基于统计分析及LSTM模型的产品订单需求量分析及预测

摘　要

当前，企业的供应链面临着严峻的挑战。经济环境的变化、市场行情等因素，都会对产品的需求量产生影响，而供应链的有效运作则需要准确的需求预测作为支撑。因此，为保证企业健康发展，能够准确预测产品的需求量已经变得尤为重要。只有通过精准的需求预测，企业才能够更好地规划采购计划、制定生产计划，以确保供应链的高效运转，应对市场的变化和挑战。本文基于对历史数据的统计分析，深度剖析历史订单需求量的内在统计学特征，并建立LSTM预测模型对企业未来3个月的需求量进行预测，以期为企业的正常运转提供数据支撑。

针对问题一，首先对历史订单数据进行数据清洗，进行列名替换、查找缺失值、处理重复值并对日期数据进行深入挖掘，包括提取工作日、节假日、促销日等信息，以对后续分析做铺垫。接着，通过绘制产品价格和订单需求量的散点图矩阵，发现两者呈现L形分布，斯皮尔曼相关系数为-0.29，并不存在显著的线性相关性，于是计算价格数据的分位点并将价格数据离散化，绘制价格区间-需求量的柱状-折线组合统计图，发现在中低价格区间内，订单需求量随着价格的升高而上升，在高价格范围内，订单需求量随着价格的升高而降低，为了更加准确地将价格进行分类，基于价格和订单需求量两个特征，采用k-means算法将价格分为3个档次，分别对应低价格、中价格和高价格，依据聚类结果，进一步进行统计分析。由于除价格外的其他变量均为分类变量，因此从区域、销售方式、时间段、节假日、促销日、季节等角度，应用统计学等基本方法对订单需求量进行组合统计分析，得出订单需求量受到各个因素的影响情况，并对数据做出合理解释。

针对问题二，首先对待预测数据进行分析，发现待预测数据特征均为分类变量，因此采用深度学习模型对订单需求量进行预测分析，由于训练数据存在很强烈的时间连续性，故采用LSTM模型对后续订单的需求预测为合理方法。遍历每个待预测数据，提取出相同的训练数据作为训练集和测试集，通过对训练数据和预测数据的表格连接等操作，发现部分待预测数据不包含于训练数据中，对于这类数据，采用相似的数据做为对该数据模型建立的训练集，通过不断的调整训练参数，最终确定LSTM层单元数为100，迭代次数为300，滑动窗口数量为5时模型精度较高，本文以产品编码20002，销售区域101，产品大类编码303，产品细类编码405为例，按照日粒度说明了模型训练预测的全过程，最终得出测试集的RMSE为3.284，说明模型具有较高精度，且具有鲁棒性，接着按照周粒度、月粒度进行同样的操作，对比每种方法的预测精度，并将结果整合到result1中。

本文由数据驱动，分别运用统计分析、LSTM等深度学习模型对企业的历史订单数据进行深入的挖掘、探究数据的潜在关系，按照天、周、月的粒度对2019年1、2、3月的订单需求量进行预测，通过准确度评估验证了模型的有效性及预测的准确性，该分析和预测结果能够应用于企业的订单需求预测中，并具有良好的推广性。

关键词：订单需求量；统计分析；数据挖掘；LSTM；k-means

目录

一、 绪论 3
1.1 背景 3
1.2 问题重述 3
1.3 研究意义 3
二、 数据描述和预处理 3
2.1 数据来源和特点 3
2.2 数据预处理 4
2.2.1 对缺失值的处理 4
2.2.2 对重复值的处理 4
2.2.3 数据扩展 4
三、 对训练数据进行深入挖掘 4
3.1 产品不同价格对需求量的影响 4
3.1.1 对整体分布的分析 4
3.1.2 相关性分析 5
3.1.3 整体数据统计量 6
3.1.4 对分位点的分析 6
3.1.5 聚类分析 7
3.2 区域对需求量的影响 9
3.2.1 单因素方差分析 9
3.2.2 分组统计 10
3.3 不同销售渠道的产品需求特性 12
3.4 不同品类之间的需求量 13
3.5 不同时段的产品需求量特性 14
3.6 节假日与促销日对需求量的影响 14
3.7 季节因素对需求量的影响 15
四、 基于 LSTM 的需求量的预测 16
4.1 问题分析 16
4.2 LSTM 神经网络 16
4.3 问题求解 17
4.3.1 数据预处理 17
4.3.2 模型训练 19
4.4 预测 20
4.5 精度比较 20
五、 总结 21
六、 团队管理 21

一、 绪论

1.1 背景

在现代商业环境中，准确地预测产品订单的需求变化对企业的运营和决策具有关键的重要性。通过深入分析产品订单的数据，企业可以更好地了解市场趋势、顾客需求和竞争动态，从而制定有效的生产计划、库存管理策略和市场营销活动。

1.2 问题重述

在这项研究中，我们将专注于产品订单的数据分析和需求预测问题。具体来说，我们的目标是利用可用的订单数据集，通过应用数据分析和预测建模技术，解决以下问题：

市场趋势分析：如何通过分析历史订单数据来识别和理解市场趋势，包括季节性变化、渠道变化以及区域变化等，以便更好地调整企业的生产和供应链策略。

需求预测：我们将研究和开发基于历史订单数据的预测模型，以帮助企业合理规划生产计划、库存管理和资源分配。

通过深入研究以上问题，我们可以提供给企业有关产品订单数据分析和需求预测的实用见解和决策支持，帮助企业提高运营效率、降低成本，并满足客户需求的变化。

1.3 研究意义

本研究对企业产品的生产分配以及效率提升具有重要的实践和理论意义。

1、提高企业运营效率：通过深入分析产品订单数据，企业可以更好地了解市场趋势和顾客需求，从而制定准确的生产计划、库存管理策略和资源分配。通过精确的需求预测，企业能够避免过量库存和缺货情况，提高生产效率，降低成本，并最大程度地满足客户需求。

2、优化供应链管理：产品订单数据的分析和需求预测有助于优化供应链管理。通过准确预测需求，企业可以与供应商建立更有效的合作关系，提前调整采购计划，减少库存积压和缺货风险。优化供应链管理将提高整体运作效率，缩短供应链的响应时间，并增强企业在市场中的竞争力。

二、 数据描述和预处理

2.1 数据来源和特点

数据集包含 order_train1.csv 和 predict_sku1.csv 两个文件。

其中 order_train1.csv 文件提供了某制造企业在 2015 年 9 月 1 日至 2018 年 12 月 20 日的出货数据，其中包含一下信息：

1、订单日期（order_date）
2、销售区域编码（sales_region_code）
3、产品编码（item_code）
4、产品大类编码（first_cate_code）：与产品细类编码是一对多关系
5、产品细类编码（second_cate_code）
6、销售渠道名称（sales_chan_name）：分为线上（online）和线下（offline）
7、产品价格（item_price）
8、订单需求量（ord_qty）

predict_sku1.csv 文件中提供了需要预测产品的相关信息，包含以下字段：

1、销售区域编码（sales_region_code）
2、产品编码（item_code）

3、产品大类编码（first_cate_code）

4、产品细类编码（second_cate_code）

2.2 数据预处理

2.2.1 对缺失值的处理

首先将 order_train1.csv 导入，利用 isnull()方法检测表单中存在的缺失值，发现表单中并不存在缺失值，故不进行缺失值处理的工作。

2.2.2 对重复值的处理

利用 duplicated()方法检测表单中的重复值，发现表单中存在 312 条有重复值的数据，由于 order_train1 表单数据量庞大，我们认为重复值是人为记录所造成的，并且重复值体量占比甚微，对数据整体分布的影响可忽略，故利用 drop_duplicates()将其去除，去除重复值的表单有 597382 条数据。

2.2.3 数据扩展

由于表单中只给出了订单日期、销售区域等基本信息，这些信息不足以支持对后续数据的分析，因此我们对订单日期数据进一步进行挖掘，并依据订单日期扩展了年、月、日、是否是促销日、季节、星期、是否是工作日、上中下旬、是否是节假日等信息，扩展后的表格每个元素有 17 维特征，后续分析全部基于此表，示例数据如下表所示。

表 1 数据扩展后的数据示例

索引	订单日期	销售区域编码	产品编码	产品大类编码	产品细类编码	销售渠道名称	产品价格	订单需求量	年	月	日
0	2015-09-01	104	22069	307	403	1	1114.0	19	2015	9	1
1	2015-09-01	104	20028	301	405	1	1012.0	12	2015	9	1
2	2015-09-02	104	21183	307	403	0	428.0	109	2015	9	2
3	2015-09-02	104	20448	308	404	0	962.0	3	2015	9	2
4	2015-09-02	104	21565	307	403	1	1400.0	3	2015	9	2

表 1 续表

索引	是否促销日	季节	星期	是否工作日	上中下旬	是否节假日
0	0	3	2	1	1	0
1	0	3	2	1	1	0
2	0	3	3	1	1	0
3	0	3	3	1	1	0
4	0	3	3	1	1	0

三、 对训练数据进行深入挖掘

3.1 产品不同价格对需求量的影响

3.1.1 对整体分布的分析

首先绘制价格和需求量的散点图矩阵，直观感受价格和需求量的关系，发现两个变量的分布在图像中呈现近似 L 型分布，从直方图中可以看出大部分产品价格处于中低价位，且需求量大，只有少量数据处于高价位且需求量很少。

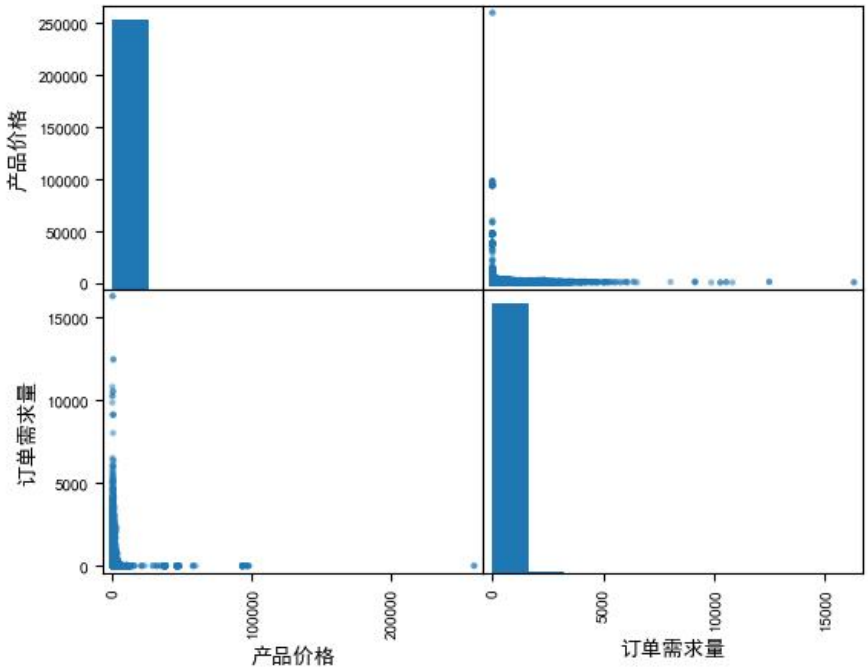


图 1 产品价格和订单需求量关系散点图

按照产品价格分组，并计算需求量平均值和总和，发现有 14365 种价格，和整体分布相似，该分布也呈现近似 L 型分布，如下图所示。

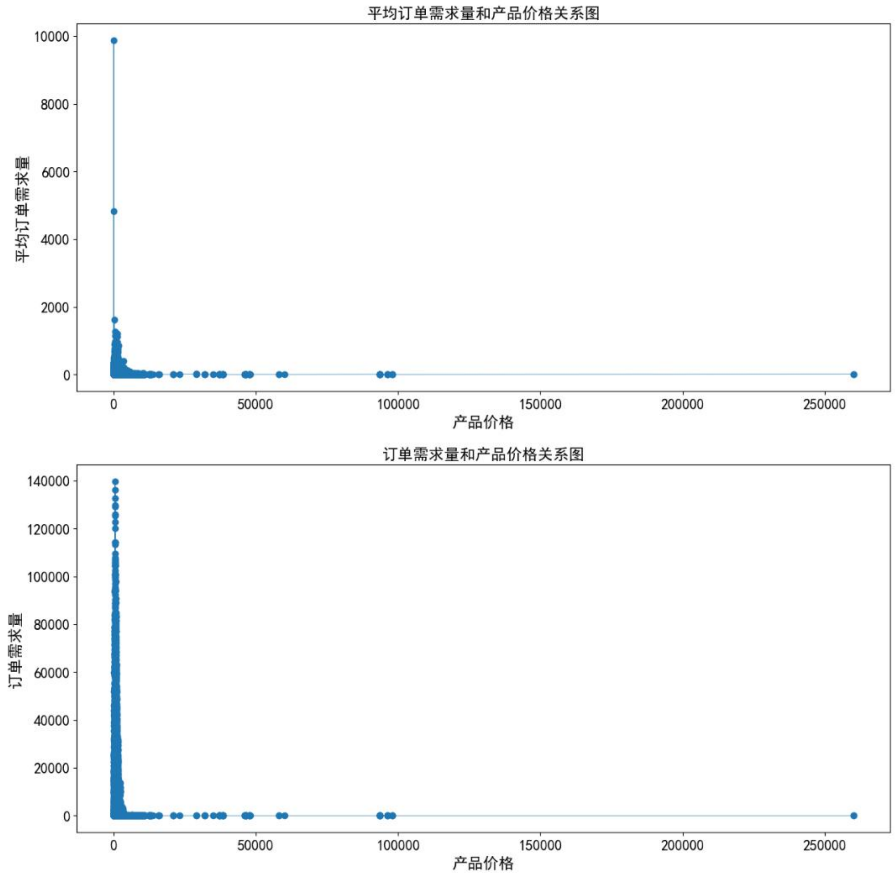


图 2 按照价格分组后的产品价格-需求量关系图

3.1.2 相关性分析

斯皮尔曼相关性分析是一种非参数方法，用于评估两个变量之间的单调关系。它使

用等级而不是原始数据值，因此对于不满足正态分布的数据也是适用的。斯皮尔曼相关性的取值范围在-1 到 1 之间，其中-1 表示完全逆向的单调关系，1 表示完全正向的单调关系，0 表示无单调关系。斯皮尔曼相关性可以通过计算等级之间的差异来确定。

产品价格和需求量均为连续变量，且根据直方图，订单需求量和产品价格均不呈现正态分布，因此用斯皮尔曼相关系数来探究两者之间的相关性为合理方法，两者的相关性矩阵热图如下图所示。

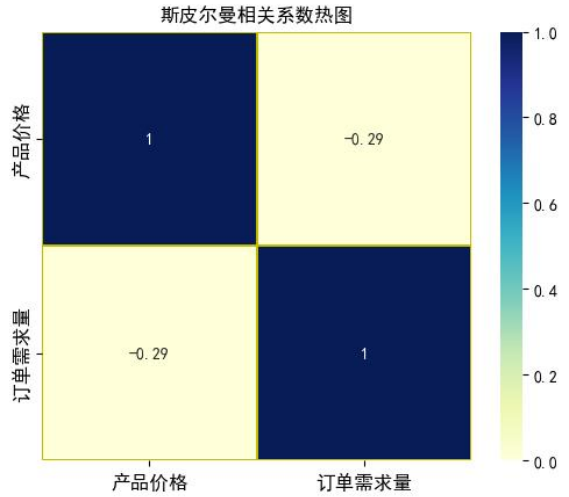


图 3 产品价格和订单需求量斯皮尔曼相关系数热图

通过计算得知两者的相关系数为-0.29 且 p 值为 0(<0.05)说明两者呈现显著的弱负相关关系，进一步说明了产品价格越高，需求量越低的特点，因此对数据进行基于统计的建模分析较为合理。

3.1.3 整体数据统计量

首先，利用 describe()方法得出产品价格和订单需求量的基本统计量，如下表所示。

表 2 产品价格和订单需求量基本统计量

	count	mean	std	min	25%	50%	75%	max
产品价格	597382	1076.14	1167.57	1	598	883	1291	260014
订单需求量	597382	91.68	199.88	1	10	29	101	16308

从该表中可看出，对于产品价格，均值为 1076.14，标准差为 1167.57，最小值为 1，最大值为 260014，说明产品价格波动性较大，且区间跨度较大，从但从均值和分位点可看出只有少量产品与最高价格为同一数量级，大部分产品价格小于 1291。

对于订单需求量，均值为 91.68，标准差为 199.88，最小值为 1，最大值为 16308，订单需求量的波动性和价格区间均小于产品价格，说明产品价格的波动没有对订单需求量造成显著影响，且大部分订单需求量小于 101。

3.1.4 对分位点的分析

计算经过分组后产品价格的 10 个分位点，依据分位点将产品价格分成 10 个区间，计算每个区间的订单需求量如下图所示。

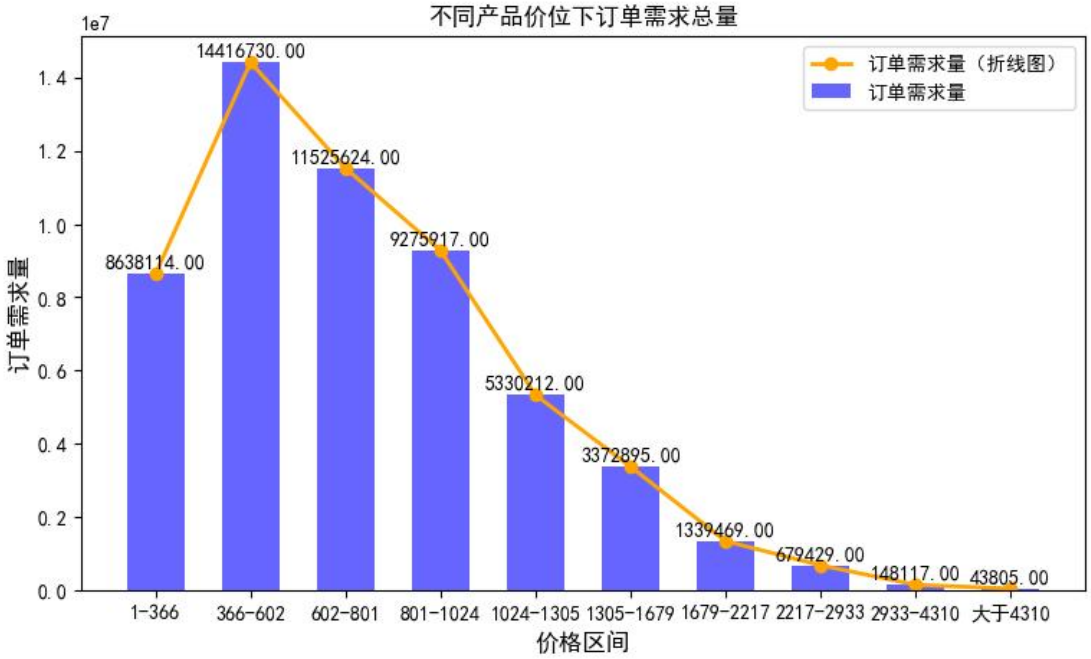


图 4 不同区间下订单需求总量统计图

可以看出在区间[1,602]，订单需求量随着价格的上升而增加，且在区间[366,602]上，订单需求总量达到峰值，相对于区间[1,366]，需求量显著增加，这可能是因为在低价格区间内，价格相对较高的产品能够弥补产品质量的不足，性价比较高，口碑较好。而当价格大于 602，随着价格的升高，订单需求量逐步减少，且在 1024 分位点处显著下降，推测该点是区分价格高低的临界点，该统计图呈现出的趋势符合市场规律。

3.1.5 聚类分析

聚类分析是一种无监督学习方法，用于将数据样本分组成具有相似特征的类别或簇。它的主要目标是发现数据中的内在模式、结构和关系，从而提供对数据的洞察和理解，基于上述统计学的分析，我们只能从宏观上窥探数据的内在联系，聚类分析能够帮助我们进一步挖掘数据，因此我们以产品价格和订单需求量为特征进行 k-means 聚类分析，将样本分成 3 类，聚类结果如下图所示。

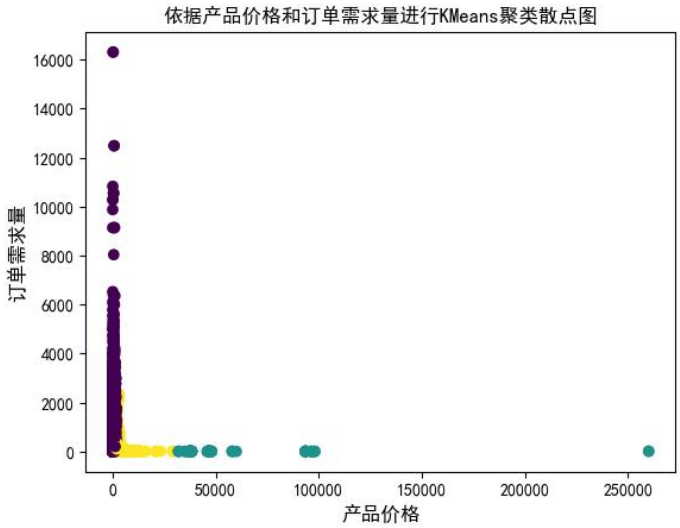
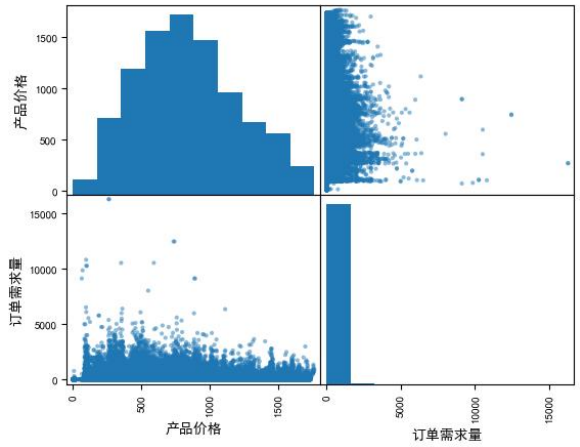
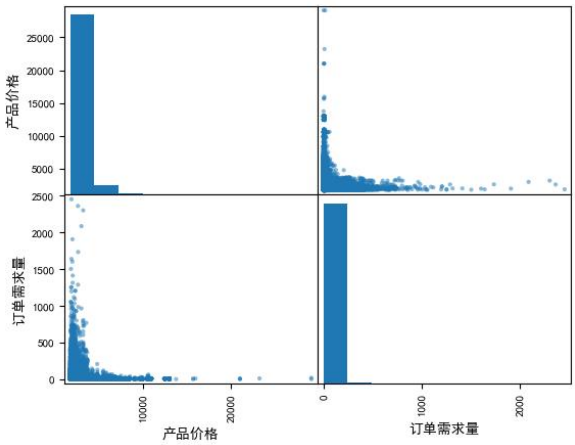


图 5 k-means 聚类结果图

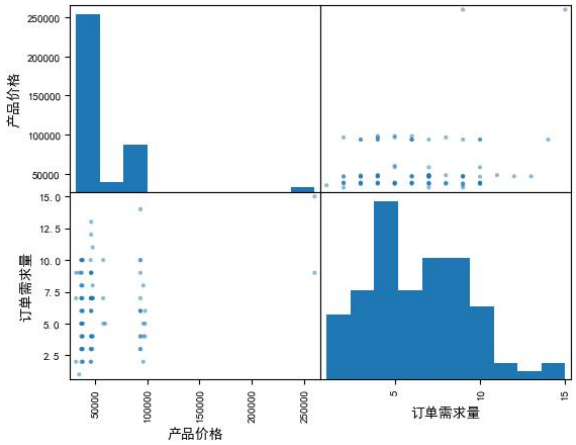
从聚类结果图中可以看出，基于价格可以将产品分为 3 类，分别为紫色的低价格，黄色中价格，以及绿的高价格，针对聚类结果绘制的散点图矩阵如图 6 所示。



(a)低价格散点图矩阵



(b)中价格散点图矩阵



(c)高价格散点图矩阵

图 6 不同价格散点图矩阵

此外，不同价格的基本统计值分析如表 3-5 所示。

表 3 低价格产品基本统计量

	count	mean	std	min	25%	50%	75%	max
产品价格	522642	842.99	373.57	1	553	809	1100	1768
订单需求量	522642	100.97	210.93	1	11	35	105	16308

表 4 中价格产品基本统计量

	count	mean	std	min	25%	50%	75%	max
产品价格	74644	2636.39	1157.79	1737.4	2013	2312	2752	29118
订单需求量	74644	26.77	58.29	1	7	11	24	2452

表 5 高价格产品基本统计量

	count	mean	std	min	25%	50%	75%	max
产品价格	96	57198.64	36831.86	32003	37219	46506	58016	260014
订单需求量	96	6.26	2.92	1	4	6	8.25	15

从表格和基本统计量可以看出，绝大部分产品处于低价位，部分产品处于中价位，只有少量产品处于高价位，且价位越高，需求量越少。

基于上表，将产品价格分为 3 个区间，分别为[1,1737]、[1737,29118]、[32003,260014]，再次绘制平均订单需求量，结果如下图所示。

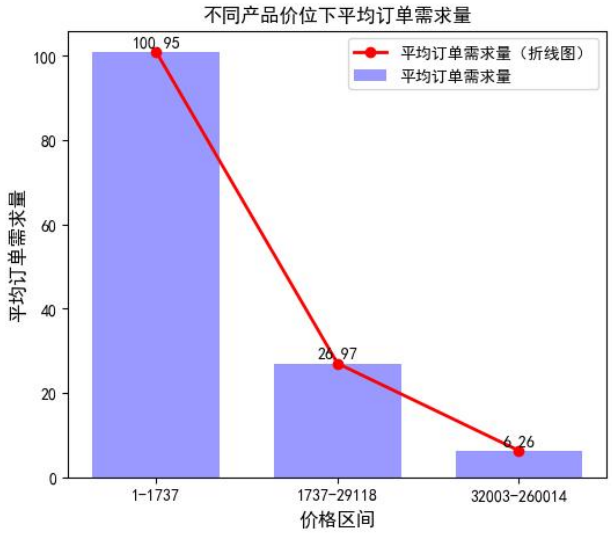


图 7 不同产品价位平均订单需求量统计图

从图中可以看出，随着价位的升高，平均订单需求量在逐步降低，低价位是产品的主要定价方向，低价位对客户的吸引度更强。

此外，通过绘制不同价位记录分布饼图和不同价位订单数量分布图可以更加直观的看出市场对于各价位层级的需求度。

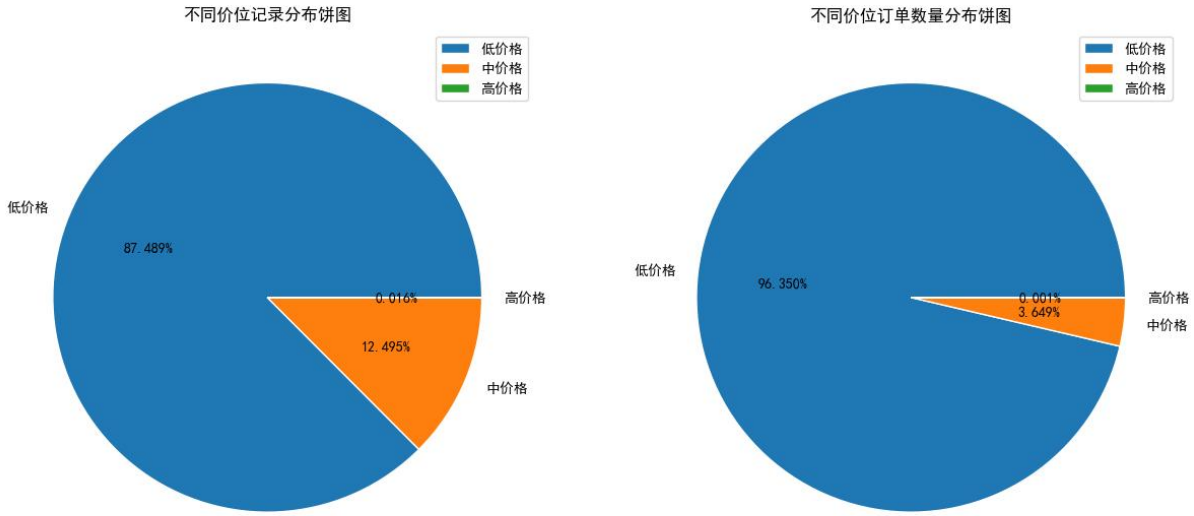


图 8 不同价位记录分布饼图

3.2 区域对需求量的影响

3.2.1 单因素方差分析

单因素方差分析（One-Way ANOVA）是一种用于比较多个组或条件之间平均数差异是否显著的统计方法。它适用于研究一个自变量对一个连续的因变量的影响，当有一个自变量有多个组别时，可以使用单因素方差分析，在该问题中，区域为分类变量，而需求量为连续变量，因此使用单因素方差分析来探究区域对需求量的影响是合理的。

经过对数据进行正态性检验得出 p 值为 0(<0.05)，因此需求量不满足正态分布，所

以对数据进行 Kruskal-Wallis 检验，得出 p 值为 0，说明区域对需求量有显著影响。

3.2.2 分组统计

将所有数据按照区域进行分组，并求出分组后订单需求量的所有统计值，如下表所示，各区域需求量总和占比及基本统计量柱状图如图 9、10 所示。

表 6 不同区域订单需求量基本统计值

销售区域编码	max	min	mean	sum	count	median	std
101	16301	1	98.54	12400949	125844	33	233.213446
102	5055	1	85.29	13966622	163738	33	154.03123
103	9140	1	99.19	11519878	116139	32	210.304888
104	9874	1	95.28	2387342	25055	34	190.206931
105	16308	1	87.00	14495521	166606	21	206.042987

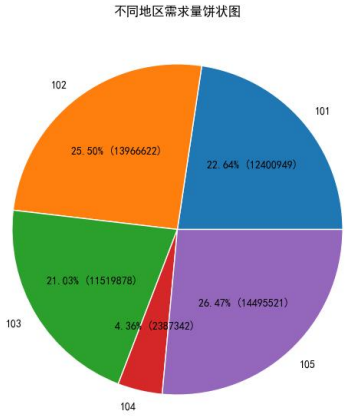


图 9 不同地区需求量饼状图

从饼状图中可以看出，101、102、103 和 105 地区的订单需求量在总订单需求量的占比均超过了 20%，其中 105 地区的订单需求量最高，为 26.47%，104 的订单需求量仅有 4.36%，说明 101、102、103 和 105 为该企业的主要售往地，应针对这些地点进行针对性的销售策略。

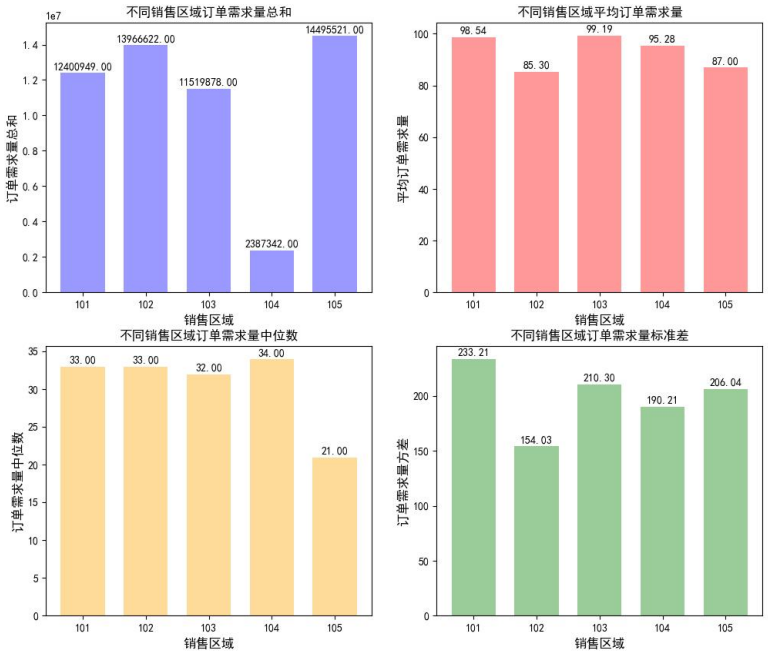


图 10 不同地区订单需求量基本统计量

从以上四幅柱状图中可以看出 103 地区的平均订单需求量最多，为 99.19 单，102 最少，为 85.30 单，各地区的平均订单需求量无明显差异，并不受订单总量的影响，对于中位数，105 地区为 21 单，其余地区均在 32 单及以上，方差代表了各地区订单需求量的波动程度，尽管 101 地区的订单需求量较大，但是其方差达到了 233.21，说明该地区对订单的需求量存在比较大波动，这可能是受到该地区的经济情况和市场行情影响，而方差最低的地区为 102，为 154.03，说明不仅 102 地区的订单需求量高，且较为稳定，因此可以将其列为重点关注区域。

此外，统计不同区域按照年份、产品编码、销售方式分类的需求量是有必要的，通过这些数据能够更深入地挖掘出区域之间的差距，其柱状图如图 11-13 所示。

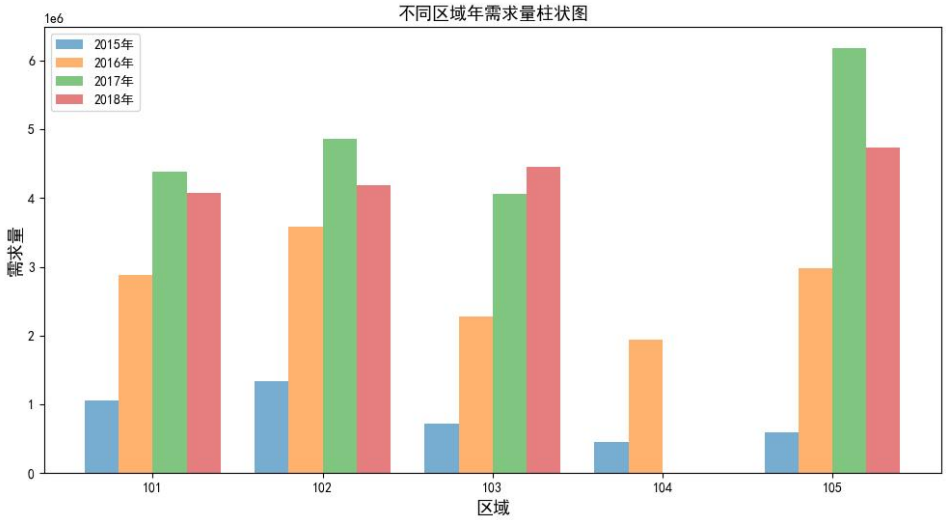


图 11 不同地区年需求量柱状图

从图中可以看出，2015 到 2017 年间，各个区域的需求量均逐年增长，说明市场行情，企业知名度在逐年提高，105 地区 2016 年到 2017 年需求量增长最为迅猛，说明产品备受 105 区域的青睐，但是 2017 和 2018 年，104 地区不再购买产品，说明 104 地区对企业的产品需求量下降或引进了更加优质且价格低廉的平替产品，企业应该加强产品在 104 地区的宣传力度或者增加创新点以吸引 104 地区的市场，2017 年到 2018 年，除 103 地区以外，各地区的订单需求量均出现下降趋势，且 105 地区的下降幅度最大，说明市场竞争非常激烈，企业应稳住 105 地区的优势，加强对其他地区的宣传力度，更应提高自己的产品质量，降低成本。

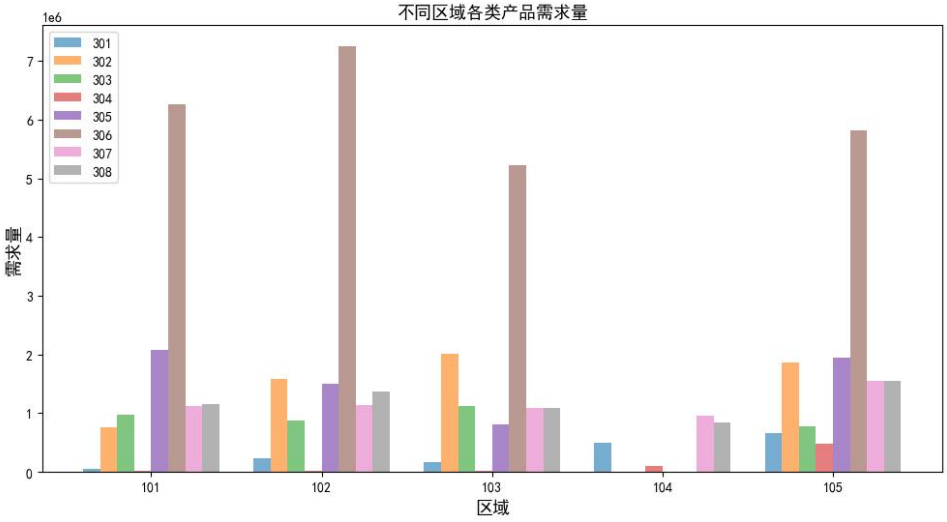


图 12 不同区域各类产品需求量柱状图

从图中可以看出，除了 104 地区以外，各个地区对 306 号产品的需求量最高，306 号产品在各个地区成为了明星产品，其次是 302 号 305 号产品，而 304 号产品在各个地区几乎没有订单，所以企业应该考虑减少对 304 号产品的投入，将更多的成本投入到 306 号产品，加强优势产品，扬长避短，以获得市场竞争高地。

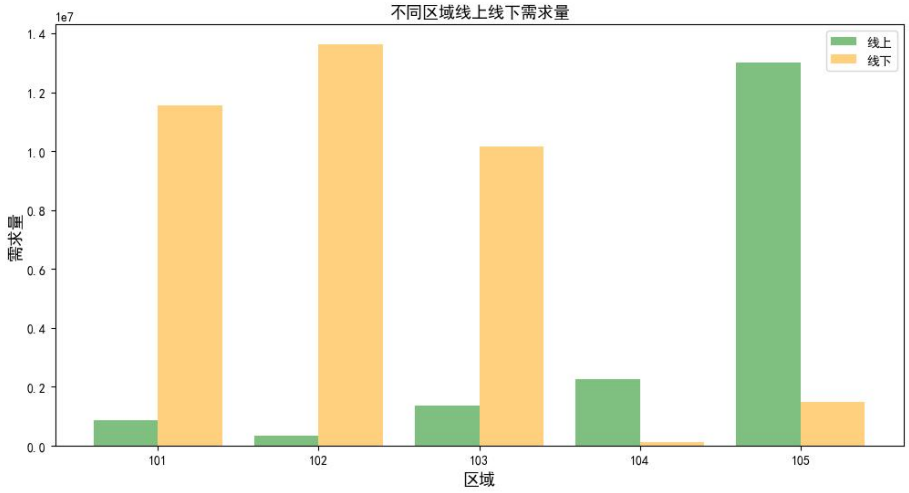


图 13 不同区域线上线下需求量柱状图

从图中可以看出，101、102 和 103 区域的大部分订单均为线下订单，而 104 和 105 的主要销售渠道为线上，所以针对 101、102、103 地区，在保证对这些地区线下优势的前提，应加强发展线上销售渠道，以获得更多利润，相反，对于 104 和 105 地区，在保证线上优势的同时深入挖掘线下销售渠道，拓宽更多的销售渠道能为企业带来更多的利润。

3.3 不同销售渠道的产品需求特性

针对不同的销售渠道（线上和线下），利用 groupby()将不同产品按照销售渠道进行分组，求平均值，结果如表 7 所示。

表 7 不同销售渠道年平均需求量

	2015	2016	2017	2018	总需求平均（万）
线上	155.9	102.7	116.0	112.6	112.5
线下	107.8	92.4	93.4	66.9	84.2

可以发现，线上的需求量平均多于线下，对于几年来商品的销售，线上渠道是更好的销售方式，并且从每年的变化趋势来看，线下平均需求量总体呈递减的趋势，所以不论哪种产品，发展线上销售是非常必要的，同时也不能忽视 2015 年的平均需求量大幅度超过后续年份的需求，下面的直方图可以更直观的展示。

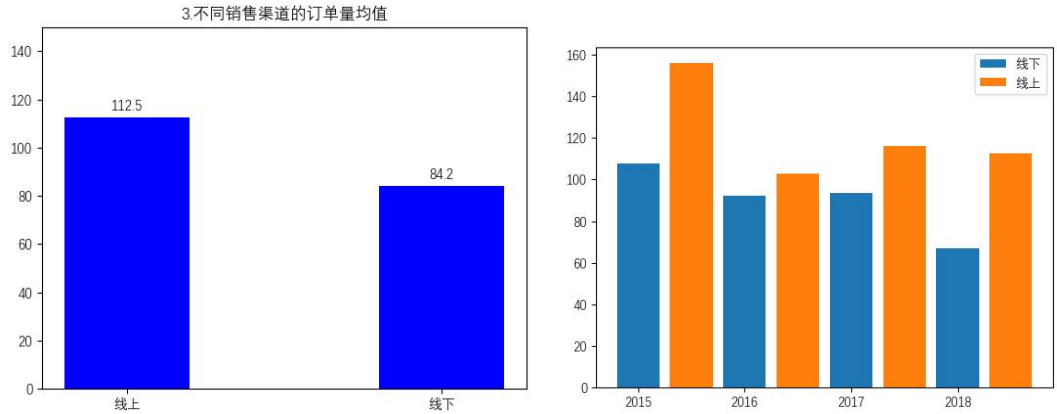


图 14 不同销售渠道订单量均值

3.4 不同品类之间的需求量

经过对数据的观察分析后发现，产品编码分为大类编码与细类编码，针对不同层次的分类，分别对需求量做统计和分析。按细类编码分类，分析需求量和价格的关系：

表 8 不同分类层次下的平均价格于需求量

大类编码	细类编码	平均价格	细类需求量	大类需求量
301	405	849.6	1586723	1586723
302	408	1328.2	6221334	6221334
303	401 406 410 411	1597.7 1796.6 4878.3 3598.2	3605644 47022 81654 13705	3748025
304	409	1726.0	618444	618444
305	412	735.5	6324256	6324256
306	402 407	326.5 906.2	2530770 22017667	24548437
307	403	588.4	5728696	5728696
308	404	1283.3	5994397	5994397

以上表格归纳出了细类与大类的从属关系，以及他们各自的平均价格和需求量。可以发现，大类需求最多的几类商品，均价都是低于所有商品的均价，所以共同点为，商品的需求量基本与价格呈负相关，下图直观地显示了销量和均价的关系。

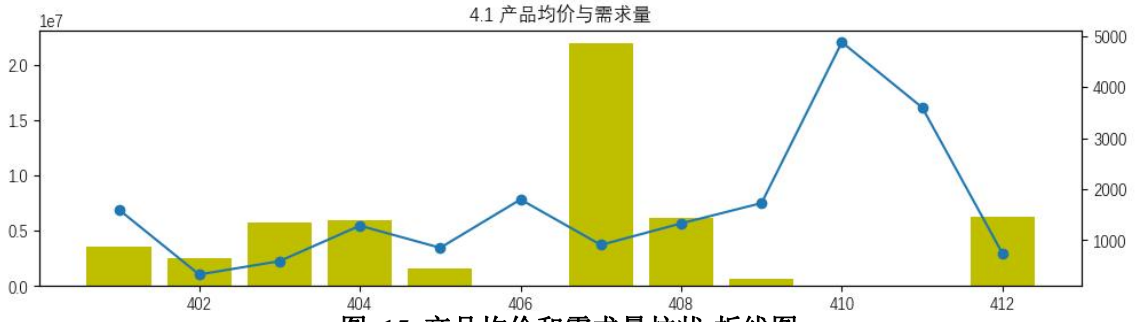


图 15 产品均价和需求量柱状-折线图

但是也有例外，如 405 产品，即使价格低于均价，销量依然很低，可见此类商品仍需要加大宣传，争取市场份额，不同点在于不同商品所占的市场份额有一定差距，即使为同一大类下，如 303 大类下，各细类份额仍有很大差距，绘图如下。

图 16 各类产品需求比例饼状图

3.5 不同时段的产品需求量特性

针对不同时段，分别对月份（上中下旬）、四季（春夏秋冬）不同时段的需求进行统计和分析。

利用 groupby()函数，对月份的上中下旬分组，得到每个细类的需求平均值，可以发现 303 大类（401，406，410，411）以及 412 和 408 在一个月内需求的变化并不明显，而 403，404，405 更倾向于月初和月末销量更客观，402、407 细类随着月份的推移，平均需求量会递减，409 细类递增，如下图所示。

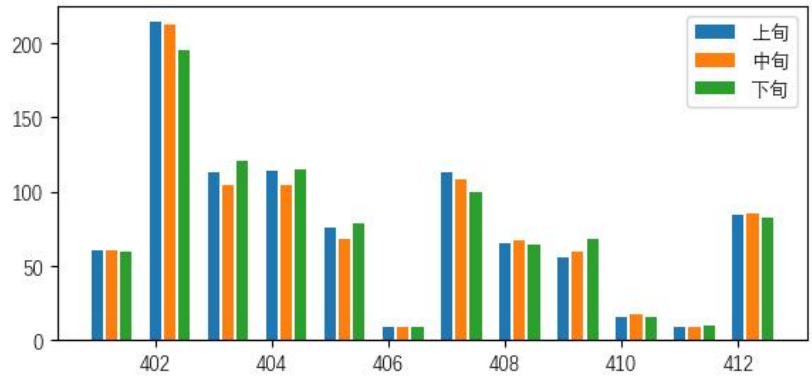


图 17 每月不同时段产品需求特性柱状图

对于季节因素，同样利用 groupby()函数对季节分组，得到结果直方图如下。

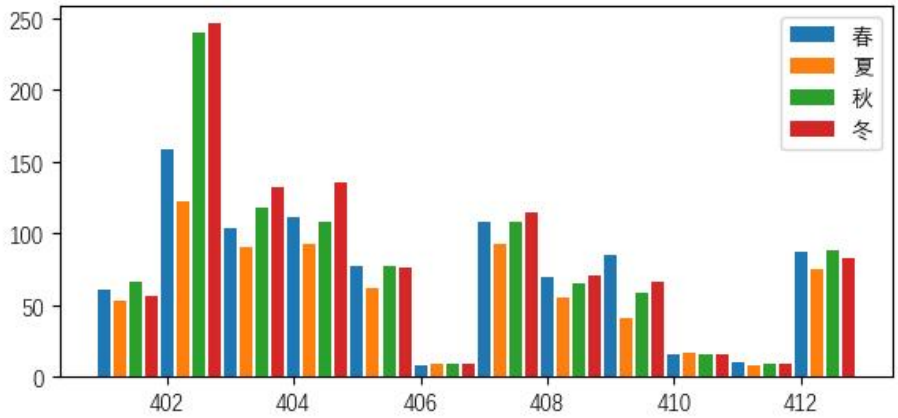


图 18 不同季节产品需求特性柱状图

可以明显看到 402、403、404 的需求受季节因素影响大于其他类别，详细来说秋冬季的需求大于春夏季，由此可以推断此类商品应为秋冬应季商品，而其他类别受季节影响并不明显。

3.6 节假日与促销日对需求量的影响

节假日与促销日都处于日期这一维度上，因此可以绘制日期-需求量曲线，观察节假日和促销日前后的需求量变化。

先将所有数据按年份分组，再按销售渠道分组，使每组的样本数减少到合理范围，最后将同一天的不同商品的需求量相加，绘制日期-需求量曲线，标出属于节假日和促销日的点如下图所示。

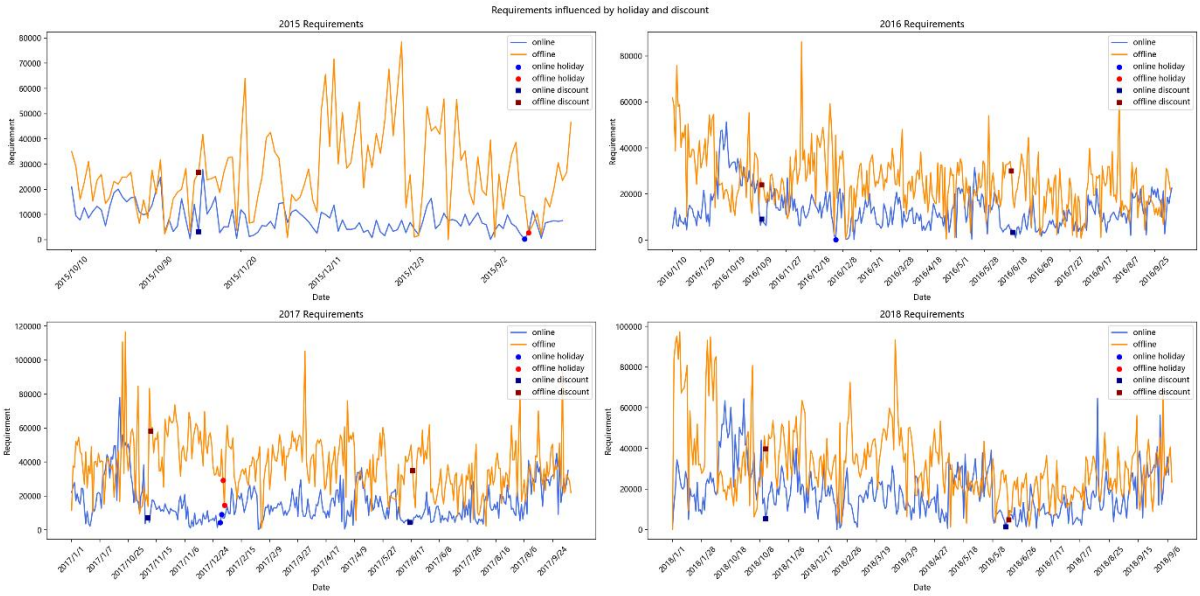


图 19 年线上-线下需求曲线图(节假日促销日单独标注)

从上图可以看出，在大部分节假日和促销日位于曲线的极小值点，在节假日和促销日之后需求量上升，且上升幅度较大。原因是经销商可能在节假日和促销日当天将已有的商品倾销出去，随后根据客流量安排之后的订货量，所以企业可在节假日和促销日之后再增加生产，节约不必要的成本。

3.7 季节因素对需求量的影响

将所有数据按年份分组，再按季节分组，将每组的需求量相加，绘制季节-需求量柱状图。

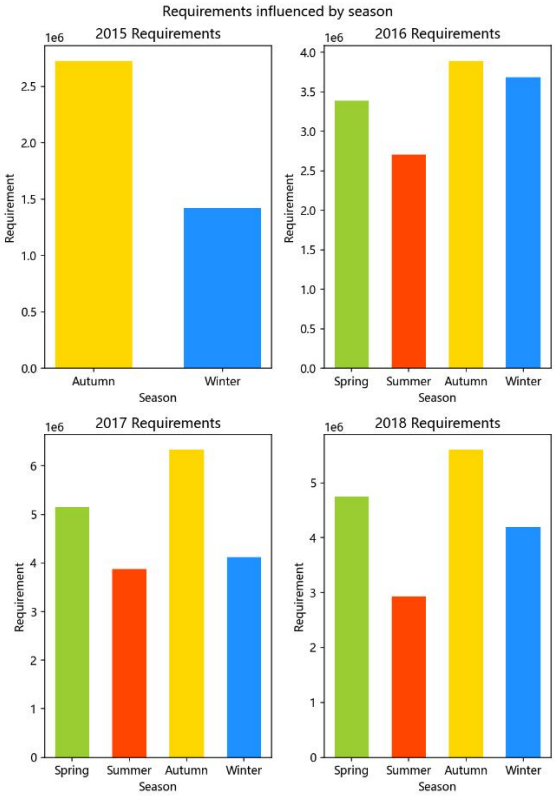


图 20 季节-需求量柱状图

从上图可知，所有年份中春季和秋季的需求量均高于夏季和冬季的需求量，其中秋季的需求量为全年最高，夏季的需求量为全年最低，这说明该企业的产品属于秋冬季节的产品，应该在旺季增加生产供应，或者开发夏季产品，增加商品种类的多样性。

四、 基于 LSTM 的需求量的预测

4.1 问题分析

在本任务中，需要预测 predict_sku1 表单中 2619 个样本的 2019 年前 3 个月的需求量，观察表单数据，发现特征仅有销售区域，产品编码，产品大类编码以及产品细类编码，这些特征均为分类变量，且预测时间较长，因此选用 LSTM 模型进行预测。

4.2 LSTM 神经网络

长短期记忆网络(LSTM)是一种用于处理长时间序列的特殊循环神经网络(RNN)，其为解决长期依赖问题而设计，LSTM 在 RNN 的基础上增加了对过去状态的筛选，从而可以有效选择更有影响的状态，并通过从长序列数据中提取长期依赖信息，有效避免了梯度消失和爆炸等问题的出现[1]。LSTM 神经网络在语音建模、翻译、识别和图片描述等方面取得了一定成功。

LSTM 结构相对于 RNN 更为复杂，在 RNN 结构中，其将过去的输出和当前的输入相连接，并通过 Tanh 函数来控制两者的输出，但是它只考虑最近时刻的状态。在 RNN 中有两个输入和一个输出。其结构如下图所示。

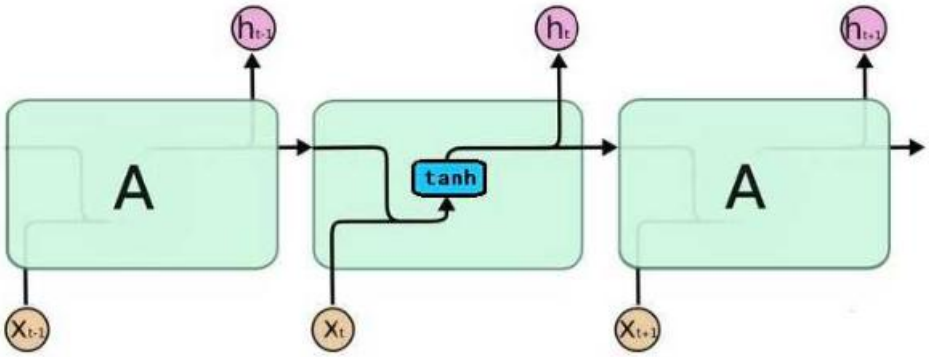


图 21 RNN 内部结构图

LSTM 同样具有相似的链状结构，但在重复模块结构上有显著差异，LSTM 增加了三个神经网络层，分别代表遗忘门、输入门和输出门，其结构如下图所示。

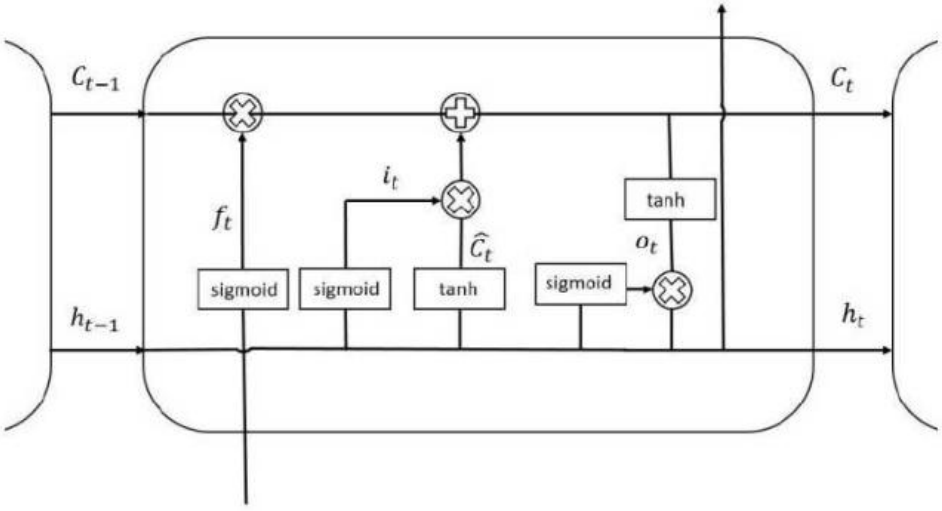


图 22 LSTM 内部结构图

遗忘门的任务是筛选细胞状态中的信息，并将其有选择性的遗忘，通过接受上个单元模块的输出模块C_{t-1}并决定要保留和遗忘C_{t-1}哪个部分，其公式如下：

$$f_t = \text{sigmoid}\left(W_f \times [x_t, h_{t-1}] + b_f\right) \tag{1}$$

$$C_t = i_t \times \widehat{C}_t + f_t \times C_{t-1} \tag{2}$$

x_t为时刻t神经元的输入，h_t为时刻t神经元的输出，代表权重矩阵，b代表阈值向量，sigmoid 为神经元激活函数，公式 2 中f_t为遗忘门输出值。

输入门的功能是有选择地记录新的信息到细胞状态，并决定储存哪种新信息到细胞状态(单元模块)中。输入门包括 Sigmoid 层和 Tanh 层，Sigmoid 层确定何值的更新，Tanh 层生成新的候选记忆，添加补充丢弃的属性信息。其公式如下：

$$i_t = \text{sigmoid}\left(W_i \times [x_t, h_{t-1}] + b_i\right) \tag{3}$$

$$\widehat{C}_t = \tanh\left(W_c \times [x_t, h_{t-1}] + b_i\right) \tag{4}$$

公式 4 中$\widehat{C}_t$为输入门结果，i_t决定$\widehat{C}_t$是否加入 t 时刻状态。最终将二者进行乘积，从而获得最终输出信息。其公式如下：

$$o_t = \text{sigmoid}\left(W_o \times [x_t, h_{t-1}] + b_o\right) \tag{5}$$

$$h_t = o_t \times \tanh\left(C_t\right) \tag{6}$$

公式 6 中h_t为输出门最后输出。

LSTM 最大的创新就是三个门的引入，它简化了神经网络的训练过程、解决了梯度爆炸与消失等问题。LSTM 成为了一种具有自学习、联想存储和指导搜索优解反馈神经网络。但 LSTM 模型对数据的周期和趋势的变动不敏感，也具有计算费时、不能并行运算等缺点，同时，当序列长度超过一定限度后，LSTM 仍会出现梯度消失的问题[2]。

4.3 问题求解

为了更加形象地介绍模型求解过程，以对产品编码为 20002，产品大类编码为 303，细类编码为 406，售往 101 区域的产品为例按照日粒度进行预测，求解整体流程图如下。

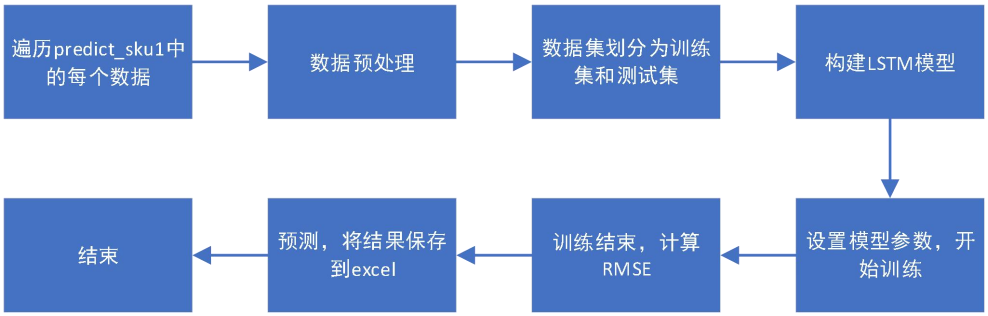


图 23 模型求解流程图

4.3.1 数据预处理

●step1 提取相同数据

首先，在 python 中通过表格连接等操作，发现部分在 predict_sku1 表单中所存在的数据并未在 order_train1 表单中所发现。所以，针对这类数据，选择用特征相似的数据进行预测，例如在 predict_sku1 中，产品 20011 并未在 order_train1 中存在，因此将相同的销售区域编码、产品大类编码和产品细类编码作为特征在 order_train1 中寻找相似样本进行训练，对于 20002 号产品，在 order_train1 寻找出 66 个相同的数据，这些数据将作为预测 20002 号产品需求量的 LSTM 模型训练数据。

订单日期	销售区域编码	产品大类编码	产品细类编码	产品编码	订单需求量
2017-08-04	101	303	406	20002	4
2018-03-14	101	303	406	20002	2
2018-03-16	101	303	406	20002	3
2018-03-25	101	303	406	20002	3
2018-03-31	101	303	406	20002	9
...	...	...	...	...	...
2018-11-30	101	303	406	20002	9
2018-12-13	101	303	406	20002	18
2018-12-14	101	303	406	20002	24
2018-12-20	101	303	406	20002	8
2018-12-20	101	303	406	20002	6

66 rows × 5 columns

图 24 order_train1 中的 20002 号产品数据

●step2 补全时间序列

对于某一个产品，存在某些天没有出售的情况，所以按照天进行分组并求出订单需求量的均值，再通过前向插值法补全订单需求量，得到连续的订单需求量，如图 25 所示，补全后 20002 号产品销量走势图如图 26 所示。

订单日期	订单需求量
2017-08-04	4.0
2017-08-05	4.0
2017-08-06	4.0
2017-08-07	4.0
2017-08-08	4.0
...	...
2018-12-16	24.0
2018-12-17	24.0
2018-12-18	24.0
2018-12-19	24.0
2018-12-20	7.0

504 rows × 1 columns

图 25 补全时间序列后的 20002 号产品数据

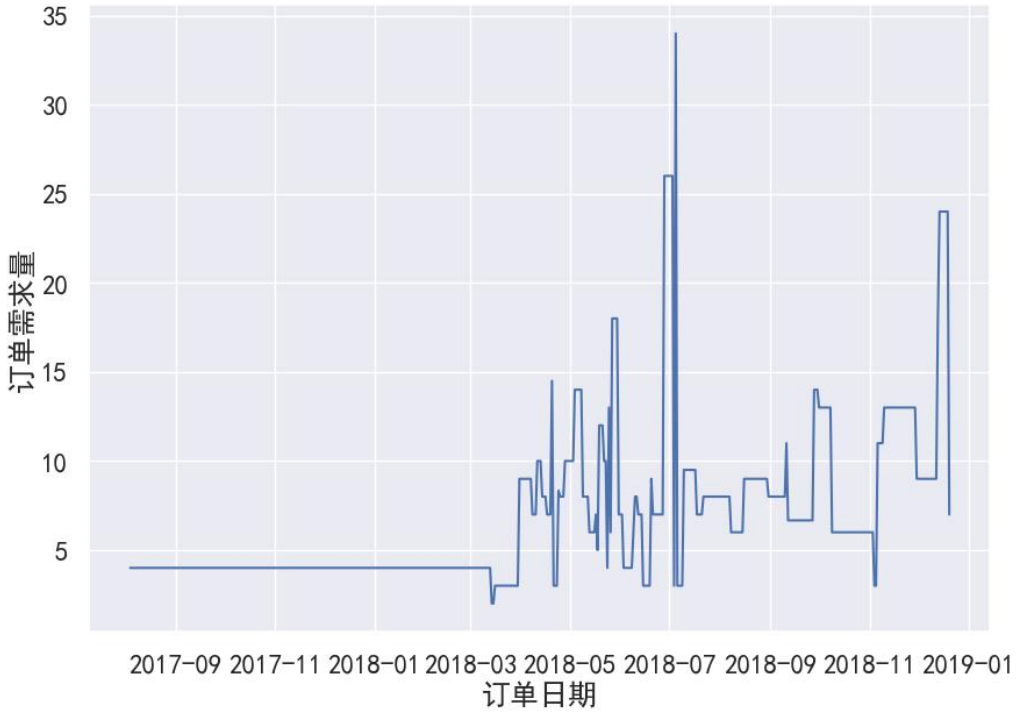


图 26 20002 号产品销量走势图

4.3.2 模型训练

●step1 构造训练集和测试集

首先先将所有数据归一化以提高算法的收敛速度，其次将数据的 90%作为训练集，10%作为测试集，来进行模型的训练，针对 20002 号产品，训练集的个数为 453，测试集的个数为 51。

●step2 设置模型参数并进行模型的训练

设置滑动窗口为 5，LSTM 层的单元数为 100，全连接层数为 1，迭代次数为 300，损失函数为 mse，进行模型的训练，迭代过程中的损失函数值如下图所示。

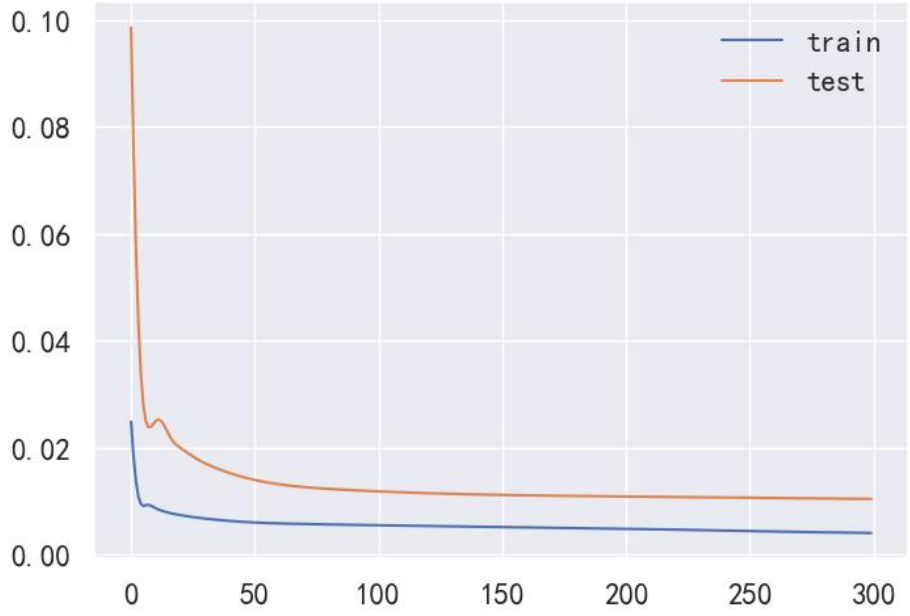


图 27 迭代过程中的测试集和训练集损失函数值

从图中可以看出，在训练的过程中，随着迭代次数的提高，损失函数逐渐下降，几

乎近似于 0，说明模型训练的效果很好，将数据进行反归一化后，得到的测试集实际值和预测值对比图如下图所示。

图 28 测试集实际值和预测值对比图

从图中可以看出该模型预测效果良好，可以应用于后期的预测需求，且 RMSE 为 3.284，模型精度较高。

4.4 预测

接着构造待预测数据的时间序列，用训练好的模型进行预测，预测结果图如下图所示。

图 29 预测结果图

最后，将每天的数据按月分组求和，得到每月的订单需求量，将结果存入 excel 表格中，得到按照天预测的 20002 号产品 1 月需求量为 248，2 月为 224，3 月为 248。

根据周和月预测的结果和上述流程类似，将数据按周和月分组即可，所有预测数据均存入 result1.xlsx 中。

4.5 精度比较

针对 20002 号产品的预测，按照日、周、月的方式进行预测，预测结果和 RMSE 如

下表所示。

表 9 不同粒度预测结果分析

	日粒度	周粒度	月粒度
2019 年 1 月	248	261	260
2019 年 2 月	224	228	232
2019 年 3 月	248	251	257
RMSE	3.284	3.437	4.573

从结果中可以看出随着采样日期细致程度的降低，RMSE 逐渐上升，按照日粒度进行预测的结果最为准确，周粒度次之，月粒度最差，因此企业如果想要精准预测未来某件产品的需求量，按照日粒度进行预测为最佳选择。

五、 总结

基于上述分析，我们能够得出以下结论：

1、该企业大部分产品价格为中低价位，且随着价格的升高，平均订单需求量逐渐下降，说明中低价位产品为企业的主要营销市场。

2、区域对订单需求量有显著影响，且不同区域的订单需求量差异较大，每个区域的偏好产品以及销售方式也有所不同，企业应通过分析结果依据区域特征制定有针对性的靶向营销策略。

3、不同销售渠道之间的商品需求量有所差异，具体来说，线上销售渠道需求量大于线下，所以各品类商品发展线上渠道是非常必要的。

4、不同品类的商品之间，需求量不尽相同，即使在相同大类下的商品，需求量也会有较大差异，相同点在于价格更低的商品，需求量倾向于更大，呈负相关.

5、在不同的时段，各品类产品销售也有差别，按月划分可以看到部分商品呈现月初月末需求大于月中，有些则随着月份深入需求递减，同时在按季节区分时，不同商品也呈现不同的需求特点，例如 402，403，404 明显在秋冬季节需求量大于其他商品以及自己在春夏季的需求量。

6、产品的需求量在节假日和促销日之后会大幅上涨，因此可以在节假日和促销日之后增加生产供应，满足经销商需求。

7、产品的需求量在春季和秋季较高，说明该企业生产的商品属于秋冬季节的产品，可以在秋冬季节增加生产，在夏季减少生产，节约成本。

六、 团队管理

刘泽坤：第一题第 1、2 小问、第二题求解以及相应部分报告撰写、摘要撰写、视频录制及剪辑、论文排版(贡献度：40%)。

韦振宇：第一题第 6、7、8 小问以及相应部分报告撰写、绪论撰写(贡献度：30%)。

李嘉伟：第一题第 3、4、5 小问以及相应部分报告撰写(贡献度：30%)。

参考文献

[1]常昊.基于LSTM神经网络的地铁短时客流量预测研究[D].西京学院,2022.DOI:10.27831/d.cnki.gxjxy.2022.000059.

[2]杨旭,黄雪梅.基于 LSTM 神经网络的饲料企业财务风险预警模型构建[J].中国饲料,2022(14):135-138.DOI:10.15906/j.cnki.cn11-2975/s.20221433.