上 海 财 经 大 学

Shanghai University of Finance and Economics

数据科学导论期末项目报告

基于东方财富网 A 股市场 2023 年一季报业绩数据进行市场研究

项目 GitHub 链接：https://github.com/X-backk/wangran.git

院（系）：	信息管理与工程学院
班 级：	大数据 2 班
学 号：	2021110980
姓 名：	王 冉
日 期：	2023 年 5 月 27 日

摘　要

上市公司的业绩报告作为一种重要的信息披露方式，不仅是投资者了解公司经营情况和投资风险的重要依据，也是监管机构对公司经营状况进行监督和管理的重要手段。同时，上市公司的业绩报告还可以为市场参与者提供有价值的信息，帮助他们做出更明智的投资决策，同时也可以为政策制定者提供有关经济发展和政策调整的参考。上市公司的业绩报告是上市公司向公众披露自身经营状况的重要途径之一，通过业绩报告，公司可以向投资者、监管机构、媒体和其他市场参与者披露其财务状况、业务运营情况、风险管理情况和未来发展计划等重要信息。这些信息对于投资者来说非常重要，可以帮助他们评估公司的价值、风险和前景，从而做出更明智的投资决策。对于监管机构来说，业绩报告可以帮助他们监督上市公司的经营状况和风险管理情况，保护投资者的利益。对于媒体和其他市场参与者来说，业绩报告可以为他们提供有价值的信息，帮助他们了解市场趋势和公司发展动态，从而做出更加准确的预测和分析。研究上市公司的业绩报告不仅有助于投资者做出更明智的投资决策，也可以为政策制定者提供重要参考。通过分析上市公司的业绩报告，政策制定者可以获得有关经济发展和产业结构变化的信息，从而更好地制定宏观经济政策和产业政策。

本项目旨在使用 Python 逐年爬取东方财富网上市公司的业绩报告数据，并进行数据存储、数据分析与可视化展示。该研究意义在于通过对业绩报告数据进行爬取、存储、分析和可视化展示，从利用最新一年的数据展示根据不同行业进行数据分析，探寻行业之间的差异与逻辑、对单个行业内的上市公司进行分析，探寻行业内的数据奥秘、利用时间序列数据，探寻市场的现状与未来走势、利用时间序列数据，探寻行业的发展现状与未来发展预测等方面进行分析与预测，以了解不同行业后疫情时代的经济复苏情况，揭示出哪些行业受到了严重影响，哪些行业表现出了强劲的复苏态势，从而指导投资决策和政策制定；此外，通过分析时间序列数据，判断当前市场的发展现状，并使用线性回归模型、随机森林等预测模型对之后市场的发展做出预测。通过以上一系列数据处理分析过程，提高对 A 股市场上市公司业绩报表的了解与市场发展现状的认知。

关键词：时间序列，行业发展，个股分析

目 录

一、绪论

第一节、项目背景

随着全球疫情的爆发和蔓延，全球经济和金融市场都受到了巨大冲击。中国作为世界第二大经济体，也受到了疫情的影响。为了应对这一挑战，中国政府采取了一系列积极的财政和货币政策，包括减税降费、加大基础设施投资、放宽货币政策等措施，以支持企业和促进经济发展。

在这个背景下，A 股市场作为中国股票市场的重要组成部分，也承受着巨大的压力和挑战。为了应对这种情况，中国证监会和上交所等机构采取了一系列措施，以稳定市场和促进市场发展。其中包括：加快上市公司的审核和发行速度，推出一系列支持政策，如减免手续费、推出降低门槛的新三板等，以及对上市公司的信息披露、财务监管和市场监管等方面的加强。在这种背景下，本项目旨在基于后疫情时代的 A 股市场所有行业的上市公司的业绩报告数据，进行数据爬取、数据存储与数据分析。通过对这些数据的分析，我们可以更好地了解上市公司的财务状况和经营状况，为投资者提供更加明智的投资决策。同时，我们也可以为市场分析师提供更加深入的行业分析和市场洞察，帮助他们更好地了解市场的整体状况和未来趋势，为市场参与者提供更加精准的市场预测和风险提示。

总之，本项目旨在通过对后疫情时代 A 股市场所有行业的上市公司的业绩报告数据的分析，为投资者和市场分析师提供更加全面、准确的市场信息和决策参考，同时也为政策制定者提供数据支持和市场反馈，以促进市场健康发展和经济稳定增长。在这个过程中，我们将充分利用现代数据爬取和数据分析技术，探索数据背后的规律和趋势，为市场参与者提供更加精准的市场预测和决策支持，为数据分析和应用技术的研究和发展做出贡献。

第二节、项目内容

本项目基于东方财富网数据中心平台展开研究，使用网络爬虫技术从数据中心的年报季报板块获取项目分析所需要的业绩指标数据，记为业绩报告数据集，接着利用各类数据分析技术对行业、个股进行多维度分析，总结分析当前 A 股市场的发展现状。

第三节、项目组织框架

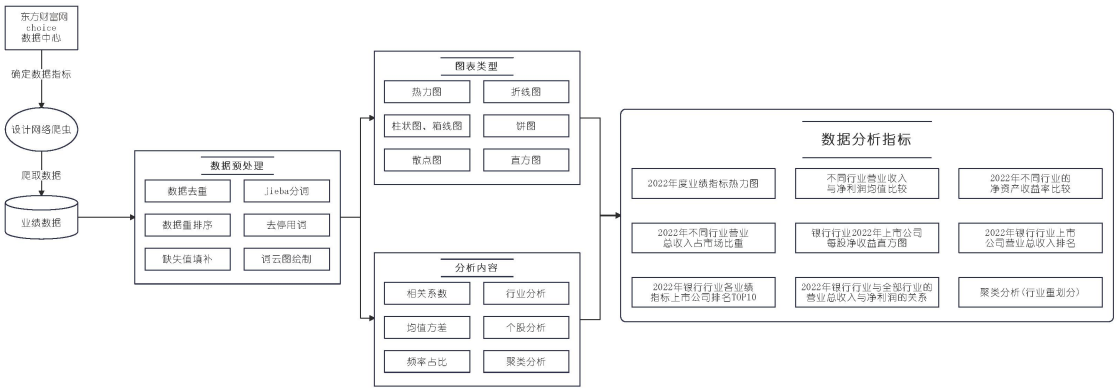


图 1.1 项目组织框架图

本报告内容将按如下顺序进行安排：

第一章：绪论，介绍本项目的项目背景以及研究内容；

第二章：数据采集部分，详述本项目所使用数据集的采集过程与部分爬虫代码示例；

第三章：对本项目使用的数据集进行介绍，描述数据的预处理过程；

第四章：详细介绍本项目的实验过程，并对实验结果进行解释；

第五章：总结项目。

二、数据采集与存储

第一节、数据采集

首先进行数据爬取，数据爬取是数据分析的第一步，它的作用是获取所需的原始数据。在这一章节中，我们将介绍如何使用 Python 爬虫工具从东方财富网上获取各类财务报表数据。数据爬取可以获取大量的实时和历史数据，为后续的数据分析和建模提供充足的数据基础。

图 2.1 东方财富网数据中心网页界面

通过网站(网站地址：https://data.eastmoney.com/bbsj/yjyg.html)。我们可以看到，在东方财富网的数据中心平台上包含了国内股票市场的业绩报表、业绩快报、资产负债表、利润表、现金流量表等。每个表中又包含了每个报表的各个指标。我们就将对各个报表的数据进行提取。核心代码如下：

首先导入必要的包，其次建立 DataScraper 类，作为数据抓取的一个类，其中包含 init 自定义函数，定义了需要提取的报表及指标，以及定义 url 作为数据爬取的网址，以及东方财富网的其他相关指标。

```
import csv
import json
import requests
from lxml import etree
class DataScraper:
    def __init__(self):
        self.pagename_type = {
            "业绩报表": "RPT_LICO_FN_CPD",
            "业绩快报": "RPT_FCI_PERFORMANCEE",
            "业绩预告": "RPT_PUBLIC_OP_NEWPREDICT",
            "预披露时间": "RPT_PUBLIC_BS_APPOIN",
            "资产负债表": "RPT_DMSK_FN_BALANCE",
            "利润表": "RPT_DMSK_FN_INCOME",
            "现金流量表": "RPT_DMSK_FN_CASHFLOW"
        }
        self.pagename_en = {
            "业绩报表": "yjbb",
            "业绩快报": "yjkb",
            "业绩预告": "yjyg",
```

```
            "预约披露时间": "yysj",
            "资产负债表": "zcfz",
            "利润表": "lrb",
            "现金流量表": "xjll"
        }
        self.en_list = []
        self.url = 'https://datacenter-web.eastmoney.com/api/data/v1/get'
        self.headers = {
            'Accept': '*/*',
            'Accept-Language': 'zh-CN,zh;q=0.9',
            'Connection': 'closed',
            'Referer': 'https://data.eastmoney.com/',
            'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/111.0.0.0 Safari/537.36',
            'sec-ch-ua': '"Google Chrome";v="111", "Not(A:Brand";v="8", "Chromium";v="111"',
            'sec-ch-ua-mobile': '?0',
            'sec-ch-ua-platform': '"Windows"'
        }
```

定义 get_table 函数、get_header 函数、write_header 函数、write_table 函数作为读取表与列名、写入表与列名的函数。

```
    def get_table(self, page):
        params = {
            'sortTypes': '-1,-1',
            'reportName': self.table_type,
            'columns': 'ALL',
            'filter': f'(REPORT_DATE=\'{self.timePoint}\')'
        }
        if self.table_type in ['RPT_LICO_FN_CPD']:
            params['filter'] = f'(REPORTDATE=\'{self.timePoint}\')'
        params['pageNumber'] = str(page)
        response = requests.get(url=self.url, params=params, headers=self.headers)
        data = json.loads(response.text)
        if data['result']:
            return data['result']['data']
        else:
            return
```

```
    def get_header(self, all_en_list):
        ch_list = []
        url = f'https://data.eastmoney.com/bbsj/{self.pagename_en[self.pagename]}.html'
        response = requests.get(url)
        res = etree.HTML(response.text)
        for en in all_en_list:
            ch = ''.join(
                [i.strip() for i in
res.xpath(f'//div[@class="dataview"]//table[1]//th[@data-field="{en}"]//text()')])
            if ch:
                ch_list.append(ch)
                self.en_list.append(en)
        return ch_list
    def write_header(self, table_data):
        with open(self.filename, 'w', encoding='utf-8', newline='') as f:
            writer = csv.writer(f)
            headers = self.get_header(list(table_data[0].keys()))
            writer.writerow(headers)
    def write_table(self, table_data):
        with open(self.filename, 'a', encoding='utf-8', newline='') as csvfile:
            writer = csv.writer(csvfile)
            for item in table_data:
                row = []
                for key in item.keys():
                    if key in self.en_list:
                        row.append(str(item[key]))
                print(row)
                writer.writerow(row)
```

Get_timelist 函数用于获取时间辍数据，在这里是每个公司报表披露公告时间

```
    def get_timeList(self):
        headers = {
            'Referer': 'https://data.eastmoney.com/bbsj/202206.html',
        }
        response = requests.get('https://data.eastmoney.com/bbsj/202206.html', headers=headers)
        res = etree.HTML(response.text)
        return res.xpath('//*[@id="filter_date"]//option/text()')
```

Run 函数用于运行时输入想要获取哪年哪季度的数据，以及获取什么样的报表类型的数据。

```
    def run(self):
        self.timeList = self.get_timeList()
        for index, value in enumerate(self.timeList):
            if (index + 1) % 5 == 0:
                print(value)
            else:
                print(value, end=' ; ')
        self.timePoint = str(input('\n 请选择时间（可选项如上）:'))
        self.pagename = str(
            input('请输入报表类型（业绩报表;业绩快报;业绩预告;预约披露时间;资产负债表;利润表;现金流量表）:'))
        assert self.timePoint in self.timeList, '时间输入错误'
        assert self.pagename in list(self.pagename_type.keys()), '报表类型输入错误'
        self.table_type = self.pagename_type[self.pagename]
        self.filename = f'{self.pagename}_{self.timePoint}.csv'
        self.write_header(self.get_table(1))
        page = 1
        while True:
            table = self.get_table(page)
            if table:
                self.write_table(table)
            else:
                break
            page += 1
if __name__ == '__main__':
    scraper = DataScraper()
    scraper.run()
```

本次爬取的数据是 2015-2023 年的业绩报表，数据集中包含了所有上市公司(5484 家)制定年份区间年报中有关业绩的指标，主要包括：每股收益(元),营业总收入(元),净利润(元), 净资产收益率(%),同比增长(%),同比增长(%).1,每股净资产(元),每股经营现金流量(元),销售毛利率(%),季度环比增长(%),季度环比增长(%).1,所处行业。

第二节、数据存储

这一节我们将把获取到的数据存储到 MySQL 数据库中，存储环节的代码如下，我们首先导入必须的 pymysql 包，然后创建一个表以及列名要与 csv 文件的列名一一对应，设置数据类型，之后读取 csv 文件并存入 mysql，最后形成 sql 文件。

```
import pandas as pd
import pymysql
# 创建表
sql_create_table = '''
    CREATE TABLE stockDataBase (
        stock_code VARCHAR(10),
        stock_name VARCHAR(50),
        announce_date DATE,
        earnings_per_share FLOAT,
        total_revenue BIGINT,
        net_profit FLOAT,
        roe FLOAT,
        yoy_growth FLOAT,
        yoy_profit FLOAT,
        net_assets_per_share FLOAT,
        operating_cash_flow_per_share FLOAT,
        gross_profit_margin FLOAT,
        qoq_growth1 FLOAT,
        qoq_growth2 FLOAT,
        profit_distribution TEXT,
        industry VARCHAR(50)
    )
'''
cursor.execute(sql_create_table)
# 将数据存入 MySQL 数据库中
for i in range(len(df)):
    values = tuple(df.iloc[i].values)
    sql = 'INSERT INTO stockDataBase VALUES ' + str(values)
    cursor.execute(sql)
```

通过上述代码即可将 csv 文件存储到 MySQL 数据库中，部分存储结果如下，以及保存的 db 文件见附件：stockDataBase.sql。

```
# 输出查询结果
for row in rows:
    print(row)
```

Python

```
('1', '平安银行', datetime.date(2023, 4, 25), 0.65, 45098000000, 14602000000.0, 3.39,
('2', '万科A', datetime.date(2023, 4, 29), 0.1251, 68474007399, 1445810000.0, 0.6, 9
('4', 'ST国华', datetime.date(2023, 4, 29), -0.0559, 21255771, -7427190.0, -2.13, 88
('5', 'ST星源', datetime.date(2023, 4, 29), -0.0026, 28171813, -2802010.0, -0.23, -3
('6', '深振业A', datetime.date(2023, 4, 29), -0.009, 265640113, -12085900.0, -0.15,
('7', '*ST全新', datetime.date(2023, 4, 29), -0.0007, 49531240, -232199.0, -0.26, -1
('8', '神州高铁', datetime.date(2023, 4, 28), -0.0121, 431343231, -32949000.0, -1.35,
('9', '中国宝安', datetime.date(2023, 4, 29), 0.0928, 8232161260, 239346000.0, 2.58,
('10', '美丽生态', datetime.date(2023, 4, 28), -0.0259, 123962807, -27469400.0, -4.05
('11', '深物业A', datetime.date(2023, 4, 28), 0.0222, 411469619, 13216800.0, 0.3, -67
('12', '南玻A', datetime.date(2023, 4, 26), 0.13, 4070673784, 396406000.0, 3.04, 46.
('14', '沙河股份', datetime.date(2023, 4, 28), 0.67, 389370140, 161574000.0, 12.95, 1
('16', '深康佳A', datetime.date(2023, 4, 29), 0.0633, 4600647852, 152514000.0, 1.98,
('17', '深中华A', datetime.date(2023, 4, 28), 0.0041, 151527918, 2848660.0, 0.98, 201
('19', '深粮控股', datetime.date(2023, 4, 26), 0.0915, 1338942952, 105445000.0, 2.19,
('20', '深华发A', datetime.date(2023, 4, 25), 0.0227, 175868642, 6432410.0, 1.8, 2.37
('21', '深科技', datetime.date(2023, 4, 29), 0.0647, 3933775539, 101013000.0, 1.02, 7
('23', 'ST深天', datetime.date(2023, 4, 29), -0.1344, 42802915, -18652900.0, -14.27,
('25', '特力A', datetime.date(2023, 4, 27), 0.0586, 339838493, 25274100.0, 1.66, 142
('26', '飞亚达', datetime.date(2023, 4, 25), 0.2505, 1200095569, 103189000.0, 3.23, 2
('27', '深圳能源', datetime.date(2023, 4, 28), 0.1085, 8092758141, 634034000.0, 1.81,
('28', '国药一致', datetime.date(2023, 4, 26), 0.85, 18686817787, 362205000.0, 2.22,
('29', '深深房A', datetime.date(2023, 4, 29), -0.0352, 109155516, -35653800.0, -0.89,
```

图 2.2 部分存储结果显示

三、数据集介绍与预处理

本项目按照第二章所介绍的数据采集方法从东方财富网数据中心获取了本项目所需的数据集。本章将对数据集进行介绍，并说明数据的预处理过程。

第一节、数据集介绍

在爬取的数据集中，共包含了 5484 家上市公司，包括了国内所有股票市场的股票，例如 A 股市场、科创板、创业板以及 B 股和 ST 股。在数据分析过程中，因为沪市、深市的主板、中小板市场更能够代表市场的发展行情，且股票的变动较为稳定，因此排除掉 B 股、ST 股、科创板、创业板等板块的个股，保留主板、中小板的股票进行后续分析。

表 3.1 业绩报表数据集字段说明

字段名	说明
stock_code	股票代码
stock_name	股票简称
announce_date	最新公告日期
earnings_per_share	每股收益(元)
total_revenue	营业总收入(元)
net_profit	净利润(元)
roe	净资产收益率(%)
yoy_growth	同比增长(%)
yoy_profit	同比增长(%).1
net_assets_per_share	每股净资产(元)
operating_cash_flow_per_share	每股经营现金流量(元)
gross_profit_margin	销售毛利率(%)
qoq_growth1	季度环比增长(%)
qoq_growth2	季度环比增长(%).1
profit_distribution	利润分配
industry	所处行业

第二节、数据预处理

数据预处理是数据分析的重要环节，它的作用是对原始数据进行清洗、转换、合并等操作，使得数据适合进行后续的数据分析和建模。在这一章节中，我们将介绍如何进行数据清洗、数据转换、数据合并等预处理操作。数据预处理的好处在于，可以提高数据的质量和准确性，减少数据分析和建模中的误差和偏差。在数据预处理中共进行了三个数据预处理步骤，分别是去重、排序和缺失值处理。

首先进行去重，通过使用 drop_duplicates()函数，根据指定的列名去除重复的行，保留其中的第一行。这一步骤可以确保数据中不存在重复的记录，避免在数据分析和建模中出现误差和偏差。

其次是排序，通过使用 sort_values()函数，按照指定的列名对数据进行排序。这一步骤可以使数据更加有序，便于后续的数据分析和可视化展示。最后进行缺失值处理，通过使用 replace()函数，将数据中的缺失值替换为 0，以便在后续的数据分析和建模中进行计算和处理。这一步骤可以避免由于缺失值导致的计算错误和偏差，确保数据分析和建模的准确性和可靠性。通过使用 reset_index()函数，可以将数据的索引重置为默认的从 0 开始的整数索引。这一步骤可以确保数据的索引与数据本身一致，使得后续的数据操作更加方便和准确。

数据预处理后的部分数据集如下：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	股票代码	股票简称	最新公告日期	每股收益(	营业总收入	净利润(元)	净资产收益	同比增长(%	同比增长(%	每股净资产	每股经营现	销售毛利率	季度环比增	季度环比增	利润分配	所处行业
2	1	平安银行	2023/4/25 0:00	0.65	4.51E+10	1.46E+10	3.39	-2.400069	13.6	19.416727	5.6248583	0	8.3305	64.8639	0	银行
3	2	万科A	2023/4/29 0:00	0.1251	6.847E+10	1.446E+09	0.6	9.2663215	1.16	20.620089	0.5868302	15.447326	-58.7916	-74.0306	0	房地产开发
4	4	ST国华	2023/4/29 0:00	-0.0559	21255771	-7427186	-2.13	88.709844	40.92	2.5975119	-0.225304	60.87122	-74.1534	98.6205	0	软件开发
5	5	ST星源	2023/4/29 0:00	-0.0026	28171813	-2802013	-0.23	-3.716186	-279	1.1428906	-0.039592	5.6297168	-74.2662	98.3133	0	环保行业
6	6	深振业A	2023/4/29 0:00	-0.009	265640113	-12085882	-0.15	-24.40352	-127.2	5.8442907	-0.478643	37.565558	-87.9851	-105.5304	0	房地产开发
7	7	*ST全新	2023/4/29 0:00	-0.0007	49531240	-232199	-0.26	-10.15071	94.77	0.260604	0.1148717	21.906363	83.9268	-106.4916	0	房地产服务
8	8	神州高铁	2023/4/28 0:00	-0.0121	431343231	-32949017	-1.35	100.08013	35.82	1.59962	-0.038406	36.759884	-47.7689	95.0904	0	交运设备
9	9	中国宝安	2023/4/29 0:00	0.0928	8.232E+09	239346190	2.58	44.584768	44.94	3.6586576	0.6697792	18.531014	-13.5792	-58.8076	0	综合行业
10	10	美丽生态	2023/4/28 0:00	-0.0259	123962807	-27469373	-4.05	-32.60061	-1490.51	0.6406086	0.0052628	7.0529074	131.0061	94.8441	0	工程建设
11	11	深物业A	2023/4/28 0:00	0.0222	411469619	13216751	0.3	-67.23477	-93.98	7.4276567	-0.703296	24.590337	-27.6057	-83.8662	0	房地产开发
12	12	南玻A	2023/4/26 0:00	0.13	4.071E+09	396406087	3.04	46.126992	3.32	4.3148109	-0.09262	22.235428	-7.378	2.5168	0	玻璃玻纤
13	14	沙河股份	2023/4/28 0:00	0.67	389370140	161573955	12.95	172.88397	929.34	5.4875331	1.6754022	67.205078	-0.8574	153.8056	0	房地产开发
14	16	深康佳A	2023/4/29 0:00	0.0633	4.601E+09	152513923	1.98	-44.70465	46.12	3.2295458	-0.491176	3.2350531	-8.2044	109.4906	0	家电行业
15	17	深中华A	2023/4/28 0:00	0.0041	151527918	2848657.7	0.98	201.56639	415.12	0.4251079	-0.010949	5.5118519	-48.7387	132.2677	0	交运设备
16	19	深粮控股	2023/4/26 0:00	0.0915	1.339E+09	105444875	2.19	-38.70507	-23.28	4.2241947	-0.162095	18.520849	-31.2554	3.1696	0	贸易行业
17	20	深华发A	2023/4/25 0:00	0.0227	175868642	6432411.8	1.8	2.3794765	8.91	1.2714323	0.1186166	13.834652	19.8079	301.2683	0	光学光电子
18	21	深科技	2023/4/29 0:00	0.0647	3.934E+09	101012991	1.02	7.7513461	-58.3	6.664833	0.3281589	10.747005	-4.2026	20.7128	0	消费电子
19	23	ST深天	2023/4/29 0:00	-0.1344	42802915	-18652947	-14.27	-44.41144	45.3	0.874838	-0.019204	-2.585033	-39.5032	89.5126	0	水泥建材
20	25	特力A	2023/4/27 0:00	0.0586	339838493	25274085	1.66	142.6206	-25.24	3.5515216	-0.100678	14.085356	105.4178	18.2218	0	汽车服务
21	26	飞亚达	2023/4/25 0:00	0.2505	1.2E+09	103189489	3.23	2.2488567	19.5	7.7887418	0.191965	36.028188	23.892	179.7974	0	珠宝首饰
22	27	深圳能源	2023/4/28 0:00	0.1085	8.093E+09	634034448	1.81	4.1950124	22.45	6.0741526	0.3906079	22.149936	-24.6278	49.9252	0	电力行业
23	28	国药一致	2023/4/26 0:00	0.85	1.869E+10	362205346	2.22	8.9555924	43.53	38.512152	0.4904747	11.41608	-0.7413	-17.505	0	医药商业
24	29	深深房A	2023/4/29 0:00	-0.0352	109155516	-35653799	-0.89	-59.95688	-121.97	3.9247241	-0.181127	7.3983633	1.0414	-224.673	0	房地产开发
25	30	富奥股份	2023/4/28 0:00	0.05	2.803E+09	81870998	1.09	3.2885569	243.12	4.3405988	-0.356405	8.288931	-15.4785	-60.9446	0	汽车零部件
26	31	大悦城	2023/4/29 0:00	0.04	6.085E+09	171381769	1.1	-25.61487	269.9	3.6573821	0.6411803	30.28666	-58.6603	105.7638	0	房地产开发
27	32	深桑达A	2023/4/26 0:00	-0.0497	1.293E+10	-56570897	-0.96	13.327525	54.7	5.4389176	-1.949405	10.324007	-27.8272	-114.7081	0	通信设备
28	34	神州数码	2023/4/29 0:00	0.3193	2.712E+10	209104952	2.71	-5.183051	11.63	11.702812	0.1219489	3.6770771	-13.5892	-35.3422	0	计算机设备
29	35	中国天楹	2023/4/29 0:00	0.0192	1.117E+09	45912936	0.44	-7.27819	7.62	4.1076145	-0.070843	23.637462	-60.682	2126.5396	0	环保行业
30	36	华联控股	2023/4/29 0:00	0.02	169313506	29710063	0.56	-16.3827	-15.04	3.5951341	-0.218475	49.069682	-90.8075	-91.8292	0	房地产开发

业绩报表(处理后)

图 3.1 预处理后部分数据集

此外，根据所处行业生成了词云图，利用所处行业的所有值字段绘制，在词云图中即可体现 A 股市场所有上市公司的行业信息。同时，字体较大的行业表示 A 股市场中该行业的上市公司数量占比大。通过词云图可以看到，在所选范围的市场中，专用设备、汽车零部件等行业的个股数量较多。

图 3.2 行业占比词云图

四、实验设计与实验结果

本章将设计实验在业绩报表数据集上实现多维度指标分析与聚类预测。

第一节、实验设计

数据分析可以探索数据中隐藏的规律和关系。在这一章节中，将介绍如何使用数据分析工具进行数据挖掘、特征工程等操作。在本节进行了 10 余类的数据分析操作，这些类又分为四大类，首先是利用最新一年的数据展示根据不同行业进行数据分析，探寻行业之间的差异与逻辑；第二类是对单个行业内的上市公司进行分析，探寻行业内的数据奥秘；第三类是利用时间序列数据，探寻市场的现状与未来走势；第四类是利用时间序列数据，探寻行业的发展现状与未来发展预测。数据分析与可视化的代码由于篇幅过长存放在附件中，接下来将对数据分析的结果进行不同维度的展示与分析。

第二节、实验结果

2.1.1 2022 年度不同行业数据对比分析

（一）2022 年度业绩指标热力图

相关系数热力图是一种基于相关系数来展示数据关系的可视化工具。它的作用在于，可以更加直观地了解数据中的变量之间的相关关系，以及进行变量的比较和分析。相关系数是衡量两个变量之间关联程度的指标，其取值范围为-1 到 1。相关系数为 1 表示两个变量完全正相关，为-1 表示两个变量完全负相关，为 0 表示两个变量没有相关关系。相关系数热力图通常将相关系数的大小和方向映射到色彩上，使用渐变色进行填充，颜色的深浅表示相关系数的大小和方向，从而能够直观地显示变量之间的相关关系。

下图绘制了所有业绩指标之间的相关系数热力图，颜色深的表示相关性大，反之表示不相关，从图中可以看出净利润与营业收入，每股净资产与每股收益的相关性均大于 0.5，表示两两相关性较高，可以在以后作为一种指标去考量。

图 4.1 2022 年度业绩指标热力图

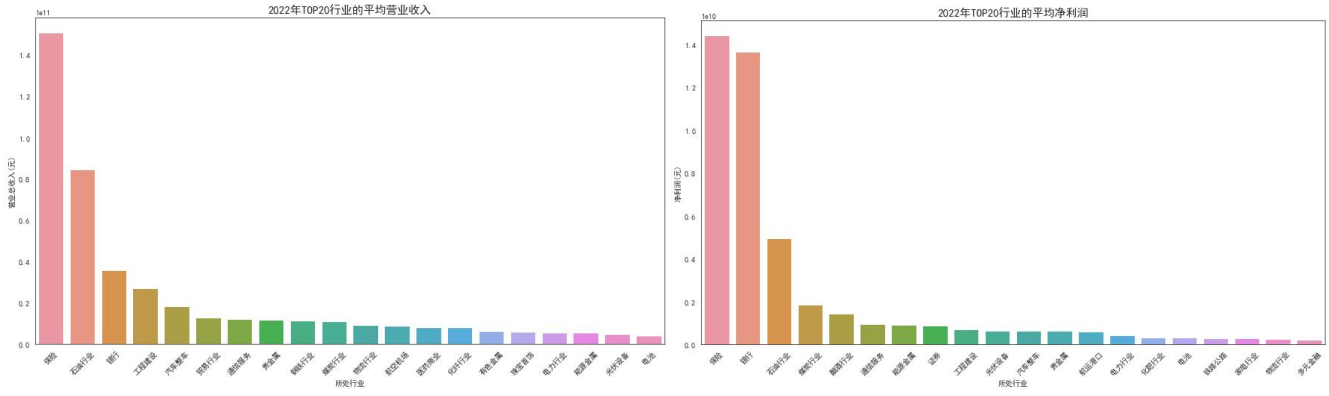
（二）不同行业营业收入与净利润均值比较

不同行业营业收入与净利润均值比较的柱状图和箱线图可以更加直观地了解不同行业之间的经营状况和经济效益，以及进行行业的比较和分析。

柱状图通常用于展示不同类别之间的数量或者比率关系，它将每个类别的数值映射到柱形的高度上，从而形成了一个直观易懂的图像。柱状图的优点在于，可以方便地比较不同类别之间的数值大小，快速了解各行业之间的经营状况和经济效益。箱线图通常用于展示数据的分布情况和离散程度，它将数据的中位数、下四分位数、上四分位数和异常值等信息映射到图像中，从而形成了一个直观易懂的图像。在这种比较中，可以使用箱线图来展示不同行业的营业收入和净利润均值的分布情况、中位数和离散程度，并进行比较和分析。箱线图的优点在于，可以直观地展示数据的分布情况和离散程度，发现数据中的异常值和趋势。

通过将柱状图和箱线图结合使用，可以更好地展示不同行业之间的经营状况和经济效益。通过比较不同行业之间的营业收入和净利润均值，可以了解各行业之间的盈利能力和市场竞争情况；通过比较不同行业之间的营业收入和净利润的分布情况和离散程度，可以了解各行业之间的经营风险和变化趋势。这些信息可以帮助做出更加准确的决策和战略规划，例如选择合适的投资行业、优化企业经营模式、制定市场营销策略等。

观察柱状图可以发现，保险、石油、银行三大行业的营业收入和净利润均位列前三名，其中保险业、石油行业的平均营业收入以及保险业、银行业的平均净利润均远超市场，成为市场的领军行业。观察箱线图可以发现，保险和银行行业的均值较高，其他行业排名靠前是因为有一些数据值较大的异常点带动行业发展，我们称之为行业龙头，之后我们将对单一行业的内部上市公司进行分析，判断离群点公司的发展与走势。



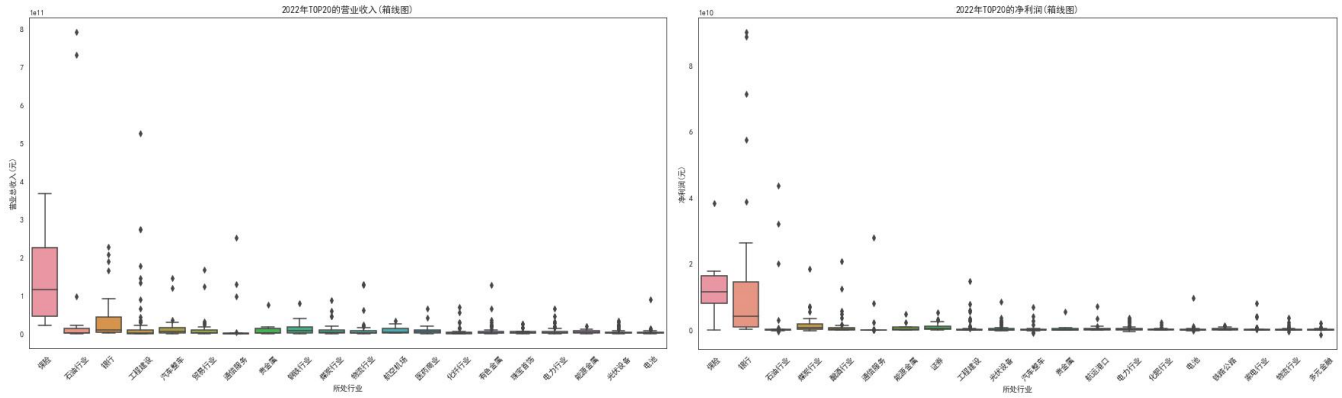


图 4.2 不同行业营业收入与净利润均值比较的柱状图\箱线图

（三）2022 年不同行业的净资产收益率比较

净资产收益率是公司净利润与净资产之比，用来衡量公司利润与投入资本之间的关系。净资产收益率值越高，说明公司的盈利能力越强，每一元的投资都能获得更多的收益。相反，净资产收益率值越低，说明公司的盈利能力较弱，需要更多的投资才能获得同等的收益。折线图的优点在于，可以直观地展示数据的变化趋势，帮助用户发现数据中的异常值和趋势。例如，通过比较不同行业之间的净资产收益率变化趋势，可以了解各行业之间的盈利能力和经济效益的变化情况。

通过观察折线图可以发现，煤炭行业、酿酒行业及保险业的盈利能力较强。

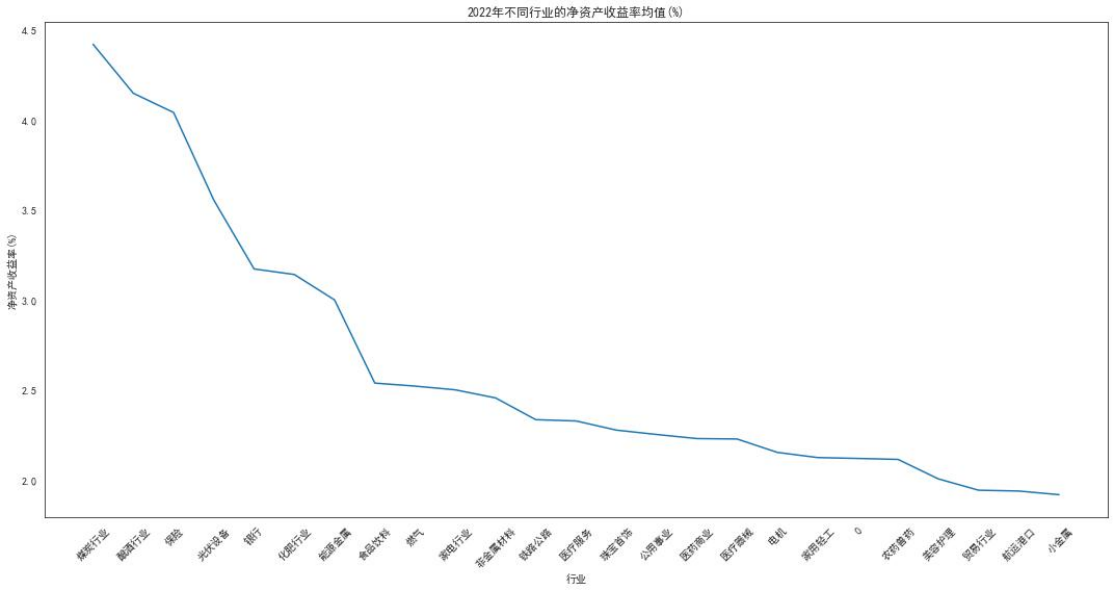


图 4.3 2022 年不同行业的净资产收益率比较

（四）2022 年不同行业营业总收入占市场比重

饼图可以帮助用户更加直观地了解不同行业之间的市场份额和市场竞争状况，以及进行行业的比较和分析。饼图通常用于展示各类别之间的数量或比例关系，它将每个类别的数值映射到圆饼的扇形区域上，从而形成了一个直观易懂的图像。在这种比较中，可以使用饼图来展示不同行业的营业总收

入占市场比重，并进行比较和分析。通过观察饼图上的扇形区域大小，可以直观地了解各个行业之间的市场份额和市场竞争状况。饼图可以方便地比较不同类别之间的数值大小，快速了解各行业之间的市场份额和市场竞争状况。

通过观察下面的饼图可以发现，工程建设、石油行业与银行业的占比较大，说明这些行业在市场中占据了较大的份额，具有较强的市场竞争力和品牌影响力，具有较高的市场需求和投资价值。通过比较不同行业营业总收入占市场比重的饼图，可以了解各行业之间的市场份额和市场竞争状况的差异。这些差异可能是由行业内部的市场竞争、技术创新、政策环境等因素所导致的。通过了解这些差异，可以制定相应的投资策略和市场营销策略，以获得更好的投资回报和市场竞争优势。

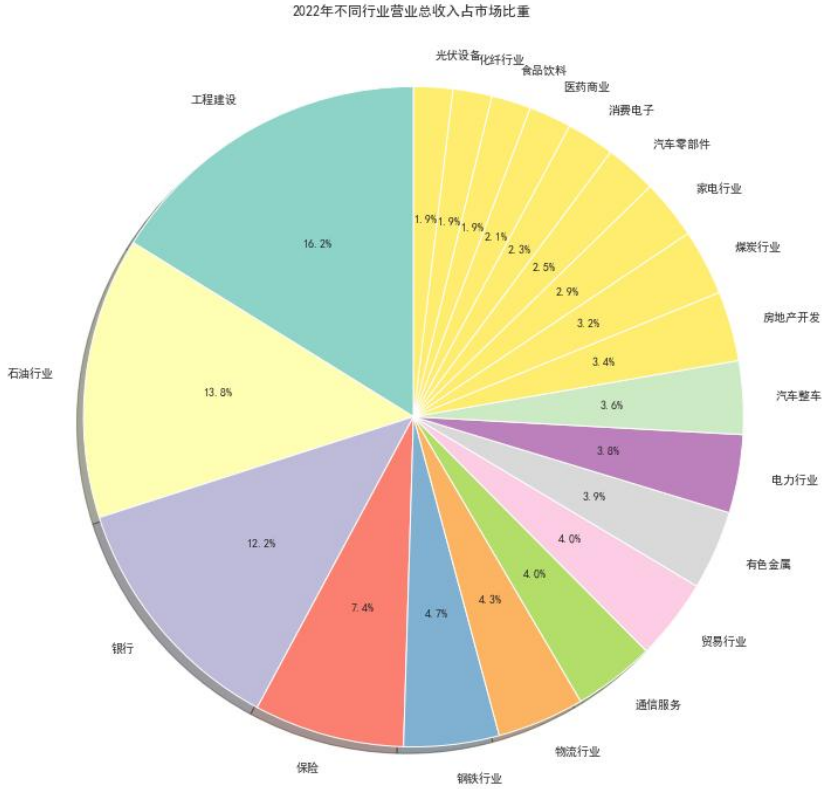


图 4.4 2022 年不同行业营业总收入占市场比重

接下来，以单个行业（银行业）为例，研究行业内部的上市公司的发展状况。

2.2.2 银行行业个股数据对比分析

（五）银行行业 2022 年上市公司每股净收益直方图

直方图通常用于展示数据的分布情况，它将数据分成若干个区间，统计每个区间内数据的数量或频率，并将其映射到直方图上。在这种比较中，可以使用直方图来展示银行行业上市公司每股净收益的分布情况，并进行比较和分析。通过观察直方图上的柱形高度和宽度，可以直观地了解银行行业上市公司每股净收益的分布情况和集中度。

通过观察直方图可以了解到银行行业上市公司每股净收益的分布范围。如果分布范围较小，可能

说明银行行业上市公司的每股净收益相对稳定，风险较低；如果分布范围较大，可能说明银行行业上市公司的每股净收益波动较大，风险较高。读图发现分布呈现左偏形态，分布范围较小，风险较低。

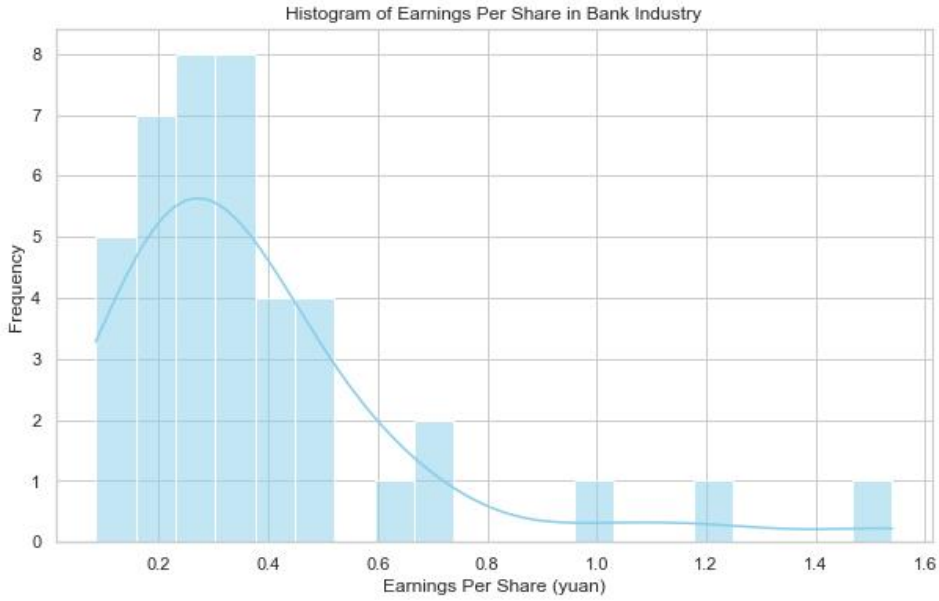


图 4.5 银行行业 2022 年上市公司每股净收益直方图

（六）2022 年银行行业上市公司营业总收入排名

下图展示了银行行业中上市公司 2022 年一年营业总收入的排名，通过柱状图可以很清楚比较值的大小，从图中可以看出工商银行、建设银行、农业银行、中信银行分列前四位，属于银行行业的龙头企业。

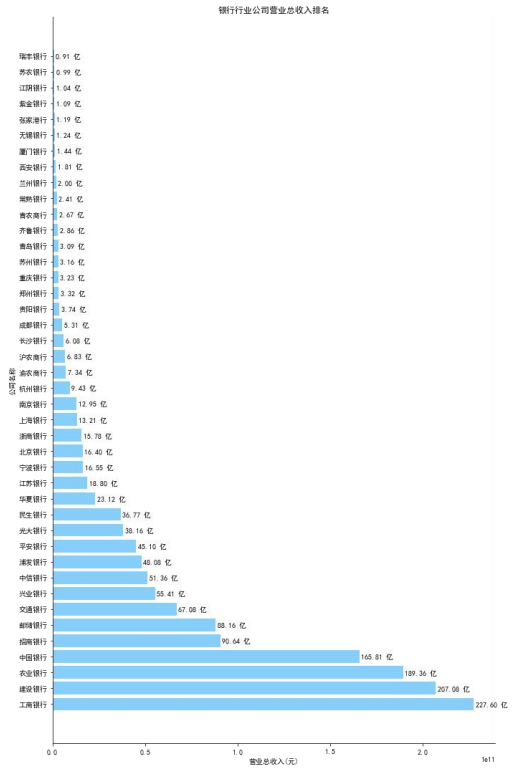


图 4.6 2022 年银行行业上市公司营业总收入排名

（七）2022 年银行行业各业绩指标上市公司排名 TOP10

通过多图折线图，可以展示银行行业各业绩指标上市公司排名，该图展示了 6 个业绩指标的上市公司排名，可以看到不同指标的排名，分析龙头公司与其他公司的发展差异。

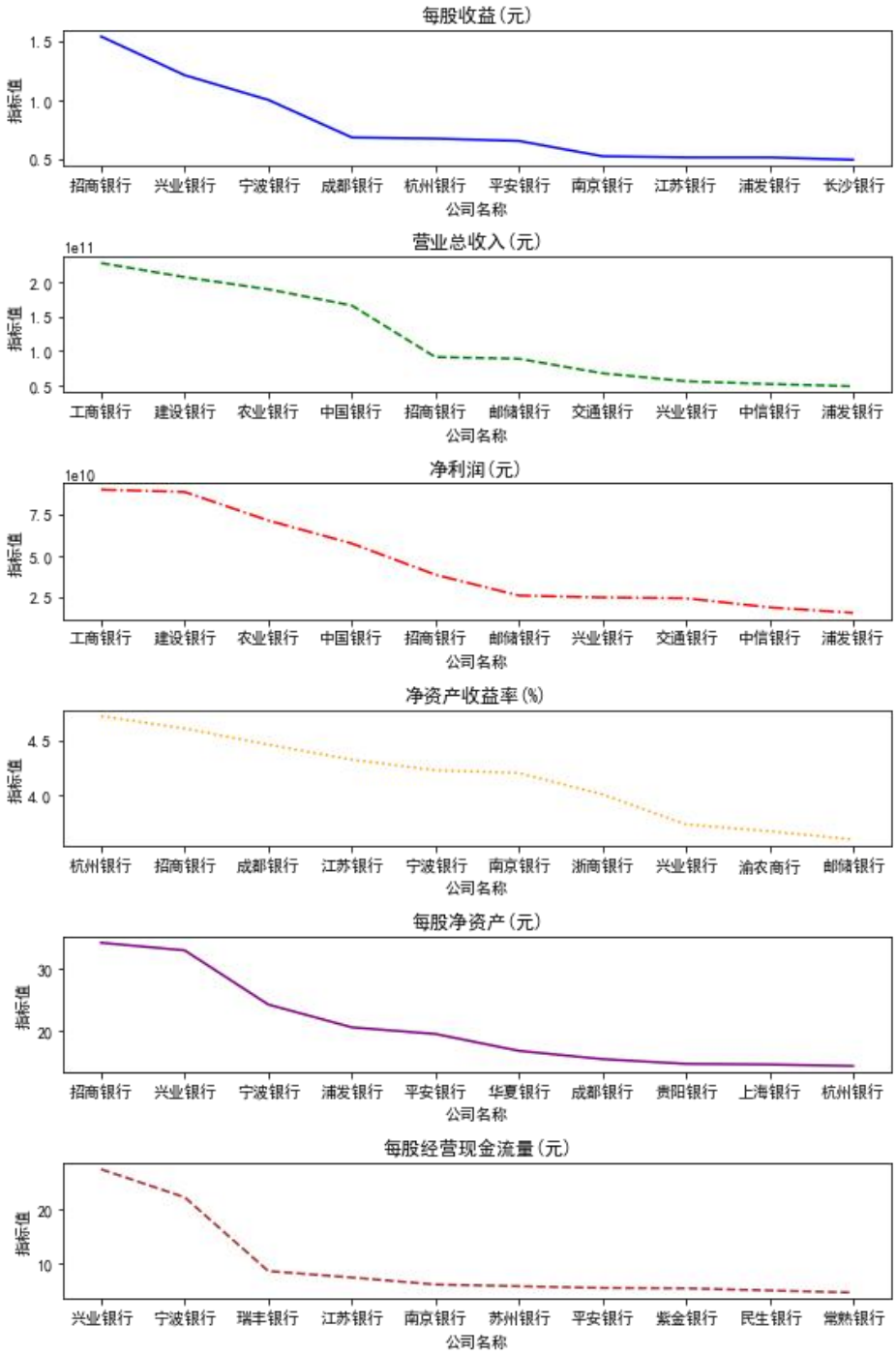


图 4.7 2022 年银行行业各业绩指标上市公司排名 TOP10

（八）2022 年银行行业与全部行业的营业总收入与净利润的关系

散点图可以用来展示两个变量之间的关系，其中一维通常被称为自变量，另一维被称为因变量。在本图中，营业总收入是自变量，净利润是因变量。通过绘制营业总收入与净利润的关系散点图，可

以帮助用户更好地了解这两个变量之间的关系。散点图可以帮助用户更好地了解行业内不同公司的表现。通过绘制银行行业的散点图，可以看出不同银行之间的表现差异；同样地，通过绘制全部行业的散点图，可以看出不同行业之间的表现差异。这有助于用户了解行业内各个公司的优劣势，并进行比较和分析。

散点图可以帮助用户更好地了解营业总收入与净利润之间的关系，并进行行业间或行业内的比较和分析。通过绘制银行行业和全部行业的散点图进行比较，可以更好地了解不同行业之间的关系和表现差异。通过识别异常值和预测趋势等分析，用户可以制定更加准确的投资策略和风险管理策略。

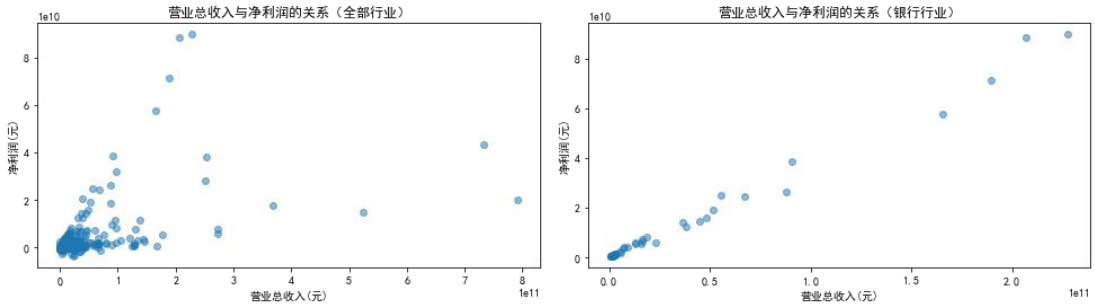


图 4.8 2022 年银行行业与全部行业的营业总收入与净利润的关系

2.2.3 市场发展现状与未来发展预测研究

（九）全市场近八年发展现状分析——营业收入

图 4.9 2016-2023 年全市场营业收入与增长率

通过观察 2016-2023 年全市场营业收入与增长率可以得出，全市场的营业收入在稳定上涨，在 2021 年受疫情影响出现小幅回调，2022 年后进行快速复苏阶段，增长率达峰值。

（十）全市场近八年发展现状分析——净利润

图 4.10 2016-2023 年全市场净利润与增长率

净利润的走势基本同营业收入一致，未来发展态势仍看好，有所期待。

（十一）全市场近八年发展现状分析——净资产收益率

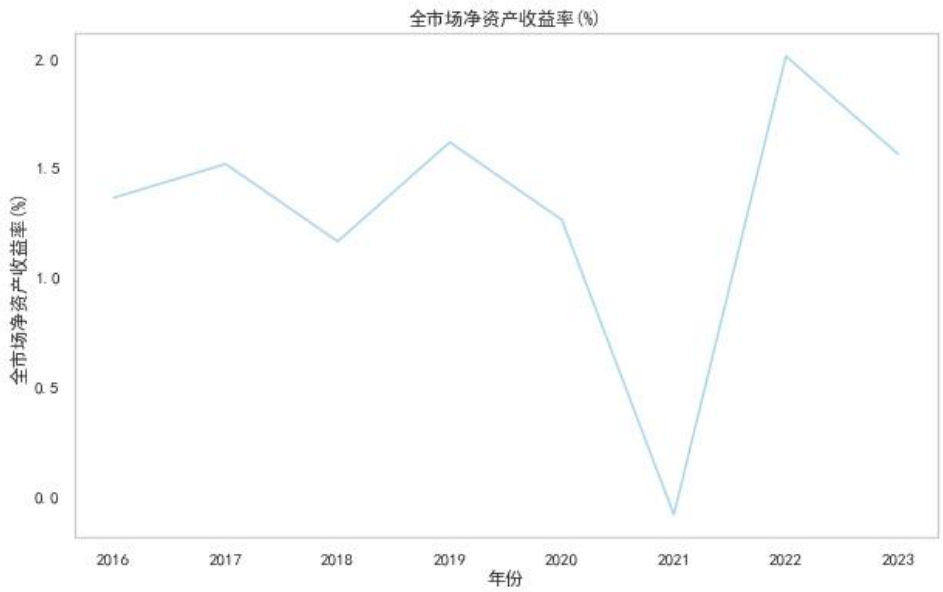


图 4.11 2016-2023 年全市场净资产收益率变化情况

净资产收益率是公司净利润与净资产之比，用来衡量利润与投入资本之间的关系。净资产收益率值越高，说明盈利能力越强，每一元的投资都能获得更多的收益。相反，净资产收益率值越低，说明盈利能力较弱，需要更多的投资才能获得同等的收益。通过图标看出，净资产收益率的变化总体在波动上升，对盈利能力持有积极态度。

（十二）线性回归模型预测营业总收入变化

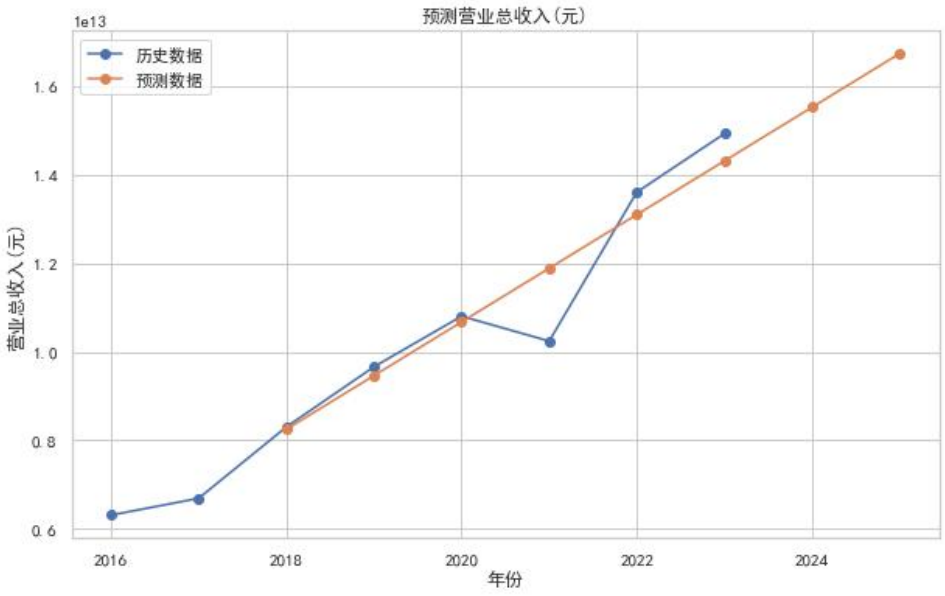


图 4.12 市场营业总收入预测

利用 python scikit_learn 中的线性规划模型预测市场营业总收入走势可以看出，总体对市场的发展持有经济复苏态度，营业总收入会逐年增长。

（十三）随机森林模型预测每股收益变化

	股票代码	股票简称	年份	每股收益(元)	营业总收入(元)	净利润(元)	净资产收益率(%)	同比增长(%)	每股净资产(元)	每股经营现金流量(元)	销售毛利率(%)	季度环比增长(%)	所处行业	预测每股收益
8	1	平安银行	2023	0.5600	4.620700e+10	1.285000e+10	3.27	10.574806	17.326136	7.580000	0.000000	9.5134	银行	0.303400
17	2	万科A	2023	0.1229	6.266707e+10	1.429295e+09	0.60	0.647209	20.375278	-1.140538	18.874781	-65.4369	房地产开发	0.081495
26	6	深振业A	2023	0.0329	3.513922e+08	4.443195e+07	0.58	-61.733943	5.696326	-0.090588	46.982195	-55.3106	房地产开发	0.039475
35	8	神州高铁	2023	-0.0189	2.155852e+08	-5.134018e+07	-1.11	44.408420	1.853004	0.022485	41.554463	-81.6051	交运设备	-0.043944
44	9	中国宝安	2023	0.0640	5.693657e+09	1.651342e+08	2.04	53.260302	3.173143	-0.183807	24.694564	5.4469	综合行业	0.118216
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30318	605580	恒盛能源	2023	0.2000	1.893398e+08	4.056710e+07	4.68	9.500287	4.438915	-0.062906	30.038930	-27.9219	电力行业	0.191075
30320	605588	冠石科技	2023	0.4200	3.716025e+08	2.294661e+07	5.92	89.576198	7.283276	-1.355555	13.151340	6.6762	半导体	0.344999
30325	605589	圣泉集团	2023	0.1600	2.226248e+09	1.254836e+08	1.57	16.526084	10.418595	-1.064828	21.284541	-9.6345	化学制品	0.091776
30327	605598	上海港湾	2023	0.6100	2.434481e+08	1.052070e+08	7.89	35.578116	8.710901	1.167310	60.964504	66.4467	工程建设	0.288666
30329	605599	菜百股份	2023	0.2300	3.469665e+09	1.796549e+08	7.87	17.549340	4.368773	0.969067	12.642746	10.2919	珠宝首饰	0.222654

真实值

预测值

4118 rows × 14 columns

```
均方误差(MSE): 0.21727788617390983
```

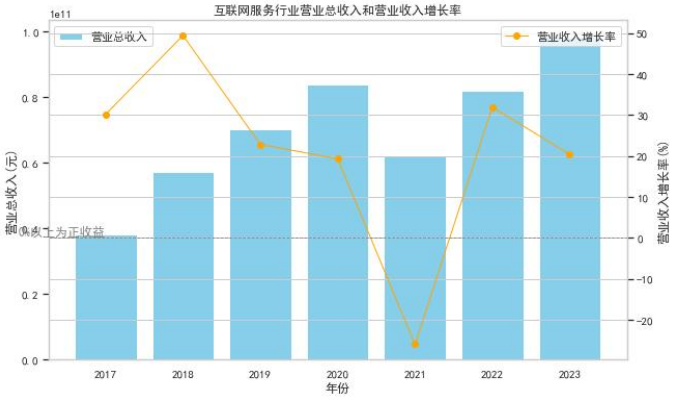
使用 python scikit_learn 中的随机森林模型，预测每股收益变化，且最大均方误差达到 0.21，说明模型在预测每股收益方面的误差相对较小。较小的均方误差表示模型的预测值与实际值之间的差异较小。可以用于下一步的预测与测试。

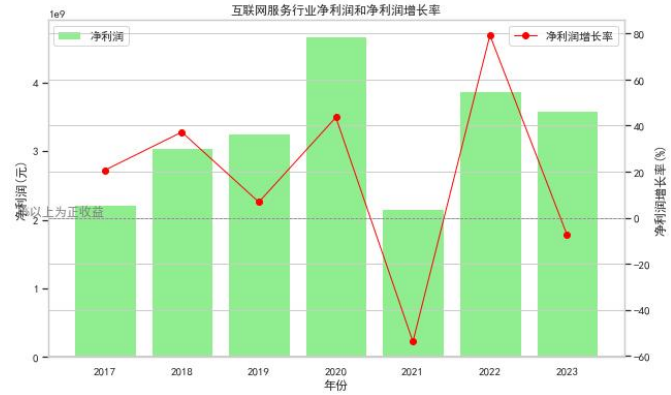
综合上面的发展现状分析可知，市场将具有较好的发展前景，且未来有望继续保持上升趋势。根据预测分析结果，可以给出投资者与企业一些建议：

- 加强战略规划：在市场发展前景看好的情况下，建议企业加强战略规划，制定长期发展计划，以适应未来市场变化和竞争格局的变化。
- 加强技术创新能力：随着行业的不断发展，技术创新能力越来越成为企业获得市场竞争优势的重要因素。建议企业加强技术研发和创新，以提高企业的技术实力和竞争力。
- 加强人才培养和管理：在行业发展快速的情况下，人才的角色越来越重要。建议企业加强人才培养和管理，吸引和留住优秀人才，为企业的长期发展提供有力支持。

2.2.4 互联网行业发展现状与未来发展预测研究

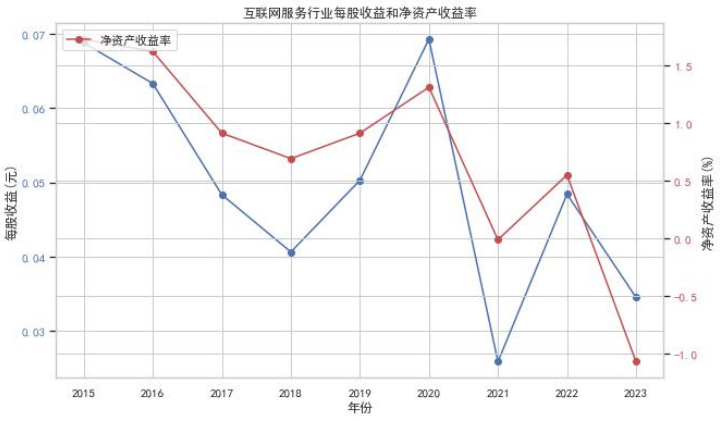
（十四）互联网行业发展现状分析——营业收入与净利润

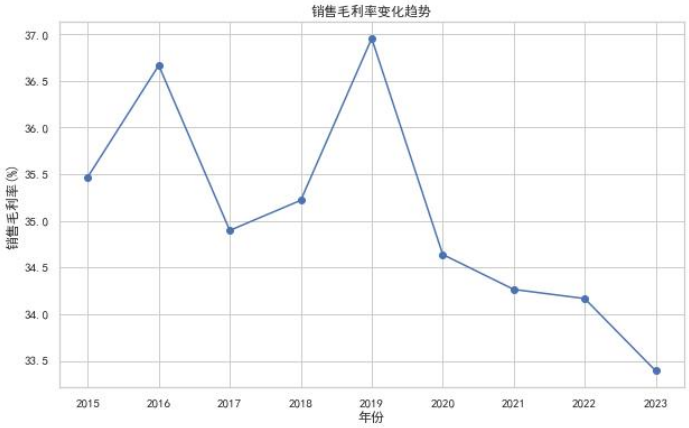




通过观察营业收入和净利润的变化图表可以看出，互联网服务行业的营业总收入与净利润指标呈现出逐年增长的趋势。尽管有些年份的增长幅度较小，但总体上表明该行业在过去几年里实现了稳定增长。且与刚才的市场走势有相似性。

（十五）互联网行业发展现状分析——其他指标





通过图表可以得出如下结论：

（1）每股收益变化趋势：从 2015 年到 2018 年，每股收益呈现逐年下降的趋势，这可能意味着互联网服务行业整体的盈利能力在这段时间内受到了一定的压力。从 2018 年到 2020 年，每股收益急速上涨，回到了 2015 年的水平。这可能是由于行业整体的盈利能力得到了提升，可能是由于市场需求的增长或者公司自身的战略调整所致。从 2020 年到 2021 年，每股收益急速下降到历史低点。这种下降可能与全球经济形势、行业竞争加剧或行业公司内部问题有关。之后，每股收益开始波动上涨，这可能意味着行业整体的盈利能力在逐渐恢复或增长。

（2）净资产收益率变化趋势：净资产收益率呈现波动下降的趋势，这表明行业内公司的投资回报率在整个时间段内有所下降。值得注意的是，净资产收益率在 2023 年呈现负值，这表示行业内公司在该年度的净利润无法覆盖其净资产。这可能是由于行业内公司面临的财务困境、高额成本或其他经营问题所致。

（3）销售毛利率是一个衡量企业销售产品或服务后所获得的毛利润占销售收入的比例，它是盈利能力的一个重要指标。如果毛利率逐年下降，这可能表明行业内企业的成本上升，或者竞争加剧，导致售价下降，从而导致盈利能力变差。

（4）每股经营现金流量均值变化可以解释为每股经营活动产生的现金流量数值变化较大，导致均值在一定时期内出现震荡的情况。这种情况可能说明行业的现金流量变化较为不稳定。

综合上述判断可得出预测结论，互联网行业的发展态势总结为震荡上行，还是不够稳定的，意味着整个行业正在经历波动性上升的趋势，但总体趋势是向上的。这种趋势可能表明行业中的竞争加剧，市场格局发生变化，但同时也意味着行业前景仍然看好，有发展潜力。在这种情况下，以下是几点建议：

- 加强市场研究：由于互联网行业的竞争加剧，市场格局可能发生变化。因此，建议公司加强市场研究，了解市场上的新趋势和变化，以便及时调整策略。
- 提高产品服务质量：在竞争激烈的市场环境中，产品或服务的质量是企业获得市场份额的关键因素。因此，建议公司注重提高产品或服务的质量，以满足消费者的需求。
- 提高技术创新能力：互联网行业的发展离不开技术创新。建议公司加强技术研发，提高技术创新能力，以满足市场需求。
- 加强资金管理：由于互联网行业的发展具有一定的不确定性，建议公司加强资金管理，控制风险，确保资金的安全和合理利用。

五、项目总结

本报告基于 Python 编程语言，使用爬虫技术从东方财富网上获取财务报表数据，并进行数据预处理、存储到 MySQL 数据库、数据分析和可视化等过程。通过对数据的处理和分析，得出以下结论：通过对财务报表数据的分析，可以得出公司的营业收入、净利润、总资产、负债总额等关键指标的趋势和变化，从而揭示公司的经营状况和财务状况。数据分析的结果可以为投资决策、经营决策提供重要的参考和依据。通过将数据以图表、图形等形式进行可视化呈现，可以更加直观地展示数据的趋势和变化，帮助用户更好地理解和分析数据。数据可视化可以帮助用户从多个角度去观察和分析数据，从而发现更多的信息和规律。

综上所述，本报告通过对东方财富网上的财务报表数据进行数据爬取、预处理、存储、分析和可视化等过程，得出了公司的关键指标的趋势和变化，为投资决策和经营决策提供了参考和依据。但是，本报告的分析结果可能存在一些局限性，例如数据获取的不完整性和准确性、数据分析方法的局限性等。为了改进分析结果的准确性和可靠性，可以采取以下改进方向：

- 可以增加资产负债表、净利润表、现金流表等财务报表，提高数据分析的全面性与覆盖性。
- 数据分析方面，可以采用更加先进的数据分析方法和技术，如机器学习、深度学习等，以挖掘更加深层次的信息和规律。数据可视化方面，可以使用更加多样化的图表和图形，以更好地呈现数据的特点和趋势。

参考文献

[1] 姚华威. 互联网企业证券业务发展策略研究[D].浙江大学,2017.

[2] 潘莹. 互联网社交媒体与股价崩盘风险[D].西南财经大学,2019.DOI:10.27412/d.cnki.gxncu.2019.001490.

[3] 唐东明. 聚类分析及其应用研究[D].电子科技大学,2010.

[4] 姜文. 基于 Hadoop 平台的数据分析和应用[D].北京邮电大学,2011.

[5] 冯伟. 聚类分析在金融数据分析中的应用研究[D].辽宁师范大学,2009.

[6] 段江娇,刘红忠,曾剑平.投资者情绪指数、分析师推荐指数与股指收益率的影响研究——基于我国东方财富网股吧论坛、新浪网分析师个股评级数据[J].上海金融,2014,No.412(11):60-64.DOI:10.13910/j.cnki.shjr.2014.11.012.