



武汉大学  
WUHAN UNIVERSITY

# 武汉大学数据探索与可视化 结课报告

《中国上市公司的分布与疫情下发展状况  
的探究——对中国 A 股 3000 余家上市公司  
的数据探索与机器学习》

专业名称 : 管理科学

课程名称 : 数据探索与可视化

指导教师 : 李永泉

学生学号 : 2021301052056

学生姓名 : 陈凯强

二〇二三年十二月

# 目录

一、前言.....	3
1. 选题动机 .....	3
2. 数据来源.....	4
3. 数据选择与处理.....	5
二、数据探索与可视化.....	7
2.1 上市公司分布特征探索.....	7
2.2 疫情对不同行业影响的探索 .....	9
三、机器学习.....	12
3.1 PCA.....	12
3.2 SOM.....	13
3.3 K-means .....	13
3.4 Diana .....	15
3.5 SNE .....	16
四、总结与反思.....	16
五、附录.....	18
4.1 前言 .....	18
4.2 数据探索与可视化.....	18
4.3 机器学习.....	20

# 一、前言

## 1. 选题动机

前阵子在网上看到许多讲疫情对企业影响的相关文章（如疫情下有些行业实现爆发式增长；疫情促使企业反思数字化转型等），激发了选题兴趣，想探究一下疫情下不同类型企业的发展状况。由于众多企业数据具有隐私性非公开，而上市公司有义务在年报中披露业绩情况，因此本文从上市公司数据着手，探究上市公司行业分布以及公司规模、财务信息等维度的时间序列数据的特点与关系，并试图以上市公司为着眼点探究不同行业在疫情影响下（19-22年）的发展情况。

### 疫情下爆发式增长的20大行业现状和前景分析

HOLLY · 关注  
2020-03-09  
3 评论 · 51471 浏览 · 43 收藏 · 77 分钟  
释放双眼，带上耳机，听听看~!  
00:00 · 00:00

这次的疫情会对中国经济转型和行业发展方向及前景将会有所改变，甚至是深刻改变。本篇文章将分析，中国在疫情影响下可能爆发的20个行业。20个行业可大体分为医疗篇、办公篇、娱乐篇、生活篇、科技篇。

图1 一篇介绍疫情下爆发式增长的行业的文章  
(来源：人人都是产品经理)

## 郝志强：疫情迫使企业反思数字化转型的必要性与迫切性

财经杂志  
2020-09-28 10:09:55 发布于北京 财经杂志官方账号  
+关注

“在新冠疫情的影响下，不少制造企业面临着组织管理困境、资金链断裂、市场需求缩水等危机。与此同时，也有部分企业表现出强韧的生命力，在危机处理、响应能力与恢复速度等方面交出满意的答卷。疫情使得越来越多的传统企业重新审视数字化转型的必要性和紧迫性，一批企业逆势而上，化危为机，在敏捷响应市场的同时，为应对疫情做出贡献。”9月27日，国家工业信息安全发展研究中心研究员、纪委书记郝志强在以“强韧、创新、突破”为主题的“2020年中国企业数字转型指数报告发布论坛”上如此表示。该论坛由埃森哲中国与国家工业信息安全发展研究中心联合主办。

图2 一篇介绍疫情促使企业反思数字化转型的文章  
(来源：财经杂志)

## 2. 数据来源

根据选题动机和兴趣，主要想获取的数据是上市公司的行业分布、公司规模（想要通过雇员数量和公司收入来表征）以及一些财务信息（如净利润等，用于衡量公司的经营发展情况）。

首先考虑利用 Tushare API 接口，通过 Python 来实现<sup>1</sup>（代码见附录 1.1），可以成功获取上市公司的证券代码(ts\_code)、地域(area)、行业(industry)、上市日期(list\_date)、雇员数等数据。但是由于 Tushare 有接口调用的限制，非 VIP 用户难以获取所有数据。经过研究和询问，发现可以通过 iFind、Wind、Choice 等金融终端获取企业的财务数据十分方便。由于 Choice 有教育优惠，大学生可以免费使用，因此最终选择使用 Choice 的 Excel 插件来获取所有数据。



图 Tushare 官网（链接：[Tushare 数据](#)）

	ts_code	symbol	name	area	industry	list_date
0	000001.SZ000001		平安银行	深圳	银行	19910403
1	000002.SZ000002		万科A	深圳	全国地产	19910129
2	000004.SZ000004		国华网安	深圳	软件服务	19910114
3	000005.SZ000005		ST星源	深圳	环境保护	19901210
4	000006.SZ000006		深振业A	深圳	区域地产	19920427
5	000007.SZ000007		*ST全新	深圳	其他商业	19920413
6	000008.SZ000008		神州高铁	北京	运输设备	19920507
7	000009.SZ000009		中国宝安	深圳	电气设备	19910625
8	000010.SZ000010		美丽生态	深圳	建筑工程	19951027
9	000011.SZ000011		深物业A	深圳	区域地产	19920330
10	000012.SZ000012		南玻A	深圳	玻璃	19920228

图 通过 Tushare 获取的部分数据截图

```
C:\Users\10136\Desktop\Python\Scripts\python.exe C:\Users\10136\PycharmProjects\pythonProject\1206_MachineLearning.py
Traceback (most recent call last):
  File "C:\Users\10136\PycharmProjects\pythonProject\1206_MachineLearning.py", line 11, in <module>
    filtered_df['2019_revenue']= pro.income(ts_code=i, period='20191231', fields='total_revenue')['total_revenue'] #n_income 净利润(含少数股东损益)
                                                               ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "C:\Users\10136\Desktop\Python\Lib\site-packages\tushare\pro\client.py", line 45, in query
    raise Exception(result['msg'])
Exception: 抱歉, 您每分钟最多访问该接口200次, 权限的具体详情访问: https://tushare.pro/document/1?doc\_id=108.
```

图 Tushare 有调用接口限制，难以获取所有数据

<sup>1</sup> 代码见附录 1.1

图 通过 Choice 的 Excel 插件可以批量获取企业的各种信息与财务数据

### 3. 数据选择与处理

#### 1. 时间维度选择

由于需要探究疫情的影响，所以需要选取公司在疫情前和后的数据，疫情与 19 年底爆发，考虑到 19 年刚上市的公司业绩数据可能存在较大的不稳定性，故选择 18 年及之前上市的所有上市公司。

#### 2. 变量选择与处理

Choice 中导出的数据字段包括：

- Industry 行业
- Employee# 雇员数量
- Income\_2019(~2022) 2019-2022 年的收入
- profit\_2019(~2022) 2019-2022 年的净利润

新增字段

- Income 方面（均通过计算获得）
  - incomeYOY%\_2020(~2022) 收入年增长率
  - incomeYOY%\_even #20-22 年 YOY 收入年增长率 20-22 年均值
  - incomeYOY%\_var
- Profit 方面
  - profitYOY%\_2020(~2022) 净利润年增长率
  - profitYOY%\_even 净利润年增长率 20-22 年均值
  - profitYOY%\_var 平均净利润年增长率 20-22 年标准差
- 其他
  - area\_hubei 是否在湖北注册
  - transIndex 数字化转型紧迫度（参考埃森哲的相关研究<sup>2</sup>）
  - status 发展状态 =19-22 年收入 YOY 均值是否为正 (0/1 变量, YOY 为正则为 1) × 净利润 YOY 均值是否为正 (0/1 变量, YOY 为正则为 1)

最终获得的数据集如下图所示，数据形状为  $3331 \times 26$ 。

<sup>2</sup> 来源：[China Digital Transformation Index 2023 | Accenture](#)，将在第三点中具体说明

	ts_code	symbol	name	area	industry	industry_index	list_date	employee <sup>*</sup>	income_2019	income_2020	incomeYOY%, 2020	income_2021	incomeYOY%, 2021	income_2022	incomeYOY%, 2022	incomeYOY%, even	incomeYOY%, var	profitYOY%, 2020	profitYOY%, 2021	profitYOY%, 2022	profitYOY%, even	profitYOY%, var			
0	000001.SZ	SZSE0001	深证成指	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	11.98	2.25	2.40	29.11	(25.80)	(17.31)	29.11	29.11	19				
1	000002.SZ	SZSE0002	深证综指	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	18.92	2.05	2.40	29.11	(25.80)	(17.31)	29.11	29.11	22.95				
2	000004.SZ	SZSE0004	深证100	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	105.73	1820.04	(320.41)	(17.18)	294.15	196.50	294.15	294.15	22.95				
3	000005.SZ	SZSE0005	深证300	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	152.29	152.29	(248.49)	(17.11)	231.00	231.00	231.00	231.00	22.95				
4	000006.SZ	SZSE0006	深证500	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	15.20	15.20	(19.65)	(19.65)	14.00	14.00	14.00	14.00	22.95				
5	000007.SZ	SZSE0007	深证A股	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	1.23	20.66	7.10	(38.70)	(19.65)	22.97	22.97	22.97	22.97	22.95			
6	000008.SZ	SZSE0008	深证B股	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	21.82	108.77	161.55	132.79	(110.66)	(212.15)	401.92	401.92	22.95				
7	000009.SZ	SZSE0009	深证创业板	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	12.00	12.00	(11.67)	(12.00)	30.15	30.15	172.00	172.00	22.95				
8	000010.SZ	SZSE0010	深证中小板	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	22.59	49.55	49.55	64.74	40.35	51.68	12.31	12.31	22.95				
9	000011.SZ	SZSE0011	深证沪A股	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	24.71	45.00	45.00	51.00	(44.67)	(44.67)	401.92	401.92	22.95				
10	000012.SZ	SZSE0012	深证沪B股	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	11.16	11.16	15.14	34.93	34.93	(47.40)	(47.40)	41.30	41.30	22.95			
11	000013.SZ	SZSE0013	深证医药	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	10.29	10.29	27.24	53.00	53.00	40.35	40.35	22.95	22.95	22.95			
12	000014.SZ	SZSE0014	深证金融	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	55.12	16.94	19.95	61.25	49.37	(310.41)	216.44	216.44	216.44	216.44			
13	000015.SZ	SZSE0015	深证公用事业	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	42.23	48.15	88.13	70.50	155.00	(200.00)	(102.70)	217.73	217.73	22.95			
14	000016.SZ	SZSE0016	深证地产	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	14.00	(18.02)	13.04	5.01	8.06	(8.11)	2.98	6.54	6.54	6.54	22.95		
15	000017.SZ	SZSE0017	深证银行	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	4.13	(22.22)	12.06	25.09	4.44	40.65	23.72	23.72	23.72	23.72	22.95		
16	000018.SZ	SZSE0018	深证保险	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	1.81	1.81	1.81	1.81	1.81	1.81	1.81	1.81	1.81	22.95			
17	000019.SZ	SZSE0019	深证证券	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	(71.47)	(31.69)	38.59	(70.61)	(98.62)	(421.51)	490.69	490.69	490.69	490.69	22.95		
18	000020.SZ	SZSE0020	深证能源	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	18.62	45.00	(34.04)	(34.04)	(34.04)	(34.04)	27.71	27.71	27.71	27.71	22.95		
19	000021.SZ	SZSE0021	深证医药	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	14.56	23.57	(54.57)	7.05	21.29	36.23	31.87	(31.24)	12.28	37.76	37.76	37.76	22.95
20	000022.SZ	SZSE0022	深证交通	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	45.00	54.54	16.17	22.82	28.64	135.30	(50.57)	4.29	29.01	34.39	34.39	22.95	
21	000023.SZ	SZSE0023	深证公用事业	深圳	综合类	1	2010-01-01	251.00	567.99	15.32	5.04	17.27	13.00	14.00	7.64	11.68	3.75	11.25	11.25	7.01	7.01	15.94	15.94	22.95	

图 经过添加字段等处理后使用的数据集

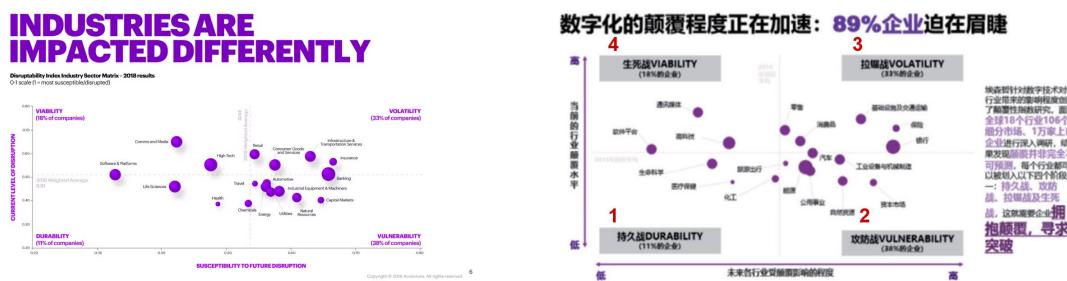
### 3. 特别说明：关于行业的划分

进行了很多关于行业分类的探索，但是无论是按申万一级行业分成 10+类，还是按照三大产业分，效果都不理想。

偶然看到全球最大的上市咨询公司埃森哲把行业分成四类：

- “生死战”数字化颠覆已经基本完成的行业，比如高科技、软件平台等行业，那些没有实现数字化转型的企业已经消失不见
- “攻防战”数字化颠覆即将发生的行业，比如公共事业、自然资源、资本市场等行业，它们需要为即将到来的数字化转型做好准备
- “拉锯战”数字化颠覆会持续发生的行业，比如零售、银行、保险、交通运输等行业，除了一些个性化服务还需要有人支持外，大量流程性或重复性的工作都会被数字技术所取代
- “持久战”数字化颠覆影响较低的行业，比如生命科学、医疗保健和化工等行业，数字化更多是前端提升客户体验，后端提升运营效率

参考这一划分标准，我们添加变量 `transIndex`，用于反映行业面临数字化转型的紧迫程度，数值越大越紧迫。



资料来源：埃森哲 Gauging Business Disruption with the Disruptability Index | Accenture

17

图 埃森哲从数字化角度将行业分成四类  
(左图为英文版，右图为对应中文版)

## 二、数据探索与可视化

### 2.1 上市公司分布特征探索

#### 2.1.1 不同行业上市公司数量的探索

通过 ggplot 绘制柱状图和玫瑰图<sup>3</sup>可以看出，截至 2018 年，上市公司分布最多的行业是医药生物、机械设备、基础化工。

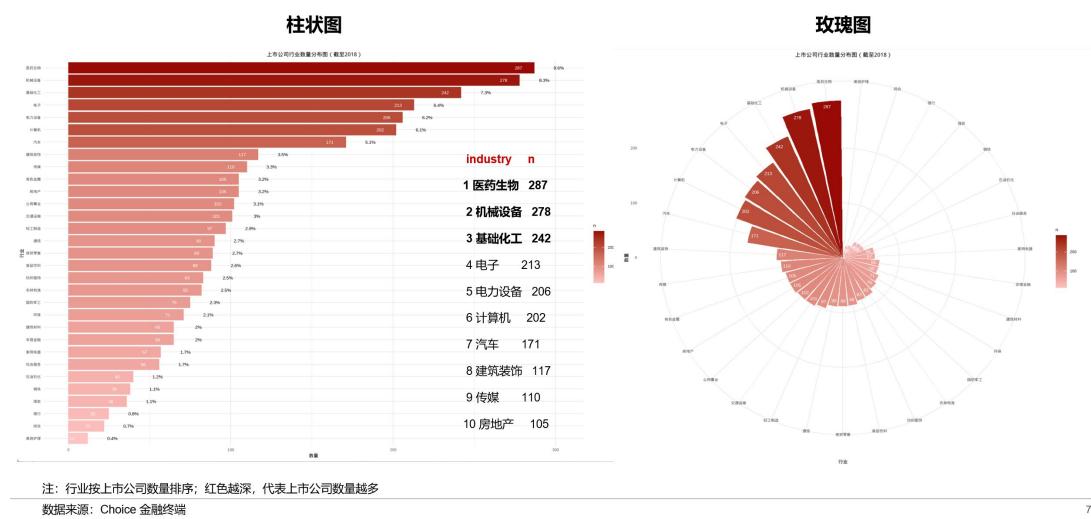


图 上市公司行业数量分布（截至 2018）

#### 2.1.2 不同行业雇员数探索

行业上市公司雇员总数：截至 2018 年上市公司雇员总数最多的行业是银行、建筑装饰、汽车、电子。

<sup>3</sup> 代码见附录 2.1

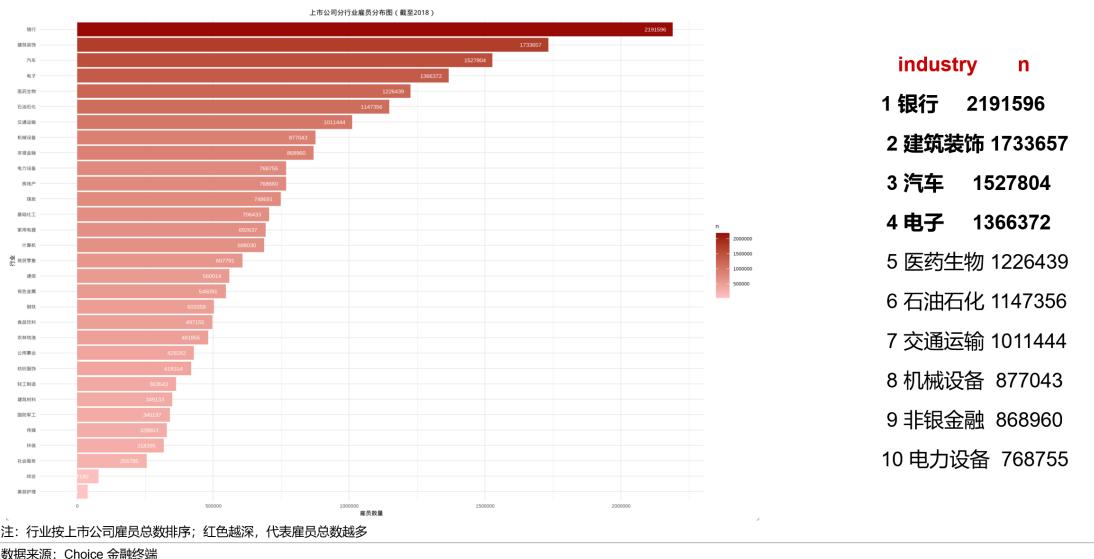


图 上市公司雇员总数分布（截至 2018）

单个公司雇员数：单上市公司雇员最多的行业是银行、煤炭、钢铁、石油石化。

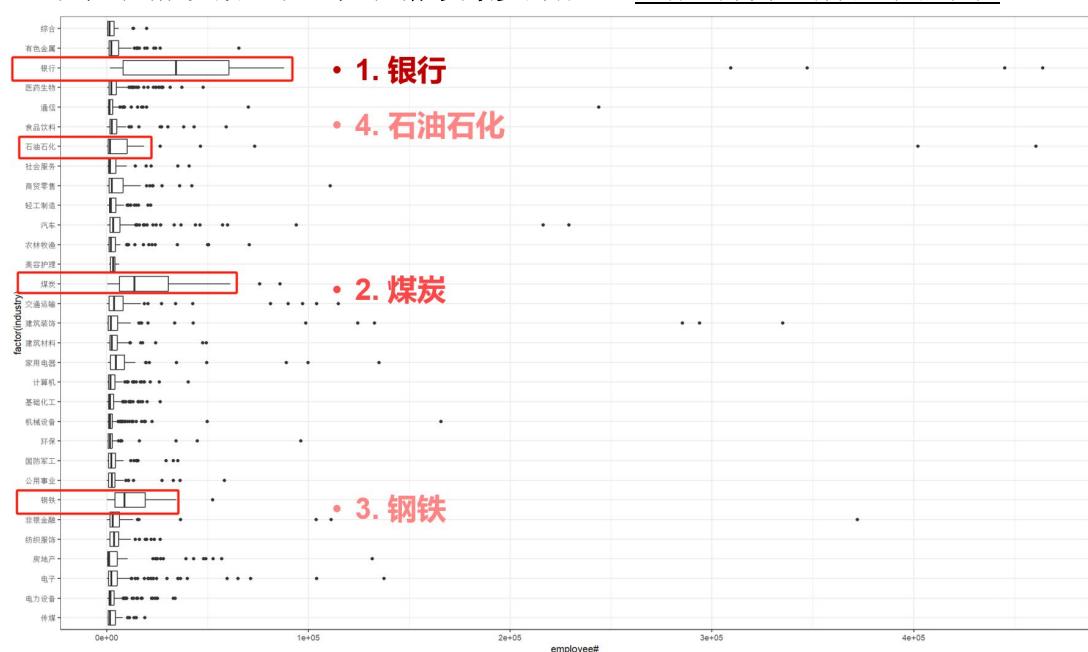


图 单个上市公司雇员数量分布（截至 2018）

### 2.1.3 不同行业上市公司营收探索

截至 2018 年，上市公司营收最多的是银行、煤炭、石油石化、钢铁行业

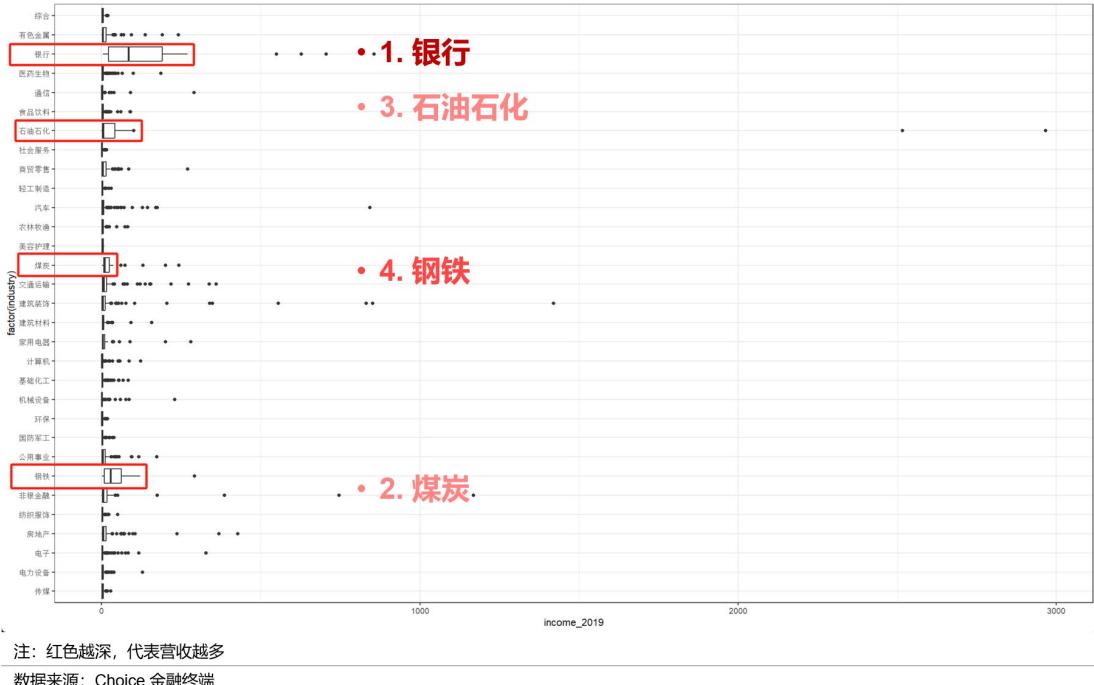


图 单个上市公司营收分布（截至 2018）

## 2.2 疫情对不同行业影响的探索

### 2.2.1 疫情期间所有上市公司营收与净利润增长率变化

疫情期间（2019-2022 年）上市公司营收总体仍呈增长态势，但净利润基本没有增长，亏损的企业更多。

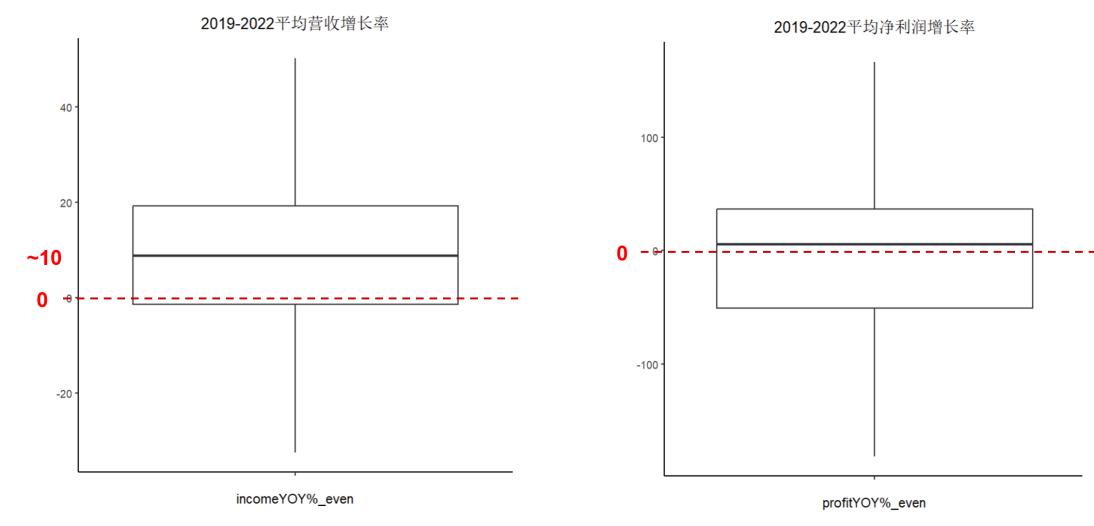


图 2019-2022 年平均营收与净利润增长率箱线图

### 2.2.2 湖北经济受疫情影响探究

(1) 作为疫情爆发地，湖北的经济在 2020 年增长弱于全国其他地区

- 20年湖北GDP和上市公司平均营收YOY均低于全国（尤其是GDP）
- 21年或因20年基数小/短暂解封后复苏而有所反弹，高于全国水平
- 22年和20-22年平均，湖北GDP和上市公司平均营收YOY均与全国持平。考虑到3年平均跨度较长以及湖北经济发展情况，应当是正常水平

## (2) 只看上市公司

- 营收方面，2019、20年湖北上市公司中位数不及其他地区，但21年和22年的营收超越其他地区，可能是因为19、20年受疫情影响比较大
- 净利润方面，除了19年湖北上市公司净利润中位数远小于其他地区外，其余年份均大于其他地区，其受疫情影响不明显

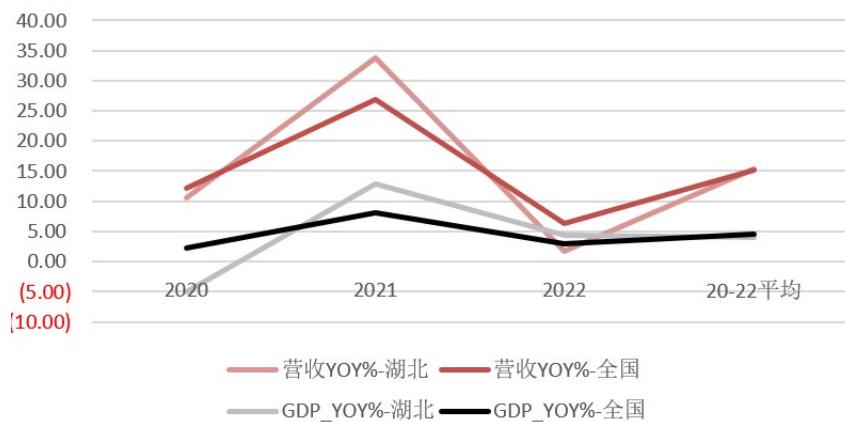


图 2020-2022 年湖北 vs 全国上市公司营收与 GDP YOY 对比折线图

	area_hubai	median('incomeYOY%_2019')	median('incomeYOY%_2020')	median('incomeYOY%_2021')	median('incomeYOY%_2022')
1	0	6.783215	3.382819	16.93507	2.373352
2	1	3.207125	1.316620	19.92516	4.433679

	area_hubai	median('profitYOY%_2019')	median('profitYOY%_2020')	median('profitYOY%_2021')	median('profitYOY%_2022')
1	0	9.5587934	8.569837	14.22280	-4.754754
2	1	0.5084914	9.237971	19.49196	1.248639

图 2020-2022 年湖北 vs 其他地区上市公司营收与净利润 YOY 中位数对比

在比较上市公司营收与净利润 YOY 时，最开始采用的是平均数，但是由于发现湖北地区的平均 income\_YOY 异常地高，通过 summary() 观察统计量特征后发现因为极高异常值的存在可能很大程度上拉高了平均值，所以改用中位数衡量；并且考虑到异常值的影响，后续只选择 20-22 年间平均 income 和 profit YOY<200（即 200%）的企业进行探究。

```

area_hubei `mean(``incomeYOY%_2020``)`  

<db> <db>  

1 0 7.09  

2 1 184.  

>  

> #`mean(``incomeYOY%_2020``)`异常高, 找原因  

> summary(datatab$'incomeYOY%_2020')  

Min. 1st Qu. Median Mean 3rd Qu. Max.  

-97.022 -11.105 3.245 12.091 16.564 16849.847  

> #Max值竟然达到16849.847, 说明异常值存在可能极大地影响了平均值, 因此应该用中位数衡量

```

图 极大异常值的存在可能导致平均值过大

### 2.2.3 不同行业收入变化情况的对比, 以石油石化行业为例深入探究

分行业来看:

- 环保、石油石化、机械设备等行业的收入增长比较大
- 石油石化、综合和房地产行业的收入波动比较大

石油石化行业: 从下图可以看出, 石油行业确实起伏比较大, 总体呈现增长趋势, 但 20 年下降比较多

- 房地产波动大是符合预期的

图: 2020-2022年各行业平均收入增长率的均值

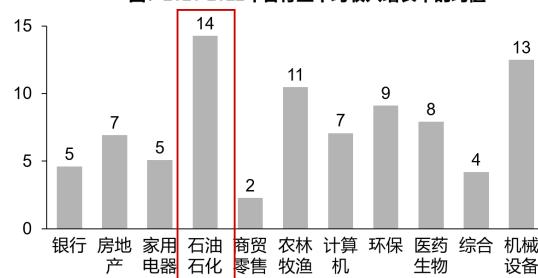
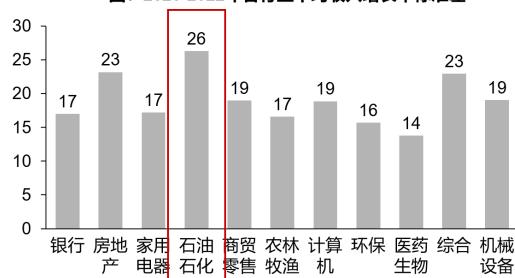


图: 2020-2022年各行业平均收入增长率标准差



资料来源: Choice, 石油和化学工业联合会, 国家统计局

图 2020-2022 年各行业平均收入增长率的均值和标准差

表 2020-2022 年石油石化行业收入 YOY 情况

Income YOY	2020	2021	2022	20-22平均
上市公司 (median)	0.56	33.9	9.66	14
整个行业 (average)	8.7	30	14.4	17.6

# 三、机器学习

## 3.1 PCA

留下 `transIndex`、`employee#`、`income_2019`、`netprofit%_2019` 四个变量，进行 PCA 降维和聚类

- `transIndex` 代表数字化紧迫度, `employee#` 和 `income` 代表公司规模, `profit` 代表盈利能力, 猜想通过这几个或许与公司在疫情期间的发展有关

异常值: `status=-1` 的, 横轴跨度最大;  $|x|$  很大的公司一般 `status` 为 1

从图 1 可以看出这几个变量与 PCA 新生成的变量都高度相关

- `Employee#` 和 `income_2019` 基本重叠, 也印证前面的猜想
- 确实大概可以分为三个影响方向: 公司规模、数字化转型程度和盈利能力

从图 2 可以看出 PCA2 维可以达到 70% 解释度, 三维达到 95% 的解释度

图 3 中聚类效果不好, 因为异常值太多

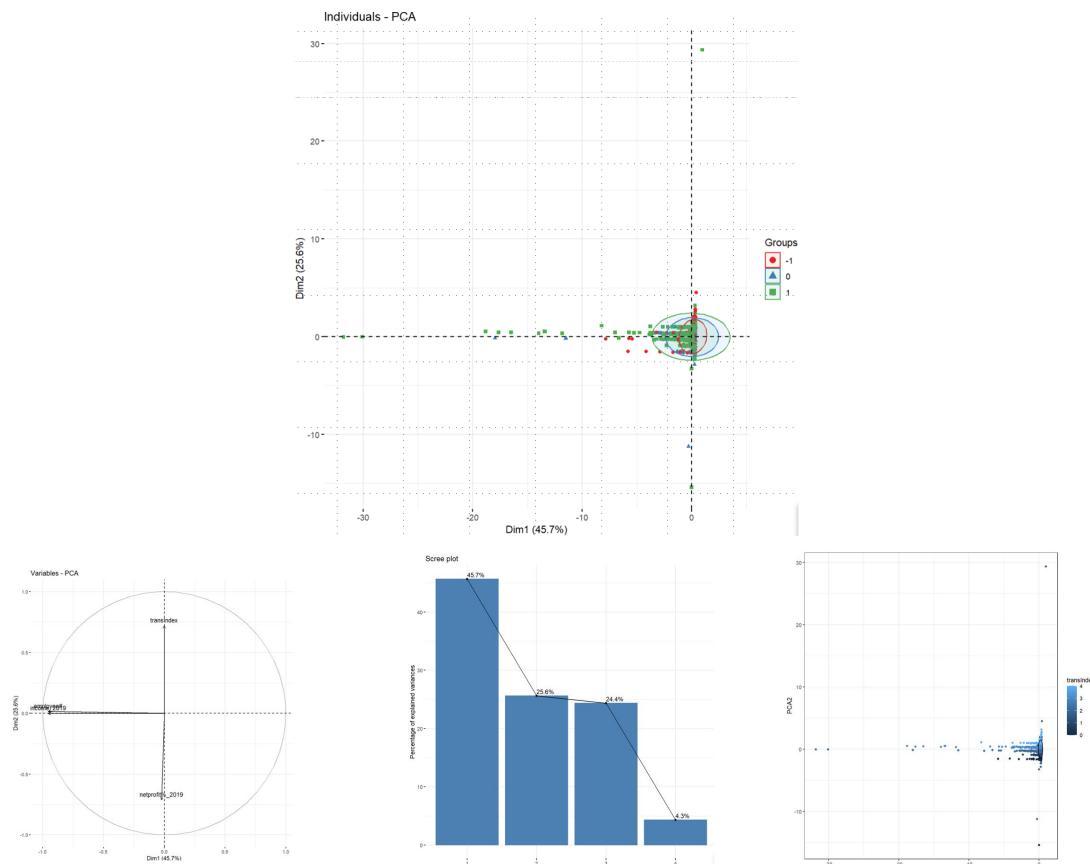


图1

图2

图3

图 PCA 结果

## 3.2 SOM

对照图 1、图 2，可以看出规模大的公司相对发展比较好

- a) 雇员数量、收入特别多的（左图中黄色和橘色扇形大）对应 status 为 1，说明规模大的公司发展比较好；
  - b) net profit% 和 tansindex 在这里影响不明显
- 根据图 3，可以看出基本全部聚成一类，聚类效果不好

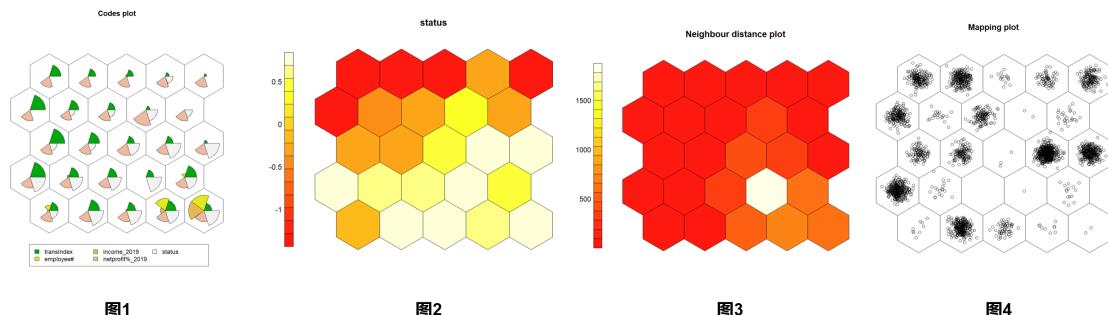


图1

图2

图3

图4

图 SOM 结果

## 3.3 K-means

根据 search 结果，K-means 聚类结果共分为 4 类最佳 (between\_SS / total\_SS = 53.3%，效果不是特别好)，利用 Excel 进行可视化处理后可以看出：

- **Status 为 1**，也就是状态最好的两个 Cluster 分别处于“持久战”和“攻防战”中，也就是数字化转型需求尚不十分紧迫
- **Status 为 0 和-1**，在疫情中并未实现明显的良性发展，对应处于“拉锯战”和“生死战”的数字化转型进程中。这可能是因为他们本就在紧迫地进行数字化转型，疫情的催化作用不明显。

```
K-means clustering with 4 clusters of sizes 1035, 11, 785, 882
cluster means:
employee_num income_2019 netPorfitRate2019 status transIndex
1 -0.063819836 -0.06484294 0.06502799 0.6075353 -0.78403537
2 13.061547448 12.26911855 0.09492419 0.6348553 0.15726711
3 -0.089787855 -0.05744783 -0.10348036 -1.4323070 -0.01571397
4 -0.008095266 -0.02579514 0.01460765 0.5539440 0.93206591
```

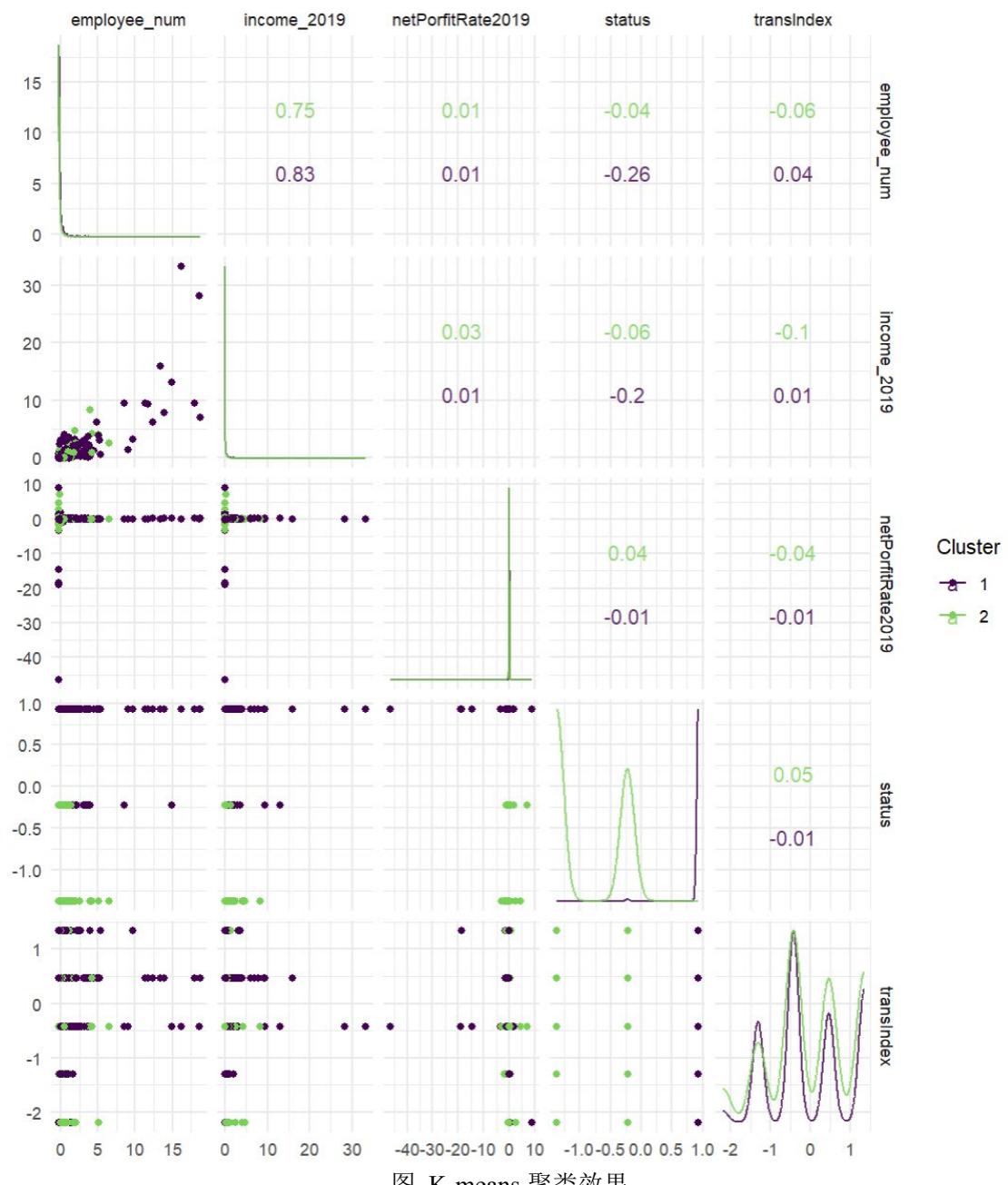


图 K-means 聚类效果

Cluster	employee_num	income_2019	netPorfitRate2019	transIndex	status	status	transIndex
1	-0.06	-0.06	0.07	-0.78	0.61	1	1
2	13.06	12.27	0.09	0.16	0.63	1	3
3	-0.09	-0.06	-0.10	-0.02	-1.43	1	2
4	-0.01	-0.03	0.01	0.93	0.55	0	4

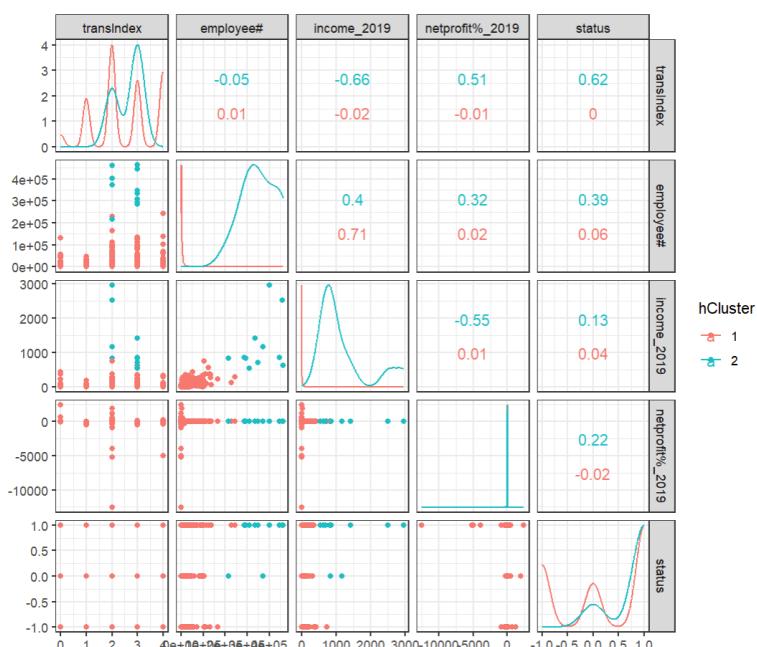
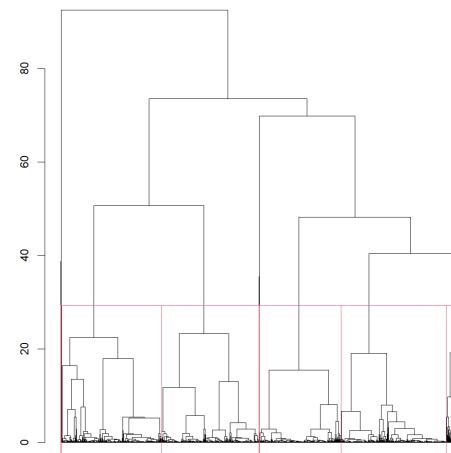
注：数据条长度代表数值长度（同一列中对比）；数值条颜色为绿色代表数值为正，红色代表数值为负

数据来源：Choice 金融终端；机器学习结果

18

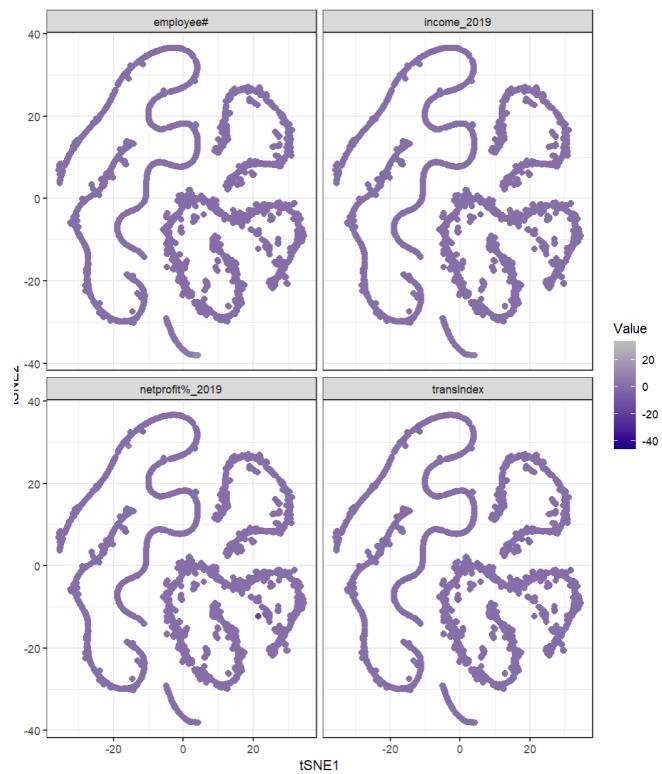
图 根据聚类结果，关注 status 和 transIndex 两个指标

### 3.4 Diana



### 3.5 SNE

```
# A tibble: 13,320 x 5
  status[,1] tsNE1 tsNE2 Feature      value[,1]
  <dbl>   <dbl>  <dbl> <chr>        <dbl>
1 0.929 -4.58 -30.0 transIndex 0.459
2 0.929 -4.58 -30.0 employee# 1.12 
3 0.929 -4.58 -30.0 income_2019 1.40 
4 0.929 -4.58 -30.0 netprofit%_2019 0.0873
5 -1.38  1.94 -37.9 transIndex -2.18 
6 -1.38  1.94 -37.9 employee# 5.10 
7 -1.38  1.94 -37.9 income_2019 3.99 
8 -1.38  1.94 -37.9 netprofit%_2019 0.0671
9 0.929 31.2 -12.9 transIndex 1.34 
10 0.929 31.2 -12.9 employee# -0.265
# i 13,310 more rows
# i Use `print(n = ...)` to see more rows
```



## 四、总结与反思

- 上市公司信息:
  - 上市公司分布最多的行业是医药生物、机械设备、基础化工，雇员总数最多的行业是银行、建筑装饰、汽车、电子
  - 上市公司营收和雇员数基本呈正相关，都是银行、钢铁、石油石化、煤炭这几个行业最多
- 疫情对上市公司的影响:

- 疫情几年营收总体仍呈增长态势，但净利润基本没有增长，亏损企业更多
- 2019、20 年湖北经济受疫情影响较大，上市企业营收也有所受挫
- 相比之下，规模大的公司在疫情期间发展更好
- 疫情可能更多推动原本数字化进程较为缓慢的行业更加迅速地步入数字化
- 问题和改进措施：
  - 异常值的处理（筛选 2019 年的 Income，去掉量级太小的数据，收入量级大的公司异常值；或者对数量级大的取对数）
  - 疫情对不同行业的影响暂时难以用机器学习探究
- 其他细节
  - 变量命名尽量不要出现 "%”、“#” 特殊符号，会带来麻烦

# 五、附录

## 4.1 前言

### 1.1.1 利用 Python 调用 Tushare API 接口获取公司数据

```
import tushare as ts
import openpyxl
import pandas as pd
#1. 筛选股票，找出地域、行业、上市时间
pro = ts.pro_api('ca0cd5409f62a54869ffa4ad1f81f99e0a3d649f4e87cf034dc0009')
df = pro.stock_basic(exchange='', list_status='L',
fields='ts_code,symbol,name,area,industry,list_date')
filtered_df = df[df['list_date'] <= '2018-12-31']
filtered_df.to_excel("filtered_stock_info.xlsx")
filtered_df =
pd.read_excel('C:/Users/10136/PycharmProjects/pythonProject/filtered_stock_info.xlsx')
#2. 从利润表中找出 19-22 年收入、利润
for i in filtered_df['ts_code']:
    filtered_df['2019_revenue']= pro.income(ts_code=i, period='20191231',
fields='total_revenue')['total_revenue'] #n_income 净利润(含少数股东损益)
```

## 4.2 数据探索与可视化

### 2.1.1 上市公司的行业分布

柱状图：

```
data %>%
group_by(industry) %>%
summarize(n=n()) %>%
mutate(pct = n / sum(n)) %>%
ggplot(aes(x = reorder(industry, n), y = n, fill = n)) +
geom_bar(stat = "identity") +
geom_text(aes(label = n, y = n-10), color="white") +
geom_text(aes(label = paste0(round(100 * pct, 1), "%"), y = n + 15)) +
scale_fill_gradient(low ="#FFC2C2", high = "#990000") +
coord_flip() +
theme_minimal()
xlab("行业") + ylab("数量") +
```

```
ggtitle("上市公司行业数量分布图（截至 2018）") +  
  theme(plot.title = element_text(hjust = 0.5))
```

### 玫瑰图：

```
data %>%  
  group_by(industry) %>%  
  summarize(n=n()) %>%  
  mutate(pct = n / sum(n)) %>%  
  ggplot(aes(x = reorder(industry, n), y = n, fill = n)) +  
  scale_fill_gradient(low ="#FFC2C2", high = "#990000") +  
  geom_bar(stat = "identity") +  
  geom_text(aes(label = n, y = n-10), color="white") +  
  coord_polar() +  
  theme_minimal() +  
  xlab("行业") + ylab("数量") +  
  ggtitle("上市公司行业数量分布图（截至 2018）") +  
  theme(plot.title = element_text(hjust = 0.5))
```

### 排名前十的公司：

```
data %>%  
  group_by(industry) %>%  
  summarize(n=sum(`employee#`)) %>%  
  arrange(desc(n)) %>%  
  head(10)
```

## 2.2.1 疫情期间所有上市公司营收与净利润增长率变化

```
p <- ggplot(df, aes(x=x, lower=lower, upper=upper, middle=middle, ymin=ymin,  
                     ymax=ymax)) +  
  geom_boxplot(stat="identity") +  
  labs(title="2019-2022 平均营收增长率", x="incomeYOY%_even") +  
  theme_classic() +  
  theme(plot.title=element_text(hjust=0.5))  
p
```

## 2.2.2 湖北经济受疫情影响探究

```
i2019 <- datatb %>% summarize(median(`incomeYOY%_2019`))  
i2020 <- datatb %>% summarize(median(`incomeYOY%_2020`))  
i2021 <- datatb %>% summarize(median(`incomeYOY%_2021`))  
i2022 <- datatb %>% summarize(median(`incomeYOY%_2022`))  
income_change <- i2019 %>% left_join(i2020) %>% left_join(i2021) %>% left_join(i2022)  
view(income_change)
```

```

p2019 <- datatb %>% summarize(median(`profitYOY%_2019`))
p2020 <- datatb %>% summarize(median(`profitYOY%_2020`))
p2021 <- datatb %>% summarize(median(`profitYOY%_2021`))
p2022 <- datatb %>% summarize(median(`profitYOY%_2022`))
profit_change <- p2019 %>% left_join(p2020) %>% left_join(p2021) %>% left_join(p2022)
view(profit_change)

```

## 4.3 机器学习

### PCA

```

pca1 <- datatb %>% dplyr::select(-`incomeYOY%_even`, -`incomeYOY%_var`, -`3index`, -industry, -`industry_index`, -`profitYOY%_2019`, -`11index`, -`incomeYOY%_2019`, -area_hubei, -name, -area, -ts_code, -list_date, -symbol, -`incomeYOY%_2020`, -`incomeYOY%_2021`, -`incomeYOY%_2022`, -`profitYOY%_2020`, -`profitYOY%_2021`, -`profitYOY%_2022`, -`profitYOY%_even`, -`profitYOY%_var`)
view(pca1) #注意这里不加 dyplr 的话会跟别的包里的 select 混起来，所以报错
datatb<-as.matrix(scale(datatb))
pca_data <- datatb %>% dplyr::select(-status, -`incomeYOY%_even`, -`incomeYOY%_var`, -`3index`, -industry, -`industry_index`, -`profitYOY%_2019`, -`11index`, -`incomeYOY%_2019`, -area_hubei, -name, -area, -ts_code, -list_date, -symbol, -`incomeYOY%_2020`, -`incomeYOY%_2021`, -`incomeYOY%_2022`, -`profitYOY%_2020`, -`profitYOY%_2021`, -`profitYOY%_2022`, -`profitYOY%_even`, -`profitYOY%_var`)
pca_data<-as.matrix(scale(pca_data))

pca <- pca_data %>%
  prcomp(center = T, scale = T) #作用相当于转置。center 平移， scale 拉伸
pca
summary(pca) #proportion of variance 方差占样本总方差的比例

pcaDat <- get_pca(pca)
fviz_pca_ind(pca, label = "none", habillage = datatb$status, addEllipses = T, ellipse.level = 0.9) +
  scale_color_brewer(palette = "Set1")

fviz_pca_biplot(pca, label = "var")
fviz_pca_var(pca)
fviz_screeplot(pca, addlabels = T, choice = "eigenvalue") #特征值
fviz_screeplot(pca, addlabels = T, choice = "variance") #方差

##ggplot
stockPCA <- datatb %>%

```

```

    mutate(PCA1 = pca$x[,1], PCA2 = pca$x[,2],PCA3 = pca$x[,3])
ggplot(stockPCA, aes(PCA1, PCA2, col = transIndex)) +
  geom_point() +
  stat_ellipse(level = 0.9)
ggplot(stockPCA, aes(PCA1, PCA3, col = transIndex)) +
  geom_point() +
  stat_ellipse(level = 0.9)

```

## K-means

```
pacman:::p_load(mlr3,mlr3verse,mlr3cluster,mlr3viz,GGally,mlr3tuning,mlr3learners)
```

gvhdCtrlScale <- as.data.table(scale(pca1))#标准化操作将数据转换为均值为 0、标准差为 1 的形式，有助于消除不同变量之间的量纲差异。

ggscatmat(gvhdCtrlScale)#使用 GGally 包中的 ggscatmat 函数绘制了标准化后的散点矩阵图。

ggpairs(gvhdCtrlScale,

```

  upper = list(continuous = "density"),
  lower = list(continuous = wrap("points", size = 0.4)),
  diag = list(continuous = "densityDiag"))
#使用 GGally 包中的 ggpairs 函数绘制了成对图，更详细地显示了变量之间的关系。
#具体地，上半部分是核密度图，对角线上是每个变量的密度图，下半部分是散点图。
#这有助于更深入地了解变量之间的相关性和分布
```

#选择 task 和 learner 设置参数

```

gvhdCtrlScale <- gvhdCtrlScale %>% rename(employee_num='employee#')
gvhdCtrlScale <- gvhdCtrlScale %>% rename(netPorfitRate2019='netprofit%_2019')
taskC = TaskClust$new("gvhdCtrlScale", gvhdCtrlScale) #如果是监督学习还要加上目标，非监督就没有
class(taskC)
autoplot(taskC)
taskPos = TaskClust$new("gvhdCtrlPos", as.data.table(scale(GvHD::pos)))

```

```

mlr_learners$keys("clust")
kmeans_lrn = lrn("clust.kmeans")
class(kmeans_lrn)
kmeans_lrn$param_set$params #超参数集合
kmeans_lrn$param_set$values = list(centers = 3, iter.max = 100, nstart = 10)

```

```

#训练模型及预测
kmeans_model <- kmeans_lrn$train(taskC)
print(kmeans_model$model)
kmeans_lrn$stategyhdCtrlScale

```

```

kmeans_lrn$assignments

kmeans_lrn1 = kmeans_lrn$clone()
kmeans_lrn1$predict(taskC)
autoplot(kmeans_lrn1$predict(taskC), taskC, type = "scatter" )

kmeansPos = kmeans_lrn$predict(taskPos)

autoplot(kmeansPos, taskPos, type = "pca", frame = T) # 还有 type="sil"是剪影图
 autoplot(kmeansPos, taskPos, type = "sil", frame = T)

```

## SOM

```

pacman::p_load(tidyverse, mclust, Rtsne, umap, kohonen, lle, GGally, plot3D, plot3Drgl)
theme_set(theme_bw())

```

```

somGrid <- somgrid(xdim = 5, ydim = 5, topo = "hexagonal",
                     neighbourhood.fct = "bubble", toroidal = F)
#表示按六边形将神经网络单元排成 5 列、5 行，共计 25 个单元。

```

#设定 topo=rectangular,toroidal=T， 重新运行 SOM 比较。

```

flea2 <- pca1%>%
  scale()

```

```

fleaSom <- som(flea2, grid = somGrid, rlen = 50000, alpha = c(0.05, 0.01))
fleaSom1 <- som(flea2, grid = somGrid, rlen = 100000, alpha = c(0.1, 0.001))
#rlen 迭代次数， alpha 学习率
par(mfrow = c(2,3))
plotType <- c("codes", "changes", "counts", "quality", "dist.neighbours", "mapping")
walk(plotType, ~plot(fleaSom, type = ., shape = "straight"))

```

```

walk(plotType, ~plot(fleaSom1, type = ., shape = "straight"))

```

# 可以让这些图分别显示

```

## plot(fleaSom,type="codes") 里面半径大小说明权重大小，但是并不知道是半径越大还是越小

```

```

## plot(fleaSom,type="changes")

```

## trainning 表示训练过程 看迭代后是否达到稳定

## count 和 mapping 其实表达的是一个意思 都是看数量

## quality 看点和碗有多匹配

# neighbor 领域距离图 计算每个碗中的点和周围的碗中的点的距离的平均值 越红距离越小

```

getCodes(fleaSom) %>%

```

```

as_tibble() %>%
  iwalk(~plot(fleaSom, type = "property", property = .,
             main = .y, shape = "straight"))

getCodes(fleaSom1) %>%
  as_tibble() %>%
  iwalk(~plot(fleaSom1, type = "property", property = .,
              main = .y, shape = "straight"))

```

### Diana

```

gvhdDist <- dist(gvhdCtrlScale, method = "euclidean")
gvhdHclust <- hclust(gvhdDist, method = "ward.D2")

plot(as.dendrogram(gvhdHclust), leaflab = "none")
gvhdCut <- cutree(gvhdHclust, k = 2)
plot(as.dendrogram(gvhdHclust), leaflab = "none")
rect.hclust(gvhdHclust, k = 9)

gvhdTib <- dplyr::mutate(pca1, hCluster = as.factor(gvhdCut))
ggscatmat(gvhdTib, color = "hCluster")

```

### SNE

```

pacman::p_load(tidyverse, mclust, Rtsne, umap, kohonen, lle, GGally, plot3D, plot3Drgl)
noteTsne <- pca1 %>%
  dplyr::select(-status)
noteTsne <- Rtsne(noteTsne, perplexity = 30, theta = 0.3, max_iter = 500, verbose = T)
noteTsne #Y 是降维后的数据, origD 是 original dimension
noteTsne1 <- pca1 %>%
  mutate_if(.funs = scale, .predicate = is.numeric) %>%
  mutate(tSNE1 = noteTsne$Y[,1], tSNE2 = noteTsne$Y[,2]) %>% # 增加 tsne 中的新的坐标
  pivot_longer(names_to = "Feature", values_to = "Value", c(-tSNE1, -tSNE2, -status))
noteTsne1
#按照 facet_wrap 进行分类——解释原始的特征跟新的维度之间的关系
ggplot(noteTsne1, aes(tSNE1, tSNE2, col = Value)) +
  facet_wrap(~ Feature) +
  geom_point(size = 2) +
  scale_color_gradient(low = "darkblue", high = "grey")

```