

**2019 届毕业设计(论文)**

**课 题 名 称：谷歌应用市场数据分析**

**课 题 名 称（英文）：Google Appstore Data Analysis**

**学 生 姓 名： 陈前 学 号： 2016082112**

**专 业 名 称： 计算机科学与技术**

**指 导 教 师： 朱绍军 职 称： 讲师**

**所 在 学 院： 信息工程学院**

**完 成 日 期： 2020 年 5 月 1 日**

**教务处制表**

**谷歌应用市场数据分析**

**摘要：** 随着华为、小米等手机品牌的崛起，安卓系统已经成为了移动端的主流操作系统，其开源的特性，使安卓应用得到快速发展，安卓应用的市场也随之猛增。本文通过探索性数据分析以及聚类分析的方法，对以谷歌应用市场为代表的安卓应用市场进行数据分析。EDA能够在数据不符合传统统计模型的条件下，从数据出发，从总体到个体、从普遍到局部，多层次、多视角的发现谷歌应用市场数据集内在的信息。 聚类分析能够在数据集没有较大相关性的属性间得到启发式的分析结果。

**关键词：**安卓应用市场，探索性数据分析，kmeans聚类

**Google Appstore Data Analysis**

**Abstract**: With the rise of mobile phone brands such as Huawei and Xiaomi, the Android system has become the mainstream operating system on the mobile terminal, and its open source feature has enabled the rapid development of Android applications, and the market for Android applications has also soared. This article uses exploratory data analysis and cluster analysis to analyze the Android application market represented by the Google application market. EDA can discover the internal information of the Google application market data set from the data, from the overall to the individual, from the general to the local, at multiple levels and from multiple perspectives when the data does not conform to the traditional statistical model. Cluster analysis can obtain heuristic analysis results between attributes that have no great correlation in the data set.

**Keywords:** Android application market, exploratory data analysis, kmeans clustering

**目 录**

[第一章 绪 论 4](#_Toc39519706)

[1.1 选题的意义 4](#_Toc39519707)

[1.2 研究现状 4](#_Toc39519708)

[1.3 发展趋势 4](#_Toc39519709)

[第二章 数据处理 5](#_Toc39519710)

[2.1 数据加载 5](#_Toc39519712)

[2.2 数据清洗 6](#_Toc39519713)

[第三章 探索性数据分析 6](#_Toc39519714)

[3.1 基础认知 6](#_Toc39519716)

[3.2 基础探索性数据分析 8](#_Toc39519717)

[3.3 应用品类策略 9](#_Toc39519718)

[3.4 应用大小策略 11](#_Toc39519719)

[3.5 应用价格策略 11](#_Toc39519720)

[第四章 聚类分析 13](#_Toc39519721)

[4.1 算法简介 13](#_Toc39519723)

[4.2 算法原理 13](#_Toc39519724)

[4.3 算法应用 13](#_Toc39519725)

[4.4 聚类分析总结 15](#_Toc39519726)

[参考文献 16](#_Toc39519727)

[结束语 17](#_Toc39519728)

[致谢 18](#_Toc39519729)

# 绪 论

## 选题的意义

随着华为、小米等手机品牌的崛起，安卓系统已经成为了移动端的主流操作系统，据统计资料表明，安卓系统在移动端所占比例达到了58.8%，而ios只占了32.2%。其开源的特性，使安卓应用得到快速发展，安卓应用的市场也随之猛增。和任何其他市场一样，有赚取暴利的产品，也有无人问津的产品，但不同于实体商品可以通过样品或者小规模生产销售的方式，来进行调研及市场的分析，安卓应用的开发在项目启动之前，就必须做好完备的市场分析。我们希望通过探索性数据分析及聚类分析，给安卓应用开发商一定的方向和建议。

## 研究现状

探索性数据分析从提出到现在40余年，作为统计分析的十分重要的一个环节，在各个领域都有着广泛的应用。

聚类是在能够有效识别数据的内在结构，在图像处理、语音识别等方面有着大量的应用。聚类的另一个主要应用是数据挖掘以及统计科学，对生物学、地质学、商学都有着重要的作用。

## 发展趋势

安卓应用的发展越来越快速且复杂，移动设备硬件的不断增强，云应用等技术的不断涌现，让安卓应用市场起伏不断，因此对于安卓应用市场的数据分析，不能停滞，需要不断的通过分析来调整应用开发的战略。

# 数据处理



## 数据加载

经过对基础知识的学习与回顾，文献[1-5]介绍了python语言的用法，numpy、pandas、scipy等包，以及seaborn、plotly包对于数据的可视化的不同图表的绘制，例如箱型图、茎叶图、双向直方图等。

在jupyter notebook中，导入pandas包，加载数据集，由图 2.1-1 数据综合情况所示，整个数据集一共10840行数据，13列特征值，其中Rating属性是float类型的，其余属性均为object类型。

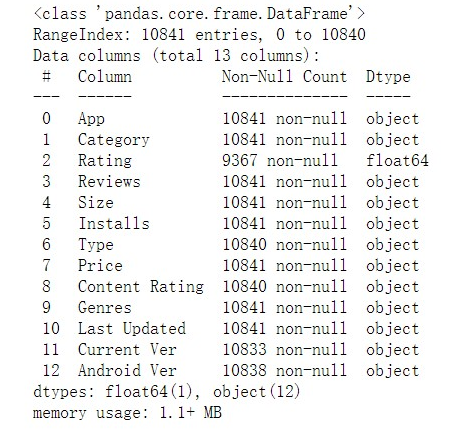


图 2.1‑1数据综合情况

具体数据如图2.1-2 数据具体情况所示，我们可以看到Size列有单位M，并且有非数字的噪音存在，Installs列中有‘+’号等。我们需要将Reviews、Size、Installs、Price转换为float类型并且将Size统一单位。因为NaN在整个数据集出现的不多，所以我们直接将有NaN的样例删除。



图 2.1‑2 数据具体情况

## 数据清洗

1. 我们对app的名字进行去重。
2. 经观察Type列只存在Free和Paid两种值，所以通过Type对数据集进行一遍筛选，筛去一部分数据错位等不正常的样例。通过对各列的值去重打印后，除了NaN，没有异常值出现。
3. 最后通过dropna()对整个DataFrame进行缺失值的过滤。
4. 因为探索性数据分析对于数据的野值（离群值）不敏感，所以在EDA过程中不做处理，但是野值（离群值）会严重影响聚类的结果，在聚类前需要对野值进行过滤。
5. 最后我们将Size列中的数据单位统一为K，将M为单位的数据乘以1000，去掉Installs列中的’+’号，并将Reviews、Size、Installs、Price转化为float型。

# 探索性数据分析



## 基础认知

自17世纪中叶统计学开始萌芽，到上世纪三、四十年代，统计学逐渐发展完善，但也越来越陷入为追求“不确定性的数学度量”而过分迁就于“正态模型”而不去挖掘数据背后的深层结构与意义。随着数理统计者的概率论和数学水平的不断上涨导致，统计结果变得高深但不直观，且提供的有用的方法极少。

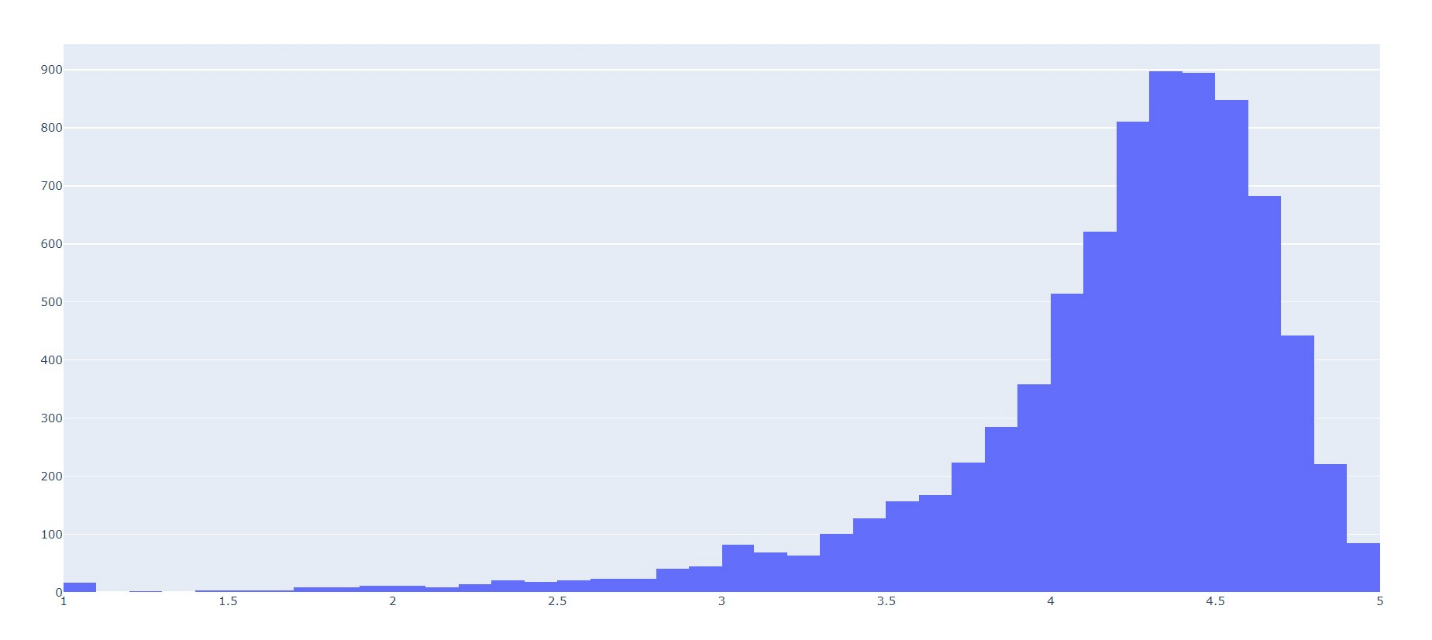
探索性数据分析（exploratory data analysis，EDA）意在让数据说话，与传统的统计学从一个设定的模型（正态模型）出发不同，EDA不需要先验假设，而是在数据分析过程中逐步产生假设，并且解决。在数据找不到对应的传统统计学模型时，如图3.1-1 Rating分布，EDA就给分析者提供了稳健、高效、直截了当地导向数据的结构的分析方法。

图 3.1‑1Rating分布

EDA不注重方法的概率论根据，对于“不精确度”的度量不过分追求，更多的是使用一种比较“松弛”的、“非正式”的方法，相较于效率，EDA更加注重方法的“抗干扰性”。

在探索性数据分析的过程中，有四大主题：耐抗性（Resistance）、残差（Residuals）、重新表达（Re-expression）及图形启示（Revelation）。

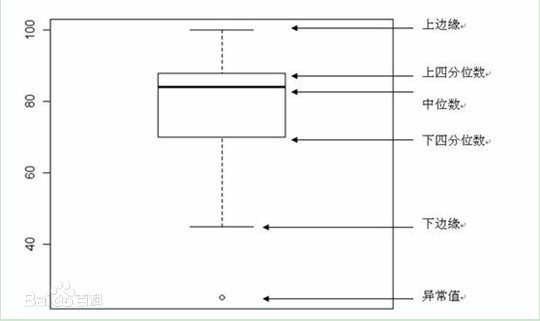
耐抗性（Resistance）是指，对于数据的野值（离群值）不敏感，野值不会对于分析结果有过大的影响。即使用与原数据差距极大的数据替换原有一小部分的数据，分析结果也不会有明显的变化。5数总括值很好的体现了EDA的耐抗性。为说明5数总括值，先介绍深度的概念：在一组有序数据中，某一数据的深度为其到数据两端的距离的最小值。5数总括值既中位数（Median）、4分数（Furth）、极端值。他们在数据中的位置又其深度决定。中位数：数据样本的中心；其深度为（n+1）/2，当n为奇数时，中位数为数据批中间的数，当n为偶数时，中位数为数据批中间两个数的平均值。极端值：是数据批的两个端点值，深度分别为1和n。4分数：4分数分为上4分数Fu和下4分数Fl，分别位于上极端值和中位数以及中位数和下极端值的中间，其深度分别为（n+3）/4、（2n+1）/4。箱型图（如图 3.1-2 箱型图）是5数总括值的一种图形化表达，广泛用于探索性数据分析。

图 3.1‑2 箱型图

残差（Residuals）是指，原始数据减去一个总括统计量或者拟合结果后的残余部分。残差作为数据偏离程度的重要指标，不仅在探索性数据分析过程中有着重要意义，也是对与聚类好坏的评价指标之一。残差分析提供了一个放大镜，使得模式和基础结构很容易被分析者观察到，为进一步的分析，提供了方向。

重新表达（Re-expression）是指，运用何种尺度能简化分析。比如对于跨度过大的数据取对数等操作缩短其尺度使数据更加集中便于观察。探索性数据分析强调：应该尽早地考虑尺度是否合适得问题。尺度得选择影响着对称性、关系直线性等性质。

图形启示（Revelation）是探索性数据分析的重要方法，通过画图给予分析者以启示作用。图能形象的反映数据、拟合以及残差等行为，从而更加容易抓住数据的特点。常用的图形有茎叶图、箱型图、小提琴型图以及传图统计图形如柱状图、散点图等。

正是因为探索性数据分析注重直观和图形显示，随着科技的不断进步，家庭电脑的普及，各种可视化技术的高度发展如matplotlib、seaborn、plotly等库，探索性数据分析越来越有优势，即使不是专业的统计分析者也能够通过python、spss等轻松探寻残、拟合中的奥秘，获得意想不到的发现，为其所在领域的分析提供启发性思想及发展方向。

## 基础探索性数据分析

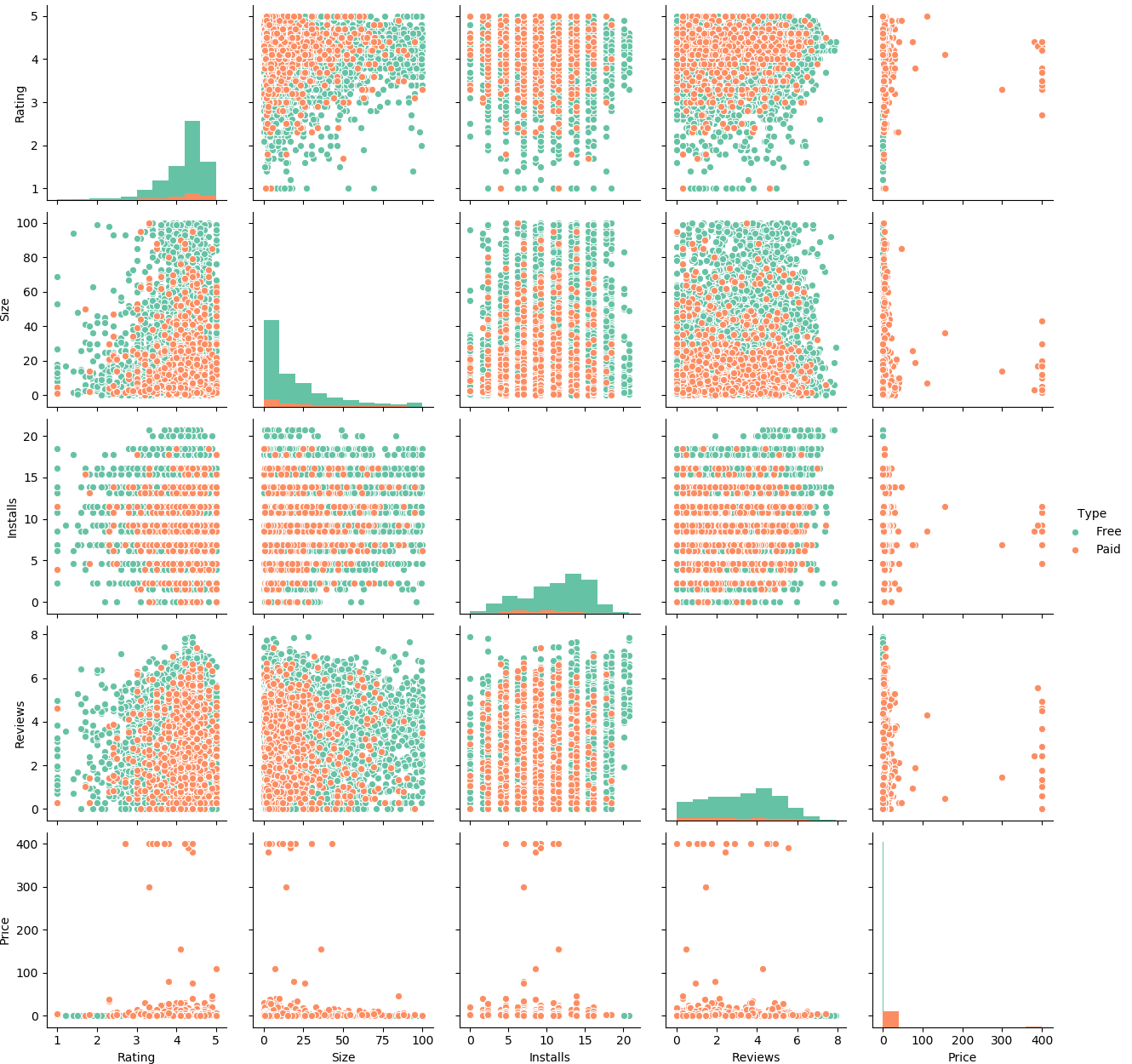
在数据分析的起始阶段，我们对于数据的特点一无所知。虽然通过属性列的名称，我们能够对数据有一点的了解，但直接提出假设，不符合探索性数据分析的分析方法。因此，我们要对数据整体有个总体的观察。我们用seaborn库中的pairplot类对各属性两两间的关系进行可视化，以寻找特征之间的明显模式或关系，如图3.2-1 basic EDA。其中对于Installs和Reviews取对数，改变了其尺度，不然他们的尺度过大，不适合绘图。

图 3.2‑1 basic EDA

首先，是对角线的柱状图，x轴是各个属性，y轴式样本数量，可以看出5种属性样本数量的分布情况，与正态分布相去甚远，所以难以用传统的统计模型进行假设的验证与分析。对于此数据集探索性数据分析是更好的分析方法。

从Rating的分布图来看，无论付费应用还是免费应用总体都有较高的评分，可想而知，目前安卓市场中的应用都是十分成熟的应用，普遍都是能使用户满意的。再结合Rating-Installs散点图，能够发现评分在4以上的应用，下载量却有高有低，说明评分和下载量不线性相关。应用技术的普遍成熟，使得评分不再是用户选择一款应用的最主要的指标。

从Rating-Size图中我们可以发现付费应用大量分布在整个图的右下角，对此我们可以提出零假设：对于付费应用，软件的大小会的对评分产生影响，且较小的应用会有较高的评分。这个假设不仅会用EDA分析，还会在之后的聚类分析中验证。

从Rating-Price图中，传统的散点统计图被离群值干扰严重，因此我们还需通过过滤离群值和箱型图的方法进一步观察价格对于评分的影响。

## 应用品类策略

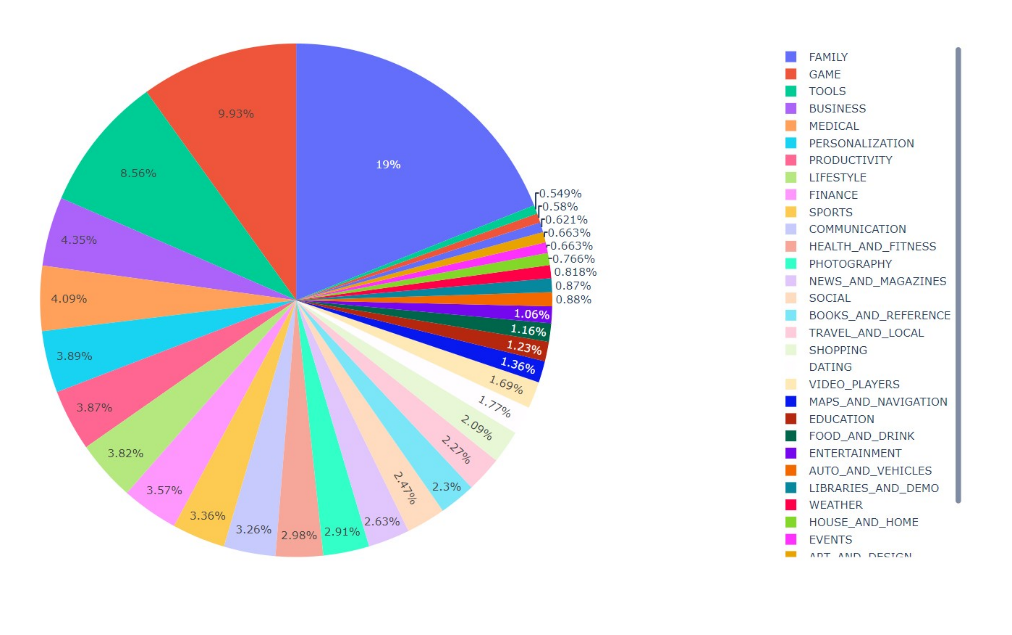
用饼图统计各个分类的应用数量。从图3.3-1 可知，最受欢迎的应用分别是FAMILY、GAME、TOOLS、BUSINESS、MEDICAL，分别占了19%、9.93%、8.56%、4.35%和4.09%。

图 3.3‑1 各应用分类数量统计

图 3.3‑2 各分类应用数量

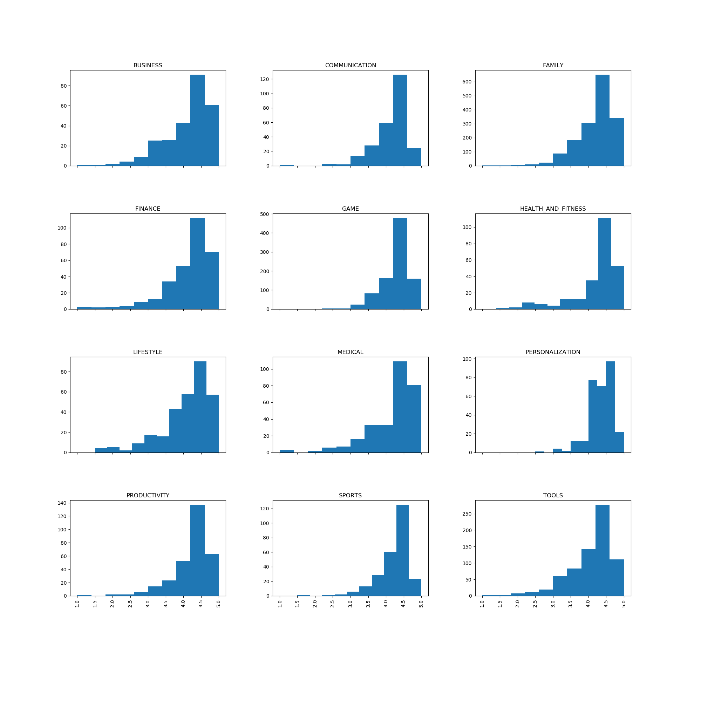
为了探究应用的分类是否会影响评分，我们统计前几的应用的评分分布并绘制柱状图。

图 3.3‑2 评分分布

对于单个因素对实验指标的影响是否显著，传统统计方法是通过单因素方差分析，但是其先决条件是数据符合正态分布，从图3.3-2可以明显发现，所有数据都是左偏态分布。这时，将多个分类的评分分布直方图放到一块的探索性数据分析就能明显的体现出，各个分类的评分有着显著的差异。

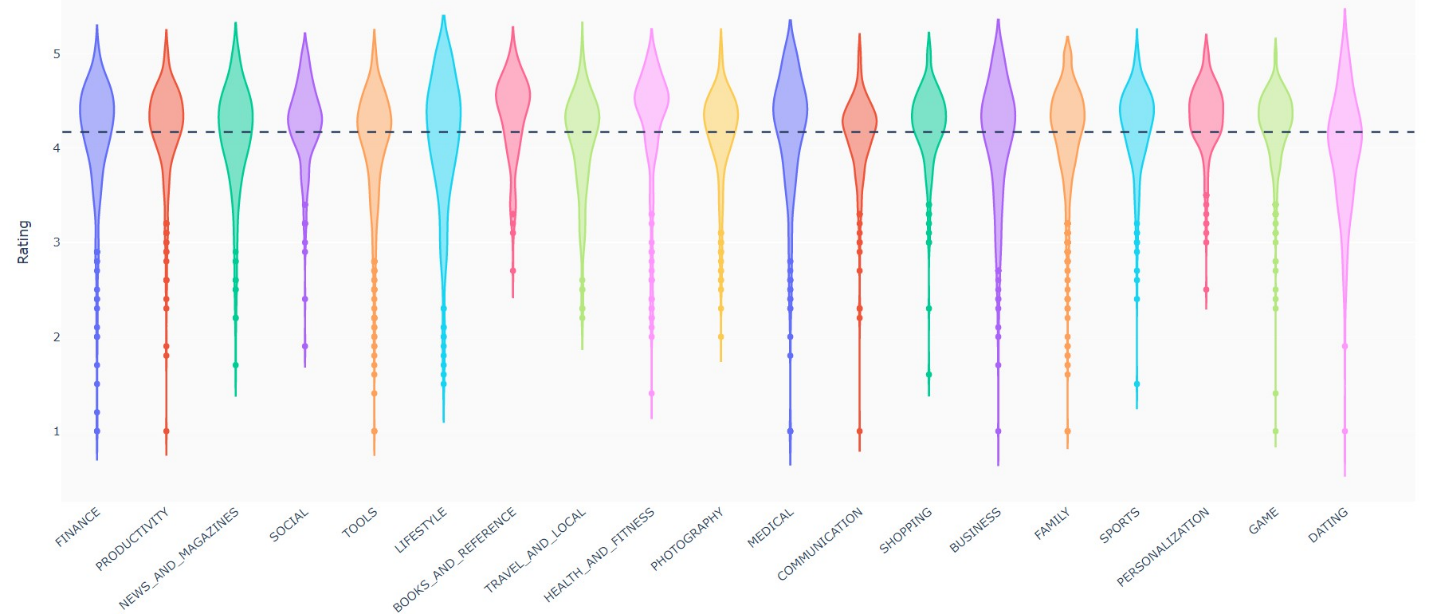
为了进一步分析，哪种分类最受用户喜爱，我们通过seaborn中的violin类，绘制各分类的小提琴型图。小提琴型图的宽度代表该评分下的样例频率，最下面的点是离群值，中间的凸起部分，分别是上4分数Fu和下4分数Fl。

图 3.3‑3 分类-评分小提琴型图

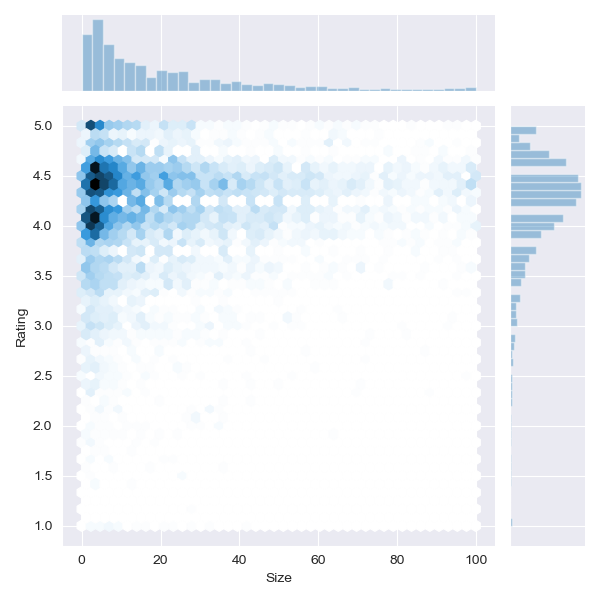
从图3.3-3中我们可以得出这样的结论：几乎所有类型的应用都有不错的表现，其中Health and Fitness和Books and Reference表现的比较突出，50%以上的应用都有着4.5分以上的分数。相反的，Dating分类的应用有50%的应用在平均分以下。安卓应用开发商在选择开发哪种分类的软件时，可以参考图3.3-3的统计结果。

图 3.4‑1

## 应用大小策略

我们在原有的Rating-Size图的上方和右侧附上Rating和Size的数量分布直方图，如图3.4-1，以便更加直观的表现评分和应用大小之间的特征关系。图中颜色的深浅代表着密集程度。很直观地就能发现40M以下，4分到4.5分之间聚集着大量应用。

从目前的情况看来，用户大多在自己的手机里下载几十甚至上百个应用，在硬件条件有限的情况下轻量级的应用理应更加受到用户们的青睐。但是随着硬件技术的不断提升，手机的存储256G、512G都是各品牌最基础的配置了，再加上网速的不断提升，人民不会因为应用大小过大，而导致下载等待较长的时间，为了更为丰富的内容，应用大小的变大也是能够接受的。同时5G、wifi6的到来，为云应用提供了可行的必要条件，未来的移动端，也许更加注重显示效果、震动反馈、网络连接技术等，对于移动端本身的运算要求降低，应用的大小也不会再有所上涨。

## 应用价格策略

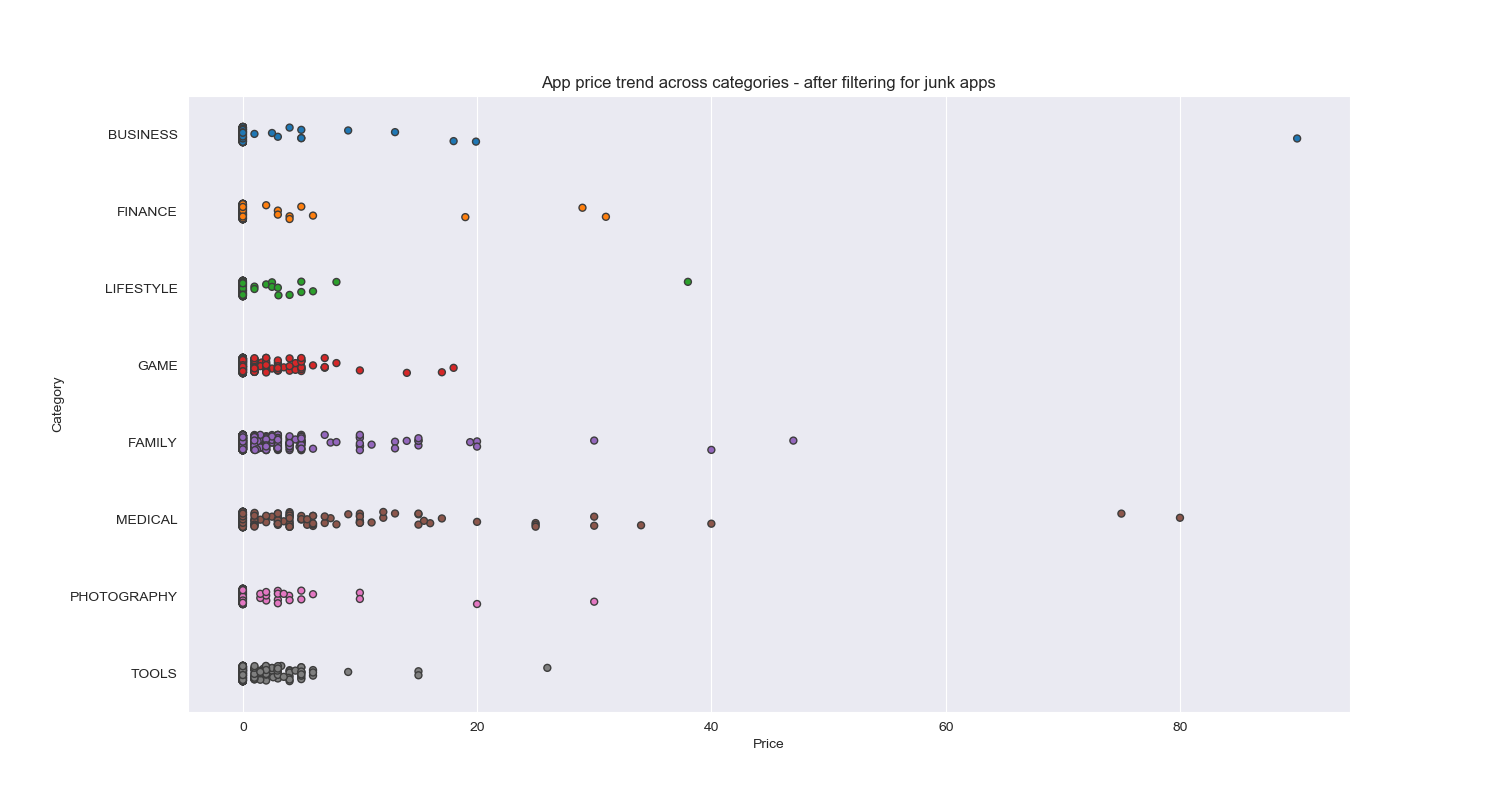
从图3.2-1 Basic EDA中我们可以看到，只有很少一小部分的应用价格超过了$100，但是这少部分的应用会影响到整个价格区间的分布，所以我们将这部分值作为离群值筛除。之后绘制Price-Category散点图3.5-1。

图 3.5‑1

从图3.5-1中我们可以看出MEDICAL、FAMILY类别的付费应用价格整体偏高，但是GAME类别的均在$20以下。从我们比较熟悉的游戏类别角度看，移动端应用的价格和电脑端的应用比起来更为低廉，这个受限于移动端硬件水平和操作方式。不少pc端的游戏移植到移动端，表现型和游戏体验都会有所下降，价格也随之下降。倘若用户想要购买专业性强、价格高的应用时，首选的肯定是pc端而不是移动端。因此，应用开发商想要让自己的应用有一个高定价，就必须使他们的产品有着移动端独有的特色并且使pc端难以做到的特色。

然而相较于付费应用，免费应用才是应用市场的主流。如图3.5-2所示，图中的纵坐标是对10取对数后的结果。

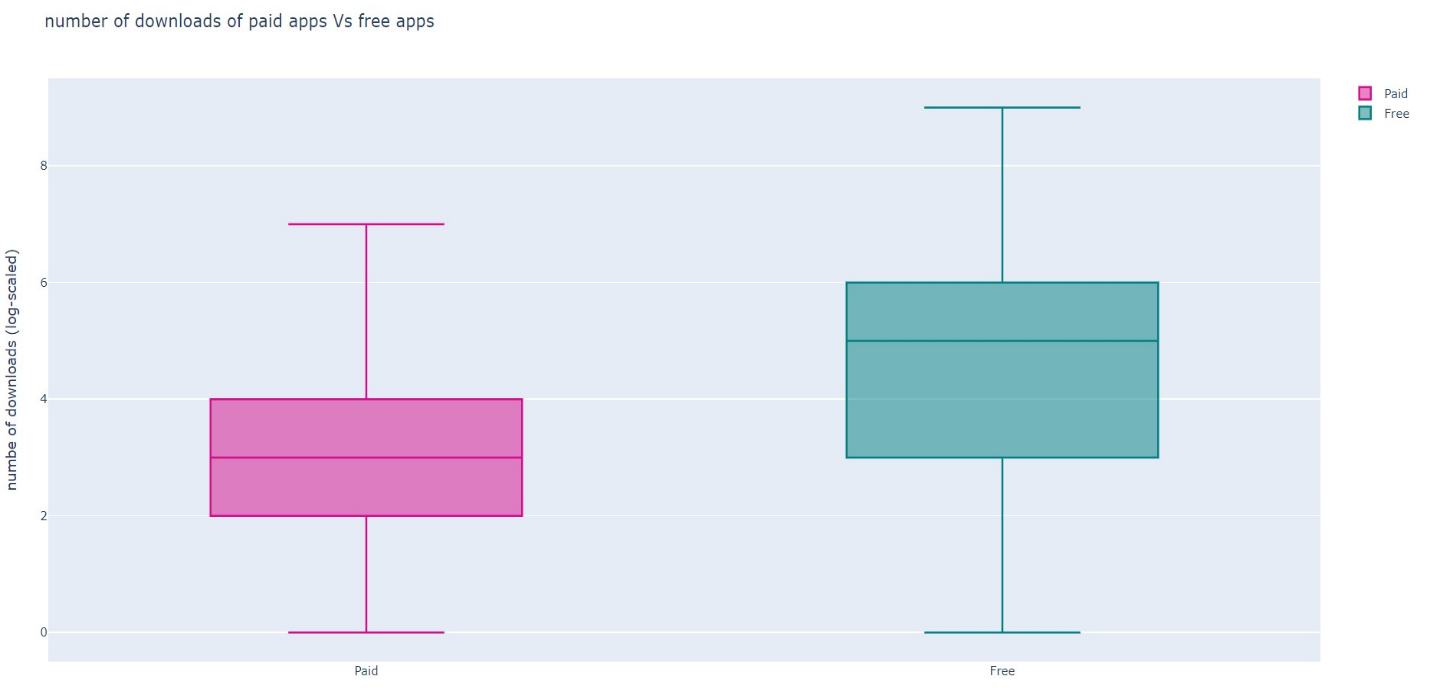


图 3.5‑2

我们对比上图的箱型图的4分数可以清楚的看到，付费软件的下载量主要集中在2和4之间，免费软件的下载量主要集中在3到6之间，因为纵坐标对10取了对数，所以纵坐标相差1，就是相差了一个数量级，免费软件远比付费软件更受用户的喜爱。以我们比较熟悉的视频软件为例，应用免费但是各个级别的vip，以及超前付费点播是较为主流的付费模式。涓涓细流的收费模式，给用户更多的选择的权力，也给依附于各个应用的内容作者更多的动力。

# 聚类分析



## 算法简介

聚类分析（clustering），无监督学习的一种，是作为人工智能的重要领域之一，被广泛用于机器学习、模式识别、数据挖掘等各大领域。聚类的目的是寻找数据内部的结构，聚类会根据样例的相似程度将样例归类到一起，形成簇，簇内的样例具有较高的相似性，相反簇间的样例相似性较低。并且用簇内相似度的总和还评价聚类的效果。

相似度则用距离这个概念来衡量，距离近的对象会被划分到同一簇中。对于距离的计算，科学家们给出来许多计算方式，如欧式距离、曼哈顿距离、余弦相似性距离等等。其中欧式距离是最简单也是使用最为广泛的一种。对于点X（坐标为（x1,x2,..,xn））和点Y（坐标（y1,y2,…,yn）），其欧式距离定义如下：

Kmeans聚类是聚类方法中使用最为广泛的一种，它能够发现数据集的k个簇，Kmeans中的means是指这k个簇的中心是由每个簇的均值决定的。Kmeans无需监督，无需训练集，简单且解释性好，但是因为初始中心是随机选择的，容易造成局部最优，同时Kmeans对离群点十分敏感。

为了防止局部最优，David Arthur 提出Kmeans++算法，对随机选择初始中心造成的弊端给予改进。Kmeans++通过计算概率的方式选择初始聚类中心，当选择了一些初始聚类中心后，离这些点较远的点会有更高的概率被选为下一个初始聚类中心。通过这样的方法，使初始聚类中心更为分散，而不会扎堆，以达到全局最优。

## 算法原理

选取初始聚类中心：

Step1：随机选取第一个聚类中心C1，加入聚类中心集合C

Step2：计算数据集中每个点到C中所有聚类中心的距离，取最小值，作为该点到C的距离D(x)，计算该店被选为聚类中心的概率

Step3：用轮盘法选取下一个聚类中心加入C，重复Step2，知道C中有k个聚类中心

Kmeans：

Step1：选取初始聚类中心，聚类中心集合C

Step2：计算数据集中每个点到k个聚类中心的距离，选最近的聚类中心，加入该簇。若所有点都保持原有簇，则跳出循环。

Step3：用簇中的所有点的平均值更新该簇的聚类中心，重复Step2。

## 算法应用

经过之前的一系列探索性数据分析，我们发现Raing、Size、Price之间存在着一定的关系，所以我们可以提出这样的假设：应用的大小和价格在一定程度上会影响用户对于该应用的评分。我们可以通过 Kmeans++对Rating、Size、Price组成的数据集进行聚类，并将聚类完成的结果用3D的散点图绘制出来，直观的检验我们的假设。

在进行聚类之前，我们发现这三个特征值的单位不同，使用欧式距离计算时量纲较大的特征值就有较大的权重，但我们认为这三个特征值应该有相同的权重，如图4.3-1所示，Size的量纲远大于其他两个特征值，聚类的结果就相当于将Size划分为三个区间，Price的影响过小。因此需要对数据进行归一化，归一化不会影响数据的分布情况，同时对三维都进行归一化不会影响各特征值之间的相关性。对于特征值X（x1,x2,…,xn）归一化公式为：

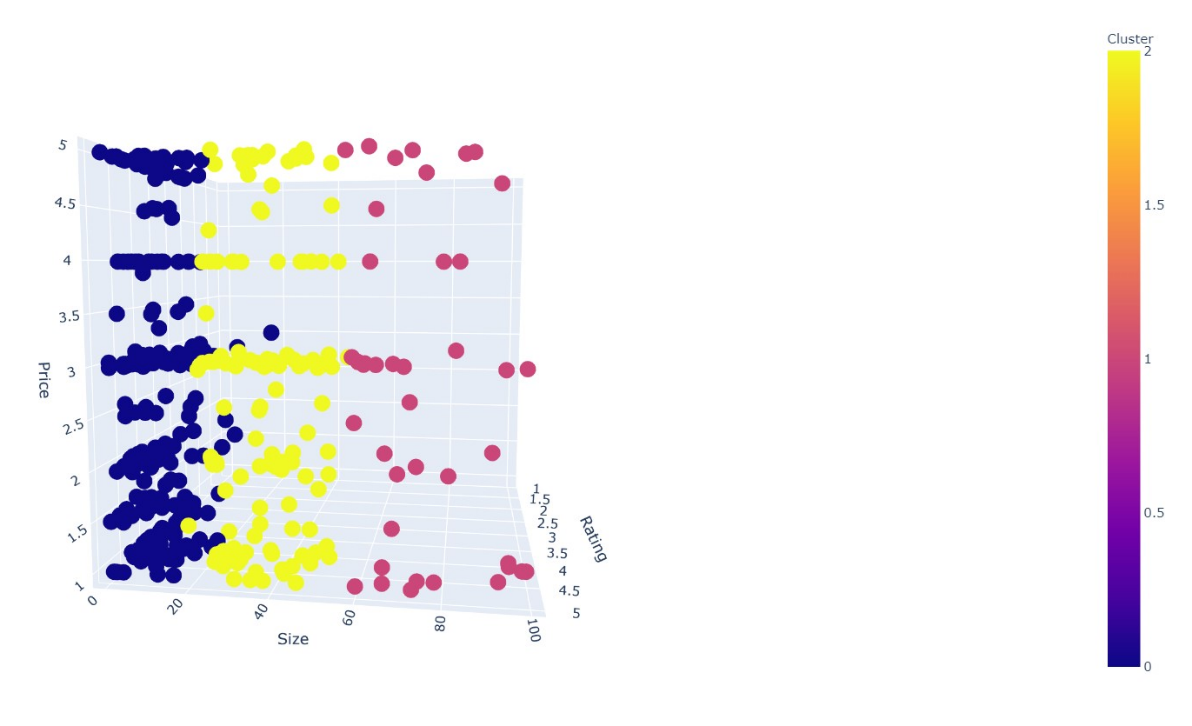


图 4.3‑1 未归一化聚类

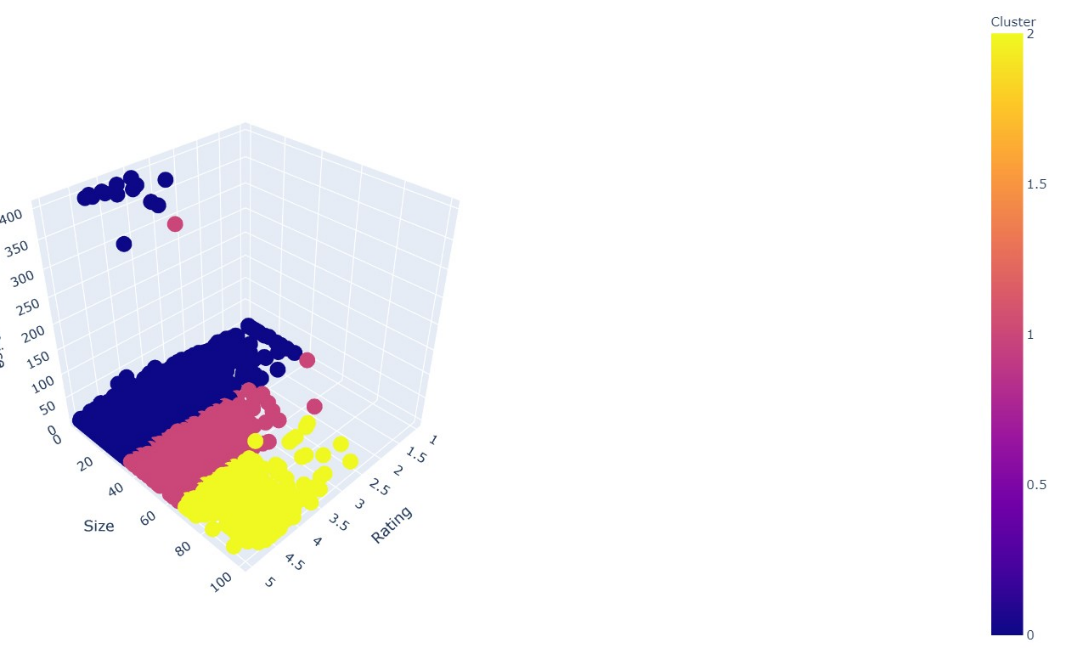
Kmeans聚类对于离群值十分敏感，如图4.3-2所示，离群值“压迫”了正常值，使聚类效果不佳。

图 4.3‑2 存在离群值

对离群值进行筛除，付费应用中75%的应用都在$5以下，所以我们将$5以上的样例筛除。

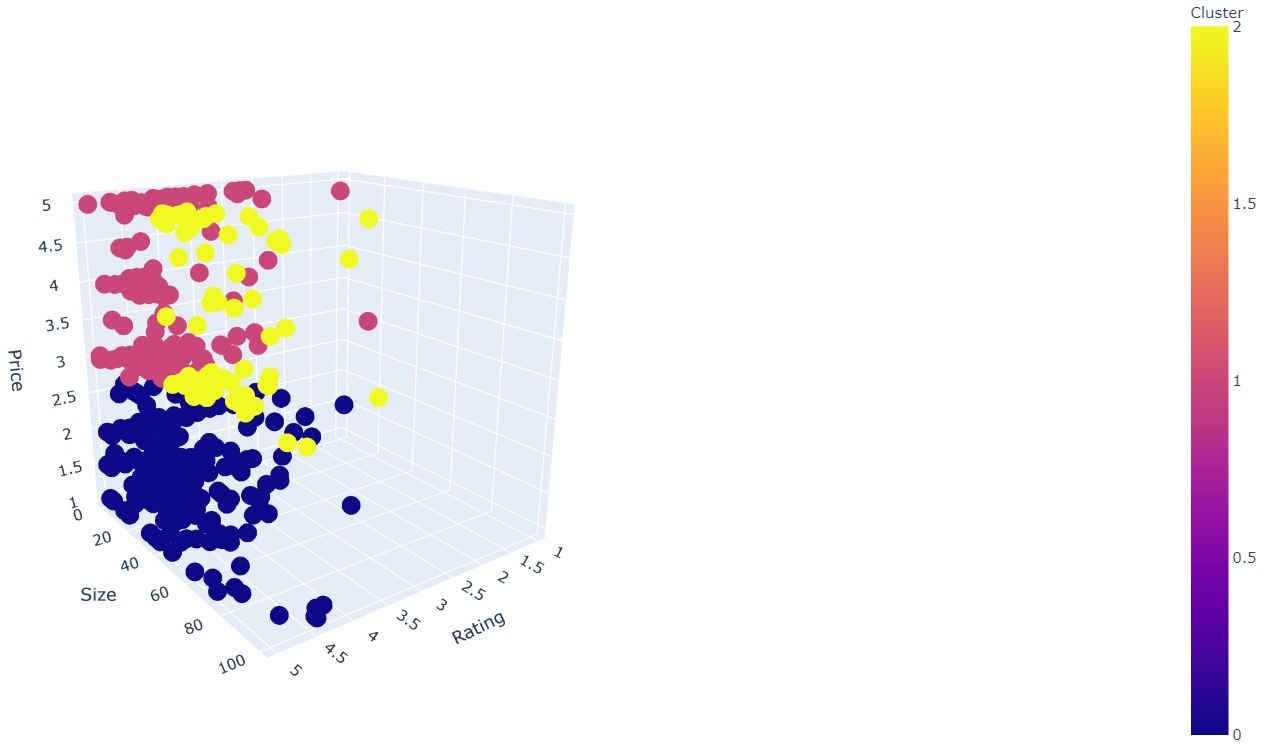


图 4.3‑3

图4.3-3 是用plotly库绘制，能在网页上进行交互，通过旋转3D散点图我们能够非常直观的看到黄颜色标记的簇集中在高价格、高存储、高评分的区间，而价格高、应用大小比较小的红色簇在评分的分布就良莠不齐，低价格的蓝色簇，在评分和大小的分布都更为的分散。这表明高价格、应用大小较大在一定程度上能够保证该应用的质量。与之相对的蓝色簇虽然有大量的高评分聚集，但评分极端低的也同样出自蓝色簇。可以见得，在低价格甚至免费的应用中，存在着许多成本低廉的垃圾软件。

## 聚类分析总结

上文并未提到K值得选择，因为缺乏严格的数学推导，学者们提出了多种启发式的、贪婪的K值选取的方法。其中较为主流的是手肘法，手肘法通过SSE（sum of the squared errors，误差平方和）作为指标，随着K值得变大，SSE会逐渐减小，当K增大到一点程度时，SSE减小得幅度会明显变小，此时得K为最优得K值选择。然而，Kmeans聚类算法因其初始聚类中心时随机选择的，算法具有不稳定性有时SSE并不会因为K值的增大乖乖减小。因为本次数据分析的数据量并没有特别大，当K = 3时，我们已经能够看出各簇的特征，因此我们认为K = 3是较好的选择。

通过此次实践，在深入学习聚类算法的过程中，让我对机器学习有了一点程度上的了解，对于分类、回归、聚类、频繁项集、降维工具等进行了简单的学习。在阅读文献的过程中，认识到了机器学习与多个领域的结合，在多个领域的重要作用，从中我还明白了一些道理：

1. 本科的学习实在是太过浅薄，还有大量的知识需要我们去学习和深入。
2. 随着科技的不断发展，对于我们年轻人的要求会越来越高，当短短三行代码就绘制出聚类结果，并有着非常好的交互时，时代和社会对于我们的要求就不再是如何在网页上画出一个3D立体图，而是用这个3D立体图去创造更高的价值。

# 参考文献

[1] Olivier B G,Rohwer J M,Hofmeyr J H S. Modelling cellular processes with Python and Scipy.[J].  Molecular biology reports. 2002(1-2)

[2] 黄琪.基于Python的数据可视化方法和系统实现[J].信息与电脑(理论版). 2019(14)

[3] 叶惠仙.Python 在学院招生数据分析中的应用 [J]. 计算机时代,2018(11):102-106.

[4] 杨昆,汪兴东.Python程序员指南[M].中国青年出版社, 2001

[5] 余肖生, 周宁, 张芳芳. 高维数据可视化方法研究[J]. 情报科学, 2007,25 (1): 117-120.

[6]张庆丰,王锋.探索性数据分析方法在成绩分析中的应用[J].安阳工学院学报,2011,10(04):102-103+122.

[7]高文秀,朱俊杰,侯建光.探索性数据分析在土地利用数据分析中的应用[J].武汉大学学报(信息科学版),2009,34(12):1502-1506.

[8]孙丽君. 探索性数据分析方法及应用[D].东北财经大学,2005.

[9]苟成玲.探索性数据分析方法的应用——沪深股市股票价格与流通盘的定量关系[J].统计与信息论坛,2003(02):26-29.

[10]王淑红,李英龙,戈保梁,李文瑶.探索性数据分析在选矿中的应用[J].金属矿山,2002(07):22-25.

[11]陈忠琏.《探索性数据分析》[J].中国统计,1997(06):47.

[12]鲁道刚,李强,孙佳.基于改进K-means和簇结构优化的非均匀分簇路由算法[J].传感器与微系统,2020,39(04):144-147.

[13]卞永明,高飞,李梦如,李乔,马逍阳.结合Kmeans++聚类和颜色几何特征的火焰检测方法[J].中国工程机械学报,2020,18(01):1-6.

[14]杨俊闯,赵超.K-Means聚类算法研究综述[J].计算机工程与应用,2019,55(23):7-14+63.

[15]柏宇轩.Kmeans应用与特征选择[J].电子技术与软件工程,2018(01):186-187.

[16]程艳云,周鹏.动态分配聚类中心的改进K均值聚类算法[J].计算机技术与发展,2017,27(02):33-36+41.

[17]刘越. K-means聚类算法的改进[D].广西师范大学,2016.

[18]章宦记.改良的kmeans与K近邻算法特性分析[J].电子产品世界,2016,23(01):79-80.

[19]赵莉,候兴哲,胡君,傅宏,孙洪亮.基于改进k-means算法的海量智能用电数据分析[J].电网技术,2014,38(10):2715-2720.

[20]王千,王成,冯振元,叶金凤.K-means聚类算法研究综述[J].电子设计工程,2012,20(07):21-24.

[21]车丽美,肖洋,王甦易,姜倩倩.Kmeans聚类分析在形音字表音度中的应用[J].计算机技术与发展,2011,21(02):223-225+233.

[22]陈寿文,李明东.基于面向对象思想KMeans算法实现[J].滁州学院学报,2008(03):42-44.

# 结束语

通过此次设计，不仅熟悉了python语言，并且通过查询API reference 的方式熟悉pandas、numpy、scipy包中函数的用法。学习了概率论中的方差、协方差、相关系数等知识。

# 致谢