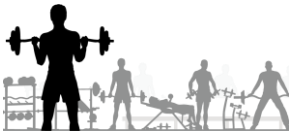
云健身战“疫”

在健身人群渗透率不断提升的中国，居民的版权意识也在提升，对线上优秀内容的付费意愿增强，对智能健身设备和配套运动产品的购买欲望增加为营造健身生态打下了坚实的基础。

疫情后在线健身平台的使用及偏好情况

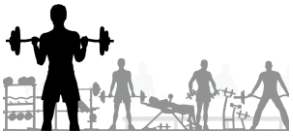


摘要

2021 年，国务院印发《全民健身计划（2021—2025 年）》，为了使全民健身更高水平发展、更好满足人民群众的健身和健康需求，提出了 5 年目标和 8 个方面的主要任务。早在 2014 年，“全民健身”就已上升为国家战略之一，随着中国特色社会主义进入新时代，我国社会主要矛盾转变，人们对精神文化生活的追求，在健康方面表现为对自己的心理和身体健康的关注度逐渐提升。政府对于加大全民健身场地设施供给、广泛开展全民健身赛事活动、提升科学健身指导服务水平、激发体育社会组织活力、促进重点人群健身活动开展、推动体育产业高质量发展、推进全民健身融合发展、营造全民健身社会氛围等提出要求，从加强组织领导、壮大全民健身人才队伍、加强全民健身安全保障、提供全民健身智慧化服务等方面提出了保障措施。为了进一步探究全民健身智慧化服务的普及情况，和新冠疫情对健身行业的影响，本文以南京市为例，对有健身需求的群体和部分健身行业从业者展开调查。

首先，我们通过对当前健身市场的观察与思考设计了科学的调查方案：通过网络和文献搜集并归纳了中国健身行业当前的发展现状，特别是线上健身行业的运营情况，为后期问卷设计调查奠定基础；同时对粤西地区（湛江市和茂名市）的居民及健身房工作人员进行访谈，将线下健身与线上健身的情况进行对比分析得出影响线上健身需求的相关信息与因素。在配合国家疫情防控的基础上，对工作人员采取线下访谈方式；此外，通过 python 软件对 bing 浏览器中“线上健身”“线下健身”相关网页进行爬取，绘制了词云图，得出当前大众的关注热点。在抽样设计方面，先将南京市各行政区通过 GDP 总值指标分为经济较发达地区与经济欠发达地区，并对各层进行三阶段 PPS 抽样。在正式调查中，严格控制问卷质量，共发放网络问卷 500 份，回收有效问卷 418 份，有效回收率为 83.60%，并对回收的有效问卷进行信效度检验，发现数据质量符合标准。

其次，本团队通过整理收集的问卷数据，从在线健身产品市场现状、文本挖掘、客户特征以及影响因素四大方面展开分析。市场现状主要通过描述统计分析方法，用直方图、饼图、聚类图等图形与表格进行可视化展现；文本挖掘则按照

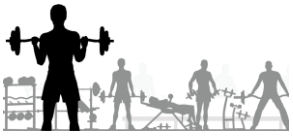


网页数据收集、文本分词、频次排序、绘制词云图四步依次进行；基于用户特征，我们从在线健身平台使用现状、营销市场现状和潜在用户挖掘三方面入手，剖析了在线健身平台的发展前景。针对 418 份在线健身样本采用 K-means 聚类算法，基于月收入和消费能力将样本分为目标用户、普通用户、高消费用户、节俭用户、谨慎用户五大群体。构建了结构方程模型，对潜变量的影响关系进行研究，保证了测量关系的高质量；分别构建 Logistic 回归模型、决策树模型对年龄、学历、月收入等因素对在线健身使用付费功能顾客特征、付费意愿进一步分析；基于朴素贝叶斯的云健身评价进行情感分析，以期从健身用户中获取更多的有效信息。

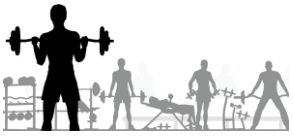
通过文献和数学模型分析，本文分别从线上健身的了解程度和接受程度、消费者特征、消费者决策影响因素和营销策略四个方面得到如下主要结论：年龄小、受教育水平高、月消费月收入高的群体对在线健身的接受程度更高，推广渠道与受众错位导致在线健身了解程度在不同类型人群中偏差较大；调查群体普遍愿意接受“优质”的线上健身服务。

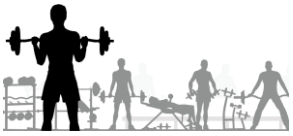
最后，针对所得结论，我们提出以下建议：第一，线上健身宣传及推广方面的建议，要加大宣传辐射面，激活潜在消费群体，宣传推广线上健身形式时注意选取突出形式新颖、内容创新、思想先进等创新角度，也要凸显线上健身方式的核心竞争力，并对线上健身接受群体进行精准营销；第二，线上健身发展趋势和产业结构方面的建议，要抓住疫情导致的线下健身用户转向线上健身的契机，推进线上健身或“线上+线下”健身，健全多元信息服务，提升用户“认知+思考”体验，发展智慧健身社交，提升用户“关联+情感”体验。

关键词：后疫情时代；在线健身；全民健身；消费偏好；决策树；Logistic 回归分析；K-means 聚类算法；结构方程模型

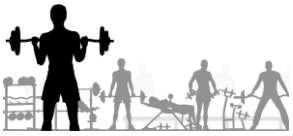


目录





表目录



图目录

一、引言

（一）研究背景

十九大报告中明确指出，中国特色社会主义进入新时代，我国社会主要矛盾已经转化为人民日益增长的美好生活需要和不平衡不充分的发展之间的矛盾。在党的百年华诞上，我国实现了全面建成小康社会的百年目标。周文彰和岳凤兰（2018）从物质生活、精神生活、生活环境、主体自身四个方面，论述了“美好生活”的定义，人们不再满足于单纯的物质生活改善，开始追求更高层次的精神文化生活。表现在健康方面，人们对自己的心理和身体健康的关注度逐渐提升。

2014 年，“全民健身”就已上升为国家战略之一，习近平总书记作为全民健身的倡导者和践行者，早在 2014 年南京青奥会就指出“要提高全民族身体素质与健康”；2016 年，习总书记提出要“落实全民健身国家战略，普及全民健身运动，促进健康中国建设”；同年 8 月 26 日，中共中央政治局审议通过《“健康中国 2030”规划纲要》，将全民健身纳入其中。2021 年，国务院印发《全民健身计划（2021—2025 年）》，强调加大设施供给，提升健身辅导水平、体育组织活力，实现体育产业现代化，政府提出 “互联网+健身”和“物联网+健身”的概念，营造全民健身的社会氛围，并提供全民健身智慧化服务。

2015 年 2 月 4 日，北京卡路里科技有限公司上线了 Keep 线上健身 APP，正式拉开了线上健身的序幕。2022 年 2 月 25 日，Keep Inc.于香港证券交易所提交招股书，申请 IPO 上市。招股书中披露了 Keep 的收入来源，包括会员订阅及线上付费内容、自有品牌产品，以及广告和其他服务，其中自有品牌产品包括在自营商城和第三方电商平台上销售的智能健身设备、健身装备、服饰和食品等商品，该部分业务贡献了 Keep 50%以上的收入。据悉，2019 年 Keep 曾面临裁员关店危机。紧接着 2020 年新冠肺炎疫情爆发，疫情在为服务类行业带来冲击的同时，也为线上健身平台提供了无限机遇：一是疫情的到来使得人们更关注自身健康问题；二是线下健身房被迫停业后，线上健身成为了用户唯一的可选择的渠道。与 Keep 同样享受到疫情带来红利的，还有美国健身平台 Peloton，其市值在 2020 年

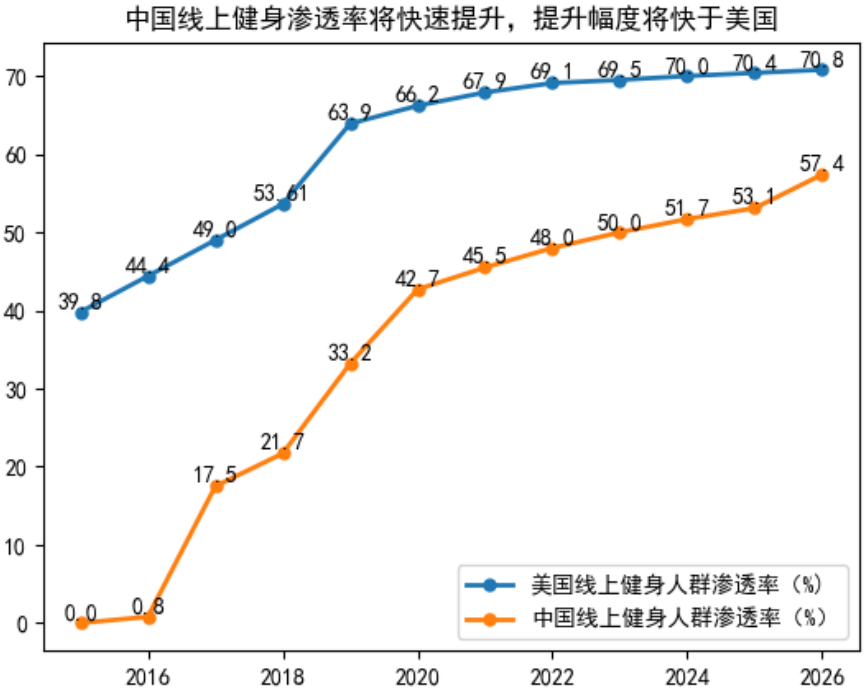
疫情爆发期间一度接近 500 亿美元。上市公司的市值能够较为直观地反应该公司乃至一个行业的发展前景，从科技公司华为、小米等加入线上健身行业的行为中更能看出，线上健身平台前景可观。

从健身行业的上中下游企业来看，处于上游的健身器材行业市场规模稳定增长，前瞻产业研究院预测数据显示，2025 年全球健身器材市场规模将达到 148 亿美元，预计 2019 年-2025 年的年复合增速为 4.01%。健身器材行业销售收入受消费者健身观念转变、政策激励、技术进步等因素，呈波动增长状态。我国虽是健身器材制造大国，但制造的健身器材主要销往欧美等国，国内健身器材的普及率却较低，渗透率增长潜力大，家用健身器材需求将会继续提升；随着我国人口老龄化加深，适合老年人使用的健身器材有望快速发展。随着互联网普及和科技水平提升，健身器材智能化趋势确定。电商行业的发展促进健身器材渠道从线下转向新零售模式。处于中游的线下健身和线上健身行业中，健身会员数量和健身房数量已经超过美国，但是健身渗透率和行业集中度仍较低。随着运动健康观念的兴起，居于下游的运动营养食品市场规模也在持续增长，但我国市场规模与发达国家间差距仍较大，虽然市场集中度较大，但渗透率较低。说明中国很可能已经存在较为成熟的健身营养供给市场，但坚持运动营养的需求市场占比不大。

在新冠疫情为背景下，长期隔离、禁足使得居民内心产生烦闷情绪，工作学习产生的紧张情绪不能得到合理排解；缺乏运动易使我们摄入的热量大于消耗的热量，同时肌肉量会逐渐减少，对形成健康的体魄不利，所以在线健身平台的普及是居民身体健康的心理和生理需求使然。对于健身自律的群体，从某种程度上来说，线上健身平台比线下健身更加方便快捷且经济。线下健身虽然有齐全的运动设施和专业教练，但健身房普遍实行会员制，会费与线上健身 APP 在线课程、会员相比是一笔不小的开支。一旦物价稍有上涨，居民生活成本提升，线下健身会费很可能成为一项负担。线上健身平台的课程大多采用录播的一对多形式，只要有足够多的受众，平均成本与线下健身相比可以忽略不计。这就是线上健身 APP 提供大量免费课程，即使是付费课程也收费较低的原因之一；当前线上健身 APP 种类繁多，竞争激烈，为了吸引和巩固用户，占据市场份额，越来越多的平台甚至已经不靠在线课程盈利，转而引导消费者购买个人定制服务、健身餐和健

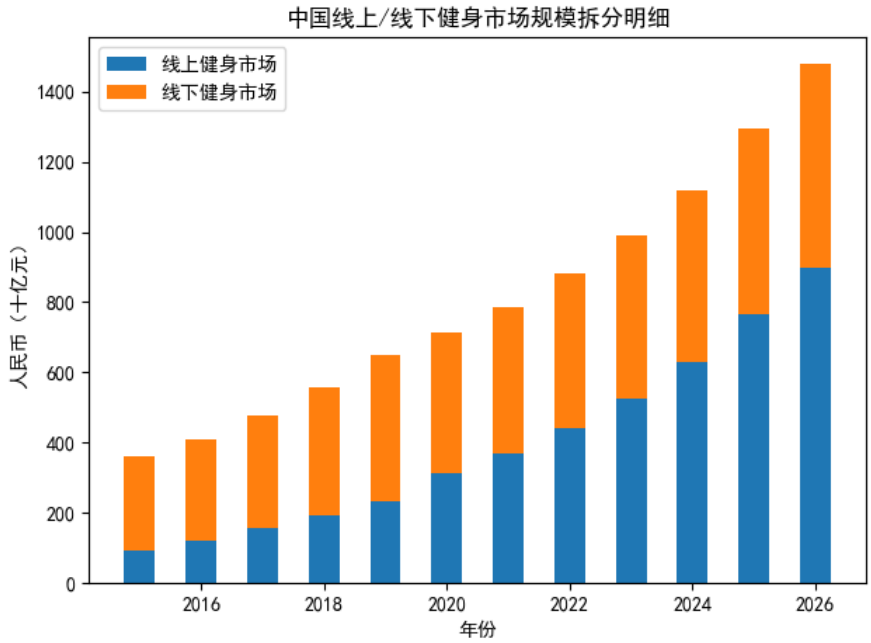
身器材等的健身经济。所以对于只有普通健身需求的受众，免费项目已能满足大部分普通工薪阶层和学生的健身需求。据推测，健身的主流群体是 20-40 岁的青壮年，他们既有一定的经济实力，又有身材管理需求，但一部分人因难以平衡家庭工作和运动健身的关系而放弃运动健身，导致中年“发福”；当线上健身平台普及之后，这部分受众不仅可以利用碎片化时间可以线上运动，还能与家人同事一起享受健身福利。即使在后疫情时代，线下健身的市场份额也难以恢复到疫情前的水平，线下健身路程较远，且不同来源地的顾客聚集在一个相对封闭的健身房存在病毒传播风险；而线上健身除了不能享受到专业的健身器材，线上健身平台几乎全部具备线下健身的功能；另外，线上健身平台的网络属性还赋予其健身分享的云社区功能。

根据灼识咨询的行业分析报告，我国线上健身行业正处于快速发展阶段，还有很大发展空间。2015 年中国线上健身行业刚刚起步，仅用两年时间，线上健身人群渗透率就实现了从零到 17.5%的历史性突破，从图 1 中可以看出，2015-2021 年中国线上健身人群渗透率共上升 45.5%，对比同期美国线上健身人群渗透率，从 2015 年的 39.8%上涨至 2021 年的 67.9%，可见中国线上健身人群渗透率增长率明显高于美国。因为中美线上健身业的发展起点不同，美国线上健身渗透率预计在 70%左右时达到饱和，而中国的线上健身渗透率才至 47%，发展潜力十足。2015-2021 年中国健身市场规模呈上升态势，复合增长率为 14.3%，预计 2021-2026 年中国健身市场规模复合增长率为 13.5%，增长率保持的主要原因是线上健身市场规模的稳步增加以及从线下健身市场用户的转化。图 2 显示中国线上健身市场规模占中国健身市场总规模比例从 2015 年的 26.9%到 2021 年的 47.0%，足足增加了 20.1%；从量上看，中国健身市场总规模从 2015 年的 935.45 亿元增加到 2021 年的 3697.02 亿元。据估计，我国线上健身市场规模到 2026 年将上升至 8964.558 亿元，占比增至 60.6%。



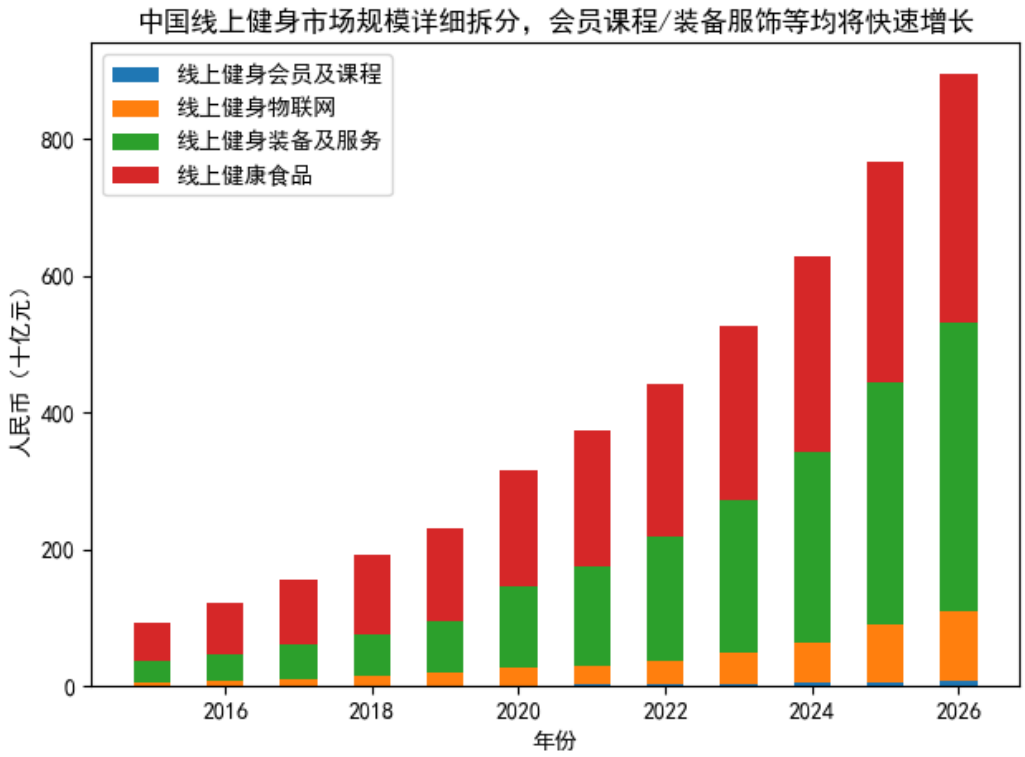
（数据来源：灼识咨询）

图 1　中美线上健身人群渗透率



（数据来源：灼识咨询）

图 2　中国线上线下健身市场规模比较



（数据来源：灼识咨询）

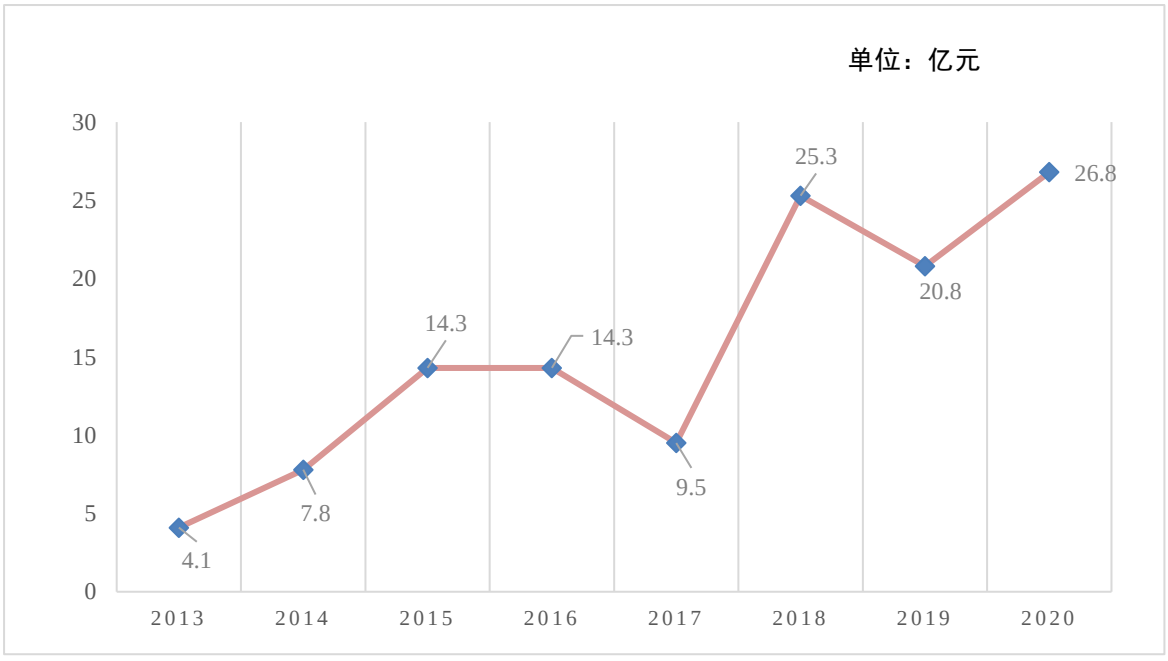
图 3　中国线上健身市场规模详细拆分

在健身人群渗透率不断提升的中国，居民的版权意识也在提升，对线上优秀内容的付费意愿增强，对智能健身设备和配套运动产品的购买欲望增加，为营造健身生态打下了坚实的基础。前面说到，当前的线上健身 APP 的盈利模式已经从健身课程导向型转化为健身生态导向型，健身课程大多免费开放，或是对 VIP 免费开放，平台甚至还会与健身教练签约，用优质的服务吸引更多市场份额。因其健身生态导向型的转变，在线健身平台亦可归为电商门类。“5G”互联网的普及为“互联网+健身”打通了网络传播渠道，视频播放更加流畅，受众体验感更佳。应该明确，我们的重点应该在于健身习惯的养成，而不仅是单次的健身行为，否则多贵的私教班和健身课程都不能让我们坚持运动、终身受益。

线上健身的火热除了疫情的加持，从众心理也不容小觑，通过对“线上健身”“fitness”等词条的文本分析，我们发现“姐妹”一词的词频较高，据此推测女性在健身时存在多人陪同现象。对于自律程度不足的人来说，坚持健身确实不易，

但多人组成团体互相督促，能够促使目标的达成。随着自媒体的发展，“小红书”等 APP 上出现了一大批分享瘦身经历的博主，营造身材焦虑，反而推动了线上健身业的发展。

不可否认的是，在线健身平台的普及也可能导致线下健身需求的增加，但通过对粤西地区（湛江市、茂名市）的实地调查，我们推测线上用户转化为线下用户的比例远远小于线下用户转化为线上用户，所以对线上健身平台进行用户粘性分析比分析线下健身的意义更加深远。根据我们对在线健身平台的定义，线上健身平台同时具备文化、体育和娱乐属性，从盈利模式分类看可以归为电商类，所以我们使用“文化、体育和娱乐业电子商务采购额”来衡量线上健身市场的发展情况。图 4 展示了 2013-2020 年文化、体育和娱乐业电子商务采购额的年度变化情况，可以看出线上健身平台目前处于快速发展阶段。



（数据来源：国家统计局）

图 4　文化、体育和娱乐业电子商务采购额

（二）研究目的和创新点

我们发现虽然我国线上健身 APP 的诞生已有八年，但因互联网普及、线下

健身的习惯思想的限制，近五年才迎来线上健身平台的爆红。

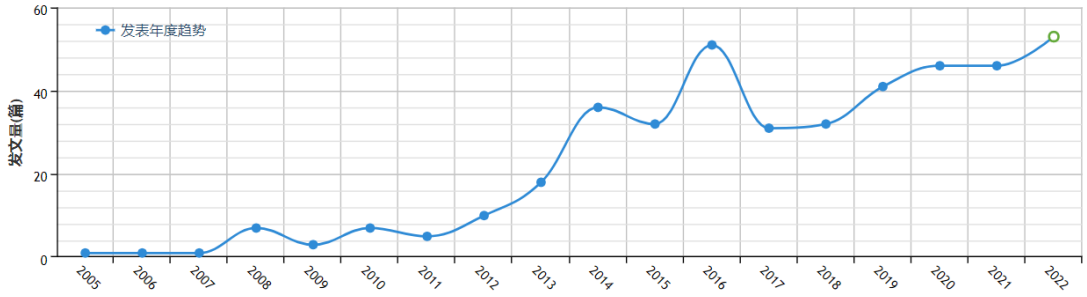
这里给出本文对“在线健身平台”的定义，可以分为广义和狭义两种。广义的在线健身平台既包括我们熟知的 Keep、薄荷健康等运动健身、身材管理 APP，又包括例如 Bilibili、抖音、微博、小红书这类以社交为主的包含健身元素的平台；而狭义的在线健身平台只包含主营业务为健身教学、实时监控、健身社交和健身用品健身餐售卖等覆盖健身行业上中下游的平台。

目前市面上从事健身教学、实时监控、健身社交和健身用品健身餐售卖方面的健身 APP 众多，产品设计的同质化严重，从而导致健身 APP 领域竞争激烈，存续期短。作为相应全民健身和健康中国口号的健身行业，在遭遇新冠疫情冲击后，线下健身行业市场占有率出现下滑，线上健身业从市场占比来看已经赶上线下健身，未来有占据市场主体的趋势。但目前学界对于线上健身平台的研究较少，目前尚缺乏结合新冠疫情背景从购买行为视角对线上健身平台盈利模式的研究，本文将弥补这一空白，致力于增加在线健身平台用户粘性，进一步提升市场份额，对平台的经营模式改进提出合理化建议。

二、文献综述

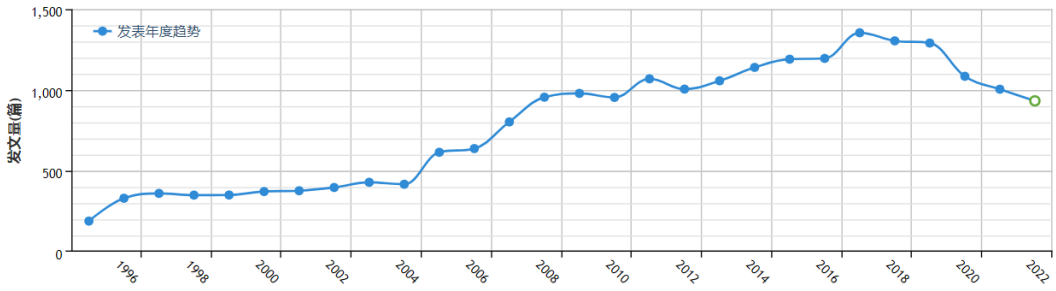
（一）现有文献的可视化分析

在中国知网对主题中包含“在线健身”或“线上健身”的中英文文献进行计量可视化分析，可以看出当前对在线健身的研究热度正在上涨，如图 5 所示。对主题为“全民健身”的知网文献总体趋势分析发现，该主题的年度发文量已经连续十年大于 1000 篇，如图 6 所示。



（数据来源：中国知网）

图 5　包含“在线健身”或“线上健身”的知网文献总体趋势分析



（数据来源：中国知网）

图 6　主题为“全民健身”的知网文献总体趋势分析

我们在知网中对主题包含“全民健身”“在线健身”“线上健身”进行检索，并筛选出所有 2016-2022 年发表的来源类别为“SCI”“EI”“北大核心”“CSSCI” “CSCD”中文学术期刊共 1108 篇，利用 Citespace 软件进行可视化分析。

表 1 和表 2 从作者和机构发文量角度进行研究，体育类高校的发文量排名居于高位，且列示的机构大多位于北京、上海，在地理分布上是否具有虹吸效应需要进一步研究。

表 1　发文量最高的前十二家机构

排名	Freq	Author
1	10	首都体育学院
2	8	北京体育大学
3	7	安徽师范大学体育学院
4	6	上海体育学院休闲学院
5	6	华南师范大学体育科学学院
6	6	上海体育学院
7	4	天津体育学院
8	3	上海体育学院经济管理学院
9	3	东北师范大学体育学院
10	3	华东理工大学体育科学与工程学院
11	3	哈尔滨工业大学建筑学院
12	3	华中师范大学体育学院

（数据来源于中国知网，经作者处理得到）

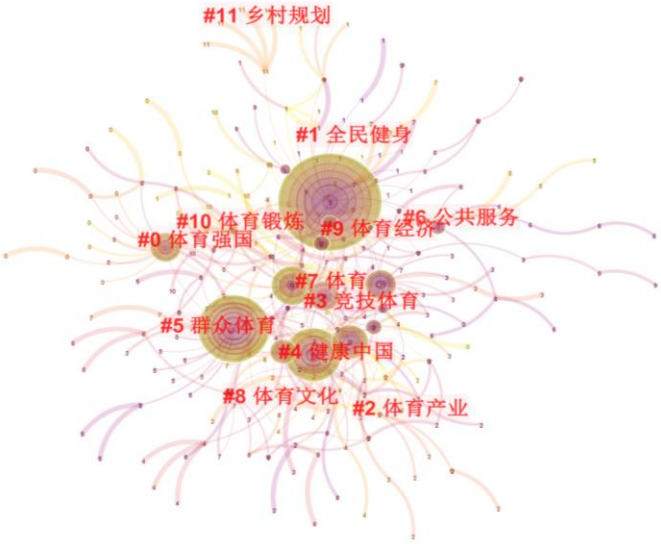
表 2　发文量排名前十七的作者

排名	Freq	Author
1	7	于鸿雁
2	7	刘红建
3	6	郭修金
4	5	马思远
5	5	岳建军
6	5	卢文云
7	4	张凤彪
8	4	施家瑜
9	4	戴健
10	4	于善旭
11	4	邓星华
12	4	陈德旭
13	4	王凯珍
14	4	史小强
15	4	舒盛芳
16	4	孙科
17	4	钟秉枢

（数据来源于中国知网，经作者处理得到）

通过对文献关键词共线图（图 7）的分析，可以得知当前最受学界关注的关键词前五名为“全民健身”“体育产业”“竞技体育”“健康中国”“群众体育”，其中“体

育产业”的受关注度已经跻身第二，从研究热点来看具备研究的意义和必要性。图 8 展示了排名前 11 的关键词突现情况，其中“乡村振兴”“公园设计”“消费行为”自 2021 年成为热点。本文将紧密围绕“消费行为”，结合全民健身现状，对健身行业进行调研分析。



（数据来源于中国知网，经作者处理得到）

图 7　关键词共线图

（数据来源于中国知网，经作者处理得到）

图 8　关键词突现图

（二）全民健身的文献综述

自 2014 年“全民健身”上升为国家战略，学界对于健康中国、全民健身的实施路径和进行了深入探讨，多地的新闻报刊对本地的全民健身普及方式和进程做出报道。周德书和黄元骋（2022）从全民健身国家战略角度，论述了全民健身的政治格局、逻辑内涵、精神实质、时代特征，认为全民健身和健康中国行动的各项方针政策最终目的都是满足人民对美好、健康、幸福生活的追求。张鹏等（2021）和侯光定等（2021）从“新时代”和“后疫情”两个历史背景进行了全民健身战略落实的路径研究。刘红建等（2022）梳理了我国全民健身政策体系的演进历程，认为当前的全民健身政策已经达到了全区域、全人群、全周期的供给，论述了推进全民健身政策体系向治理效能转化、全民健身从理论到实践的路径。曾繁荣（2022）从赛事运营、教育培训、私教服务、礼仪服务、商业表演及文化宣传等方面对全民健身背景下的大学生体育服务模式进行探究，分别制定了运营方案。田春兰和张秋婷（2021）在城市低收入群体的角度，分析了该群体健身的必要性、现存问题，对健身建立包含法制建设、沟通协商、需求反馈、宣传引导的“四位一体”保障机制。李冬梅（2022）结合空间经济学的有效供给理论，针对大众公共体育空间的不足提出了设计理念、结构布局、场所文化、存量空间、资源配置五方面的改进意见。王爽等（2021）从健身设施的数量与健身需求不够匹配的角度，提出运用地理信息系统 GIS 高效处理体育健身设施的布局问题，解决全民健身设施空间的分布不合理问题。

（三）有关健身需求的文献综述

当前，多方面因素驱动大众产生了健身需求。大众审美的多元化趋势是导致全民健身的原因之一。赖锐（2022）从哲学的角度阐述了身体美学，认为它既是实践美学的具体化，又将审美活动中的“身-心”因素合并至“身体”中。所以大众审美偏好的变化是实用需求和心理需求的综合结果。

通过对网络和文献资料的搜索，我们发现导致健身需求增加的因素还包括健

康养生观念的普及、政策引导、择偶偏好的变化、高品质生活的追求、极端疫情的限制、职业歧视等。从社会对健康养生和运动健身的偏好来看，现有研究主要集中于老年人、病人和大学生等群体，对其他年龄段和属性的研究并不透彻，不具有雷雪梅（2022）研究了健身对老年人认知功能的影响，余玲等（2021）实验发现传统体育养生运动处方可以改善轻度抑郁女大学生情绪和提高其睡眠质量，且停训后仍有维持作用，艾冬梅等（2022）发现了健身气功对大学生动态平衡能力的改善作用，黄开颜和李乃适（2022）临床研究得出中国传统养生运动对糖尿病预防起积极作用。从国家政策引导方向来看，全民健身已经成为国家战略，从城乡、学生和地区体育经济角度同步推进，上一小节已从全民健身角度进行了详细的文献综述，此节中不做赘述。从社会性别认同角度来看，赵玉婷（2020）站在男性角度，认为单一的性别刻板意识正在转变，传统的社会价值体系中对于男性的标签代表了父权制社会的审美倾向，娱乐文化中阴柔化的选星偏好使得男性拥有追求个性化审美的权力。樊梦吟（2021）和刘怡（2021）站在女性角度，对“A4 腰”、“反手摸肚脐”、“BM 风”等热潮进行分析，认为社交媒体时代的女性身材审美仍然单一甚至畸形，受到父权制、苗条文化和现代消费的规训。田芊（2013）对中国女性的择偶倾向的研究发现，男性的社会经济条件、性格特质、生理条件在女性择偶时受到青睐，其中物质条件与生理条件相关性较大；而 Buss 和 Barnes（1986）的研究表明，男性更偏好有生理吸引力的女性。从大众对高品质生活的追求来看，线下健身房的健身器材大多较为丰富专业，常见的选址有高档小区、购物中心、企业大楼、学校附近，以中高端健身房为主，能够满足一般上班族的碎片化锻炼需求。从乐刻、一兆韦德、超级猩猩等健身连锁品牌门店快速扩张中可以看出，大众对于健身的需求远未饱和，渗透率较发达国家来说处于较低水平。从职业歧视层面来看，赵耀（2006）研究发现劳动力市场存在容貌和身材歧视，不可否认在一些职业（例如：主持人、模特、演员）中对容貌和身材的特殊偏好难以消除，不符合主流审美的容易受到非常严重的排挤。从新冠肺炎疫情对居民健身的影响来看，在疫情全面爆发之时，线下健身房暂停门店业务，居民生活处于封闭状态，对线下健身经济产生了巨大冲击。由于健身需求的持续性，消费者将目光转向提供陪练视频、专业化服务的线上健身平台，“美丽芭蕾”、“帕梅拉”

等系列视频在 Bilibili 弹幕网上的总播放量已经远超 1 亿次，在线健身平台的受众群体显著增加。方子隽（2021）对 B 站健身视频中的弹幕文本进行研究，认为在线健身已经成为社会交往的一种形式。

（四）线上健身的文献综述

线上健身在我国的发展较迟，但近几年国内的线上健身行业进展势头迅猛，大众对于碎片化健身的需求增加，市场上出现了不少提供在线健身服务的竞品，在线健身也引起了学界关注。陆佳莉（2017）早在 2017 年就对体育运动健身类 APP 的概念、功能、发展现状以及产品竞争等内容进行分析。钟丽萍等（2020）分析了线上健身在疫情下的发展态势，总结了疫情下在线健身的制约因素，并提出了后疫情时代的在线健身策略。刘建武等（2021）提出健身服务线上线下融合发展的可能性，并对其机理与路径进行了管理运营方面的研究。刘高福和李永华（2021）通过问卷调查，分析了在线健身平台中用户的求助、社交、反馈、助人对价值共创行为的影响。牟粼琳等（2021）以某线上健身 APP 为例，探究了区块链底层技术在线上健身中的贡献，对监管体系完善、线上线下模式结合、健身公共信息服务水平提升、智力支持等方面提出改进策略。钟丽萍等（2021）以 B 站为例，对新冠疫情下的在线健身视频网站进行了运营研究。刘东锋和傅钢强（2020）在新冠疫情的背景下，对在线健身服务持续使用意愿的影响因素进行探究，提出了增加用户持续使用意愿的相关建议。

三、调查方案设计

（一）调查目的与内容

在本次调查研究中，我们从被调查者对线上及线下运动健身平台的选择偏好、在线运动健身媒介平台偏好选择的主要依据、在线运动健身媒介平台产品消费情况、潜在购买意愿情况、影响因素等多角度出发，针对不同年龄层群体进行区分调查，为当下各大在线运动健身媒介平台从用户粘性、潜在客户挖掘、营销策略、付费产品定位等多方面提供建议与改进方向。本次调查主要包括以下四个方面：一是被调查者的基本信息，二是被调查者选择在线运动健身媒介平台主要依据，三是被调查者于在线运动健身媒介平台的消费情况，四是被调查者于在线运动健身媒介平台的产品购买意愿情况。针对以上四个方面，并根据以下主要目的进行了调查研究，具体问卷及问卷编码表详见附录。

第一，为了解当前市场用户运动健身现状，搜集被调查者对于线上及线下运动健身平台选择偏好、消费情况。

第二，为分析具有健身需求的用户特征并描绘用户画像，搜集每一位用户的性别、年龄、职业、学历、常住地区、月收入、月消费等。

第三，为分析当前市场各大在线运动健身媒介平台用户粘性的背后影响因素并挖掘潜在客户，搜集用户选择在线运动健身媒介平台的偏好及主要依据。

第四，为了解当前市场在线运动健身媒介平台各产品的消费情况，搜集用户付费产品种类及付费意愿等，从而分析付费产品定位，并给予各大在线运动健身媒介平台建议与改进方向。

（二）调查对象

此次调查选定的调查对象为具有健身需求的用户，由于我们采用线下健身与线上健身的对比分析，线下健身的调查对象为粤西地区（湛江市和茂名市）和南京市线下健身房的付费用户，线上健身的调查对象为粤西地区（湛江市和茂名市）和江苏省（主要集中于南京市和南通市）具有健身需求的在线运动健身媒介平台

使用用户。

（三）调查费用

此次调查费用预算如下：

通讯费：200 元 交通费 1000 元

（四）调查方式和方法

本调研小组首先通过文献分析法得出当前有关全民健身、线上健身的热点话题，对前人的理论和实证研究结果进行总结，提出了具有时效性的研究方向。我们观察到健身行业在近几年发展迅速，且“全民健身”国家战略和“健康中国”行动的提出使得该类主题符合政策导向。

接着我们从目前新冠疫情的实际情况出发，对线上健身行业快速发展进行预测与事实验证，最终确定了“疫情后在线健身平台的使用及偏好情况”的研究主题。然后参考行业热点，制定了相对完善的问卷，进行小范围的预调研，发放了 200 份问卷，收回问卷 184 份，有效问卷共 168 份。

之后再根据反馈结果对设问进行微调，重新发放问卷 500 份，收回 443 份，经筛选共 418 份有效问卷，并对问卷进行了质量控制、数据预处理和信度效度分析，之后运用数据分析方法建立了决策树模型、Logistic 模型等模型，运用 Pearson 相关分析、K-means 聚类等方法分析收集到的调查数据。

最后根据模型分析结果，我们对在线健身行业本身、营销现状和用户挖掘方式进行分析，并提出了具有针对性的结论和建议。

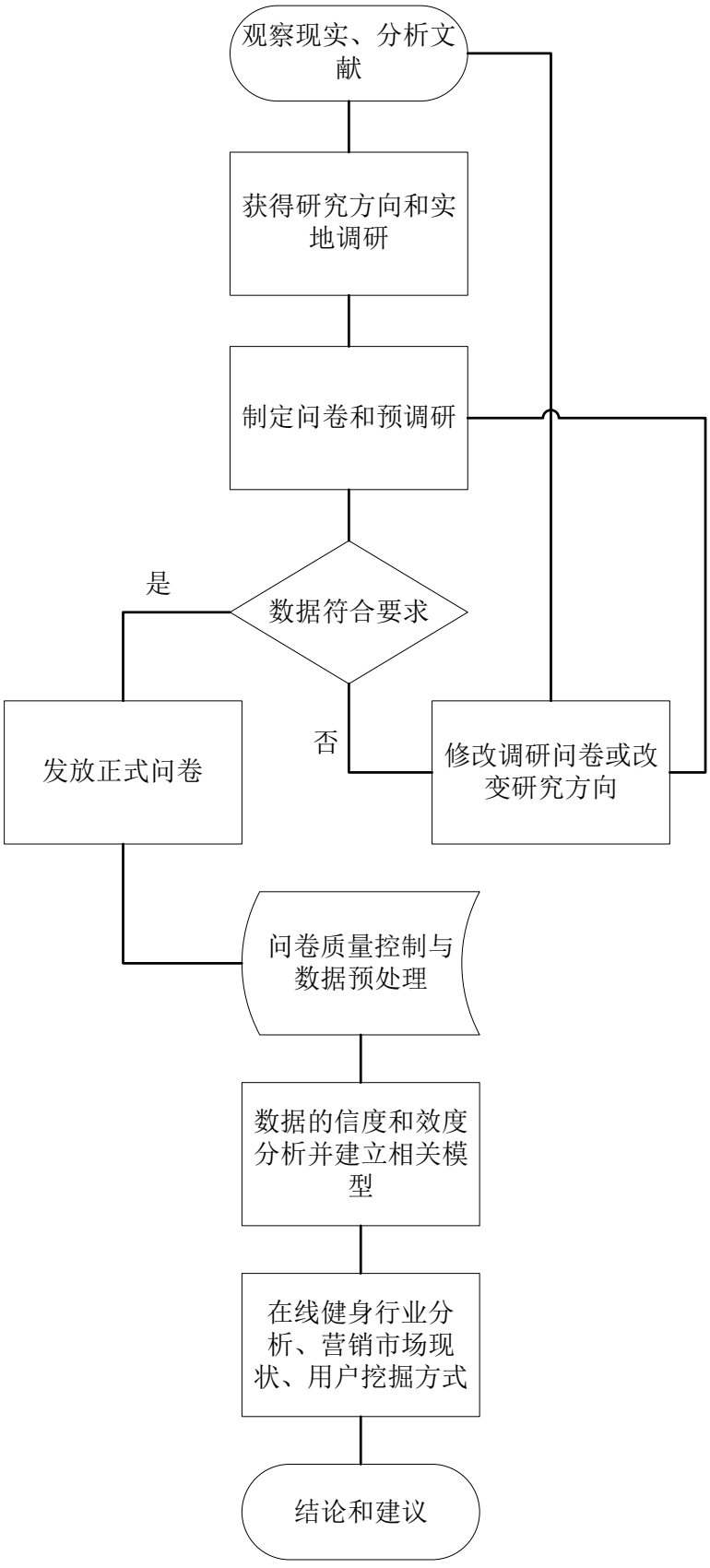

观察现实、分析文献
获得研究方向和实地调研
制定问卷和预调研
数据符合要求
是
否
发放正式问卷
修改调研问卷或改变研究方向
问卷质量控制与数据预处理
数据的信度和效度分析并建立相关模型
在线健身行业分析、营销市场现状、用户挖掘方式
结论和建议

图 9　项目流程图

1、调查方式

本次调查中，我们的主要研究精力集中于线上运动健身，但由于运动健身存在线上和线下两种形式，我们将对线下健身做一定的调查，用于和线上健身做对比分析；我们不仅期望从非付费角度挖掘在线运动健身平台的用户粘性的主导因素，也期望从付费角度挖掘在线运动健身媒介平台各付费产品的潜在消客户及产品定位。粤西地区及江苏南京、南通作为 GDP 水平较高的几个地区，居民生活水平高，对消费的需求高，因此线上健身及付费产品在这几个地区存在发展潜力和发展空间。因此，我们从科学性、完整性和严谨性三个角度出发，主要采取以下四种调查方式：

（1）文案调查

在问卷调查前期，为了对在线运动健身有一定的初步了解，我们先利用计算机在互联网平台中进行关键词搜索；另外，查阅已有文献，查看相关同类型问题的调查研究方法、数据分析方法，并进行资料整理。

（2）问卷调查

我们对线上运动健身的相关数据主要采用问卷调查的方法搜集。随着科技水平的迅猛发展，线上网络问卷的发放效率高、成本低的同时，也能满足我们对样本容量的需求。另外，对于不同分类问题的不同回答，可以设置不同的跳转，缩短被调查者填写时间、提高发放效率，也能提高问卷质量，保证其严谨性。

该部分问卷调查主要分两个步骤，一是预调研，二是正式调查。预调研阶段我们将事先设计的问卷采用线下填写（线下拦截扫描二维码，无需纸质问卷）的方式发放出去，根据回收结果的合理性以及填写者的反馈，修改相关问题，确定最终问卷；正式调查阶段，我们采用线上网络发放问卷的方式，搜集数据，提高搜集效率。

（3）访问调查

我们对线下运动健身的相关数据主要采用访问调查的方法搜集。本次我们小组在粤西地区开展了访问调查。我们的本次访问具有较好的灵活性。由于调查者

和被调查者双方面对面交流，交谈的主题可以突破时间限制，调查者可以采取灵活委婉的方式,迂回提问,逐层深入；实地方位调查在市场推广的各个阶段都有着很大的作用，做好目标客户的调查，针对于目标市场的高校学生、体育训练馆、健身俱乐部等地方。可以在访问过程中使用图解材料，直观明了。

2、调查方法

（1）Logistic 回归分析

Logistic 回归分析是一种广义的线性回归分析，Logit 是二元选择模型中的一种，常用于研究影响关系分析，即 X 对 Y 的影响情况。其中，Y 为分类数据，X 可以为分类数据，也可以是定量数据。

Logistic 回归与线性回归的模型形式有部分类似，都存在W^TX+b的部分，其中 W 和 b 为待估参数，而不同的是两个模型的因变量，线性回归模型直接将 Y 作为因变量，即$Y=W^TX+b$，而 Logistic 回归模型的因变量是通过函数 L 将将W^TX+b与隐状态p相对应，即

$$p=L(W^TX+b)$$

其中 L 为 logistic 函数，为$L=\frac{1}{1+e^{-x}}$。则最终得到的 Logistic 模型的表达式为：

$$p=\frac{1}{1+e^{-(W^TX+b)}}$$

之后，再根据p与 1-p的大小决定因变量的值。

本文将运用 Logistic 回归分析相关变量，如“是否选择于线上平台使用过和运动健身相关的付费功能”，分析影响因素，从而得出结论。

（2）Person 相关分析

Pearson 相关分析用于研究定量数据之间的关系情况，以及判断他们的紧密程度情况。

Pearson 相关分析常用 Pearson 相关系数来衡量两个数据集合是否在一条线

上，来衡量定距变量间的线性关系，是最常用的相关系数，又称积差相关系数，取值在-1 到 1 之间，绝对值越大，说明相关性越强。该系数的表达式为：

$$r=\frac{N\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{N\sum x_i^{\,2}-\left(\sum x_i\right)^2}\sqrt{N\sum y_i^{\,2}-\left(\sum y_i\right)^2}}$$

通常情况下通过以下取值范围判断变量的相关强度：相关系数 0.8-1.0 表明极强相关，0.6-0.8 表明强相关，0.4-0.6 表明中等程度相关，0.2-0.4 表明弱相关，0.0-0.2 表明极弱相关或无相关。

对于 x,y 之间的相关系数 r：当 r 大于 0 小于 1 时表示 x 和 y 正相关关系，当 r 大于-1 小于 0 时表示 x 和 y 负相关关系，当 r=1 时表示 x 和 y 完全正相关，r=-1 表示 x 和 y 完全负相关，当 r=0 时表示 x 和 y 不相关。

本文将运用 Pearson 相关分析分析线上平台运动健身相关的消费总额与各个因变量之间的相关系数，分析影响因素，从而得出相关结论。

（3）K-means 聚类

K-Means 是聚类算法中的最常用的一种，对于给定的样本集，计算样本之间的距离，按照距离的大小，将样本集划分为 K 个簇，使簇内的点距离尽可能得小，而簇间的距离尽量得大。

其聚类步骤是：预将数据分为 K 组，即随机选取 K 个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

（4）文本挖掘

本文利用 Python 软件爬取 bing 浏览器的相关主页，获取文本数据进行数据

预处理，并进行文本分词，对处理后的文本做可视化及搭建模型分析。绘制“线上健身”、“线下健身”等相关关键词的词云图，进行舆情分析。

（5）决策树

分类决策树模型是一种描述对实例进行分类的树形结构。

决策树由结点和有向边组成。结点有两种类型：内部结点和叶结点。内部结点表示一个特征或属性，叶结点表示一个类。用决策树分类，从根结点开始，对实例的某一特征进行测试，根据测试结果，将实例分配到其子结点；这时，每一个子结点对应着该特征的一个取值。如此递归地对实例进行测试并分配，直至达到叶结点。最后将实例分到叶结点的类中。

（6）结构方程模型

结构方程模型（SEM）是一种基于变量的协方差矩阵来分析变量之间关系的多元数据分析方法，它在估计一组观察变量与其代表的潜变量、因子的关系的同时，分析各潜变量之间的关系，这样潜变量之间的关系估计不受测量误差的影响。它通过路径范式来描述自变量与因变量之间的关系并用线性方程式来表述自变量和因变量的数目。结构方程模型（SEM）包含测量模型和结构模型两个基本模型。

测量模型表示潜在变量与观测变量间的共变关系，可看作一个回归模型，由观测变量向潜在变量回归。用方程可以表示为：

$$y_i = \Lambda w_i + \varepsilon_i, i = 1, \cdots, n$$

其中，y_i表示$\mathrm{p} \times 1$的观测向量；Λ为观测向量对应的$\mathrm{p} \times \mathrm{q}$的因子矩阵；$w_i$为$\mathrm{p} \times 1$的因子得分向量；$\varepsilon_i$是与$w_i$独立的误差项。

结构模型部分表示潜变量间的结构关系，也可看作一个回归模型，由内生潜在变量对若干内生和外生潜在变量的线性项作回归。用方程表示为：

$$\eta_i = \prod \eta_i + \Gamma \xi_i + \delta_i, i = 1, \cdots, n$$

其中，η_i和ξ_i分别是$q_1 \times 1$和$q_2 \times 1$的潜在变量；Π和Γ表示未知的相关参数矩阵；δ_i为误差项。

（7）情感分析

文本情感分析是指对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。

互联网上产生了大量的用户参与的、对于诸如人物、事件、产品等有价值的评论信息。这些评论信息表达了人们的各种情感色彩和情感倾向性,如喜、怒、哀、乐、批评、赞扬等。基于此,潜在的用户就可以通过浏览这些主观色彩的评论来了解大众舆论对于某一事件或产品的看法。

（五）抽样设计

为了保证抽样的科学性，我们采用概率抽样的方式，即依据随机原则，总体中每一个样本都有一定的概率被抽中。为了减少抽样误差，综合考虑，我们采取分层抽样与三阶段抽样相结合的方式，选取入样样本。抽样的具体过程如下：

1、分层抽样

首先，南京市下辖鼓楼区、玄武区、建邺区、秦淮区、栖霞区、雨花台区、浦口区、六合区、江宁区、溧水区、高淳区共 11 个行政区，南京市统计年鉴各地区将总体按照最新国内生产总值 GDP 排序，如表 3 所示。由于 GDP 水平的不同可能会导致消费水平的不同，导致付费产品的发展潜力和发展空间也不同，将总体以国内生产总值以 1000 亿元为分界划分为两层，为国内生产总值 GDP 水平 1000 亿元以上（经济较发达区）和国内生产总值 GDP 水平 1000 亿元以下（经济欠发达区）。

表 3　南京市分区 GDP 及社区基本情况表

市辖区	国内生产总值 GDP(亿元)	社区数	居民户数	GDP 排名
江宁区	2509.32	201	462130	1
鼓楼区	1772.60	120	134946	2
栖霞区	1569.15	120	196590	3
秦淮区	1286.60	105	262932	4
建邺区	1121.53	60	134946	5
玄武区	1108.66	58	155606	6
雨花台区	947.14	63	121106	7
溧水区	911.51	112	162857	8
六合区	514.39	146	336775	9
高淳区	513.13	145	162857	10
浦口区	443.57	111	281590	11

2、不等概率三阶段抽样

第一阶段的 PPS 抽样：这一阶段采取概率比例规模抽样，以国内生产总值 GDP 水平 1000 亿元以上（经济较发达区）这一层为例，入样概率以居民户数为比例计算。以居民户数为基本单位编写抽样框，全市共 2605092 户，给每户赋予编码，编码依次累加，对应编码及行政区如下表。利用计算机生成 3 个 1~2605092 中的随机数，将随机数编码对应的居民户所在行政区纳入样本，这就得到了经济较发达区的初级抽样单元。本次抽取的为江宁区、栖霞区和秦淮区。经济欠发达区的初级抽样单元与经济较发达区抽样方法完全相同。因此，这一步中一共抽取 6 个行政区。

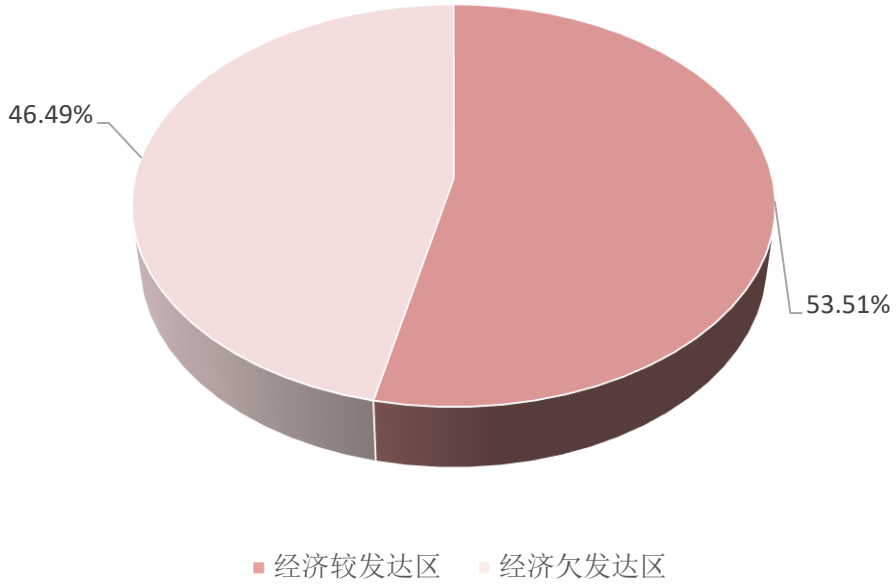


图 10　分层社区数比重

表 4　经济较发达区抽样表

市辖区	居民户数	累计户数	编码范围	生成随机数	抽中	再编码
江宁区	462130	462130	1~462130	56609	是	1
鼓楼区	134946	597076	462131~597076		否	
栖霞区	196590	793666	597077~793666	724195	是	2
秦淮区	262932	1056598	793667~1056598	988358	是	3
建邺区	134946	1191544	1056599~1191544		否	
玄武区	155606	1347150	1191545~1347150		否	

第二阶段的分层抽样：对于第一阶段的 PPS 抽样得到的行政区再进行分层抽样，以经济较发达区为例，将抽取到的 3 个行政区的所有社区编码，进行简单随机抽样。我们决定抽取经济较发达区的 6 个社区和经济欠发达区的 4 个社区，共 10 个社区。经济欠发达区的抽样方式与经济较发达区抽样方式完全相同。

第三阶段的系统抽样：通过网络调查搜索，如二手房 APP，并咨询南京市各行政区“社区行政服务中心”获取得到抽到的入样社区的小区及其居民楼栋数的数据。得到数据后对居民楼编码，通过生成随机数进行简单随机抽样，每个社区抽取 5 栋居民楼。经济欠发达区的抽样方式与经济较发达区抽样方式完全相同。

四、调查实施与质量控制

（一）调查组织分工

在此次市场调研开始前，我们首先熟悉本次调研的目的、性质、具体要求等，事先学习访谈技巧与调查知识，并对可能遇到的问题与难点设置应急预案，在充分准备下，再进行调研与访谈。

考虑到新冠疫情防控现状，在预调研阶段，我们采取线下发放问卷的方式，在抽取到的六个区采取街头拦截式发放问卷，在征得被调查者同意后进行调查与访谈。在正式调查部分，采取线上发放问卷的方式，提高发放效率。

根据各个成员的综合能力、专业特长，分配各自的任务，并相互监督，最终确定本次市场调查工作的具体安排如表 5：

表 5　调查任务分配图

成员	工作安排
成员一	问卷设计、调查方案设计、抽样设计、数据收集、论文排版
成员二	问卷设计、数据收集、研究现状分析、论文排版
成员三	数据收集、文本挖掘、数据分析
成员四	访谈设计、数据收集、市场现状分析、数据分析
成员五	数据收集、用户特征分析、数据分析、PPT 制作

（二）调查实施进度

根据本次调查各项工作的具体安排及难易程度，我们设定任务调查实施计划如图 11：

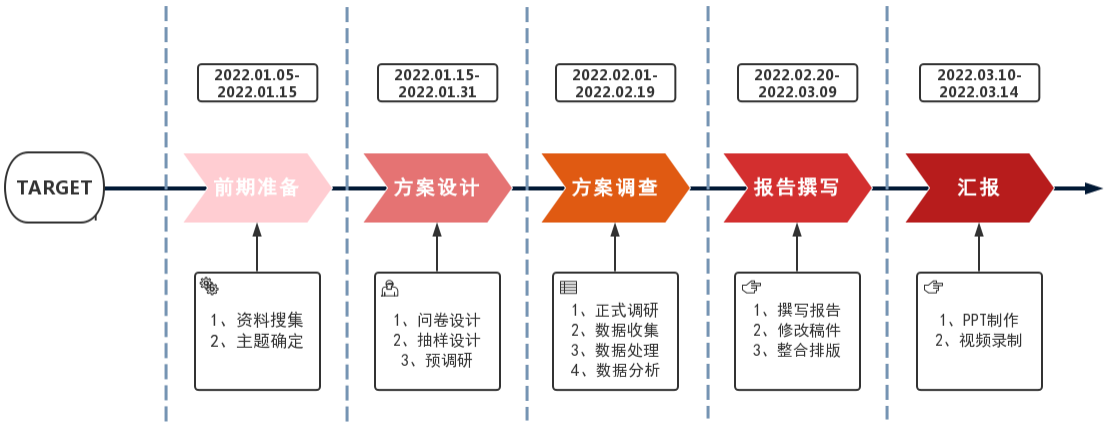


图 11　调查进度流程图

（三）质量控制

为了保证最后得到数据的科学性、准确性和合理性，我们要对整项调查的全程做好质量控制。对于不合理的问卷做剔除处理，若样本量不足，再进行补发，以达到需求样本量。

1.调查前的质量控制

在调查实施前，考虑到质量问题的主要来源为问卷与调查员本身。

对于问卷而言，我们多次调整设问，并对一些问题做备注性解释，做到设问清晰、没有歧义、方便理解。

对于调查员而言，我们事先对调查员进行相关的培训，对每个问题的设定做具体的讲解，做到每一位成员对各个问题有相同的理解，从而预防调查过程中由于调查员对问题不同的理解使结果造成的导向性偏差。

2.调查中的质量控制

在调查实施的过程中，对于被调查者提出的疑问，调查人员要做好尽可能详尽仔细的解释。同时，各个调查员之间随时保持联络，对于自己无法解决的疑惑，可以寻求队友帮助，给与解答。另外，对被调查者提出的疑问做好详细记录，方便后续问卷整理时的核对。

3.调查后的质量控制

在调查结束后，对得到的数据做好审核、筛选工作，对不符合要求的问卷做剔除处理，从而保证问卷的可靠性。对于收集到的问卷，若有以下问题，即可剔除：

（1）在问卷设计中，若出现前后矛盾的情况，如 Q2 和 Q6 起到筛选问卷的作用，Q2 搜集被调查者的年龄数据，Q6 搜集被调查者的出生年月数据，两个问题结合起来，判断是否有误，若两个问题回答不相符，即可剔除问卷。

（2）问卷填写时间小于 30 秒，即可认定为没有认真读题，可做剔除处理。

（3）矩阵单选题中，若出现一份问卷全部填写相同答案，没有变化，或出现极端数据情况，即可认定为乱填，做剔除处理。

另外，对收集到的问卷筛选后，对搜集数据也要进行质量控制，包括以下内容：

（1）记录与审核

对于收集到的一手数据，我们首先对其进行了预处理，包括审核、筛选、排序等，以保证数据准确性。根据上文中的质量控制处理方法，审核问卷、筛选得到有效问卷。

（2）编码与录入

对于回收的问卷进行编码处理，编码与问卷序号保持一致，具体编码见附录 2，为 Q1、Q2、Q3、……Q25，选中均编为“1”，未选中为“0”，整合数据，最终导出为 EXCEL 格式数据。

（3）缺失值处理

对于存在缺失值的问卷，我们需要对其进行处理，使得问卷更有效。若一份问卷中存在大量的缺失值，我们可以直接剔除问卷，若一份问卷存在少量的单个缺失值，我们则用该变量的均值作为填补，最终得到最终问卷数据。

另外，我们做了问卷的信度分析和效度分析。

（1）信度分析

信度分析用于研究定量数据的回答可靠准确性，我们通常用 Cronbach α 系数来衡量。

首先分析 Cronbach α 系数，公式为：

$$\alpha = \frac{k}{k-1} * (1 - \frac{\sum S_i^2}{S_T^2})$$

其中，k 为量表中题项的总数，S_i^2为第i题得分的题内方差， S_T^2为全部题项总得分的方差。

如果 Cronbach α 系数的值高于 0.8，则说明信度高；如果此系数介于 0.7-0.8 之间，则说明信度较好;如果此系数介于 0.6-0.7，则说明信度可接受；如果此系数小于 0.6，说明信度不佳。

我们对本次的调研做了相应的信度分析，根据表 6 可以看出 Cronbach α 系数在 0.6 至 0.7 范围之内，即可说明该问卷的信度检验尚可接受。

表 6　信度分析表

名称	校正项总计相关性(CITC)	项已删除的 α 系数	Cronbach α 系数
您的年龄是？	0.372	0.509	0.696
您的最高学历（含目前在读）是？	0.530	0.687	
2020 年 1 月 1 日后，您使用在线运动媒介平台的频率是？	0.642	0.568	
您的月收入？	0.104	0.231	
您的月消费？	-0.022	-0.092	
2020 年 1 月 1 日之后，您于线上平台运动健身相关的消费总额？	0.540	0.571	

（2）效度分析

效度分析指尺度量表达到测量指标准确程度的分析。效度相当于是对于问卷质量的一个前置条件，如果问卷的效度比较好，证明问卷的数据可靠性比较高，问卷数据内部一致性比较高，固然可用来做后续的建模分析。

通俗来讲，效用分析就是要确定设计的题项是否合理，是否能有效反应本次我们调研小组的研究目标——探究影响在线健身 APP 等平台的使用因素以及这些因素带来偏好情况。本次我们的调研对重要自变量做了相应的效度分析，得到了较理想的 KMO 值、巴特球形的卡方值以及 p-value。

KMO(Kaiser-Meyer-Olkin)-value 是用于比较变量间简单相关系数和偏相关系数的指标，主要应用于多元统计的效度分析等。它的公式为：

$$KMO = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} r_{ij\bullet 1,2\cdots k}^2}$$

KMO 值越接近于 1，意味着变量间的相关性越强，越适合做效度分析。统计学家 Kaiser 给出了常用的 KMO 度量标准：0.9 以上表示非常适合；0.8 表示适合；0.7 表示一般；0.6 表示不太适合；0.5 以下表示非常不适合。

而如表 7 所示，本次调研中，我们对采集的有效数据做了相应的效度分析，可以得到的 KMO 值是 0.872，比较趋近于理想结果，这也证明了他们相关性较强。

表 7　效度分析表

项目	因子 1	因子 2	因子 3	因子 4	因子 5	因子 6	因子 7	因子 8	共同度
Q1	-0.02	0.15	0.74	-0.00	-0.05	-	-	-	0.568
Q2	0.78	0.20	-0.04	-0.13	-0.19	-	-	-	0.703
Q3	0.42	-0.37	0.07	0.25	0.29	-	-	-	0.462
Q4	-0.85	0.09	-0.01	-0.06	-0.07	-	-	-	0.733

Q10	0.07	-0.02	0.41	0.63	0.14	-	-	-	0.598
Q11	-0.02	0.50	-0.53	0.08	0.14	-	-	-	0.554
Q13	-0.04	-0.00	-0.10	-0.04	0.83	-	-	-	0.700
Q14	-0.01	0.62	0.20	-0.06	0.33	-	-	-	0.531
Q17	-0.08	0.06	-0.31	0.75	-0.18	-	-	-	0.703
特征根值(旋转前)	1.54	1.23	1.10	1.07	1.03	1.00	1.00	1.00	-
方差解释率%(旋转前)	41.87	29.43	8.45	8.21	7.93	7.69	7.69	7.69	-
累积方差解释率%(旋转前)	41.87	21.31	29.76	37.97	45.90	53.59	61.29	68.98	-
特征根值(旋转后)	1.51	1.20	1.14	1.07	1.04	1.00	1.00	1.00	-
方差解释率%(旋转后)	31.62	19.26	8.80	8.23	7.98	7.69	7.69	7.69	-
累积方差解释率%(旋转后)	31.62	20.89	29.69	37.93	45.90	53.59	61.29	68.98	-
KMO 值		0.872				-			
巴特球形值		118.173				-			
df		78.000				-			
p 值		0.002				-			

五、基于文本挖掘的云健身评价情感分析

近年来，科技的飞速发展和智能手机的迅速普及，对人们的生活方式产生了巨大的影响，云健身开始进入了大众的视野，在线健身用户的规模在逐年递增。尤其是疫情诞生之后，在线健身用户数量得到了显著的提升，这说明云健身已经是大众所向。这里为了证实大众对云健身的评价态度，我们爬取了 2000 条中国大学生慕课、腾讯课堂及网易云课堂 3 个 app 对在线健身平台网课的评价，制作了情感分析系统。具体内容如下：

（一）评论数据获取

本研究选取运动健身作为关键词进行中国大学生慕课筛选，分别抓取到用户对于“在线健身评价”、“健身网课”、“云健身”等评价的文本数据，对初始数据进行初步清洗后，利用中文分词将文本内容中的词语进行分割并统计词频，除去单字以及虚词后，得到热度最高的词语，绘制如下词云图 12。

图 12　运动健身关键词词云图

（二）文本分词

1、数据清洗

中国大学生慕课等健身网课平台中用户可以删除已发表评论，开发者也可以删除恶意刷屏的评论，而这些评论无论是否被删除都会在爬取的数据中显示。我们认为这部分评论不具有充分的客观性，无法代表用户的真实态度，因此利用“七麦网”本身具备初步筛选功能，下载按时间倒序排列，且未被删除的数据。在此基础上，针对所获取的 3 个 APP 在线评论均为可正常显示的文本数据。针对慕课 APP 在线评论的特点，具体问题具体分析，数据清洗工作主要分 3 步进行。

首先，删除带有明显广告性质的评论。观察发现中国大学生慕课等健身网课平台中的评论会存在部分“水军”发布的带有带有明显广告性质的评论，如：“瑞银信，中国十大 POS 机排行交易第一，大品牌，找我办理，直接开后台，返点给客户，终身售后，绝对放心，怕封卡，怕降额就找我装瑞银信。完美支付养卡提额神器 3 个月准时提额”，该类在线评论数据与 APP 本身无关，不具有反映用户态度的作用。因此，使用 Excel 表格的查找、筛选等功能对此类评论进行删除。

然后，删除过短或表意不明评论。对于一个完整的评论，若其中只包含诸如“@”、“#”、“【”、“】”、“.....”等符号或“的”、“了”、“是”、“aaaaa”等字符，而不具有反映用户态度、产品自身特性的词语，则认为该评论不具有实际意义、表意不明的评论，后续情感极性无法判别。最后得到有效文本数据库。

最后，将所有中文文本数据处理为“UTF-8”的编码形式。UTF-8 它是一种针对 Unicode 的可变长度字符编码，能表示更多的语言文字，更加通用，尽可能减少后续 Python 操作时的乱码问题。

2、在线评论关键词提取

通过以上数据库清洗，我们所获得的文本数据仍然为句子或者小篇章级，而进行情感分析的对象应该是能够表达中文语义的最小单元——词语，因此我们需要对清洗后的有效评论数据进行文本分词，将每个句子处理为若干词语组成的集

合。中文分词作为自然语言处理的研究领域之一，也是文本情感分析工作中的重要环节，它是指采用计算机技术对中文文本进行分词处理。

不同的分词方法对每个句子进行分词得到的分词结果也不尽相同，分词效果的好坏对后续建立模型及其他工作有着重要的影响。自然语言处理这一领域发展至今，人们广泛使用的中文分词技术主要基于以下三种：词典、统计、语法和规则。在情感分析的过程中，我们需要将词语作为最小语义单元开展研究工作，中文文本分词是十分必要的一步。基于现有分词理论，整个开源领域，陆陆续续做中文分词的人有很多，但是仍在维护的且质量较高的中文分词工具并不多。下表为比较常见的几款分词工具：jieba 分词、HanLP、SnowNLP 等。

Python 第三方工具包 Scikit-learn 提供了 TF-IDF 算法的相关函数，本文主要用到了 sklearn.feature_extraction.text 下的 TfidfTransformer 和 CountVectorizer（）函数。其中，CountVectorizer 函数用来构建语料库的中的词频矩阵，TfidfTransformer 函数用来计算词语的 tf-idf 权值。TfidfTransformer 函数有一个参数 smooth_idf，默认值是 True，若设置为 False，则 IDF 的计算公式为$\mathrm{IDF} = \log(\mathrm{D_n}/\mathrm{D_t}) + 1$。本文采用默认参数 smooth_idf=True。由于此前已进行过文本预处理，在文本数据清洗、分词、去停用词之后，得到分词后的文件“cut_commenti.txt”，然后进行基于 TF-IDF 方法的文本关键词抽取工作。以表 8 中的文本为例，具体的代码执行步骤及输出示例如下：

表 8　分词表

这门 课 帮助 我 锻炼 了 身体 的 柔韧性 和 协调性 ， 在 学习 生活 中 让 我 ...
练习 瑜伽 最大 的 享受 就是 感受 身心 的 高度 契合 ， 感受 自己 身体 的 每 ...
我家 宝贝 非常 满意 老师 的 课程 ， 这个 课程 生动有趣 ， 对 身体 有 很大 的 用处
老师 知识 渊博 ， 实际 经验丰富 ， 讲课 内容 条理清晰 ， 最 重要 的 颜值 高 ...
个人感觉 不 满意 ， 很多 细节 点 都 没 讲清 或者 是 说 摆 个 姿势 就 完事 了...

（1）遍历 cut_commenti.txt 中文本记录，将预处理完成的文本按行放入文档集 corpus 中。

（2）利用 CountVectorizer（）函数得到词频矩阵，矩阵由各行的词向量组成，a[j][i]的意义是第 j 个词在第 i 篇文档中的词频，具体见表 9。

表 9　分词表

	专业	健身	动作	协调	学习	帮助	想要	满意	生活	练习	老师	课程	身体	非常
0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0	1
2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	1	1	4	3	0	1
4	0	0	0	0	0	0	0	1	0	0	1	2	1	1

（3）使用 TfidfTransformer（）函数计算每个词于的 tf-idf 权值；

（4）得到词袋模型中的关键词以及对应的 tf-idf 矩阵；

（5）遍历 tf-idf 矩阵，打印每篇文档的词汇以及对应的权重，并输出至 TF-IDF.txt 文件中；

（6）对每条评论，按照词语权重值降序排列，选取排名前 top N 个词最为文本关键词，并写入数据框中。

（7）将最终结果写入文件 topwords.txt 中。

通过以上操作后我们提取到了能够代表每条评论的关键词，但是提取的关键词词 性混乱，而我们想要选取的关键词仅仅是与外卖 APP 本身相关的名词，因此需要对提取的 TF-IDF 值较高的词进行筛选。

3、词性标注

由于汉语的特殊性，很多情况下同一个词语在不同语境中所表示的意义不同，因此这个词语的词性也就不同，例如“游泳”一词，在场景“我最喜欢的运动是游泳”，词性为名词，而在场景“天气好热，在泳池里游泳好舒服”中，词性为

动词，这某种程度上为词性标注带来不固定性。但实际上，汉语中大多数的词语，尤其是实词，通常只有一个或两个词性，并且其中一个词性相对于另一个是高频的，将高频词性作为该词词性进行标注，标注的准确率也会较大，因此基本可以满足度精确度的需求。

本文采用的是已经标注好词性的北大词性标注语料库，在进行词性标注的时候，对语料库中每个单词的对应词性进行，然后将每个单词的高频词性作为这个词语的词性。本文对提取的代表句子基本观点的关键词做词性标注，使用 Python 的 jieba.posseg 模块，jieba 词性标注和其分词流程相似，都是基于规则和统计的方法，换句话说就是在词性标注的过程中，字典匹配和 HMM 一起工作。

（1）首先通过正则表达式判断是否为汉字

（2）如果不是汉字，将继续通过正则表达式进行类型判断。

（3）如果是汉字，则基于前缀词典构建有向无环图，然后基于有向无环图计算最大概率路径。同时，在前缀词典中找出它所分出的词性，如果找不到，则将“x”分配给他，代表未知。如有未登录词，则会通过 HMM 进行词性标注。HMM 是就是在分词任务中，使用“B”、“M”、“S”、“E”四中标签，与句子中的每个字符分别一一对应，在词性标注中 jieba 采用联合模型的方法，将基于字标注的方法和词性标注相结合，使用复合标注集。最终得到 topwords 词性标注结果[pair(‘长胖’，‘a’), pair(‘钱财’，‘nr’), pair(‘浪费’，‘n’), pair(‘健身’，‘v’), pair(‘深圳市’，‘ns’), pair(‘孤单’，‘a’), pair(‘夜晚’，‘t’), pair(‘越来越’，‘d’), pair(‘希望’，‘v’)等。我们把以上带有标注的词语做成字典，方便进行下面的词频统计。

4、词频统计

表 10　中文词性对照表

词性编码	词性名称	词性编码	词性名称	词性编码	词性名称	词性编码	词性名称

Ag	形语素	g	语素	ns	地名	u	助词
a	形容词	h	前接成分	nt	机构团体	vg	动语素
ad	副形词	i	成语	nz	其他专名	v	动词
an	名形词	j	简称略语	o	拟声词	vd	副动词
b	区别词	k	后接成分	p	介词	vn	名动词
c	连词	l	习用语	q	量词	w	标点符号
dg	副语素	m	数词	r	代词	x	非语素字
d	副词	Ng	名语素	s	处所词	y	语气词
e	叹词	n	名词	tg	时语素	z	状态词
f	方位词	nr	人名	t	时间词	un	未知词

获取到提取关键词的词性后，对每个单词在所有 topwords 里面出现的频率进行统计，输出标注了词性以及统计好频率的表格，首先筛选出词性为 n（名词）、nz(其他专名)的词，然后舍去出现频率低于 3 次的词以及与 APP 自身特征不相关的词，最终得到 56 个与 APP 特征有相关的词语，我们选取词频前十的特征词语绘制了如下的直方图。

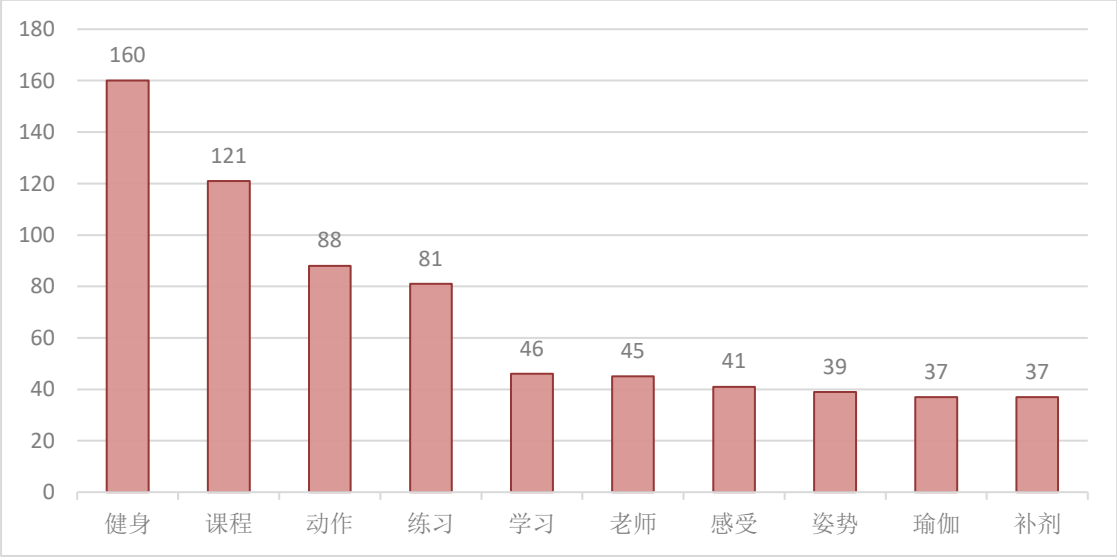


图 13　每 20 条评论中出现次数前十词频图（次）

这些结果说明很多用户使用在线健身更在在乎的是在线健身能否像线下健身一样，课程足够全面，能够多角度的看到指导老师的动作，在学习过程中与老师进行更多的交互。

（三）用户满意度分析

顾客满意(Customer Satisfaction)是以购买者知觉到的产品实际状况和购买者的预期相比较来决定的。菲利普·科特勒认为，顾客满意“是指一个人通过对一个产品的可感知效果与他的期望值相比较后，所形成的愉悦或失望的感觉状态”亨利·阿塞尔也认为，“当商品的实际消费效果达到消费者的预期时，就导致了满意，否则，则会导致顾客不满意”。顾客满意率的计算公式为：

$$\text{顾客满意率} = \frac{\text{满意顾客数}}{\text{顾客总数}} * 100\%$$

综合满意率是各分项的满意率乘以各分项的权值（就是重要程度），然后除以各分项的权值之和。尽管该指标适用于单项简单指标的顾客满意测量，但是足以满足本文研究的需求。就本文而言，我们将顾客满意率作为衡量用户对在线健身的满意程度的指标，定义为用户满意度。如果用户购买某一健身平台的网课后，用户对该产品某一属性特征的实际知觉高于使用前心理预期，该用户发表在平台中与这一属性的相关评论必然呈现积极情感倾向，则认为该用户为对这一特征表示满意一个用户。

我们对文本进行分词、清洗、去除停用词，再使用朴素贝叶斯的方法对结果进行情感分析。如果文本显示的情感为积极，则记为 1，否记为 0，结果存放在 sentiment 这一列。作为对比，我们参考淘宝评分系统，通过评论星级来标注该条评论的情感极性，一星、二星和三星标注为负向情感，四星和五星标注为正向情感训练数据，结果保存在 snip-result 中。所采集数据格式如下表所示：

表 11　分词表

	comment	star	sentiment	snlp_result
0	这门课帮助我锻炼了身体的柔韧性和协调性，在学习生活中让我学会了放松身心，集中精力关注自身。感...	5	1	1
1	练习瑜伽最大的享受就是感受身心的高度契合，感受自己身体的每一个部分。它的好处不仅在于锻炼身体...	4	1	1
2	我家宝贝非常满意老师的课程，这个课程生动有趣，对身体有很大的用处	5	1	1
3	老师知识渊博，实际经验丰富，讲课内容条理清晰，最重要的颜值高，给个大大的好评	5	1	1
4	个人感觉不满意，很多细节点都没讲清或者是说摆个姿势就完事了，摆个姿势我也会啊，我去找球星动作...	2	0	1

这里我们采用了两种方式进行打分，第一种 sentiment 是基于客户使用在线健身网课之后给的对应的星级，统计后占比 94.5%，说明绝大多数的客户对在线健身的网课是很满意的。第二种我们考虑到部分网课可能是学校要求选课，出于是自己学校的老师，打分的星级偏高，但是评价反应自己并没有在网课中学到东西，或者认为网课有需要改善的地方。于是我们基于朴素贝叶斯的方法，对文本信息做了情感分析，其中评论为积极情感的占比 87.5%。这说明大部分学生对网课的积极性很高，虽然网课存在些许不足，但是总体还是得到了大多数学生的认可。

本次分析也反应了在线健身课程带来的不足，例如“个人感觉不满意，很多细节点都没讲清或者是说摆个姿势就完事了，摆个姿势我也会啊，我去找球星动作视频不就好了？感觉篮球的付费视频都不值得购买，某站 up 主的教学视频比他专业多了，还不收钱，感觉他完全就是个门外汉，什么都不懂，就是为了赚钱。”、“课堂回应较少，回答区答案很多人都是复制粘贴没有参考价值，期末考试主观性太强”、“这分明就是在水时间吧？照本宣科一样的讲 PPT 怎么学？完全实在浪费时间，每到重点的地方就一笔带过，怀疑好评都是刷的。”等差评反应出在线健身的网课往往存在以下优缺点：

首先优点有以下：

（1）线上教学突破了地域限制。教练和学生只要通过一台电脑、一部手机就可以开展教学活动，解决了因疫情防控而不能出行、聚集的问题，师生在家就

可以完成教学任务。

（2）丰富了学习方式。现在网络资源很丰富，可供选择的平台也很多，很多平台都是直播加录制的方式，都有课程回放功能，如果在直播时间无法参与学习，学生可以通过回放上课视频，自主学习，使得学生学习知识的途径更广泛，尤其是对一些能够严格自律的学生来说，通过线上学习，效果会更好。

（3）丰富了教学方式，促进了教练的“再学习”。一个优秀的教练除了具备很好的健身技能以外，信息技术手段成为必须。这次开展的网络教学也是对教练信息技术应用水平的一次检验和提高，促使健身教练们勇于上阵，共同学习。同时，教师在探索的过程中，通过互联网获取到一些优质的教育资源，例如微课、短视频、教学课件等，丰富了教学资源，对教学水平的提升有很好的推动作用。

（4）方便沟通。网络教学环境下，必定用到较平时更先进的辅助手段，可以更快速的捕捉到一些动作的细节，再加上学生端可以使用录屏、截图等方式记录课堂笔记，效率无疑是提高了。

（5）没有干扰。网络教学“穿越”了空间距离，无论身处何方，网络教室有多少人，只要是一个人静静地健身，就可以排除许多干扰，对于自控能力强的学生来讲，在线健身会极大的提高效率，可以完全按照自己的节奏锻炼身体。

（6）可以回放。网络教学的确照顾到了那些因各种缘故漏掉了线上教学内容的学生，只要是想健身，总会有办法找到没听到的部分或没理解透的内容进行自觉学习、巩固。

缺点：

（1）教练的角色变成了单纯的"网络主播"。由于教练多数是初次接触线上授课，对于平台的使用处于摸索阶段，教练的角色变成了单纯的"主播"，很多时候都是教练一人在讲，学生在听，缺少现场教学的师生互动，即便教练讲得绘声绘色，但学生看久了会产生疲倦心理，教学效果大打折扣。

（2）缺乏对学生的有效管控。因为没有了现场教学，教练无法约束学生，如果学生的自我约束能力较差，听课的时候注意力不集中，甚至出现中途离开课

堂的现象，也无法及时制止，这样的教学效果注定很差。

（3）线上教学课堂学习氛围不足。线上教学教练无法在课堂上进行巡视，无法直观面对学生的健身状态，更无法开展现场教学，师生之间、学生与学生之间的互动与交流远没有现场教学效果好，这样的课堂显得冷冰冰，没有教室里那种热烈的学习氛围。

（4）教学效果得不到有效保证。线上教学过程中，学生缺乏教练的现场指导，教练也不能及时发现学生的错误动作并及时做出指导，学生容易走弯路。同时，课后反馈效果差，授课教师布置了任务，很少有人能够完成，这样的课后反馈是无效的，教练也不能通过学生的反馈及时发现问题并作出调整。

（5）对学生视力影响大。线上教学应用电脑、手机等设备授课，对学生视力健康构成威胁。会对视力甚至身体健康造成不可小觑的负面影响。

（6）网络卡顿，短时间内集中大批用户上线，同时各教育 app 占据了手机内存，许多性能不足的智能手机就会出现切换不及时，掉线等现象。尤其是教师端，如果网络不稳定，出现掉线，或教师操作不当，导致没有声音，整个课堂都会受影响。

（7）课堂气氛淡。没有集体学习的环境，许多相互促进的法子便不太灵，没有眼见邻桌认真学习，便也没了榜样，不能相互讨论，许多方法只能靠自己去悟。

（8）互动不便。由于云课堂本身的技术限制，暂时没有开发出提交图片、音频和视频的入口，只能通过第三方软件实现，增加了学生操作次数，降低了课堂效率。答疑不及时，听课有疑惑，只能通过线上交流的方式，信息一来一往，远不如面对面讲解。

（9）无法全方位的展示动作。相比线下教学，线上只能依靠教师的摄像头看到一个角度的动作，再加上沟通困难，学生无法看到立体的动作，增加了学习难度。

六、在线健身平台使用及付费市场现状分析

（一）在线健身平台使用现状分析

1.在线健身用户持续增多，用户粘性不断增强

伴随着当今中国经济、科技、教育、文化的日新月异，人们对身体体质、精神状态、身心健康、物质文化水平等生活质量的追求也逐步提高。因此，在线健身用户的持续增多已成为在线健身平台使用情况的一个趋势。

在体育训练方面，在线健身 APP 的推出将主攻爱好健身，缺乏体育锻炼，兴趣瑜伽、舞蹈的人群。“全民健身”已成为了当今国民所呼吁的口号之一。这便涌现了大批量的运动以及健身爱好者。许多在线健身 APP 通过仿用人体工学原理，贴近人体运动路径，降低了运动室的器材运动设备、教练人员的成本。通过手机或电脑的处理，在用户练习的过程中不断显示用户的动作状态，同时显示出需要纠正的部位，给出纠正意见。这样可以达到在家里也能学习的目的。这也能反应出用户对在线健身产品的用户粘性不断增强。

开发庞大而充满活力的用户群及活跃社区的能力是中国线上健身市场的主要进入壁垒之一。因共同话题聚集的用户之间基于社区专业内容的互动可以产生巨大影响力，能够提升品牌亲和力并有助于吸引和留住更多用户。若线上健身社区拥有大量活跃用户，则其表现可超越同行。市场参与者可通过适当激发用户在线上健身平台的热情并结合优质内容，逐步培育充满活力的社区文化，从而提高用户忠诚度，吸引更多潜在用户，进而形成正面反馈循环。

国家体育局倡导居家科学健身，各健身企业线上直播课程开启，千万家庭开始参与云健身。可以预见的是，随着疫情的影响、人们健康意识的增强以及科学技术的不断完善与创新，未来我国在线体育类健身APP用户规模还将持续增长。

2.在线健身行业链正处在逐步完善的阶段

目前的在线健身行业链正处在逐步完善的阶段，分工不断细化。根据《2018年健身行业白皮书》，健身行业中游主要包括线下健身中心、020 服务平台和健

身 APP，覆盖线上线下健身场景，为健身群体提供多元化服务。从健身行业的中游来看，我国在线健身产品竞争较激烈，利用公域流量和私域流量进行推广，可以快速提高健身参与人群数量，同时打破健身时间，更灵活便捷，我们值得思考的是其盈利模式的优化。

3.在线健身平台“长尾效应”较为突出

在市场博弈中，优势企业往往拥有较大的市场份额，而当市场份额达到某个阈值时，就可能出现“顶端聚集”现象。作为一个新兴行业，以“互联网＋健身”的在线健身行业正经历着相似的发展历程。对各平台排名前五的健身应用下载量进行统计，统计其占平台内全部健身应用下载量的比例（数据来源于 2020 艾媒咨询）：Googleplay 为 56.3％，腾讯应用宝为 65.4％，360 手机助手为 55.8％；百度手机助手为 45.7％。说明健身应用的“长尾效应”现象突出。究其原因，由于用户在有限时间里没有能力和精力去比较优劣、分辨好坏，而宁愿相信和默选大众化选项，这种基于羊群效应的取舍方式，加剧了“长尾效应”的发生。处于“长尾”中末端的应用，需要在这样的困局中总结经验教训，去揭示和把握更真实的用户需求，心存追求极致产品和服务的愿望，打造出更完美的产品与更完善的服务。

4.后疫情时代下，在线健身平台使用用户的运动参与度更高

疫情提升居民健康生活意识，云健身快速兴起，驱动了在线运动品类发展。从长远来看，经过此次疫情，大众的体育消费观念还会增强，我们国家的公共管理体制将更加完善，在线健身行业前景较好，近年来快速发展的势头不会逆转。疫情居家、线上健身直播与动作矫正成为健身爱好者的头号选择。根据第五章对在线健身平台的描述性分析，我们可以看到，在线运动健身行业的发展已到了成熟阶段，需求趋于饱和。后疫情时代下，户外场地与线下运动健身场所的安全性相较以前偏低，以前受限制较高。不仅如此，疫情唤醒了大众健身意识，许多人对自身的健康满意度下降，预示着运动健康行业的市场规模仍有比较大的发展空间。疫情的出现，让“云健身”成为很多中国人的体育必修课。让全民健身链接云、拥抱云，是不少健身企业和体育从业者的选择。疫情期间，健身类软件用户暴涨；不少健身机构推出线上服务；健身教练在直播平台当起了“网红”；很多退役和现

役体育明星也登上“云端”，录制视频指导大众健身。疫情可以加速体育健身休闲产业的互联网基因植入，基于线下场景的健身休闲产业将加强线上教育和培训，多元化的经营模式正逐步呈现。

5.在线健身平台使用走向智慧化、场景化和社区化。

在线健身平台使用智慧化的分析：随着家庭健身的崛起，用户对于在线健身 APP 的内容与服务体验要求越来越高，目前智能设备可通过互联网互动建立运动社区、提供制定优质内容等，粘合用户，搭建健身应用场景，实现用户从设备到场景端的全闭合体育生态。未来在民众健身新需求下，产品如何创新，解决痛点，一站式满足用户新诉求是运动健身企业面对的挑战。

在线健身平台使用场景化的分析：随着移动互联网、数据及智能技术的发展，运动政策的支持，运动健身与民众生活越来越紧密，运动健身不仅仅局限于运动场，更拓宽于家庭、办公、室外、户外等更多空间中，让运动健身成为一种新的生活方式。

在线健身平台使用社区化的分析：对于有健身需求的人来说，居家健身可能是一种更好的选择，健身行业的转型应该是在器材的互动属性上，引入社区的概念，提高互动性、娱乐性。

6.子领域仍有较大发展空间

线上健身平台可以利用公域流量和私域流量进行推广，可以快速提高健身参与人群数量，同时打破健身时间，更灵活便捷，但盈利模式仍待探索。以“社交+健身”的模式积累用户，再通过销售硬件和付费增值服务来商业化的模式，却并没有逃离“增收不增利”的怪圈。看似也符合“烧钱换增长”、“战略亏损换成长空间”这一互联网企业的经典方式，但这需要的是一个合适的商业模式。以 Keep 等在线健身产品为例，它们的问题就出在这里，其商业化的业务远没有达到预期，它们需要更多地考虑 ROI。

7.在线健身平台优质创新内容能力有待提高

全面互动内容一直且预计将仍是线上健身社区开发的关键要素。内容包括了

多种旨在提高中枢肌肉力量、减肥、康复的线上培训课程，且应当具有多维性、专业性及系统性。能够不断制作并升级各类互动健身内容的市场参与者能够更好地吸引并获取用户。

8. 线上健身产品供应及品牌知名度方面的问题

线上平台很难建立品牌知名度、信任度以及与用户之间深度连结。这将需要公司在内容与产品供应以及迭代、深入的客户洞察、差异化价值主张和战略品牌定位方面持续保持卓越。成为中国线上健身市场，尤其是在越来越注重个人健康保健的年轻一代中的领先知名品牌的能力，将带来强大而持久的品牌价值，可支持市场参与者的长期增长。

为确保线上健身社区的稳定发展，线上平台提供种类繁多的健身相关产品以满足用户的各种需求，这也是进一步增加公司收入的一种方式。尤其是在各类产品的辅助下，用户不仅注重健身活动的参与，而且还重视锻炼的效果。为支持广泛的产品供应，建立强大的供应链系统以简化众多生产线、组织库存单位并以有效的方式履行管理职能及提供物流服务同样至关重要。

（二）在线健身产品营销市场现状

1.在线健身逐步成为后疫情时代下的常态化健身方式

线下健身俱乐部作为新时代的产品，虽然其带来了健康科学的运动方式和丰富多彩的生活方式，但其新的问题也层出不穷，如何提高运动过程中学习的姿势纠正效率，通过团队的走访调查，健身房俱乐部的教练普遍反馈学员在学习过程中学习新动作的效率及其低下，需要长时间的复习巩固中进行动作的多次纠正；相应地，大部分学生也对课程成本高、课时减少、缺乏实际运动知识等问题感到不满。自 2020 年的新冠肺炎疫情爆发之后，全国人民经历了长时间的居家抗疫。我国对新冠肺炎疫情制度的处理与抗击彰显出了社会主义制度的优越性。我国对新冠疫情的防控独有的功能，满足了群众的体育活动需求，起到了舒缓紧张、焦虑的情绪作用，成为众人瞩目的“运动处方”。在疫情的影响下，后疫情时代势必会催生一波全民健身新高潮。但疫情何时结束犹未可知，疫情结束后，人际交往

也会保持一定空间上的距离。这在一定程度上引起了群众体育活动驱动机制的变化，促使群众体育由组织化向碎片化转变。目前的大背景下，人民群众不运动或不坚持运动的理由主要归结于时间与运动场地的限制，而在线健身的场景化可以充分利用比较方便的场地与碎片化时间进行体育锻炼，具有简单、经济、省时、高效的优势。与此同时，疫情下体育活动空间的狭小使得群众对健康的需求越来越大，在线健身作为当前状态下一种有效可行的手段，正成为越来越多人的首选方法，它的接受程度正在迅速提高，人们对它的认知正在从最初的兴趣转向需求。有不少专家和业内人士指出："群众在防疫期间养成的运动习惯会有一定保留，定期运动有望成为常态。"此外，受疫情影响，在线健身产品的普及率有所增加，包括健身房、健身教练和健身内容提供商在内的供应方更多地参与到了在线模式。因此，在后疫情时代中国体育健身产业整体结构面临线上与线下相结合的调整，这在一定程度上推动了在线健身的常态化发展。

2.线上与线下融合的健身将成为健身行业营销的基本模式

由于新冠疫情的影响，大批健身企业转移到线上，线上教学虽然有不能面对面沟通、反馈不及时等缺点，但是它在引流、聚人、降低出差成本等方面有得天独厚的优势。随着"互联网＋"、5G 技术的快速发展、元宇宙概念的推出以及国家的政策推动，线上与线下融合的健身已经是大势所趋。线下受地域影响，主要靠高校促销、口口相传、排放传单等地推的形式。获客，而线上面向全国，其潜在客户将突破地域限制无限扩大。健身企业需要粉丝，互联网平台需要内容，两者正好形成互补共赢关系。一方面，通过在线内容的输出（独立输出或合作输出），健身机构有机会建立更大的用户覆盖面和影响力，并为课程预订提供在线渠道，从而更合理地利用线下空间资源；另一方面，由于健身房的物理空间和用户个人时间的限制，线下健身用户每周在健身房健身的时间并不充裕。通过在线互动方式，能有效增加健身用户与健身房的接触时间、内容和服务。且线上＋线下全场景健身模式可以有效地巩固用户的健身习惯，增强品牌与用户之间的黏性，增加潜在客户套现的可能性。因此，线上线下融合健身将成为行业运营的基本模式，且线上用户将成为线下客户获取的重要渠道。

（三）在线健身产品潜在用户挖掘方式

近年来，在线健身一直在寻找变现方式。然而，在疫情之前，很少有详细的在线用户关注在线健身空间中课程内容系统的质量。运动健身类在线移动 APP 盈利模式主要依赖广告投放，变现效果不理想。随着在线用户的激增以及半专业和专业用户的增长，现有的在线内容已无法满足用户对质量、数量和“多样化”日益增长的需求。此外，我们对优质健身内容和服务的识别能力不断提高，在线健身内容供应商评级更加客观，优质健身内容和服务提供商更加突出。Keep 等在线健身平台一直从专业化、多元化的角度积极发力。课程按难度分为零基础阶段——初级阶段——高级阶段——强化阶段——挑战阶段。用户的健身设置进一步细化，对课程内容进行分层和推进，同时推进单一课程内容迭代功能，为用户提供持续的健身新鲜感。此外，在疫情期间，Keep 推出了家庭健身创意大赛等挑战。因此，在后疫情时代，我们可以看到，随着在线健身的不断发展，在线健身课程的内容创新已经成为一项核心竞争力，增加线上课程内容的创新与趣味性势必摆在各公司产品层面的一大问题。

（四）基于描述性统计的在线健身平台使用及付费市场现状分析

（1）平行坐标图

平行坐标图(Parallel coordinates plot)用于多元数据的可视化，它将高维数据的各个属性(变量)用一系列相互平行的坐标轴表示。纵向表示属性值，横向表示属性类别，如图 14 所示。

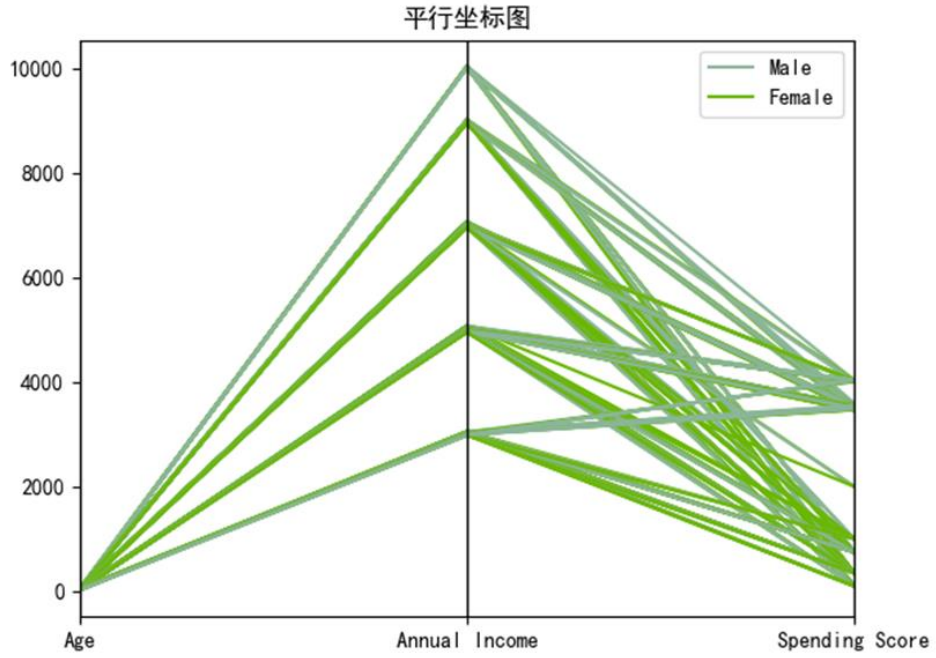


图 14　平行坐标图

（2）年龄、月收入、消费的分布

本文使用了直方图和核密度图，如图 15。（注：核密度图看的是(x<X)的面积，而不是高度）

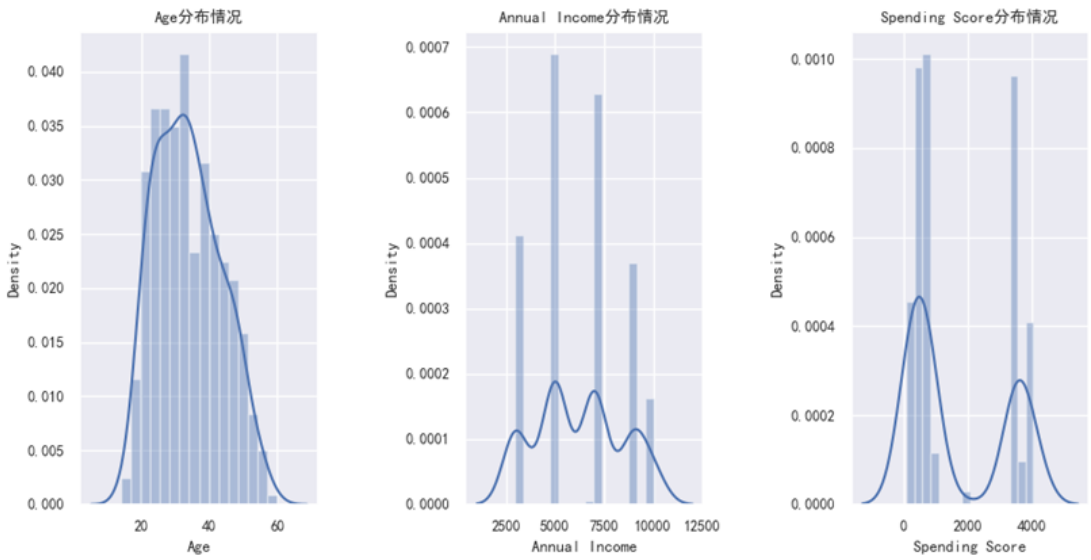


图 15　年龄、月收入、消费的直方图和核密度图

如图 15 所示，从左到右分别为年龄、月收入和消费的分布情况。我们可以发现，在年龄方面，20 岁到 40 岁范围的客户是最多的，另外，在 40 岁至 50 岁也不少，但是 60 岁以上的老年人是最不常来消费的，整体呈现偏态分布。在月收入方面，大部分的客户的月收入集中在 4000 至 6000 元范围里，月收入在 10000 元以上的很少。在消费方面，大部分客户的月消费大多在 2000 元以下，符合我们的线下采访情况。

（3）年龄分布柱状图

这里使用的是柱状图，和直方图不同的是，轴上的每一个刻度对应的是一个离散点，而不是一个区间。

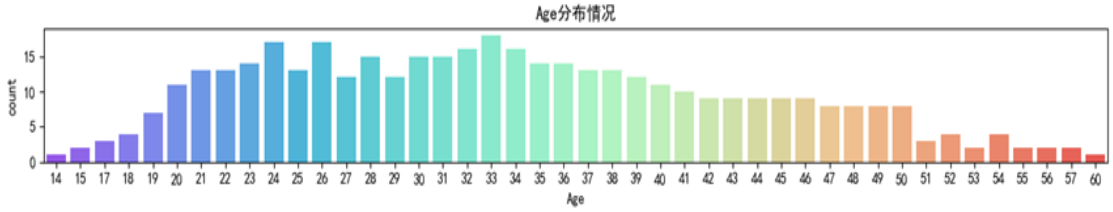


图 16　年龄分布柱状图

如图 16，这张图上可以看到调查样本的年龄分布情况。我们可以发现，年龄方面 27 岁至 40 岁范围的客户居多。其中，24、26、33 岁的客户是健身的常客，50 岁以上的用户却很少。总的来说，年龄较大的人群较少，年龄偏小的人群较多。考虑到年龄较大的健身群体对于手机的依赖程度、问卷调查配合程度与年轻群体不同，导致接收并填写问卷的概率不同，所以也可能导致本研究产生年龄分布的差异。

（4）不同性别用户占比

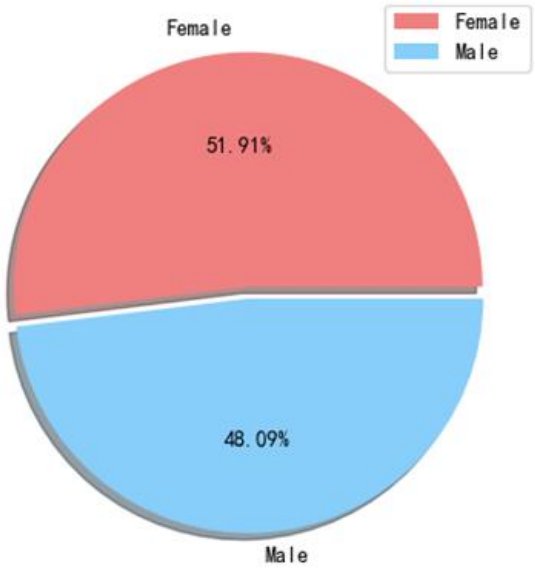


图 17　不同性别用户占比

如图 17 的饼图。可以发现，女性以 51.91％的份额居于领先地位，而男性则占整体的 48.09％。特别是当男性人口相对高于女性时，这是一个比较大的差距，这说明女性相比男性更热衷于使用在线健身相关的产品，这里印证了为何“姐妹”在词频统计中排名靠前。

（5） 两两特征之间的关系

Pairplot 主要展现的是属性（变量）两两之间的关系（线性或非线性，有无较为明显的相关关系）。本文以性别为变量作图，结果如下图 18 所示：

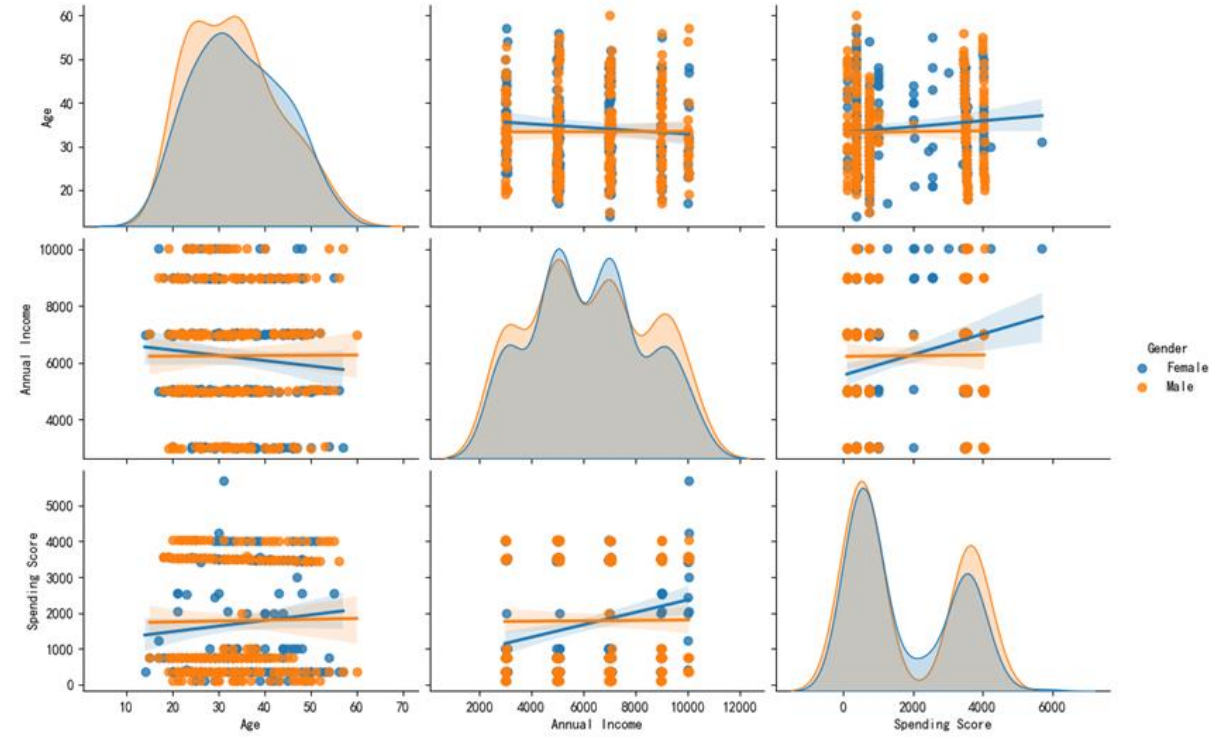


图 18　两两特征关系图

对角线上的图是各个属性的核密度分布图，非对角线的图是两个不同属性之间的相关图。我们可以看出，月均收入和消费能力之间有较为明显的相关关系。将 kind 参数设置为 reg 会为非对角线上的散点图拟合出一条回归直线，更直观地显示变量之间的关系。

（6） 两两特征之间的分布

如图 19 使用了增强箱图，适用于大数据中。我们可以通过绘制更多的分位数来提供数据分布的信息，能够看出男性的消费能力略低于女性，一定程度上说明了女性对在线健身的热衷程度要高于男性。

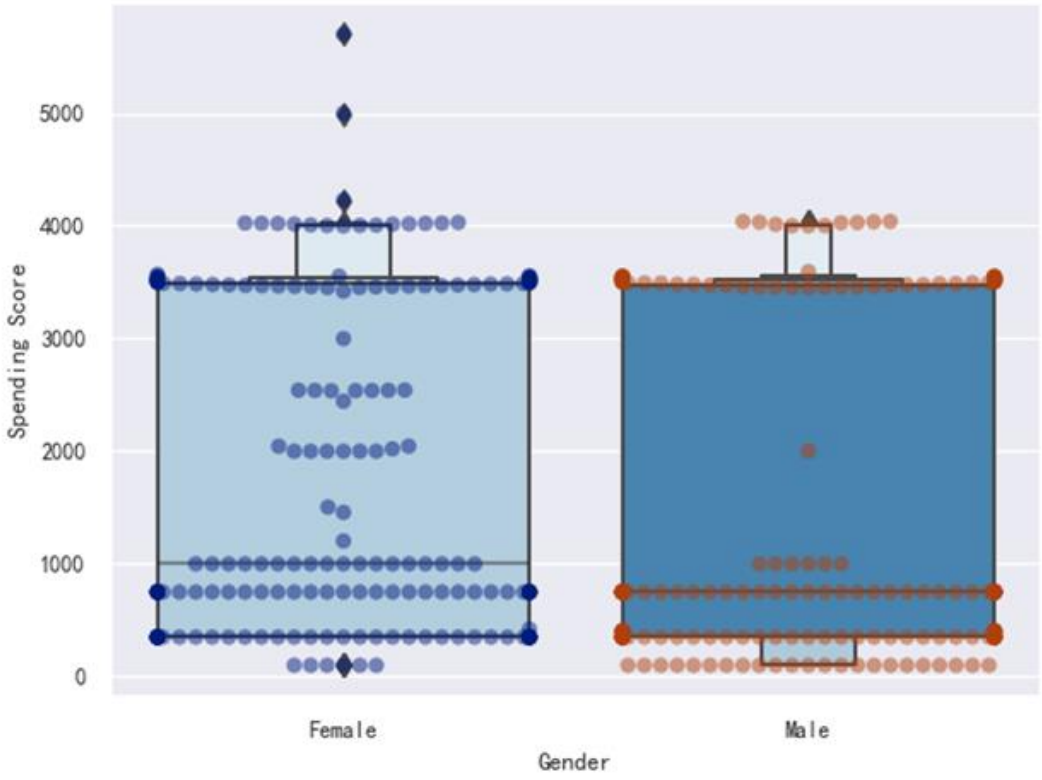


图 19　男、女性消费能力比较图

（7）可视化舆情分析

我们利用 Python 软件爬取 bing 浏览器的相关主页，图 20 和图 21 分别是绘制 “线上健身”、“线下健身”等相关关键词的词云图。

图 20　线上健身词云图

图 21　线下健身词云图

七、在线健身平台使用及付费影响因素分析

（一）基于决策树模型的在线健身平台用户粘性影响因素分析

为了了解不同因素对用户使用在线健身平台意愿的影响因素情况，通过调查各个影响因素，我们运用决策树的分类树直观地表现用户对于在线健身平台的使用意愿所受到的影响因素以及影响程度。

1、模型简介

分类决策树模型是一种描述对实例进行分类的树形结构。决策树由结点和有向边组成。结点有两种类型：内部结点和叶结点。内部结点表示一个特征或属性，叶结点表示一个类。用决策树分类，从根结点开始，对实例的某一特征进行测试，根据测试结果，将实例分配到其子结点；这时，每一个子结点对应着该特征的一个取值。如此递归地对实例进行测试并分配，直至达到叶结点。最后将实例分到叶结点的类中。

本文使用的是 CART 算法分类树模型，CART 是在给定输入随机变量 X 条件下输出随机变量 Y 的条件概率分布的学习方法。CART 假设决策树是二叉树，内部结点特征的取值为“是”和“否”，左分支是取值为“是”的分支，右分支是取值为“否”的分支。这样的决策树等价于递归地二分每个特征，将输入空 间即特征空间划分为有限个单元，并在这些单元上确定预测的概率分布，也就是在输入给定的条件下输出的条件概率分布。

决策树的生成就是递归地构建二叉决策树的过程。对回归树用平方误差最小化准则， 对分类树用基尼指数（Gini index）最小化准则，进行特征选择，生成二叉树。分类问题中，假设有 K 个类，样本点属于第 k 类的概率为p_k，则概率分布的基尼指数定义为:

$$Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

对于二类分类问题，若样本点属于第 1 个类的概率是p，则概率分布的基尼指数为:

$$Gini(p) = 2p(1-p)$$

对于给定的样本集合 D，其基尼指数为:

$$Gini(D) = 1 - \sum_{k=1}^{K} \left(\frac{|C_k|}{|D|} \right)^2$$

这里，C_k是 D 中属于第 k 类的样本子集，K 是类的个数。

如果样本集合 D 根据特征 A 是否取某一可能值 a 被分割成D_1和D_2两部分，即:

$$D_1 = \{(x,y) \epsilon D | A(x) = a\}, \qquad D_2 = D - D_1$$

则在特征 A 的条件下，集合 D 的基尼指数定义为:

$$Gini(D,A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

基尼指数 Gini(D)表示集合 D 的不确定性，基尼指数 Gini(D,A)表示经 A＝a 分割后集合 D 的不确定性。基尼指数值越大，样本集合的不确定性也就越大，这一点与熵相似。

CART 分类树的生产：

输入：训练数据集 D，停止计算的条件； 输出：CART 决策树。

根据训练数据集，从根结点开始，递归地对每个结点进行以下操作，构建二叉决策树：

（1）设结点的训练数据集为 D，计算现有特征对该数据集的基尼指数。此时，对每一个特征 A，对其可能取的每个值 a，根据样本点对 A＝a 的测试为“是”或“否”将 D 分割成 D1 和 D2 两部分，利用公式计算 A＝a 时的基尼指数。
（2）在所有可能的特征 A 以及它们所有可能的切分点 a 中，选择基尼指数最小的特征及其对应的切分点作为最优特征与最优切分点。依最优特征与最优切

分点，从现结点生成两个子结点，将训练数据集依特征分配到两个子结点中去。（3）对两个子结点递归地调用（1），（2），直至满足停止条件。（4）生成 CART 决策树。

算法停止计算的条件是结点中的样本个数小于预定阈值，或样本集的基尼指数小于预定阈值（样本基本属于同一类），或者没有更多特征。

2、模型建立

本文将针对研究“用户是否使用在线健身平台”这个输出结果构建决策树，我们按如下表格确立变量：

表 12　决策树I变量符号

变量	变量符号	变量定义
是否使用线上健身平台	Y	1，是
		2，否
性别	X[0]	1，男
		2，女
年龄	X[1]	1，20 岁以下
		2，21-30 岁
		3，31-40 岁
		4，41-50 岁
		5，51 岁以上
职业	X[2]	1，健身行业从业者
		2，在校学生
		3，政府/机关干部/公务员

		4，企业管理者
		5，普通职员
		6，专业人员（如医生/律师/文体/记者/老师等）
		7，普通工人（如工厂工人/体力劳动者等）
		8，商业服务业职工（如销售人员/商店职员/服务员等）
		9，个体经营者/承包商
		10，自由职业者
		11，农林牧渔劳动者
		12，退休
		13，暂无职业
		14，退休
最高学历	X[3]	1，小学及以下
		2，初中
		3，高中
		4，大学本科
		5， 大学专科

如表 12 所示，我们将“是否使用线上健身平台”作为最终的分类标签，将年龄，性别，职业和最高学历作为影响因素变量来进行第一棵决策树的构造与建模。

3、结果分析

变量确定完毕后，我们对于我们搜集的问卷数据进行了 CART 决策树的分类并进行了可视化，我们得出第一棵决策树图为：

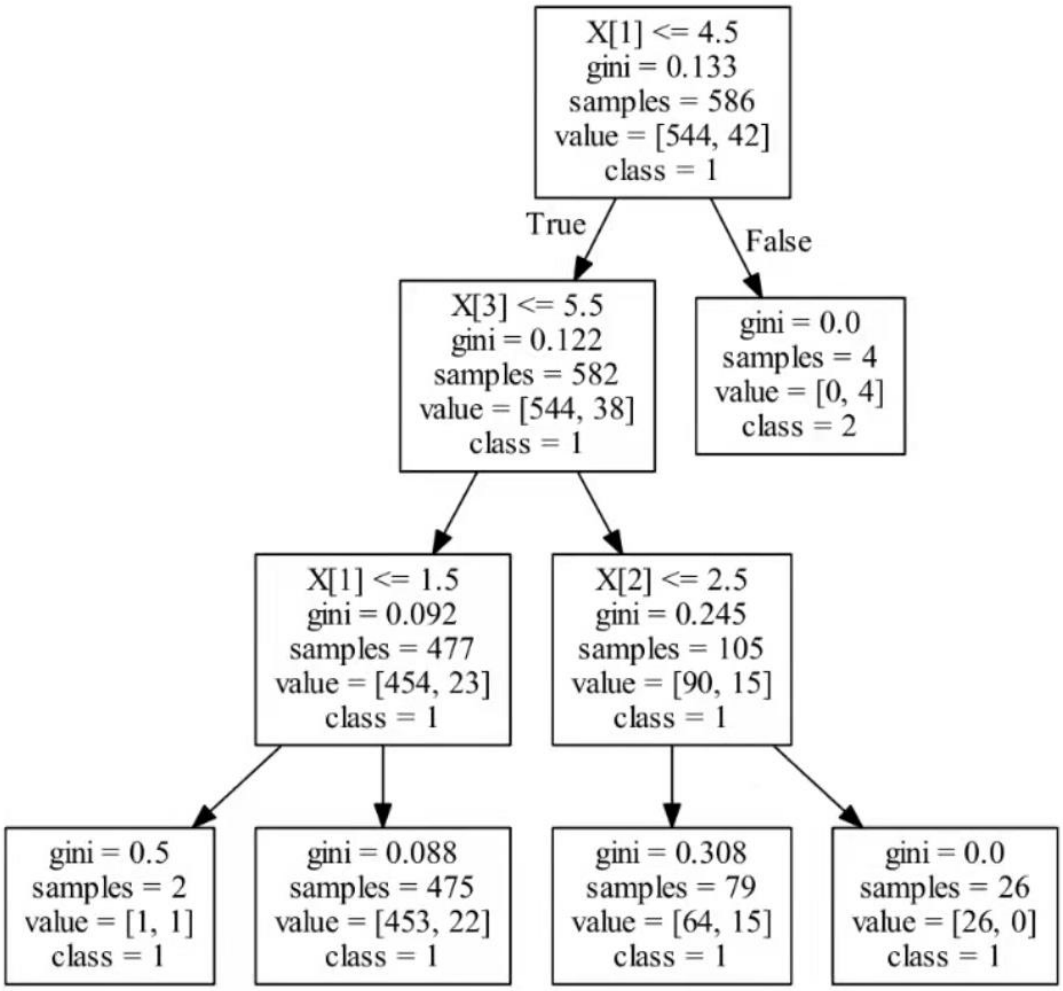


图 22　决策树图I

从这棵决策树的可视化结果可以发现，对于用户是否使用在线健身平台这一结果的影响因素中用户的年龄最为重要的影响因素，年龄这一因素是首要影响因素，从图中可以看到它的基尼指数经计算是 0.133，说明通过年龄这个因素进行分段分类能够使分类后的样本的异质性达到最低，能使分类最有效，也即说明年龄是对结果产生影响最大的一个影响因素。

其次是用户的学历和职业这两个影响因素，它们也分别对样本分类产生了影响；可以发现，性别这一因素对是否使用在线健身平台的分类几乎不产生影响。

（二）基于决策树模型的在线健身平台付费功能使用情况影响因素分析

为了了解不同因素对用户对于在线健身平台付费功能消费意愿的影响情况，通过调查各个影响因素，我们运用决策树的分类树直观地表现用户消费意愿受到的影响因素以及影响程度。

1、模型建立

本文将针对研究“用户是否在健身平台进行消费”这个输出结果构建再次决策树。第二棵决策树是针对“是否在线上健身平台使用付费”这一标签的影响因素的研究，我们按照以下表格来确定变量和建模：

表 13　决策树II变量符号

变量	变量符号	变量定义
是否使用线上健身平台付费功能	Y	1，是 2，否
健身数据记录和健康监测功能（量化为 1-6 个水平，数字越大代表功能越全面）	X[0]	1 2 3 4 5 6

系统稳定性（量化为 1-6 个水平，数字越大代表系统越稳定）	X[1]	1
		2
		3
		4
		5
		6
平台提供个性化定制服务水平（量化为 1-6 个水平，数字越大代表水平越高）	X[2]	1
		2
		3
		4
		5
		6
平台提供信息的及时性(量化为 1-6 个水平，数字越大代表信息越及时)	X[3]	1
		2
		3
		4
		5
		6
平台便捷程度(量化为1-6个水平，数字越大代表越便捷)	X[4]	1
		2
		3

		4
		5
平台提供信息的准确性（量化为 1-6 个水平，数字越大代表信息越准确)	X[5]	6
		1
		2
		3
		4
		5
		6

如上表 13 所示，我们将是否使用线上健身平台付费功能作为最终分类标签，将线上健身平台的健身数据记录和健康监测功能、系统稳定性、平台提供个性化定制服务水平、平台提供信息的及时性、平台便捷程度、平台提供信息的准确性这六个变量作为影响因素变量并进行量化。

2、结果分析

变量确定完毕后，我们对于我们搜集的问卷数据进行了 CART 决策树的分类并进行了可视化，我们得出第二棵决策树图为：

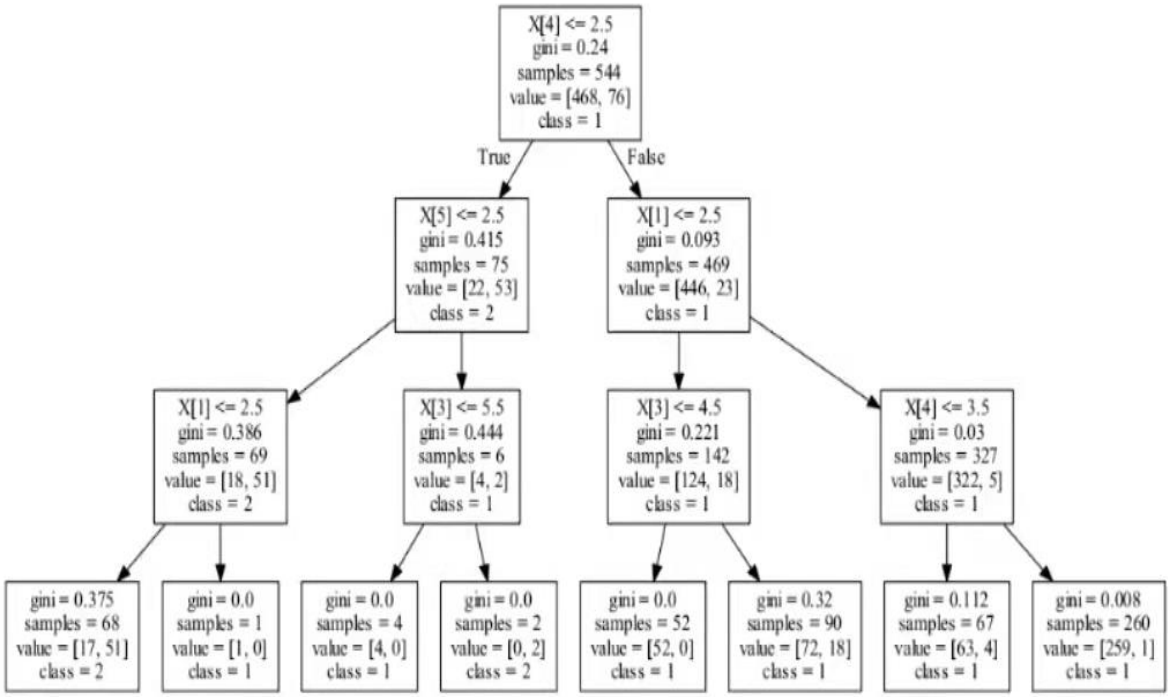


图 23　决策树图II

从第二棵决策树的可视化结果，我们可以发现促使用户在线上健身平台产生消费意愿的排序前三的重要影响因素为平台使用的便捷性，平台系统的稳定性以及平台提供信息的准确性。

而影响程度最不明显的为平台的健身数据记录和健康监测功能以及平台提供个性化定制服务水平。这在一定程度上说明在线健身平台的专业性和用户的基础体验对购买行为影响程度较大，良好的口碑、系统的稳定和便捷分别从心理、感官等方面影响着消费者的价值判断，而优良的质量更能提升产品价值在消费者内心的评估。

（三）基于 Pearson 相关分析的在线健身平台付费情况影响因素分析

相关分析是一种简单易行的测量定量数据之间的关系情况的分析方法，可以分析包括变量间的关系情况以及关系强弱程度等。相关分析，常用的方法类别有：

简单相关分析、偏相关分析、距离相关分析等。其中，Pearson 相关分析是简单相关分析最常用的方法之一，它能够直接计算两个变量的相关程度。

Pearson 相关分析用于研究定量数据之间的关系情况，以及判断他们的紧密程度情况。它能够对相关性以及显著性程度进行判断，是用来衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系。在本次的调研中也具有一定的意义。

本次调研中，我们拟讨论线上平台运动健身相关的消费总额是否与年龄、最高学历、疫情后使用频率、月收入、月消费有关。我们猜测：线上平台运动健身相关的消费总额与上述因素有关。

使用 Pearson 相关分析时，需要考虑 4 个假设。

假设 1：变量两两之间都是连续变量。

假设 2：变量两两之间应当是配对的，即来源于同一个个体。

假设 3：两个连续变量之间存在线性关系，通常做散点图检验该假设。

假设 4：变量两两之间均没有明显的异常值。Pearson 相关系数易受异常值影响。

假设 5：变量两两之间符合双变量正态分布。

接下来，根据 spss 软件，我们判断变量之间均不存在异常值。因此，我们进行下一步（假设 5），通过 Shapiro-Wilk 检验结果，我们可以得出检验数据均服从正态分布。最后，我们便可以通过 Pearson 相关分析表（表 14）进行分析，把使用线上平台运动健身相关的消费总额为因变量，分析其影响因素。结果表明，被调查者线上平台运动健身相关的消费总额与被调查者在线运动媒介平台的频率与有较大的相关关系，Pearson 相关系数为 0.857。

其次是被调查者的最高学历和被调查者月消费总额，Pearson 相关系数分别为 0.619 和 0.569。

表 14　Pearson 相关分析表

	2020 年 1 月 1 日之后，您于线上平台运动健身相关的消费总额？
您的年龄是？	0.541
您的最高学历（含目前在读）是？	0.619
2020 年 1 月 1 日后，您使用在线运动媒介平台的频率是？	0.857
您的月收入？	-0.017
您的月消费？	0.569

（四）基于结构方程模型的在线健身平台使用及付费影响因素分析

结构方程模型（SEM）是一种多元数据分析方法，其可用于研究多个潜变量之间的影响关系情况。它通过路径范式来描述自变量与因变量之间的关系并用线性方程式来表述自变量和因变量的数目。

结构方程模型共包括两部分结构，分别是测量关系和影响关系。如下图（结构方程模型图 24）所示，在本次调研中所采用的结构方程模型，包括了四个潜变量，分别是 Factor1 用户基础信息、Factor2 用户收支信息、Factor3 用户运动媒介使用频率和 Factor4 用户消费额。

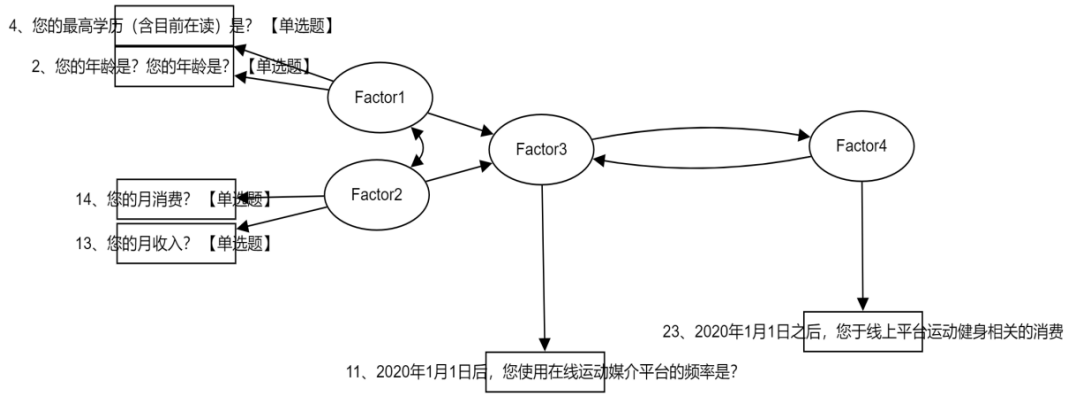


图 24　结构方程模型图

从测量关系来看：Factor1 用户基础信息由问卷中的第 2 题（最高学历）、第 4 题（年龄）共 2 项测量；Factor2 用户收支信息由问卷中的第 13 题（月收入）、第 14 题（月消费）共 2 项测量；Factor3 用户运动媒介使用频率由问卷中的第 11 题（平台频率）共 1 项测量；Factor4 用户消费额由问卷中的第 23 题（平台消费）共 1 项测量。

本次我们的现有在线健身样本量为 418 份，该样本量能够保证数据拟合效果呈现较佳的效果。同时在正式分析之前，需要保障测量关系具有良好的质量，通过探索性因子分析和验证性因子分析，共两步分析后，以保测量关系的高质量。

表 15　模型拟合指标表

指标	判断阈值	模型拟合值
GFI	>0.9	0.994
RMSEA	<0.10	0.044
RMR	<0.05	0.038
CFI	>0.9	0.956
NFI	>0.9	0.917
NNFI	>0.9	0.833

卡方自由度比 χ^2/df	<3	1.81

结果表示，最终得到的模型卡方自由度比为 1.81，GFI、RMSEA、RMR、CFI、NFI、NNFI、 χ^2/df 均满足良好模型的判断条件。除了以上重要指标，其余指标也均在标准范围内，因此说明模型构建良好，模型结果可靠。

我们还做出了模型系数估计结果表（表 15），它展示了测量关系情况。其中，标准化回归系数，它是在对自变量和因变量同时进行标准化处理后所得到的回归系数，数据经过标准化处理后消除了量纲、数量级等差异的影响，使得不同变量之间具有可比性，因此可以用标准化回归系数来比较不同自变量对因变量的作用大小。通常我们主要关注的是标准化回归系数的绝对值大小，绝对值越大，可认为它对因变量的影响就越大。

因此，我们可以看到，在 SEM 结构方程模型系数估计结果表中，四个潜变量以及其对应的相关因素均具有表现出较好的回归效果，证明本次选取的潜变量以及对应相关因素都能够使得模型拟合程度较良好。

表 16　模型系数估计结果表

X	→	Y	非标准化回归系数	SE	z (CR 值)	p	标准化回归系数
Factor1	→	Factor3	0.185	0.071	1.31	0.757	0.256
Factor2	→	Factor3	0.627	0.018	10.06	0.953	0.698
Factor3	→	Factor4	0.873	0.012	15.397	0.691	0.898
Factor4	→	Factor3	0.736	1.109	14.205	0.228	0.816
Factor1	→	您的最高学历（含目前在读）	0.012	1.107	0.646	0.518	0.087

Factor1	→	您的年龄	1	-	-	-	1
Factor2	→	您的月消费	0.636	0.609	11.236	0.952	0.723
Factor2	→	您的月收入	1	-	-	-	0.951
Factor3	→	2020 年 1 月 1 日后，您使用在线运动媒介平台的频率	1	-	-	-	0.907
Factor4	→	2020 年 1 月 1 日之后，您于线上平台运动健身相关的消费总额	1	-	-	-	0.973

备注：→表示回归影响关系或者测量关系

八、在线健身平台使用付费功能顾客特征分析

（一）基于 Logistic 回归的在线健身平台使用付费功能顾客特征分析

1、模型选择

二元选择模型一共分为线性概率模型、Probit 模型、Logit 模型，其中线性概率模型假定被解释变量与解释变最间是线性关系，采用普通最小二乘法(OLS)或加权最小二乘法(WLS)进行参数估计。尽管在实际应用中,对于同一资料用 Probit 回归和 Logistic 回归的结果非常接近，但本文采用的是应用更为广泛的 Logistic 回归。原因如下:

第一, Logistic 回归中的偏回归系数可以计算其 OR 值(Odds Ratio,优势比),具有更贴近实际的解释意义,而 Probit 回归中的偏回归系数含义为其他自变量取值保持不变时自变量每改变 1 个单位，出现阳性结果的概率密度函数值的改变量，这种解释远不如前者直观有用。

第二，Probit 回归是在正态分布的理论基础上进行的，Logistic 回归则是基于二项分布的理论。由于发放的调查问卷中分类变量较多，连续变量极少，因此我们认为采用 Logistic 回归更为合理。

2、模型建立

将 1 个个体“是否选择于线上平台使用过和运动健身相关的付费功能”这种二元选择行为表示为因变量，当个体选择“是”时，y取值为 1，当选择“否”时，y取值为 0。

在回归模型中，回归系数β表示其他自变量不变，x 每改变一个单位时，所预测的 y的平均变化量，当x为连续性变量时这样解释没有问题，二分类变量 由于只存在两个类别间的比较，也可以对系数得到很好的解释，由于本次设置 的问卷调查中大量数据为分类资料，例如群体的月收入分成了七档，如果直接编码为 1、 2、3、4、5、6、7 令其作为自变量纳入分析，就等价于是假设这七档间

的差距完全相等，或者说他们对因变量的数值影响程度是均匀的，这样的假设会过于简单武断、与实际情况不符。另外对于无序多分类变量，如学历，它们之间不存在数量上的高低，因此不可能为其给出一个单独的回归系数估计值，来表示学历上升一个单位时因变量的变化趋势。此时需要使用哑变量对模型加以定义。

3、对用户群体的研究

在对哑变量进行编码时，采取了指示对比方法，并将最后一个选项设置为参照水平，分类变量编码方式见附录 。纳入所有需要考虑的变量，把最终选择作为阳性结果，即p表示选择是的概率，β表示某一变量改变一个单位时，选择是与否的概率之比的对数变化值，x表示变量取值，建立模型如下：

$$\ln\frac{P_i}{1-P_i}=\sum\beta_i X_i(i=1,2,3\cdots)$$

选用输入方法，显著性水平给定为 0.05，得到的回归结果经过筛选后如表17：

表 17　回归结果表

变量	B	标准误差	瓦尔德	显著性	EXP(B)
Age_20	1.692	0.463	12.710	0.000	5.430
Age21_30	-0.562	0.467	3.871	0.294	0.570
Age31_40	1.441	0.413	10.966	0.000	4.225
Age41_50	1.538	0.393	15.187	0.005	4.655
Primary	-3.983	0.694	39.284	0.000	0.019
Junior	-3.169	0.591	41.761	0.000	0.042
High	-2.096	0.475	20.981	0.000	0.123
Undergraduate	-2.003	0.461	15.912	0.000	0.135
College	-1.063	0.399	11.046	0.011	0.346

Income1	3.651	0.529	41.251	0.000	38.513
Income2	3.084	0.515	19.285	0.000	21.846
Income3	2.549	0.614	18.700	0.000	12.794
Income4	2.097	0.483	21.762	0.000	8.142
Income5	1.692	0.462	15.962	0.000	5.430
Income6	0.904	0.424	3.429	0.201	2.469

4、模型结果分析

$EXP(B)$表示其他变量保持不变，自变量变化一个单位，变化后的选择线上平台使用运动健身相关的付费功能与不选择线上平台使用运动健身相关的付费功能人数之比是变化前的选择线上平台使用运动健身相关的付费功能与不选择线上平台使用运动健身相关的付费功能人数之比的$EXP(B)$倍。在结果中观察可以发现，性别、月消费、居住地的总的 P 值大于 0.05，超过了显著性水平，表明从总体上讲，这三个自变量应当对因变量无影响，此时所有的哑变量都不用再纳入分析了。

（1）年龄

模型结果显示，用户的年龄状况通过模型系数的显著性检验，这说明不同的年龄段影响用户选择线上平台使用运动健身相关的付费功能的意愿。从$EXP(B)$值来看，20 岁以下的年轻群体选择线上健身的积极性是基准类 50 岁以上的老龄人群的 5.430 倍。并且随着年龄的增长，选择在线上平台使用运动健身相关的付费功能的概率在逐渐降低。

（2）学历

学历的回归系数为负，且通过了显著性检验，说明具有硕士研究生及以上学历的群体对在线上平台使用运动健身相关的付费功能的选择意愿更强。从$EXP(B)$值来看，其他学历群体的选择意愿都低于硕士研究生及以上学历。

（3）月收入

月收入的回归系数为正，且通过了显著性检验，其中无收入（学生属于无收入）的群体相比与基准类 10000 元及以上人群，对在线上平台使用运动健身相关的付费功能的选择意愿更强。出现这个原因也许是因为调查对象大多数为学生，属于无收入群体。

（二）基于 K-means 聚类的在线健身平台使用付费功能顾客特征分析

对于已经搜集的用户基本数据，如客户 ID，年龄，性别，月均收入和健身消费。我们想对这些接触在线健身的用户进行聚类，分析经常使用在线健身的用户具有什么样的特征。

由于本次调查的消费者数据均采用离散的分类数据，因此对其进行群体特征分析时适合采用 K-means 聚类算法。本文选取健身的采访者中使用健身 APP 等在线方式的用户，并选取这些用户的性别、年龄、月均收入和在线健身的支出等维度，通过 python 编程语言实现 K-means 聚类算法对其进行特征分析，分类。K-means 算法首先随机选取 k 个点作为聚类中心，然后计算每个样本到聚类中心的汉明距离，将样本归到距离最近的那个点形成簇，然后根据簇内的样本重新形成聚类中心，重复这个过程，直到聚类中心不再改变。

1、K-means 聚类模型简介

以在线健身的特征分析为例。现有在线健身样本 418 份，每个样本具有性别、年龄、收入和在线健身的支出等 6 个属性且全是离散的，簇的个数为 k。

步骤一：随机确定 k 个聚类中心 C1,C2,C3……Ck，Ci 是长度为 6 的向量，Ci={Ci1 ,Ci2 ,Ci3…Ci6}，Ci 那么由计算机随机给出。

步骤二：对于样本 x（ii=1,2,3…537），分别比较其与 k 个中心之间的距离。

这里的距离为汉明距离，即不同属性值的个数，例如 x1={2,7,4,3,3,1}，C1={1,7,3,3,3,2}，那么 x1C1 之间的汉明距离就为 3。

步骤三：将 xi 划分到距离最小的簇，在全部的样本都被划分完毕之后，重新确定簇中心，向量 Ci 中的每一个分量都更新为簇 i 中的众数。

步骤四：重复步骤二和三，直到总距离（各个簇中样本与各自簇中心距离之和）不再降低，返回最后的聚类结果。

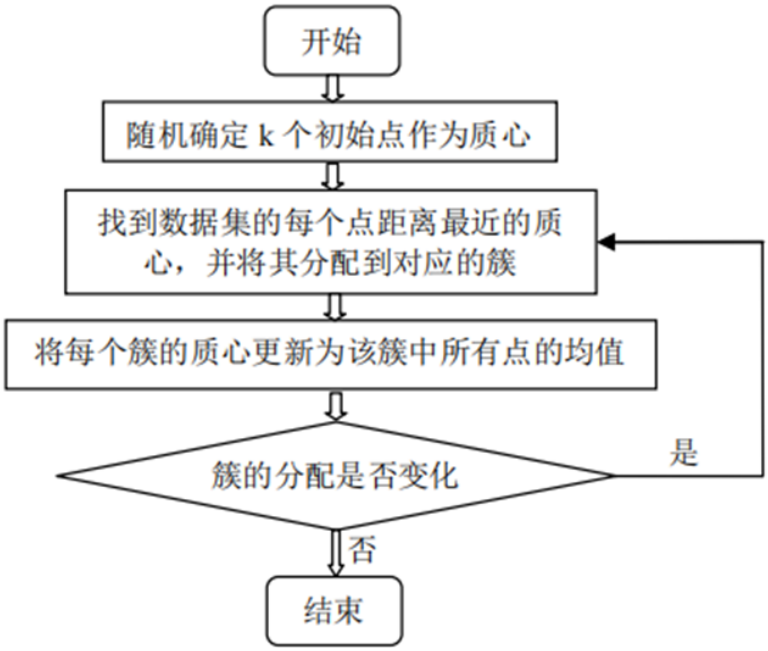


图 25　K-means 聚类算法步骤

2、“肘部法则”聚类分析

本文采用“肘部法则”来选择最优的簇数量，即 k 值。“肘部法则”通过比较不同 k 值的 SSE 值（成本函数值即各类别中样本点到中心点的距离之和）来进行选择。随着 k 值的增大，平均畸变程度会减小；每个簇包含的样本数会减少，于是样本离其重心会更近。但是，随着 k 值继续增大，平均畸变程度的改善效果会不断减低。k 值增大过程中，畸变程度的改善效果下降幅度最大的位置对应的 k 值就是肘部。由 python 绘制的 SSE 值与 k 值的函数图像。

（1）“肘部法则”简介

误差平方和(sum of the squared errors，SSE)是所有样本的聚类误差反映了聚类效果的好坏，公式如下：

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

其中，C_i 是第 i 个簇，p 是 C_i 中的数据点，m_i 是 C_i 的质心。

随着聚类数 k 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么 SSE 会逐渐变小。

当 k 小于真实聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，故 SSE 的下降幅度会很大。

当 k 到达真实聚类数时，再增加 k 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减。然后随着 k 值的继续增大而趋于平缓，也就是说 SSE 和 k 的关系图是一个手肘的形状，而这个肘部对应的 k 值就是数据的真实聚类数。

肘部法则 K-means 算法的目标函数是使得样本与质心的平方误差最小化，该算法通过计算各簇内的质点与簇内样本点的距离误差平方和，作为衡量分类效果的指标，称为畸变程度(distortions)。具体来看，如果簇内部的畸变程度较低，则认为簇内各数据点的距离较小，此时聚类效果也较好好，反之，则认为该簇内部各数据点相似较低，聚类效果较差。尽管畸变程度会受聚类数目 K 的值影响，通常 K 越大，畸变程度越小，但一般情况，对于参与聚类的可区分数据，在达到某个点时畸变程度会迅速降低，之后缓慢下降，这个临界点就可以判定为聚类性能较好的点。

（2）基于年龄和消费分数的聚类

所需要的数据有‘Age’和‘Spending Score’。使用手肘法确定最合适的 k 值，绘图如图 26 确定 k 值，这里将 k 确定为 4。

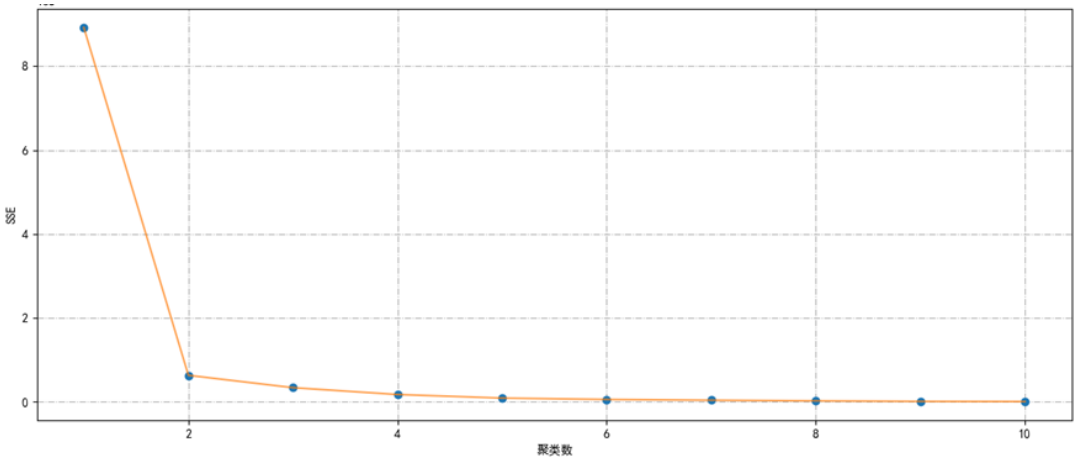


图 26　基于年龄和消费分数的手肘法图

确定 k＝4 后。重新构建 k=4 的 K-means 模型，并且绘制聚类图。效果如图 27，基于年龄和消费能力这两个参数，可以将用户划分成 4 类。

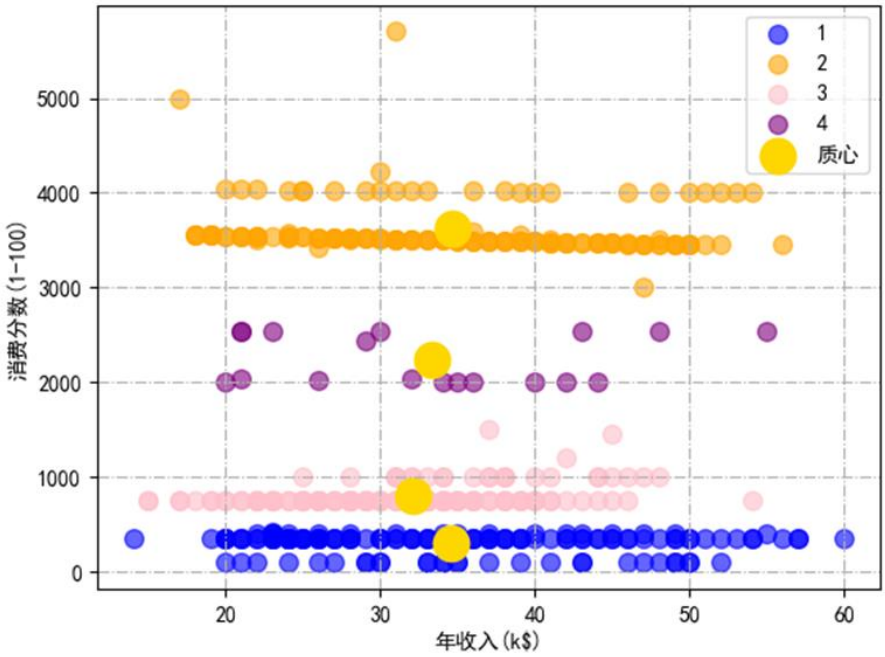


图 27　基于年龄和消费分数的聚类图（k=4）

（3）基于月收入和消费分数的聚类

同理，使用手肘法确定合适的 k 值，这里取 k=5。

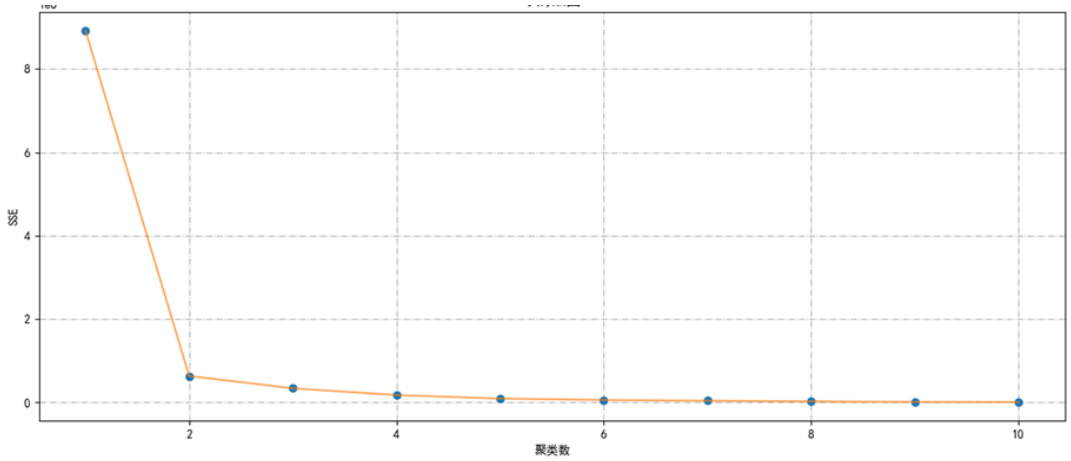


图 28　基于月收入和消费分数的手肘法图

效果如图 28，基于月收入和消费能力这两个参数，可以将用户划分成为 5 类，用户特征画像如图 30 所示共分为 5 个群体，分别为目标用户（这类客户月收入高，而且高消费）、普通用户（月收入与消费得分中等水平）、高消费用户（月收入水平较低，但是却有较强烈的消费意愿，舍得花钱）、节俭用户（月收入高但是消费意愿不强烈）、谨慎用户（月收入和消费意愿都较低）。

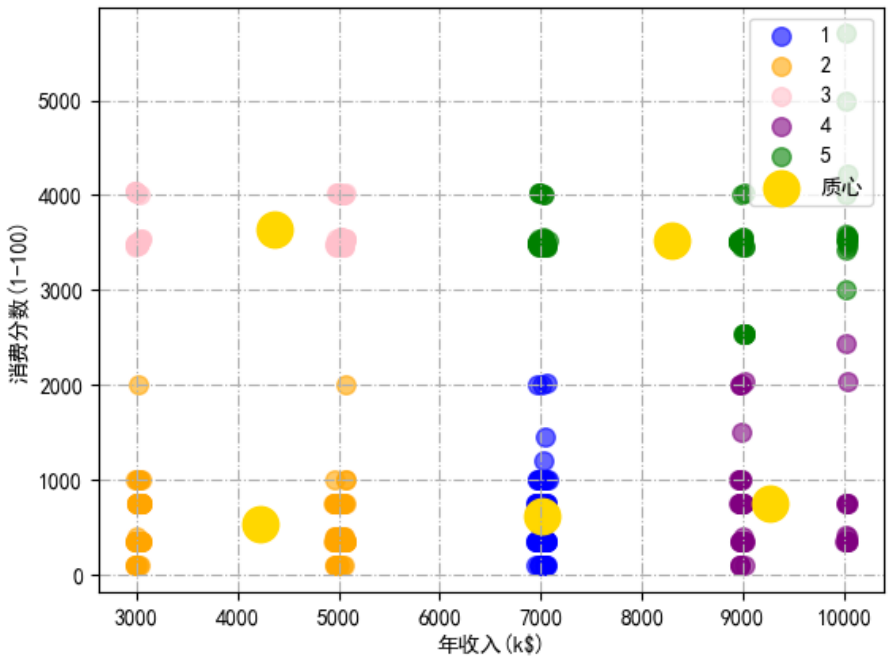


图 29　基于月收入和消费分数的聚类图（k=5）

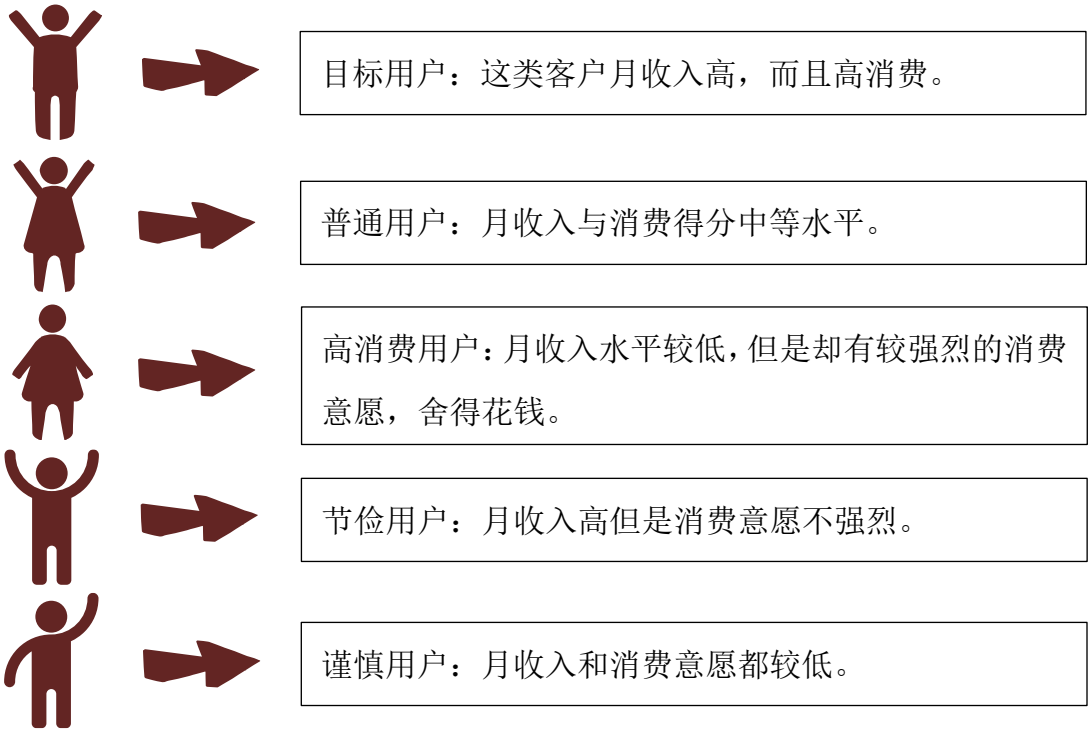

目标用户：这类客户月收入高，而且高消费。
普通用户：月收入与消费得分中等水平。
高消费用户：月收入水平较低，但是却有较强烈的消费意愿，舍得花钱。
节俭用户：月收入高但是消费意愿不强烈。
谨慎用户：月收入和消费意愿都较低。

图 30　用户特征画像图

九、结论与建议

（一）研究结论

1、对线上健身了解程度和接受程度方面的结论

（1）推广渠道与受众错位导致在线健身了解程度在不同类型人群中偏差较大。有 92.83%的调查对象使用过在线健身媒介或平台，总体了解程度较好，中青年人群、高学历人群以及城市居民对于在线健身较为了解， 但作为消费主体的中高收入群体对在线健身方面的消费意愿和平均消费额较低。基于上述结论，我们认为在线健身当前的推广宣传方式存在问题，从而导致传播面存在缺失。

（2）优质的在线健身服务接受程度更高

调查群体对优质的线上健身服务接受程度更高，数据显示有 76.64%的人愿意为优质的在线健身服务进行付费，对于线上健身平台的特色，消费者更倾向于平台的便捷程度、平台提供信息的准确性和及时性以及平台提供个性化定制服务；不可否认的是，有很大一部分人愿意尝试和选择，说明这些线上健身仍有发展的潜在可能，并不一定会被市场淘汰。

2、线上健身消费者特征方面的结论

从聚类算法分析的结果来看，线上健身的典型接受者为以下五类人群：

群体 1 ⇒ 目标用户：这类客户月收入高，而且高消费。

群体 2 ⇒ 普通用户：月收入与消费得分中等水平。

群体 3 ⇒ 高消费用户：月收入水平较低，但是却有较强烈的消费意愿，舍得花钱。

群体 4 ⇒ 节俭用户：月收入高但是消费意愿不强烈。

群体 5 ⇒ 谨慎用户：月收入和消费意愿都较低。

从二元选择的结果来看，当前不同年龄的消费群体对于线上健身的接受程度

是完全不同的，随着年龄的减小，接受程度越高。老年人相对条件更有限一些，受传统习俗的影响更深，较难接受和进行线上健身；相反，年轻人的教育程度、健身意识和对新事物的接受度都比老年人更高，对于新鲜事物的欲望更加强烈。

此外，高学历人群的接受程度比之低学历接受程度更高。更广阔的知识面、更高的健康意识、更自由开放的思想、更发达的信息传播环境和较高的对于新鲜事物的接受和探究能力都更有助于对线上健身的接受因此将线上健身需要推广到低学历人群和农村、郊区居民也是提高线上健身使用程度的有效途径之一。

3、线上健身消费者决策影响因素方面的结论

随着当前经济的发展与社会的变迁，大众体育商业化的健身行业作为领军者领跑前沿。伴随人们健康意识的提高与健身热度兴起，当代青年忙于工作与商务，挤出线下健身房的时间少之又少。这一系列背景便线上健身行业的健身行业发展提供了广阔的空间。通过对调查者们的访问调查，我们小组归纳出线上健身消费者决策主要影响因素包括如下：年龄、学历、月消费收入。其中，21-40 岁的青年与中年是线上健身媒介的受众与参与度最广泛的群体；较高学历的人更加注重对身材管理与身体健康的维护；月消费与月收入较高的人更加注重对身材管理与身体健康的维护。

4、线上健身营销策略方面的结论

通过上述在线健身行业营销市场的分析，我们可以了解到，线上与线下融合的健身将成为健身行业营销的基本模式。线上与线下全场景健身模式，能够有效地巩固用户的健身习惯，增强品牌与用户之间的黏性，增加潜在客户套现的可能性，增加用户对产品的认可度。把线下健身课程移植到线上，并加以创新与改良，为不方便去线下健身房锻炼的用户提供便利，提升健身房收益，然而，公司要想实现向现代企业转型，持续保持销量的快速增长和公司的健康稳步发展，进行品牌运作是当务之急，利用现有市场优势，进行品牌战略规划、品牌传播，使公司有效实现品牌和销量的提升，建立起阶段性产品策略也十分必要。各线上健身 APP 都可以阶段性产品策略分为市场投入——黄金竞争——成熟效益期。当达到实现资源最优化和效益最大化时，线上平台推广便可到全国各地，能够达到争

取线上健身市场高度占有率的目的。

（二）提出建议

1、线上健身宣传及推广方面的建议

（1）加大宣传辐射面，激活潜在消费群体

在推广线上健身时，宣传辐射面应该要更广，除了现有对线上健身较为了解的中青年人群、高收入人群以及城市人群外，线上健身的宣传应覆盖到中老年群体、较低收入和低收入人群以及农村、郊区居民。其中年和青年群体是未来将有可能购买线上健身服务的潜在消费者，同时低收入人群、农村、郊区居民也都是线上健身市场的潜在消费者，因此需要激活潜在消费者和主力消费者，就需要对这些人群进行线上健身的推广，在了解程度提升的同时，未来将会购买线上健身服务的潜在能力也会有所提升。

除此以外，在进行宣传时可以采取“渐进式吸引”的方式，逐步推进，引导其改变传统健身观念带来的束缚，愿意尝试线上健身，但也不能盲目引导和强加，真正的选择权还是应该留给消费者，因此“渐进式吸引”这一方式会温和的改善部分消费者的抵触心理。

（2）宣传推广线上健身形式时注意选取突出角度

在对线上健身进行宣传时，多对其形式新颖、内容创新、思想先进等创新角度进行渲染，新鲜事物的出现往往能够直击消费者猎奇心理，通过突出展示 创新点来吸引消费者视线，刺激购买欲望，唤醒购买意识。

另外也要突出线上健身方式的核心竞争力，线上健身相较于传统健身形式来说，方便快捷和贴合时代背景无疑是最大的亮点，对线上健身这两个方面进行大力宣传，突出优势，牢牢抓住消费者群体。

（3）对线上健身接受群体进行精准营销

在精准定位五类接受群体的基础上，对其制定专属营销策略，使其从高认可度群体转变为未来的实际消费者，并愿为线上健身进行宣传。保持平台和客户的

密切互动沟通，从而不断满足客户个性需求，建立稳定的忠实顾客群，实现客户链式反应增值，从而达到的长期稳定高速发展的需求。

2、线上健身发展趋势和产业结构方面的建议

（1）受疫情影响，居家健身人士增加，健身 APP、在线健身集中爆发。

据百度指数数据，2020 年起“郑多燕减肥操”、“健身”、“瑜伽”等关键词的搜索热度飙升，搜索热度平均上涨超 100%。据七麦数据显示，2020 年 1 月-4 月，健身 APP 的搜索指数由 5831 上升到了 6744。2020 年 2 月，运动健身 APP 行业用户月活 8928 万，同比+93.3%。大量线下健身用户大批量向线上健身转移，健身 APP、在线健身集中爆发。

（2）健全多元信息服务，提升用户“认知+思考”体验

一个健身 APP 课程、一个训练方案、一个简单的奖励是远远不够的，用户需要健身、健身附属产品、相关的运动讯息等多元体验。（1）在线健身需要对用户进行精准定位与智能分层，不断完善体育新闻、体育健身视频、体育技巧书籍、体育赛事通道、志愿者服务、健身分享社区、休闲养生、运动医护以及健身商店等综合信息服务的融合。（2）在线健身在各平台的评论是用户与健身品牌的重要桥梁，重视信息反馈的兑现与修订才是提升竞争力的重要途径，提升用户健身信息主动思考体验是关键，加大人工客服咨询、健身用户信息特征、界面功能的人性化设置和交互设计情感关怀等。（3）重视用户健身信息主动“思考”成果，增设用户新闻、用户设计、用户健身活动创设区，充分调动用户的主观能动性与创造性。

（3）发展智慧健身社交，提升用户“关联+情感”体验

在线健身的用户表现出强烈的运动社交意愿，所以我们应当营造“关联+情感”健身社交环境，从而改善用户体验。首先，在线健身平台需要加大与微信、微博、QQ、抖音、今日头条等社交媒体合作，全面连接网络用户，提升健身文化的影响力，让健身交流更便捷有效。其次，打造健身明星，积极参与社区话题和社区活动的引导，实现情感资本向健身动力的转化。并且，要打造健身社交模式，完

善线上讨论线下健身活动开展，线下装备体验线上购买，线上运动方案制定线下健身互动等，线上健身需加强与线下商店、运动跑团、塑身馆、健身广场等合作，通过线上平台的人群延伸到线下集体健身活动。

参考文献

[1] Buss D. M., Barnes M. F., Preferences in human male selection[J]. Journal of Personality andSocial Psychology, 1986, 50: 559-570.

[2] 艾冬梅,郝锋,曹青青,邹安弟,董瀚月,徐勰.健身气功-易筋经训练对大学生动静态平衡能力的影响[J].河南中医,2022,42(03):462-467.

[3] 樊梦吟.青年女性的身材焦虑:B M 风流行现象试解[J].山西青年职业学院学报,2021,34(03):23-27.

[4] 方子隽.在线健身中的社交仪式——以 b 站健身视频中的弹幕文本为例[J].文化产业,2021(17):149-150.

[5] 侯光定,孙民康,杨水金,余磊,李锐.后疫情时代下全民健身的功能、任务与路径研究[J].辽宁体育科技,2021,43(03):1-5.

[6] 黄开颜,李乃适.中国传统养生运动与糖尿病预防的初步探讨[J].中华健康管理学杂志,2022,16(03):196-198.

[7] 赖锐.身体美学与实践美学的对话[J].南昌大学学报(人文社会科学版),2022,53(01):115-124.

[8] 雷雪梅.健身运动对老年人认知功能的影响及其作用机制[J].体育科技文献通报,2022,30(02):149-151.

[9] 李冬梅.我国大众健身公共体育空间供给现实窘境与发展对策[J].沈阳体育学院学报,2022,41(01):83-89.

[10] 刘东锋,傅钢强.新型冠状病毒肺炎疫情期间在线健身服务用户持续使用意愿的影响因素研究[J].体育学研究,2020,34(02):41-50.

[11] 刘高福,李永华.用户互动对价值共创行为的影响研究——以线上健身社区为例[J].江西社会科学,2021,41(12):197-207+256.

[12] 刘红建,高奎亭,徐百超.中国全民健身政策体系演进历程、优势特征及效能转化研究[J].体育学研究,2022,36(01):91-102.

[13] 刘建武,钟丽萍,张凤彪.健身服务业线上线下融合发展的机遇、机理与路径[J].体育文化导刊,2021(09):86-92+104.

[14] 刘怡.规训与挣扎:社交媒体语境下的女性身材焦虑与自我管理[J].视听,2021(11):153-154.

[15] 陆佳莉.体育运动健身类 APP 的现状及对策研究[J].辽宁体育科技,2017,39(06):20-23.

[16] 牟粼琳,孙笑,沈克印,李红辛,李欣芮.区块链技术赋能体育健身产业的理论阐释、应用实例与推进策略——以上海角马私教 APP 为例[J].武汉体育学院学报,2021,55(07):72-79+87.

[17] 田春兰,张秋婷.全民健康背景下城市低收入群体健身问题研究[J].当代体育科技,2021,11(36):115-118.

[18] 田芊. 中国女性择偶倾向研究[D].复旦大学,2012.

[19] 王爽,张磊,张俊勇,王一乐.GIS 在全民健身中的应用特征研究[J].自然资源遥感,2021,33(04):265-271.

[20] 余玲,易国忠,夏君玫,李越,张伟伟.传统体育养生运动处方对轻度抑郁症女大学生情绪和睡眠质量的影响[J].四川体育科学,2021,40(06):31-34.

[21] 张鹏,杨涛,刘艳娜.新时代全民健身促进乡风文明建设的发展路径研究[J].沈阳体育学院学报,2021,40(05):76-81+89.

[22] 赵耀. 中国劳动力市场雇用岐视研究[D].首都经济贸易大学,2006.

[23] 赵玉婷. 男性媒介形象“阴柔化”与社会性别认同研究[D].吉林大学,2020.

[24] 曾繁荣.全民健身背景下大学生体育服务模式与运营[J].当代体育科技,2022,12(02):69-72+76.

[25] 钟丽萍,刘建武,范成文,周进.新冠肺炎疫情下在线健身的实践逻辑、发展态势与推进策略[J].武汉体育学院学报,2020,54(09):34-41.

[26] 钟丽萍,万佳慧,钟江南,刘建武,周进,龙寰宇,尹廷廷,方庆军.新冠疫情下的在线健身视频网站运营研究——以 bilibili 健身视频网站为例[J].体育研究与教育,2021,36(03):16-20+34.

[27] 周德书,黄元骋.习近平全民健身论述的逻辑内涵与时代特征[J/OL].广州体育学院学报,2022(01):30-42[2022-03-13].

[28] 周文彰,岳凤兰.新时代从四个层面创造美好生活[J].前进,2018(12):19-20.

附录 1 调查问卷

线上健身平台的使用情况及偏好调研

尊敬的先生/女士：

您好！

我们是来自 xx 大学的研究生，感谢您能抽出 2-5 分钟时间来参加本次调研。本次调研旨在了解用户对在线运动媒介平台的使用情况及偏好。

答案没有对错之分，本次调研采用匿名方式，严格遵守保密原则，所获资料仅供学术研究使用，您的意见对本次研究意义重大，恳请您根据自己的真实情况如实填写，感谢您的支持！

1、您的性别 【单选题】

○ 男

○ 女

○ 其他

2、您的年龄是？ 【单选题】

○ 20 岁以下

○ 20~30 岁

○ 31~40 岁

○ 41~50 岁

○ 51 岁及以上

3、您目前的职业是？ 【单选题】

- ○ 健身行业从业者
- ○ 在校学生
- ○ 政府/机关干部/公务员
- ○ 企业管理者（包括基层及中高层管理者）
- ○ 普通职员（办公室/写字楼工作人员）
- ○ 专业人员（如医生/律师/文体/记者/老师等）
- ○ 普通工人（如工厂工人/体力劳动者等）
- ○ 商业服务业职工（如销售人员/商店职员/服务员等）
- ○ 个体经营者/承包商
- ○ 自由职业者
- ○ 农林牧渔劳动者
- ○ 退休
- ○ 暂无职业
- ○ 其他职业人员（请注明）

4、您的最高学历（含目前在读）是？ 【单选题】

- ○ 小学及以下
- ○ 初中
- ○ 高中/中专/技校
- ○ 大学专科

○ 大学本科

○ 硕士研究生及以上

5、您的常住地区 【填空题】

省份

城市

6、您的出生日期是？ 【多项填空】

日期_____

7、您是否正在使用或曾经使用过在线运动媒介平台？ 【单选题】

包括在线运动教程 APP（如 keep）和包含健身视频的播放平台（如 Bilibili）

○ 是

○ 否

8、您正在使用或曾经使用过的最熟悉的一款或几款在线运动媒介平台 【多选题】

□ keep

□ 咕咚

□ 小米运动

□ 悦跑圈

□ 悦动圈

□ 火辣健身

□ 每日瑜伽

□ 春雨计步器

□ 即刻运动

□ 乐动力

□ 糖豆

□ 步多多

□ Fittime

□ 薄荷健康

□ Bilibili

□ 抖音

□ 快手

□ 微博

□ 小红书

□ 其他

9、您获知这个（这些）平台的途径 【多选题】

□ 朋友推荐或身边有人使用

□ 网红、大 V 推荐

□ 线下渠道（楼宇广告、室内电子屏、电梯广告、地铁广告等）

□ 传统媒体渠道（电视、广播、纸媒等）

□ 其他

10、2020 年 1 月 1 日前，您使用在线运动媒介平台的频率是？ 【单选题】

○ 每周 1 次及以下

○ 每周 1-3 次

○ 每周 3-5 次

○ 每周 5 次及以上

11、2020 年 1 月 1 日后，您使用在线运动媒介平台的频率是？ 【单选题】

○ 每周 1 次及以下

○ 每周 1-3 次

○ 每周 3-5 次

○ 每周 5 次及以上

12、您使用在线运动媒介平台的目的是？ 【多选题】

□ 增肌塑形

□ 减脂

□ 增强体质

□ 社交

□ 解压

□ 其他

13、您的月收入？ 【单选题】

○ 暂无收入

○ 2000 以下

○ 2001-4000 元

○ 4001-6000 元

○ 6001-8000 元

○ 8001-10000 元

○ 10000 以上

14、您的月消费？ 【单选题】

○ 1500 元及以下

○ 1501-3000 元

○ 3001-4500 元

○ 4500 元以上

15、您对在线运动媒介平台的选择偏好？ 【矩阵单选题】

您选取常用平台的主要依据，请尽量避免填写中立选项

	非常重要	比较重要	无所谓	不太重要	与其无关
健身数据记录和健康监测功能	○	○	○	○	○

平台的便捷程度	○	○	○	○	○
平台对手机、平板等的适配性	○	○	○	○	○
系统稳定性	○	○	○	○	○
平台提供信息的准确性	○	○	○	○	○
平台提供信息的及时性	○	○	○	○	○
平台提供的信息易于理解	○	○	○	○	○
平台提供个性化定制服务	○	○	○	○	○
平台的问题反馈渠道畅通	○	○	○	○	○
平台具有社交互动功能	○	○	○	○	○

16、您是否于线上平台使用过和运动健身相关的付费功能？ 【单选题】

包括购买运动器材、相关营养补剂以及线下健身课程等

○ 是

○ 否

17、2020 年 1 月 1 日之后，您于线上平台运动健身相关的消费总额？ 【单选题】

○ 200 元及以下

○ 201-500 元

○ 501-1000 元

○ 1001-2000 元

○ 2001-5000 元

○ 5000 元以上

18、您使用线上平台消费过以下哪些产品？ 【多选题】

□ 运动装备类

□ 运动 APP 会员（不包括在线视频类平台会员）

□ 营养补剂类

□ 健身课程类

□ 其他

19、请根据您的实际感受，选择最符合的选项 【矩阵单选题】

请尽量避免填写中立选项

	非常不同意	不同意	无所谓	同意	非常同意
和支付的金钱相比，购买运动健身类 APP 会员是划算的	○	○	○	○	○
和付出的时间和精力相比，购买运动健身类 APP 会员是值得的	○	○	○	○	○
总的来说，购买运动健身类 APP 会员为我带来了很多好处	○	○	○	○	○
购买运动健身类 APP 会员可以让我更好地达到运动健身的目的	○	○	○	○	○

购买运动健身类 APP 会员可以让我享受到更多特权（定制化健身计划、食谱、会员专属课程）	○	○	○	○	○
购买运动健身类 APP 会员可以让我更轻松地养成运动健身的习惯	○	○	○	○	○

20、请根据您的实际感受，选择最符合的选项 【矩阵单选题】

请尽量避免填写中立选项

	非常不同意	不同意	无所谓	同意	非常同意
媒体宣传或者社交平台上的分享会促使我购买运动健身类 APP 会员	○	○	○	○	○
我会在家人、朋友的推荐下购买运动健身类 APP 会员	○	○	○	○	○
我认为购买运动健身类 APP 会员比免费使用更能达到运动健身目的	○	○	○	○	○
我愿意为优质的互联网内容或服务付费	○	○	○	○	○
如果运动健身类 APP 会员特权是我所需要的，我愿意购买	○	○	○	○	○

21、您对在线健身平台未来发展的建议？ 【多选题】

□ 完善服务平台管理和问题反馈机制

□ 拓展服务项目种类

□ 提高服务质量与水平

□ 增强用户体验性

□ 加强用户间交流

□ 其他

22、您是否在线下有过运动健身相关的支出？ 【单选题】

包括购买运动器材、相关营养补剂以及线下健身课程等

○ 是

○ 否

23、2020 年 1 月 1 日之后，您于线上平台运动健身相关的消费总额？ 【单选题】

○ 200 元及以下

○ 201-500 元

○ 501-1000 元

○ 1001-2000 元

○ 2001-5000 元

○ 5000 元以上

24、您在线下有过运动健身相关的哪些支出？ 【多选题】

□ 健身卡

□ 运动装备类

□ 营养补剂类

□ 健身课程类

□ 其他

25、请根据您的实际感受，选择最符合的选项 【矩阵单选题】

请尽量避免填写中立选项

	非常不同意	不同意	无所谓	同意	非常同意
我愿意为优质服务付费	○	○	○	○	○
我目前没有运动健身需求	○	○	○	○	○

附录 2 问卷编码表

Q1	您的性别	性别题
Q2	您的年龄是？	单选题
Q3	您目前的职业是？	单选题
Q4	您的最高学历（含目前在读）是？	单选题
Q5	您的常住地区	城市题
Q6	您的出生日期是？	日期题
Q7	您是否正在使用或曾经使用过在线运动媒介平台？	单选题
Q8	您正在使用或曾经使用过的最熟悉的一款或几款在线运动媒介平台	多选题
Q9	您获知这个（这些）平台的途径	多选题
Q10	2020 年 1 月 1 日前，您使用在线运动媒介平台的频率是？	单选题
Q11	2020 年 1 月 1 日后，您使用在线运动媒介平台的频率是？	单选题
Q12	您使用在线运动媒介平台的目的是？	多选题
Q13	您的月收入？	单选题
Q14	您的月消费？	单选题
Q15	您对在线运动媒介平台的选择偏好？	矩阵单选题
Q16	您是否于线上平台使用过和运动健身相关的付费功能？	单选题
Q17	2020 年 1 月 1 日之后，您于线上平台运动健身相关的消费总额？	单选题
Q18	您使用线上平台消费过以下哪些产品？	多选题

Q19	请根据您的实际感受，选择最符合的选项	矩阵单选题
Q20	请根据您的实际感受，选择最符合的选项	矩阵单选题
Q21	您对在线健身平台未来发展的建议？	多选题
Q22	您是否在线下有过运动健身相关的支出？	单选题
Q23	2020 年 1 月 1 日之后，您于线上平台运动健身相关的消费总额？	单选题
Q24	您在线下有过运动健身相关的哪些支出？	多选题
Q25	请根据您的实际感受，选择最符合的选项	矩阵单选题

附录 3 可视化分析代码

```
import numpy as np

import pandas as pd

from pandas import plotting

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.graph_objs as go

import plotly.offline as py

from sklearn.cluster import KMeans

import warnings

warnings.filterwarnings('ignore')

io = '.../Mall_Customers.csv'

df = pd.DataFrame(pd.read_csv(io))

# 修改列名

df.rename(columns={'Annual Income (k$)': 'Annual Income', 'Spending Score (1-100)': 'Spending Score'}, inplace=True)

print(df.head())

print(df.describe())

print(df.shape)

print(df.count())

print(df.dtypes)
```

```
plotting.parallel_coordinates(df.drop('CustomerID', axis=1), 'Gender')

plt.title('平行坐标图', fontsize=12)

plt.grid(linestyle='-.')

plt.show()

sns.set(palette="muted", color_codes=True)       # seaborn 样式

#  配置

plt.rcParams['axes.unicode_minus'] = False       #  解决无法显示符号的问题

sns.set(font='SimHei', font_scale=0.8)           #  解决 Seaborn 中文显示问题

#  绘图

plt.figure(1, figsize=(13, 6))

n = 0

for x in ['Age', 'Annual Income', 'Spending Score']:

    n += 1

    plt.subplot(1, 3, n)

    plt.subplots_adjust(hspace=0.5, wspace=0.5)

    sns.distplot(df[x], bins=16, kde=True)       # kde  密度曲线

    plt.title('{}分布情况'.format(x))

    plt.tight_layout()

plt.show()

plt.figure(1, figsize=(13, 6))

k = 0
```

```
for x in ['Age', 'Annual Income', 'Spending Score']:
    k += 1
    plt.subplot(3, 1, k)
    plt.subplots_adjust(hspace=0.5, wspace=0.5)
    sns.countplot(df[x], palette='rainbow', alpha=0.8)
    plt.title('{}分布情况'.format(x))
    plt.tight_layout()
plt.show()
df_gender_c = df['Gender'].value_counts()
p_lables = ['Female', 'Male']
p_color = ['lightcoral', 'lightskyblue']
p_explode = [0, 0.05]
# 绘图
plt.pie(df_gender_c, labels=p_lables, colors=p_color, explode=p_explode, shadow=True, autopct='%.2f%%')
plt.axis('off')
plt.legend()
plt.show()
# df_a_a_s = df.drop(['CustomerID'], axis=1)
sns.pairplot(df, vars=['Age', 'Annual Income', 'Spending Score'], hue='Gender', aspect=1.5, kind='reg')
plt.show()
```

```
# 根据分类变量分组绘制一个纵向的增强箱型图
plt.rcParams['axes.unicode_minus'] = False       # 解决无法显示符号的问题
sns.set(font='SimHei', font_scale=0.8)           # 解决 Seaborn 中文显示问题
sns.boxenplot(df['Gender'], df['Spending Score'], palette='Blues')
# x:设置分组统计字段，y:数据分布统计字段
sns.swarmplot(x=df['Gender'], y=df['Spending Score'], data=df, palette='dark', alpha=0.5, size=6)
plt.title('男女性的消费能力比较', fontsize=12)
plt.show()
plt.rcParams['axes.unicode_minus'] = False       # 解决无法显示符号的问题
sns.set(font='SimHei', font_scale=0.8)           # 解决 Seaborn 中文显示问题
m = 0
for feature in ['Age', 'Annual Income', 'Spending Score']:
    m += 1
    plt.subplot(1, 3, m)
    plt.subplots_adjust(hspace=0.3, wspace=0.3)
    sns.violinplot(x=feature, y='Gender', data=df, palette='Blues')
    sns.swarmplot(x=feature, y='Gender', data=df, palette='dark', alpha=0.5, size=4)
    plt.ylabel('性别' if m == 1 else '')
plt.show()
```

附录 4　Logistic 回归模型代码

```
import numpy as np

def sigmoid(x):

    z = 1 / (1 + np.exp(-x))

    return z

def initialize_params(dims):

    W = np.zeros((dims, 1))

    b = 0

    return W, b

### 定义逻辑回归模型主体

def logistic(X, y, W, b):

    输入：

    X: 输入特征矩阵

    y: 输出标签向量

    W: 权值参数

    b: 偏置参数

    输出：

    a: 逻辑回归模型输出

    cost: 损失

    dW: 权值梯度

    db: 偏置梯度
```

```
    # 训练样本量
    num_train = X.shape[0]
    # 训练特征数
    num_feature = X.shape[1]
    # 逻辑回归模型输出
    a = sigmoid(np.dot(X, W) + b)
    # 交叉熵损失
    cost = -1/num_train * np.sum(y*np.log(a) + (1-y)*np.log(1-a))
    # 权值梯度
    dW = np.dot(X.T, (a-y))/num_train
    # 偏置梯度
    db = np.sum(a-y)/num_train
    # 压缩损失数组维度
    cost = np.squeeze(cost)
    return a, cost, dW, db
### 定义逻辑回归模型训练过程
def logistic_train(X, y, learning_rate, epochs):
    输入：
    X: 输入特征矩阵
    y: 输出标签向量
    learning_rate: 学习率
```

epochs: 训练轮数

输出：

cost_list: 损失列表

params: 模型参数

grads: 参数梯度

```
# 初始化模型参数
W, b = initialize_params(X.shape[1])
# 初始化损失列表
cost_list = []


# 迭代训练
for i in range(epochs):
    # 计算当前次的模型计算结果、损失和参数梯度
    a, cost, dW, db = logistic(X, y, W, b)
    # 参数更新
    W = W -learning_rate * dW
    b = b -learning_rate * db
    # 记录损失
    if i % 100 == 0:
        cost_list.append(cost)
    # 打印训练过程中的损失
```

```
        if i % 100 == 0:
            print('epoch %d cost %f' % (i, cost))
    # 保存参数
    params = {
        'W': W,
        'b': b
    }
    # 保存梯度
    grads = {
        'dW': dW,
        'db': db
    }
    return cost_list, params, grads
def predict(X, params):
    输入：
    X: 输入特征矩阵
    params: 训练好的模型参数
    输出：
    y_prediction: 转换后的模型预测值
    # 模型预测值
    y_prediction = sigmoid(np.dot(X, params['W']) + params['b'])
```

```
        # 基于分类阈值对概率预测值进行类别转换
        for i in range(len(y_prediction)):
            if y_prediction[i] > 0.5:
                y_prediction[i] = 1
            else:
                y_prediction[i] = 0
        return y_prediction
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets.samples_generator import make_classification
class logistic_regression():
    def __init__(self):
        pass
    def sigmoid(self, x):
        z = 1 / (1 + np.exp(-x))
        return z
    def initialize_params(self, dims):
        W = np.zeros((dims, 1))
        b = 0
        return W, b
    def logistic(self, X, y, W, b):
```

```
        num_train = X.shape[0]
        num_feature = X.shape[1]
        a = self.sigmoid(np.dot(X, W) + b)
        cost = -1 / num_train * np.sum(y * np.log(a) + (1 - y) * np.log(1 - a))
        dW = np.dot(X.T, (a - y)) / num_train
        db = np.sum(a - y) / num_train
        cost = np.squeeze(cost)
        return a, cost, dW, db
    def logistic_train(self, X, y, learning_rate, epochs):
        W, b = self.initialize_params(X.shape[1])
        cost_list = []
        for i in range(epochs):
            a, cost, dW, db = self.logistic(X, y, W, b)
            W = W - learning_rate * dW
            b = b - learning_rate * db
            if i % 100 == 0:
                cost_list.append(cost)
            if i % 100 == 0:
                print('epoch %d cost %f' % (i, cost))
        params = {
            'W': W,
```

```
            'b': b
        }
        grads = {
            'dW': dW,
            'db': db
        }
        return cost_list, params, grads
    def predict(self, X, params):
        y_prediction = self.sigmoid(np.dot(X, params['W']) + params['b'])
        for i in range(len(y_prediction)):
            if y_prediction[i] > 0.5:
                y_prediction[i] = 1
            else:
                y_prediction[i] = 0
        return y_prediction
    def accuracy(self, y_test, y_pred):
        correct_count = 0
        for i in range(len(y_test)):
            for j in range(len(y_pred)):
                if y_test[i] == y_pred[j] and i == j:
                    correct_count += 1
```

```
            accuracy_score = correct_count / len(y_test)

            return accuracy_score

        def create_data(self):

            X, labels = make_classification(n_samples=100, n_features=2, n_redundant=0,
n_informative=2,

                                            random_state=1, n_clusters_per_class=2)

            labels = labels.reshape((-1, 1))

            offset = int(X.shape[0] * 0.9)

            X_train, y_train = X[:offset], labels[:offset]

            X_test, y_test = X[offset:], labels[offset:]

            return X_train, y_train, X_test, y_test

        def plot_logistic(self, X_train, y_train, params):

            n = X_train.shape[0]

            xcord1 = []

            ycord1 = []

            xcord2 = []

            ycord2 = []

            for i in range(n):

                if y_train[i] == 1:

                    xcord1.append(X_train[i][0])

                    ycord1.append(X_train[i][1])

                else:
```

```
                xcord2.append(X_train[i][0])
                ycord2.append(X_train[i][1])
        fig = plt.figure()
        ax = fig.add_subplot(111)
        ax.scatter(xcord1, ycord1, s=32, c='red')
        ax.scatter(xcord2, ycord2, s=32, c='green')
        x = np.arange(-1.5, 3, 0.1)
        y = (-params['b'] - params['W'][0] * x) / params['W'][1]
        ax.plot(x, y)
        plt.xlabel('X1')
        plt.ylabel('X2')
        plt.show()
if __name__ == "__main__":
    model = logistic_regression()
    X_train, y_train, X_test, y_test = model.create_data()
    print(X_train.shape, y_train.shape, X_test.shape, y_test.shape)
    cost_list, params, grads = model.logistic_train(X_train, y_train, 0.01, 1000)
    print(params)
    y_train_pred = model.predict(X_train, params)
    accuracy_score_train = model.accuracy(y_train, y_train_pred)
    print('train accuracy is:', accuracy_score_train)
```

```
y_test_pred = model.predict(X_test, params)

accuracy_score_test = model.accuracy(y_test, y_test_pred)

print('test accuracy is:', accuracy_score_test)

    model.plot_logistic(X_train, y_train, params)
```

附录 5　Logistic 回归模型选择变量

变量	变量符号	变量定义
是否选择线上平台使用运动健身相关的付费功能	Y	1，是
		0，否
性别	Male	1，　男
		0，女
年龄	Age_20	1，20 岁以下
		0，其他
	Age21_30	1，21-30 岁
		0，其他
	Age31_40	1，31-40 岁
		0，其他
	Age41_50	1，41-50 岁
		0，其他
最高学历	Primary	1，　小学及以下
		0，其他
	Junior	1，　初中
		0，其他
	High	1，　高中
		0，其他
	Undergraduate	1，　大学本科

		0，其他
	College	1， 大学专科
		0，其他
收入	Income1	1， 无收入
		0，其他
	Income2	1，2000 以下
		0，其他
	Income3	1，2001-4000 元
		0，其他
	Income4	1，4001-6000
		0， 其他
	Income5	1，6001-8000
		0，其他
	Income6	1，8001-10000
		0，其他
居住地	City	1， 城市
		0，其他
	Countryside	1， 农村
		0，其他
月消费	Expenditure1	1，1500 元以下
		0，其他

	Expenditure1	1，1501-3000 元
		0，其他
	Expenditure1	1，3001-4500 元
		0，其他

附录 6　K-Modes 聚类算法代码

```
df_a_sc = df[['Age', 'Spending Score']].values

# 存放每次聚类结果的误差平方和

inertia1 = []

for n in range(1, 11):

    # 构造聚类器

    km1 = (KMeans(n_clusters=n,          # 要分成的簇数，int 类型，默认值为 8

                  init='k-means++',      # 初始化质心，k-means++是一种生成初始质心的算法

                  n_init=10,             # 设置选择质心种子次数，默认为 10 次。返回质心最好的一次结果（好是指计算时长短）

                  max_iter=300,          # 每次迭代的最大次数

                  tol=0.0001,            # 容忍的最小误差，当误差小于 tol 就会退出迭代

                  random_state=111,      # 随机生成器的种子 ，和初始化中心有关

                  algorithm='elkan'))   # 'full'是传统的 K-Means 算法，'elkan'是采用 elkan K-Means 算法

    # 用训练数据拟合聚类器模型

    km1.fit(df_a_sc)

    # 获取聚类标签

    inertia1.append(km1.inertia_)

plt.figure(1, figsize=(15, 6))
```

```
plt.plot(np.arange(1, 11), inertia1, 'o')

plt.plot(np.arange(1, 11), inertia1, '-', alpha=0.7)

plt.title('手肘法图', fontsize=12)

plt.xlabel('聚类数'), plt.ylabel('SSE')

plt.grid(linestyle='-.')

plt.show()

km1_result = (KMeans(n_clusters=4, init='k-means++', n_init=10, max_iter=300,

                     tol=0.0001,  random_state=111, algorithm='elkan'))

# 先 fit()再 predict()，一次性得到聚类预测之后的标签

y1_means = km1_result.fit_predict(df_a_sc)

# 绘制结果图

plt.scatter(df_a_sc[y1_means == 0][:, 0], df_a_sc[y1_means == 0][:, 1], s=70, c='blue', label='1',
alpha=0.6)

plt.scatter(df_a_sc[y1_means == 1][:, 0], df_a_sc[y1_means == 1][:, 1], s=70, c='orange', label='2',
alpha=0.6)

plt.scatter(df_a_sc[y1_means == 2][:, 0], df_a_sc[y1_means == 2][:, 1], s=70, c='pink', label='3',
alpha=0.6)

plt.scatter(df_a_sc[y1_means == 3][:, 0], df_a_sc[y1_means == 3][:, 1], s=70, c='purple', label='4',
alpha=0.6)

plt.scatter(km1_result.cluster_centers_[:, 0], km1_result.cluster_centers_[:, 1], s=260, c='gold',
label='质心')

plt.title('聚类图(K=4)', fontsize=12)

plt.xlabel('年收入(k$)')
```

```
plt.ylabel('消费分数(1-100)')
plt.legend()
plt.grid(linestyle='-.')
plt.show()
df_ai_sc = df[['Annual Income', 'Spending Score']].values
# 存放每次聚类结果的误差平方和
inertia2 = []
for n in range(1, 11):
    # 构造聚类器
    km2 = (KMeans(n_clusters=n, init='k-means++', n_init=10, max_iter=300, tol=0.0001,
random_state=111, algorithm='elkan'))
    # 用训练数据拟合聚类器模型
    km2.fit(df_ai_sc)
    # 获取聚类标签
    inertia2.append(km2.inertia_)
# 绘制手肘图确定 K 值
plt.figure(1, figsize=(15, 6))
plt.plot(np.arange(1, 11), inertia1, 'o')
plt.plot(np.arange(1, 11), inertia1, '-', alpha=0.7)
plt.title('手肘法图', fontsize=12)
plt.xlabel('聚类数'), plt.ylabel('SSE')
plt.grid(linestyle='-.')
```

```
plt.show()

km2_result = (KMeans(n_clusters=5, init='k-means++', n_init=10, max_iter=300,

                     tol=0.0001,  random_state=111, algorithm='elkan'))

# 先 fit()再 predict()，一次性得到聚类预测之后的标签

y2_means = km2_result.fit_predict(df_ai_sc)

# 绘制结果图

plt.scatter(df_ai_sc[y2_means == 0][:, 0], df_ai_sc[y2_means == 0][:, 1], s=70, c='blue', label='1',
alpha=0.6)

plt.scatter(df_ai_sc[y2_means == 1][:, 0], df_ai_sc[y2_means == 1][:, 1], s=70, c='orange', label='2',
alpha=0.6)

plt.scatter(df_ai_sc[y2_means == 2][:, 0], df_ai_sc[y2_means == 2][:, 1], s=70, c='pink', label='3',
alpha=0.6)

plt.scatter(df_ai_sc[y2_means == 3][:, 0], df_ai_sc[y2_means == 3][:, 1], s=70, c='purple', label='4',
alpha=0.6)

plt.scatter(df_ai_sc[y2_means == 4][:, 0], df_ai_sc[y2_means == 4][:, 1], s=70, c='green', label='5',
alpha=0.6)

plt.scatter(km2_result.cluster_centers_[:,  0],  km2_result.cluster_centers_[:,  1],  s=260,  c='gold',
label='质心')

plt.title('聚类图(K=5)', fontsize=12)

plt.xlabel('年收入(k$)')

plt.ylabel('消费分数(1-100)')

plt.legend()

plt.grid(linestyle='-.')
```

plt.show()

附录 7 决策树算法代码

```
plt.rcParams['font.sans-serif']=['SimHei']   #解决中文显示乱码问题

plt.rcParams['axes.unicode_minus']=False

import sklearn.linear_model as LM

from sklearn.metrics import classification_report

from sklearn.model_selection import cross_val_score,train_test_split

from sklearn.datasets import make_regression

from sklearn import tree

from sklearn.preprocessing import LabelEncoder

from sklearn.tree import export_graphviz

import graphviz

data=pd.read_excel('alldata2.xlsx')

data=data.replace(0,np.NaN)

data=data.dropna()

X_train=data.iloc[:,1:]

y_train=data['是否付费']

print(y_train.value_counts())

print("输入变量:\n",X_train.columns)

modelDTC = tree.DecisionTreeClassifier(max_depth=3,random_state=123)

modelDTC.fit(X_train, y_train)

print(tree.export_text(modelDTC))
```

dot_data=export_graphviz(modelDTC,out_file=None,class_names=['1','2'])

graph=graphviz.Source(dot_data)

graph.render('决策树可视化 1')

附录 8 实地调研影像记录


很危险 生命很宝贵