超市零售数据分析与预测

1. 问题背景分析

零售业务是当今商业环境中至关重要的一部分，对企业的运营和成功具有重要影响。通过对零售业务数据进行深入分析，我们可以揭示出一系列关键问题，这些问题涉及零售业务的特征、商品销售的主要影响因素以及未来的销售趋势。针对零售业务的需求，本文提出了以下四个问题：

1. 零售业务的特征

零售业务的特征涵盖了多个方面，包括但不限于订单数量、利润率、制造商、产品名称、发货日期、地区、客户信息等。通过深入分析这些特征，我们可以了解到不同制造商、产品类别在销售中的表现，掌握不同地区或时间段的销售状况，从而为企业制定更精准的销售策略提供依据。

2. 商品类别畅销特点

了解商品类别的畅销特点对于库存管理、市场定位至关重要。通过采用聚类算法，我们可以将商品进行分类，发现不同商品类别之间的相似性和差异性。同时，通过决策树和集成算法，可以深入分析每个商品类别的畅销特点，帮助企业更好地了解消费者的需求和喜好。

3. 销量主要影响

分析销售量的主要影响因素是优化销售策略的关键一步。通过回归分析等方法，我们可以揭示出订单日期、制造商、折扣、地区等因素对销售量的影响程度。这有助于企业更加精准地制定促销活动、调整供应链策略，从而提高销售效益。

4. 销售趋势预测

随着市场和消费者需求的不断变化，对未来销售趋势的准确预测对企业制定战略计划至关重要。采用时间序列分析和神经网络等方法，可以基于历史销售数据进行未来销售趋势的预测。这有助于企业提前调整生产计划、库存管理，并更好地迎接市场的挑战和机遇。

通过深入分析以上问题，我们可以为企业提供全面的业务建议，帮助其更好地理解市场、提高竞争力。

2. 预处理

2.1 数据转换

2.1.1 错误公式替换

观察原始数据，发现在 excel 软件打开后，利润一列中出现了错误的公式：

订单 Id	利润率	记录数	制造商	产品名称	利润	发货日期	国家	地区	城市	子类别	客户名称	折扣	数量	省/自治区	类别	细分	订单日期	邮寄方式	销售额
US-2015-	-47%	1	Fiskars	Fis	=NAME?	2016/4/29	中国	华东	杭州	用品	曹惠	40%	2	浙江	办公用品	公司	2016/4/27	二级	￥130
CN-2015-	34%	1	GlobeWei	GlobeWei	￥43	2016/6/19	中国	西南	内江	信封	许安	0%	2	四川	办公用品	消费者	2016/6/15	标准级	￥125
CN-2015-	13%	1	Cardinal	Cardinal	￥4	2016/6/19	中国	西南	内江	装订机	许安	40%	2	四川	办公用品	消费者	2016/6/15	标准级	￥32
US-2015-	-8%	1	Kleencut	Kleencut	=NAME?	2016/12/13	中国	华东	镇江	用品	宋良	40%	4	江苏	办公用品	公司	2016/12/9	标准级	￥321
CN-2014-	40%	1	KitchenA	KitchenA	￥550	2015/6/2	中国	中南	汕头	器具	万兰	0%	3	广东	办公用品	消费者	2015/5/31	二级	￥1,376
CN-2013-	34%	1	柯尼卡	柯尼卡	￥3,784	2014/10/31	中国	华东	景德镇	设备	俞明	0%	9	江西	技术	消费者	2014/10/27	标准级	￥11,130
CN-2013-	36%	1	Ibico	Ibico	￥173	2014/10/31	中国	华东	景德镇	装订机	俞明	0%	2	江西	办公用品	消费者	2014/10/27	标准级	￥480
CN-2013-	31%	1	Safco	SAFCO	￥2,684	2014/10/31	中国	华东	景德镇	椅子	俞明	0%	4	江西	家具	消费者	2014/10/27	标准级	￥8,660
CN-2013-	8%	1	Green Ba	Green Ba	￥47	2014/10/31	中国	华东	景德镇	纸张	俞明	0%	5	江西	办公用品	消费者	2014/10/27	标准级	￥588
CN-2013-	22%	1	Stockwel	Stockwel	￥34	2014/10/31	中国	华东	景德镇	系固件	俞明	0%	2	江西	办公用品	消费者	2014/10/27	标准级	￥154
CN-2012-	1%	1	爱普生	爱普生	￥4	2013/12/24	中国	西北	榆林	设备	谢雯	0%	2	陕西	技术	小型企业	2013/12/22	二级	￥434
CN-2015-	27%	1	惠普	惠普 墨水	￥640	2016/6/6	中国	东北	哈尔滨	复印机	康青	0%	4	黑龙江	技术	消费者	2016/6/1	标准级	￥2,369
CN-2013-	13%	1	Jiffy	Jiffy 局	￥89	2014/6/9	中国	华东	青岛	信封	赵婵	0%	3	山东	办公用品	消费者	2014/6/5	标准级	￥684
CN-2013-	26%	1	SanDisk	SanDisk	￥344	2014/6/9	中国	华东	青岛	配件	赵婵	0%	5	山东	技术	消费者	2014/6/5	标准级	￥1,327
CN-2013-	48%	1	诺基亚	诺基亚 充	￥2,849	2014/6/9	中国	华东	青岛	电话	赵婵	0%	2	山东	技术	消费者	2014/6/5	标准级	￥5,937
US-2014-	-38%	1	KitchenA	KitchenA	-￥3,963	2015/11/25	中国	华东	徐州	器具	刘斯云	40%	7	江苏	办公用品	公司	2015/11/22	一级	￥10,336
US-2014-	45%	1	Novimex	Novimex	￥38	2015/11/25	中国	华东	徐州	标签	刘斯云	0%	3	江苏	办公用品	公司	2015/11/22	一级	￥85
CN-2015-	46%	1	Memorex	Memorex	￥1,071	2016/10/4	中国	华东	上海	配件	白鹃	0%	7	上海	技术	消费者	2016/10/2	二级	￥2,330
CN-2015-	28%	1	Acme	Acme 尺子	￥24	2016/10/4	中国	华东	上海	用品	白鹃	0%	1	上海	办公用品	消费者	2016/10/2	二级	￥86
CN-2015-	2%	1	Avery	Avery 孔	￥2	2016/10/4	中国	华东	上海	装订机	白鹃	0%	5	上海	办公用品	消费者	2016/10/2	二级	￥138
CN-2015-	32%	1	Cardinal	Cardinal	￥127	2016/10/4	中国	华东	上海	装订机	白鹃	0%	6	上海	办公用品	消费者	2016/10/2	二级	￥397
CN-2015-	45%	1	三星	三星 办公	￥959	2016/10/4	中国	华东	上海	电话	白鹃	0%	7	上海	技术	消费者	2016/10/2	二级	￥2,133

在表中错误公式显示为“#NAME?”，点击查看其内容，发现为“=-@￥61”，于是全局替换“=-@￥”为“-￥”。

2.1.2 数值数据转换

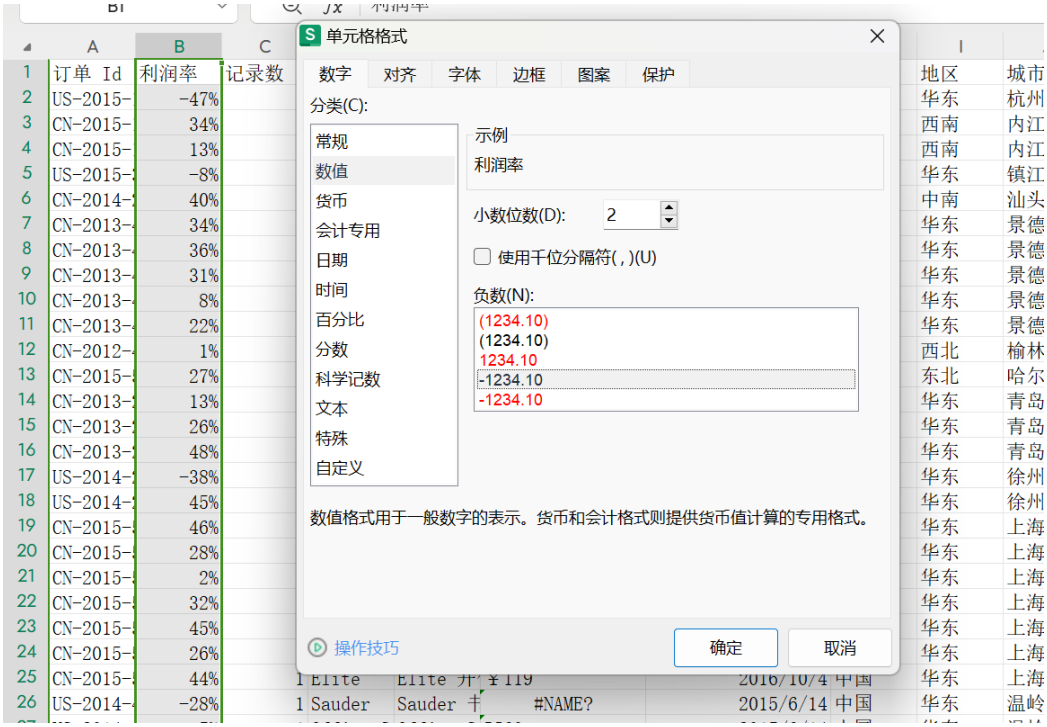
本文使用的数据的字段和描述如下：

订单 Id: 订单的唯一标识符。
利润率: 销售额中的利润占比。
记录数: 数据记录的数量。
制造商: 生产或提供商品的公司或制造商。
产品名称: 商品的名称或描述。
利润: 销售所产生的利润。
发货日期: 订单发货的日期。
国家: 订单所属的国家。
地区: 订单所属的地区。
城市: 订单所属的城市。
子类别: 商品的子类别。
客户名称: 下单的客户名称。

折扣：订单中的折扣金额或折扣率。
数量：订单中商品的数量。
省/自治区：订单所属的省份或自治区。
类别：商品的类别。
细分：商品的详细分类。
订单日期：订单生成的日期。
邮寄方式：订单的邮寄方式。
销售额：订单的总销售额。

其中销售额、利润、利润率、折扣等为数值类型，但由于销售额和利润都添加了“￥”符号，为字符串类型，从而影响数值数据的读取，全局删除掉该符号。

而对于利润率和折扣，在 excel 中选中，设置单元格格式，将百分号的数据转换为小数的数值数据：



2.2 数据清洗

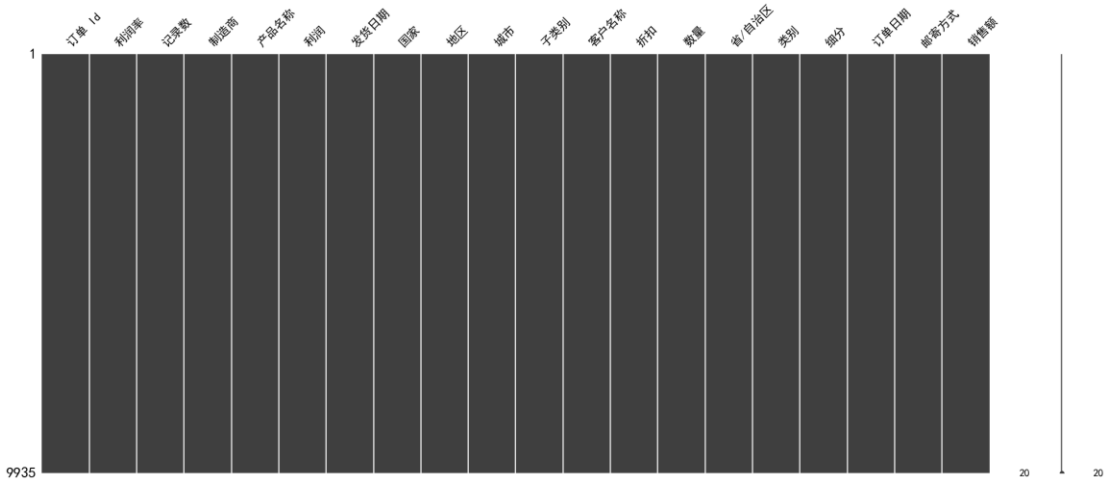
数据清洗是指对原始数据进行处理，以确保数据的质量和可用性。在实际的数据分析和机器学习项目中，原始数据可能包含各种问题，如缺失值、错误数据、重复数据等，这些问题会影响到分析的结果和模型的性能。因此，数据清洗是数据预处理的一个重要步骤，旨在净化数据，提高数据的可信度和可用性。

2.2.1 缺失值处理

缺失值是指数据集中某些字段或观测值缺少实际数值或信息的情况。在实际数据中，缺失值是比较常见的情况，可能是由于测量误差、设备故障、用户未填写等原因导致的。缺失矩阵可以帮助分析数据中缺失值的分布情况，了解哪些特征或观测值更容易出现缺失。本文通过缺失矩阵来帮助数据清洗和分析的可视化，使用了 python 中的 missingno 库，可视化代码如下：

```
# 缺失矩阵
import pandas as pd
import numpy as np
import missingno as msno
import matplotlib.pyplot as plt

# 设置全局字体为宋体
plt.rcParams['font.sans-serif'] = ['SimHei']

# 读取 CSV 文件
file_path = 'superMarket.csv'
data = pd.read_csv(file_path)

# 构建缺失矩阵
missing_matrix = msno.matrix(data)
plt.show()
```

本文使用的超市销售样本数据的缺失值情况通过缺失矩阵展示，见下图。从图中可以看出，样本数据共 9935 条记录中的 20 个特征值都没有出现缺失值。通过可视化的缺失矩阵分析，可以快速、直观的分析数据的缺失值密度情况，确保数据的完整性。在发现缺失值后，可以依据实际情况，采用删除缺失值，使用均值、中位数等合理数值填充缺失值，采用线性插值、多项式插值等插值方法，通过模型来预测缺失值等方法来填充缺失值。



2.1.2 去除重复数据

重复数据可能会对数据分析和建模产生误导，影响结果的准确性。在统计分析或机器学习模型中，重复样本可能导致模型对某些特征或类别过度拟合。去除重复数据有助于提高分析的准确性。本文使用下面代码来识别重复数据：

```
# 重复值处理

duplicates=cleaned_data.duplicated()

# 打印重复行的数量和具体的重复行

print("重复行数量:", duplicates.sum())

print("重复行:")

print(cleaned_data[duplicates])
```

结果如下图，未发现有重复数据。

```
重复行数量: 0
重复行:
Empty DataFrame
Columns: [订单 Id, 利润率, 记录数, 制造商, 产品名称, 利润, 发货日期, 国家, 地区, 城市, 子类别, 客户名称, 折扣, 数量, 省/自治区, 类别, 细分, 订单日期, 邮寄方式, 销售额]
Index: []
```

2.1.3 处理错误数据

处理错误数据是数据清洗阶段的一项重要任务，旨在识别和纠正数据集中存在的错误或异常值。错误数据可能是由于记录错误、测量误差、输入错误或其他不合理的数据值而引起的。这些错误可能导致对数据的不准确分析和模型建设，因此在进行数据分析和机器学习前，需要对错误数据进行处理。

通过观察数据的统计摘要，能够初步了解是否存在异常数据。describe 是 Pandas 库中的一个用于生成描述性统计信息的方法。该方法对数据集的每一列进行统计分析，返回包括计数、均值、标准差、最小值、25%，50%（中位数）、75%分位数和最大值在内的摘要统计信息。获取数据统计信息的代码如下：

```
data.describe()
```

观察对超市销售数据的统计信息（见下图），可以发现在多个数值特征中，存在一些数据异常值，如利润中，最大值为 10108，远远大于平均值 213.63 以

及上四分位点 276，于是推测数据中存在异常值需要处理。

	利润率	记录数	利润	折扣	数量	销售额
count	9935.000000	9935.0	9935.000000	9935.000000	9935.000000	9935.000000
mean	0.085460	1.0	213.632612	0.107111	3.768093	1608.892099
std	0.465521	0.0	856.568912	0.187930	2.236421	2630.089315
min	-4.000000	1.0	-7978.000000	0.000000	1.000000	13.000000
25%	0.020000	1.0	7.000000	0.000000	2.000000	249.500000
50%	0.170000	1.0	74.000000	0.000000	3.000000	637.000000
75%	0.340000	1.0	276.000000	0.200000	5.000000	1785.000000
max	0.500000	1.0	10108.000000	0.800000	14.000000	35621.000000

本文的异常值识别和处理方法为，对于多个数值特征，通过设定阈值来检测那些偏离均值较远的样本，如果存在某个特征上偏离超过阈值，就标记为异常值，处理数据的代码如下：

```
# 异常值处理
import pandas as pd

# 读取数据
data = pd.read_csv('superMarket.csv')

# 选择数值特征
numeric_features = ['利润率', '记录数', '利润', '折扣', '数量', '
销售额']

# 计算每个数值特征的均值和标准差
means = data[numeric_features].mean()
std_devs = data[numeric_features].std()

# 定义阈值，例如，均值加减 2 倍标准差
threshold = 2

# 标识异常值
outliers = ((data[numeric_features] - means).abs() > threshold *
std_devs).any(axis=1)

# 打印异常值
# print(outliers)
print("异常值数量:", outliers.sum())
print("异常值:")
print(data[outliers])
```

```
# 去除异常值
cleaned_data = data[~outliers]
cleaned_data.to_csv("preprocessed.csv",index=False)
```

数据异常值的处理结果如下：

```
异常值数量: 371
异常值:
                订单 Id   利润率  记录数             制造商                    产品名称
5     CN-2013-4497736  0.34    1             柯尼卡             柯尼卡 打印机, 红色  \
15    US-2014-2511714 -0.38    1      KitchenAid      KitchenAid 冰箱, 黑色
46    US-2014-5956361 -3.75    1          Boston          Boston 速写本, 蓝色
49    CN-2015-2396895  0.11    1              思科             思科 充电器, 全尺寸
52    CN-2014-2828982  0.28    1  Hamilton Beach  Hamilton Beach 炉灶, 黑色
...                ...   ...  ...             ...                       ...
9706  CN-2013-4950934  0.42    1              宜家              宜家 书库, 传统
9777  US-2013-1436231 -0.40    1              三星             三星 充电器, 蓝色
9810  CN-2014-4271574  0.00    1          Hoover          Hoover 炉灶, 白色
9891  CN-2012-2508741  0.08    1          Hoover          Hoover 冰箱, 银色
9934  CN-2012-3557528 -0.48    1        Breville        Breville 冰箱, 白色

          利润        发货日期  国家  地区   城市 子类别 客户名称   折扣  数量 省/自治区    类别   细分
5     3784.0  2014/10/31  中国  华东  景德镇  设备   俞明  0.0   9    江西    技术  消费者  \
15   -3963.0  2015/11/25  中国  华东   徐州  器具  刘斯云  0.4   7    江苏   办公用品   公司
46   -1021.0   2015/6/18  中国  西南   内江  美术   邹涛  0.8   6    四川   办公用品   公司
49    1340.0   2016/6/22  中国  东北   蛟河  电话   薛磊  0.0   4    吉林    技术  消费者
52    3539.0   2015/5/26  中国  华东   青岛  器具  苏晒明  0.0   5    山东   办公用品  消费者
...       ...         ...  ..  ..  ...  ..  ...  ...  ..   ...   ...  ...
9706  4273.0   2014/1/21  中国  中南   广州  书架   涂丽  0.0   6    广东    家具  消费者
9777 -5701.0   2014/12/6  中国  中南   襄樊  电话   贺立  0.4   8    湖北    技术   公司
9810     0.0   2015/6/22  中国  中南   南阳  器具   赖虎  0.0   5    河南   办公用品   公司
...
9891   2013/9/26   一级  12297.0
9934   2013/12/2  标准级   7245.0
```

3.探索性数据分析

通过探索性数据分析，对数据集的整体结构进行初步了解。使用统计指标和可视化工具，深入分析利润、销售额等特征的分布情况，并对订单日期、发货日期等时间特征进行详细研究，以发现潜在的销售趋势。

3.1 探究问题一：零售业务的特征

零售业务的特征是指在零售行业中常见的、具有代表性的一些业务属性或数据方面的特点。这些特征可以涵盖多个方面，反映了零售业务的经营状况、市场表现、客户行为等各个方面的信息。订单数量，销售额等是一些零售业务的典型特征。

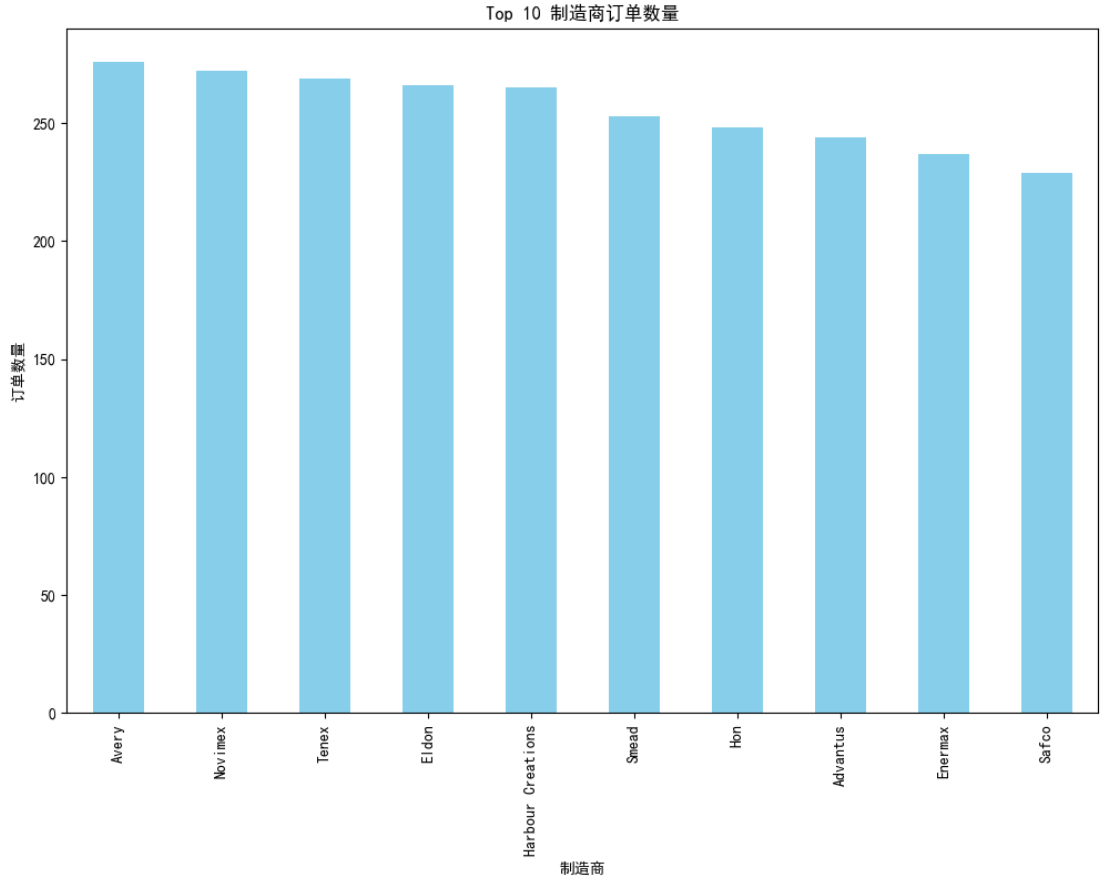
3.1.1 订单数量统计

零售业务通常涉及大量的订单。订单数量的分布和趋势反映了企业的销售活动繁忙程度，从多方面统计订单数量有助于全面了解零售业务的运营情况，并提供更全面、深入的业务洞察。通过按制造商、商品类别、消费者群体等多

个维度统计订单数量，可以将市场细分为不同的部分。这有助于识别不同市场细分的销售状况，找到潜在的市场机会和挑战。

1. 制造商订单数量可视化

不同制造商提供的产品在市场上的受欢迎程度和销售状况可能存在差异。通过分析制造商的订单数量，企业可以了解哪些制造商的产品更受欢迎，哪些制造商的产品销售较差，有助于制定合作策略、优化供应链，提高盈利能力。

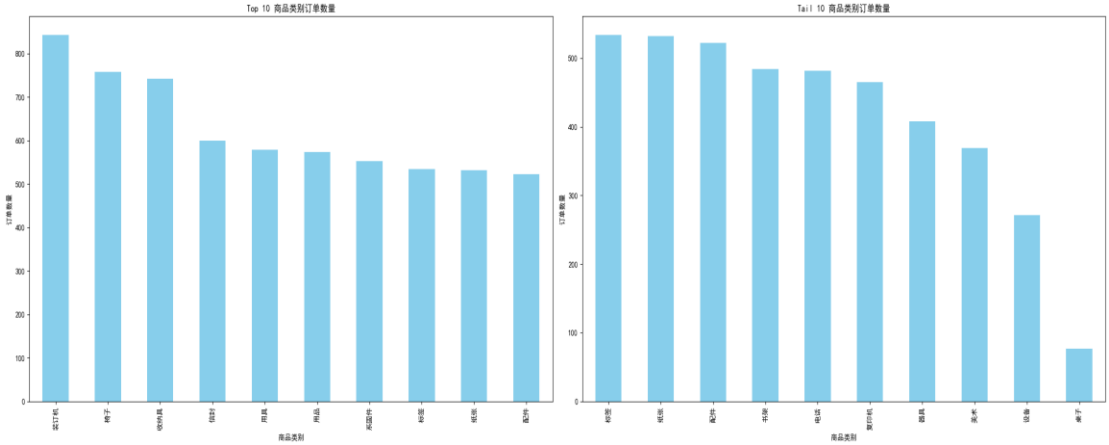


通过上图可以看出，在本文使用的超市的数据中，制造商的订单数量的分布差异不是很大，没有订单数量明显较高的制造商，制造商中订单数量最高为276 单，由此推测制造商对订单数量的影响不大。

2. 商品类别与订单数统计

研究商品类别与订单数统计对于零售业务的经营和策略制定具有重要意义，通过对不同商品类别的订单数进行统计，可以了解哪些商品类别在市场上更受欢迎，具有更高的销售量，从而确定畅销商品。反之，对于订单较少的商品类别，制造商或者商家可能需要重新评估其市场表现，考虑是否需要调整库存策略或进行促销活动。本文分别绘制了 top10 和 tail10 的商品类别和对应订单数

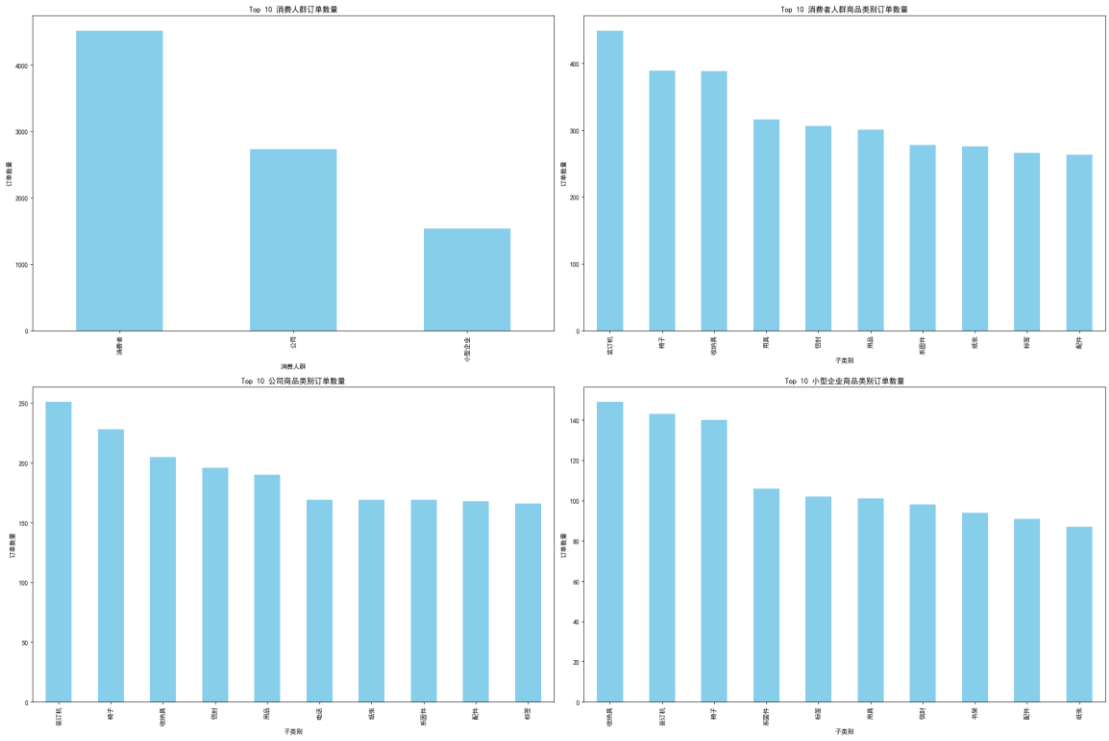
量的柱状图：



结合上图分析发现装订机、椅子、收纳具三种商品的订单数量相比其他商品明显较多，分别为 843 单、757 单、742 单。这三种商品可以认为是该超市的畅销商品。在订单数量排行尾部的商品类别中，发现桌子的订单数量很少，只有 77 单，结合成本等因素，超市可以考虑减少或者停止销售部分品牌的桌子。

3. 消费人群与订单数统计

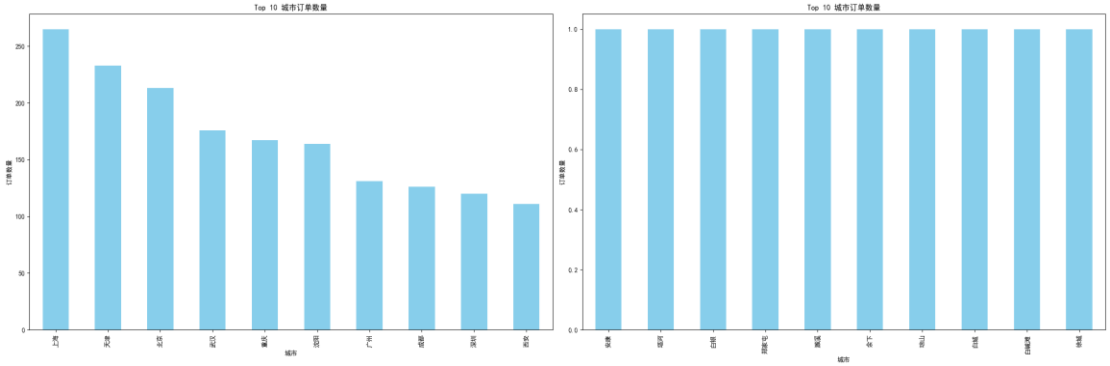
不同消费者类别（如普通消费者、公司、小型企业）的购买习惯和订单数量可能存在显著差异。分析不同类别消费者的订单数量有助于企业理解不同客户群体的需求，制定个性化的营销策略，提高客户满意度，促进客户忠诚度。本文采用柱状图的方式，可视了不同消费人群的订单数量情况，每个消费人群中不同商品类别的订单数量分布情况：



结合上图分析得知，本超市面向的主要消费人群为普通消费者，其总订单数为 4519 单，其次为公司，其总订单数为 2734 单，最后是小型企业，其总订单数为 1538 单。这三个消费群体对商品的类别的喜好有所不同，如公司的订单数量 top10 的商品类别中的电话并没有在普通消费者喜好的 top10 商品中出现。在前三的商品类别中，虽然顺序不一致，但三类消费群体都为装订机、椅子、收纳具，这是不同消费群体间的共同点。

4. 城市与订单数统计

通过统计不同城市的订单数量，超市可以了解产品在不同地区的市场分布情况。这有助于确定销售的重心区域，使超市能够更有针对性地制定地区性的销售策略，提高效益。不同城市的人口规模、经济发展水平以及消费水平存在差异，订单数量的统计也可以帮助制造商评估不同城市的市场潜力。对于订单数较高的城市，制造可以考虑加大市场投入。本文通过柱状图统计了在订单数量上 top10 和 tail10 的城市：

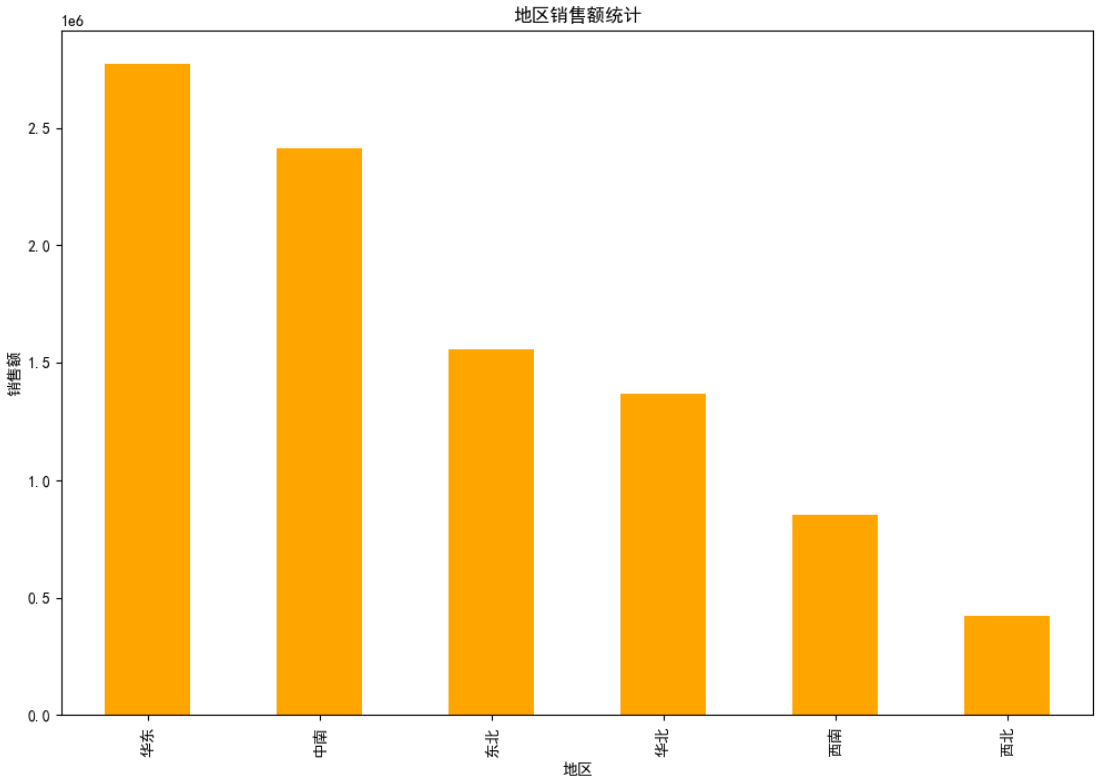


结合上图分析，可以发现前几个城市的订单数量领先幅度很大，top3 的城市为上海（订单数量 265 单）、天津（订单数量 233 单）、北京（订单数量 213 单），而尾部的城市订单数量均只有一单。由此可见城市间的消费能力，市场潜力的差异是很大的，超市和制造商等可以考虑制订相应的销售和资本投入策略。

3.1.2 销售额统计

5. 地区销售额统计

考虑城市的数量较多、分布比较分散，我们进一步研究地区的消费能力和市场潜力的差异。本文采用柱状图来可视化不同地区的销售额：

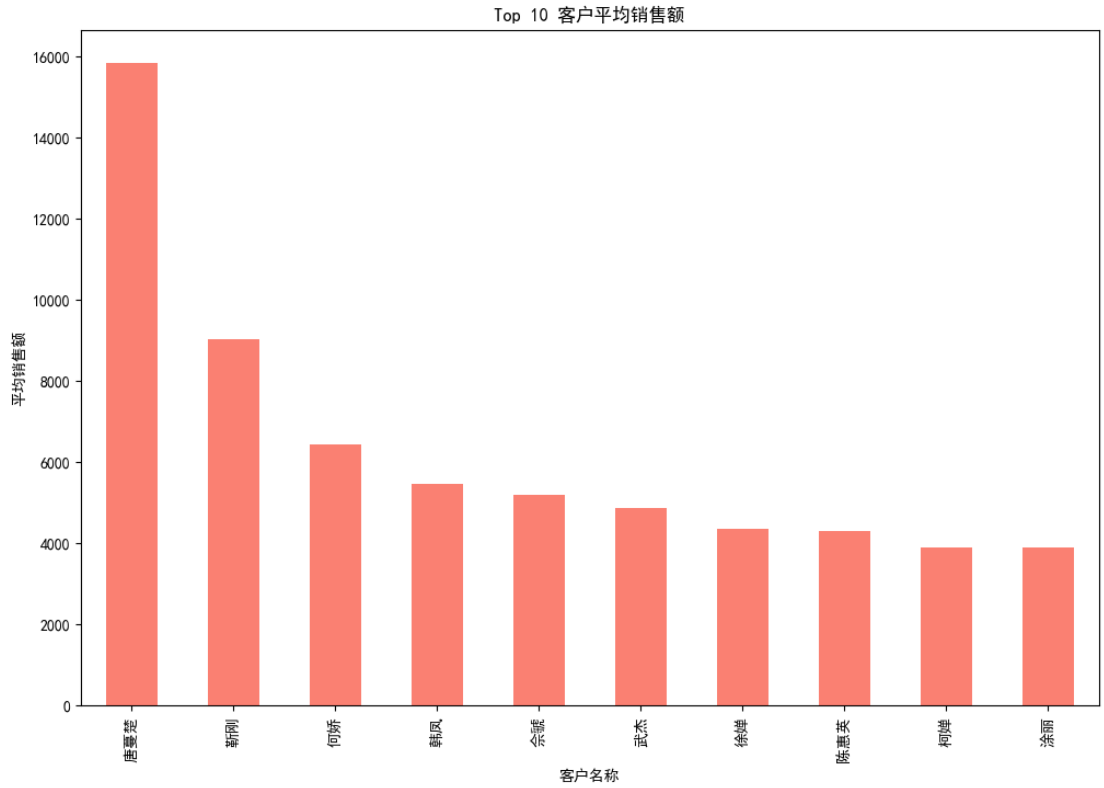


结合上图可以发现，商品销售额的地区差异极大。消费额较高的地区为华东和中南地区，分别为 277 万元和 241 万元。而西南和西北的销售额较低，分别为 85 万元和 42 万元。可以得出结论，该超市的主要销售区域为长江流域中下游地区。

6. 客户销售额统计

通过统计不同客户的销售额，企业可以了解哪些客户对业务的贡献较大。一些高价值客户可能会对企业的收入产生显著影响，因此，了解客户销售额有助于识别和重点关注高价值客户。本文通过柱状图来展示 top10 的高价值用户

与其贡献的销售额：



结合上图可以发现有几位客户提供销售额比较大，尤其是前三的客户，本文进一步提取出了这三位客户的订单信息：

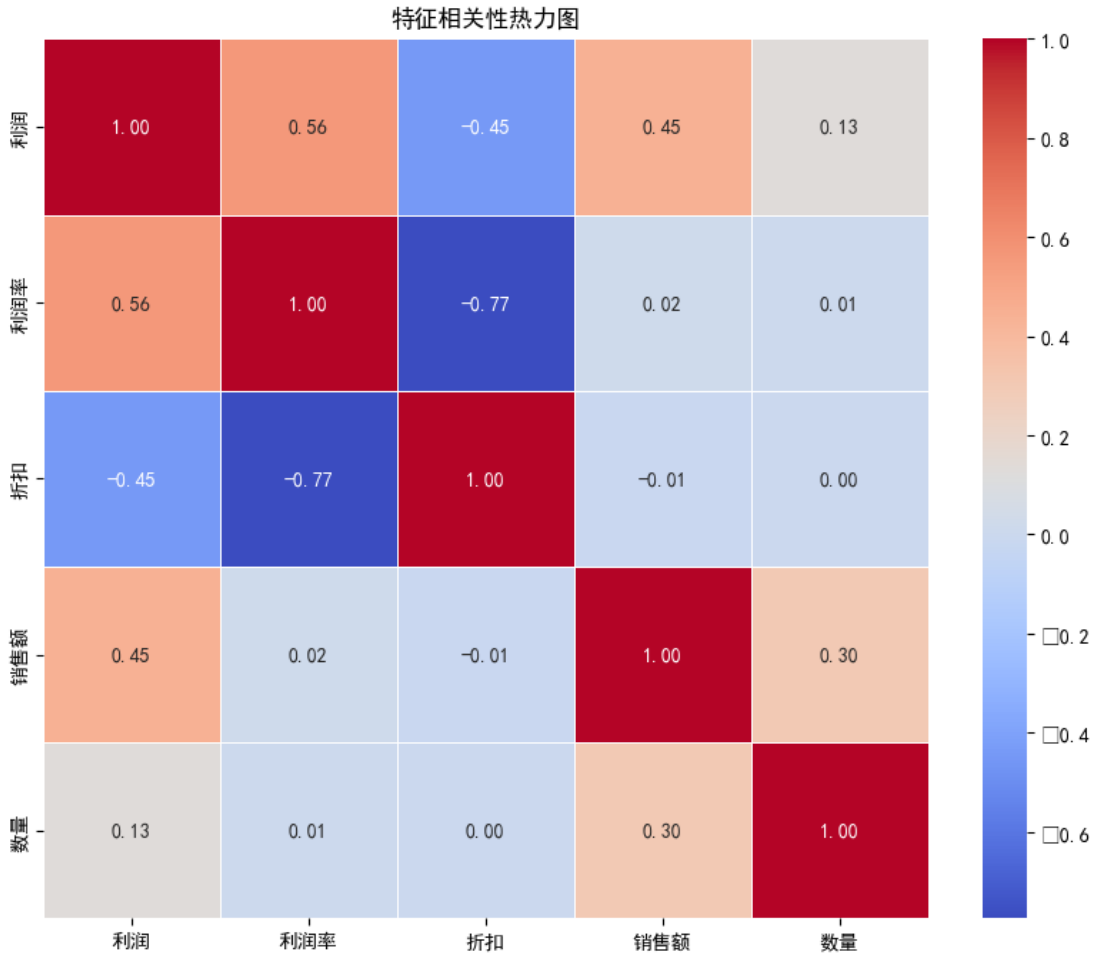
	产品名称	子类别	客户名称	细分	销售额
330	Stiletto 修剪器，锯齿状	用品	唐蔓楚	公司	368.0
331	Bevis 会议桌，组装	桌子	唐蔓楚	公司	35621.0
332	Hoover 微波炉，银色	器具	唐蔓楚	公司	11551.0
4446	Hon 运输标签，可调	标签	何娇	公司	260.0
4647	Hewlett 无线传真机，彩色	复印机	何娇	公司	3562.0
8352	SAFCO 扶手椅，红色	椅子	何娇	公司	15504.0
9358	摩托罗拉 充电器，全尺寸	电话	靳刚	公司	9016.0

可以发现这些客户均为公司，购买的商品类别也很分散，包含会议桌、微波炉、扶手椅等。

3.1.3 特征分析

7. 特征关系分析

特征关系分析是数据探索和建模过程中不可或缺的一部分。通过深入了解特征之间的关系，可以更好地引导特征选择、特征工程和模型优化的过程，提高机器学习模型的性能和可解释性。本文通过热力图来分析数值型特征间的关系：



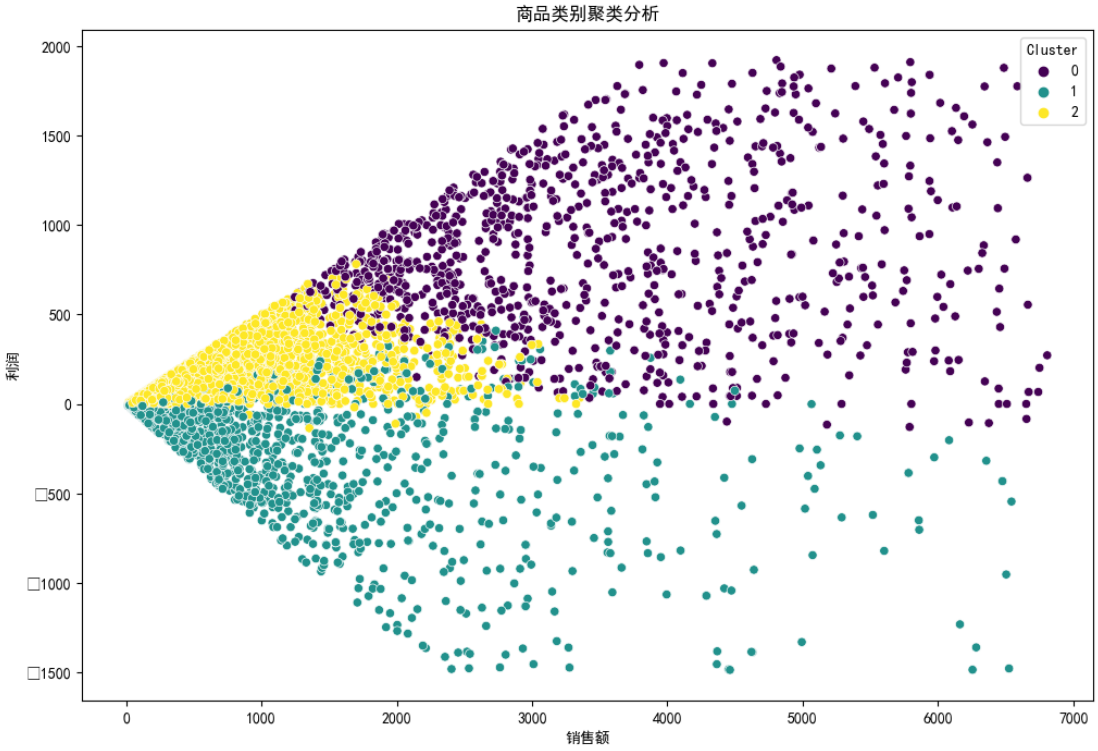
我们重点关注销售额的影响因素。结合上图，发现销售额和利润以及数量的相关性比较强，且都是正相关，相关系数分别为 0.45 和 0.30，而销售额和利润率以及折扣的相关性弱，接近 0。

3.2 探究问题二：商品类别畅销特点

3.2.1 畅销商品聚类

经过热力图分析后，发现销售额和利润的相关性比较强，本文选取这两个因

素进行了聚类分析，探究畅销商品和利润的关系。

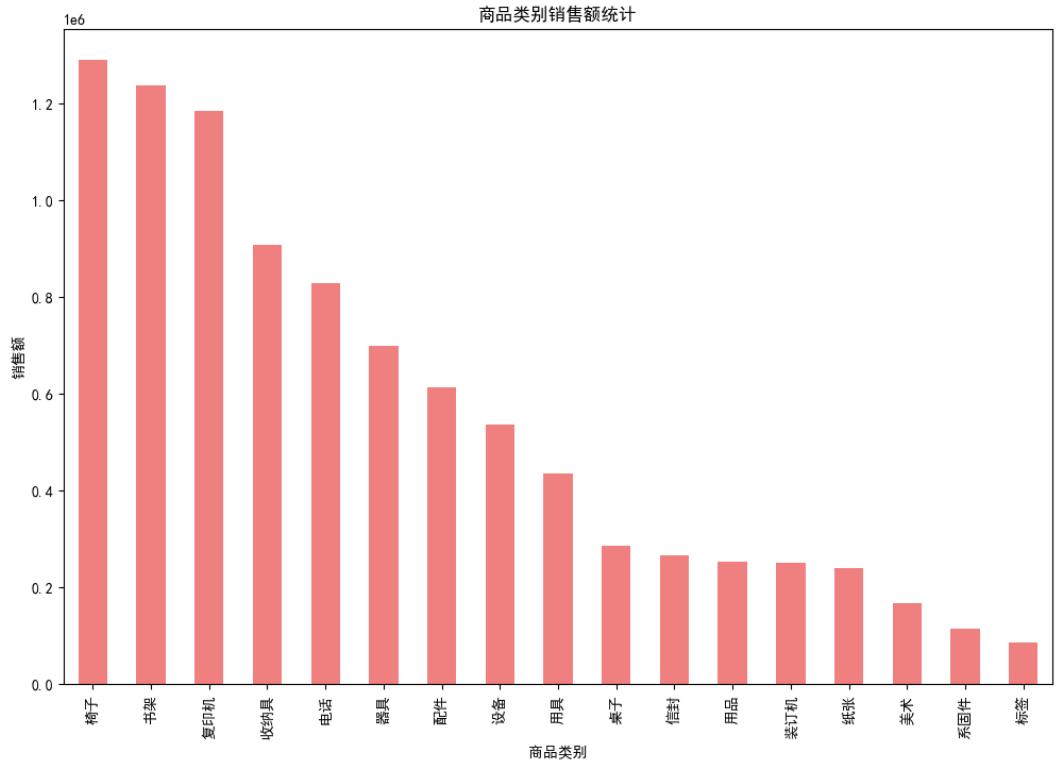


如图，发现聚类的簇为三。其中销售额较高和利润较高的一类是超市最期待的商品，这类商品的占比高于销售额高而利润为负的商品，说明总体上该超市的热销商品利润较高，该超市的销售商品选择比较合理。

3.2.2 商品类别销售额统计

通过统计每个商品类别的销售额，超市可以了解到哪些商品类别在市场上具有更高的销售额。这有助于发现畅销商品，从而调整库存策略、优化采购计

划，确保常销商品的充足供应。



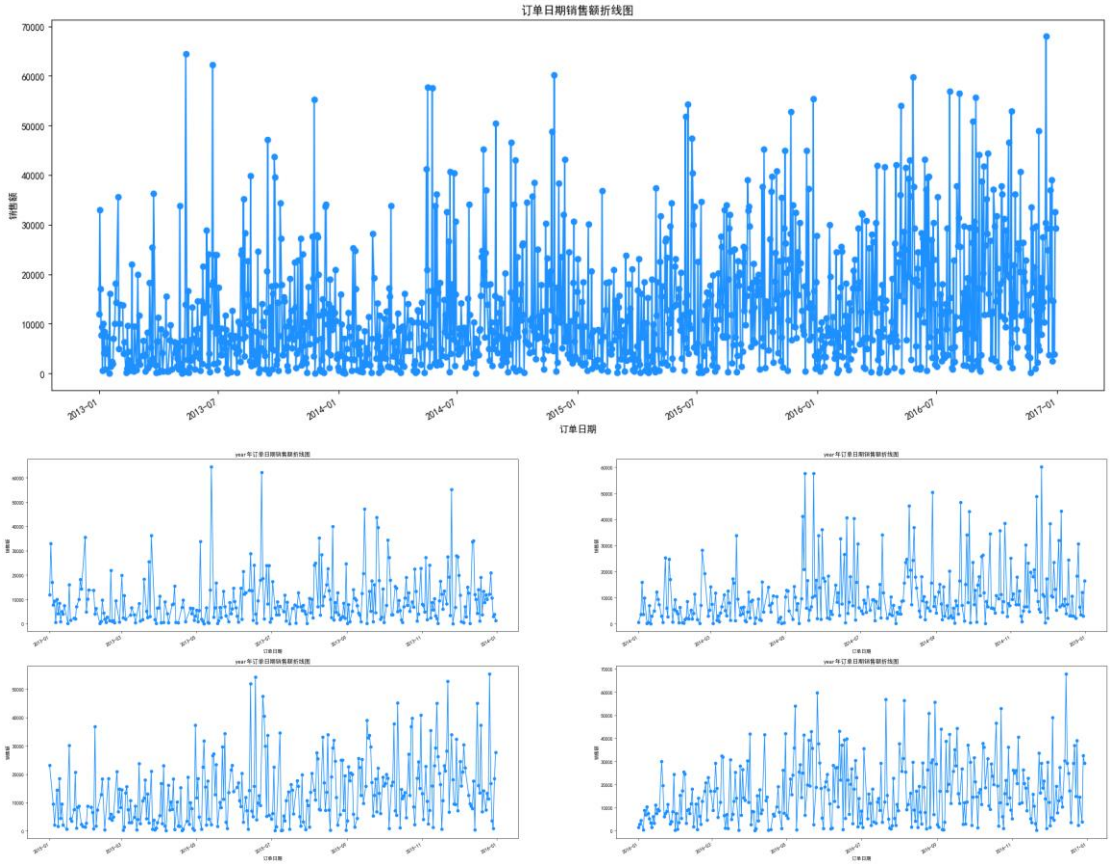
由上图可以发现，不同种类的商品销售额的差别极大，在该超市中椅子、书架、复印机的销售额与其他商品相比有明显优势，分别为 129 万元、123 万元、128 万元。而美术、系固件、标签的销售额较低，分别为 17 万元、11 万元、9 万元。

3.3 探究问题三：销量主要影响

研究销量的主要影响因素对于企业制定销售策略、优化供应链、提高销售效益具有重要意义。通过了解销量的主要影响因素，企业可以更加精准地制定促销活动、库存管理策略，以及优化产品供应和定价策略。

3.3.1 订单日期影响分析

分析订单日期可以帮助企业了解销售在不同季节、月份或节假日之间的变化。在销售额高峰期和低谷期，制造商和销售商等都可有针对性地制定策略，更有针对性地进行市场营销活动。例如，在销售淡季时可以推出促销活动来刺激销售，而在销售旺季时可以加强库存和物流准备。



由总订单日期销售额折线图和各年的订单日期销售额图可以发现，各年的订单日期与销量的走势是相对一致的，几个销售额的小高峰主要集中在 6、7 月份的暑假期间，9 月份的开学季还有 12、1 月份的寒假期间。

3.3.2 制造商和折扣对销量的影响

折扣是促使消费者购买的一种常见手段。分析折扣对销量的影响可以揭示在不同折扣条件下，消费者对购物的反应。企业可以通过制定灵活的折扣政策，调整折扣力度和时机，以达到最大化销售的目的。本文通过聚类分析来研

究制造商和折扣对销量的影响：

如图分析可以得到，折扣对于销售额的影响不是绝对的，聚类的效果不是很明显，但是可以发现多个制造商的高销售额的商品折扣为 0。

4. 探究问题四：销售趋势预测建模

销售趋势预测建模在零售业务中具有重要的作用，通过销售趋势预测，企业可以更准确地了解不同产品的销售走势，帮助企业合理制定库存策略，更好地应对市场波动。这样可以避免过量库存导致的资金浪费，也可以确保有足够的库存以满足潜在的需求。

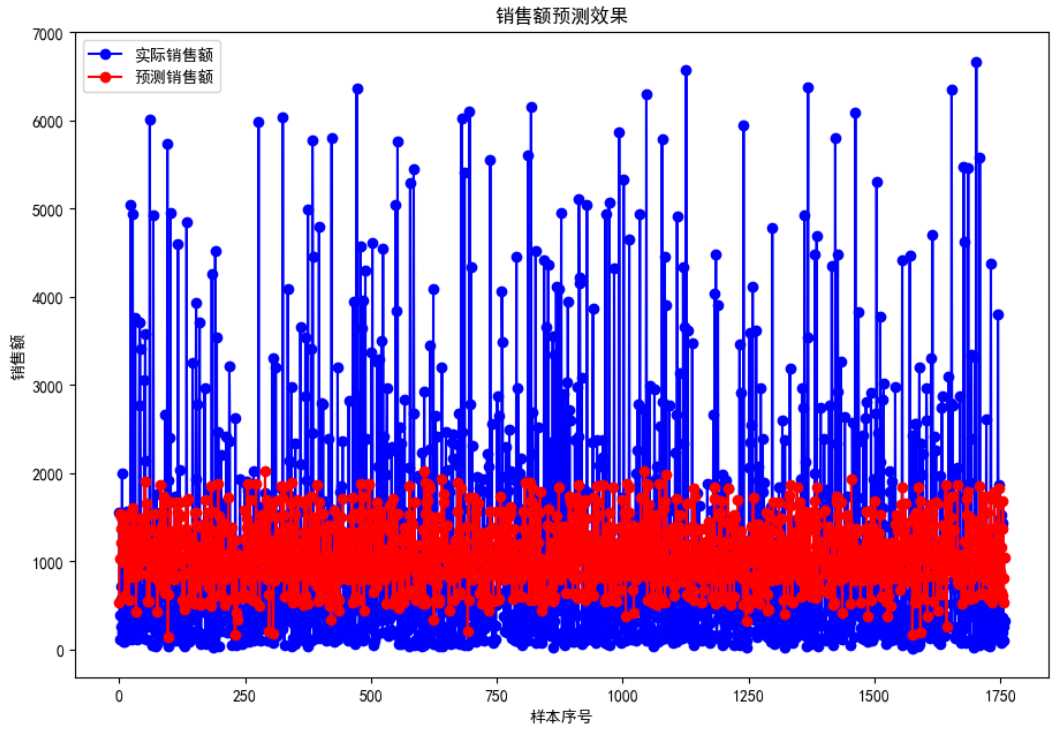
4.1 神经网络预测

神经网络是一种强大的非线性模型，能够捕捉和建模复杂的非线性关系。在销售数据中，各种影响销售的因素可能存在复杂的相互作用，神经网络能够更好地适应这种复杂性。销售趋势通常是具有时间序列结构的数据，包含随时间变化的模式。神经网络能够学习并捕捉这些时间序列模式，从而更准确地预

测未来的销售趋势。神经网络在处理大规模数据时表现优异。对于零售业务来说，可能涉及大量产品、客户和交易记录，神经网络可以更好地处理这种大规模、高维度的数据。本文使用的神经网络训练代码如下：

```
# 神经网络预测
from sklearn.neural_network import MLPRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

# 准备数据
X = data[['利润率', '折扣', '数量']]
y = data['销售额']

# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# 构建神经网络模型
model = MLPRegressor(hidden_layer_sizes=(100,), max_iter=1000,
random_state=42)

# 拟合模型
model.fit(X_train, y_train)

# 预测销售额
y_pred = model.predict(X_test)

# 评估模型性能
mse = mean_squared_error(y_test, y_pred)
print("均方误差(MSE):", mse)

# 可视化实际销售额和预测销售额
plt.figure(figsize=(12, 8))
plt.plot(y_test.values, label='实际销售额', color='blue',
marker='o')
plt.plot(y_pred, label='预测销售额', color='red', marker='o')
plt.title('销售额预测效果')
plt.xlabel('样本序号')
plt.ylabel('销售额')
plt.legend()
plt.show()
```

神经网络的均方误差(MSE):为 1424507.10，预测效果的可视化图像如下：



4.2 决策树预测

决策树在销售趋势预测中具有一些优势。决策树模型易于理解和解释，它们生成一系列简单的规则，描述了不同特征如何影响预测结果。对于业务决策者和非技术人员来说，了解决策树模型的规则更为直观。决策树能够自适应地捕捉数据中的非线性关系，而无需事先对数据进行复杂的变换。这对于销售数据中可能存在的非线性关系是一个优势。决策树能够处理混合类型的特征，包括数值型和分类型特征。这在销售数据中很常见，因为可能同时存在不同类型的特征。

但是决策树也有一些限制，例如对噪声敏感、容易过拟合等。在本文中，同时使用了集成方法——随机森林来弥补这些缺陷。

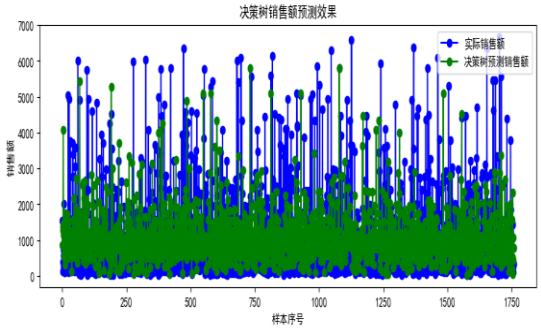
训练模型的代码如下：

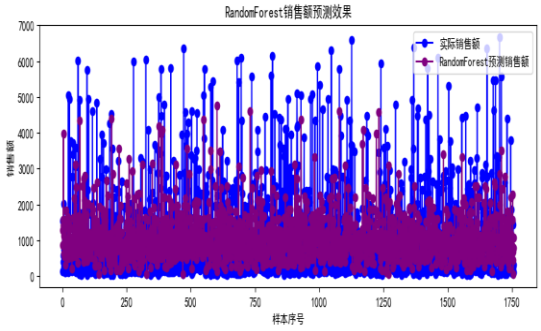
```
# 决策树
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

# 准备数据
X = data[['利润率', '折扣', '数量']]
y = data['销售额']
```

```
# 划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# 决策树模型
decision_tree_model = DecisionTreeRegressor(random_state=42)
decision_tree_model.fit(X_train, y_train)

# RandomForest 模型
random_forest_model = RandomForestRegressor(n_estimators=100,
random_state=42)
random_forest_model.fit(X_train, y_train)

# 预测销售额
y_pred_tree = decision_tree_model.predict(X_test)
y_pred_forest = random_forest_model.predict(X_test)

# 评估模型性能
mse_tree = mean_squared_error(y_test, y_pred_tree)
mse_forest = mean_squared_error(y_test, y_pred_forest)

print("决策树均方误差(MSE):", mse_tree)
print("RandomForest 均方误差(MSE):", mse_forest)

# 可视化实际销售额和预测销售额（决策树）
plt.figure(figsize=(12, 8))
plt.plot(y_test.values, label='实际销售额', color='blue',
marker='o')
plt.plot(y_pred_tree, label='决策树预测销售额', color='green',
marker='o')
plt.title('决策树销售额预测效果')
plt.xlabel('样本序号')
plt.ylabel('销售额')
plt.legend()
plt.show()

# 可视化实际销售额和预测销售额（RandomForest）
plt.figure(figsize=(12, 8))
plt.plot(y_test.values, label='实际销售额', color='blue',
marker='o')
plt.plot(y_pred_forest, label='RandomForest 预测销售额',
color='purple', marker='o')
plt.title('RandomForest 销售额预测效果')
plt.xlabel('样本序号')
```

```
plt.ylabel('销售额')
plt.legend()
plt.show()
```

在模型的预测结果上，决策树均方误差为 1527053.00，而随机森林的均方误差:为 1480370.02。预测结果的可视化图形如下：





4.3 线性回归模型

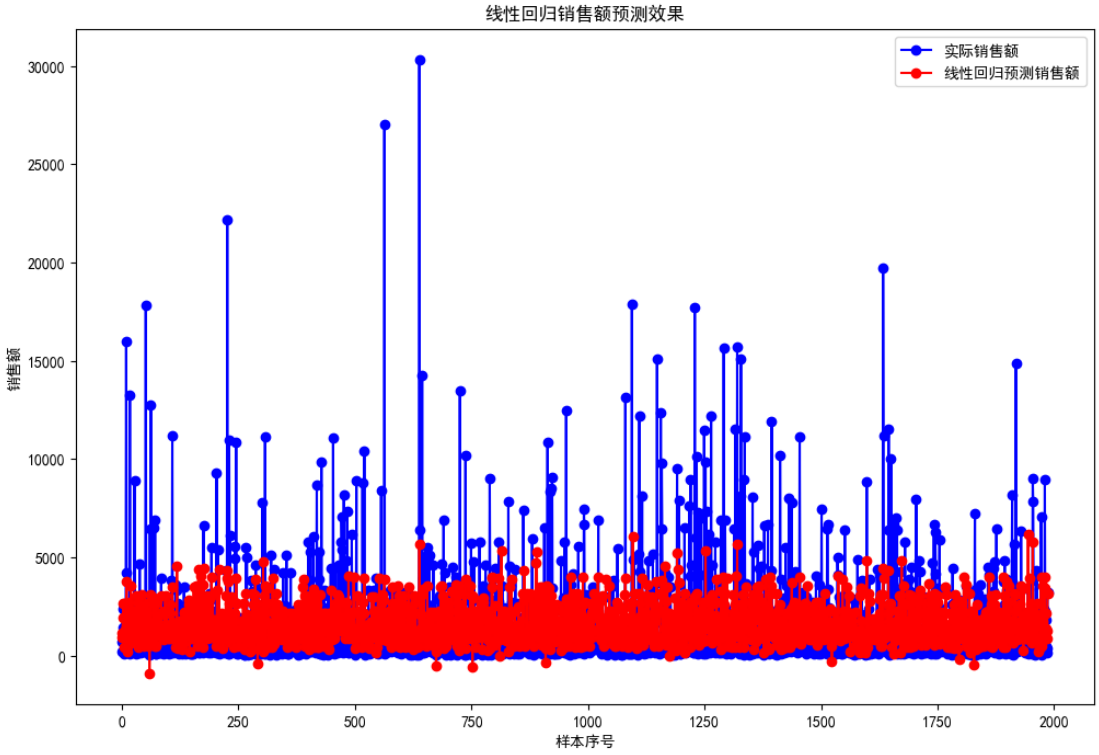
线性回归是一种简单而直观的模型，易于理解和解释。它建立了自变量和因变量之间的线性关系，通过斜率和截距来描述这种关系。同时，线性回归模型的参数直接反映了自变量对因变量的影响程度。通过检查模型的系数，可以了解不同特征对销售趋势的贡献，线性回归模型可以用于特征选择，帮助识别对销售趋势影响较大的特征。

本文使用的线性回归代码如下：

```
from sklearn.linear_model import LinearRegression

# 线性回归模型
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)

# 预测销售额
y_pred_linear = linear_model.predict(X_test)

# 评估模型性能
mse_linear = mean_squared_error(y_test, y_pred_linear)

print("线性回归均方误差(MSE):", mse_linear)
# 可视化实际销售额和预测销售额（线性回归）
plt.figure(figsize=(12, 8))
plt.plot(y_test.values, label='实际销售额', color='blue',
marker='o')
```

```
plt.plot(y_pred_linear, label='线性回归预测销售额', color='red',
marker='o')
plt.title('线性回归销售额预测效果')
plt.xlabel('样本序号')
plt.ylabel('销售额')
plt.legend()
plt.show()
```

模型的预测均方误差为 1435280.71，而预测效果的可视化图形如下：



5. 模型评估

模型评估通过一系列指标和度量来衡量模型在解决特定问题上的效果。在机器学习的过程中，有多种不同的模型和算法可供选择。通过模型评估，可以比较不同模型在同一任务上的表现，选择最适合问题的模型，有助于优化整个机器学习流程。

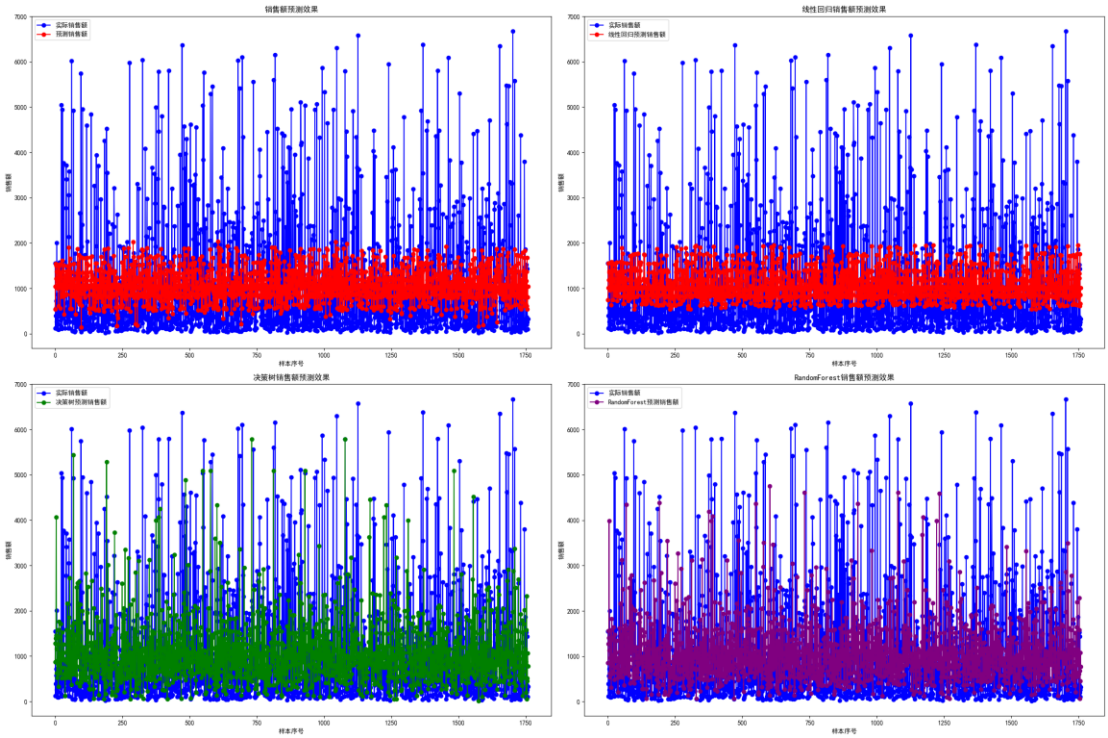
通过模型评估还可以帮助发现潜在问题，如过拟合（模型在训练集上表现好，但在测试集上表现差）或欠拟合（模型无法很好地拟合训练数据）。及早发现并解决这些问题有助于提高模型的稳定性和可靠性。

本文首先汇总了模型在测试数据集上的均方误差：

模型	神经网络	决策树	随机森林	线性回归
均方误差（MSE）	1424507.1041	1672177.2651	1617333.0408	1435280.7138

可以发现，神经网络和线性回归的模型均方误差较小，在 140 万左右，模型的表现较好，而决策树和随机森林模型的均方误差相对较大，在 160 万到 170 万之间，相对来说模型表现较差。但是注意到本次建模使用的数据集样本量并不大，提供的信息相对有限，所以几种模型的结果总的来说是接近的。

可视化预测结果的汇总如下：



6. 模型优化

根据分析可以发现，虽然决策树与随机森林的预测结果均方误差较大，但是在趋势上和真实数据更为接近，在一些数据峰值上和真实数据更为接近。基于此，本文考虑进一步优化随机森林的参数，从而既能有较好的趋势拟合，又能获取较小的均方误差。

参数网格搜索（Grid Search）是一种通过遍历给定的参数组合来优化模型性能的方法。本文采用参数格搜索的办法来优化随机森林的参数，减少误差，具体步骤为：

1. **定义参数网格：** 针对模型的每个超参数，定义一组可能的取值。这形成了一个参数网格，每个点都是一个超参数组合。
2. **遍历参数组合：** 遍历所有可能的参数组合，对每一组参数进行模型训练和评估。
3. **选择最佳参数组合：** 通过比较模型性能（通常使用交叉验证得到的评估指标）选择最佳的参数组合。

在 scikit-learn 等机器学习库中，通常提供了 GridSearchCV 类来简化参数网格搜索的实现。本文使用了 scikit-learn 的机器学习库，并通过 intel 的加速来缩短模型训练时间，具体代码如下：

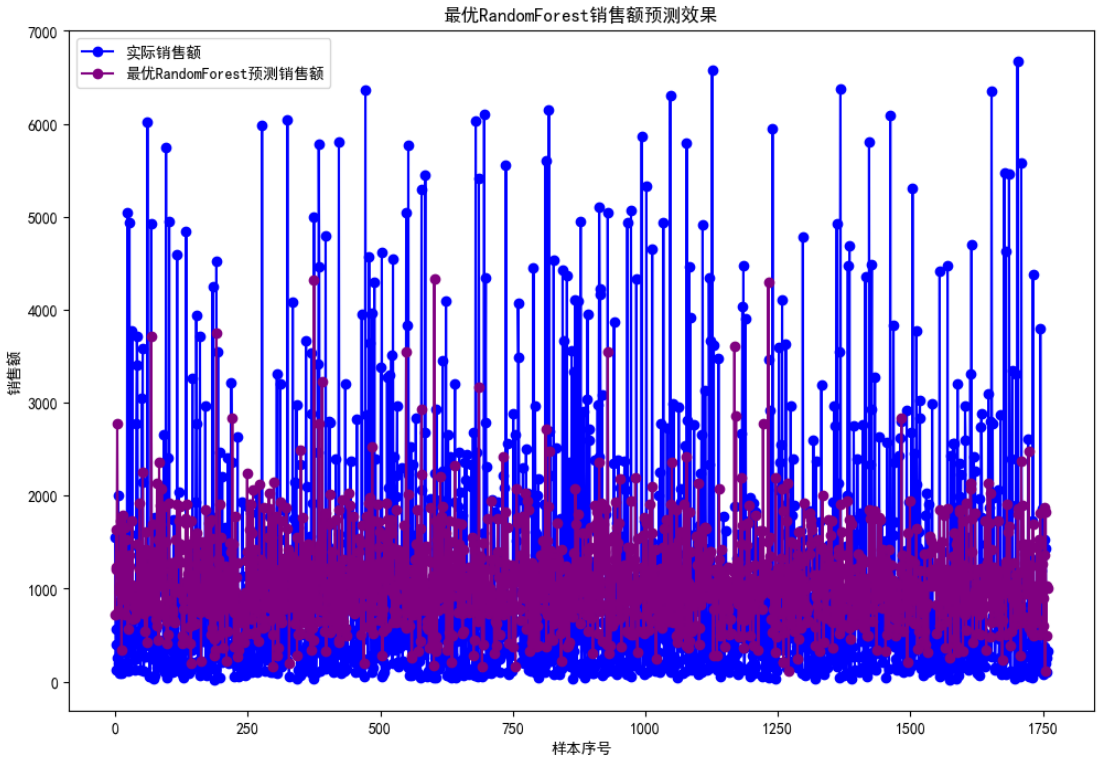
```
from sklearn.model_selection import GridSearchCV
from sklearnex import patch_sklearn
# 应用 Intel 的 scikit-learn 扩展
patch_sklearn()
# 定义参数网格
param_grid = {
    'n_estimators': [50, 100, 150],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}

# 创建 RandomForest 模型
random_forest_model = RandomForestRegressor(random_state=42)

# 创建 GridSearchCV 对象
grid_search = GridSearchCV(random_forest_model, param_grid, cv=5,
scoring='neg_mean_squared_error', n_jobs=-1)

# 在训练数据上进行网格搜索
grid_search.fit(X_train, y_train)

# 输出最优参数
print("最优参数:", grid_search.best_params_)

# 使用最优参数的模型进行预测
best_forest_model = grid_search.best_estimator_
y_pred_best_forest = best_forest_model.predict(X_test)

# 评估最优模型性能
mse_best_forest = mean_squared_error(y_test, y_pred_best_forest)
print("最优 RandomForest 均方误差(MSE):", mse_best_forest)
plt.figure(figsize=(12, 8))
```

```
# 可视化实际销售额和预测销售额（最优 RandomForest）
plt.plot(y_test.values, label='实际销售额', color='blue',
marker='o')
plt.plot(y_pred_best_forest, label='最优 RandomForest 预测销售额',
color='purple', marker='o')
plt.title('最优 RandomForest 销售额预测效果')
plt.xlabel('样本序号')
plt.ylabel('销售额')
plt.legend()
plt.show()
```

结果如下：

```
最优参数: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 100}
最优RandomForest均方误差(MSE): 1369267.062665363
```

可视化预测结果如图：



于是，通过模型的评估，我们选择随机森林模型，并且为了减小误差，使用了格搜索优化，获取了最佳参数。优化后的随机森林模型无论在预测均方误差上，还是可视化结果的趋势上，都达到了最好的效果。

7. 总结

在现代商业环境中，零售业务数据的分析和预测对于制定有效的业务策略至

关重要。本文基于提供的超市数据，采用数据可视化、决策树、神经网络、聚类、回归分析等多种机器学习算法，深入分析了零售业务的特征、商品类别畅销特点、销售量主要影响因素以及未来销售趋势，并对比了不同模型的预测效果，优化出了当前问题下性能更佳的模型。

针对分析结果，本文也相对应提出了业务决策上的建议，并对分析方法的合理性和可行性进行了探讨。通过这些深入分析，企业可以更好地了解市场需求，优化商品管理，提高销售效益。