队伍编号	MCB2303455
赛道	B

电商零售商家需求预测及库存优化问题

摘 要

随着互联网技术高速发展，线上购物成为了越来越多消费者的选择，电商零售商家无论在规模还是在种类上同样飞速增长。与此同时，对客户需求的分析以及对自身产品存储策略的制定对商家的盈亏显得尤为重要。因此商家若想在提高服务水平的同时尽可能节约成本，对客户需求的精准预测与库存方案的优化是关键所在。

针对问题一预测 2023 年 5 月中下旬需求。本题在读取数据并预处理后首先进行 ADF 检测和白噪声检测，以此来观测数据的平稳性与价值。由于未出现白噪声且检测结果显示数据的一阶差分总体平稳。因此本题的模型选择为**ARIMA 模型**。关于模型参数的确定工作，我们根据最佳 aic 和 bic 来进行确定。对于商家、仓库、商品的分类问题，我们将表 2，3，4 全部运用 merge 方法拼接起来，选取出相关的分类，利用**K-means 算法**：基于样本的相似性对样本进行分群；对特征值进行归一化，大量循环测试选择最优 sse 评分的情况得出 K-means 最佳参数。先直接将参数适配聚类模型，然后利用**pca 主成分分析**将数据降维到三维进行分析，根据聚类分析后的分布结果来对商家、仓库、商品进行分类。

由于第二题要求在增加相应“商家+产品＋仓库”维度的情况下预测五月中下旬的需求量。因此需要对比附件中的各个序列，查找历史数据中与之相似的时间序列。计算其相似度。相似性可以通过多种方式度量，例如欧氏距离，相似性矩阵，相关系数等。找到最相似的历史序列，用作预测新维度的参考。同时参考 ARIMA 模型在上一题的表现，该模型仍是本题的选用模型之一。但因为部分序列的非线性关系和季节性，故本题同样采用了**线性回归**模型进行预测。

关于第三题，此题的复杂度相比前两题有了较大的提高，且给出的数据集与需要预测的时间节点需求量相隔较大（需要预测 2023 年 6 月促销销售情况），故本题在读取附件 1-4 与附件 6 的同时借助了问题二的预测结果。在对数据进行 groupby 分组等预处理工作后，本题选用三次**指数平滑模型**进行预测，但结果过于稀疏。

关键字：需求预测、时间序列、ARIMA 模型、线性回归、指数平滑模型

目录

一、问题重述与问题背景

1. 问题背景

随着互联网购物的高速发展与普及，线上购物成为了越来越多消费者的选择。然而在线上购物的规模不断扩大的同时，对客户需求的分析以及对自身产品存储策略的制定对商家的盈亏显得尤为重要。对此一些商家选择降低员工薪酬、福利甚至裁减经营规模，但效果往往事倍功半。因此商家若想在提高服务水平的同时尽可能节约成本，对客户需求的精准预测与库存方案的优化是关键所在。

2. 问题重述

问题一

使用附件 1-4 中的数据，预测出各商家在各仓库的商品 2023-05-16 至 2023-05-30 的需求量，请将预测结果填写在结果表 1 ，对模型性能进行评估，并根据数据分析及建模过程，这些由商家、仓库、商品形成的时间序列如何分类，使同一类别在需求上的特征最为相似？

问题二

针对一些新出现的“商家+仓库＋商品维度”（导致这种情况出现的原因可能是新上市的商品，或是改变了某些商品所存放的仓库。）通过参考附件 1 的数据到相似序列并完成这些维度在 2023-05-16 至 2023-05-30 的预测值。

问题三

每年 6 月会出现规律性的大型促销，为需求量的精准预测以及履约带来了很大的挑战。附件 6 给出了附件 1 对应的商家+仓库+商品维度在去年双十一期间的需求量数据，请参考这些数据，给出 2023-06-01 至 2023-06-20 的预测值。并把预测结果填写在结果表 3。

二、问题分析

1. 第一题分析

根据 ADF 检测和白噪声检测，能判断此序列属于非平稳序列，但其一阶差分属于平稳序列,那么对于平稳序列进行预测活动,我们运用了 ARIMA 模型----我们根据最佳 aic 和 bic 来确定参数，预处理中发现数据中没有缺失值，于是我们将 dataframe 按照商家，仓库，商品，时间的顺序来排序。

依据预处理结果，我们发现，数据中对于唯一的（商家，仓库，商品）组合，有严格的 166 天数据，于是我们运用 arima 模型去分别适配每个组合的时间序列，然后我们将结果拼起来，获得指定的预测结果。

另外对于商家、仓库、商品的分类问题，我们将表 2，3，4 全部运用 merge 方法拼接起来，选取出相关的分类，利用 K-means 算法：基于样本的相似性对样本进行分群；对特征值进行归一化，大量循环测试选择最优 sse 评分的情况得出 K-means 最佳参数。先直接将参数适配聚类模型，然后利用 pca 主成分分析将数据降维到三维进行分析，根据聚类分析后的分布结果来对商家、仓库、商品进行分类。

2. 第二题分析

新的维度可能代表新上市的商品或改变了商品存放的仓库。因此需要对比附件中的各个序列，查找历史数据中与之相似的时间序列。计算其相似度。相似性可以通过多种方式度量，例如欧氏距离，相似性矩阵，相关系数等。找到最相似的历史序列，用作预测新维度的参考。

通过选定的历史序列，使用时间序列模型对时间段为 2023-05-16 至 2023-05-30 的序列进行训练和预测，可视化相应的相关系数指标。这里可以沿用第一题中的 ARIMA 模型，同时对模型精度等性能进行评价。

在具体操作中，我们将附件 1 和 5 中的数据进行预处理和合并操作后，识别出历史数据中不存在的商家+仓库+商品维度。根据附件 5 中的新维度，在附件 1 的数据中找到相似的历史序列。计算相似度矩阵和欧式距离，找到最相似的历史序列。使用 ARIMA 对序列进行训练，需要进一步调整参数以输出理想的 qty 的预测值。最后对该模型的性能进行评价。

3.第三题分析

问题三中要求读取 2022 年双十一促销期间附件 1 各商家+仓库+商品的需求量，以此估计 2023 年-06-01 至 2023 年-06-20 促销期间的需求量。针对此情况本题决定合并附件 1-附件 4 与附件 6，同时读取第二题的预测结果方便。利用指数平滑模型（三次指数平滑模型）预测 2023 年 6 月 1 日至 6 月 20 日的需求量。

三、模型假设

（1） 不同商品、商家之间彼此独立。
（2） 所有商品在任何仓库存放的每日存储成本均为 h。
（3） 不存在商家与客户之间因各种原因取消交易的情况

四、符号说明

附件 1：商家历史出货量表

字段名	字段类型	描述
seller_no	String	商家编码
product_no	String	商品编码
warehouse_no	String	仓库编码
date	String	日期
qty	Float	出货量（可看做需求量）

附件 2：商品信息表

字段名	字段类型	描述
product_no	String	商品编码
category1	String	商品一级分类

category2	String	商品二级分类
category3	String	商品三级分类

附件 3：商家信息表

字段名	字段类型	描述
seller_no	String	商家编码
seller_category	String	商家分类
inventory_category	String	库存分类
seller_level	String	商家规模

附件 4：仓库信息表

字段名	字段类型	描述
warehouse_no	String	仓库编码
warehouse_category	String	仓库类别
warehouse_region	String	仓库区域

五、问题一的模型建立与求解

5.0 在工作之前的一些猜测

在开始工作之前，我们阅读题目，观察数据，初步断定我们的工作是根据历史时间上的规律来预测 qty 数据的变化，于是我们想到了 ARIMA 模型和 ARMA 模型。此外我们也考虑到运用线性回归模型，Prophet 模型、LSTM 模型进行匹配。在本阶段中，我们逐步发现其他模型的适用性不佳，并在 ARIMA 和 ARMA 之间做了大量的比对，最后我们选择了 ARIMA 模型。

5.1、数据的预处理

由题干可知此问题涉及附件 1-附件 4 的数据，因此在建立模型前先后对附件 1 至附件 4 四个数据集进行读取。由于观察到数据中“交易日期”、“商家编号”、“商品编号”，“仓库编号”等属性均杂乱无章，但并未发现缺失值与离群点。在观察序列的特征后，我们发现，本序列可以根据当前时间出货量来预测未来商品出货量，故预处理工作主要为将数据分类，并且按照每一类数据的时间先后进行排序。

首先在将时间数据转换为 datatime 格式后便仅依照日期进行排序，但发现所有相同日期内出现的“商家、仓库、商品”信息依旧混乱。为解决此问题，将数据排序的方式修改为利用 Pandas 下辖的 Dataframe 以“商家、仓库、商品”维度作为主键按照时间进行升序排序。之后，为了观察数据在时间上的变化情况，我们构建时间窗口特征，对数据进行分析。预处理完成后，可获得相应描述性统计信息。（包括但不限于附件 1 中记载的商家数量与商品数量等）。

5.2、数据平稳性估计与参数估计

经观察，预处理后的数据中主键（商家+仓库＋商品维度）的有严格的 166 天数据。基于此情况，首先每 166 行数据选取了五组作为例子进行检验。在时间序列检验流程中，对需求量数据采用了滚动统计的方式。将窗口值设置为 30，随后检测均值与方差是否发

生明显变化并绘制滚动统计结果。

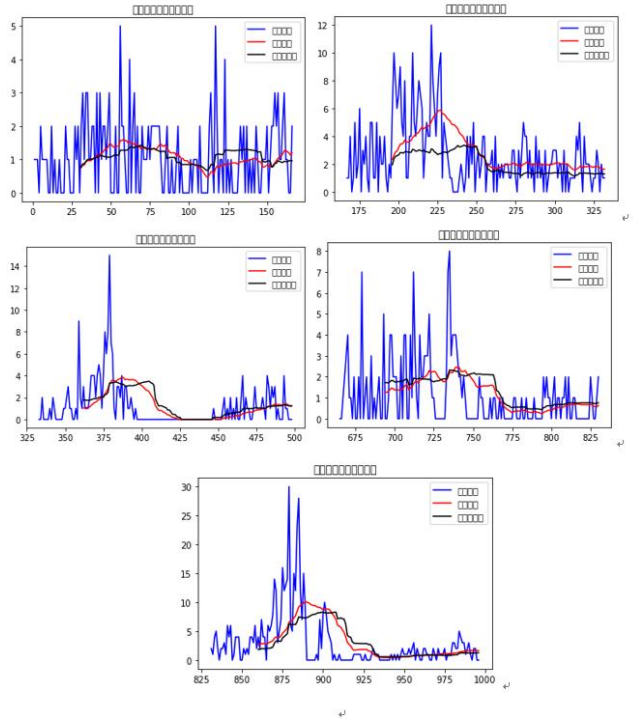


图 1：所选分类的滚动统计序列检测图（蓝色：原始数据；红色：滚动均值；黑色：滚动标准差，图二同理）

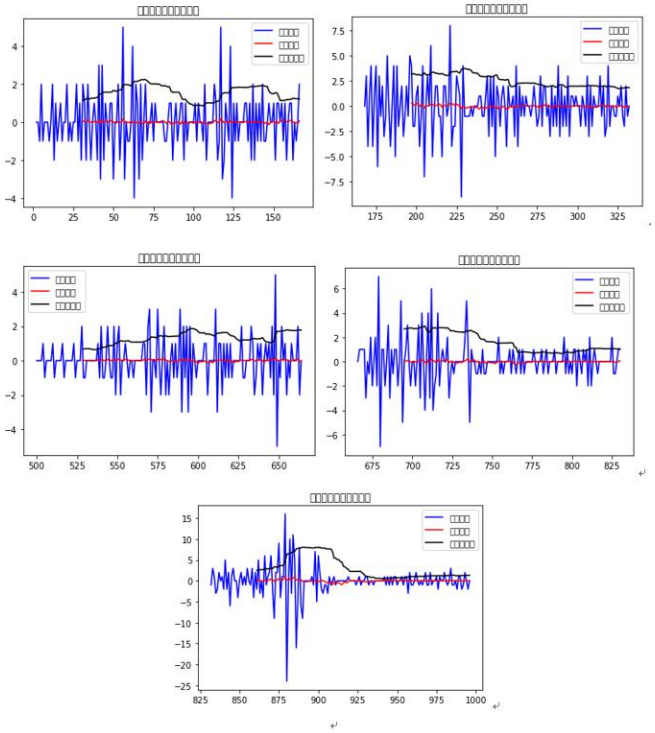


图 2：所选分类的一阶差分滚动检测图

我们根据滚动检测图可以发现，原时间序列的时间平稳性比较差，但是当我们对时

间序列做了一阶差分之后，其非常接近平稳序列，我们猜测，其可能适用于 ARIMA 模型，于是我们需要接着对其进行下一步的检验。

针对数据的平稳性检验问题，最终采用了 Dicky-Fuller 检验法（滞后选择方法为 BIC），其公式（时序回归方程）如下。

$$X_t = k * (X_{t-1} - 1) + b$$

表 1：Dicky-Fuller 检验结果

Test Statistic	-5.780271e+00
p-value	5.145248e-07
#Lags Used	2.000000e+00
Number of Observations Used	1.630000e+02
Critical Value (1%)	-3.471119e+00
Critical Value (5%)	-2.879441e+00
Critical Value (10%)	-2.576314e+00

Test Statistic	-3.347066
p-value	0.012901
#Lags Used	3.000000
Number of Observations Used	162.000000
Critical Value (1%)	-3.471374
Critical Value (5%)	-2.879552
Critical Value (10%)	-2.576373

Test Statistic	-1.821247
p-value	0.369965
#Lags Used	7.000000
Number of Observations Used	158.000000
Critical Value (1%)	-3.472431
Critical Value (5%)	-2.880013
Critical Value (10%)	-2.576619

表 2：一阶差分后 Dicky-Fuller 检验结果

Test Statistic	-9.790684e+00
p-value	6.365719e-17
#Lags Used	4.000000e+00
Number of Observations Used	1.600000e+02
Critical Value (1%)	-3.471896e+00
Critical Value (5%)	-2.879780e+00
Critical Value (10%)	-2.576495e+00

Test Statistic	-5.967114e+00
p-value	1.975730e-07
#Lags Used	1.000000e+01
Number of Observations Used	1.540000e+02
Critical Value (1%)	-3.473543e+00
Critical Value (5%)	-2.880498e+00
Critical Value (10%)	-2.576878e+00

Test Statistic	-8.797529e+00
p-value	2.156135e-14
#Lags Used	6.000000e+00
Number of Observations Used	1.580000e+02
Critical Value (1%)	-3.472431e+00
Critical Value (5%)	-2.880013e+00
Critical Value (10%)	-2.576619e+00

对于原数据的 ADF（Augmented Dickey-Fuller）检验结果分析如下：

Test Statistic（检验统计量）：ADF 检验的核心目标是检验时间序列数据是否具有单位根（非平稳性）。检验统计量的值与临界值的比较有助于确定是否可以拒绝单位根存在的假设。在这种情况下，原数据的 Test Statistic 表现普遍小于置信水平为 1%、5%和 10%的临界值，但它接近于临界值，表明结果并不非常明确。

p-value（p 值）： p 值表示在单位根存在的假设下，观察到检验统计量或更极端值的概率。在这种情况下，原数据 p 值普遍相对较小，但仍然大于通常的显著性水平（例如 0.05），这意味着此数据不能非常强烈地拒绝单位根存在的假设。

#Lags Used（滞后阶数）： 这个值为 3.000000，表示在进行检验时使用了 3 个滞后阶数。

Number of Observations Used（观测值数量）表示在检验中使用的观测值的个数。

Critical Values（临界值）： 这些是用于比较检验统计量的临界值，以确定是否可以拒绝单位根存在的假设。在所有三个置信水平下，检验统计量的值都普遍接近或略

低于相应的临界值，但并没有明显低于它们。

综合考虑这些结果，我们可以得出结论，虽然检验统计量的值略低于一些临界值，但 p 值相对较高，因此不能强烈拒绝单位根存在的假设。这可能表明您的时间序列数据在某种程度上具有一些非平稳性，但结果不够明确。进一步的分析可能需要考虑其他方法或模型来处理时间序列数据。

而对于差分后数据的情况，我们可以分析得到如下结论：

Test Statistic（检验统计量）：它用于检验时间序列数据是否具有单位根。在 ADF 检验中，如果这个值比临界值更小（绝对值更大），则表明数据具有平稳性。在这种情况下，差分后数据的检验统计量的值普遍非常小，表明数据不具有单位根，可能是平稳的。

p-value（p 值）：它用于判断检验统计量的显著性。在统计学中，通常选择显著性水平（例如 0.05），如果 p 值小于显著性水平，则可以拒绝原假设。在这里，差分后数据的 p 值普遍非常接近零，远远小于通常的显著性水平，表明可以拒绝原假设，即数据不具有单位根，可能是平稳的。

#Lags Used（使用的滞后阶数）： 这里使用了 4 个滞后阶数，这是在检验中引入的滞后项以考虑时间序列中的自相关性。通常，滞后阶数的选择取决于一些统计准则。

Number of Observations Used（使用的观测数）： 有 160 个观测值被用于进行 ADF 检验。

Critical Values（临界值）： 这些值是在不同显著性水平下的临界值。差分后数据的检验统计量普遍小于这些临界值，这也支持了拒绝原假设的结论，即数据可能是平稳的。

综合来看，根据检验统计量非常小和极小的 p 值，以及检验统计量小于显著性水平下的临界值，可以得出结论，差分后数据的时间序列数据在 ADF 检验下被认为是平稳的，不具有单位根。这是一个积极的结果，为我们否定 ARMA 模型，决定使用 ARIMA 模型提供了理论依据。

针对数据的分布情况，本题利用白噪声模型（具体方式为 Ljung-Box 检验,延迟数为 1），其基本公式如下。

$$X_t = e_t \mid e_t \sim WN(0,\sigma^2)$$

白噪声检测可以帮助我们判断数据的随机性和噪声水平。如果数据符合白噪声的统计特性，那么我们可以假设数据是随机的，并且可以使用一些经典的统计方法进行建模。如果数据包含非随机成分或噪声较大，我们可能需要采取进一步的处理措施，例如去除噪声或寻找其他适当的模型。

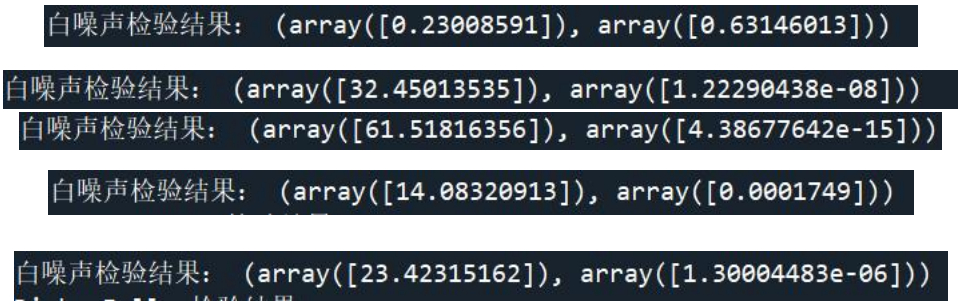


图 3 原数据白噪声检测

第一个数组是检测到的白噪声的统计量。通常，在白噪声检测中，我们关注的是统计量是否接近于零。因此，这个结果表明检测到的白噪声统计量非常接近于零，这是一个好的指标。

第二个数组是与白噪声统计量相关的 p 值。p 值表示观察到的结果在假设条件下发生的概率。在这种情况下，非常小的 p 值（接近于零）意味着我们可以拒绝零假设并得出结论：检测到的数据不符合白噪声特征。这可能表明数据中存在一些非随机的成分或其他类型的噪声。

总体来看，根据提供的结果，我们可以得出结论：原数据不符合白噪声的统计特性。即原数据存在一些规律性变化，不符合纯随机数据。于是，我们可以利用 arima 模型对其进行预测。

在确定了模型之后，我们可以通过 AIC 和 BIC 的值确定模型参数

$$BIC=k\ln(n)-2\ln(L)$$

其中 k 为模型参数的个数，n 为样本的数量，L 为似然函数。
根据程序计算的结果，取每一个时间序列 bic 值最低的情况综合考虑，并结合 ADF 检测的结果，我们决定选择参数为（1，1，1）

5.3、时间序列预测实现

我们建模的基本思想是利用数据本身的历史信息来预测未来。一个时间点上的标签值既受过去一段时间内的标签值影响，也受过去一段时间内的偶然事件的影响。为了获得更精确的预测结果，在此部分工作中，我们使用 arima 模型来预测时间序列。
ARIMA 模型的数学表达式：
首先我们可以分析 AR 和 MA 模型的数学表达式：

$$AR:\ Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \xi_t$$

$$MA:\ Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

如果我们暂时不考虑差分（即假设 d=0），那么 ARIMA 模型可以被看作是 AR 模型和 MA 模型的直接结合，形式上看，ARIMA 模型的公式可以表示为：

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

上述模型被称之为 ARIMA(p,d,q)模型，其中 p 和 q 的含义与原始 MA、AR 模型中完全一致，且 p 和 q 可以被设置为不同的数值，而 d 是 ARIMA 模型需要的差分的阶数

我们将附件一原数据中的出货量设置为浮点数，然后适配模型，得出预测数据一。

5.4、关于数据的分类

在这个部分，我们主要使用了 k_means 算法来将相近的类分配在一起。
K-means 算法，也称为 K-平均或者 K-均值，是一种无监督的聚类算法。对于给定的样本集，按照样本之间的距离大小，将样本划分为 K 个簇，让簇内的点尽量紧密的连接在一起，而让簇间的距离尽量的大。他的工作流程如下：
①.给定一个待处理的数据集；
②.记 K 个簇的中心分别为 c 1 , c 2 , . . . , c k c1,c2,…,ck c1,c2,…,ck；每个簇的样本数量为 N 1 , N 2 , . . . , N 3 N1,N2,…,N3 N1,N2,…,N3；
③.通过欧几里得距离公式计算各点到各质心的距离，把每个点划分给与其距离最近的质心，从而初步把数据集分为了 K 类；
④.更新质心：通过下面的公式来更新每个质心。就是，新的质心的值等于当前该质心所属簇的所有点的平均值。

$$c_j = \frac{1}{N_j} \sum_{i=1}^{Nj} x_i, y_i$$

⑤.重复步骤 3 和步骤 4，直到质心基本不再变化或者达到最大迭代次数。

我们先根据附件二，三，四的内容，根据商家，商品，仓库的属性，利用 k_mean 聚类算法进行分析，得到如下分类结果：

根据商家进行分类得到数据：

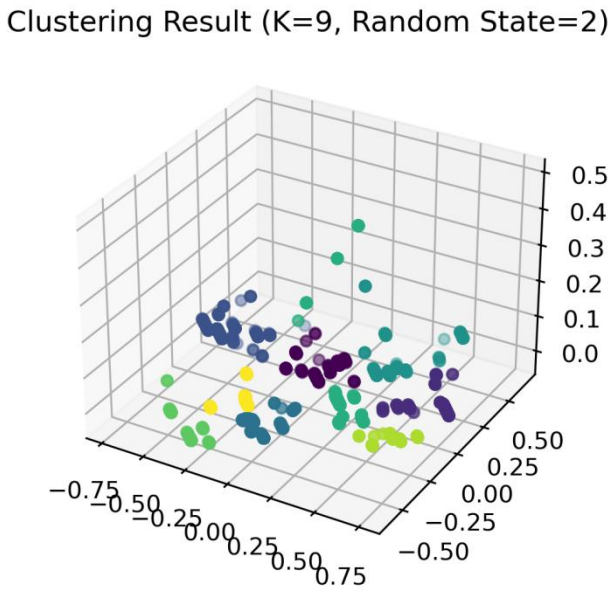


图 4 商家簇分类结果可视化

分类结果如下（图表先后为按照仓库分类与按照商品分类）：

warehouse_no	warehouse_category	warehouse_region	cluster
wh_50	1	6	0
wh_43	1	6	0
wh_48	1	6	0
wh_38	1	6	0
wh_38	1	6	0
wh_14	1	6	0
wh_37	1	6	0
wh_23	1	6	0
wh_19	1	6	0
wh_2	1	6	0
wh_5	1	2	1
wh_22	1	2	1
wh_45	1	2	1
wh_11	1	2	1
wh_7	1	2	1
wh_17	1	2	1
wh_15	1	2	1
wh_30	0	4	2
wh_13	1	4	3
wh_6	1	4	3
wh_32	1	4	3
wh_39	1	4	3

wh_16	1	4	3
wh_36	1	4	3
wh_46	1	4	3
wh_26	1	4	3
wh_51	1	4	3
wh_52	1	4	3
wh_47	1	1	4
wh_4	1	1	4
wh_53	1	1	4
wh_8	1	1	4
wh_49	1	1	4
wh_31	1	1	4
wh_29	1	1	4
wh_29	1	1	4
wh_44	1	1	4
wh_27	1	1	4
wh_18	1	1	4
wh_35	1	1	4
wh_33	1	1	4
wh_54	1	3	5
wh_9	1	3	5
wh_41	1	3	5
wh_40	1	3	5
wh_34	1	3	5
wh_3	1	3	5
wh_25	1	3	5
wh_24	1	3	5
wh_20	1	3	5
wh_12	1	3	5
wh_10	1	3	5
wh_1	0	1	6
wh_21	1	5	7
wh_42	1	5	7
wh_28	1	0	8

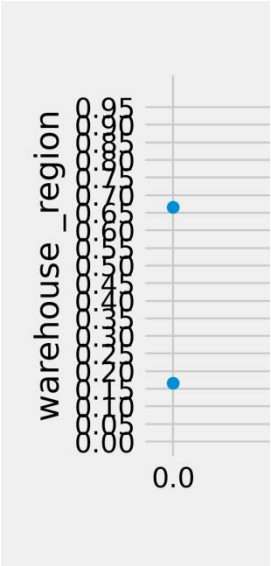


图 5 仓库分类结果

可视化

warehouse_no	warehouse_category	warehouse_region	cluster
wh_50	1	6	0
wh_43	1	6	0
wh_48	1	6	0
wh_38	1	6	0
wh_38	1	6	0
wh_14	1	6	0
wh_37	1	6	0
wh_23	1	6	0
wh_19	1	6	0
wh_2	1	6	0
wh_5	1	2	1
wh_22	1	2	1

wh_45	1	2	1
wh_11	1	2	1
wh_7	1	2	1
wh_17	1	2	1
wh_15	1	2	1
wh_30	0	4	2
wh_13	1	4	3
wh_6	1	4	3
wh_32	1	4	3
wh_39	1	4	3
wh_16	1	4	3
wh_36	1	4	3
wh_46	1	4	3
wh_26	1	4	3
wh_51	1	4	3
wh_52	1	4	3
wh_47	1	1	4
wh_4	1	1	4
wh_53	1	1	4
wh_8	1	1	4
wh_49	1	1	4
wh_31	1	1	4
wh_29	1	1	4
wh_29	1	1	4
wh_44	1	1	4
wh_27	1	1	4
wh_18	1	1	4
wh_35	1	1	4
wh_33	1	1	4
wh_54	1	3	5
wh_9	1	3	5
wh_41	1	3	5
wh_40	1	3	5
wh_34	1	3	5
wh_3	1	3	5
wh_25	1	3	5
wh_24	1	3	5
wh_20	1	3	5
wh_12	1	3	5
wh_10	1	3	5
wh_1	0	1	6
wh_21	1	5	7
wh_42	1	5	7
wh_28	1	0	8

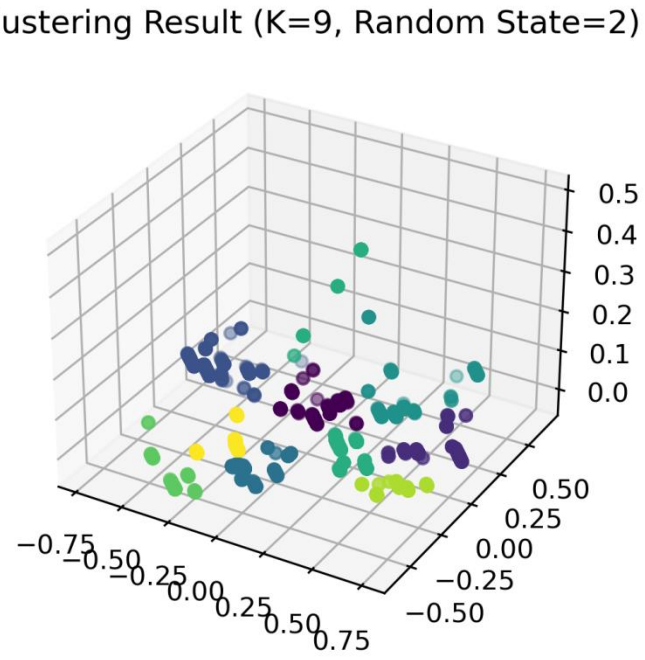


图 6 商品分类簇可视化

分类结果：

product_no	category1	category2	category3	cluster
product_2228	15	37	131	0
product_2323	11	40	149	0
product_2356	11	40	148	0
product_2262	15	37	128	0
product_2069	13	51	179	0
product_2340	14	39	135	0
product_2389	16	36	125	0
product_1436	11	40	148	0
product_2224	15	37	131	0
.....				
product_1478	7	0	5	1
product_661	8	12	40	1
product_567	7	0	0	1
product_1397	8	12	40	1
product_1497	7	0	8	1
product_1467	7	0	5	1
product_673	8	12	41	1
product_655	8	12	41	1
product_2171	7	14	53	1

product_561	7	0	8	1

.....

之后，我们将附件二，附件三，附件四中的数据连接起来，利用 k_eans 进行总的分类,预处理步骤如下：

1.将 df_sales 数据表中的"date"列转换为时间序列格式。

2.将 df_sales 和 df_product 两个数据表根据"product_no"列进行合并,生成 df_merge 数据表。

3.将 df_merge 和 df_seller 两个数据表根据"seller_no"列进行合并，更新 df_merge 数据表。

4.将 df_merge 和 df_warehouse 两个数据表根据"warehouse_no"列进行合并，最终得到 df_merge 数据表。

5.从 df_merge 数据表中选择包含以下特征列：'seller_no', 'product_no', 'warehouse_no', 'category1', 'category2', 'category3', 'seller_category','inventory_category','seller_level','warehouse _category', 'warehouse _region', 'qty'，并赋值给 df_feature 数据表。

6.对于指定的分类特征列（cat_cols），使用 LabelEncoder 对象进行数值编码，并更新 df_feature 数据表中的相应列。

7.创建一个 MinMaxScaler 对象 scaler，用于将数据进行归一化处理。

8.从 df_feature 数据表中删除'seller_no'、'product_no'和'warehouse_no'列，并将结果赋值给 df_feature_2 数据表。

9.使用 scaler 对 df_feature_2 进行归一化处理，得到归一化后的数据 df_feature_normalized

然后，我们创建一个空列表 sse_scores，用于存储不同 K 值下的 SSE 得分并定义待测试的 K 值范围 k_values 和最佳 K 值和随机种子变量 best_k、best_random_state 和 best_sse。我们循环遍历不同的 K 值和随机种子，使用 KMeans 模型进行聚类和计算 SSE 得分，并比较得分，更新获得最佳的 K 值和随机种子。最后，我们可视化最佳 K 值和随机种子下的聚类结果。对 df_feature_normalized 应用 PCA 进行降维处理，得到降至 3 维的 df_feature_3d 数据。调用 scatter 函数，在三维图形上绘制散点图，颜色通过 labels 进行映射。

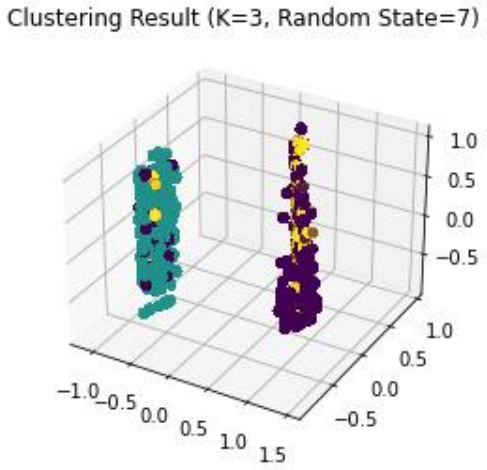


图 7 总数据的聚类结果可视化

得到数据如下：

category1	category2	category3	seller_category	......	cluster
0.421052632	0.360655738	0.354066986	0.4		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0
0.368421053	0.098360656	0.129186603	0.533333333		0

......

我们发现，相似类别的商品，商家，仓库往往具有相似的特征，据此，我们可以根据聚类结果表附件 5.4.1 对他们进行分类，使同一类别在需求上的特征最为相似。

总分类簇结果.xlsx

附件 5.4.1 聚类结果表

六、问题二的模型建立与求解

6.1、问题二的回顾

针对新出现的商家+仓库+商品维度进行需求预测，本文拟采用以下思路。

（1）识别出历史数据中不存在的新维度，对数据进行预处理，是时间规范化。

（2）因此需要对比附件中的各个序列，查找历史数据中与之相似的时间序列，计算其相似度。

（3）相似性可以通过多种方式度量，例如欧氏距离，相似性矩阵，相关系数等。

（4）找到最相似的历史序列，用作预测新维度的参考

（5）通过选定的历史序列，使用时间序列模型对应时间段的序列进行训练和预测，可视化相应的相关系数指标。

6.2、模型的选择与建立

时间序列有多种不同的模型，本题需要根据附件一中的各个维度预测附件五中的数据，可以根据计算的相似度，使用不同的模型进行训练。

线性回归模型

回归分析是一种统计学方法，用于研究自变量和因变量之间的关系。它是一种建立关系模型的方法，可以帮助我们预测和解释变量之间的相互作用我们将所有的特征放在向量中并将所有权重放在向量 w ∈ R^d 中，d 是特征维度，并用 y 来表示模型的预测结果，从而我们可以用点积形式来简洁地表达模型.同时在线性回归模型中，成本函数（Cost function）通常采用最小二乘法（Least Square Method）来定义。

$$\hat{y} = \boldsymbol{w}^T \boldsymbol{x} + b.$$

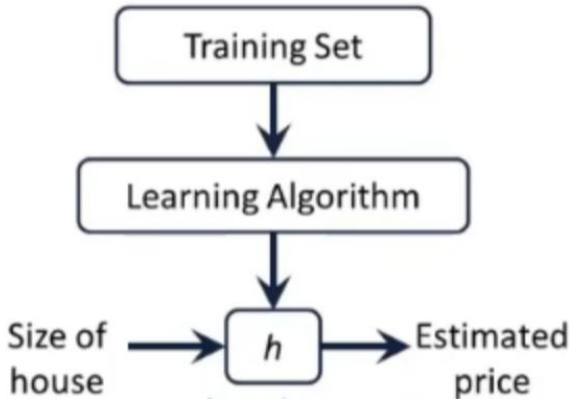


图 8 线性回归的预测模型流程

成本函数就是所有训练样本的预测值与实际值之间的误差平方和。

$$J(w,b)=\frac{1}{2m}\sum_{i=0}^{m-1}(f_{w,b}(x^{(i)})-y^{(i)})^2$$

ARMIMA 模型

ARIMA 模型全称为自回归差分移动平均模型（Autoregressive Integrated Moving Average Model）。ARIMA 模型主要由三部分构成，分别为自回归模型（AR）、差分过程（I）和移动平均模型（MA）。ARIMA 模型的公式可以表示为：

$$Y_t=c+\varphi_1 Y_{t-1}+\varphi_2 Y_{t-2}+...+\varphi_p Y_{t-p}+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+...+\theta_q\epsilon_{t-q}+\epsilon_t$$

在本题中可以根据数据的相似性预测新维度的需求，而不需要依赖完全没有的新维度。

6.3、模型求解

6.3.1 数据预处理

将附件中一和五中的日期规格化，并进行排序，使用 grouby 函数对两个 dataframe 按照 seller_no、warehouse_no、product_no 分组，得到每个组内的数据。

6.3.2 相似度计算

相似度有多种度量标准，比如欧氏距离、余弦相似度、jaccard 相似性系数等，在本题中，我们使用余弦相似度和欧式距离相结合的方式计算时间序列的相似性。

余弦相似度计算

余弦相似度就是通过一个向量空间中两个向量夹角的余弦值作为衡量两个个体之间差异的大小。公式如下：

$$sim(i,j) = \cos(i,j) = \frac{i \cdot j}{||i|| \cdot ||j||}$$

one-hot 编码

one-hot 编码又称独热码，该编码使用 n 位状态寄存器对 n 个状态进行编码，而且只有一个比特位为 1，其他位全为 0。本题中我们根据两个附件中相应的列标签,生成相应的向量，计算文本的相似度。

生成相似度矩阵

根据两个附件中对应商品的特征向量，生成相似度矩阵，再转化为数组形式，提取平均相似度。

表 3：相似度矩阵

	样本 j 属性取值为 1	样本 j 属性取值为 0	属性总个数
样本 i 属性取值为 1	a	b	a + b
样本 i 属性取值为 0	c	d	c + d
属性总个数	a + c	b + d	p

欧氏距离计算

欧氏距离是最常见的两点之间或多点之间的距离表示法，又称之为欧几里得度量，它定义于欧几里得空间中，如点 x = (x 1 ，… ，x n) 和 y=(y1,……,yn)之间的距离为

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

二维平面上两点 a(x1,y1)与 b(x2,y2)间和两个 n 维向量的欧氏距离：

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$d_{12} = \sqrt{\sum_{k=1}^{n} (x_{1k} - x_{2k})^2}$$

本题中可以使用欧氏距离计算两个附件中的进货量的相似度，和余弦相似度一起识别最为相关的序列。

6.3.3 序列识别与输出

结合欧氏距离和余弦相似度，使用循环遍历序列，提取相似度最高的序列值作为接下来使用模型预测的数据。

	seller_no_df1	warehouse_no_df1	product_no_df1	seller_no_df5	\
0	seller_10	wh_1	product_1914	seller_1	
1	seller_10	wh_1	product_1914	seller_1	
2	seller_10	wh_1	product_1914	seller_1	
3	seller_10	wh_1	product_1914	seller_1	
4	seller_10	wh_1	product_1919	seller_1	

	warehouse_no_df5	product_no_df5
0	wh_1	product_2073
1	wh_1	product_2073
2	wh_1	product_2073
3	wh_1	product_2073
4	wh_1	product_2073

部分序列结果

6.3.4 模型应用求解

ARIMA 模型

基于问题一，可得一阶差分过后序列呈平稳分布，因此可以沿用该模型来对相关的时间序列进行预测，而模型的参数会影响到结果的分布。使用自相关系数(ACF)和偏自

相关系数(PACF)计算当前时间点上的观测值与历史时间点观测值之间的相关性。对于任意的滞后（lag）k，我们都计算出在时间 t 和时间 t+k 的数据点之间的协方差，然后除以该时间序列的方差。

$$\rho(k) = Cov(X_t, X_{t+k})/Var(X_t)$$

自相关系数

表 4：部分预测结果

	seller_no	product_no	warehouse_no	date	forecast_qty
0	seller_1	wh_1	product_2073	2023-05-16	1.217873
1	seller_1	wh_1	product_2073	2023-05-17	1.219666
2	seller_1	wh_1	product_2073	2023-05-18	1.219598
3	seller_1	wh_1	product_2073	2023-05-19	1.219601
4	seller_1	wh_1	product_2073	2023-05-20	1.219601

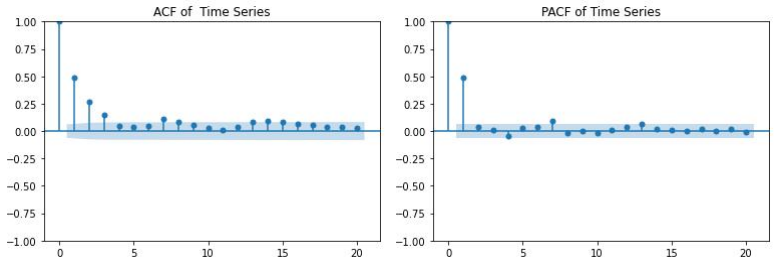


图 9：序列 ACF 与 PACF 值

线性回归模型

由于部分序列的非线性关系和季节性，线性回归无法捕捉所有的模式，存在欠拟合的现象，只有部分数据表现出较好的线性预测结果。

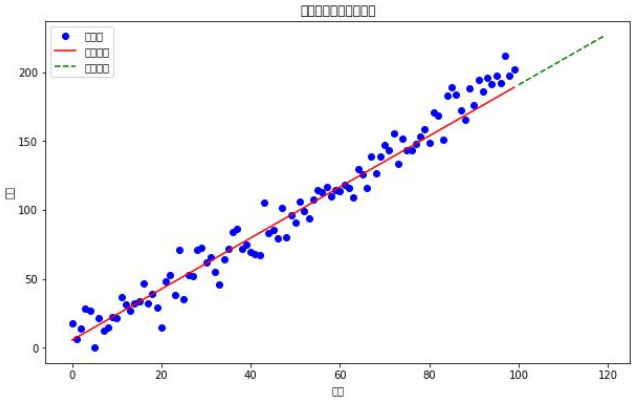


图 10：线性回归预测结果

七、问题三的模型建立与求解

7.1、问题三的前置工作

通过阅读题干可知此题得主要目标是预测 2023 年 6 月 1 日至 2023 年 6 月 20 日促销季的需求量。基于预测目标“促销季”这一属性，本题同时需要观测 2022 年双十一**促销**期间附件 1 中所记载商家的交易情况（附件 6）。同时由于促销的**时间节点**较为接近 2023 年 5 月中下旬且考虑 5 月份出现了一些新的“商家+仓库+商品”维度，因此在解决本问题中也需要一定程度参考问题 2 的预测结果。

根据以上分析结果，本问题首先读取附件 1，2，3，4，6 以及问题 2 的预测结果。在将所读取数据日期转化为 datatime 形式后合并历史需求量数据。至此，本问题的数据预处理工作基本完成。

表 5：合并后历史需求量数据

	seller_no	product_no	warehouse_no	date	qty
0	seller_19	product_448	wh_30	2023-05-09	10.0
1	seller_19	product_448	wh_30	2023-04-17	14.0
2	seller_19	product_448	wh_30	2023-01-09	2.0
3	seller_19	product_448	wh_30	2023-01-20	1.0
4	seller_19	product_448	wh_30	2023-02-13	22.0
...	...	...	...	...	...
355836	seller_9	wh_15	product_2122	2023-05-26	0.0
355837	seller_9	wh_15	product_2122	2023-05-27	0.0
355838	seller_9	wh_15	product_2122	2023-05-28	0.0
355839	seller_9	wh_15	product_2122	2023-05-29	0.0
355840	seller_9	wh_15	product_2122	2023-05-30	0.0

355841 rows × 5 columns

7.2、记住需求量计算与特征信息提取

为获取商品的基准需求量，在进行数据预处理后本题采用按照“日期+商家编号+商品编号+仓库编号“的方式对数据进行分组，并计算出分组后平均值作为基准需求量。

由于本题仍然存在按照“商家+仓库+商品”这一维度进行预测这一根本要求，因此特征信息的提取仍围绕“商家+仓库+商品”维度。首先将先前读取的“历史需求”与“产品信息”进行合并作为第一张“合并表”随后用该“合并表”与“商家信息”合并生成新合并表，最终在合并所有所读取的信息（产品信息，商家信息与仓库信息，分别对应最开始读取的附件 2-附件 4）后生成特征信息总表。

表 6：特征信息总表（节选）

	seller_no	product_no	warehouse_no	date	qty	demand	category1	category2	category3	seller_category	inventory_category	seller_level	warehous
0	seller_19	product_448	wh_30	2023-05-09	10.0	10.0	手机通讯	手机配件	手机配件_12	数码	C	Large	
1	seller_19	product_448	wh_30	2023-04-17	14.0	14.0	手机通讯	手机配件	手机配件_12	数码	C	Large	
2	seller_19	product_448	wh_30	2023-01-09	2.0	2.0	手机通讯	手机配件	手机配件_12	数码	C	Large	
3	seller_19	product_448	wh_30	2023-01-20	1.0	1.0	手机通讯	手机配件	手机配件_12	数码	C	Large	
4	seller_19	product_448	wh_30	2023-02-13	22.0	22.0	手机通讯	手机配件	手机配件_12	数码	C	Large	
...	...	...	...	...	...	...	...	...	...	...	...	...	

生成特征信息总表后，使得商家、商品与仓库的属性（比如表中现实的 category1-category3 为产品的三级分类）更为直观，以此希望能够提高预测的精确程度。

7.3、建立指数平滑模型进行预测

7.3.1、模型原理

指数平滑模型的一大核心思想即认为时间序列的态势具有稳定性或规则性，所以时间序列可被合理地顺势推延。在模型的具体操作中，预测值为先前观测值的加权和（不同数据，不同权重），新数据的权重往往高于旧数据。同时根据序列的趋势性和季节性，该模型被细分为一次指数平滑模型（无趋势与季节性）；二次指数平滑模型（有趋势但无季节性）以及三次次指数平滑模型（序列有趋势也有季节性）。其基本公式如下,其中 S_t 与 y_t 分别为时间 t 的平滑值与实测值 S_{t-1} 为时间为 t-1 时的平滑值，而 a 为平滑常数（在[0,1]之间取值）。

$$S_t = a * y_t + (1-a)S_{t-1}$$

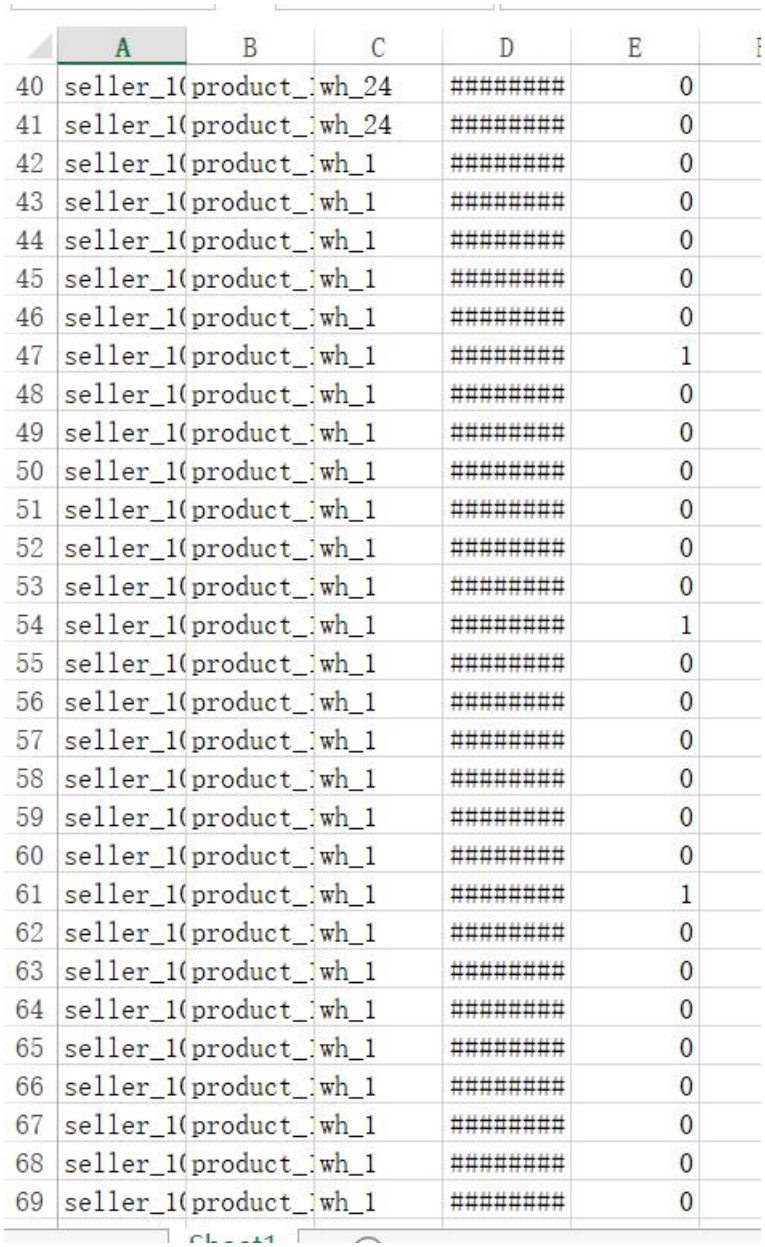
通过先前获取到相关描述性统计信息可知，促销期间的数据具有趋势与季节性，因此本题最终选择三次指数平滑模型（累加法）。预测模型如下。

$$\hat{y}_{t+T}=a_t+b_tT+c_tT^2$$
$$a_t=3S_t^{(1)}-3S_t^{(2)}+S_t^{(3)}$$
$$b_t=(\alpha/(2(1-\alpha))[(6-5\alpha)S_t^{(1)}-2(5-4\alpha)S_t^{(2)}+(4-3\alpha)S_t^{(3)}]$$
$$c_t=(\alpha^2/(2(1-\alpha)^2)[S_t^{(1)}-S_t^{(2)}+2S_t^{(3)}]$$

7.3.2 建立与求解

提取合并特征信息总表中的日期，商家,产品与仓库属性并以此分类。以分出的类遍历数据集。遍历期间需确保按日期升序排列。若发生日期重复的情况，则采取将重复的日期分组求平均值的方式处理。处理完成即可直接调用 ExponentialSmoothing 模块进行预测（趋势、季节均为累加）。

在循环完成后，由于出现部分预测结果小于零的情况，因此将值为复数的预测结果全部赋值为 0 后方可写入结果表。



	A	B	C	D	E
40	seller_1(	product_1	wh_24	########	0
41	seller_1(	product_1	wh_24	########	0
42	seller_1(	product_1	wh_1	########	0
43	seller_1(	product_1	wh_1	########	0
44	seller_1(	product_1	wh_1	########	0
45	seller_1(	product_1	wh_1	########	0
46	seller_1(	product_1	wh_1	########	0
47	seller_1(	product_1	wh_1	########	1
48	seller_1(	product_1	wh_1	########	0
49	seller_1(	product_1	wh_1	########	0
50	seller_1(	product_1	wh_1	########	0
51	seller_1(	product_1	wh_1	########	0
52	seller_1(	product_1	wh_1	########	0
53	seller_1(	product_1	wh_1	########	0
54	seller_1(	product_1	wh_1	########	1
55	seller_1(	product_1	wh_1	########	0
56	seller_1(	product_1	wh_1	########	0
57	seller_1(	product_1	wh_1	########	0
58	seller_1(	product_1	wh_1	########	0
59	seller_1(	product_1	wh_1	########	0
60	seller_1(	product_1	wh_1	########	0
61	seller_1(	product_1	wh_1	########	1
62	seller_1(	product_1	wh_1	########	0
63	seller_1(	product_1	wh_1	########	0
64	seller_1(	product_1	wh_1	########	0
65	seller_1(	product_1	wh_1	########	0
66	seller_1(	product_1	wh_1	########	0
67	seller_1(	product_1	wh_1	########	0
68	seller_1(	product_1	wh_1	########	0
69	seller_1(	product_1	wh_1	########	0

图 ：预测结果节选（过于稀疏）

八、结束语

本问题致力于对时间序列的研究，从根据一段时间的情况预测未来走势，根据相似货物走势预测新货物走势，根据已知货物走势和新的货物走势来预测特殊时期货物的售卖情况三个方面，体现了对数据在一段时间上的规律的研究的重要性。

我们在众多模型总对比试验，最终选择了 arima 模型来完成这个任务，并且，我们通过一系列测试来最终确定最适合的参数和距离计算方法。并在问题三中完成最终的预测。

感谢 mathercup 杯的工作人员精心设计了这一赛题，严格把关，给了我们一个在解决问题中超越自我的机会，感谢指导老师的关切和指导，促成了我们这一次比赛的圆满完成，感谢各位队员不分日夜的奋斗，终究使汗水凝结成瑰丽的成果。谢谢大家

参考文献：

[1] 智能供应链:预测算法理论与实战[M]，北京：电子工业出版社，2023. [2] Makridakis S , Spiliotis E , Assimakopoulos V .The M5 Accuracy competition: Results, findings and conclusions[J].International Journal of Forecasting, 2020, 36(1):224-227. [3] Makridakis S , Spiliotis E , Assimakopoulos V ,et al.The M4 Competition: 100,000 time series and 61 forecasting methods[J].International Journal of Forecasting, 2020, 36(1):54-74..
[2] Makridakis S , Spiliotis E , Assimakopoulos V .The M5 Accuracy competition: Results, findings and conclusions[J].International Journal of Forecasting, 2020, 36(1):224-227.
[3] Makridakis S , Spiliotis E , Assimakopoulos V ,et al.The M4 Competition: 100,000 time series and 61 forecasting methods[J].International Journal of Forecasting, 2020, 36(1):54-74..
李志超， 刘 升． 基于 ＡＲＩＭＡ 模型、 灰色模型和回归模型
的预测比较 ［Ｊ］． 统计与决策， ２０１９， ３５（２３）： ３８－４１．
［3］ 蔺富明， 史代敏． 整值自回归模型下的单位根检验可靠性研
究 ［Ｊ］． 统计与决策， ２０２０， ３６（１２）： ４４－４９．
［4］ 张 莹， 王立洪． 基于残差的非线性自回归模型的拟合优度检验
［Ｊ］． 南京大学学报（数学半年刊）， ２０１２， ２９（１）： ９３－１０４．
［5］ 曾 艳， 郝志峰， 蔡瑞初， 等． 基于时序隐变量模型的因果

关系发现算法 ［ Ｊ］． 计算机工程与设计， ２０２２， ４３ （ ５ ）：
１４２８－１４３４．
［6］ 胡青松， 钱建生， 李世银， 等． 基于自回归移动平均算法的
产能预 测 方 法 研 究 ［ Ｊ］． 中 国 石 油 和 化 工 标 准 与 质 量，
２０２１， ４１（２３）： １３１－１３２．
［7］ 王 慧， 吴茜茜． 空间自回归模型中系数变量及误差项的贝
叶斯估计 ［Ｊ］． 合肥工业大学学报（自然科学版）， ２０２１， ４４
（９）： １２９１－１２９６．
［8］ 刘 涛， 杨炜明， 胡瑞婷． 基于 ＡＩＣ 准则的混频数据季 ＧＤＰ
预测实证研究 ［Ｊ］． 统计理论与实践， ２０２１， （６）： ２６－３３．
［9］ 王振龙， 胡永宏． 应用时间序列分析 ［Ｍ］． 北京： 科学出版
社， ２００８．
［10］ 聂淑媛． 沃尔德与离散平稳时间序列 ［ Ｊ］． 咸阳师范学院学
报， ２０１２， ２７（２）： ７２－７５．
［11］ 傅惠民， 刘成瑞， 马小兵． 时间序列均值和方差函数的确定
方法 ［Ｊ］． 机械强度， ２００４， ２６（２）： １６４－１６９