第十四届全国大学生
电子商务“创新、创意及创业”挑战赛

项目报告书

作品名称：情感洞察——跨境电商女装店铺消费数据分析

团队名称：代码不报错队

团队负责人：陈欣宇

团队成员：王锐恒、杨雨洁、李飞扬、罗小雪

负责人学校：南京邮电大学

日期：2024.4.3

FINE GRANULARITY OF WOMEN'S CLOTHING

目　录

摘要

本次研究的目的是为企业设计针对性的营销策略，助力企业销售增长。**根据用户购买行为和评价中所反映的用户体验和情感进行营销**，是电商平台中开展营销活动的有效途径。

本报告首先进行**市场分析**，包括**整体市场分析，宏观环境分析，竞争分析和目标市场定位**。利用某企业在跨境电商平台运营的女装店铺近半年的消费数据分别从用户画像、企业运营以及消费者细粒度情感三个方面对用户的（购买前）关注点——购买行为——（购买后）评价与推荐的每个环节中蕴含的体验和情感因素进行分析，根据用户的体验和情感因素分析结果为该企业设计了数字营销计划。

在**用户画像**部分，从年龄段、女装分类、尺码分类三个维度对于产品销量和评分进行分析，从而对该企业电商平台进行的**客户画像分析**；并基于**K-means++聚类分析**实现消费者画像模型的构建，将消费者分为**重点稳固、重点培养、精准培养三**类，帮助企业针对不同消费者群体**精准推送**女装产品。

在**企业运营**部分，根据用户评分定义了满意度指标，分析消费者对于不同服装品类的满意度，并利用**基于熵权法的 TOPSIS 模型**评价出该企业最受欢迎的女装产品是 Tops 大类和 Dresses 小类，对企业运营情况进行整体分析，以便后续进行产品设计、生产等决策。

在**细粒度情感分析**部分，基于**传统情感分析**和 **Logistic 回归模型**对顾客推荐产品的意愿进行分析，确定推荐机制，例如成年或大龄女装更容易被推荐。利用**词云图、LDA 主题挖掘算法**以及 **Python 算法**的对客户的产品评价进行**细粒度情感分析**，发现客户评价主要关注**样式、尺寸、品质、色彩**四类主题；多维度的评分分析方便了企业评估和改进女装产品，把握顾客需求和市场趋势，积极调整营销策略，并以 Dress 分类为例分析了导致客户正面评价的因素为独特**款式**与极致**舒适度**，而负面评价的主要因素主要为尺寸问题。

根据上述分析结果，为该企业制定的**数字营销计划**的要点包括：依据细粒度情感分析抓取商品问题和顾客需求、结合行业热搜关键词、优化非私人贴身衣物生产重心，针对不同年龄段消费者推荐合适产品，以提高销量并扩大客户群体。

本报告使用**数据可视化大屏**展示数据分析结果，采用**图文并茂的创意手法**诠释分析数据。同时根据数据分析结果提供相应数字营销策略，提高搜索**关键词**排名，洞察热门关键词和方向，以及制定跨平台的**广告营销**策略，旨在帮助企业精准定位市场，拓展品牌知名度，呈现项目的**创意**性和实用价值。

本项目不仅对该企业客户和销售情况从情感角度进行了深入分析，发现了消费者购买需求和偏好方面规律，并设计了针对性的营销计划，期望能为该品牌女装销售起到促进作用。分析过程和方法也可用于其他同类企业的销售分析和营销计划制定的工作，具有实际**创业**意义。

1 前言

1.1 研究背景与研究目标

随着互联网的蓬勃发展，跨境电商行业迅猛崛起，作为其中重要分支的女装跨境电商行业也处于高速发展期。近年来，女性消费者对时尚、品质的追求不断提升，对国际品牌女装的需求不断增加，这促使着女装跨境电商市场持续发展。然而，女装跨境电商行业呈现稳健增长态势的同时，其市场竞争格局日趋激烈。因此，女装电商企业惟有深入洞察消费者心理、不断满足消费者需求，持续获得消费者青睐，才能在这个机遇与挑战并存的市场中脱颖而出。

以往的营销以商品推荐为主，较少关注用户的心理和情感因素，但其实用户的情感和态度对于购买决定很重要，所以从用户情感因素的角度出发、制定营销方案是当前网络营销中的一大趋势。因此**本次研究的目标是为了对用户的情感因素进行深入分析，从而制定营销计划、促进销售**。

电商平台中用户的喜好和情感通常可以（购买前）从关注点——购买行为——（购买后）评价和推荐中获得信息，所以本报告利用某企业在跨境电商平台运营的女装店铺近半年的消费和评价数据，对顾客进行画像分析并基于K-means++聚类模型将顾客分类，帮助电商企业针对不同消费者群体精准推送女装产品；利用 LDA 主题挖掘算法的对顾客的产品评价进行细粒度情感分析，了解消费者不同的服装需求背后的情感因素，多维度的评分分析方便了企业评估和改进女装产品，把握顾客需求；最后会通过可视化大屏展现本次数据分析的结果，同时也会相应提供数字营销策略，帮助企业进行市场营销。本报告主要目的是实现以下目标：

(1) 帮助企业根据消费者特征和偏好进行客户分类，针对不同顾客实现精准营销；

(2) 对企业运营情况进行整体分析，确定最受欢迎的服装产品类别，以便后续进行产品设计、生产等决策；

(3) 对顾客推荐产品的意愿进行分析，有助于企业专注于提高客户的购物满意度，促进顾客推荐企业服装产品；

(4) 对顾客评价进行细粒度情感分析，帮助企业全面评估服装产品，采取相应的措施提高服务和改进服装产品。

本报告通过针对用户体验和情感从多个角度进行分析，依靠精准把握顾客喜好和情感体验，并及时反馈和调整，在激烈的服装市场竞争中取得成功，拓宽企业产品的知名度。

1.2 分析思路

本报告具体逻辑框架图如图 1.1 所示：

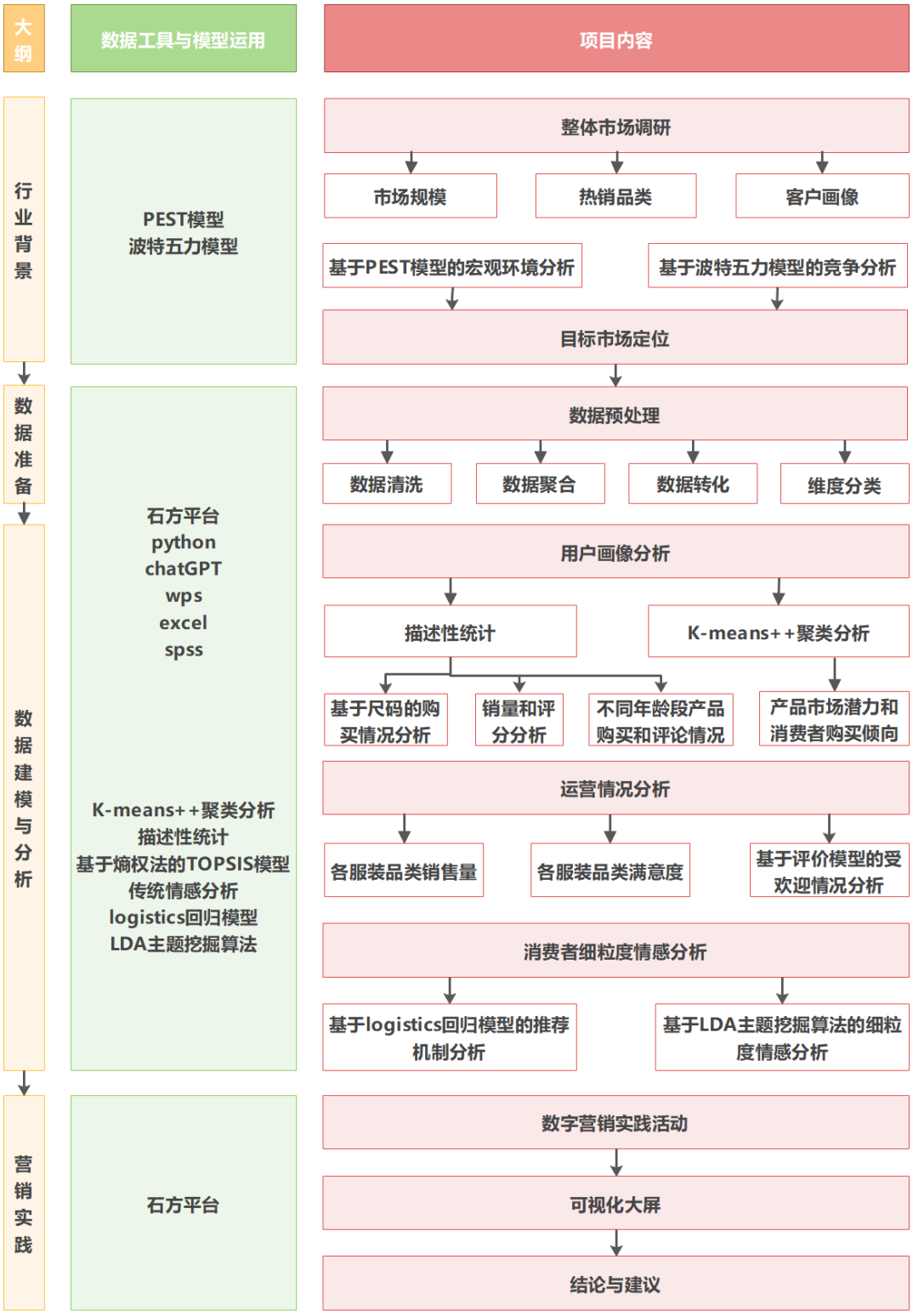

图 1.1 逻辑框架图

2 市场分析

2.1 整体市场调研

通过国家统计局以及其他网站资料，搜集中国服装市场销售量以及女性消费者对女装的购买偏好的数据，进行整体市场调研，充分了解女装市场状况，帮助企业准确定位并作出合理决策。

2.1.1 中国服装市场规模

从国家统计局、中国服装协会网等网站统计数据得到 2006-2022 年中国服装商品的销售额以及同比增长率，绘制出统计图。由于疫情对经济的冲击显著，从近年的销售额可以看出，社会销售额有明显的下降趋势，服装市场经济状况收到疫情的严重打击。而后国家发布《纺织业“十四五”发展纲要》等政策促进服装市场发展，大力推广线上销售服装道路，服装市场的销售情况得到明显改善。

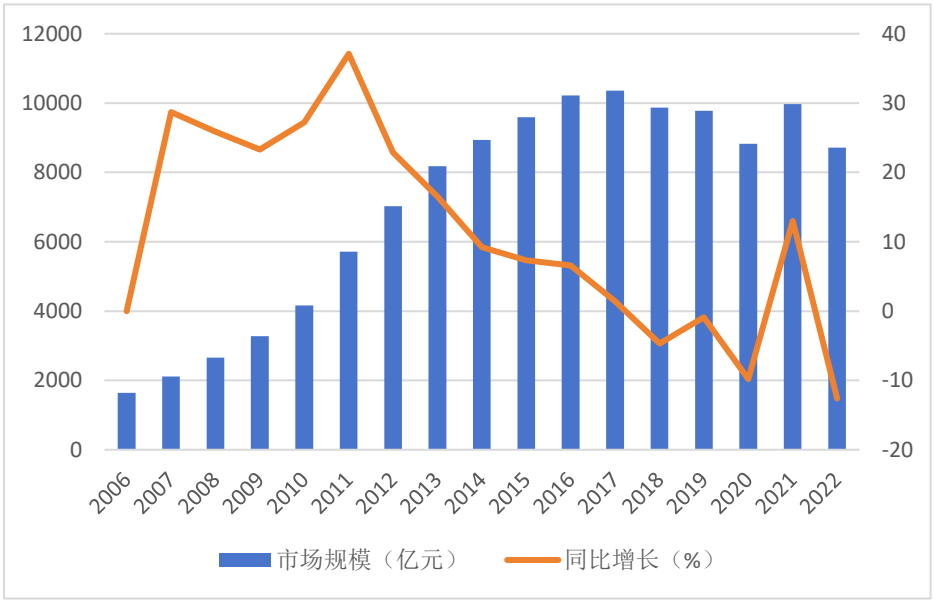

图 2.1 2006-2022 年中国服装类商品零售额

由于女性就业比例的提高和社会地位不断提升，女性有经济上的独立性，其消费需求也随之不断提高，这推动女装行业发展。2020 年，女装市场规模受疫情影响有所下降，而在 2021 年疫情好转后，中国女装市场的销售额也增长至 2019 年的水平。

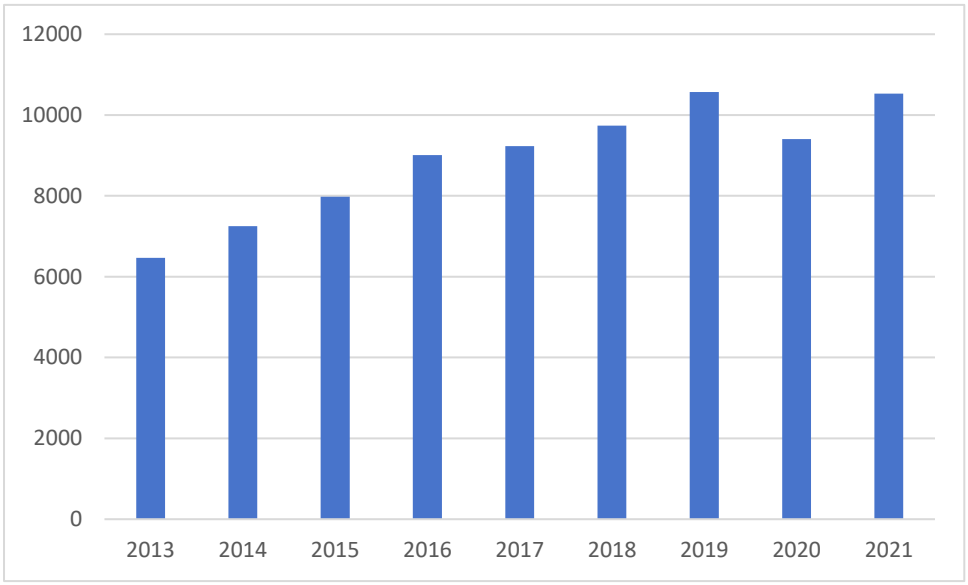

图 2.2 女装市场规模

2.1.2 热销女装品类分析

根据服装相关网站统计的 2022 年女装热销的品类数据，可以得到最受欢迎的品类是连衣裙，而 T 恤和毛衣也是比较热门的服装，企业应当增加投入此类女装的生产，迎合市场需求。

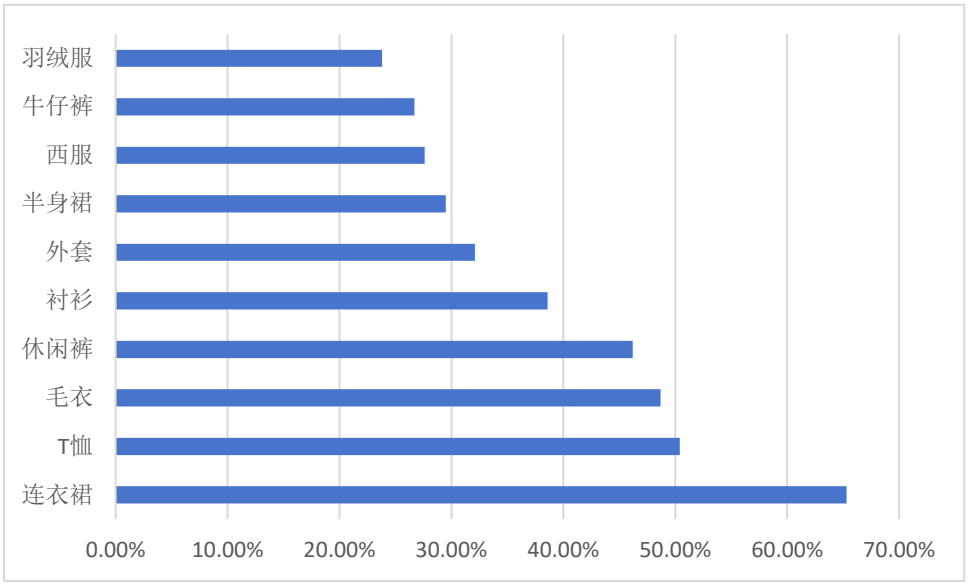

图 2.3 2022 年热销女装品类

2.1.3 消费者画像调研

统计女性消费者每月的服装购买金额以及购买频次，可以看出购买金额集中于 201-600 元的范围，而每月购买次数大多是 2-3 次，女性消费者在服饰上消费水平比较高，企业应该投入更多资源，推出更多具备吸引力的女装产品。

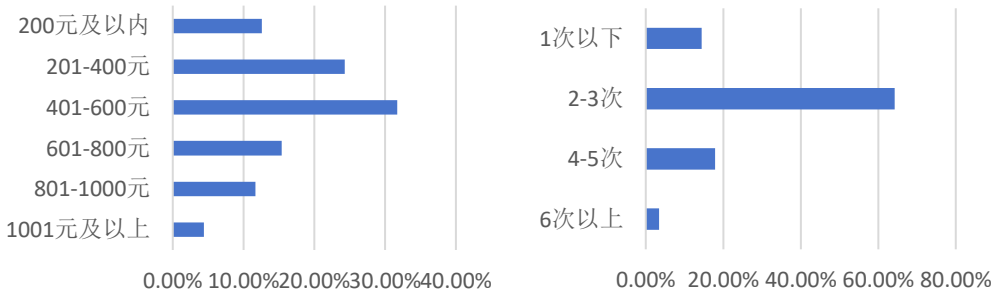

图 2.4 女性消费者每月服装购买金额及频次

2.2 宏观环境分析

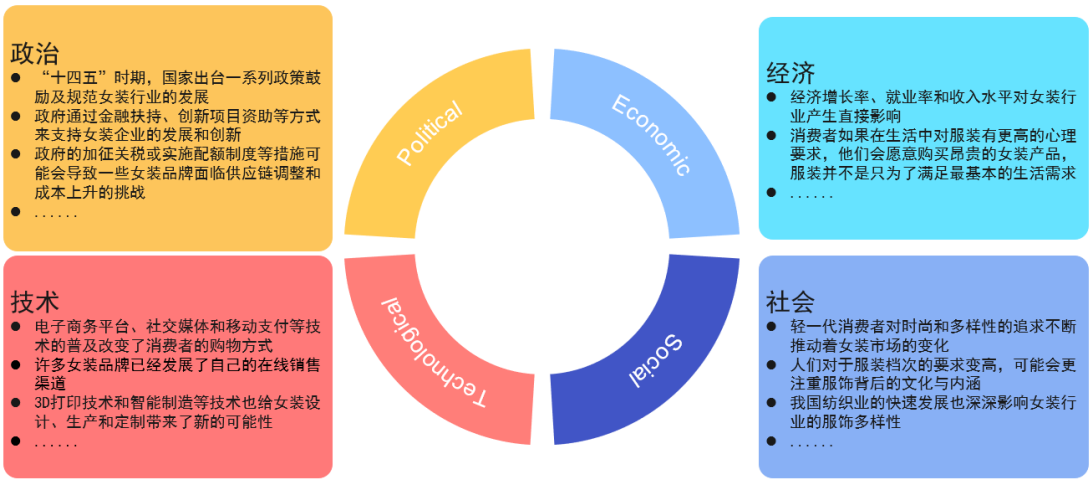

图 2.5 PEST 模型分析

2.2.1 政治因素

在“十四五”时期，我国出台了一系列关于女装行业的政策，鼓励和规范女装行业的健康发展，完善了中国女装行业政策体系，促进了我国女装行业发展。政府通过金融扶持、税收优惠、创新项目资助等方式来支持女装企业的发展和创新。而贸易政策调整、劳动力法规和税收政策等都可能直接或间接地影响到女装行业的发展。政府的贸易政策和关税变化可能对女装行业的进出口产生重要影响，加征关税或实施配额制度等措施可能会导致一些女装品牌面临供应链调整和成本上升的挑战。

2.2.2 经济因素

经济因素对女装行业的影响主要表现在两个方面，一是经济状况是否快速增长，二是消费者的生活观念是否转变。经济增长率、就业率和收入水平对女装行业产生直接影响。当经济繁荣时，人们更愿意购买高品质的女装，并倾向于追求时尚和多样性。然而，当经济不景气的时候，人们可能会减少对昂贵品牌的购买，转而选择价格更为实惠的产品。而消费者在生活中对服装有更高的心理要求时，他们会愿意购买昂贵的女装产品，服装并不是只为了满足最基本的生活需求。

2.2.3 社会因素

社会对美的定义和价值观可能导致服装需求的变化，潮流和文化趋势对女装行业产生深远影响。年轻一代消费者对时尚和多样性的追求不断推动着女装市场的变化。此外由于社会教育水平的提高等情况，人们对于服装档次的要求变高，可能会更注重服饰背后的文化与内涵。我国纺织业的快速发展也深深影响女装行业的服饰多样性。

2.2.4 技术因素

技术的迅速发展对女装行业带来了新机遇。电子商务平台、社交媒体和移动支付等技术的普及改变了消费者的购物方式。互联网广泛应用让电子商务平台成为女装消费的重要渠道。近年来，许多女装品牌已经发展了自己的在线销售渠道，并通过社交媒体平台与消费者进行互动。同时，3D 打印技术和智能制造等创新技术也给女装设计、生产和定制带来了新的可能性。3D 打印技术可以加速样衣制作过程，而智能制造则提供了更高效和精确的生产方式。

2.3 竞争分析

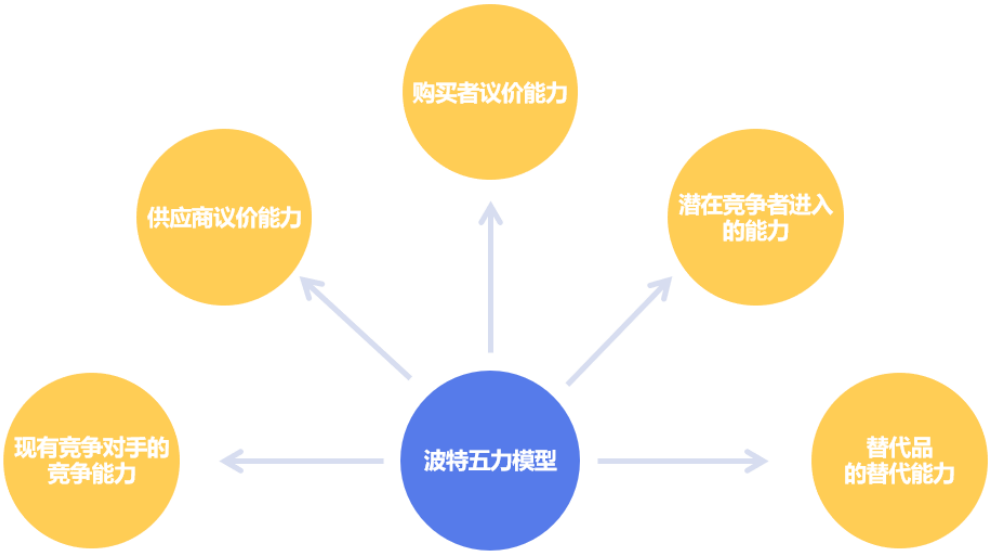

图 2.1 波特五力模型

2.3.1 供应商议价能力分析

在女装行业中，供应商主要是材料供应商和生产工厂。女装企业通常依赖于多个材料供应商，包括面料、纽扣以及其他原材料。如果市场上的供应商数量较少，或者某些供应商享有市场垄断地位，他们可能具有较高的议价能力。此外，供应商的规模、专业化程度和产品质量也会有所影响，大型供应商通常能够从规模经济中获益，并且在谈判中更有优势。

此外，女装企业通常将生产外包给制衣工厂，而这些工厂对价格和交货期具有一定的谈判权。规模较大、技术先进的工厂可以利用其资源和专业知识来提供更高质量的产品和更快的交货速度。然而如果与工厂建立长期合作关系还可以增加议价能力，因为双方建立了互信和稳定的合作关系。

2.3.2 购买者议价能力

女装市场竞争激烈，消费者面临多样化的选择。在众多品牌和产品之间，消费者倾向于比较价格和质量，并寻求最有价值的购买选项。因此，如果企业无法提供与竞争对手相比的附加价值或独特性，消费者可能会通过议价来获取更好的交易。线上购物平台也帮助消费者轻松比较不同品牌和商家的价格，并根据自己的预算做出选择。此外，如果女装品牌无法提供足够的附加值或与其他品牌的区别，批发商和零售商可能会要求更高的折扣，从而削弱产品的利润率。

2.3.3 现有竞争者竞争能力

在竞争激烈的市场中，建立和维护良好的品牌知名度对于吸引消费者、增加市场份额至关重要。拥有较高的品牌知名度和市场声誉的企业通常能够赢得消费者的信任和忠诚度，从而提高销售和市场份额。除此之外，通过设计独特、时尚且符合消费者需求的产品，女装企业可以与竞争对手区分开来。

此外，有效管理生产成本、运营成本和供应链成本，降低企业的总体成本，可以为企业提供更具竞争力的定价策略。通过提高生产效率、优化供应链以及寻找成本效益更高的材料和制造商，女装企业可以降低成本并提高利润率。

2.3.4 潜在竞争者进入能力

潜在的竞争者可能有一些新的想法，对于行业而言会有更加丰富服饰涌入市场，而对企业而言是新的冲击和竞争。然而一些女装行业的龙头品牌在市场上建立了强大的品牌壁垒，包括知名度、忠诚度和声誉，所以新竞争者也面临更大的挑战。因此，品牌建设对于女装企业来说是建立竞争优势的关键策略。

2.3.5 替代品的替代能力

替代品的存在增加了消费者的选择范围，从而可能降低对女装产品的需求。但是如果替代品价格更低且质量相当，消费者可能更倾向于购买替代品。此外，替代品的设计、风格、功能等特点也会一定程度影响消费者对替代品的接受程度。不过，如果消费者对某个品牌有较高的忠诚度和情感认同，他们可能不容易被替代品吸引，反而会一直购买所信任的品牌。

2.4 目标市场定位

STP 分析即市场细分(Segmenting)、目标市场(Targeting)和市场定位(Positioning)，是现代市场营销战略的核心，企业将市场进行细分，以此确定目标市场，再根据目标市场设计相对应的产品，明确产品定位，确定产品相关细节，开展市场营销活动。

2.4.1 市场细分

将整个女装市场细分，尝试使用不同的细分变量或组合变量，以便更好地满足消费者的需求和偏好。根据年龄、风格、价格和地理位置四个变量进行市场细分。年龄方面可以根据不同年龄段的消费者需求开展划分，可将市场细分为青少年、青年人、中年人和老年人等。在风格方面可以将市场细分为不同风格和时尚取向的消费者群体，例如休闲、正式、运动或潮流风格等。价格因素

方面，将市场细分为高端奢侈品市场、中高档市场和大众消费市场等，以满足不同消费者对价格的要求。地理因素方面，根据地理位置的差异，将市场细分为城市、乡村或不同国家或地区的市场。

2.4.2 目标市场

不同女装企业应该仔细研究各个市场细分，并评估每个细分市场的潜力和适应性。根据企业的资源、能力和品牌定位等因素，选择一个或多个目标市场进行市场开发和营销活动。一家女装企业可以选择青年人这个细分市场作为其目标市场。在此目标市场中，企业可以专注于提供时尚、多样化且质量可靠的产品，以满足年轻消费者对时尚潮流的需求。

2.4.3 市场定位

市场定位是女装企业将自己与竞争对手区分开来，在目标市场中塑造独特的品牌形象和价值主张的过程。一个成功的市场定位战略，必须要确保四个关键因素的成立：竞争性，可信性，简明性，一致性，即市场定位的 4C 关键因子。通过市场定位，企业能够凸显自己的优势，并建立消费者心目中的独特地位。

3 数据探索与预处理

3.1 数据清洗

3.1.1 缺失值处理

本次研究使用的数据集来自三创平台，共包含 11 列 19999 行，各列含义如表 3.1 所示。

表 3.1 数据字段及其含义

字段	含义	字段	含义
Class_Name	是服装小类，例如 Pants（裤子）、Skirts（半身裙）、Jeans（牛仔裤）等，它们都属于 Bottoms 类。	Age	客户年龄
Clothing_ID	衣服编号	F0	评论编号
Department_Name	服装分类，例如 Tops（上装）、Bottoms（下装）、Dresses（裙装）等	Positive_Feedback_Count	其他买家对该评论的反馈，可以理解为“认为评论对自己有用”。
Division_Name	代表服装款式，它分为普通款、小号普通款和贴身服饰。	Rating	星级，1-5 级，更高级别表示更高的满意度。
Review_Text	客户评论	Recommended_IND	表示是否愿意推荐该商品，为布尔类型变量
Title	客户评论主题		

首先针对 11 个特征变量利用 python 筛选出缺失数据，发现 Class Name, Department_Name , Division_Name, Review_Text, Title 这五个特征变量有缺失值，筛选结果如下：

表 3.2 缺失值筛选结果

特征变量	存在缺失的评论数
Class_Name	9
Clothing_ID	0
Department_Name	9
Division_Name	9
Review_Text	711
Title	3248
Age	0
Positive_Feedback_Count	0
Rating	0

Recommended_IND	0

因为评论内容和服装信息在后续的数据分析中评论尤为重要，故我们针对 Class_Name，Department_Name，Division_Name，Review_Text 这四项存在缺失的数据进行**去除**，且数据量较少，去除后对数据分析**影响较小**。本文假设用户因自身的某种原因忽略或不想写评论标题，故**保留** Title 存在缺失的数据集，字符串为 null。

3.1.2 **异常值处理**

(1) 评论文本（review_text）中的异常值处理

我们根据单词数量进行筛选出**异常评论**，将单词量不超过 3 个的判为无效评论。首先，通过对文本进行分割，利用空位分割符对文本中的评论分割后，通过增加对分隔符的个数判据即分隔符大于等于 2，从而筛选出单词个数大于等于 3 的有效评论。同时，考虑到存在刷单的行为，我们**去除重复评论**。

处理之后的部分数据如下图所示：

Class_Name	othing_	rtment_	vision_Na	Review_Text	Title	Age	FO	e_Feedbac	Rating	ommended_IND
Blouses	847	Tops	General	This shirt is very flattering to a	Flatterir	47	4	6	5	1
Blouses	853	Tops	General	Took a chance on this blouse and s	Looks gre	41	17	0	5	1
Blouses	847	Tops	General	If this product was in petite, i v	Cute, cri	33	20	2	4	1
Blouses	847	Tops	General	I love this shirt because when i i	Versatile	55	24	0	5	1
Blouses	823	Tops	General	Very comfortable, material is good,	cut out	52	47	0	5	1
Blouses	845	Tops	General F	This top is so cute, but it is ma	Cute, but	38	70	10	4	1
Blouses	822	Tops	General	Why do designers keep making crop	Short and	36	71	0	2	0
Blouses	850	Tops	General	I have a short torso and this worl	Beautiful	27	72	4	5	1
Blouses	845	Tops	General F	I am so drawn to baby doll and bo	Very very	48	75	5	5	1

图 3.1 评论异常值处理过后的部分数据展示

可见处理过后的评论都为有效数据。

(2) 顾客年龄（age）中的异常值处理

我们对年龄进行描述性统计分析，我们将年龄数据导入 sspspro,如图：

表 3.3 顾客年龄描述性统计结果

类型	定量	最大值	99
样本量	19281	最小值	18
缺失值	0	中位数	41
去重量	77	变异系数	0.284
平均值	43.339	方差	151.188
标准差	12.296	S-W 正态检验	不满足 (P=0.000***)

在描述性统计分析结果中，可以看到，最大年龄为 99 岁，最小年龄为 18 岁，平均年龄为 43 岁，中位数为 41 岁，年龄的均值和中位数的差距不大，所以，年龄分布基本符合正态分布。下面的直方图可以更直观地看到年龄的分布情况：

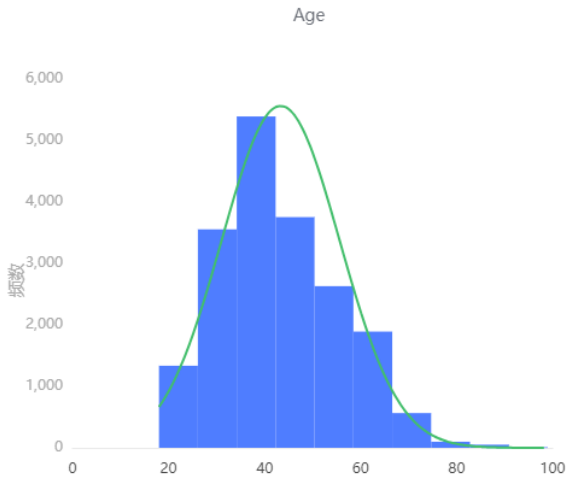

图 3.2 年龄分布直方图

通过对年龄数据进行正态性检验，检验结果如下表：

表 3.4 正态性检验值

变量名	样本量	中位数	平均值	标准差	偏度	峰度	S-W 检验	K-S 检验
Age	19281	41	43.339	12.296	0.518	-0.125	0.976(0.000***)	0.086(0.000***)

注：***、**、*分别代表 1%、5%、10%的显著性水平

发现显著性 P 值为 0.000***,水平上呈现显著性，拒绝原假设，因此数据不满足绝对的正态分布，通常现实情况下很难满足检验，但观察发现其峰度（-0.125）绝对值要小于 10 并且偏度（0.518）绝对值要小于 3，结合正态性检验直方图，发现正态图基本上呈现中间高，两端低，因此用户年龄基本可以接受为正态分布，由于年龄高于 80 岁的数据很少，而且可能存在用户乱填的情况，考虑到下面要进行用户画像分析，故采用 3sigma 原则去剔除该区域以外的数据，通过计算数据边界为[6.422,80.227],将区域外的数据剔除，从而得到合理的用户年龄值。

3.2 数据聚合

消费者对购买的女装给予了相应的星级评价 Rating，其中星级范围为 1-5 星，通过对商品评论中星级这一数据进行聚合，求出星级中位数为 3，则星级评级大于 3 为好评，等于 3 为中评，低于 3 为差评。

图 3.3 通过聚合求出星级中位数

3.3 数据转化

数据来源：三创赛平台所给数据提取的用户评论

(1) 消费者情感分析的原理

为了引入客户评价因素，考虑到客户评价指标值数据形式为文本格式，为了能够进行定量分析，首先应对客户评价文本进行数值转化，通过采用传统情感分析方法，评价各条评论的情感倾向，进一步计算出相应的情感得分，根据情感值确定情感标签。考虑到分析对象为女装，因此使用 VADER 工具，VADER 是使用一组预定义的情感极性字典和一套规则，来估计文本的情感倾向，能够处理文本中的情感表达，包括表情符号，大写字母，标点符号和否定词语等，VADER 分类器更适用于网络文本的情感分析。具体过程是：

- 将用户评价的文本分句，把句子切分为词；
- 统计情感词及其位置，利用 python 中的 NLTK vader 情感分类器对所给数据进行转化，并进行加权计算，统计整段评价文本的情感值；
- 将情感标签分为消极、中性、积极三类，根据整段评价的情感值确定情感标签。

(2) 消费者情感分析的结果

所得结果如下表所示：

表 3.5 用户评价情感分析

情感分类标签	客户评价总额百分比	客户评价频数
消极	5.65%	1090
中性	2.05%	397
积极	92.3%	17794

从表 3.5 和图 3.4 可以看出：当情感分数接近 1 时，密度较高，说明整体

评论的积极性分布更为广泛，同样当情感分数小于 0 时分布密度较低，说明评论的消极性分布非常稀疏，通过对分数进行编码，得到如图 6.1 右所示的饼状图，其中积极文本占比 92.3%，消极文本占比 5.7%，中性文本占比 2.1%。绝大多数是积极评价，但消极评价的占比高于中性评价。将客户评论进行传统情感分析所得到的情感分数作为指标 **score**，代表客户评论情感。

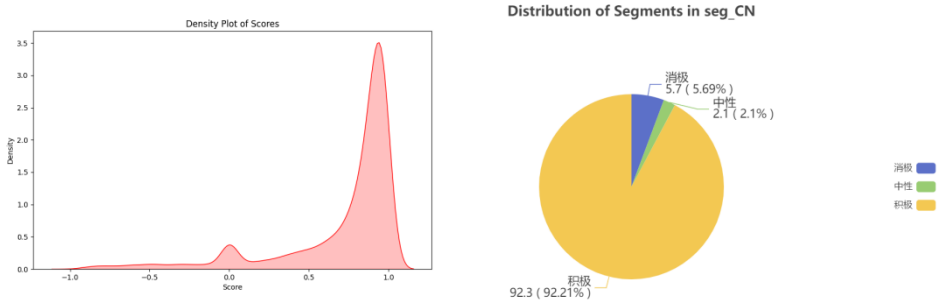

图 3.4 请感分数分布密度以及情感关键词分类占比

3.4 数据维度分类

本次使用的数据中包含 4 个维度变量和 5 个目标变量，如表 3.6 所示。

表 3.6 4 个维度变量和 5 个目标变量

特征变量	分类情况	变量类型
Class_Name	具体分类和 department_name 有关，如表 3.6 所示	维度变量
Clothing_ID		
Department_Name	分为 6 类，如表 3.6 所示	维度变量
Division_Name	分为 3 类： General 指正常大小， General Petite 指为身材娇小的女性提供的服饰， Initmates 则指私人贴身服饰	维度变量
Review_Text		目标变量
Title		目标变量
Age	分为 5 类： 30 岁以下、 31 至 40 岁、 41 至 50 岁、 51 至 60 岁、 61 岁及以上	维度变量
Positive_Feedback_Count		目标变量
Rating		目标变量
Recommended_IND		目标变量

女装数据的 Department_Name 与 Class_Name 对应关系如以下矩形树状图 3.5 所示

图 3.5 Class_Name 与 Division_Name 对应关系矩形树状图

4 用户画像分析

4.1 描述性统计

4.1.1 基于尺码的女装购买情况分析

各年龄段消费者频数和百分比具体如下表所示：

表 4.1 基于尺码分类下消费者年龄分布频数和百分比表

Age \ Division	General	Petite	Initmates	总计
30 岁以下	1581	862	231	2674
	13.80%	13.20%	19.30%	13.90%
31 至 40 岁	3810	2223	433	6466
	33.30%	33.90%	36.10%	33.70%
41 至 50 岁	2946	1683	264	4893
	25.70%	25.70%	22.00%	25.50%
51 至 60 岁	1914	1107	169	3190
	16.70%	16.90%	14.10%	16.60%
61 岁及以上	1200	677	102	1979
	10.50%	10.30%	8.50%	10.30%
总计	11451	6552	1199	19202

通过频数和百分比表分析不同年龄段的购买情况，得出以下结论：

(1) 非内衣类别中常规尺码比偏小尺码更受欢迎

General 类产品最受欢迎，购买量共计 11451 件，约占总购买量的 60%。General Petite 类购买量达到 6552 件，表明对于非私人贴身服饰，选择正常尺码的消费者占多数，但小个子市场的潜力不容忽视。

(2) 31-40 岁的年龄段是消费主力

某产品下各年龄段的购买量占该产品的总购买量之比，可反映该产品的消费者年龄分布情况，而产品消费者年龄分布情况在一定程度上可视为该产品在各年龄段的市场潜力。

根据表格数据，三类产品的消费主力均集中在 31 至 40 岁。General 类和 Petite 类消费者年龄分布非常相近，即在各年龄群体市场潜力基本一致。对比来看，Initmates 类（贴身私人服饰）30 岁以下的消费者占比 19.3%，显著高于其他两类，其在年轻市场中市场潜力更大。

(3) 不同尺码分类的消费者评分有一定差异

女装市场整体评分较高，消费者评分均值为 4.18。不同品类的评分呈现出较大差异。

表 4.2 Division Name 分类下市场总评分

类别	评分
General	4.16
General Petite	4.19
Intimate	4.28

结果显示，General Petite（小个子服装）市场评分高于 General(常规)市场，Intimate 市场份额最小，但产品评分高。评分和销量之前呈现反比趋势，销量越高，评分越低。商家应严格把控产品质量，避免因销量上升而品控不过关而影响消费者购物体验。

(4) 不同年龄段的消费者评分有较大差异

为进一步分析各产品的市场评分，本文统计了各年龄段的评分均值，如下表所示：

表 4.3 各年龄段段对不同产品的评分均值

age	General	General Petite	Initmates
30 岁以下	4.16	4.16	4.36
31 至 40 岁	4.14	4.14	4.22
41 至 50 岁	4.12	4.18	4.33
51 至 60 岁	4.23	4.26	4.25
61 岁及以上	4.25	4.33	4.19

结果分析：

General 类：在 41 至 50 岁群体评分最低（该年龄段的销量第二），61 岁以上群体的评分最高（该年龄段的销量最低）。

General Petite 类：在 31 至 40 岁群体评分最低（该年龄段的销量最高），61 岁以上群体的评分最高（该年龄段的销量最低）。

上述两类中评分和销量有明显的负相关关系，可能是和产品的品质有关，建议提高产品质量。

Initmates 类：在 61 岁及以上群体评分最低（该年龄段的销量最低），30 岁以下群体评分最高（该年龄段的销量中等，排在第三）。该分类中，评分和销量没有明显的相关关系。原因可能是目前市场上 Initmates 类面向年轻人的设计占多数，对老年人的需求关注少，因而年轻消费者的评分普遍更高。

由于 61 岁以上群体的生活场景相较于年轻群体更简单，倾向于“一衣多用”，对 Initmates 类的需求少。年轻群体倾向于“专衣专用”，对睡衣（Sleep）、家居服（Lounge）、泳衣(Swim)等有区别于普通衣物的需求。并且 Initmates 类在年轻市场的销量和评分都较高，所以商家应将 Initmates 类销售重心放在 50 岁以下群体，根据消费者购物反馈，不断改良产品。

4.1.2 基于大类的销量和评分分析

(1) 各大类的销量对比

Department_Name 下产品共分为 6 类,Tops 类最为畅销，其次是 Dresses 类和

Bottoms 类。各品类销售量帕累托图如下：

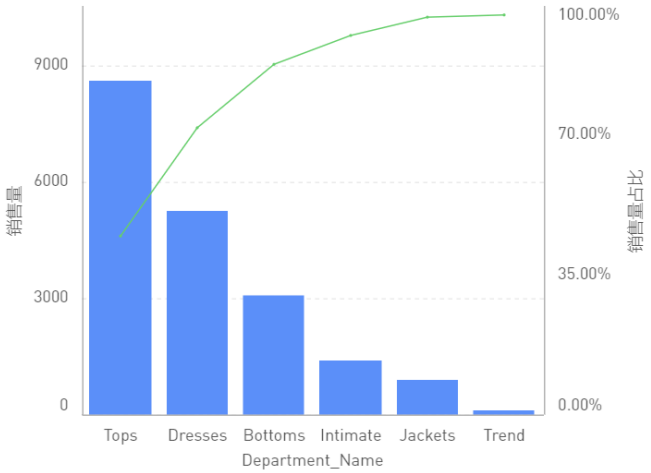

图 4.1 Department_Name 下各品类销售量帕累托图

(2) 各大类消费者年龄分布情况

Department_Name 下各类别消费者年龄分布如下图所示，图例 1 至 5 依次表示 30 岁以下、31 至 40 岁、41 至 50 岁、51 至 60 岁、61 岁及以上。

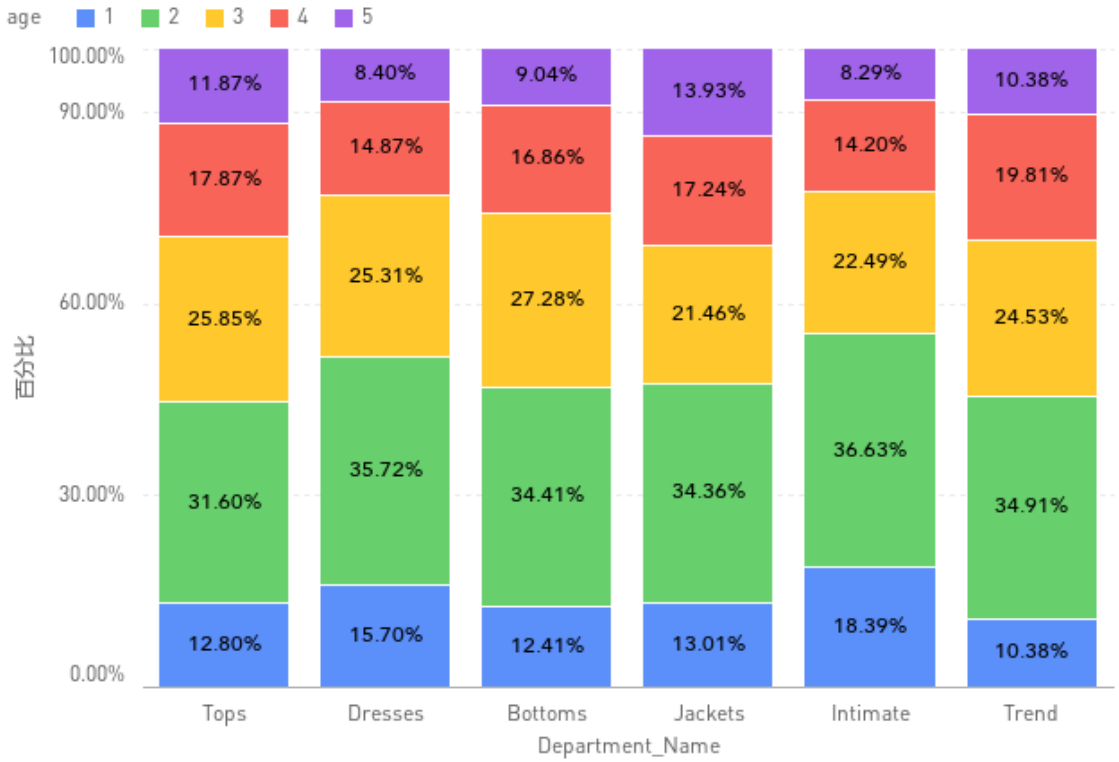

图 4.2 Department_Name 分类下各年龄段百分比堆叠图

- **畅销品在 41 至 50 岁消费市场表现良好**

根据百分比堆叠图，总的来看，各类产品的消费者年龄分布情况一致。畅销品（Tops、Dresses、Bottoms)中，Dresses 的消费者年轻化特征明显，40 岁以下消费者占比约 40%。而相较于畅销品，三类非畅销品（Jackets、Intimate、Trend）在 41 至 50 岁群体表现较弱。

- **Intimate 类年轻化特征明显**

Intimate 类产品受众群体最为年轻，30 岁以下消费者占比约 18%，31 至 40 岁消费者占比约 36%。Intimate 类 30 岁以下消费者相较于 61 岁及以上占比更高，而 Jackets 类刚好相反。在 61 岁及以上群体中，市场潜力最大的是 Jackets 类。Trend 类销量最少，但其在 51 至 60 岁群体中的市场潜力大于其他品类。

(3) 各大类评分比较

表 4.4 Department_Name 分类下各类别市场评分

类别	评分
Tops	4.16
Dresses	4.15
Bottoms	4.27
Intimate	4.27
Jackets	4.26
Trend	3.87

畅销品中 Bottoms 评分最高。非畅销产品中，Intimate 和 Jackets 产品评分高，Trend 类产品评分显著低于其他品类。结合各类销售量分析，Trend 类销量最少且评分最低，其原因之一是大多数消费者都是保守型消费，不敢轻易尝试 Trend 类产品，而 Trend 类产品的品质、风格等又难以满足消费者的需求和期待，导致购买 Trend 类产品的消费者失望而给出较低评分，低评分又进一步降低其他消费者的购买意愿，由此形成恶性循环。

(4) 各年龄段评分比较

表 4.5 Department_Name 分类下各年龄段段对不同产品的评分均值

age	Tops	Dresses	Bottoms	Intimate	Jackets	Trend
30 岁以下	4.14	4.14	4.28	4.32	4.18	3.55
31 至 40 岁	4.11	4.11	4.20	4.24	4.33	4.19
41 至 50 岁	4.10	4.12	4.29	4.30	4.13	3.85
51 至 60 岁	4.23	4.21	4.33	4.24	4.30	3.52
61 岁及以上	4.28	4.23	4.32	4.21	4.32	3.73

- **31 至 40 岁群体评分偏低**

Tops、Dresses、Bottoms 三类在 31 至 40 岁群体中的评分显著低于其他年龄段，其原因可能是这三类产品为日常生活的主要穿着，需满足消费者各种生活场景的需求。且该年龄段人口基数大，消费需求多样且差异明显，产品难以一一满足。

商家需根据风格、款式、适用场景等将产品进一步细分，做到精准推荐，精准满足。Tops、Dresses、Bottoms 三类在 50 岁以上群体的评分较高，专注于中老年、老年女装服饰的商家更易获得青睐。

- **Bottoms 和 Jackets 类在老年人(51 以上)评分较高**

Bottoms 类为下装裤子，Jackets 类一般为保暖所需产品，在老年群体中评分

都较高。可见，老年群体的主要要求为实用性，对其款式、风格等其余要求不高。商家应重点关注该特点。

4.1.3 基于小类的销量和评分分析

(1) 各小类的销量对比

对于非私人贴身衣物，在 Class_Name 小类分类下，Dresses、Knits、Blouses 是市场主力产品，Sweaters、Pants、Jeans 和 Fine gangu 四类产品销售量相差不大，构成市场第二主力。下图是各品类的销售量条形图：

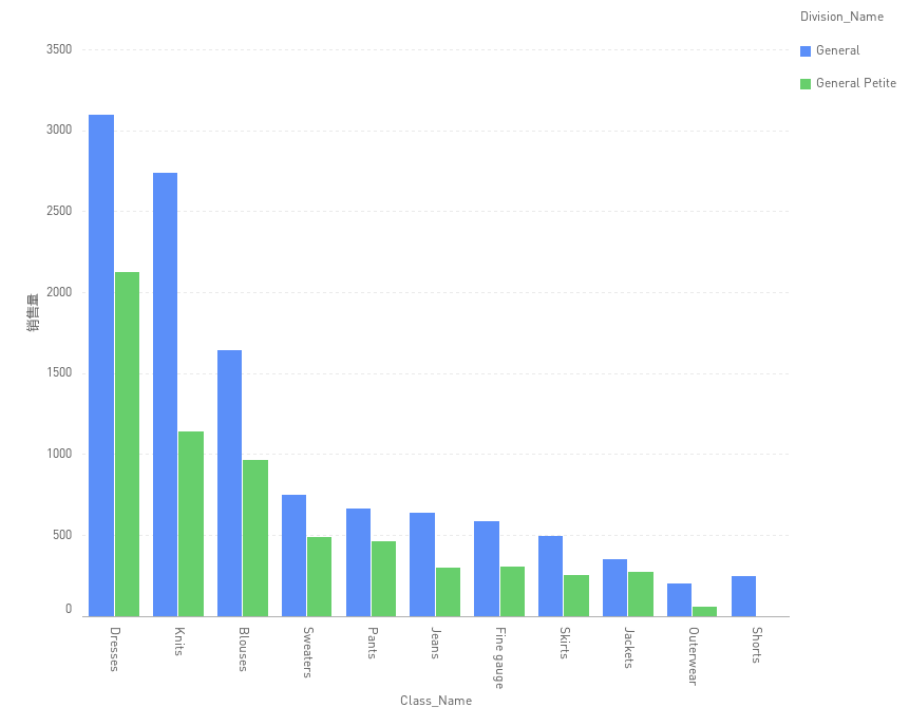

图 4.3 Class_Name 分类下各品类销售量

图形表明，Knits 类产品在 General 市场中竞争力与第一热销品 Dresses 相当，远高于第三热销品 Blouses。相反，在 General Petite 市场中，Knits 类产品竞争力与 Blouses 相近，显著低于 Dresses。对于专注于小个子市场的商家，应将 Dresses 作为主推产品，平衡 Knits 和 Blouses 的进货量和资源分配，同时视为第二主推产品。

(2) 各小类消费者年龄分布情况

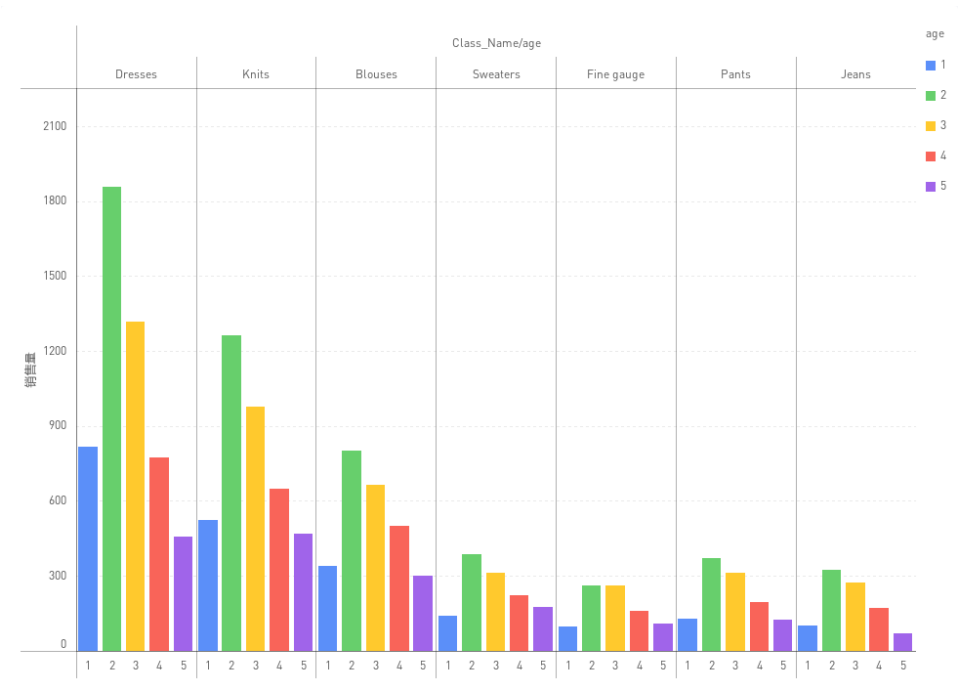

图 4.4 Class_Name 分类下各年龄段销售量柱形图

- **Dresses 类优势显著**

细分品类下各产品的年龄分布情况相近，第一消费主力为 31 至 40 岁，其次是 41 至 50 岁，51 至 60 岁。值得关注的是 Dresses 产品 30 岁以下群体消费力大于 51 至 60 岁群体。Fine gauge 产品 31 至 40 岁与 41 至 50 岁群体消费力相等。对于 61 岁及以上群体，Knits 为第一热销品，其余年龄段第一热销品均为 Dresses。

(3) 7 类热销款的市场总体评分

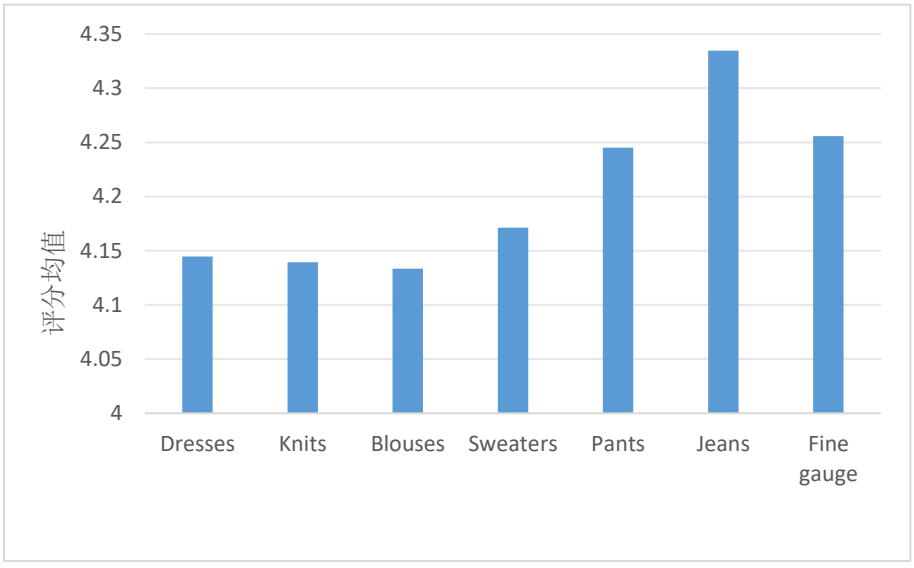

图 4.5 Class_Name 分类下市场评分均值柱形图

- **畅销款评分与销量成正比**

最畅销的三类 Dresses、Knits、Blouses 市场总体评分依次降低，与其销量成正比。分析其原因可能为消费者在购买这三类产品时，会考虑产品的好评情况做出购买决策，因此商家需采取适当策略提高好评率。综合来看，销量低的产品评

分反而高，这表明消费市场越大，消费者的需求越大、要求越高、竞争越大，主营这三类产品的商家可尝试减少产品种类，做某特定风格类产品，打造品牌特色，培养属于自己的客户群。

- **非畅销款评分相差较大**

Sweaters、Pants、Jeans 和 Fine gaugu 四类产品销售量相差不大，但 Sweaters 评分显著低于 Pants、Jeans。商家可尝试通过推出 Sweaters 与 Pants、Jeans 的成套搭配款以共同提高销量和评分。

4.1.4 不同年龄段客户产品购买和评论情况

(1) 各年龄段对大类产品的购买情况

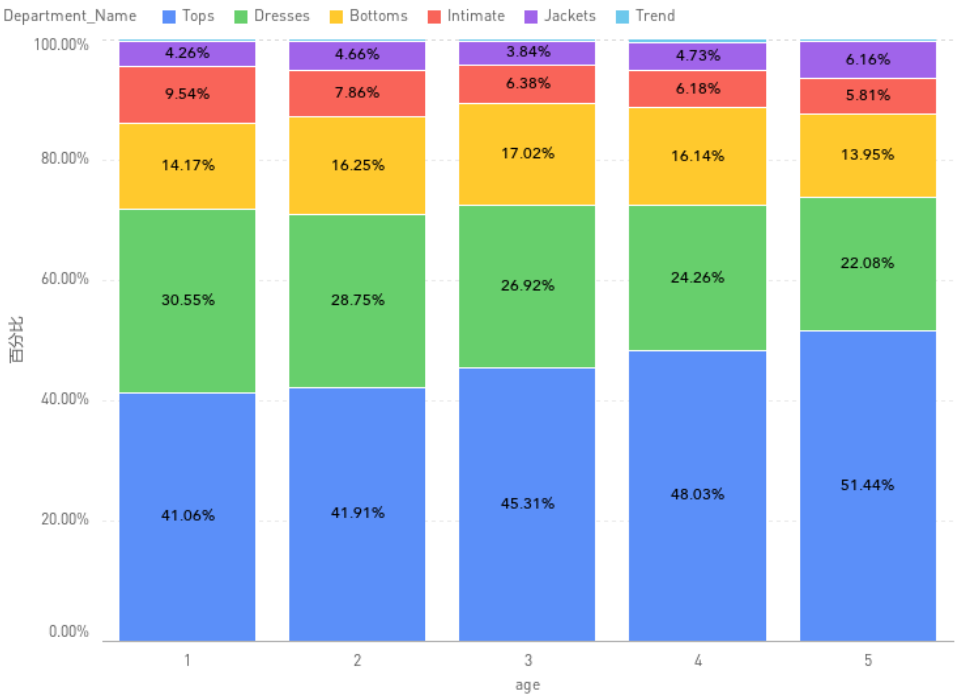

图 4.6 Department_Name 分类下各年龄段所购产品百分比堆叠图

上图展示了各个年龄段购买各产品的百分比。图形显示，Tops 类的百分比随年龄增加而逐渐增大，Dresses 类、Intimate 类的百分比随年龄增加而减小。一般来说，Tops 类比 Dresses 类更便利。随着年龄的增加，消费者对服装美观的需求降低、对便捷性的需求更高，因此年龄越大的群体中，Tops 类的百分比越高。

(2) 各年龄段对小类热销产品的购买情况

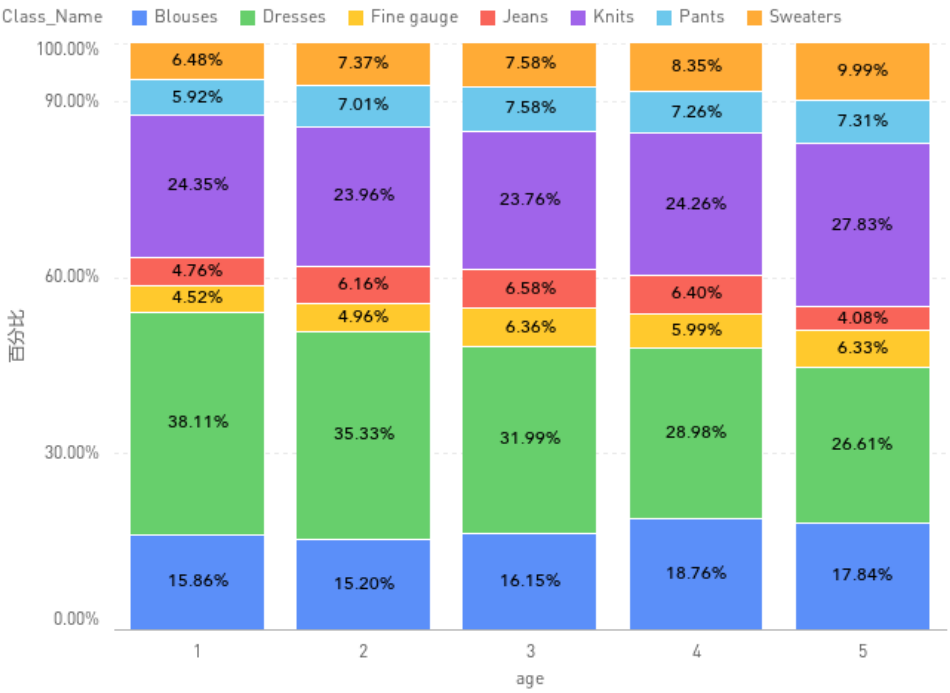

图 4.7 Class_Name 分类下各年龄段所购产品百分比堆叠图

- **50 岁以上群体 Knits 类产品购买量增加**

根据图表结果，50 岁以下群体购买 Dresses 品类的占比均在 30%以上，51 至 60 岁群体购买 Knits 品类的比例增加，61 岁及以上群体购买 Knits 品类的比例为 27.83%,高于 Dresses，同时 Sweaters 类的比例高于其他群体。因此主要目标客户群体为 50 岁以下的服装品牌需将 Dresses 作为销售重点；主要目标客户群体为 51 至 60 岁群体的服装品牌应提高 Knits 的产品比例；主要目标客户群体为 61 岁及以上的服装品牌应将 Knits 和 Dresses 二者作为主推品，并且适当提高 Sweaters 类的产品比例。

(3) 各年龄段评分偏好分析

为更详细地了解消费者打分偏好，本文分析了 Class_Name 分类下 7 类热销款的市场总体总评分和不同年龄段消费者评分，如下图所示：

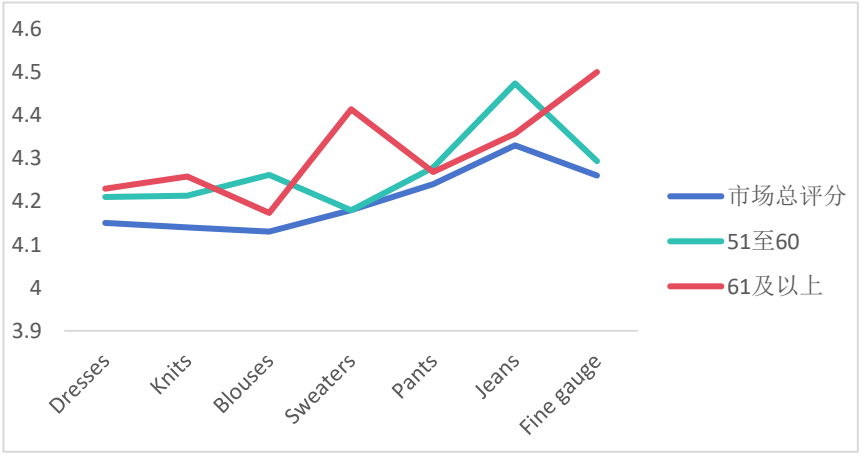

图 4.8 评分折线图

结果表明，51 至 60,61 及以上群体的满意度均高于市场总体满意度。61 岁及以上群体对 Sweaters 和 Fine gauge 的满意度最高，51 至 60 岁群体对 Jeans 满意度最高。

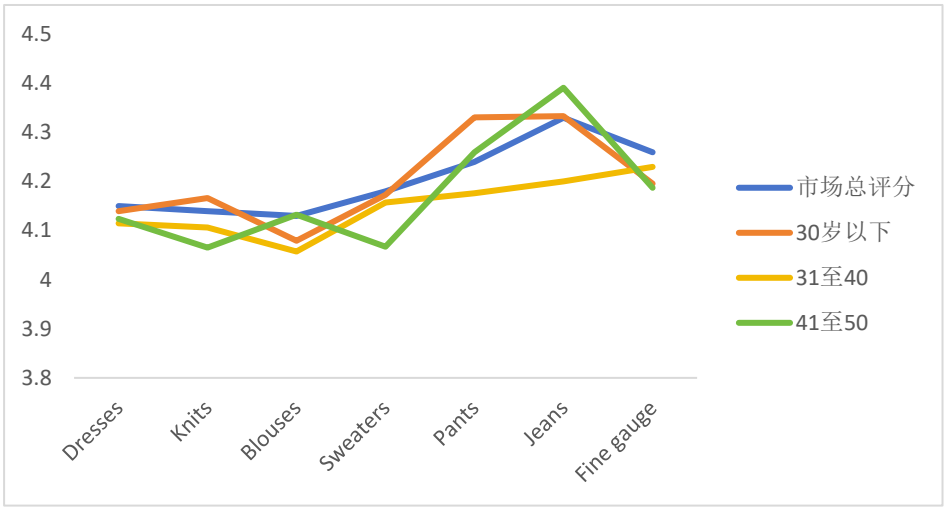

图 4.9 满意度折线图

31 至 40 群体的满意度均低于市场整体满意度，41 岁至 50 岁群体对 Knits、Sweaters、Fine gauge 的满意度较低。

(4) 各年龄段消费者评分分布情况

表 4.6 各年龄段各评分频数表

Age \ Rating	1	2	3	4	5
30 岁以下	93	176	380	539	1486
	3.50%	6.60%	14.20%	20.20%	55.60%
31 至 40 岁	232	485	856	1424	3469
	3.60%	7.50%	13.20%	22.00%	53.60%
41 至 50 岁	176	352	637	1117	2611
	3.60%	7.20%	13.00%	22.80%	53.40%
51 至 60 岁	125	188	348	659	1870
	3.90%	5.90%	10.90%	20.70%	58.60%
61 岁及以上	71	111	218	391	1979
	3.59%	5.61%	11.02%	19.76%	60.03%

- **市场总体评分高**

数据显示，大多数消费者不会轻易给出差评。各年龄段给出评分“1”的比例均低于 4%，给出评分“3”及以上的比例约占 90%。31 至 40 岁的群体相较于其他年龄段更易给出评分“2”。

- **年老者更易给好评**

50 岁以下的群体评分比较保守，给出中评“3”和良好“4”的概率较高。50 岁

以上群体更具“表扬性”，给出满评“5”的概率接近 60%。这是由于不同年龄段的评价心理不同所造成的。一般来说，年龄较大的人物欲较低，要求较低，对事物的包容程度高，对产品有成长性期待，给出好评的概率大。而年轻人评价呈现出“比较型”，会综合考虑产品的优缺点，做出的评价更中肯。基于此，商家需重点推荐老年人的评论，提高产品印象分。此外，商家应从年轻消费者的评论中发现并总结问题，不断改善产品以满足消费者期待。同时，可推出“好评有礼”活动，激励打中评的消费者提高评分。

4.2 基于 K-means++聚类分析的消费者画像模型构建

4.2.1 产品市场潜力和消费者购买倾向的定义

- **产品市场潜力**

某产品下各年龄段的购买量占该产品的总购买量之比，可反映该产品的消费者年龄分布情况，而产品消费者年龄分布情况在一定程度上可视为该产品在各年龄段的市场潜力。

- **消费者购买倾向**

某年龄对某产品的购买量占该年龄的总购买量之比，在一定程度上可视为该年龄对某产品的购买倾向。

4.2.2 产品市场潜力和消费者购买倾向的分析步骤

(1) 原始数据分组

原始数据共 19202 条，数据量大，但关于消费者的指标较少，不易研究。因此首先将同一年龄购买同一品类的消费者聚为一类，以评分均值代表各类消费者的满意度，以评分均值的方差描述每一类消费者需求的内部差异性。在这样的简化方法下，将 19202 个消费者分为 993 类。

(2) 利用聚类算法刻画消费者画像

进一步，我们采用 K-means++聚类算法刻画消费者画像。将消费者年龄、购买倾向、产品市场潜力、满意度评分、评论获赞数、需求差异性 6 个因子作为因子尝试聚类，利用 SPSS 对数据进行预处理之后得到以下结论。

表 4.7 聚类单因素方差分析

	聚类类别（平均值±标准差）			F	P
	类别 1(n=358)	类别 2(n=356)	类别 3(n=278)		
A	29.358±5.989	49.817±5.994	71.338±8.217	3092.321	0.000***
Positive_Feedback_Count	1.994±1.898	2.31±1.868	2.625±2.845	6.498	0.002***
Recommended_IND	0.834±0.156	0.831±0.169	0.853±0.244	1.196	0.303
购买倾向	0.061±0.084	0.059±0.072	0.122±0.161	33.119	0.000***
市场潜力	0.029±0.075	0.022±0.01	0.007±0.006	19.164	0.000***
平均值(Rating)	4.247±0.495	4.175±0.605	4.298±0.78	3.157	0.043**

需求差异性	1.103±0.942	1.165±1.0	0.771±1.037	13.774	0.000***

(3) 聚类结果的差异性分析

分析每个分析项的 P 值是否显著(P<0.05)，若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据均值±标准差的方式对差异进行分析，反之则表明数据不呈现差异性。

变量 Recommended IND 的 P 值为 0.303，水平上不呈现显著性，不能拒绝原假设，说明变量(Recommended IND)在聚类分析划分的类别之间不存在显著性差异；其余 5 个变量的 P 值均小于于 0.05，水平上呈现显著性，拒绝原假设，说明变量在聚类分析划分的类别之间存在显著性差异。

(4) 聚类结果的评价

表 4.8 数据评价指标

轮廓系数	DBI	CH
0.527	0.591	2714.768

轮廓系数的取值范围是[-1,1]，同类别样本距离越相近不同类别样本距离越远，分数越高，聚类效果越好。DBI 用来衡量任意两个簇的簇内距离之后与簇间距离之比，该指标越小表示聚类效果越好。CH 用来衡量聚类结果的质量，通过计算类内各点与类中心的距离平方和来度量类内的紧密度，通过计算类间中心点与数据集中心点距离平方和来度量数据集的分离度，CH 指标由分离度与紧密度的比值得到，CH 越大表示聚类效果越好。通过评价指标分析，得出该分类结果较好。

4.2.3 消费者聚类分析结果

通过聚类分析，我们将消费者分成 3 类最为合理，如下图所示：

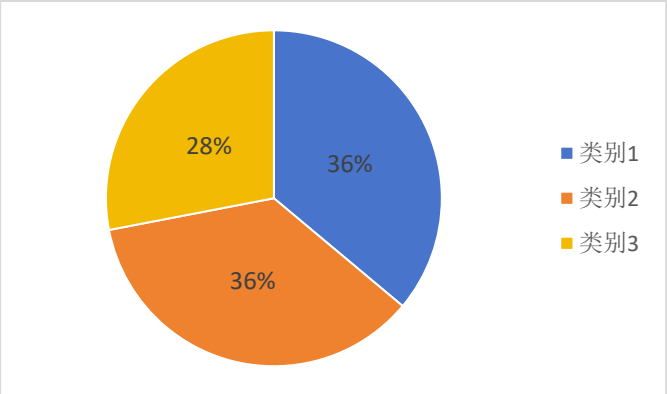

图 4.10 消费者分类汇总饼状图

(1) “重点稳固”型

类别 1 可概括为“重点稳固”型，是市场长期价值最高的客户。

这类群体主要为 24 至 35 岁的年轻人，该群体人口基数大，生活场景多样，所选择的服装品类广泛，需求量高，各类产品的在该群体的市场潜力都较高。同时，这类群体对于评论的积极性较低，目前对市场产品的满意度较高，但不同消费者的需求呈现较大的差异性。

针对该群体，需要重点关注时尚潮流，着重提高产品在款式、风格方面上的竞争力。一般情况下，该群体对服装的“换新”、“换款”行为更多。因此，商家应向该群体推荐更多种类、更多风格的产品，并且需要提高产品的性价比，以较低利润获得更大销量。其次，需要丰富产品种类，突出产品特点，避免产品同质化严重，以满足不同消费者的个性化需求。同时，可通过“晒单有礼”、优化评价流程等方法激励消费者参与评价。

(2) “重点培养”型

类别 2 可概括为“重点培养”型，是市场未来潜在价值最高的客户。

这类群体消费潜力较强，评论获赞量高，但市场满意度偏低。

针对该类群体，需要重点提高产品的品质以赢得消费者青睐。该群体主要为 45 至 54 岁的中年人，有一定经济基础，可作为中高端产品的主要目标客户，提高单品利润。在推荐策略上，可将价格较高的优质品牌产品推荐给该类群体。此外，因该群体的满意度较低，商家可以设置奖励机制，提高该群体的好评率。

(3) “精准培养”型

类别 3 可概括为“精准培养”型，是市场近期最可能吸收的客户。

这类群体对产品的选择较谨慎，选择的产品较固定，需求比较一致，购物满意度高，评论积极性高，较其他群体更容易满足。

针对该群体，需要细化产品线，做到小而精，精准服务，精准推广。在向该群体推荐时，应多推荐不同价位的同类产品以便消费者进行比较。同时，可以提高该群体评论的推荐比例，帮助消费者更好地了解产品，增加好感度。

5 运营情况分析

5.1 统计性描述分析

5.1.1 服装分类情况

我们针对筛选过后的数据进行了统计分析，根据 Department_Name 和 Class_Name 绘制成如下矩形树状图。

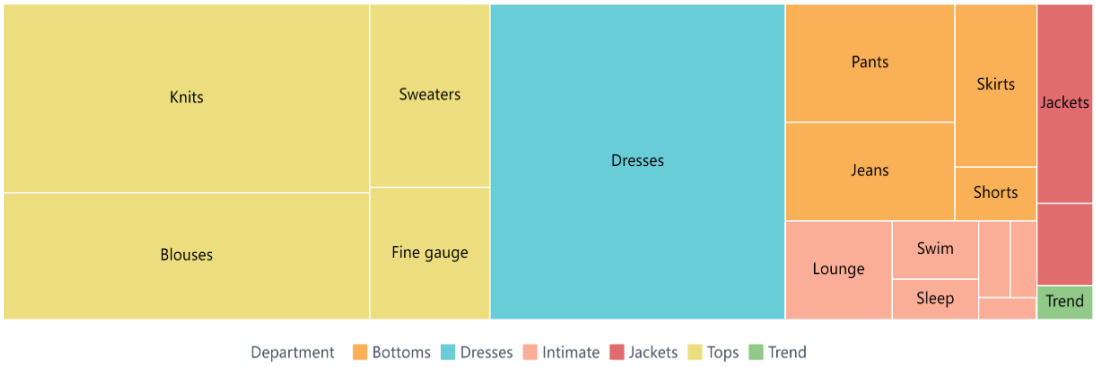

图 5.1 服装分类矩形树状图

其中，我们发现 Tops 类女装占比最大，为 44.66%，而剩下几种分别为 Dresses（27.1%），Bottoms（15.88%），Intimate（7.23%），Jackets（4.58%），Trend（0.55%）。

5.1.2 消费者对于不同大类服装的满意度

(1) 满意度的定义

女装市场整体评分较高，均值为 4.18，其中各大类的评分存在较大差异。

表 5.1 女装各大类评分均值

Tops	Dresses	Bottoms	Intimate	Jackets	Trend
4.16	4.15	4.27	4.27	4.26	3.87

对于 Tops、Dresses、Bottoms、Intimate、Jackets 和 Trend 六个大类，综合考虑用户的评分和是否推荐，由于存在部分用户评分较高，但是对服装却并不推荐和评分低却推荐服装的情况，这并不能说明用户对该服装是满意的，只有当用户自己认可服装，给予其高评分（4 和 5），并且认为该服装具有普适性和魅力，他人可能同样会喜欢该服装，愿意分享给他人时，我们才认为该用户对服装是满意的。因此我们定义评分为 4 和 5 并且愿意推荐的用户是满意的，**满意度则是满意的用户在总体用户中的占比。**

(2) 各大类的满意度

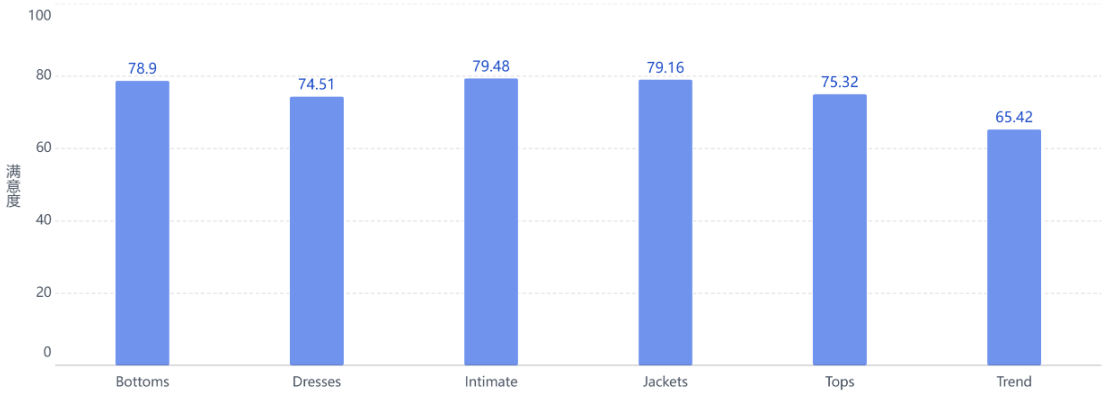

图 5.2 各大类满意度柱状图

其中，各种服装满意度大体相同，并且均在 65%以上，由此可以看出大部分女装的口碑很不错，女装市场的潜力巨大，值得我们去深入挖掘其中的数据，为企业的发展提供帮助。在六个大类中，Intimate 的满意度略高，为 79.48%，Trend 最低，为 65.42%。

5.1.3 消费者对于不同大类下细分类目服装的满意度

在不同的大类下，针对每一个细分类目，我们分别统计了客户的满意度，结果如下图所示。

(1) TOPS 大类下的满意度

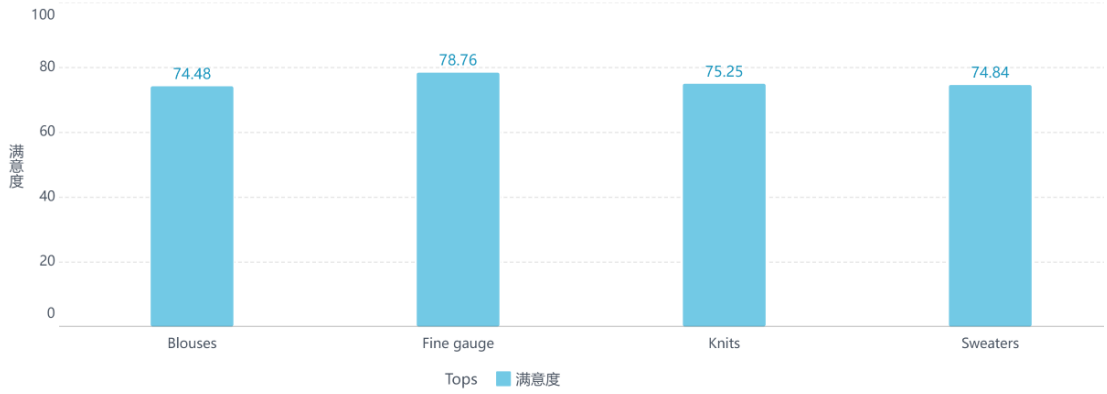

图 5.3　Tops 大类下细分类目满意度柱状图

在 Tops 大类下，Blouses、Fine gauge、Knits 和 Sweaters 四者的销量均稳居前列。其中 Blouses、Knits 和 Sweaters 的满意度相差不大，在 3/4 左右，而 Fine gauge 的满意度略高于三者，达到了 78.76%，由此可以看出，消费者对于 Fine gauge 类型的服装更加喜爱，并且更愿意向他人推荐。

(2) Dresses 大类下的满意度

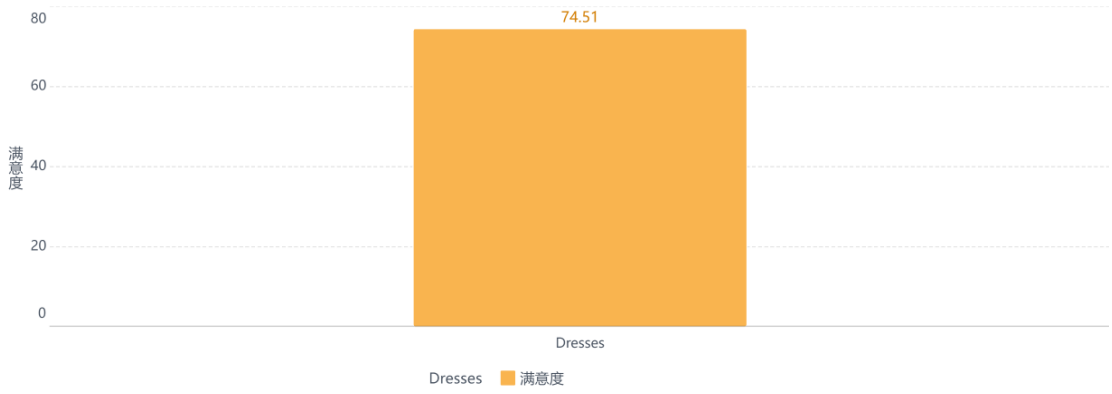

图 5.4　Dresses 大类下细分类目满意度柱状图

Dresses 大类下并没有其他的细分类目，Dresses 服装能在细分类目销量位居榜首的情况下高达 74.51%的满意度，可见 Dresses 在女性群体中的热度非常高，是女性购买衣服的不二选择。而这一类型女装也绝对是公司发展相关业务的一大至关重要的方向。

(3) Bottoms 大类下的满意度

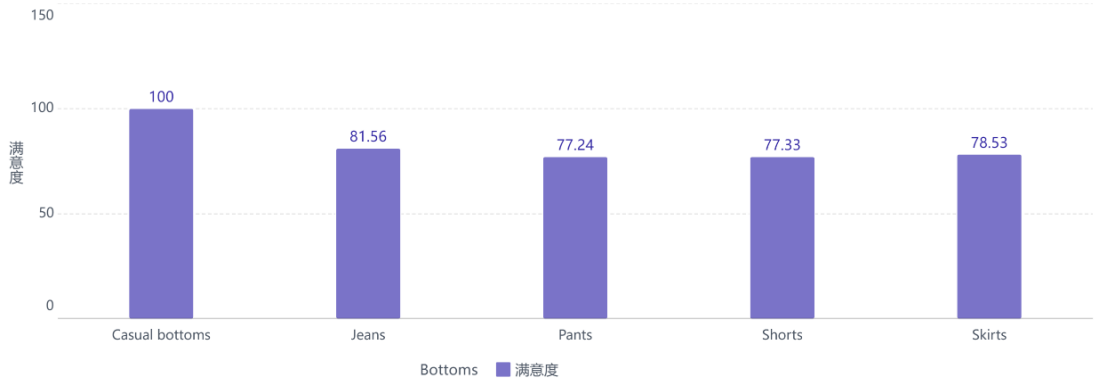

图 5.5　Bottoms 大类下细分类目满意度柱状图

Bottoms 大类下，尽管满意度是 100%，但 Casual bottoms 的销量只有一件，该评分不能代表用户的真实满意情况。剩下的四个类型满意度均在 77%以上，可谓非常之高，甚至 Jeans 的满意度突破了 80%，达到了 81.56%。

(4) Intimate 大类下的满意度

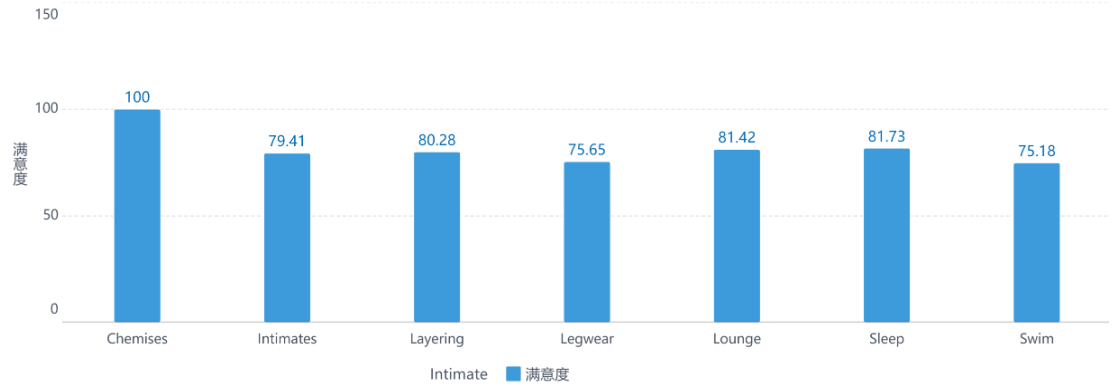

图 5.6　Intimate 大类下细分类目满意度柱状图

Intimate 大类下，和 Casual bottoms 类似，Chemises 的销量只有一件，该评分同样不能代表用户的真实满意情况。其余六个类型满意度差距较大，Sleep 的满意度在细分类目下位居榜首，高达 81.73%，而 Swim 却仅有 75.18%。

(5) Jackets 大类下的满意度

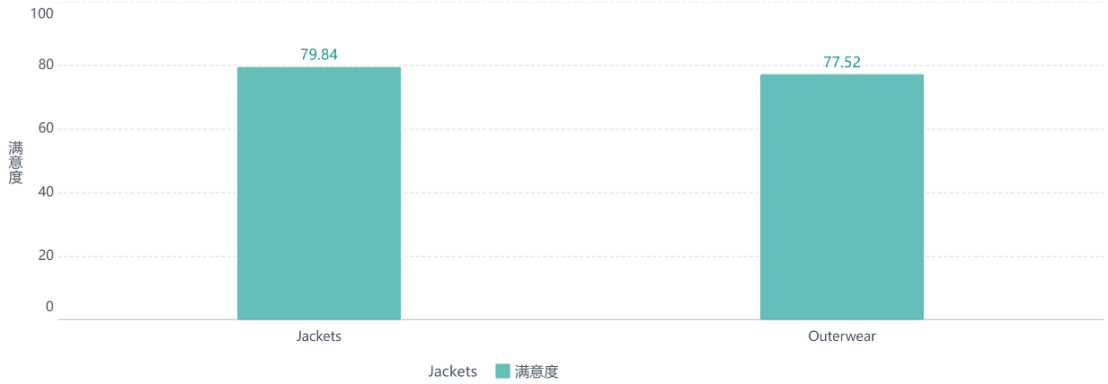

图 5.7 Jackets 大类下细分类目满意度柱状图

Jackets 大类下，Jackets 比 Outerwear 略高，分别为 79.84%和 77.52%。

(6) Trend 大类下的满意度

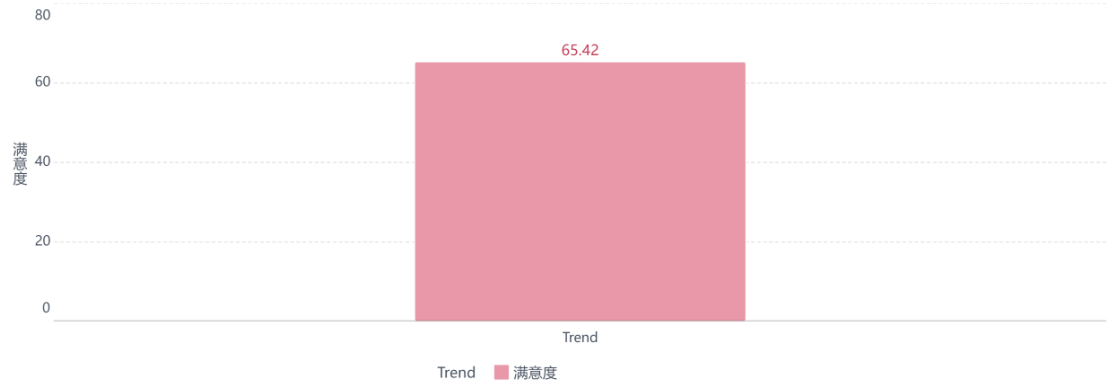

图 5.8　Trend 大类下细分类目满意度柱状图

作为销量最少的大类，Trend 的满意度也是最低的，仅仅 65.42%，是唯一一个满意度在 70%以下的。如果公司尚在经营相关的业务，有必要重新思考是否换到其他赛道。

5.2 最受欢迎服装大类及其细分类目

由于销售量和满意度的权重并未确定，为了综合考虑女装销售量和满意度对其受欢迎程度的影响，我们采用**基于熵权法的 TOPSIS 模型**。它可以有效避免数据的主观性，不需要通过检验，能够很好的刻画多个影响指标的综合影响力度，并且评估程序简单，计算过程简单易懂。

5.2.1 受欢迎程度的计算步骤

(1) 基础数据

各大类及其细分类目的销售量和满意度的数据如下表所示：

表 5.2 女装各大类及其细分类目销售量及满意度

大类	销售量	满意度	细分类目	销售量	满意度
Tops	8611	75.32%	Blouses	2606	74.48%
			Fine gauge	890	78.76%
			Knits	3879	75.25%
			Sweaters	1236	74.84%
Dresses	5225	74.51%	Dresses	5225	74.51%
Bottoms	3061	78.9%	Casual bottoms	1	100%
			Jeans	938	81.56%
			Pants	1125	77.24%
			Shorts	247	77.33%
			Skirts	750	78.53%
Intimate	1394	79.48%	Chemises	1	100%
			Intimates	136	79.41%
			Layering	71	80.28%
			Legwear	115	75.65%
			Lounge	592	81.42%
			Sleep	197	81.73%
			Swim	282	75.18%
Jackets	883	79.16%	Jackets	625	79.84%
			Outerwear	258	77.52%
Trend	107	65.42%	Trend	107	65.42%

由于 Bottoms 大类中 Casual bottoms 和 Intimate 大类中的 Chemises 的数量只有 1 件，两者的客户均为满意，满意度为 100%，但是由于数量太少，会对我们的数据处理造成严重的影响，故不将其纳入最受欢迎服装考虑范围之内。

并且，求取大类和细分类目中最受欢迎的服装类型方法相同，为了简便说明，我们下面只利用大类的数据进行演示。

(2) 基础数据的标准化处理

销售量为整数，满意度为百分数，为了消去销售量和满意度两个指标量纲的影响，我们利用标准化公式

$$Z_{ij} = X_{ij} / \sqrt{\sum_{i=1}^{n} X_{ij}^2}$$

得到二者数据标准化后的值

表 5.3 女装各大类销售量及满意度标准化后的数据

大类	销售量	满意度
Tops	0.8082	0.4066

Dresses	0.4903	0.4022
Bottoms	0.2872	0.4259
Intimate	0.1308	0.4291
Jackets	0.0829	0.4273
Trend	0.01	0.3532

(3) 满意度概率矩阵的计算

接下来，我们计算概率矩阵 P，利用公式

$$P_{ij} = Z_{ij} / \sum_{i=1}^{n} Z_{ij}$$

得到下列数据

表 5.4 女装各大类销售量及满意度概率矩阵

大类	销售量	满意度
Tops	0.4467	0.1663
Dresses	0.2710	0.1645
Bottoms	0.1587	0.1742
Intimate	0.0723	0.1756
Jackets	0.0458	0.1748
Trend	0.0055	0.1445

(4) 信息熵的计算

之后利用下列公式计算销售量和满意度两个指标的信息熵。

$$e_j = -1/\ln n * \sum_{i=1}^{n} (P_{ij} * \ln P_{ij})$$

表 5.5 女装销售量及满意度的信息熵

指标	销售量	满意度
信息熵	0.7622	0.9988

(5) 信息效用值的计算

计算信息效用值。

$$d_j = 1 - e_j$$

表 5.6 女装销售量及满意度的信息效用值

指标	销售量	满意度
信息效用值	0.2378	0.0012

(6) 熵权的计算

将信息效用值归一化，得到熵权，即销售量和满意度在本模型中的权重大小。

$$w_j = d_j / \sum_{j=1}^{m} d_j$$

表 5.7 女装销售量及满意度的熵权

指标	销售量	满意度
熵权	0.995	0.005

(7) 理想最优解和理想最劣解的定义

定义最大值Z^+即理想最优解和最小值Z^-即理想最劣解。

表 5.8 女装销售量及满意度的最大值和最小值

指标	销售量	满意度
最大值	0.8082	0.4291
最小值	0.01	0.3532

(8) 销售量及满意度与理想解的距离

计算不同服装大类的销售量和满意度与理想最优解和理想最劣解之间的距离D_i^+和D_i^-。当服装大类的销售量和满意度距离理想最优解越近，距离理想最劣解越远时，说明该种类别最受大众欢迎。

$$D_i^+ = \sqrt{\sum_{j=1}^{m} w_j * (Z_j^+ - Z_{ij})^2}$$

$$D_i^- = \sqrt{\sum_{j=1}^{m} w_j * (Z_j^- - Z_{ij})^2}$$

表 5.9 女装大类与最大值和最小值之间的距离

大类	D_i^+	D_i^-
Tops	0.0016	0.7962
Dresses	0.3171	0.4791
Bottoms	0.5197	0.2766
Intimate	0.6757	0.1206
Jackets	0.7235	0.0729
Trend	0.7962	0

(9) 受欢迎得分的计算

计算受欢迎程度的得分以及归一化后的得分

$$S_i = D_i^- / (D_i^+ + D_i^-)$$

$$S_i' = S_i / \sum_{i=1}^{n} S_i$$

表 5.10 女装大类的受欢迎程度得分及其归一化后的值

大类	S_i	S_i'
Tops	0.998	0.4557
Dresses	0.6017	0.2747

Bottoms	0.3474	0.1586
Intimate	0.1515	0.0692
Jackets	0.0915	0.0418
Trend	0	0

同样的方法计算可得女装细分类目的受欢迎程度得分及其归一化后的值如下表所示。

表 5.11 女装细分类目的受欢迎程度得分及其归一化后的值

细分类目	S_i	S_i'
Blouses	0.4918	0.1407
Fine gauge	0.1589	0.0455
Knits	0.7388	0.2114
Sweaters	0.226	0.0647
Dresses	0.9984	0.2857
Jeans	0.1681	0.0481
Pants	0.2044	0.0585
Shorts	0.0341	0.0098
Skirts	0.1316	0.0377
Intimates	0.0128	0.0037
Layering	0.0033	0.0009
Legwear	0.0088	0.0025
Lounge	0.1012	0.0290
Sleep	0.0247	0.0071
Swim	0.041	0.0117
Jackets	0.1074	0.0307
Outerwear	0.0362	0.0104
Trend	0.0069	0.0020

5.2.2 受欢迎程度的结果分析

经过(1)-(9)的计算，我们将女装大类和细分类目的受欢迎程度得分及其归一化后的值汇总如表 5.10 和表 5.11 所示。

S_i'越大，说明该类型女装越受欢迎。由表 5.10 和表 5.11 得知 **Tops**（S_i'为 0.4557）在女装大类中最受欢迎，**Dresses**（S_i'为 0.2857）在女装细分类目中最受欢迎。

6 客户评论细粒度情感分析

对客户推荐行为、客户关注点及客户各关注指标的评论情感三个方面进行分析，构成客户评论的细粒度情感分析。

6.1 基于传统情感分析和二元 Logistic 回归模型的推荐机制分析

探究哪些因素会促使消费者愿意“推荐”某个商品，对于商家理解顾客忠诚度和口碑传播具有重要意义，我们利用回归分析方法对该问题进行研究，以明确推荐机制，即消费者基于自身的满意度和体验，主动向他人推荐产品或服务的心理和行为过程。理解推荐机制对商家至关重要，它涉及识别和增强那些鼓励顾客推荐产品的因素，同时减少负面影响。这有助于商家优化产品服务，制定有效营销策略，提升顾客满意与忠诚度，促进新顾客增长和留存。

6.1.1 模型选择和变量设定

我们建立回归模型去为商家探究 Recommended IND 的影响因素，进而促使商家采取措施提高顾客推荐率。假设其受到客户评论、客户满意度、客户年龄、衣服种类、服装尺寸的影响，将这五个因素作为自变量。探究各影响因素的显著程度，判断推荐的共性特征和原因，确定推荐机制。

在五个自变量中，衣服种类 Class Name、服装款式 Division Name 属于多分类变量。如果直接编码 1、2、3.....，令其作为自变量带入模型进行计算，会导致回归结果和真实情况存在较大误差。因此,本研究采用统计学上的标准做法，将这两个定类变量转化为**虚拟变量**进行拟合,然后再进行分析。

由于 Recommended IND 只涉及两种回答，推荐或者不推荐，属于明显的二分类因变量，因此我们建立二元 Logistic 模型，用二元 Logistic 进行回归分析，探究前述变量和是否推荐的因果关系。同时二元 Logistic 回归是一个概率模型，也可以利用它根据潜在客户的特征预测推荐服装的概率。

6.1.2 模型的建立

根据上述的自变量和因变量定义，本次调研建立的二元 Logistic 回归模型下：

$$\ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right)=\beta_0+\beta_1 score+\beta_2 Age+\beta_3 Rating+$$
$$\beta_4 Class\_Name+\beta_5 Division\_Name+\varepsilon$$

其中，β_i 为二元回归的自变量系数，ε 为随机误差。

6.1.3 模型及回归系数检验

(1) 模型检验

由于二元 Logistic 回归是基于概率进行判断的模型，采用的是极大似然估计法(MLE)，也就是选择使得似然函数最大的参数估计值。我们进行模型整体的回归效果检验，结果如下表所示:

表 6.1 二元 Logistic 回归模型似然比检验结果

模型	-2 倍对数似然值	卡方值	df	p	AIC 值	BIC 值
仅截距	18260.8					
最终模型	5462.7	12798.1	5	0	5474.7	5521.9

通过 P 检验， P<0.05 模型有效。二元 Logistic 回归结果的准确度如下表:

表 6.2 二元 Logistic 回归模型预测准确度

		预测值		预测准确率	预测错误率
		0	1		
真实值	0	2841	657	81.22%	18.78%
	1	641	15142	95.94%	4.06%
汇总				93.27%	6.73%

由上表中的结果可知，不推的回归结果正确率为 81.22%，推荐的准确率为 95.94%，模型综合准确率高达 93.27%，具有高度可靠性和稳定性。

(2) 回归系数显著性检验

回归系数的输出结果表 6.3 所示，存在部分回归系数不显著。但是只要当某一个因素下的变量对应的系数显著时，就认为该因素在方程中是显著的。所以可以看出，score, Age, Rating 这 3 个因素都是显著的。

表 6.3 二元 Logistic 回归分析结果

自变量	回归系数	标准误差	z 值	*p* 值	OR 值	OR 值 95% CI
score	0.813	0.08	10.099	0	2.255	1.925 ~ 2.640
Age	0.011	0.003	3.837	0	1.011	1.006 ~ 1.017
Rating	3.112	0.061	50.709	0	22.47	19.93 ~ 25.35
Class Name	-0.004	0.007	-0.581	0.561	0.996	0.983 ~ 1.009
Division Name	0.059	0.059	1.005	0.315	1.061	0.945 ~ 1.192
截距	-10.488	0.258	-40.61	0	0	0.000 ~ 0.000

从上表可以看出 score, Age, Rating, Class Name, Division Name 可以解释 Recommended IND 的 0.70 变化原因。从上表可知模型公式为：

$$\ln\left(\frac{p}{1-p}\right) = -10.488 + 0.813 \times score + 0.011 \times Age + 3.112 \times Rating$$
$$- 0.004 \times Class\_Name + 0.059 \times Division\_Name$$

其中 p 代表 Recommended IND 为 1 的概率，$1-p$ 代表 Recommended IND 为 0 的概率。

从系数和显著性可以看出 score, Age, Rating 会对 Recommended IND 产生显著的正向影响关系。但是 Class Name, Division Name 并不会对 Recommended IND 产生影响关系。

6.1.4 二元 Logistic 回归模型结果分析

利用 SPSS 软件对进行二元 Logistic 回归，观察和分析模型的结果可以发现，从顾客角度看，顾客的年龄以及对服装的评价对是否推荐有显著的影响，从服装角度来看，服装的种类和大小对是否推荐无显著影响。

SPSS 模型回归结果中的 EXP(B)(即 OR 值）为发生比例，当顾客推荐概率较小的时候(一般认为小于 0.1)，可近似认为 EXP(B)与发生概率之比非常接近。因此，本研究可以将 EXP(B)值解释为其他变量不变的情况下，自变量每改变 1 个单位，变化后的推荐数量是变化前的 Exp(B)倍。

(1) 成年或更年长女性倾向于推荐女装

Age 的优势比(OR 值)为 1.011，意味着 Age 增加一个单位时，Recommended IND 的变化(增加)幅度为 1.011 倍，这个增长幅度相较于顾客情感和满意度的影响来说较小，但它表明成年女性更倾向于进行推荐行为，尽管这种倾向性在不同年龄段中的差异不大。这与之前客户画像分析中所提到的大龄顾客往往满意度更高的结论相符。

(2) 高评分显著增加推荐可能性

Rating 的优势比为 22.474，意味着评级每提升一个单位，推荐的可能性增加 22.474 倍。这强调了顾客评分对于推荐行为的重要性，高度满意的顾客更倾向于推荐他们购买的服装。

(3) 积极评论极大促进推荐行为

情感得分 score 的优势比为 2.255，表明情感得分每增加一个单位，推荐的可能性增加 2.255 倍。这表明顾客的正面情绪与对服装的推荐密切相关，喜爱和赞赏的情绪会激励顾客进行推荐。

- **深入挖掘**

通过再次分析销量较高的 **Dresses 类**和 **Blouses 类**服装的数据，我们深入探讨了 score, Age, Rating 对推荐影响的差异，发现不同类别间的各因素相关性指标和回归系数**差异不大**，证实了我们的初步分析揭示的规律具有普遍性。这些发现进一步深化了我们对影响服装销售因素的理解，尤其是年龄因素对消费者偏好的影响，为服装行业的营销策略提供了重要指导。

总而言之，顾客是否推荐产品受到满意度和情感态度的显著影响。高度满意且持正面情感态度的顾客更倾向于向他人推荐服装。这强调了商家为增加顾客推荐行为，需致力于提升购物满意度和正面情感体验。通过提高产品质量、优化客户服务和加快物流速度等措施，可以有效提升顾客的满意度和正向情感反馈，从而鼓励她们推荐产品。对于女装市场而言，这意味着除了注重产品设

计和质量外，还需确保顾客购物体验能满足她们的预期，以激发推荐意愿。

6.2 LDA 主题挖掘算法

6.2.1 LDA 模型构建

LDA (Latent Dirichlet Allocation)是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。通俗来说，LDA 是一种主题模型,可以将文档集中的每篇文档按照概率分布的形式给出；是一种无监督学习，在训练时无需手工标注的训练集，只需要文档集和指定主题的个数；是一种典型的词袋模型，认为一篇文档是由一组词组成的集合，词语词之间没有顺序。

评论中词的概率公式为:

$$p(x \mid d) = \sum p(x \mid t) \times p(t \mid d)$$

其中，$p(x \mid d)$表示主题词的分布，$p(t \mid d)$表示评论主题的分布，通过概率分布我们可以了解到顾客对跨境电商女装主要的关注点及建议。

6.2.2 LDA 模型分析步骤

我们团队将预处理过后的有效评论数据进行汇总，使用 **Python 的 sklearn** 自带语料库进行 LDA 主题分析，挖掘分析大众用户对直播健身的热门关注点。具体分析步骤如下，相应的 Python 代码见附录三所示。

Step 1:对文本进行清洗及分词。通过爬虫搜集到的开放题数据较为杂乱，进行文本分词后提取有效词进行分析。

Step 2:加载分词文件,构建词典及向量空间。通过字典的词频统计分出高频词和低频词，输出词典结果。

Step 3:进行 LDA 分析，这里使用主题一致性指数和困惑度指数来确定合理的主题数目。

Step 4:使用暴力搜索来确定合适的主题模型。对于评论数据，我们使用该 LDA 模型只能一次确定一个主题数，无法选择最佳的主题，通过暴力搜索确定最佳模型。最后输出可视化结果。

6.2.3 LDA 模型结果分析

(1) 顾客评价的关注点

通过 LDA 文本可视化,如图 6.1 所示,线上购买服装的顾客普遍给予高度评价，其中“love（喜爱）”和“like（喜欢）”等正面词汇占据了主要位置。在女装市场，"Dress（连衣裙）"、"shirt（衬衫）"和"sweater（毛衣）"等服装类别占据了主导地位。顾客在赞扬服装时，常关注于“color（颜色）”、“size（尺寸）”、“design（设计）”等方面，用“very cute（非常可爱）”、“true（准确无误）”和“well（质量上乘）”等词汇表达满意度。

然而，也有顾客提出建议和问题反馈，指出实际购买的服装与预期存在差异。例如尺寸问题如“large（过大）”或“long（过长）”以及“shoulder（肩部）”、

“knee（膝部）”等特定部位的设计问题。

词汇如“wish（期望）”和“better（更好）”反映了顾客对电商平台的期待，希望它们能够改进不足，不断进步。

总之，跨境电商在女装领域受到了广泛认可和喜爱。尽管存在一些问题和建议，顾客们的反馈呈现出一个蓬勃发展的市场景象，显示了顾客对于电商平台改善服务和产品的期待和信心。

图 6.1 LDA 文本可视化图

(2) 顾客评价的主题

将经过数据预处理后的有效评论数据作为数据源，利用 Python 训练 LDA 主题模型，借助困惑度和主题一致性确定主题数量，对主题进行标识并挖掘其主题关键词，对处理后的文本数据进行建模，不断调整主题个数，以得到较为独立的主题。最终选取主题个数为 4 个，相关主题气泡图如图 6.2 所示。左边为主题间距离图，即多维标度图，右边为相关性指标 a=1 时的相关术语。

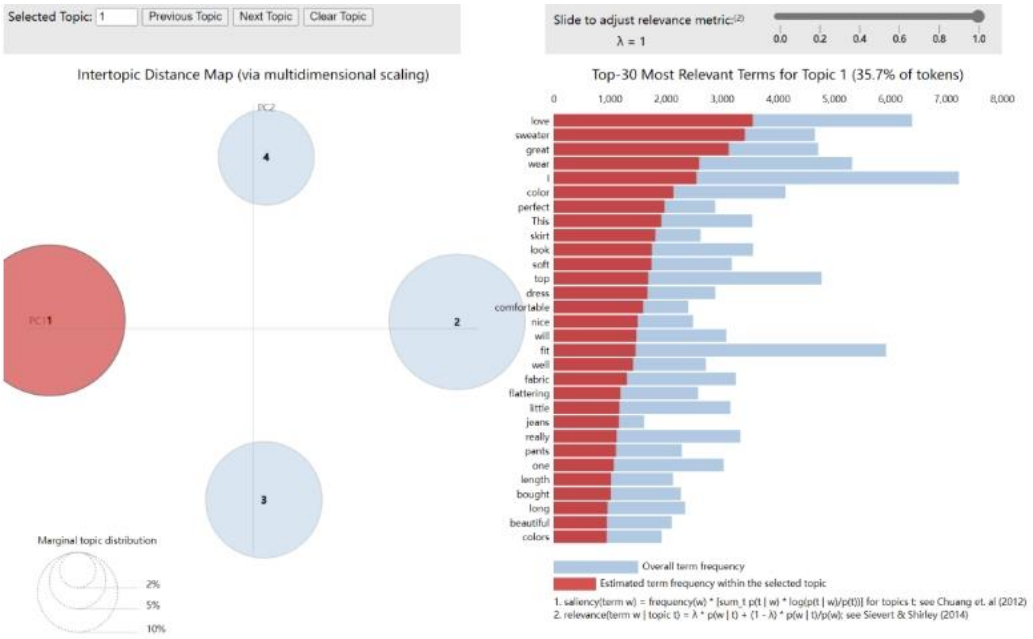

图 6.2 相关主题气泡图

在评论文本数据中可以看出，主题气泡完全分散，没有出现重叠部分，说明提取的主题代表性较好。在基于 LDA 模型的模拟训练后，得到“主题--词组”的概率分布，每个主题内的主题词组根据其概率大小排序，得到如下 4 个主题，其中核心主题词组如图 6.3 所示:

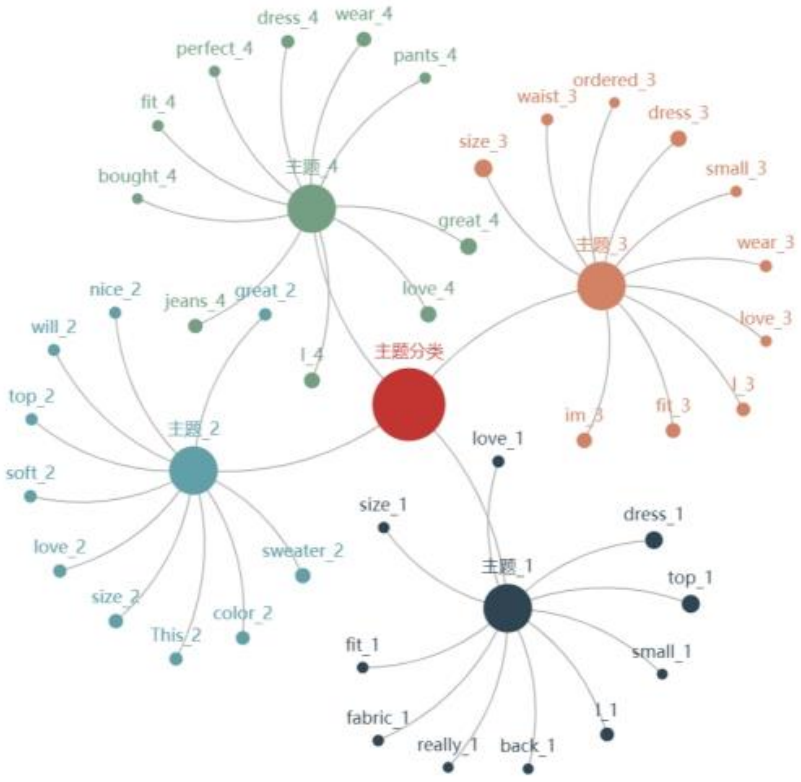

图 6.3 LDA 主题分析结果

结果分析：

客户会针对服装的具体方面进行评价，而并非简单地描述整体满意度。根据开放题评论进行潜在主题关键词提取发现，客户评价角度越来越多元，情感表达也越来越丰富。主题 1 主要描述观众对服装**样式**的评价，主题 2 主要反映了客户对所发衣服**尺寸**的评价，主题 3 主要是客户对服装**品质**的评价，主题 4 主要展示了客户对衣服**颜色**方面的评价以及网上购物的感受。

由上图可知**主题 1** 的高频特征词主要是“love（喜爱）”、“sweater（毛衣）”、“great（极好的）”、“wear（穿戴）”等，主要反映了客户对于购买衣物的**直接感受**，大多为**喜爱和赞美**等积极的情感。针对于服装品类，可以看出**毛衣提及度较高**，背后主要有两个原因，一是毛衣购买量较大，这与前文中行业背景部分热销品类分析（2.1.2）结果相符，且多为好评，例如"the sweater is perfect all the way around（这件毛衣从各个角度来看都是完美的）"很明显相比于其他服装种类，毛衣踩雷较少。二是许多服装的搭配建议也围绕毛衣展开，例如“but the sleeves aren't lined so it's cold unless you have a nice sweater underneath it if you're going to be in cold weather（但是袖子没有衬里，所以会冷，除非你在里面穿一件漂亮的毛衣）”。故商家可以根据这些顾客的建议提前将问题告知，以增加顾客的满意度。在服装的**样式和设计**方面，虽然**大多数**顾客对他们购买的产品的风格和样式表示**满意**，但也有反馈表明实际体验与网上浏览时的感受**存在差异**。例如“while the intent is to be flowy, I was overwhelmed by the bagginess and it changes the style from flowy to bag-like.（虽然设计初衷是要轻盈飘逸，但我被其过度的宽松感淹没，这使得风格从轻盈飘逸变成了袋子般的样式）”。总的来说，尽管顾客对网购衣物的样式和风格整体满意度高，评论主要围绕这一点，但他们仍希望衣物的实际样式与预期保持一致。因此，电商平台需要准确展示和描述服装的风格，让顾客能根据自己的需求做出选择，避免夸大或误导，确保衣物的实际外观与介绍相符。

主题 2 的高频特征词主要是“size（尺寸）”、“small（短小）”、“fit（合适）”、“medium（中等）”等，突显了购物者在服装**尺寸**选择上的挑战，这是在线购物中最常见的问题之一，同时也是顾客最关注的**难题**。顾客经常反映，网上提供的尺码指南与实际收到的服装尺码存在差异，这造成了不少的不便。例如"I almost always wear a size 6, but the six was too big.（我几乎总是穿 6 号尺码，但这次 6 号对我来说太大了）"。表明尺码的不一致性给顾客造成了困扰。有些顾客之所以能买到合适的服装，是因为他们在下单前预先选择了大一号或小一号的尺码。因此，顾客强烈希望商家能够提供更准确、详细的尺码信息，并在有特殊尺码情况时给出明确的提示，以减少购物时的不确定性和担忧。

主题 3 的高频特征词主要是“fit（合适）”、“soft（柔软）”、“fabric（织物）”、“quality（质量）”等，更加注重服装的**品质**。在当前的服装市场中，消费者对服装的期待已经远远超越了基本的款式和尺寸需求，转而**更加注重于服装的整体品质**，尤其是材质和舒适度方面。这种趋势体现在顾客反馈中的一系列关键

词，如“soft（柔软）”、“comfortable（舒适）”等，强调了穿着体验的重要性。显然，大多数服装能够满足消费者对高舒适度的要求，且这与服装使用的材料质地密切相关。例如，"the fabric is super soft for a chambray style（对于牛仔布风格来说，这种面料超级柔软）"这不仅表明了牛仔布服装可以同样提供出色的舒适度，而且也反映了消费者对于面料软度的提高期望。这种期望推动了服装品牌在材质选择和加工工艺上的创新，追求在保持风格独特性的同时，也能给予穿着者更舒适的体验。因此，服装制造商和设计师需要密切关注市场动态和消费者反馈，将**舒适度和材料品质**作为设计和生产过程中的重要考量因素。通过使用高质量、触感良好的面料，并结合人体工学设计，制作出既时尚又舒适的服装，以满足日益挑剔的消费者需求。

主题 4 的高频特征词主要是“color（颜色）”、“size（尺寸）”、“ordered（订购）”、“online（线上）”等，主要为**衣物颜色等方面评价以及网络购物体验分析**。颜色准确性问题是由于显示屏的差异和拍摄光线的影响等，导致消费者在网上看到的衣物颜色与实物可能会有所不同。这种**色差**导致了消费者的反响各异，有些人对于收到的衣物颜色与预期相符感到满意，而有些人则因为颜色与网页显示的图片有出入而感到“disappointed（失望）”。故商家可以通过提供更多角度的图片、不同光线下的照片，甚至是视频展示来增加颜色呈现的准确度。此外，消费者还提到了**线上购物的便利性**，尤其是其他“reviewers（评论者）”留下的**意见**对于做出购买决策非常**有帮助**。

总之，通过增强产品展示的透明度和提供全面的购买信息，电商平台和服装品牌可以有效提升消费者的购物体验，降低因颜色和尺寸问题导致的退换货率，进而增强消费者的信任和满意度。

6.3 基于 LDA 挖掘算法的细粒度情感分析

6.3.1 基于 LDA 主题挖掘算法的指标与标签词选取

我们将上述 LDA 算法挖掘出的四个 topic 作为我们细粒度情感分析的四个指标，建立服装评价体系如下：

表 6.4 服装评价体系

总指标	指标	指标解释	标签词
服装评价	样式	涉及到衣服的风格、外观、款式等	love, sweater, great, wear, perfect, skirt, look, soft, top, dress, comfortable,nice, fit, well, flattering, little, jeans, really, pants, bought, beautiful
	尺寸	服装的尺寸、大小等	size, small, fit, medium, large, ordered, wear, waist, love, look, usually, little, big,

		bit, short, petite, xs,long,much, great, tried
品质	与服装材质、舒适度等有关，进而影响服装的性价比	soft,small,fabric, quality, love, price,well, wear, fits,beautiful,comfortable, great, sale, really, bought, cute, wool,true
颜色	服装的色彩、色差等	color,size, fit, love, store, ordered,online,red,wear, looks,tried,bought,purple,blue

6.3.2 细粒度情感分析的原理

首先基于上文所用 LDA 主题词提取的主题和关键词组成细粒度情感分析的词典，即细粒度情感的指标和标签词。通过 **python 的 SenticNet 库**分析单词的情感极性，每个关键词在 SenticNet 中都有一个相关的情感极性和一个数值用来表示情感的强度。本文通过由 LDA 组成的英文词典进行细粒度情感分析，从而得到针对于每句不同的评论所涉及到的各个主题的情感评分，相较于传统情感分析整句话的情感取向，该方法更加细致的去描述了样式、尺寸、品质、颜色四个指标下评论的情感取向，从而作出针对性的政策实施。

6.3.3 细粒度情感分析的结果与讨论

(1) 细粒度情感分析的总体情况

下图展示了对顾客购买服装经历的四个不同指标进行的情感分析饼图，从中我们可以明显看出顾客对于购买过程中的各个方面整体上表达了高度的满意。具体来说，四个主题的**正面评价**比率都超过了 65%，其中对于衣服的样式和风格的满意度更是高达 80%以上，这一数据强烈表明在设计和风格选择上，服装品牌已经成功地满足了大多数顾客的期待和喜好。

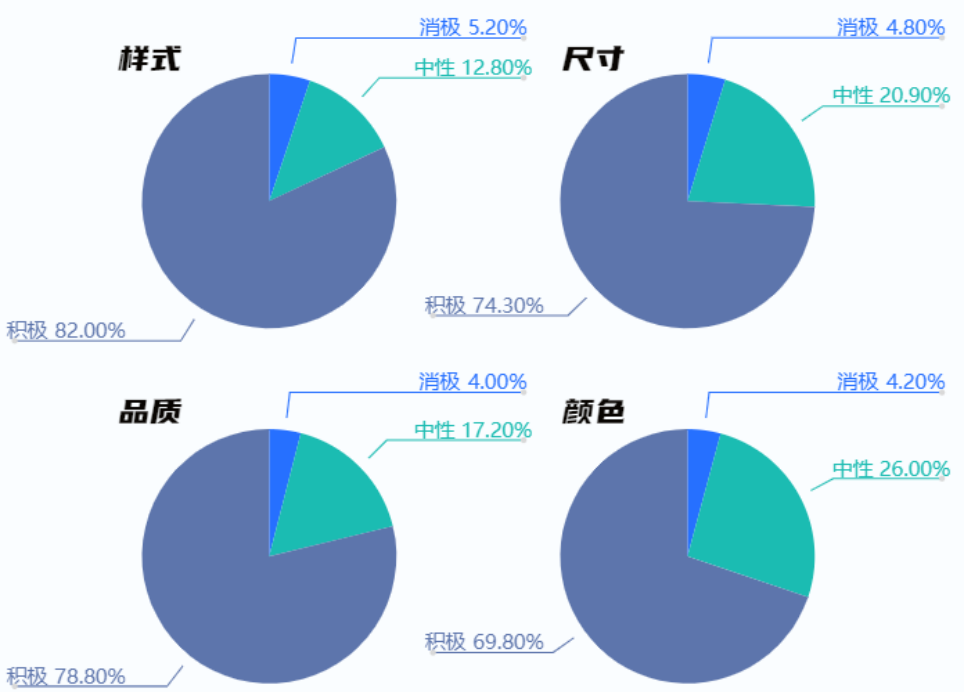

图 6.4 四个主题总体情感分析

(2) 对中性评价和消极评价的分析

除了正面评价之外，中性评价也占据了一定的比例，平均大约为 20%，高于消极评价（大约只有 4%到 5%）。前文传统情感分析说消极评价比中性评价更多，与这里的结论不同。这是因为传统情感分析和细粒度情感分析**分析精度和分类标准**不同：传统情感分析通常将情感分为正面、中性、和消极三大类，往往更倾向于简单直接地将情绪表达归为正面或消极，可能会忽略或减少中性评价的划分，因为它们不像正面或消极评价那样明确。细粒度情感分析则尝试**捕捉更微妙的情绪差异和强度**，可能将一些边缘性的消极评价划分为中性，尤其是当这些评价包含了既不完全消极也不完全正面的情绪表达时。该结果也展示我们选取的细粒度情感分析较传统情感分析的优势所在。

从内容上看，中性评价包含了顾客的一些小建议或是他们遇到的轻微不便，我们筛选出这一部分的评价，分析后原因如下：

- 针对样式方面，服装的设计虽然基本满足了穿着需求，但缺乏创新性或个性化元素，给人留下了平庸的印象。
- 针对尺寸方面服装尺寸与尺码表略有偏差，虽然能穿，但穿着感受并不理想，可能略微宽松或紧绷。
- 针对服装颜色与图片或描述略有差异，虽然总体接受，但未能完全满足顾客的期望。
- 针对服装质量一般，面料或制作工艺满足基本穿着需求，但在某些细节上存在小瑕疵，如线头、缝合不平等。

从整体上看问题还体现在物流和配送问题，客户服务和沟通障碍等。虽然这些评价并不直接影响整体的满意度，但它们提供了宝贵的反馈，商家可以通过这些建议进一步优化服务和产品。

而消极评价所透露出的问题程度更深，其对其他顾客做抉择时的影响也更大，因此商家需要着重重视消极反馈，通过**积极响应和采取改善措施来解决问题**，从而减少未来的消极评价并改善整体顾客体验。

(3) 各类女装的细粒度情感分析对比

其他种类的服装细粒度情感分析见附件，其中，**chemise 和 Causal bottoms** 这两类服装有效评论数量均小于三条，**样本量过小**，不具备单独分析意义，故**去除，不参与分析**。

综合上述分析，我们计算各服装品类各指标平均值绘制折线图如下图所示：

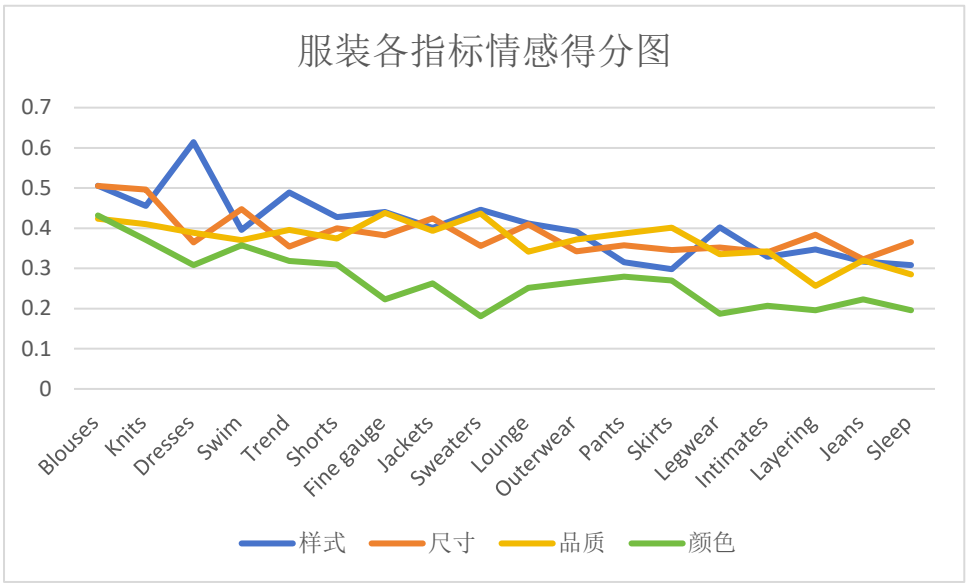

图 6.5 服装各指标的情感得分

从上图可见，色彩指标的评分在所有考量因素中始终处于较低水平。当我们排除了样本量不足以及语言模型潜在偏差等外在因素后，显然网购中的色差问题成为了消费者普遍关注的难题。

我们选取四个指标平均情感得分作为该服装种类的平均分：

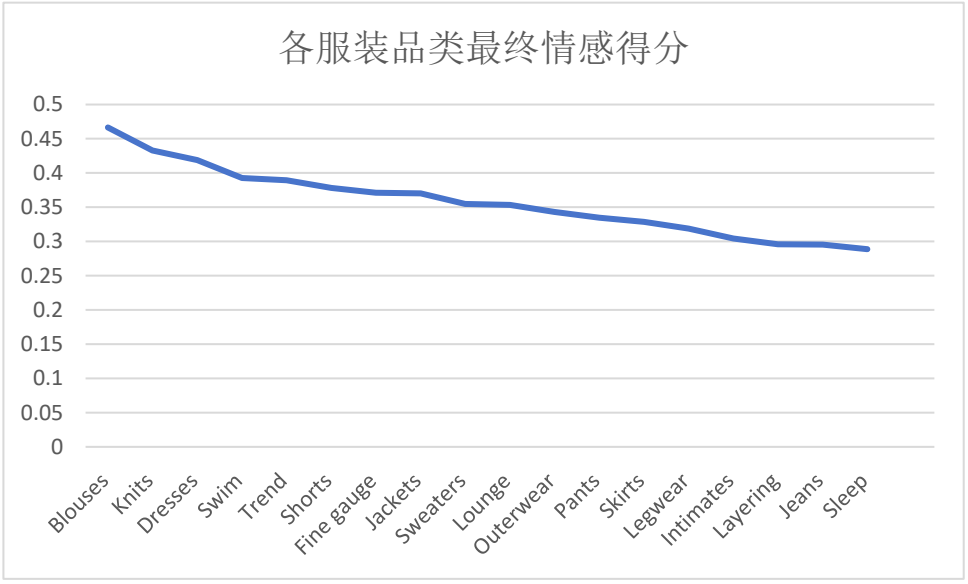

图 6.6 服装的最终得分

在对各服装品类的评分分析中，我们发现不同类别之间的评分差异较小，显示出消费者对各种类型服装的整体满意度较为均衡。

(4) Dress 类产品的细粒度情感分析结果与应用

基于上述结果，我们继续深入探究各个种类的服装评价，Dress 销量较高，我们以 Dress 为例展示细粒度分析结果，不同于传统的情感分析，我们得到了多个维度的评价，结果如图：

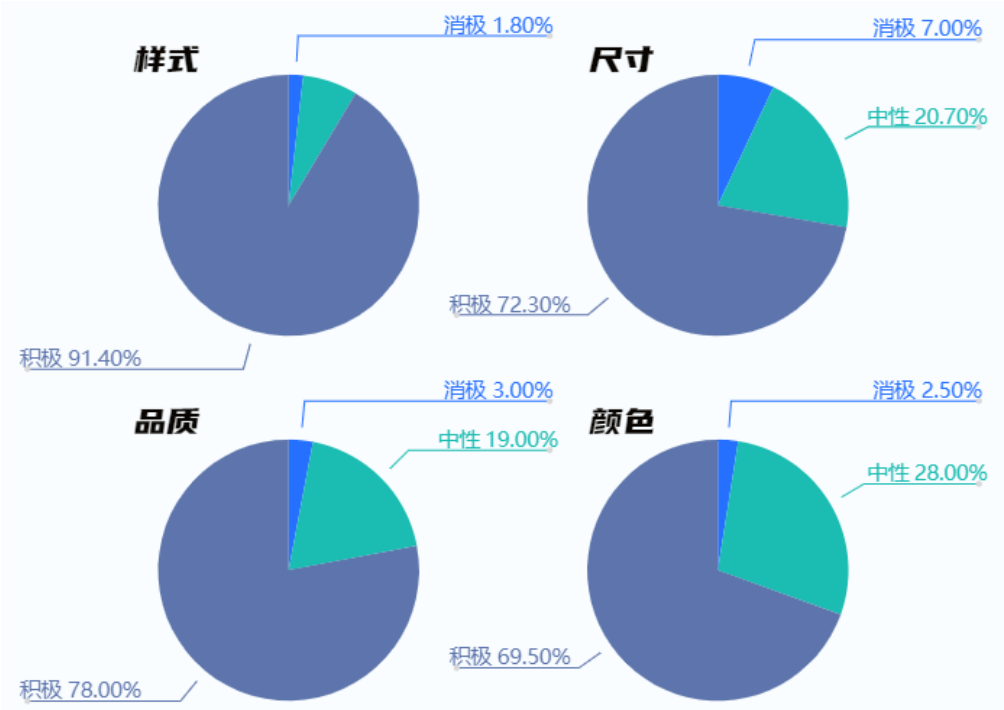

图 6.7 连衣裙细粒度情感分析结果

① 总体评价

从提供的信息中，我们可以看出，连衣裙在顾客评价中整体表现良好，特别是在款式、风格上，它们显著地赢得了顾客的青睐和好评。这表明设计团队在捕捉时尚趋势和顾客偏好方面做得非常出色，成功地创造了既时尚又能满足顾客期望的产品。

② 主要的负面因素——尺寸

然而，连衣裙尺寸大小方面的问题却成为了一个突出的负面因素，导致了较高比例的差评。尺寸不合适不仅会影响顾客的穿着体验，还可能导致退换货率的增加，从而增加运营成本并影响顾客满意度。

因此，对于这一问题，商家需要采取切实有效的措施进行改进。比如提供更详细的尺寸对照表。同时，因为连衣裙衣服本身的特性，相较于上衣、裤子等更会出现版型上的问题，故商家还可以改善产品设计和制版满足不同体型顾客的需求。

③ 主要的正面因素——款式和舒适度

同时，如果一个顾客在连衣裙的评论中表达了正面情感，比如对款式的喜爱或对舒适度的高评价，那么可以推荐具有类似特征的其他连衣裙或相关款式

的衣物。

(5) 总结

商家可以针对各种服装的最终评分进行改进。对于评分较高的服装种类，这不仅意味着它们在市场上受到了顾客的广泛欢迎，而且也标志着商家在这些方面的努力取得了成功。商家应该深入分析这些产品的成功因素，如设计特点、材料选择或者市场定位等，以便在未来的产品开发中继续采用和优化。如果顾客对某个特定特征的正面情感评价较多，那么推荐该特征下的新品或热销商品。

与此同时，对于那些评分较低的服装种类，商家需要认真分析背后的原因，是否是设计上的不足、质量问题，或是市场需求的误判。基于这些分析，商家可以采取具体的改进措施，比如优化设计、提高制造标准或调整市场策略，以提升产品的整体表现和顾客满意度。

总之，这种基于多维度评分的方法为商家提供了一个全面的视角来评估和改进他们的服装产品。通过持续地关注顾客的反馈和积极地调整产品策略，商家不仅能够提高现有产品的表现，还能够更好地把握市场趋势和顾客需求，从而在竞争激烈的服装市场中取得成功。

7 数字营销实践

7.1 基于细粒度情感分析的关键词营销

7.1.1 细粒度情感分析的重要性和应用

通过细粒度情感分析，我们可以深入了解女性消费者的需求，以关键词为引导，打造出真正打动人心的产品，让我们的营销更加精准，更加贴近消费者的内心需求。而用户评论是消费者真实反馈的直接体现，通过情感分析可以深入了解消费者的购买体验、产品满意度以及潜在需求。

通过收集和分析消费者的在线评论和反馈，企业可以更精确地确定目标市场，通过分析消费者对不同款式、颜色或价格的女装的情感反应，企业可以确定哪些产品最受特定人群的欢迎，从而制定更加精确的市场定位策略。并且深入了解消费者的需求和偏好不仅有助于产品改进，还可以为企业的营销策略提供指导。

7.1.2 基于细粒度情感分析的精准营销实施

针对本研究所分析的女装品牌，根据留言及评论，通过细粒度情感分析，得出用户对于该品牌的正面、中性和负面情感。

（1）正面情感涉及的关键词

大部分用户表示喜欢该品牌的设计，认为其款式多样、时尚，材质舒适。在本文 6.2 章节中对顾客评价进行分析得到“love（喜爱）”和“like（喜欢）”等正面词汇比较多，一般是对服装的颜色、尺寸、设计等方面赞扬较多，这些说明大多数顾客对该品牌是满意的。为了保持这些用户的忠诚度，可以定期推送产品的设计细节，请设计者详细聊聊设计理念，展示服装的特点，吸引顾客继续购买；也可以在品牌下设置不同系列，强化不同系列的设计风格，让顾客在购物中了解到自己最喜欢的系列，也愿意长期购买该品牌的服装。

（2）中性情感涉及的关键词

一些用户表示对该品牌的产品没有特别的喜好或反感，觉得衣服一些设计很喜欢，但整体感觉并不舒适，虽然如此但仍旧希望品牌能越做越好。在 6.2 章节中顾客大多提出尺寸方面“过大”或“过长”等问题，所以为了提升这些用户的满意度，可以增加服装的尺码，让顾客穿上更舒适的衣服；也可以提供更加完善的售后服务和购物体验，如快速配送、无忧退换货等便于顾客退换得到合适的衣服。此外，可以定期邀请用户参与品牌活动或调研，以便更好地了解他们的需求和期望。对于对该品牌的产品没有特别喜好或反感的用户，可以推荐该品牌的经典款式或热门商品，以吸引顾客的注意力。

（3）负面情感涉及的关键词

部分用户提到实际收到的产品与图片差距太大，不太符合预期，甚至感到“失望”。这些差距会出现在样式设计上，也会有颜色上的差异。在样式方面，品牌应该在营销时就详细描述衣服风格，展示实物照片，确保实物外观与所展示的相符。在服装色差方面，在营销时最好展示不同光线下的实物以便呈现更

准确的颜色。在商品营销时应当准确全面地展示商品信息，尽量保证顾客收到的产品满足预期，这样能有效提高消费者的购物体验，增进顾客忠实度。

总而言之，根据 6.3 章节细粒度情感分析的研究结果，针对样式、尺寸、颜色和质量这四个关键词，制定具体的**营销策略**：

a.对于样式方面：服装应当有所创新，仅仅满足穿着需求，更需要有一些特点吸引客户的注意，在推广营销时，可以更多地展现服装样式和设计，甚至可以请服装设计者介绍设计产品时的一些设计特色，提高客户对企业产品的期待和信心。

b.对于尺寸方面：为进一步提升客户满意度，建议企业不仅优化库存管理和引入在线预订服务，还应扩大尺码范围，提供更多细分的尺寸型号。这样做能更好地满足不同体型消费者的需求，减少因尺寸不合适导致的退换货问题，同时提升消费者购物体验和品牌形象。通过精确匹配消费者的尺寸需求，企业可以建立更加个性化和贴心的服务，从而吸引更广泛的客户群体，增强市场竞争力。

c.对于颜色方面：图片展示的服装可能存在视觉上的差距，这里可以通过为服饰打不同颜色的灯光对比展示色差，在客户选购的时候可以更全面地对比颜色，减少收货之后因为图文不符而造成的退货状况，更能有效提升客户的购物体验，增强购物满意度。

d.针对质量方面：虽然客户满意服装设计，但是在一些细节方面有瑕疵，比如拉链不合适、缝合不平等问题，对于企业应当更加注重生产细节，保证产品质量，避免一些细小问题而损失部分客户的信任，应当保证服装的性价比，让客户的购物感受更加舒适。

企业也可以在后续运营过程中，可以在平台上持续观察几个月来用户的购买量，满意度和评论信息，并通过调查问卷等方式了解用户的需求和意见，并根据本文研究的细粒度情感分析得到其他的客户评价指标，再次调整关键词营销策略，不断改讲产品设计和质量，并推出更符合用户期望的新品，以提高销售量。

7.1.3 女装行业热搜关键词

品牌还可以积极寻找最新的女装热搜关键词，根据关键词可以了解最新的市场趋势，从而对服装产品进行创新。我们根据 TheList.com 网站搜集到 2023 年末全球女装的最新热门关键词，如图 7.1 所示：

图 7.1 女装商品热销词云

这些关键词涵盖了多样化的时尚风格和元素，从复古到现代，从休闲到正式，企业可以根据这些关键词进行风格创新，吸引更多的顾客。例如在热门关键词中出现哥特风格和朋克，其通常以深色调和反叛精神为特点，强调个性和独特性。网眼衫和白色及膝靴则是更具有时尚感的选择，既可以单独穿着也可以作为层叠穿搭的一部分。耐克运动装代表了运动休闲风的流行，既适合运动也适合日常穿搭。早春连衣裙、多巴胺等元素，则展现了更多样化的时尚选择，满足不同场合和个人喜好的需求。这些关键词都是基于对时尚领域的综合观察和分析得到的，反映了全球女装时尚的主要趋势，有助于企业及时了解顾客需求和市场趋势。根据对这些热搜词的分析，品牌能够及时抓住热点和洞察流行趋势，并知道顾客的需求和关注点，并结合顾客评论为品牌提供新的思路。这些是近期比较热门的女装卖点关键词，企业在后期进行商品营销时，可以多多关注市场上的服装关键因素，根据其女装卖点热销词进行精准营销。

综上所述，通过运用细粒度情感分析和关键词营销策略，女装跨境电商能够精准触达目标用户，深入了解客户需求和痛点。这不仅优化了产品设计、定价策略和营销活动，还显著提高了客户满意度和忠诚度，从而增强了品牌形象和销售业绩。具体而言，这种策略使企业能够实现精准市场定位，提升用户体验，优化营销策略，提高品牌形象，增加销售和收入，降低营销成本，并促进数据驱动的决策制定。最终，这些效益将助力企业在竞争激烈的市场中实现可持续的业务增长和品牌价值提升。

7.2 基于消费者画像和运营情况的内容营销方案

在客户画像分析部分，利用数据分析软件，结合多维度的消费者数据，构建出精准的客户模型。关注的指标包括客户的年龄、购物评分、评论数据等，通过对这些指标的综合分析，制定了以下营销策略。

7.2.1 产品战略重点方案

（1）以非私人贴身衣物的生产为重心

a.非私人贴身衣物销量远高于私人贴身衣物，商家应把生产重心放在非私人贴身衣物。

b.非私人贴身衣物正常尺码和偏小尺码的生产量应控制在 5:3 左右。

c.重点生产产品为 Dresses、Knits、Blouses，其次为 Sweaters、Pants、Jeans 和 Fine gauge。

（2）以热销品类带动非热销品类的销售

a.推出上衣和下装的成套搭配设计款时，注意将热销品类和非热销品类进行搭配。

b.在页面上设置较为方便的跳转链接，同时提高非热销品 Sweaters、Pants、Jeans 和 Fine gauge 的销量。

（3）关注 intimates 类的市场推广

a.私人贴身衣物 Initmates 类的市场重心在于 40 岁以下的年轻人，在进行设计和推广时，应重点关注年轻人的喜好。

b.适当提高 Lounge、Swim、Sleep 三类产品的产量，注重不同的款式设计。

7.2.2 产品推荐方案

根据消费者年龄可以将市场消费者主要为 30 岁以下、31 至 40 岁、41 至 50 岁、51 至 60 岁、61 岁及以上五个年龄群体，消费主力集中于 31 至 40 岁，但各产品的受众与各年龄的消费习惯不同。

（1）针对各产品类设定不同的推荐群体

Dresses、Knits、Blouses 类产品受众群体较广，应扩大推荐范围。性价比较高、时尚热品主要推荐给 30 岁以下群体；气质、经典中高端产品主要推荐给 31 至 50 岁群体；舒适、轻便款主要推荐给 50 岁以上群体；Initmates 主要推荐给 40 岁以下群体；Jackets 主要推荐给 50 岁以上群体。

（2）面向各年龄段客户推荐不同的品类

对于 30 岁以下群体，可推荐适用于运动场景、工作场景、社交场景、旅游场景等多种类产品，推荐的产品应丰富多样，不拘于一种风格。对于 30 岁至 50 岁群体，主要推荐适用于工作场景和日常生活场景的服装，应注重产品品质的描述。对于 50 岁以上群体，主要推荐实用度高产品，注重产品材质、透气性、保暖性的描述。

（3）针对不同特征的消费者采用不同的营销策略

根据消费者分类分别进行个性化推送以实现精准推荐，参考本文的数据分析结果将消费者分为三类：“重点稳固”型，“重点培养”型，“精准培养”型。

对于“重点稳固”型消费者，应当多多推送时尚服饰产品，提供穿搭技巧和风格搭配的时尚资讯，丰富的服饰款式和风格更能吸引相关人群的注意力，满足不同消费者的个性化需求，企业可以通过“薄利多销”的策略提高产品销售量；

对于“重点培养”型消费者，此类消费者消费潜力较强，更倾向于购买中高端服饰产品，注重产品质量，在介绍产品时可以详细解释服装面料等细节，展现产品物有所值的一面，企业可以通过提高单品的利润从而得到较高的盈利；

而“精准培养”型消费者对于服装产品的选择比较固定，在推送时可以倾向于同类型产品，精准满足该类消费者的需求，企业对于同类服装的丰富更能吸

引“精准培养”型消费者的注意，更能将企业服饰视为常购买的品牌产品，有助于留住老客户。

7.2.3 评论管理方案

50 岁以上的消费者好评度高且评论内容丰富全面，将其设置为优质评论，以增加消费者的好感度。该群体评论积极性高，商家可以推出“好评抽奖”活动，进一步提高产品评分。31 至 40 岁的消费者群体对产品的评分低于市场平均评分，对于该群体，商家可以推出“晒单有礼”活动，鼓励消费者给出真实评价，了解消费者需求。同时推出“分享有礼”活动，鼓励消费者向家人朋友分享产品。简化评价流程，引导消费者评价行为。30 岁以下群体，可在评价流程中添加游戏元素，增加评价过程的趣味性。积极的评论对于未购买过公司产品的客户来说，可以更全面认识到产品特征，结合官方给出的详细信息，再决定是否购买，这些评论可能会让一些潜在客户也愿意购买公司产品。

7.2.4 产品推广方案

为增加产品在 50 岁以下群体的曝光度，可借助小红书、抖音双平台推广。在小红书建立优质内容池，在抖音主攻曝光，双平台相辅相成共同助力电商溢出和转化。其余平台辅助推广，如利用微博做明星向的曝光和种草，B 站做科普和测评型种草。对于 50 岁以上群体，采用传统大屏广告+合适的代言人的方法增加品牌好感度。邀请大众熟知的中老年演员或歌手代言，主要营销产品的舒适性、品质性，在电视卫视上推广。同时，定期在公园、超市、社区等老年人主要生活场景开展线下推广体验活动。推广活动可以打造公司品牌，更容易让产品被大众熟知，也是一种吸引客户购买的方式。

8 可视化大屏

为了便于掌握企业的总体运营情况，建立可视化大屏。图 8.1 中覆盖率运营情况分析，客户画像分析，客户情感分析，以及数字营销分析的相关内容。该数据大屏采用灵动型设计，根据石方 BI 数据分析平台，建立如下数据大屏，可以访问如下链接 http://bi.zjsfsz.com/dash/newScreen/share/2926?token=17087570518283afc21e2996a8fd5207b7496，分享密码：u1s1 进行观看。

图 8.1 跨境电商女装店铺销售数据可视化大屏

8.1 对客户画像分析进行的可视化呈现

对于客户年龄段分析，通过对市场数据进行描述性统计，得出所有用户得具体年龄，通过对年龄聚类为 18~30 岁，31~40 岁，41~50 岁，51~60 岁，61 岁以上，并以轮播排行条形图予以呈现，发现年龄段在 31~40 岁以及 41~50 岁的客户占比相对较高，因此企业在做相应的产品营销时应当着重关注 31~50 岁的客户人群。

对于客户类别分析，通过 K-means++聚类算法刻画消费者画像，并将消费者聚类为重点稳固客户，重点培养客户，精准培养客户。并以饼状图的形式进行呈现，通过观察相应类比客户的比例，从而确定相应的市场政策来稳固以及培养潜在客户以及已购客户。

8.2 对运营情况分析进行的可视化呈现

对于已售服装分类，通过对 Class_Name 以及 diversion_Name 分层统计，统计出各个类女装商品的销售量情况，采用矩形树状图的可视化呈现方式，可以清晰的看出各个大类以及小类的商品销售数量，从而转变市场销售模式，以

及指导企业未来选品方向，提高销售量。从大类角度来看第一类商品的销售量较高，从商品单体来看商品 Dresses 的销售量最高。未来进行采购商品时，应当着重关注这些商品，从而实现更高利润。

客户满意度分析采用栅格条形图的方式对客户满意度评分进行呈现，由于所给数据满意度为 1-5 分评分体系，首先统计出总得分数并且除以满分总数得到相应的满意度比例，通过计算主要几种商品的满意度比例呈现在栅格条形图中。从中看出客户满意度几乎位于 70%，说明该公司的货物质量等其他方面还要继续加强。

8.3 对客户情感分析进行的可视化呈现

对于 LDA 客户评论主题提取分析，通过对预处理后的客户评论进行 LDA 主题词提取，将得到的客户评论主题词按照类别以气泡图的形式进行呈现，从中可以看出各类别的关键词占比权重，通过对出现频率高的关键词进行针对性设计产品以及提高服务要求从而提高企业销售额。

对于客户细粒度情感分析，通过基于字典的细粒度情感分析进一步分析客户数据，将客户情感评分由整体情感评分，转化针对于样式，尺寸，品质，颜色各个类别的细粒度情感分析，通过构建英文情绪字典，通过细粒度情感分析将各个类别情感趋向变为消极、中性和积极并以四个饼状图呈现。企业通过各个类客户的情感倾向，能够了解目前产品设计优劣，从而更好导向未来变革。

对于各类服装客户情感评分，该分析进一步对上文细粒度情感分析按照衣服种类进行聚类统计，并以折现图的形式呈现，从而更加直观的描述各种类衣服的具体情感得分。

8.4 对数字营销分析进行的可视化呈现

对于累计营销额，通过后台调取商品售卖信息 url 进行对该企业累计销售额及时反馈，利用 python 网络编程获取相应的商品数据，利用自动翻牌器呈现，这样可以帮助企业及时获取企业销售额，从而及时获取相应市场情况。对于女装商品卖点热销分析，通过利用 python 进行爬虫，获取目前女装行业的热销卖点，并以动态词云图的形式呈现，并利用后端程序进行随时更新词云，该可视化分析可以帮助企业及时了解女装市场的热销产品，分析这些热销卖点，从而精确调整该企业的市场供货渠道以及运营政策。

9 总结

随着社会发展，女性就业比例的提高和社会地位不断提升，女性有经济上的独立性，其消费需求也随之不断提高，这推动女装行业发展。与此同时，由于市场的巨大潜力，女装品类也面临着激烈的竞争。为了帮助商家找到产品的发展方向，本文通过资料调查、描述性统计分析、情感分析、Logistic 回归模型分析、LDA 主题挖掘等方法对消费者的特征、感受、偏好、需求等进行研究，最终将消费者分为三类，评论主题分为四类并做了相应的营销方案分析，具体结论与商业建议如下所示：

A、结论

(1) 女装市场前景良好

整体市场调研可以帮助商家准确定位并作出合理决策。经济上，疫情后，经济复苏，女装市场逐渐回暖。政策《纺织业“十四五”发展纲要》支持女装行业发展。社会文化多元化推动市场多样化，技术进步如互联网、3D 打印和智能制造为女装创新提供新机会。

(2) 31 至 40 岁群体为消费主力

消费者分析显示 31 至 40 岁群体购买量最高，为市场消费主力，但满意度较低。为稳固此群体，商家需主动沟通了解及解决问题以提升满意度。重点销售连衣裙和针织品，多样化款式和风格，满足个性化需求，防止产品同质化。

(3) 差异化管理年龄评论

针对 51 岁以上群体满意度高于市场平均，商家应利用其正面评论提升产品印象。这一年龄段倾向于给出较高评分，推荐这些积极评价可以吸引更多消费者。同时，应关注年轻消费者的具体反馈，通过识别并解决问题，改善产品满足更广泛消费者的期望。

(4) 积极情感促进服装推荐

通过二元 Logistic 回归模型分析发现，顾客的高满意度和对商品的积极情感显著增加了推荐意愿。因此，商家应通过提升产品质量、和加快物流速度等措施，来增强顾客的满意度和正面体验，从而提高商品的推荐率。

(5) 消费者对产品的关注点较多

LDA 模型的分析显示，消费者在评价中关注多个方面，包括产品的样式、尺寸、品质、网购体验和颜色，揭示出消费者评价内容的丰富性和多样性。

B、商业建议

（1）针对“重点稳固型”消费者——提供多样化产品，采取优惠策略

针对 24 至 35 岁年轻人构成的“重点稳定型”消费者群体，商家需提供广泛的产品种类和风格，满足其对新款式的高频需求，并提高性价比，采取优惠策略以增加销量。鉴于该群体评价参与度低，应通过奖励性活动如“晒单有礼”和优化评价流程鼓励更多反馈，丰富评价内容。确保销售的品类多样、款式和风格丰富，满足其对个性化和多样化的需求，同时避免产品同质化。

（2）针对“重点培养型”消费者——推荐高端款式，提价增利，通过沟

通解决低评分问题

针对 45 至 54 岁的“重点培养型”中年消费者，这一群体对服装品质的关注度较高且市场评分最低。商家应推荐高价位、高品质的款式，将其定位为中高端产品的主要目标客户，并适当提高定价以增加单品利润。鉴于该群体倾向于给出较低评分，商家需要主动沟通了解原因，提供适当补偿，以此提升评分。销售品类上，应重点推荐连衣裙和针织品等类目，同时确保产品面料优质、款式经典，以满足其对高品质的追求。

（3）针对“精准培养型”消费者——提供多价位选择，聚焦品质与舒适

针对 60 岁以上的“精准培养型”老年消费者，这一群体在女装市场中表现出高度一致的消费需求和最高的市场评分。针对其谨慎的产品选择习惯，商家应提供不同价位的相似产品，使消费者能够比较选择。鉴于该群体的高评分特点，企业应突出展示这一群体的正面评论，以提升产品在其他消费者心目中的印象。销售品类应聚焦于一两个特定类目，通过精细化管理和优化产品质量来建立良好的品牌口碑和提高消费者忠诚度。同时，产品设计需重视实用性和休闲性，以符合老年消费者对舒适度的高要求。

（4）基于评论主题词“样式”——追踪时尚趋势，通过社交媒体提升款式竞争力和品牌曝光

消费者反映服装的设计较为平庸，消费者期待创新性或个性化元素。商家需敏锐捕捉时尚趋势，深度挖掘顾客偏好方面，提高产品款式、风格上竞争力。商家应利用社交媒体等平台分享丰富的时尚资讯、潮流趋势，展现穿搭效果和款式特点，引起用户的兴趣，提升品牌的曝光度，掌握时尚潮流的主动权。

（5）基于评论主题词“尺寸”——提升产品透明度和信息全面性，优化库存及尺码多样性，加强客服以改善购物体验

尺码偏差影响了消费者的购物感受。商家需增加产品展示的透明度、提供全面的购买信息以减少因下单不准造成的差评。同时，应优化库存管理和引入在线预订服务，扩大尺码范围，提供更多细分的尺寸型号。此外，需提高在线客服服务水平，及时地回应和解决消费者的问题。

（6）基于评论主题词“颜色”——增加多角度和不同光线下的产品照片及视频，使用多肤色模特，以提高颜色呈现准确度

产品“色差”影响了消费者的购物感受。商家需要提供更多角度不同光线下的照片，以及视频展示来增加颜色呈现的准确度。同时商家应聘请不同肤色的模特展示服装上身效果，减少消费者因服装颜色与肤色不合适产生的差评。

（7）基于评论主题词“质量”——提高生产细节关注度以保证产品质量，展示产品细节，激励消费者分享实物图片和使用体验

产品瑕疵影响了消费者的购物感受。商家应注重生产细节和质量，避免因小问题而损失客户信任。同时商家应向消费者详细展示产品细节，通过“晒图有礼”、“追评有礼”等活动促进消费者在评论时拍摄产品图片，描述产品使用感受等，以便其他消费者全面了解产品并做出购物决策。

附件

附件一 数据预处理代码

数据预处理

语言:python

```python
import pandas as pd

# 加载 Excel 文件
df = pd.read_excel('C:/Users/29760/Desktop/原始数据(1).xlsx')

# 使用一个函数来检查每个文本中单词的数量
def count_words(text):
    return len(str(text).split())

# 筛选出'Review_Text'列中单词数量大于等于 3 的行
filtered_df = df[df['Review_Text'].apply(count_words) >= 2]

# 将筛选后的数据保存到新的 Excel 文件
filtered_df.to_excel('筛选后的 Excel 文件.xlsx', index=False)
```

```python
import pandas as pd

# 加载 Excel 文件
df = pd.read_excel('筛选后的 Excel 文件.xlsx')  # 请替换为您的文件路径

# 检查缺失值
missing_values = df.isnull().sum()

# 打印每个字段的缺失值数量
print("每个字段的缺失值数量：")
print(missing_values)
```

附件二 LDA 主题挖掘代码

LDA 主题词选取

语言:python

```python
import pandas as pd
import re
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation

# 读取 Excel 文件
file_path = 'LDA.xlsx'  # 请根据你的文件路径进行修改
data = pd.read_excel(file_path)

# 简化的文本预处理函数
```

```
def preprocess_text_simplified(text):
    text = text.lower()  # 转换为小写
    text = re.sub(r'[^\w\s]', '', text)  # 移除标点符号
    words = text.split()  # 分词
    return ' '.join(words)

# 应用文本预处理
data_subset = data['Review_Text'].iloc[:]  # 选取评论 s
preprocessed_texts = data_subset.apply(preprocess_text_simplified)

# 构建文档-词矩阵
vectorizer = CountVectorizer(max_df=0.95, min_df=2, stop_words='english')
dtm = vectorizer.fit_transform(preprocessed_texts)

# 定义和训练 LDA 模型
n_topics =5  # 主题数量
lda_model = LatentDirichletAllocation(n_components=n_topics, random_state=0)
lda_model.fit(dtm)

# 显示每个主题的代表词汇
def display_topics(model, feature_names, no_top_words):
    for topic_idx, topic in enumerate(model.components_):
        print(f"Topic {topic_idx}:")
        print(" ".join([feature_names[i] for i in topic.argsort()[:-no_top_words - 1:-1]]))

display_topics(lda_model, vectorizer.get_feature_names_out(), 10)
```

附件三 细粒度情感分析代码

细粒度情感分析标签词聚类
语言:python

```
import pandas as pd
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib
import warnings
import statsmodels
import seaborn as sns
import matplotlib.pylab as plt
from scipy import   stats

warnings.filterwarnings('ignore')
```

```
matplotlib.rcParams['font.family'] = 'SimHei'
plt.rcParams['axes.unicode_minus'] = False
# Load the data
'''
Blouses                    Pants
Casual bottoms             Shorts
Chemises                   Skirts
Dresses                    Sleep
Fine gauge                 Sweaters
Intimates                  Swim
Jackets                    Trend
Jeans
Knits
Layering
Legwear
Lounge
Outerwear
'''
df = pd.read_excel('total.xlsx')
# Check the first few rows of the dataframe to understand its structure
df.head()
# Convert scores to categories
categories = {'积极': lambda x: x > 0, '中性': lambda x: x == 0, '消极': lambda x: x < 0}

# Apply the transformation for each topic
for topic in ['topic1', 'topic2', 'topic3', 'topic4']:
    df[topic] = df[topic].apply(lambda x: next((key for key, func in categories.items() if
func(x)), None))

# Check the transformation
df.head()
import matplotlib.pyplot as plt

# Create a figure for the pie charts
fig, axes = plt.subplots(2, 2, figsize=(12, 12))
axes = axes.flatten()  # Flatten the axes array for easy iteration
topics = ['topic1', 'topic2', 'topic3', 'topic4']

for ax, topic in zip(axes, topics):
    # Count the frequency of each category in the current topic
    counts = df[topic].value_counts()
    # Plot pie chart
    ax.pie(counts, labels=counts.index, autopct='%1.1f%%', startangle=140)
    ax.set_title(f'Topic {topic[-1]} Distribution')
```

```
plt.tight_layout()
plt.show()
```

细粒度情感分析

语言: python

```
import pandas as pd
from senticnet.senticnet import SenticNet
sn = SenticNet()
'''
Blouses                    Pants
Casual bottoms             Shorts
Chemises                   Skirts
Dresses                    Sleep
Fine gauge                 Sweaters
Intimates                  Swim
Jackets                    Trend
Jeans
Knits
Layering
Legwear
Lounge
Outerwear
'''
data = pd.read_excel('副本评论分类(1).xlsx',sheet_name="total")
features = {
    'topic1': ['love', 'sweater', 'great', 'wear', 'I', 'color', 'perfect', 'This', 'skirt', 'look', 'soft', 'top', 'dress', 'comfortable', 'nice', 'will', 'fit', 'well', 'fabric', 'flattering', 'little', 'jeans', 'really', 'pants', 'one', 'length', 'bought', 'long', 'beautiful', 'colors'],
    'topic2': ['size', 'small', 'im', 'fit', 'I', 'top', 'medium', 'large', 'ordered', 'wear', 'waist', 'love', 'look', 'really', 'usually', 'little', 'big', 'bit', 'short', 'petite', 'xs', 'runs', 'way', 'This', 'long', 'think', 'fabric', 'much', 'great', 'tried'],
    'topic3': ['size', 'fit', 'soft', 'small', 'I', 'fabric', 'im', 'quality', 'love', 'price', 'This', 'well', 'ordered', 'got', 'wear', 'shirt', 'tee', 'fits', 'one', 'beautiful', 'These', 'comfortable', 'great', 'sale', 'really', 'bought', 'cute', 'wool', 'sweater', 'true'],
    'topic4': ['I', 'color', 'size', 'fit', 'top', 'one', 'love', 'store', 'ordered', 'online', 'back', 'wear', 'looks', 'will', 'im', 'tried', 'bought', 'much', 'fabric', 'colors', 'didnt', 'really', 'little', 'see', 'person', 'dont', 'even', 'sleeves', 'pretty', 'soft']
}
# 初始化一个空的 DataFrame 来存储结果
results = pd.DataFrame(columns=['Comment'] + list(features.keys()))
# 遍历每条评论
for index, row in data.iterrows():
    comment = row['text']
    sentiment_scores = {'Comment': comment}
```

```
        # 计算每个主题的情感得分
        for feature, keywords in features.items():
            feature_sentiment_score = 0
            keyword_count = 0
            for word in comment.split():
                if word.lower() in keywords:
                    try:
                        polarity_value = float(sn.polarity_value(word))
                        feature_sentiment_score += polarity_value
                        keyword_count += 1
                    except:
                        pass
            if keyword_count > 0:
                feature_sentiment_score /= keyword_count
            sentiment_scores[feature] = feature_sentiment_score
        # 将当前评论的得分添加到结果 DataFrame 中
        results = results.append(sentiment_scores, ignore_index=True)
    # 导出结果到 Excel 文件
    results.to_excel('total.xlsx', index=False)
```

附件五 各品类服装细粒度分析结果（见“其他资源”）