江苏警官学院	07组	28	化学 或 生物	刑事科学技术(浦口校区)	127	569	23578	0.051824
南京工程学院	03组	27	不限	电气工程及其自动化(电力系统及其自动化)	126	558	31480	0.740316
江苏大学	11组	66	不限	电气工程及其自动化	124	545	42191	0.497337
南京工业大学	06组	41	不限	机械类(机械工程、过程装备与控制工程、车辆	123	548	39585	0.068224
徐州医科大学	08组	86	化学 或 生物	口腔医学	120	564	27048	0.755138
中国药科大学	02组	04	化学	药学类(药学、药物制剂、药物分析、药物化学	106	576	19304	0.688152
徐州医科大学	08组	85	化学 或 生物	医学影像学	104	547	40477	0.641375
南京工业大学	06组	50	不限	计算机科学与技术	102	556	32995	0.731081
南京师范大学	22组	A2	化学 或 生物	生物科学类(生物科学(含师范、国家理科基地班	101	589	12546	0.743093
苏州大学	15组	A6	不限	数学类	100	589	12546	0.710927
南京邮电大学	04组	45	不限	计算机科学与技术	100	575	19888	0.10483
南京林业大学	06组	39	不限	机械类(机械设计制造及其自动化、机械电子工	100	547	40477	0.395567

根据上表的概率分布结果和学生的自主意愿，可以进行最终的志愿报考的选择。

4.3 模拟报考评价模型

4.3.1 模型构建

如何填报一个既能使自己满意，又能够被录取的高校作为第一志愿，是每个考生所面临的一个相互关联、相互制约的众多因素构成的复杂系统。为此我们采用为解决这类复杂问题提供的一种新的、简洁的、实用的系统分析方法——层次分析法。基于上述分析，建立了高考志愿填报定量评价模型：



本实验中主要是选取了对高考志愿填报影响较大的几项因素，而前文提到的就业情况等信息在此处不作为决定因素考虑。

由于考生填报的理想中大学受到多个因素的限制，而对于不同的考生，每种因素的重要程度不同，由此造成了考生选择理想大学的侧重点不同。

根据上述的层次结构图，构造模糊互补判断矩阵，将所有因素进行两两对比，得到相对重要程度以及对应的数值。（下表所示为模糊标度及含义）

标度	含义
0.1	两个元素相比，后者比前者极端重要
0.3	两个元素相比，后者比前者明显重要
0.5	两个元素相比，后者比前者同等重要
0.7	两个元素相比，前者比后者明显重要
0.9	两个元素相比，前者比后者极端重要

可以看出按上述标度构建的模糊互补判断矩阵$A=(a_{ij})_{n\times n}$具有以下性质：

$$(1)a_{ij}+a_{ji}=1;\ (2)a_{ii}=0.5;\ (3)0<a_{ij}<1,i,j=1,2,...n$$

利用最小方差法（LVM）计算权重ω_k的公式为：

$$\omega_i=\frac{1}{n}\left(\sum_{j=1}^{n}a_{ij}+1-\frac{n}{2}\right),i=1,2,3,4 \quad (5)$$

从而计算出各个因素的权重大小：

$$\omega=(\omega_1,\omega_2,\omega_3,\omega_4) \quad (6)$$

4.3.2 指标处理

大学排名、专业排名可以通过上文给出的表进行查询，录取概率可以通过 4.2 给出的概率公式进行计算。

但地理位置是个模糊的概念，所以需对它进行量化处理。

	落后地区	二线城市	一线城市
落后地区	4	6	8
二线城市	2	2	4
一线城市	1	1	1

由于很大考生考虑到家庭以及未来的就业，所以如果本省较发达，很多人愿意留在本省；但如果本省比较落后，很多考生更愿意到发达地区去发展。所以规定，如果考生处于二线城市，对于该考生来说，本省可以视为一线城市。

此外，对于已经读取到数据的指标，为了后续计算的准确性，不可以直接使用，而是要先进行归一化处理（注意，其中大学排名、专业排名和地理位置的值都是逆向指标）。

进行归一化处理后可以得出，大学归一化值$a=(a_1,a_2,...,a_n)^T$，按照专业的排名为$b=(b_1,b_2,...,b_n)^T$，地理位置为$c=(c_1,c_2,...,c_n)^T$，录取的概率为$d=(d_1,d_2,...,d_n)^T$，由此得到大学的整体理想值为：

$$M=(a,b,c,d)\cdot\omega^T \tag{7}$$

考生可根据计算目标学校的理想值大小确定是否填报该大学，利用理想之大小排序确定志愿填写的顺序。

4.3.3 实例模拟

问题：考生小王处于江苏省，2022 年高考 600 分，有意报考法学专业。

根据问题可以看出，相较于其他因素，专业因素更重要于其它因素，专业的因素比较重要。由此，根据模糊标度表，得到模糊互补矩阵：

$$\begin{bmatrix} 0.5 & 0.4 & 0.7 & 0.8 \\ 0.6 & 0.5 & 0.8 & 0.9 \\ 0.3 & 0.2 & 0.5 & 0.6 \\ 0.2 & 0.1 & 0.4 & 0.5 \end{bmatrix}$$

并计算得出各个因素的权重大小：

$$\omega=(0.35,0.45,0.15,0.05)$$

从权重的大小情况来看，小王更注重专业和学校。而小王有意报考法律专业说明他报考时候兴趣在法学专业上,同时希望被法学专业录取。为此，选取学校时候法学专业的录取概率尽可能为 100%。如果录取法学的概率为 100%，那么被学校录取的概率也为 100%。利用录取概率模型，预测出 2022 年各高校在江苏省的录取分数线和法学专业录取分数线，并由此计算出高校录取概率和法学专业录取概率。

为了缩小选取范围，从高校录取概率和法学录取概率为 100%的所有大学中选取出综合排名和法学专业排名均较前的学校：

学校	综合排名	法学排名
吉林大学	13	7
中南财经政法大学	73	21
华中科技大学	10	42
中国政法大学	78	5
湖南大学	33	29

同时，对于小王来说，处于北京的高校的地理位置重要程度更大一些，设置为 1,；而长沙和长春同样作为二线城市，但是长春距离江苏的距离相对于长沙来说太远，因此将吉林大学的地理位置设置为 3，而湖南大学设置为 2。

利用预测得到的 2022 年专业录取平均分，可得到小王分数和专业录取平均分之间的分数线差。

将大学综合排名、法学专业排名、地理位置和分数线差归一化处理，得：

学校	综合排名	法学专业排名	地理位置	分数线差
吉林大学	0.538	0.429	0.333	0.830
中南财经政法大学	0.096	0.143	1.000	0.951
华中科技大学	0.700	0.071	1.000	0.684
中国政法大学	0.090	0.600	1.000	0.672
湖南大学	0.212	0.103	0.500	1.000

将四个因素指标的权重大小和各个大学的四个指标值代入式（7），可得到各个大学的整体理想值，并由此对大学选择进行排序：

学校	理想值	排名
吉林大学	0.488	1
中国政法大学	0.469	2
华中科技大学	0.446	3
中南财经政法大学	0.293	4
湖南大学	0.271	5

由此问题得以解决，为该同学的志愿报考给出了合理性建议。

五、 结论与不足

5.1 结论

本次实验，通过使用爬虫技术和官网提供的准确数据，获得了大量关于高考志愿填报的相关数据，这些数据，对于本届考生的填报具有巨大的参考价值。随后，使用 Python 对 Excel 表格的收集数据进行了去重去空处理，令数据更加符合规范，使用 Excel 表格自带的回归分析功能，预测出缺失的分数端或排名信息，加强了数据的完整性。最后，又通过将多张表格汇总的方式，得到了最终的总表格，对高考志愿填报有很大参考价值。

随后，使用差分法、层次分析法等方法，构建了分数预测、录取概率、报考理想值的数学模型，又通过实例进行验证，发现根据相应的模型，确实可以对不同类型的同学的报考需求进行评价和指导，给出的理想值更是对报考学校与专业具有很强的借鉴价值，学生和家长可以通过该模型进行模拟志愿填报指导。

5.2 不足

但是模型仍存在很多不足：

5.2.1 数据集不足

在数据收集的过程中，难免遇到数据的准确性和时效性的问题，还有一些学校没有官方的给出所需的数据集，这对报考该学校的学生的参考依据有很大的影响。其次，考虑因素不全面，因此数据集可能会出现不全面的情况，比如，在某家庭报考志愿时，对外界对大学评价比较看重，但数据集中没有类似数据，因此对该类的参考性较弱。

5.2.2 模型不足

对于分数预测和预测拟录取分数线的方法选取上存在不足，本文主要是借鉴了*高考考生志愿填报策略分析*[1]中的预测方法，但随着时间变化等其他因素，可能需要采用其他方法，考虑更多因素。

在最后的理想值计算中，仅采用了*高考志愿填报数学模型*[2]层次分析法作为权重标准，可能存在不严谨的情况，此外，不同考生对于专业、学校、环境的需求不同，不好使用单一的评价方法。因此，该模型在后期改进中，准备使用文本分析法和情绪分析法，爬取不同论坛中的关于志愿填报的建议、讨论的文本，随后进行情感分析，最后根据情感分析得出相应的比重，进行理想值的计算。这样可以帮助在填报志愿中很迷茫的学生，不知道该看重专业还是学校的学生，最大程度上的填报最优志愿。

六、 课程心得

通过本次课程的学习，在商务数据分析课程的基础上，学到了更多的数据收集和处理的方法，在实践中也遇到了各种各样的问题，通过对这些问题的求解，我深刻地感受到了自己能力的提升。

首先是数据收集方面，在课程作业中，我学习到了如何去找到可信度高、数据量大的数据源，并且在和同学们的交流过程中，得到了大量的经验，通过同学们作业的分享，收集了不同领域的数据和分析方法，在数据收集方面收货颇多，此外，还学会使用了以八爪鱼为代表的爬虫工具和数据收集网站，其中八爪鱼的使用令我最为印象深刻，从最开始的什么都不会，到最后可以自行通过软件进行目标网站的数据爬取，可以说是成就感满满。

其次是数学建模方面的成长，在商务数据分析课程所教的K-means、层次分析法等方法的基础上，我学到了更多的数学模型，可以通过这些模型解决实际问题，令我由衷的感到兴奋、快乐。同时，利用 Python 语言进行求解也使我得到了编程方面的提升。

数学建模是一个经历观察、思考、归类、抽象与总结的过程，也是一种信息捕捉、筛选、整理的过程，更是一个思想与方法的产生与选择的过程。它给学生再现了一种“微型科研”的过程。数学建模还激发我学习数学的兴趣；有利于我自觉检验、巩固所学的数学知识，促进知识的深化、发展，此外，还有利于我对数学思想方法体会和感悟。

为了使描述更具科学性，逻辑性，客观性和可重复性，人们采用一种普遍认为比较严格的语言来描述各种现象，这种语言就是数学。只有经历不断的探索过程，数学的思想、方法才能沉积、凝聚，从而使知识具有更大的智慧价值。动手实践、自主探索与合作交流是学习数学建模的重要方式。整体的学习过程也随之变成了一个主动、活泼的、生动和富有个性的过程。

此外，数学模型的搭建还需要考虑到简单性和普适性，在课程设计的过程中，参考了大量的资料，但最后可以借鉴的只有寥寥几篇，因为有很多文献只是为了使用模型而使用模型，使用了各种稀奇古怪的模型，不仅原理复杂，还没有很大的实际作用，求解得到的结果对结论没有直接关系。因此，在之后的模型搭建中，也需要注意构建简单明了的、符合生活实际情况的数学模型。

总体看来，整个课程虽然老师教学方面较少，但通过自己的学习和与同学们的交流，也获得了很大的收货，也希望之后可以继续选修徐老师的课程，获得其他方面的提升。

七、 参考文献

[1] 王冰杰,林洋,马靖敏.高考考生志愿填报策略分析[J].白城师范学院学报,2022,36(02):104-110.

[2] 高考志愿填报数学模型,高考志愿填报数学模型 - 百度文库 (baidu.com)

[3] 2016 数学建模 高考志愿填报模型, 2016 数学建模 高考志愿填报模型 - 百度文库 (baidu.com)

[4] 数学建模高考志愿填报模型 2021, 数学建模高考志愿填报模型 2021.docx-原创力文档 (book118.com)

[5] 牟锋.“3+1+2”背景下的高考志愿填报[J].考试与招生,2022(Z1):7-11.

[6] 高考志愿填报应避开哪些误区[N]. 北京日报,2022-06-08(020).DOI:10.28033/n.cnki.nbjrb.2022.003113.

[7]河海大学信息门户河海大学信息门户 (hhu.edu.cn)

[8]江苏省教育考试院江苏省教育考试院 (jseea.cn)

[9]大学生必备网河海大学专业排名 最好的专业有哪些_大学生必备网 (dxsbb.com)

[10]河海大学招生信息网首页 - 河海大学招生信息网 (hhu.edu.cn)

[11]河海大学就业创业信息网河海大学就业服务管理平台 (91job.org.cn)