英国礼品销售公司的客户与商品分析：数据驱动的运营改进策略

目录

附录

1. 包含全部代码的pdf格式分析报告：Project White.pdf
2. 包含全部代码的原始Jupyter Notebook：Project White.ipynb
3. SQL语句：SQL statements.pdf
4. 基础数据集：basic_data.xlsx

一、分析目标

（一）分析概述

我们以记录某英国礼品销售公司两年内全部交易的真实数据集作为分析起点，以通过分析数据发现业务运营中的可改进方面后提出建议作为分析目标；在完成基础性数据整理工作后，我们通过探索性数据分析确定分析方向为客户分析和商品分析。各项分析的结论和可能的改进方面如下：

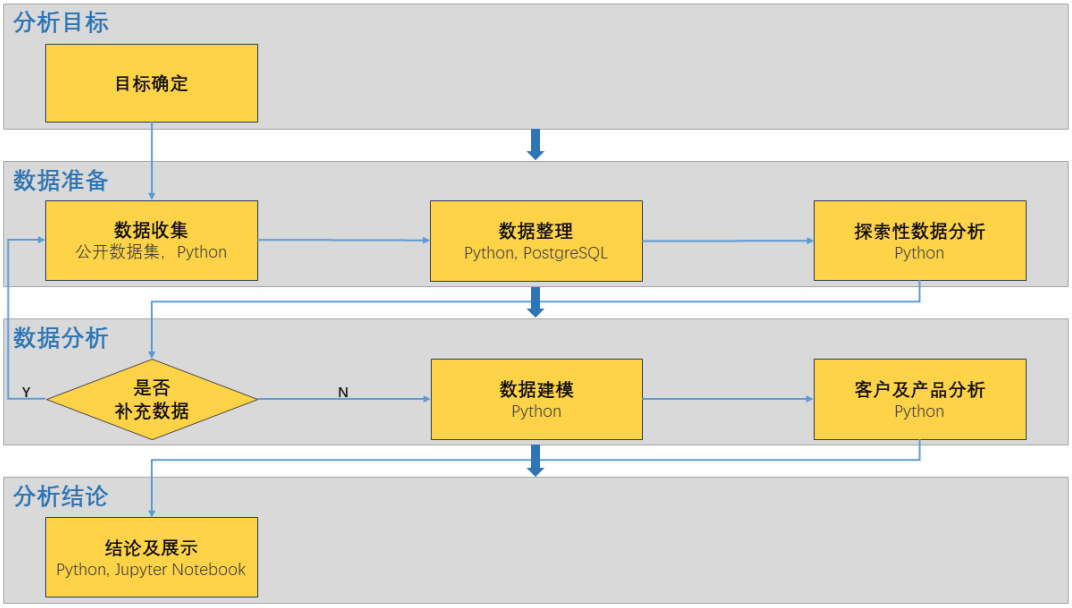
方向	结论	改进方面
客户分析	1.“高价值”客户占比较低； 2.“长期不活跃”客户占比较高； 3.客户留存率随时间推移快速下降，客户粘性较低； 4.客户留存率在年末迎来回升后进一步降低； 5.部分月份加入客户的整体留存率较高。	1.抓住新客户加入后的窗口时机进行营销活动和针对性产品推介，提供定制化服务，提高“高价值”客户的转化率； 2.利用年末销售高峰通过关联产品推介和优惠活动吸引已有客户回流和新客户加入； 3.对不活跃客户进行回访，了解其退出原因，通过优惠活动或新产品推介鼓励回流； 4.利用客户资源和物流资源开拓新的业务线，提高业绩稳定性和资源利用率； 5.了解部分月份客户留存率更高的原因，推广成功经验。
商品分析	1.商品销售集中度低，商品品类覆盖度高，商品价格优势明显； 2.商品销售趋势均全年保持大体稳定，在年末迎来销售高峰； 3.高收入商品（收入高）之间销售额存在明显差异，但未形成高差异化； 4.热销商品（销量大）之间销售量差异不显著； 5.商品销售存在明显季节性，使用专用模型预测了商品未来销售量。	1.加强商品开发。开发高价值商品，淘汰低利润商品和调整定价策略； 2.在营销活动中突出高收入商品的独特价值，并通过关联推荐提高热销商品的总体销量； 3.关注季节性的行业特点，在销售淡季利用客户资源和物流资源开发其他产品线； 4.持续更新销售预测模型，根据预测结果动态调整库存管理和营销策略。

（二）目标确定

本次分析的目标为使用真实商业数据发掘包含信息，建立信息间关系，发现业务运营中的可改进方面后提出建议。

由于数据集规模较大，我们采用PostgreSQL存储数据，使用Python中的数据整理，数据分析和可视化工具完成分析，并使用Jupyter Notebook展示。

本次分析的总体框架如下：



二、数据收集

本次分析使用的基础数据集来自加州大学尔湾分校机器学习知识库，在说明来源的情况下，该网站允许将该数据集用于任何用途。数据来源为一家英国礼品销售公司，主营业务为向批发商销售礼品及礼品类商品。数据集内容为该公司2009年12月1日至2011年12月9日之间的全部107万笔交易订单的基本信息。

数据集结构（整理前）：

Structure of the dataset(before sorting)

	INVOICE	STOCKCODE	DESCRIPTION	QUANTITY	INVOICEDATE	PRICE	CUSTOMER ID	COUNTRY
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom

在收集数据进行基础整理后，我们将数据集保存至PostgreSQL数据库中，并在数据集中建立invoice_amount列，设置其为quantity和price的乘积。

三、数据整理

我们对数据结构的了解表明：

1. 数据集共计有1,067,371条记录，包含订单编码，商品编码，商品描述，销售数量，订单日期，商品单价，客户编码和客户国别等信息；
2. 部分信息存在内容缺失，如description和customer_id，但大部分信息内容完整；
3. 部分信息重复内容较多，可以优化数据类型以节省内存或完善数据结构，如invoicedate，customerid和country；
4. 部分信息的特征表明存在特殊交易，需对这些交易的实质进行了解，确定是否需要从数据集中排除，如invoice，stockcode和quantity；
5. 部分信息命名规则不统一，应当修改以避免理解歧义，如customer_id和invoice_amount。

针对上述情况，我们通过处理缺失值、数据类型转换、处理异常值和其他处理等步骤完成数据整理：

1. 对description和customer_id进行数据填充；
2. 对invoicedate，customerid和country进行数据类型转换，分别转换为datetime格式，int格式和category格式；
3. 对于特殊交易的总体原则为了解其实质在具体分析中按需求处理。这些特殊交易包括：
    a. invoice中特殊记录为字母A开头订单和字母C开头订单，分别对应手工调整坏账订单和取消订单。其中调整坏账订单应当被移除；
    b. stockcode中特殊记录为与客户购买行为无关的订单，这些记录应当被移除；
    c. description中的空白记录不对应实际销售，应当被移除；
    d. quantity为负的记录为取消订单和因各种原因无法完成的订单，在进一步分析中具体考虑；
    e. price为零的记录为因各种原因无法完成的订单，应当被移除；
    f. quantity和price中显著偏离其他数据的异常记录影响后续分析，应当被移除；
    g. customerid为空的记录是正常交易记录，在进一步分析中具体考虑。
4. 对customer_id和invoice_amount列重命名。

数据集结构（整理后）：

Structure of the dataset(after sorting)

	INVOICE	STOCKCODE	DESCRIPTION	QUANTITY	INVOICEDATE	PRICE	CUSTOMERID	COUNTRY	INVOICEAMO
0	500108	22028	PENNY FARTHING BIRTHDAY CARD	12	2010-03-04	0.42	15358	United Kingdom	5.04
1	500133	21260	FIRST AID TIN	12	2010-03-04	3.25	16329	United Kingdom	39.00
2	500149	15039	SANDALWOOD FAN	10	2010-03-04	0.85	17931	United Kingdom	8.50
3	500232	20713	JUMBO BAG OWLS	200	2010-03-05	1.65	15769	United Kingdom	330.00

	INVOICE	STOCKCODE	DESCRIPTION	QUANTITY	INVOICEDATE	PRICE	CUSTOMERID	COUNTRY	INVOICEAMO
4	500356	21903	MAN FLU METAL SIGN	2	2010-03-07	2.10	16984	United Kingdom	4.20

四、探索性数据分析

通过对信息进行描述性统计和可视化分析，并汇总前述已获取信息，我们得到对数据集体现的信息间关系的理解，由此确定具体分析方向和方法。

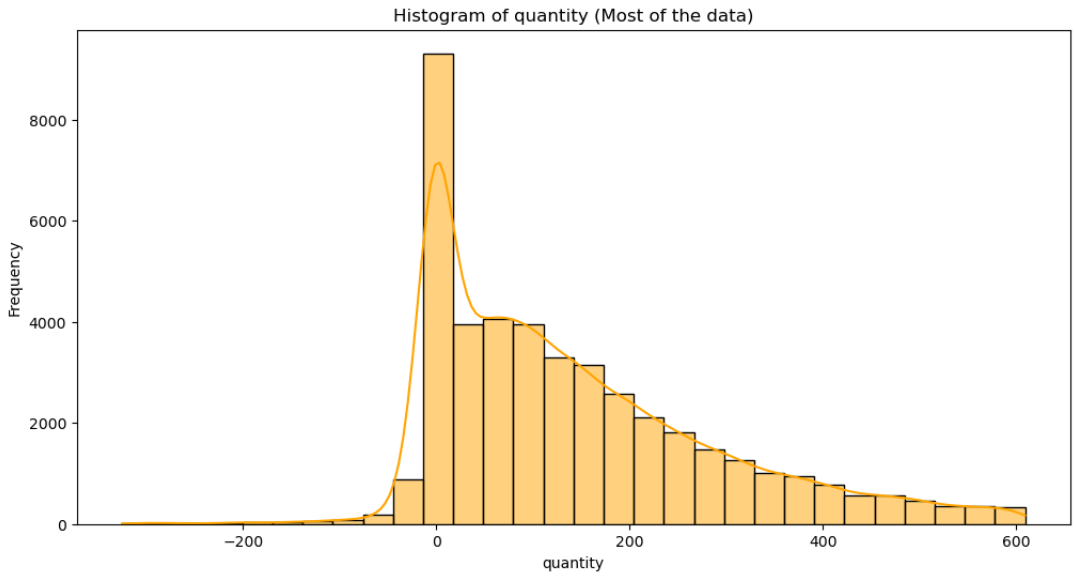
（一）描述性统计和可视化分析

在对数据集按照订单进行汇总后，我们对以数字类型格式存储的信息（quantity，price和invoiceamount）进行描述性统计分析，程序包括了解数据特征，计算数据分布，数据集中度可视化和数据分布可视化。

以数字类型存储数据的特征：

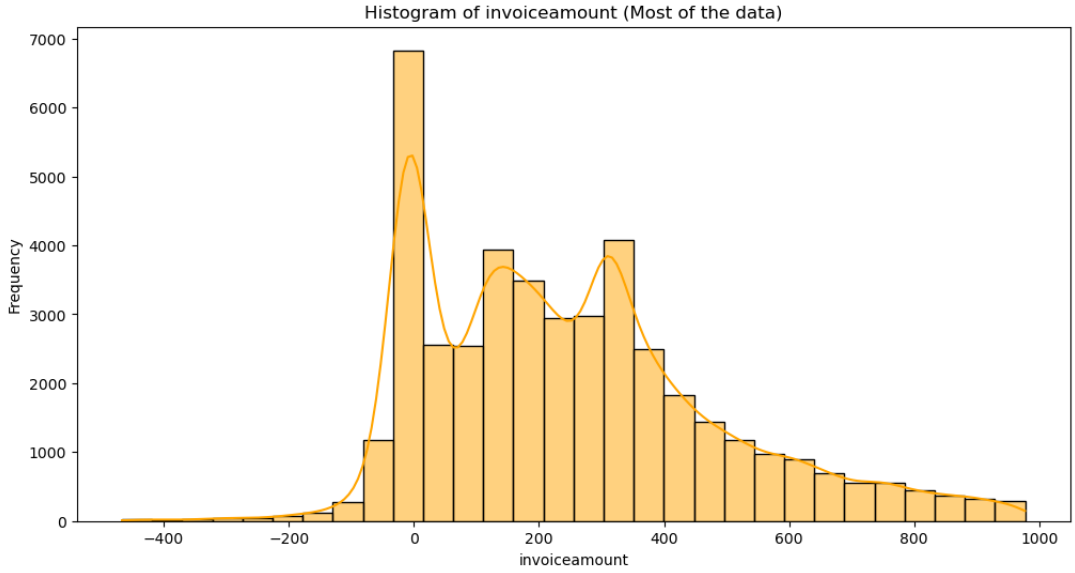
```
       quantity    price  invoiceamount
count  46946.00 46946.00       46946.00
mean     232.63     3.25         413.22
std     1078.81    11.67        1100.21
min   -87167.00     0.03      -22998.40
25%      24.00     1.47          74.10
50%     117.00     2.05         242.30
75%     258.00     3.19         435.45
max   87167.00  1157.15       52940.94
```

quantity分布：

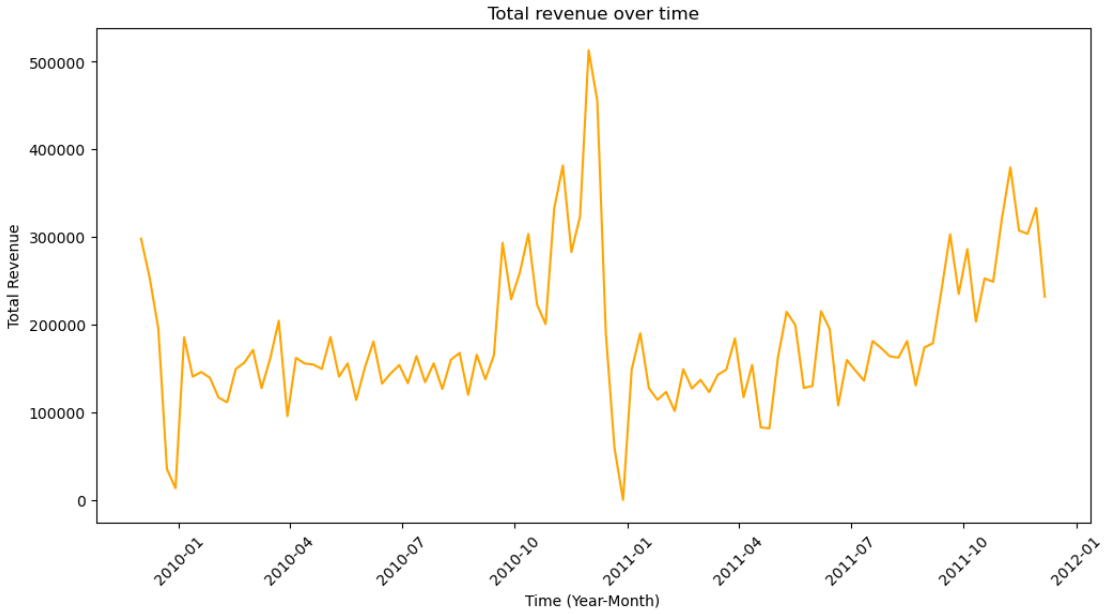


price分布：

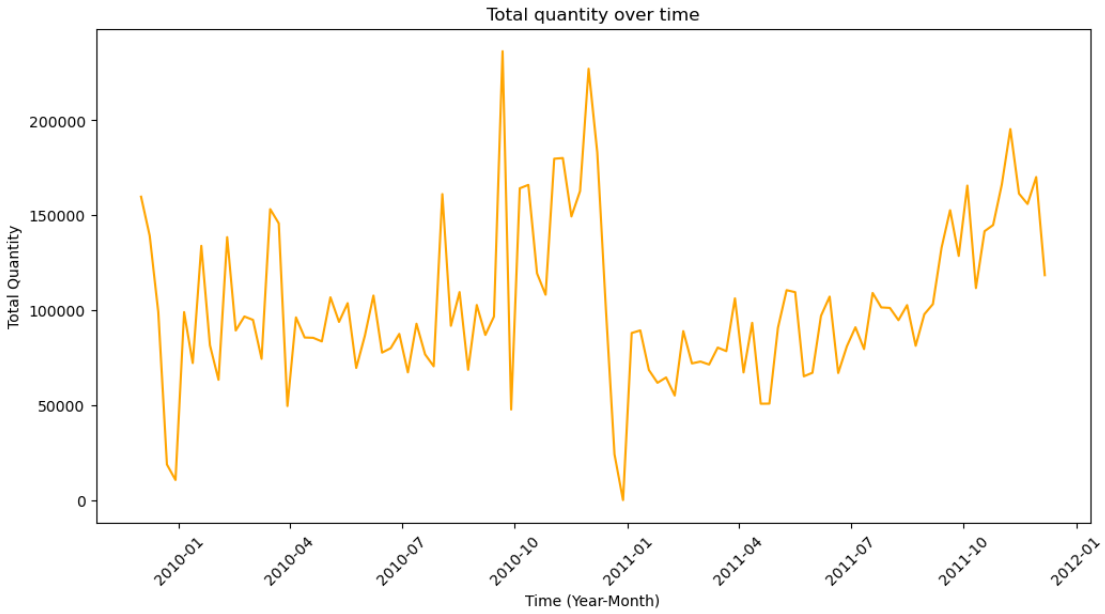
invoiceamount分布：



总销售额变动趋势：



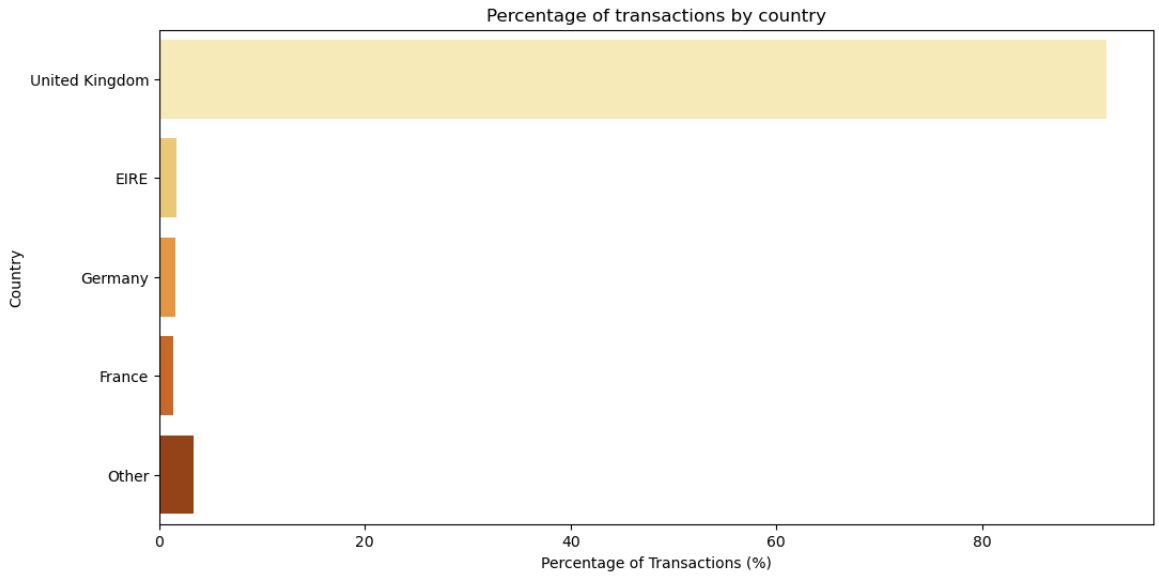
总销售量变动趋势：



对quantity，price和invoiceamount进行描述性统计发现：

1. 各列数据均呈现明显的右偏分布，大部分数据集中在较小的范围内；
2. quantity中，单笔订单采购量集中于600件以内，单笔订单退货量集中于300件以内，且总体来看采购量明显大于退货量；
3. price中，单件商品价格总体较低，集中于6英镑以内；
4. invoiceamount中，单笔订单金额集中于1000英镑以内；
5. quantity和invoiceamount的总体变动趋势类似，均在每年年末出现高峰，其他期间总体变动保持平稳。

客户国别分布：



我们对其他主要信息（country）进行可视化分析，发现来自于英国本土的订单占总订单数90%以上，表明销售的国别集中度极高。

以上发现综合表明，公司销售商品以销往英国本土的低价商品为主，且单笔订单销售量较高，商品单价较低。

（二）分析途径确定

公司客户主要为批发商，向他们提供商品服务的重点关注要素包括：

1. 合作关系。包括客户关系和服务质量；
2. 商品覆盖。包括品类丰富性和品种多样性；
3. 价格优势；
4. 供应稳定性。包括库存稳定和物流稳定；
5. 商品质量。

总体来看，这些要素要求我们分别从客户角度和商品角度出发，考虑如何提供更好商品服务。因而我们确定分析途径为客户分析和商品分析。

已获取数据能够提供商品角度和客户角度信息。客户方面，已有数据体现客户的购买行为以及随时间推移客户的活跃程度。对于客户购买行为进行分类和追踪可以有针对性地维护客户关系；对于客户退货和降低购买频率等行为进行回访则有助于改善商品质量和服务质量。商品方面，已有数据体现商品的整体结构和销售趋势。对于商品整体结构进行分析可以了解商品覆盖和价格优势的程度；对于商品销售趋势进行分析则可以通过及时调整库存而提高供应稳定性。

客户角度可能的分析方法包括将客户按照销售行为分组后进行精准营销的客户价值分析，和将客户按照加入日期后分组观察其后续留存率的客户流失预测；商品角度可能的分析方法包括对商品进行销售情况分析，和建立模型通过已有数据预测商品未来销售量的销售趋势预测。

（三）数据补充

上述分析方法所需数据绝大部分由基础数据集提供，但销售趋势预测中对于季节性因素的考虑可能需要公共假期信息，因而通过Python收集2009年12月1日至2011年12月9日之间主要销售地区（英国，爱尔兰，法国，德国）的公共假期信息，由于该部分数据结构简单，收集后即保存至PostgreSQL数据库中。

进一步分析中的销售趋势预测需要自基础数据结束后一年内的公共假期信息，同样使用Python收集并保存至PostgreSQL数据库中。

五、数据建模

根据探索性数据分析的结果，需建立如下模型：

1. RFM模型（客户分析中的客户价值分析）
   RFM模型将每个客户按照R（客户最近一次购买日期与当前日期差值），F（总的购买次数），M（购买的总金额）三个维度分类以识别高价值客户，然后将R值、F值和M值的高低进行分段，并赋予分值，最后将RFM三个分数相加，得到每个客户的RFM总分。针对已有数据集，由于其中包含退货订单，因而我们将数据集分为正常交易和退货交易。正常交易按照前述维度将客户分类后赋分，识别不同类型客户；退货交易则按照高频退货，大额退货和近期退货的维度进行分类后赋分，识别不同类型客户。
2. Cohort模型（客户分析中的客户流失预测）
   Cohort模型将拥有类似特征的客户识别为一个群体，观察其随时间的变动情况。本次分析中，我们将每月新增客户识别为一个群体，然后按月度时间单位分析新增客户的后续购买情况，以在时间推移下了解客户的留存度，进而预测客户的流失。
3. SARIMAX模型（商品分析中的销售趋势预测）
   SARIMAX 模型的作用为根据数据的过去值“解释”给定的具有季节性和外生变量的时间序列。在本次分析中，我们通过模型拟合商品销售的时间序列。由于探索性数据分析阶段已表明商品销售可能存在季节性，并且商品销售可能和节假日相关，因而我们向模型添加季节性周期长度，选择节假日作为外生变量，选择最佳参数以拟合历史销售量，进而预测数据集结束后12个月内的销售量。

六、客户及商品分析

我们通过使用前述的模型分别从客户角度和商品角度对数据集进行分析，为运营活动提供建议。

在各项分析之前，我们均需考虑在数据处理阶段识别的异常交易，包括数量为负的记录和客户编码为空的记录。数量为负的记录产生于客户取消订单，包含客户价值和销售情况信息，因而在客户价值分析和销售情况分析中保留，客户流失预测和销售趋势预测中剔除。客户编码为空的记录为未记录客户编码的正常订单，因而在客户分析中剔除，商品分析中保留。

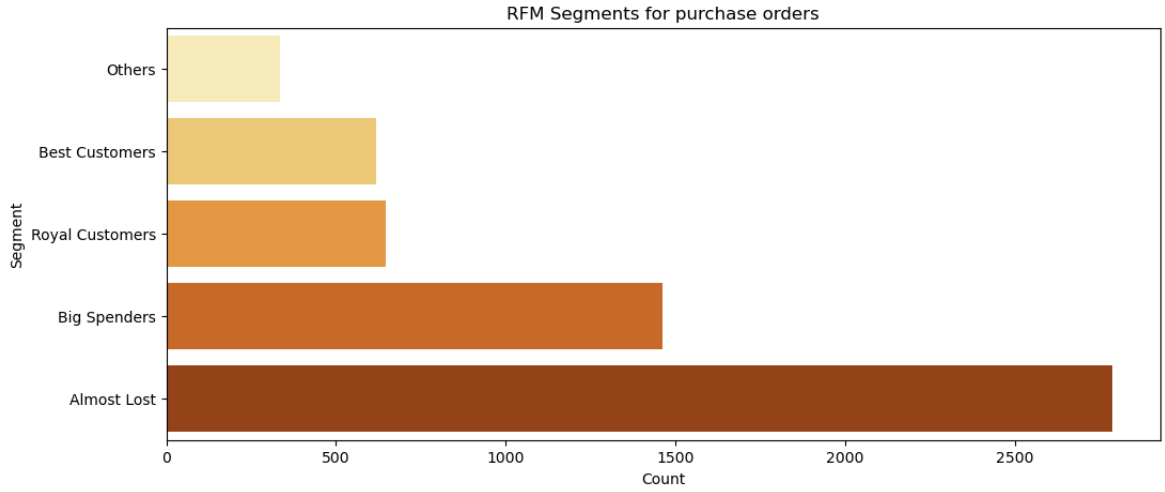
（一）客户分析

1. 客户价值分析

客户价值分析可以为客户关系维护和服务质量提升提供依据。

我们将所有记录分为正常交易和退货交易，正常交易中将R值、F值和M值分为4个分段，退货交易中将值分为2个分段。每个值中，最近购买和退货，购买和退货频率最高或者金额最多的客户的客户得最高分，最久未购买和退货，购买和退货频率最低或者金额最少的客户得最低分。得到三个分数综合考虑后，我们将客户分为不同的群体。

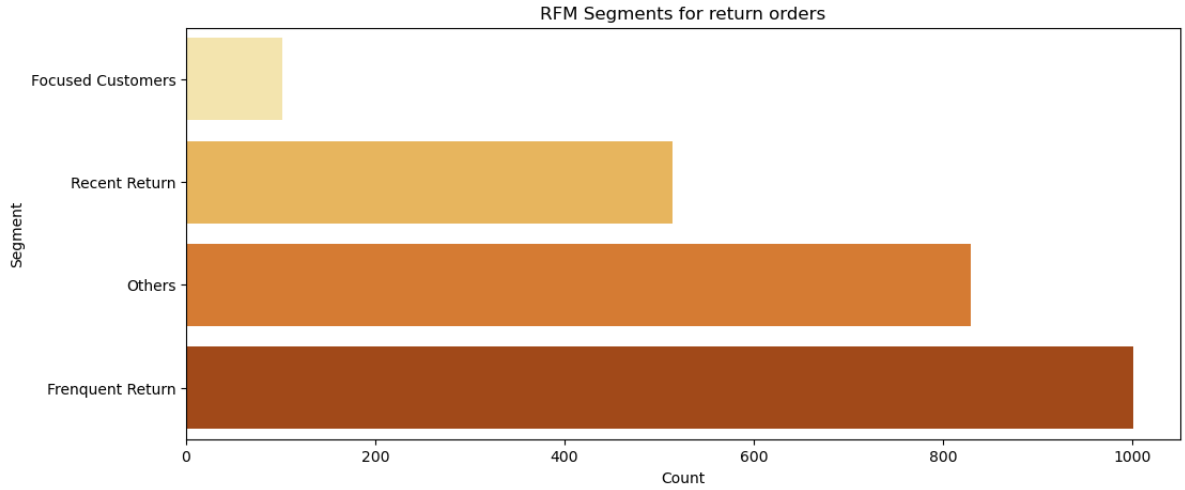
正常交易中客户分布：



在正常交易中，客户群体包括“最佳”客户（RFM均为4分），“潜力”客户（M分较高），“忠诚”客户（F分较高），“需要唤醒”客户（R分较低）等。分析结果表明，正常交易客户中“需要唤醒”客户占比较高，其次是“潜力”客户，而“最佳”客户和“忠诚”客户占比相对较低。“需要唤醒”客户占比较高可能和商品销售中存在的季节性有关，但也反映客户粘性总体较低。

针对“最佳”客户，可以考虑向他们推荐更高附加值的产品，并优先提供服务；针对“潜力”客户，可以通过优惠活动或新产品推介激励他们更频繁地购买；针对“忠诚”客户，可以考虑向他们推广有针对性的产品，并提供个性化服务；针对“需要唤醒”客户，需要通过回访等途径了解客户离开原因，进而挽回这些客户，例如提高商品覆盖，提升服务质量等。针对客户粘性总体较低，公司可以利用客户资源和物流资源开拓新的业务线，提高业绩稳定性和资源利用率。

退货交易中客户分布：



在退货交易中，客户群体包括“关注”客户（RFM均为2分），“高频”客户（F分较高），“近期退货”客户（R分较高）等。分析结果表明，退货交易客户中，“高频”客户占比较高，其次是“近期退货”客户，“关注”客户占比最低。探索性数据分析阶段发现，退货订单占订单总体比例很低，但仍需通过向退货交易客户了解退货原因，以避免对物流资源的占用。

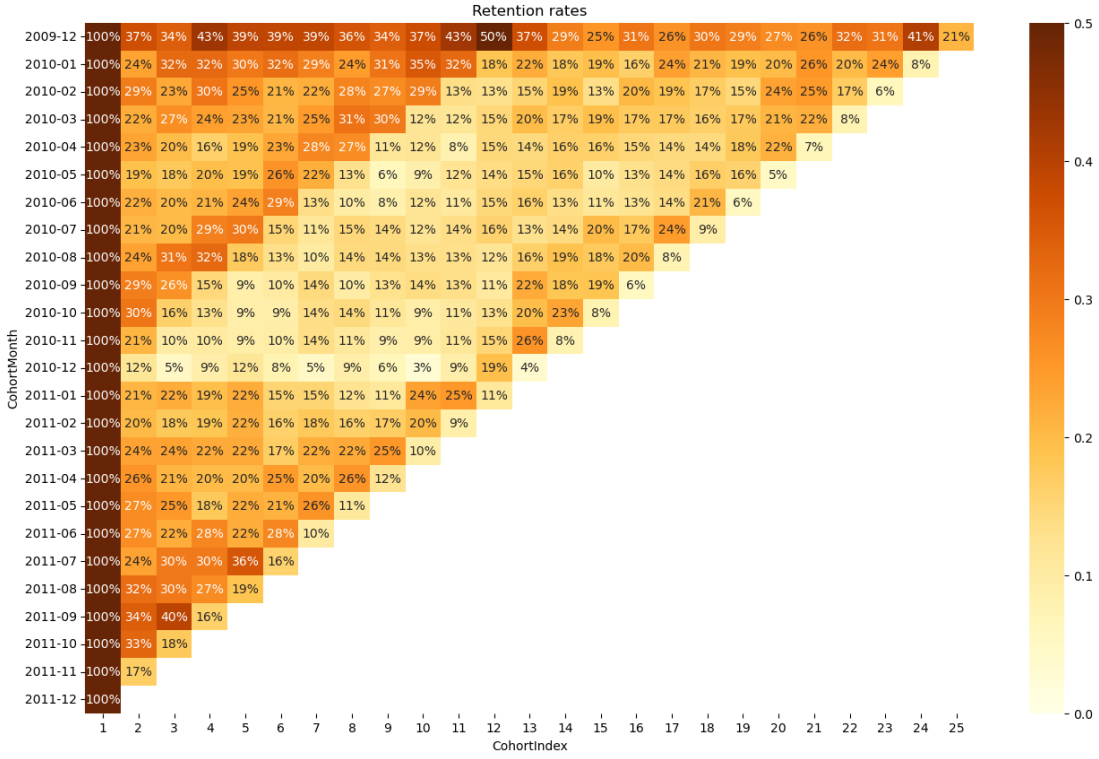
针对“关注”客户，可以考虑向他们了解对产品和服务不满的原因，以改进产品和服务质量；针对"高频"客户，可以考虑向他们提供更有针对性的推荐，避免高频退货；针对“近期退货”客户，可以及时向他们了解退货原因并进行补偿，以避免客户流失。

2. 客户流失预测

客户流失预测可以用来预测“需要唤醒”客户的流失情况，有针对性地提高商品质量和服务质量。

我们将客户按照加入的年份和月份进行分组，分别统计每组客户逐月的留存活跃客户数，并计算留存率。在得到留存率总体矩阵后，我们使用热图展示留存率的变化，其中颜色深浅表示留存率高低。

留存率热图：



分析结果表明，新客户通常在加入后第一年能保持相对较高的留存率，在接近年末时出现留存率的显著提高后，第二年的留存率会显著降低。并且，部分月份加入客户的总体留存率显著高于其他月份。同时，由于每年年末留存率均明显提高，表明商品销售可能存在季节性。

针对上述发现，公司在进行营销活动时，应抓住新客户加入后的高留存率空间，争取将更多客户保持为“最佳”客户。并且，由于在接近年末时留存率有显著提高，这时应当进行针对性营销，将其他类型客户转化为“最佳”客户。图示还表明，部分新加入客户群的留存率显著高于其他时间加入的客户，公司应当对其原因进行分析，将成功经验用于营销活动中。

（二）商品分析

1. 销售情况分析

销售情况分析主要是为了理解公司的销售组合，了解公司商品覆盖和价格优势的程度。

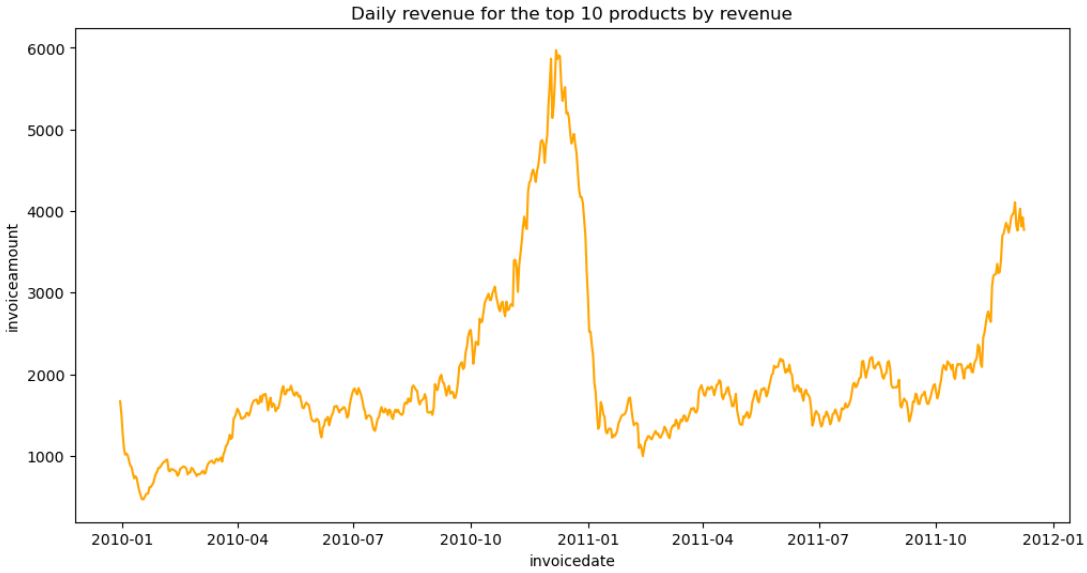
销售额数据特征：

```
count      4917.00
mean       3945.32
std       10659.12
min        -126.00
25%         233.67
50%        1027.14
75%        3361.25
max      327813.65
Name: invoiceamount, dtype: float64
```

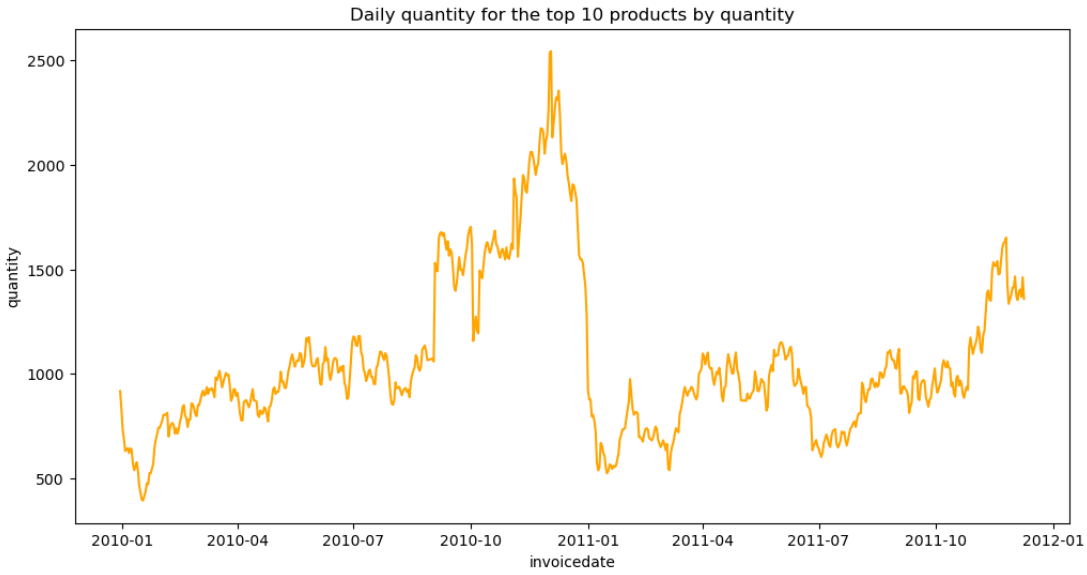
销售量数据特征：

```
count      4917.00
mean       2221.07
std        5581.90
min         -27.00
25%          97.00
50%         565.00
75%        2033.00
max      108434.00
Name: quantity, dtype: float64
```

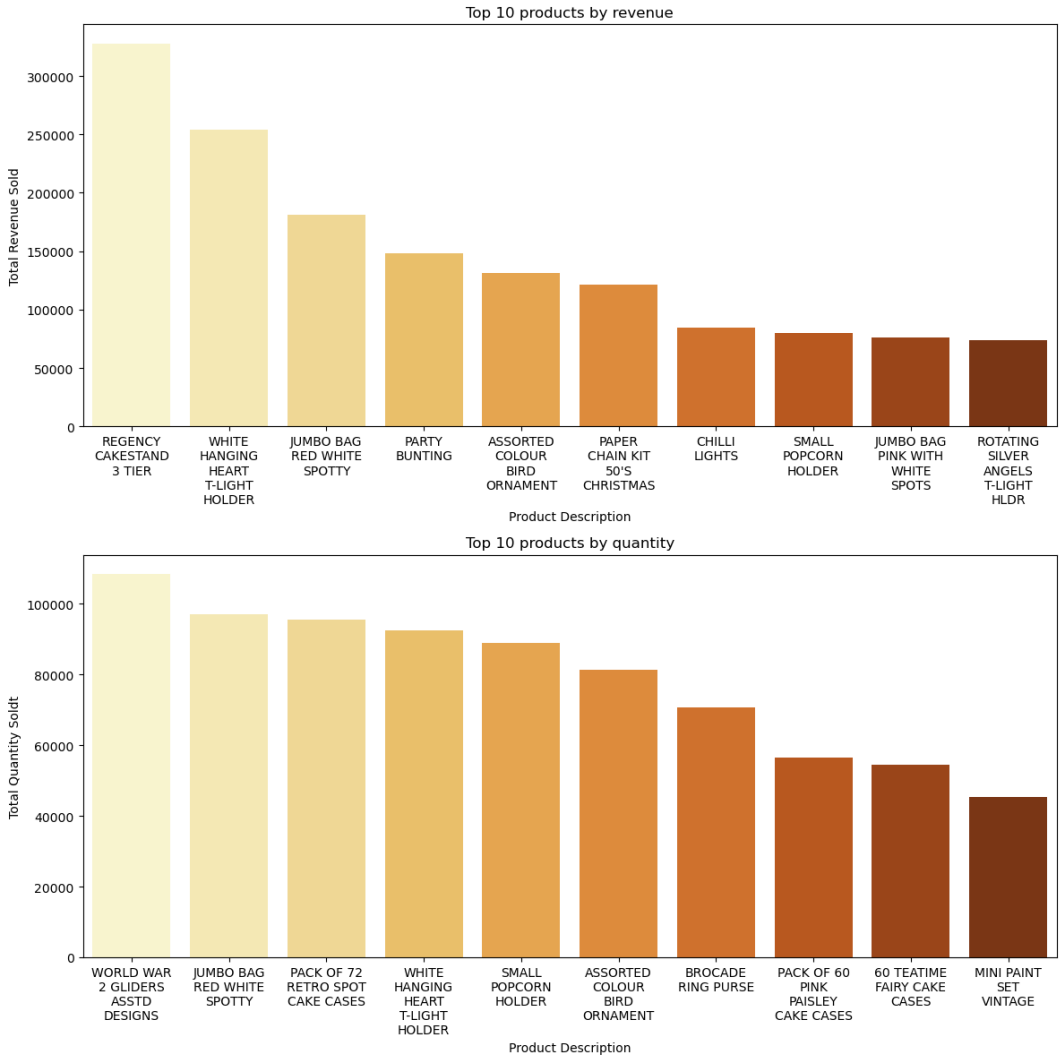
高收入商品销售额变动趋势：



热销商品销售量变动趋势：



分品种高收入商品销售额和热销商品销售量：



对销售情况的分析发现：

1. 商品总品类数为4,917，销售额前十商品（“高收入商品”）占总销售额比例和销售量前十商品（“热销商品”）占总销售量比例均达到7%，表明商品销售品类集中度低，商品覆盖度广；
2. 高收入商品和热销商品重叠品类为4种，表明不同商品间单价差异较大；
3. 高收入商品和热销商品的销售趋势均全年保持大体稳定，并在年末迎来销售高峰；
4. 高收入商品之间销售额存在明显差异，但未形成差异化；
5. 热销商品之间销售量差异不显著；
6. 商品销售存在明显季节性，存在预测未来销售额的基础。

上述分析说明公司商品覆盖程度较高，价格优势强；但也表明公司发展受限于所处行业，商品开发尚有较大空间。公司可以通过开发高价值商品，淘汰低利润商品和调整定价策略提高收入及利润。鉴于商品销售存在明显的季节性，公司可以针对性调整库存管理和营销策略。例如在年末销售高峰期前通过销售预测调整库存，并在销售较慢的时期推出促销活动以提高库存周转。

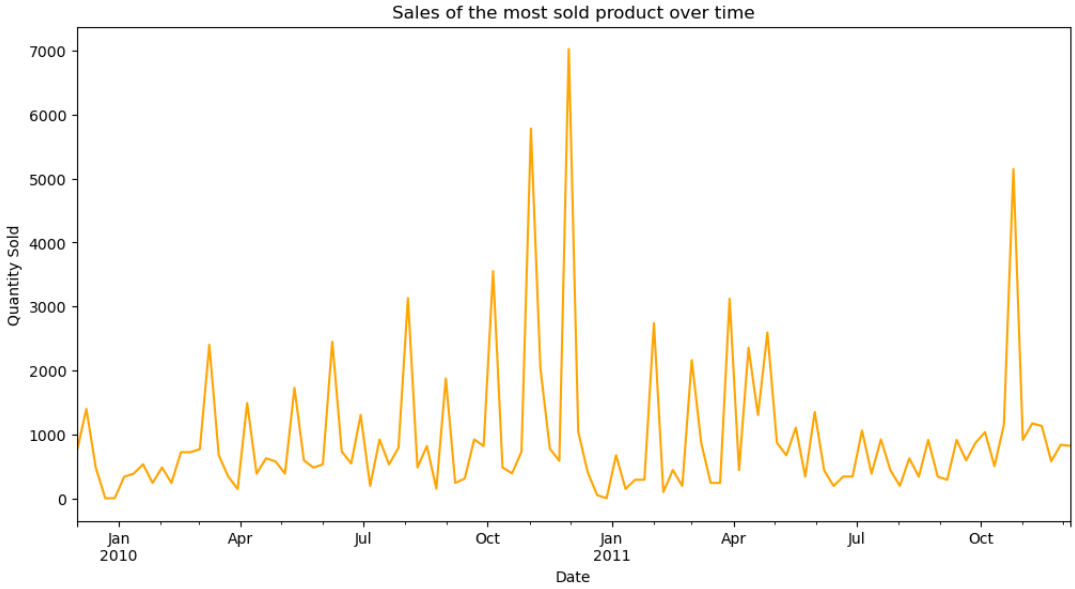
并且我们注意到，高收入商品之间销售额差异明显，因而为这些产品量身定制的营销策略可能会提高它们的销售，公司可以在营销活动中强调这些产品的独特价值和优势。而热销商品之间销售量差异不显著，表明这些商品属于基础款，客户对它们有相似的兴趣水平，可以通过关联推荐让客户关注到这些商品，进一步提高销量。

2. 销售趋势预测

销售趋势预测使用已有销售数据通过模型预测未来销售数据，提高供应稳定性。考虑到前述对销售数据的分析表明销售趋势存在明显的季节性，并且公司销售商品为礼品及礼品类商品，销售趋势的变动可能和节假日存在关系，因而我们采用能够包含上述因素的SARIMAX模型进行分析和预测。为增强分析针对性，销售趋势预测选取了在英国本土销售量最大的商品，预测其在数据集结束后未来1年内在英国本土的销售量。

我们首先确定模型的使用前提已经满足，然后将销售量最大商品的销售数据进行可视化，发现于每年年末销售量均出现明显峰值，且每年销售变动趋势总体类似，因而表明基础数据存在季节性周期，并将其长度确认为52周（1年）。构建模型过程中，我们首先将公共假期信息提前3周聚合至基础数据中，提前3周的原因为客户主要为需要较长周期备货的批发商，我们通过对比销售高峰和节假日的方式确定销售高峰通常在节假日前3周左右来临；我们将已有数据分为训练数据和测试数据，使用训练数据拟合最终使用的模型，并让模型对测试数据进行验证， 模型捕捉到销售高峰的到来，效果良好；我们使用模型进行未来销售情况预测，模型预测的销售量变动趋势符合预期。

最热销商品在英国销售量变动趋势：

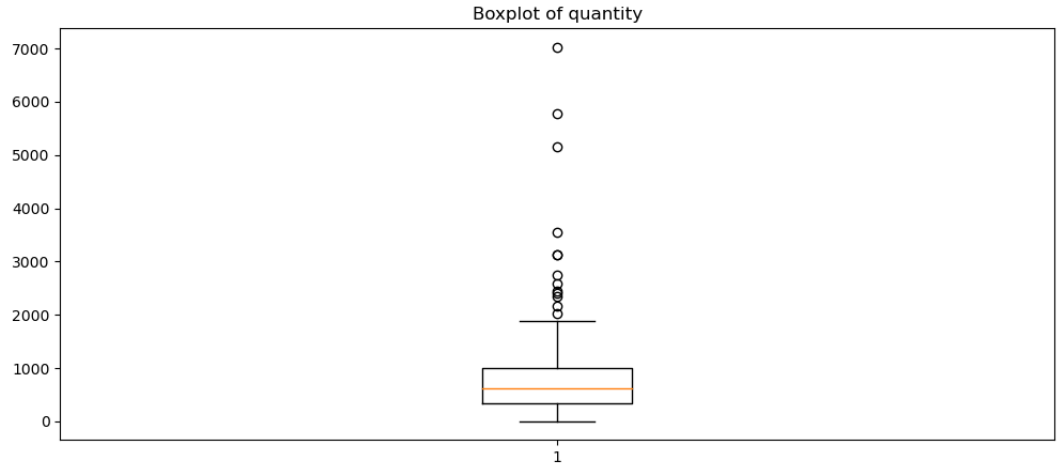


确认最热销商品在英国销售量变动满足模型使用前提条件：

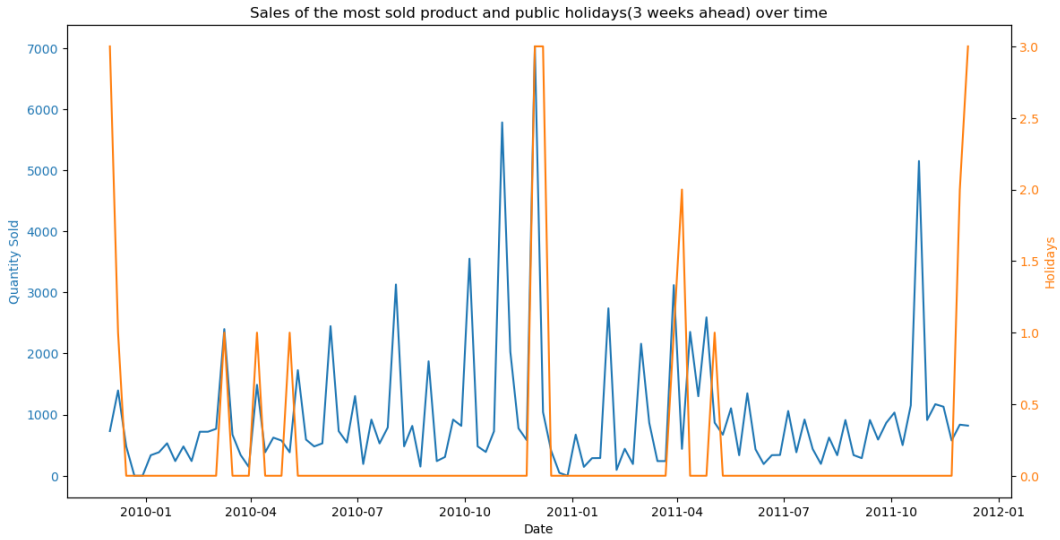
1. 数据总体平稳：

```
Results of Dickey-Fuller Test:
Test Statistic                  -3.513757
p-value                          0.007641
#Lags Used                       3.000000
Number of Observations Used    102.000000
Critical Value (1%)             -3.496149
Critical Value (5%)             -2.890321
Critical Value (10%)            -2.582122
dtype: float64
```

2. 数据无异常值：



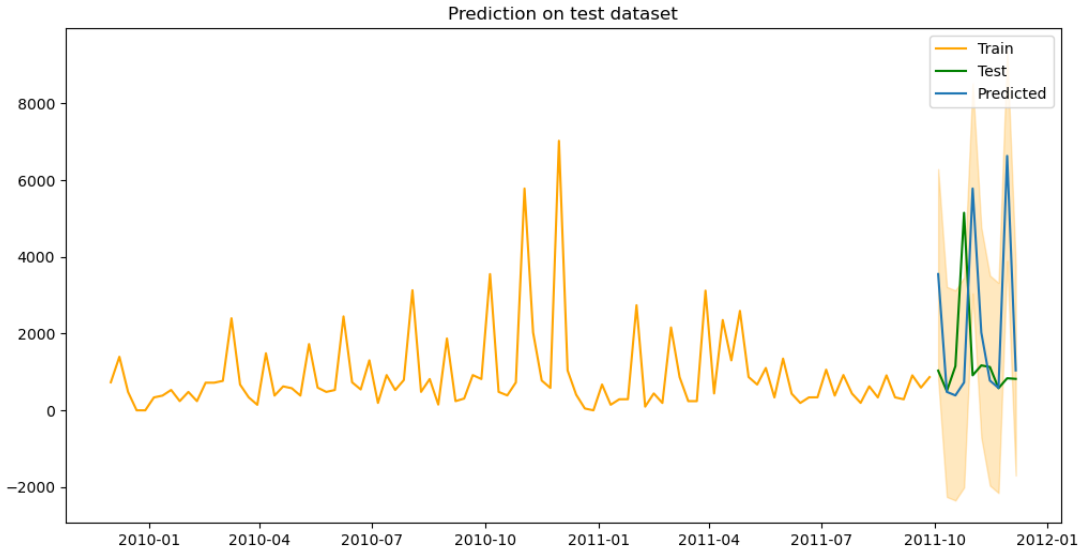
调整后的节假日高峰与销售高峰吻合：



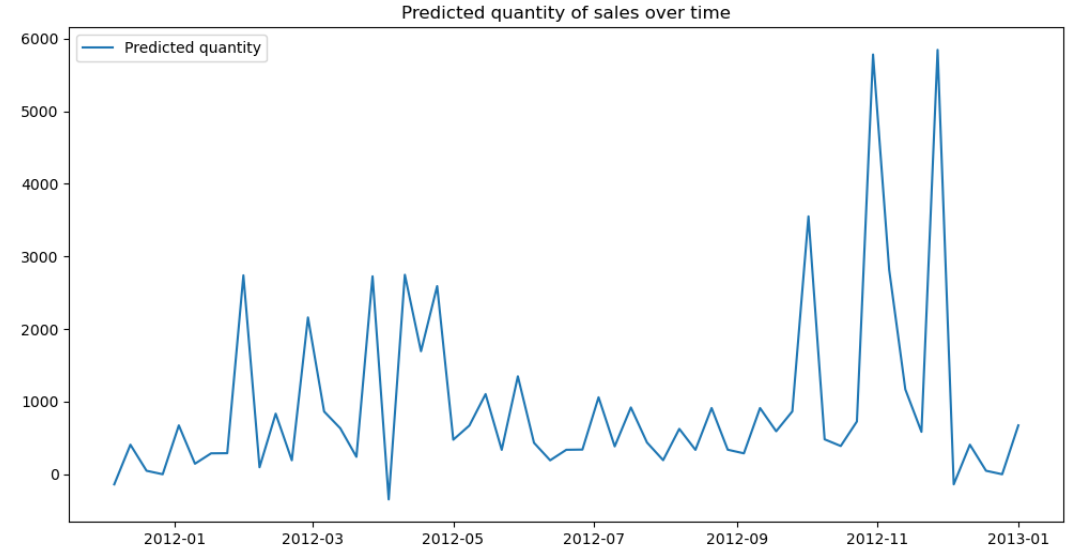
使用训练数据确定的模型参数：

```
best_pdq, best_seasonal_pdq
((0, 0, 0), (0, 1, 1, 52))
rmse
2910.5554852386963
```

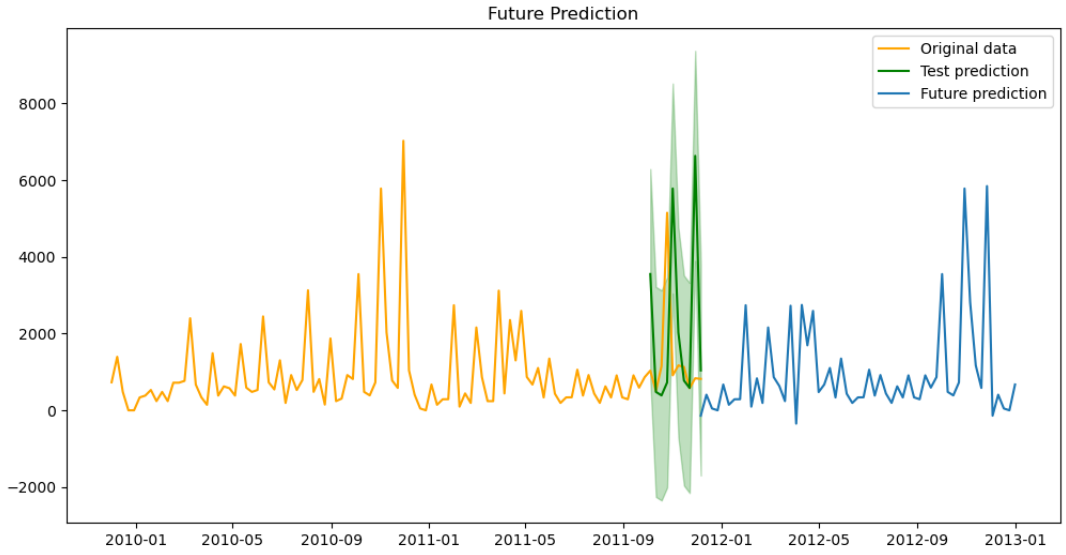
使用模型对测试数据进行验证：



使用模型对未来期间销售量进行预测：



预测全过程展示：



预测结果表明，模型成功根据可获取信息捕捉到销售情况的变化趋势及整体体现的季节性。通过对未来销售量的预测，公司可以在预测的高峰期增加备货，提前准备物流资源，以提高供应稳定性。

七、结论

受限于数据的可获取性，我们未能对于业务运营的其他方面进行分析；但针对已有数据的分析，我们获取如下主要发现：

1. 在客户分析中，我们的主要发现为客户整体粘性不高，退货订单占总订单比例较低，但其中高频率退货客户较多，随时间推移客户留存率显著下降，且每年末后客户留存率会进一步下降，以及商品销售存在明显季节性；
2. 在商品分析中，我们发现商品覆盖程度较高，价格优势强，但商品开发程度较低，高收入商品的独特价值和优势未成功带来商品差异化，而热销商品之间也未形成商品矩阵，吸引客户购买。

以上结果表明合作关系的提升空间较大，商品覆盖，供应稳定性和商品质量还有改进空间。针对上述问题，我们认为如下方面可以改进：

1. 合作关系方面，按照客户分层进行针对化营销和及时回访，提高客户粘性和服务质量；
2. 商品覆盖方面，提高高收入商品之间的差异化和营销力度，建立热销商品之间的商品矩阵，提高单笔订单的销售额；
3. 价格优势方面，在保持价格优势的基础上逐步提升商品结构；
4. 供应稳定性方面，通过不断更新销售预测模型进行更精准的库存管理，提高物流资源的使用效率；
5. 商品质量方面，按照客户层次的结果向低满意度客户（如“需要唤醒”客户和“关注”客户）及时回访。