江西财经大学
JIANGXI UNIVERSITY OF FINANCE AND ECONOMICS

# 课程名称：Python语言与数据分析

# 课 程 报 告

项目名称 考研数据分析

班　　级 金融201

学　　号 0204752

姓　　名 易玲

任课教师 肖　泉

开课学期： 2020 至 2021 学年 第 二 学期

完成时间： 2021 年 7 月 4 日

# 《考研数据分析》数据分析报告

# 目　录

## 目录

# 1 概述

随着社会的发展，科学的进步，信息化的提高，这个时代对于人才的要求越来越高，竞争越来越大。而高校连年扩招，每年毕业生人数连年增长，据教育部数据，2020 届高校应届毕业生总数将达到 874 万，较 2019 年增加 40 万人，再创历史新高。面对疫情过后经济下行的压力，工作岗位没有明显增加的现实，毕业生们的情况属实不算乐观。为了更好适应这个时代的要求，就有非常多的人选择去考研，提升自己的实力。但是在考研选择院校和专业的过程中，出现了各种各样的问题。现根据已有数据对考研历年国家线和相关信息进行数据分析，了解考研报考情况，录取情况等，有助于考生对考研院校及相关事项作出决策。

# 2 数据描述

## 2.1 数据来源、特点及包含字段

数据来源于和鲸社区
网址为 https://www.heywhale.com/mw/dataset/5fe1706383e4460030ab004f
总共 100864 条数据，13 个字段

年份
学校名称_链接
学校名称
院系名称_链接
院系名称
专业代码
专业名称_链接
专业名称
总分
政治__管综
外语
业务课_一
业务课_二

## 2.2 导入数据

下载好数据后使用 Pandas 导入数据，查看数据集的信息，快速理解数据

```
年份           100864 non-null int64
学校名称_链接    100864 non-null object
学校名称         100863 non-null object
院系名称_链接    100864 non-null object
院系名称         100725 non-null object
专业代码         100761 non-null object
专业名称_链接    100864 non-null object
专业名称         100864 non-null object
总分           100864 non-null object
政治__管综       100864 non-null object
外语           100864 non-null object
业务课_一        100864 non-null object
业务课_二        100864 non-null object
```

根据以上结果，对数据进行基本了解，13 个字段中只有 1 个字段是数字类型，其他字段的数据都是 object 类型，如需计算需要转换类型。

2.3 数据清洗

2.3.1 查看是否含有缺失值

```
各字段是否含有空值情况:
 年份           False
学校名称_链接    False
学校名称         True
院系名称_链接    False
院系名称         True
专业代码         True
专业名称_链接    False
专业名称         False
总分           False
政治__管综       False
外语           False
业务课_一        False
业务课_二        False
dtype: bool
```

由上可知，缺失值很多，需要进行处理。

### 2.3.2 处理重复值和空值

```
重复值:  58
空值:
 年份              0
学校名称_链接        0
学校名称             1
院系名称_链接        0
院系名称           139
专业代码           103
专业名称_链接        0
专业名称             0
总分               0
政治__管综           0
外语               0
业务课_一            0
业务课_二            0
dtype: int64
```

因为无法补充空值，而且数据基数很大，删除少量数据不会有太大影响，所以删除重复值和空值，返回结果如下图。

```
年份           100563 non-null int64
学校名称_链接     100563 non-null object
学校名称         100563 non-null object
院系名称_链接     100563 non-null object
院系名称         100563 non-null object
专业代码         100563 non-null object
专业名称_链接     100563 non-null object
专业名称         100563 non-null object
总分           100563 non-null object
政治__管综       100563 non-null object
外语           100563 non-null object
业务课_一        100563 non-null object
业务课_二        100563 non-null object
```

重复值： 0
空值：
年份 0
学校名称_链接 0
学校名称 0
院系名称_链接 0
院系名称 0
专业代码 0
专业名称_链接 0
专业名称 0
总分 0
政治__管综 0
外语 0
业务课_一 0
业务课_二 0

## 2.4 数据整理

### 2.4.1 删除不需要的列

因为链接不对数据分析造成影响，所以删除数据表中所有的链接列。

| | 年份 | 学校名称 | 院系名称 | 专业代码 | 专业名称 | 总分 | 政治__管综 | 外语 | 业务课_一 | 业务课_二 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2020 | 中国人民大学 | 公共管理学院 | 125200 | (专业学位)公共管理 | 175 | 88 | 44 | - | - |
| 1 | 2020 | 中国人民大学 | 公共管理学院 | 125200 | (专业学位)公共管理 | 175 | 88 | 44 | - | - |

### 2.4.2 替换删除特殊字符

因为特殊字符不利于数据分析，且对分析结果没有影响，所以删除数据表中所有的特殊字符。

| | 年份 | 学校名称 | 院系名称 | 专业代码 | 专业名称 | 总分 | 政治__管综 | 外语 | 业务课_一 | 业务课_二 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2020 | 中国人民大学 | 公共管理学院 | 125200 | 公共管理 | 175 | 88 | 44 | - | - |
| 1 | 2020 | 中国人民大学 | 公共管理学院 | 125200 | 公共管理 | 175 | 88 | 44 | - | - |

2.4.3 单独筛选出 2020 年考研信息

```
年份          65991 non-null int64
学校名称        65990 non-null object
院系名称        65857 non-null object
专业代码        65990 non-null object
专业名称        65991 non-null object
总分          65991 non-null object
政治__管综      65991 non-null object
外语          65991 non-null object
业务课_一       65991 non-null object
业务课_二       65991 non-null object
```

# 3 数据分析内容

3.1 统计 2020 考研专业数及 2020 年考研专业 Top5

专业数是一个在考研专业选择中很重要的数据，通过了解考研专业数，我们能够得知哪些专业开设比较热门，比较大众。专业数越多，就代表有越多的院校选择，越少则相反。我们可以根据 2020 考研专业数分析出每个专业大概的开设数目情况，能够依据其数据并结合自身兴趣进行考研专业的选择。此外本次分析还挑选出 2020 年考研专业前 5，作为一个热门内容，供大家参考。

| | |
|---|---|
| 工商管理 | 1059 |
| 计算机科学与技术 | 880 |
| 管理科学与工程 | 844 |
| 内科学 | 809 |
| 数学 | 808 |
| 材料科学与工程 | 740 |
| 公共管理 | 716 |
| 机械工程 | 687 |
| 控制科学与工程 | 681 |
| 马克思主义理论 | 623 |
| 化学 | 589 |
| 药学 | 585 |
| 物理学 | 568 |
| 会计 | 565 |
| 艺术设计 | 560 |
| 设计学 | 554 |
| 生物学 | 530 |
| 信息与通信工程 | 526 |
| 化学工程与技术 | 501 |

| | |
|---|---|
| 土木工程 | 465 |
| 金融 | 461 |
| 音乐 | 461 |
| 环境科学与工程 | 459 |
| 软件工程 | 457 |
| 应用经济学 | 445 |
| 统计学 | 442 |
| 临床医学 | 436 |
| 美术学 | 426 |
| 美术 | 420 |
| 音乐与舞蹈学 | 406 |
| 中国史 | 402 |
| 电气工程 | 399 |
| 法学 | 380 |
| 外科学 | 380 |
| 企业管理 | 375 |
| 电子科学与技术 | 372 |
| 动力工程及工程热物理 | 364 |
| 光学工程 | 363 |
| 法律（非法学） | 348 |
| 法律（法学） | 344 |
| 生物医学工程 | 340 |

| | |
|---|---|
| 会计学 | 330 |
| 建筑学 | 330 |
| 金融学 | 328 |
| 中药学 | 325 |
| 社会工作 | 315 |
| 应用统计 | 312 |
| 思想政治教育 | 309 |
| 旅游管理 | 308 |
| 生态学 | 303 |
| 护理 | 300 |
| 新闻与传播 | 297 |
| 课程与教学论 | 295 |
| 公共卫生 | 280 |
| 中国语言文学 | 279 |
| 外国语言文学 | 273 |
| 外国语言学及应用语言学 | 268 |
| 生物化学与分子生物学 | 267 |
| 英语语言文学 | 262 |
| 风景园林 | 261 |
| 力学 | 250 |
| 教育学 | 246 |

| | |
|---|---|
| 中西医结合临床 | 242 |
| 微生物学 | 241 |
| 口腔医学 | 241 |
| 马克思主义中国化研究 | 237 |
| 技术经济及管理 | 233 |
| 环境科学 | 227 |
| 基础数学 | 226 |
| 中国古代文学 | 220 |
| 产业经济学 | 220 |
| 影像医学与核医学 | 218 |
| 药理学 | 217 |
| 马克思主义基本原理 | 214 |
| 材料物理与化学 | 212 |
| 基础医学 | 212 |
| 行政管理 | 211 |
| 仪器科学与技术 | 211 |
| 环境工程 | 210 |
| 食品科学与工程 | 210 |
| 农艺与种业 | 209 |
| 社会学 | 206 |

```
儿科学                 204
文艺学                 203
世界史                 198
凝聚态物理              197
国际商务                194
中国现当代文学            193
植物学                 191
英语笔译                189
遗传学                 185
应用数学                181
风景园林学              180
神经病学                179
应用心理                178
计算机应用技术            178
网络空间安全              177
材料学                 175
中医内科学              173
国际贸易学              173
Name: 专业名称, dtype: int64
```

3.2 分组归纳学校对应的专业数及学校专业数排名 Top20

通过分组归纳学校对应的专业数，我们可以得知不同大学开设研究生专业的数量，能够从数据中看出大学的教育水平，一般来讲，教育水平越高的高等院校开设的研究生专业数越多。而分组归纳每一个学校开设的专业数能够让考生更好的了解心仪学校开设的专业信息，能够知道自己心仪的学校有多少专业供自己选择。进而更好地决定自己是要选择心仪的学校，还是选择心仪的专业，还是二者兼得。分析这组数据为考生选择报考院校及报考专业有一定的帮助作用。

| 学校名称 | |
|---|---|
| 北京大学 | 3093 |
| 复旦大学 | 1248 |
| 清华大学 | 1217 |
| 武汉大学 | 858 |
| 中国人民大学 | 758 |
| 厦门大学 | 742 |
| 南开大学 | 728 |
| 四川大学 | 682 |
| 山东大学 | 646 |
| 南京大学 | 646 |
| 云南大学 | 626 |
| 北京工业大学 | 625 |
| 同济大学 | 603 |
| 中国科学院大学 | 603 |
| 华东师范大学 | 592 |
| 昆明理工大学 | 554 |
| 北京师范大学 | 534 |
| 北京科技大学 | 530 |
| 首都师范大学 | 517 |
| 北京理工大学 | 515 |

福州大学 514
南京理工大学 510
东南大学 496
重庆大学 463
华中科技大学 461
天津师范大学 458
暨南大学 443
西北师范大学 426
中国农业大学 425
华南师范大学 414
...
北京中医药大学 253
华中师范大学 253
中国地质大学（北京） 250
四川农业大学 244
湖南大学 244
广西医科大学 242
浙江大学 241
华东理工大学 239
陕西中医药大学 239
电子科技大学 239

| | |
|---|---|
| 上海理工大学 | 238 |
| 哈尔滨师范大学 | 237 |
| 上海财经大学 | 235 |
| 江西财经大学 | 232 |
| 福建医科大学 | 230 |
| 武汉理工大学 | 229 |
| 湖南中医药大学 | 227 |
| 重庆师范大学 | 226 |
| 华南理工大学 | 226 |
| 首都医科大学 | 223 |
| 东北大学 | 222 |
| 汕头大学 | 222 |
| 江西农业大学 | 220 |
| 内蒙古师范大学 | 219 |
| 中国矿业大学 | 218 |
| 河南科技大学 | 211 |
| 华北理工大学 | 208 |
| 上海师范大学 | 207 |
| 吉林农业大学 | 205 |
| 沈阳农业大学 | 203 |

3.3 分组归纳各专业的最高分，最低分，平均分

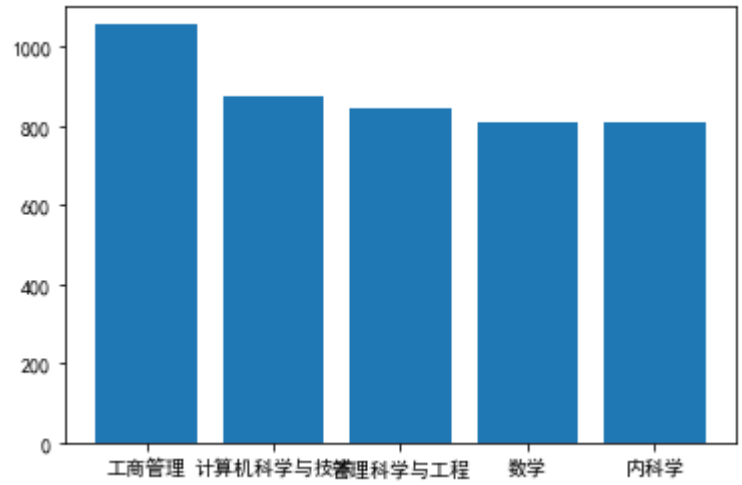
通过分组归纳各专业的最高分，最低分，平均分，能够让考生更好地了解自己处于哪个层次，以及自己想报的专业处在哪个层次，将要面临怎样的竞争，自己是否有能力能够考得上心仪的专业。同时直观数据的分析也有利于考生在选择自己考研方向作出一个理性的判断，通过思考自身能力以及状况之后作出一个适合自己的选择。

| 专业名称 | mean | amax | amin |
| --- | --- | --- | --- |
| 公共关系学 | 409 | 409 | 409 |
| 国别与区域研究 | 402 | 402 | 402 |
| 美术理论研究 | 397 | 397 | 397 |
| 信息艺术设计 | 390 | 390 | 390 |
| 国民经济动员学 | 388 | 400 | 376 |
| 高级秘书与行政助理学 | 386 | 386 | 386 |
| 公共组织与人力资源 | 385 | 385 | 385 |
| 中国政治 | 383 | 383 | 383 |
| 网络与新媒体 | 382 | 382 | 382 |
| 符号学 | 382 | 382 | 382 |
| 传媒艺术学 | 381 | 381 | 381 |
| 媒介管理学 | 380 | 380 | 380 |
| 非传统安全 | 380 | 380 | 380 |
| 公共财政与公共政策 | 380 | 380 | 380 |
| 政府经济管理 | 379 | 379 | 379 |
| 地方政府与社会治理 | 377 | 377 | 377 |
| 电子政务 | 376 | 380 | 360 |
| 犯罪心理学 | 375 | 375 | 375 |
| 城乡发展与规划 | 375 | 375 | 375 |
| 公共管理信息化理论与技术 | 375 | 375 | 375 |

| | | | |
|---|---|---|---|
| 能源与气候经济 | 375 | 376 | 375 |
| 文化研究 | 372 | 378 | 355 |
| 医院管理与卫生政策 | 372 | 372 | 372 |
| 物流管理与工程 | 372 | 372 | 372 |
| 医药信息系统 | 370 | 370 | 370 |
| 互联网金融学 | 370 | 370 | 370 |
| 法与经济学 | 370 | 370 | 370 |
| 艺术与科学 | 370 | 370 | 370 |
| 电视电影与视听传播学 | 370 | 370 | 370 |
| 创业学 | 370 | 370 | 370 |
| 军事法学 | 370 | 370 | 370 |
| 饭店管理 | 370 | 370 | 370 |
| 文化传播学 | 370 | 370 | 370 |
| 企业经济学 | 368 | 368 | 368 |
| 网络经济学 | 368 | 368 | 368 |
| 财务学 | 368 | 370 | 365 |
| 对外汉语教学 | 368 | 370 | 365 |
| 话语与传播 | 368 | 368 | 368 |
| 金融统计、保险精算与风险管理 | 367 | 367 | 367 |
| 广播电视学 | 367 | 380 | 355 |
| 市场营销管理 | 365 | 365 | 365 |
| 体育文化与管理 | 365 | 365 | 365 |
| 信息分析 | 365 | 365 | 365 |
| 比较法学 | 365 | 365 | 365 |
| 食品经济管理 | 365 | 365 | 365 |
| 新媒体 | 365 | 365 | 365 |
| 广告学 | 365 | 380 | 355 |
| 汉语国际推广 | 365 | 365 | 365 |
| 自然资源管理 | 365 | 365 | 365 |
| 党的历史与理论 | 365 | 365 | 365 |

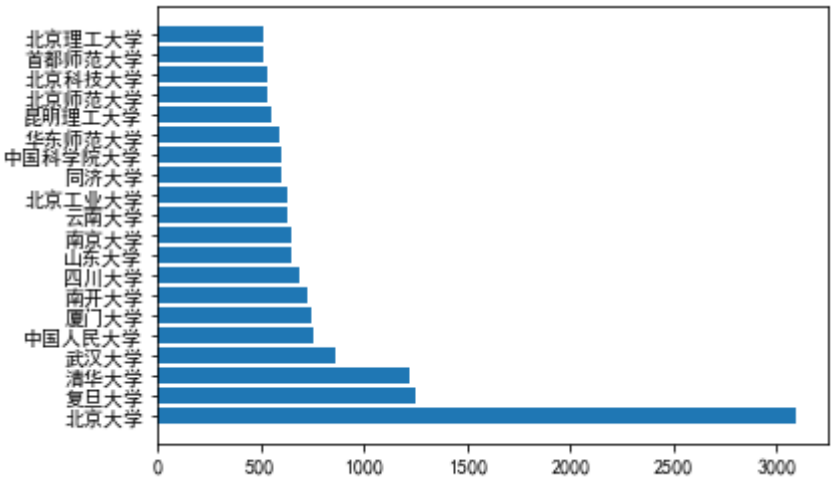
# 4 数据分析图表

4.1 考研专业 Top5

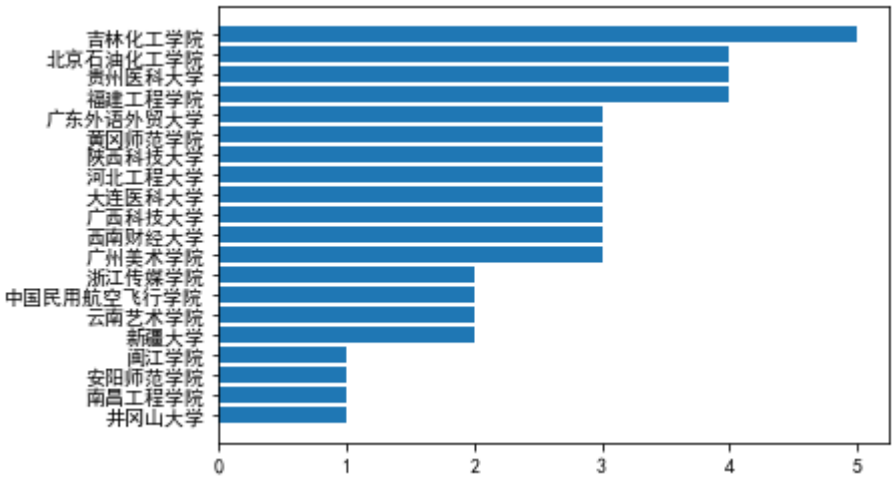



通过分析上图，我们可以得知开设数量最多的专业是工商管理，其次是计算机科学与技术，管理科学与工程，数学，内科学。这里我们可以直观的看到，几个比较热门的专业。这些专业的选择院校很多，如果想报考以上专业，会有更多的院校选择，但是开设专业多也意味着竞争更大，我们可以结合分析 3.1 看出专业开设数量的大致情况，能够结合自己的兴趣爱好以及特长，同时兼顾考研的压力和风险对自己的考研方向进行一个选择。

4.2 学校研究生专业数 Top20

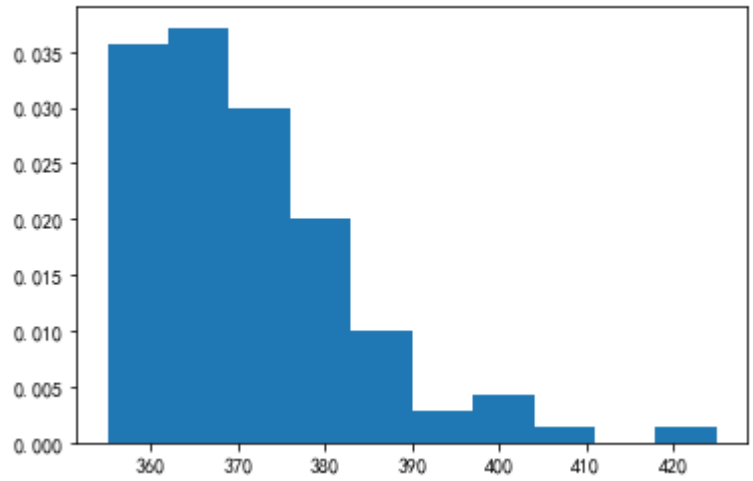



学校研究生专业数排名后 20 位




通过上两张图，我们可以直观的发现，一般来讲，拥有越多资源的学校所开设的研究生专业越多，选择也更多。越高等的学府越注重更深层次的教育，清北复交等名校虽然竞争激烈，但仍是很多考研学子的选择。同时，在北京的一些学校也因为地域和财政等优势拥有较多的研究生专业，也是除清北复交外较好的选择。而一些比较偏远的地方的学校开设专业会比较少，如果要选择这类学校要注意学校是否开设有自己想去的专业。

4.3 各专业的最高分分布频率直方图




通过分析 3.3 和上图，我们可以知道每个专业的最高分，最低分，平均分，能够结合自身的能力，对自己所能够选择的专业做一定的规划。同时各专业的最高分还是大部分集中在 360 到 380 之间，我们可以根据这个数据对自己做一个规划。同时也能够了解自己现在处在什么层次，需要多少努力才能上自己心仪的专业，而录取分数较高的一些专业一般代表着竞争较大，需要谨慎考虑。

# 5 数据分析结果

通过分析考研历年国家线的数据，我们可以了解到考研专业的热门情况，不同院校开设专业的情况，也能够了解到不同专业录取分数的情况，我们可以从这

些分析当中了解自己想要的信息。然后根据这些分析，对自身的情况也做一个对照的分析，在考研的专业和院校选择上做出一个更合理的判断。例如像工商管理，计算机科学与技术这些都是开设院校比较多的专业，如果你想选择这些专业的话，你就会有更多的院校选择。相较于此，材料学，中医内科学，国际贸易学等，开设这些专业的学校就比较少，选择院校的空间就不是很大，但是这也意味着考研的难度一般会低于热门专业，如果自己有兴趣，可以朝这些方面发展。通过分析，我们还能够看到不同大学的研究生专业开设数目情况，其中以北京大学为最，复旦大学其次。我们可以看到越好的高等教育学府学校开设的专业越多，给予我们越多的选择。最后就是通过对不同专业录取分数的分析，我们可以知道不同专业的录取分数大概是在多少。其中公共管理学的录取分数是最高的，达到了 409 分，这就与前面工商管理类所开设专业最多也有一定承担上的关联，专业越热门，考的人数相应的也会越多，竞争也会越激烈。我们在选择专业的时候，如非必要，可以适当避开热门专业。同时我们观察分数可以看到大部分的专业录取分数是在 360~380 之间，我们可以根据这个数据对自己现在处在一个什么位置也进行一个规划，能够让我们更好的规划自己的复习计划以及自己的院校选择

# 6 总结

在整个数据分析的过程中做的第一件事情就是去网站上寻找自己要做的课题的数据。此次数据分析的数据来源于和鲸社区，是关于考研的数据。然后下载数据，导入数据，进行数据清洗等一系列的基本操作，然后就是根据数据文件中有的一些信息，进行一些数据分析。例如在数据中有 2020 年考研的各种专业，于是我就统计出了 2020 年考研的专业数和研究生专业开设最多的前五个专业，在数据中还有各专业的录取分数，于是我也将他们整理出来，分为最高分、最低分、平均分，做了一个可视化的图表，便于观察，这些信息都有利于考生对报考专业作出选择。其次就是整理了一下关于不同的大学开设专业数目的问题，通过做这些分析，我对考研的这一个话题也有了一些新的了解，知道了考研的一些热门内容，和与热门相对应的就是分数会比较高，在自己做出选择的时候，可以适当的在这方面做一些规避

# 附录-数据分析代码

2.2 使用 Pandas 导入数据，查看数据集的信息：

```
import pandas as pd
df1 = pd.read_csv('考研历年国家分数线(1).csv')
df2 = pd.read_csv('考研历年国家分数线(2).csv')
df3 = pd.read_csv('考研历年国家分数线(3).csv')
df4 = pd.read_csv('考研历年国家分数线(4).csv')
df5 = pd.read_csv('考研历年国家分数线(5).csv')
df6 = pd.read_csv('考研历年国家分数线(6).csv')
df_all= pd.concat([df1,df2,df3,df4,df5,df6])
df_all.info()
```

2.3.1 查看是否含有缺失值：

```
print('各字段是否含有空值情况：\n', df_all.isna().any())
```

2.3.2 处理重复值和空值：

```
print('重复值：' ,df_all.duplicated().sum())
print('空值: \n',df_all.isnull().sum())
```

删除重复值和空值：

```
df_all = df_all.drop_duplicates()
df_all = df_all.dropna(axis=0,how='any')
df_all.info()
print(df_all.shape)
print('重复值：' ,df_all.duplicated().sum())
print('空值: \n',df_all.isnull().sum())
```

2.4.1 删除不需要的列：

```
df_all = df_all.drop(labels=['学校名称_链接','院系名称_链接','专业名称_链接'],axis=1)
df_all.head(2)
```

2.4.2 替换删除特殊字符：

```
df_all['专业名称'] = df_all['专业名称'].str.replace('\(专业学位\)','')
df_all['专业名称'] = df_all['专业名称'].str.replace('★','')
df_all.head(2)
```

2.4.3 单独筛选出 2020 年考研信息：

```
data_2020 = df_all[df_all['年份'] == 2020]
data_2020.info()
```

3.1 统计 2020 考研专业数及 2020 年考研专业 Top5:

```
data_2020['专业名称'].value_counts()[:100]
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif']=['SimHei']
data_2 = data_2020['专业名称'].value_counts()[:5]
data_x =data_2.index.tolist()
data_y = data_2.values.tolist()
plt.bar(data_x,data_y)
plt.ylim([0,1100])
plt.show()
```

3.2 分组归纳学校对应的专业数及学校专业数排名 Top20：

```
data_3=data_2020.groupby('学校名称')['专业名称'].count().sort_values(ascending = False)[:20]
data_3x =data_3.index.tolist()
data_3y = data_3.values.tolist()
plt.barh(data_3x,data_3y)
```

排名后 20：

```
data_3=data_2020.groupby('学校名称')['专业名称'].count().sort_values()[:20]
data_3x =data_3.index.tolist()
data_3y = data_3.values.tolist()
plt.barh(data_3x,data_3y)
```

3.3 分组归纳各专业的最高分，最低分，平均分：

```
def tranform_num(x):
    if '-' in x:
        return 0
    else:
        return x
data_2020['总分'] = data_2020['总分'].apply(lambda x:tranform_num(x) )
data_2020['总分'] = data_2020['总分'].astype('int')
data_1 = data_2020.groupby('专业名称')['总分'].agg([np.mean,np.max,np.min])
data_1['mean'] = data_1['mean'].astype('int')
data_1 = data_1.sort_values(by=['mean'],ascending=False)[:100]
print(data_1)
```

各专业最高分分布频率直方图：

```
data_11=data_1['amax'].tolist()
plt.hist(data_11,bins=10,density=1)
```