**数据可视化大作业——以星巴克为研究对象**

202211260028 梁思佳

目录

[**数据可视化大作业——以星巴克为研究对象** 1](#_Toc175093088)

[数据来源 1](#_Toc175093089)

[数据解读 1](#_Toc175093090)

[数据处理 1](#_Toc175093091)

[数据筛选整合 1](#_Toc175093092)

[重复值处理 2](#_Toc175093093)

[缺失值处理 2](#_Toc175093094)

[异常值处理 2](#_Toc175093095)

[数据分析 2](#_Toc175093096)

[门店分布 2](#_Toc175093097)

[不同产品的月销情况 2](#_Toc175093098)

[购物时间分析 3](#_Toc175093099)

[用户转化率 3](#_Toc175093100)

[用户画像 4](#_Toc175093101)

[基于Echarts的可视化 5](#_Toc175093102)

[大屏幕排版 5](#_Toc175093103)

[图表设计 5](#_Toc175093104)

[可视化效果 5](#_Toc175093105)

[结果总结 6](#_Toc175093106)

数据来源

为了分析国内星巴克的整体情况，我通过kaggle，和鲸社区，公司官网等网站获得了相关主数据集，并使用八爪鱼，python等软件进行爬取，尝试与这些公开数据集进行比较校正确保数据的可信度和可使用性。

数据解读

第一个数据集是星巴克门店数据（directory.csv），数据统计时间截至2017年2月，包括了星巴克旗公司旗下的所有品牌门店的地址号，电话，具体地址，持有股份类型，所在街道，城市，省份，国家，邮箱地址，营业时间，和经纬度等信息。

第二个数据集是星巴克的客户数据，是星巴克公司自主提供的公开数据集，用于模拟他们的客户和交易以便于进一步促销产品，包含三个文件：

portfolio.csv - 发送给顾客的优惠数据（10 项优惠 x 6 列）

profile.csv - 客户的人口统计数据（17,000 名客户 x 5 列）

transcript.csv - 客户对报价和交易的响应（306,648 个事件 x 4 列）

第三个数据集是国内星巴克2023年1～6月的销售数据，包含了149117条订单数据，无空值，包含了订单号，订单具体时间，购买种类与个数，购买备注，开销等信息。(coffeesales.xlsx)

数据处理

数据筛选整合

为了便于分析我先对数据集中数据进行筛选和部分整合。将全国性质的数据集通过国家筛选为只针对中国的，同时第一个数据集品牌只选择星巴克，然后将国内星巴克的数据根据分析聚焦的重点需要精简列数，整理到了一张excel表里。主要为三部分的数据：

**用户数据**（性别，年龄，注册日期，收入，接收优惠次数，使用优惠次数）

**门店数据**（公司股权类型，城市，省份，营业时间，经纬度）

**交易数据**（每单购买商品，商品单价，购买个数，总消费，月，日，小时）

接下来，对三张表中的数据进行以下处理。

重复值处理

重复值指数据集中存在完全相同的行或列，可以使用excel的去重操作直接实现；也可以先进行重复查找，看是否需要合并数据。如果每行数据有唯一标识会便于判断。

本文收集的数据集具有一定威信度且较为整齐，不太需要进行去重处理。

缺失值处理

缺失值指数据集中存在空缺或有未知的值，可以用删除法，填补法，或基于时间序列等数学公式的差值。

经过筛选后本文主要在用户数据这张表上存在用户部分信息的缺失。是因为有些用户注册了星巴克小程序进行下单，但是并没有完善自己的用户信息。所以不可以删去，可以在年龄处赋值118，标记未完善个人信息的用户。

异常值处理

异常值指数据集中明显偏离正常数据范围或有违常理的数值。如果影响较小可以进行删除；或者用均值中位数众数等特定值替换；可能要进行分箱操作；或者保留异常值分析其背后的特殊情况。

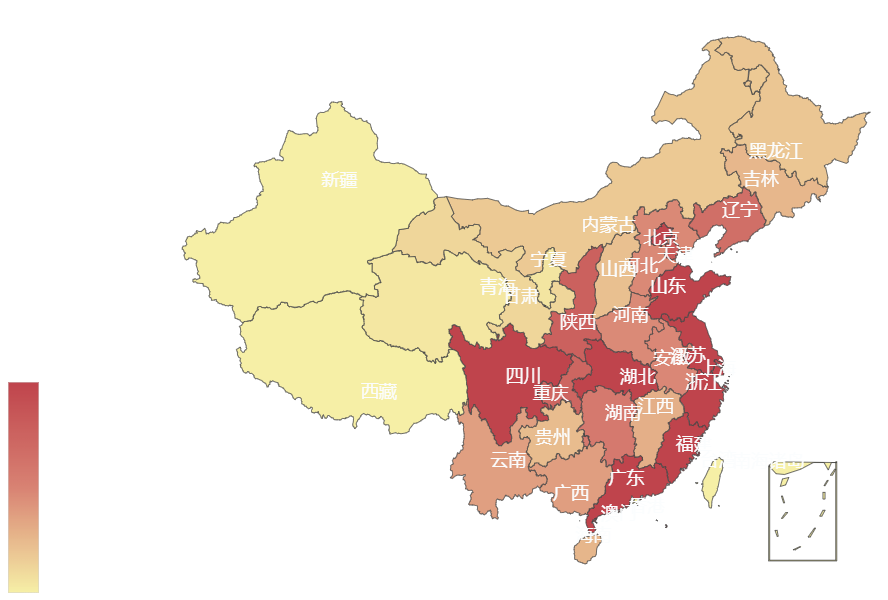
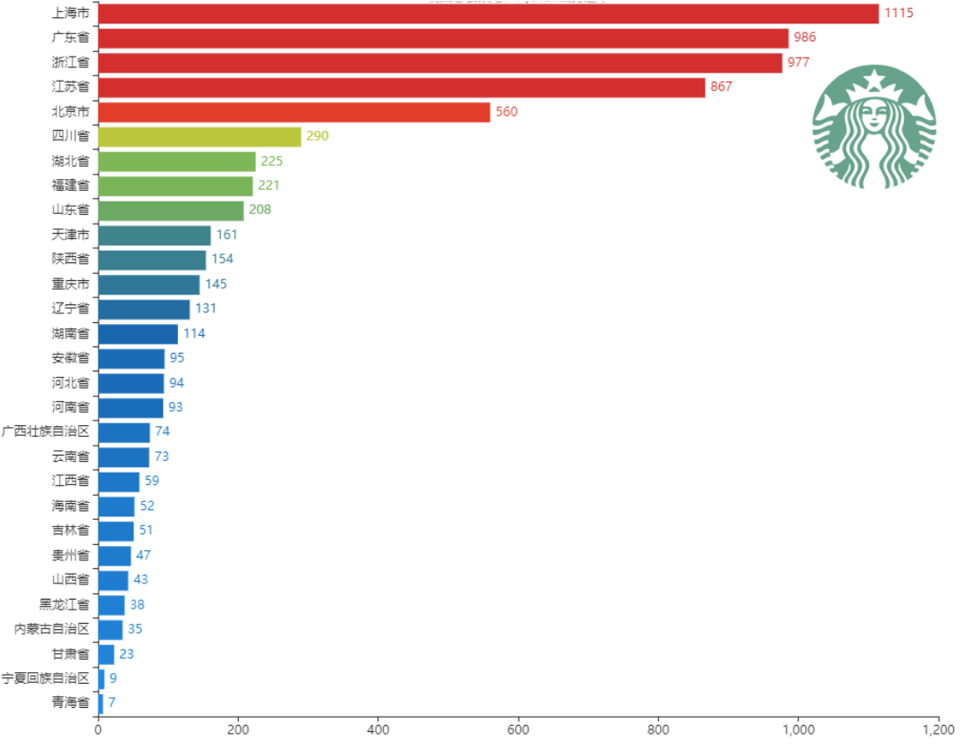
本文中的异常值主要为性别选择中的O和年龄有90多岁，这可能和星巴克官方对部分用户的信息模糊和保护有关，或者部分用户填年龄时会往偏大的填。我们保留异常值，在后续分析用户年龄这类数据时需要注意，其数据真实度有待商榷。

数据分析

门店分布

使用excel的筛选和counta函数以省为单位得到星巴克在国内的门店分布数据（无港珠澳等非内陆地区的），并且可以得到拥有门店数前十的城市。

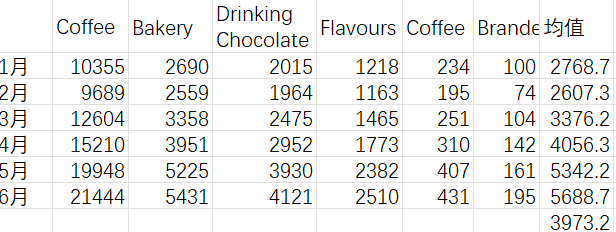
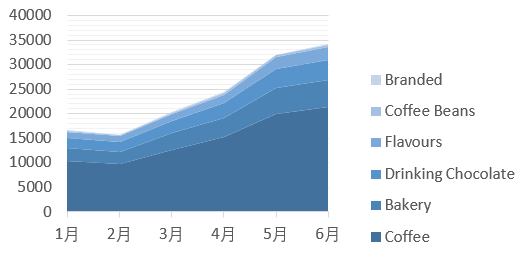
初步考虑用地图热力图显示门店在不同省的分布情况，通过颜色深浅生动形象地展展现出不同省份的门店多少，并且现在地图上用波点圆圈特殊标记前十的城市，使热销区显著。



不同产品的月销情况

已知产品类型主要有Coffee，Drinking chocolate，Bakery，Coffee beans，Branded，Flavors六种。统计他们在六个月内的销售金额得下表，可知二月是个小低谷，但是一月到六月总体获利呈增长趋势，平均每月获利13521.81美元。带来收入最多的产品是咖啡，烘培品和热可可次之，然后是咖啡豆，品牌产品，和调料包。

初步考虑用堆叠图展示，因为堆叠图可以利用面积形象展示每个月的收益“构成”并且利用边界线展示出不同产品收益随时间的变化。

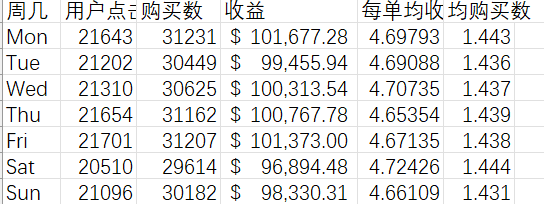


购物时间分析

接下来是购物时间分析，首先是周期性时间，即周一到周日。通过用countifs和sumifs函数可得周一到周日 用户的点击数，购买数，以及总收益，每单均收益，和均购买次数。

周一是一周内收益最多的，可能和周一是每周上班的第一天有关；周六是一周内收益最少的，可能和周六是周末的第一天有关。

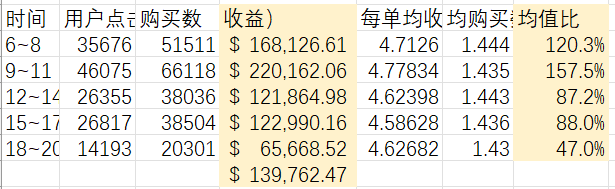
对总收益进行可视化，使用简明直观的柱形图，适当的空白会把条形图分隔成视觉组，减少视觉干扰并直观化了一周内的收益变化情况。



购物时间分析还可从一天的不同时间段看，我们将时间段分为：清晨（6—8），上午（9—11），中午（12—14），下午（15—17），晚上（18—20），五个时段。和上面一样，用countifs和sumifs函数得到用户点击数，购买数，总收益，每单均收益，和均购买数。

这五项指标都随时间推移呈先增后减的趋势，清晨和上午的收益和购买是最多的，中午、下午，和晚上 较少人购买。

在可视化时，我们将总收益用水平条形图展示，在数据面前会先看到时间，易于展示不同时间段的收益，反应一天的时间内星巴克的营业情况。



用户转化率

使用TRANSPOSE和FILTER组合可以对应起每个客户的优惠获得与使用记录（如果不加TRANSPOSE函数则需要逐一进行操作，无法批量操作，因为纵向输出会出现溢出现象），然后根据用户的注册情况，信息完备情况，优惠券领取情况和优惠完成情况，我们可以大概定性地将用户分为：总用户，可挖掘客户，受吸引客户，较稳定客户，高价值客户。

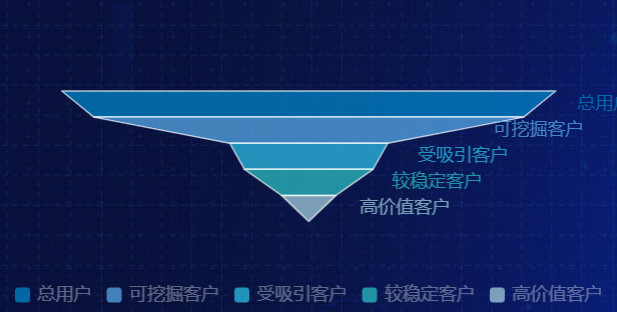
可挖掘客户：填写了完整信息，便于挖掘其需求，精准用户画像，实现对点打击，是潜在客户。并且一定程度上说明用户较为信息开明，易于捕获。

受吸引客户：会接受并且浏览优惠及广告，表现出对星巴克的兴趣但是出于价格或其他原因未使用，可以在特殊节假日等时间加强力度实现突破。

较稳定客户：即使未使用优惠等也进行了购买，是对星巴克表现出一定偏爱的客户，忠诚度较高，可定期投放较大的优惠加固其喜爱程度和用户粘性。

高价值客户：使用优惠完成交易的稳定客户，对星巴克支持度极高，并且接受品牌福利，是高价值客户，可以带动星巴克的各式各样活动。

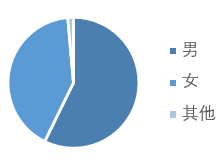
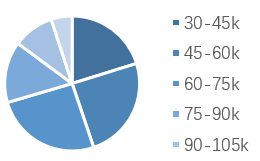
使用漏斗图，生动形象的展现了浅层用户到深层用户的转化率。

用户画像

这部分的信息分析针对填写了完备信息的用户，我们可以从性别，收入，年龄等方面进行用户画像分析。

用户有57.2%的男性，41.3%的女性和1.4%的其他。可以用中规中矩的饼图进行可视化。

年龄跨度为18到101，大约按10岁为一梯度分为八梯度后呈正态分布趋势，用户主要集中于中年群体，可能和星巴克比较贵有关，比起青少年喜欢的饮品，它更像是一种奢侈品和品牌的代表。

用户收入跨度为3w到12w，均值为65.4k。

可大致化为等差的六个梯度，各梯度的用户含量差异不是非常大，可视化时可考虑用具有点击放大效应的环形图，兼具美观性和对比饼图更进一步的量化视觉冲突。

最后，可以根据备注制作的词云图。



基于Echarts的可视化

大屏幕排版

采取典型的排版，将屏幕分为左右上下四大板块。

上方有小牌子展示总体数据：门店总数和销量总数。

下方为地图展示，并且兼用盒模型实现动态效果。

左边分割三个区域块分别展示：

以Week为单位的周期销量，客户转化率，客户收入分布

右边三个区域块分别展示：

单日不同时段盈利条形图，不同产品月销量堆叠图，用户备注词云图

可视化屏幕总色调为蓝色，不会因为使用杂色产生较大的视觉干扰，底图为渐变色但是没有喧宾夺主反而使大屏更深沉立体。

图表设计

在图表设计方面，数据分析的时候就做出了阐释，已经对数据进行了总体分析和要点突出，选择的图表都与要表达的重点信息相辅相成彼此契合。并且除了图表设计，直观的数据就直接展示了。

在制作图表的过程中，也着重注意了图表的交互功能，以实现大屏的智能化，提高可读性易读性。

可视化效果



不足之处与遇到的问题：地图的渲染有差池，用另一个html写了成功的中国地图热力图，大屏里也无法加载数据项只能显示经纬度，但是用了黄色涟漪突出门店数前十的城市。

结果总结

获取了较为全面的数据，通过各种手段得到了具体的重要的数据，并对数据进行了全面的分析和选择，兼顾使用echarts的各种功能和组件，包括但不限于tooltip,toolbox,lengend……，得到了各式各类的图表。然后将图表进行了可交互性的可视化，整个过程实现了星巴克中国市场在用户，门店数，销量三方面的整体可视化。