**京东商城物价分析**

**学生姓名 李博**

**学 号 2020080211**

**专业班级 数据科学与大数据技术2002班**

**日期**  **2022年5月3日**

# 摘要

......老师说随便写的小论文还写什么摘要，直接看正文就完事 ->

选题介绍都不用看，直接看第二章就ok

[第一章 选题介绍 - 1 -](#_Toc25548)

[1.1 选题背景 - 1 -](#_Toc27502)

[1.2 爬虫介绍 - 1 -](#_Toc16508)

[第二章 数据来源介绍与数据分析 - 3 -](#_Toc1766)

[2.1数据来源介绍 - 3 -](#_Toc1687)

[2.1.1原始数据展示 - 3 -](#_Toc25508)

[2.1.2数据处理结果展示 - 3 -](#_Toc31057)

[第三章 涉及方法介绍 - 5 -](#_Toc11837)

[3.1 网页爬虫技术爬取数据 - 5 -](#_Toc30801)

[3.1.1 网页爬虫 - 5 -](#_Toc16253)

[3.1.2 商品信息网页爬虫 - 5 -](#_Toc6635)

[3.1.3 商品信息数据清洗 - 5 -](#_Toc23102)

[第四章 数据分析 - 7 -](#_Toc19488)

[4.1 数据分析 - 7 -](#_Toc14555)

[4.1.1数据展示 - 7 -](#_Toc23219)

[第五章 讨论和心得 - 9 -](#_Toc28542)

# 选题介绍

### 选题背景

如今，我们生活在一个伟大的互联网时代，电商平台的兴起是互联网时代的典型代表之一，虽然电商平台严重的冲击了作为市场主流的线下实体经济，但它也为人们的生活提供了极大的便利。在为人们提供便利的同时，各种各样的电商平台，繁多、复杂的商品信息，也给享受了网络购物便利的人们造成了一定的困扰，所以本篇文章写这个程序的目的就是为了让大家可以更容易的各种商品间信息的对比，让人们在购物时有更好的选择。同时，分析的结果也可以给想做电商的同学提供一些建议和意见。

在人们生活的互联网时代,电商平台为人们的日常生活提供了极大的方便。但是电商平台里面繁多、复杂的购物信息，给人们在购物的同时带来了干扰。所以本篇文章写这个程序的目的就是为了让人们看到某个具体商品的详细信息以及商品之间的对比信息，让人们在选购的同时能有更多好的选择。同时分析之后的结果还可以给电商团队,给他们带来相对客观的结果。也可以说，人们所在的时代是一个大数据时代。随着电商平台的快速发展以及逐渐的完善，随之而来的是人们消费而产生的大量购物数据，各个公司都希望能够通过数据挖掘以及数据分析找到潜藏在这些数据中的隐藏价值，为公司的发展和经营以及市场竞争提供有价值的信息和正确的支持。

网站的信息是不会说话的，但是其中具有大量的研究价值等待发现和探索。怎样把电商平台的信息与数据挖掘和数据分析结合起来，是现在社会值得去讨论和研究的领域。本文通过对电商平台网页信息的抓取与相关信息的研究，帮助指导相关公司做出客观正确的决定,从而提高其在市场中的竞争力。

### 爬虫介绍

网络爬虫作为如今搜索引擎非常重要的一部分，其担当的角色是抓取网页或网站数据和内容并且保存下来，在以前，就有很多程序员和研究员开始从事网络爬虫的研究。目前，网络爬虫技术日趋成熟，相当多的搜索引擎核心就是爬虫。在爬虫开发中，python是最受欢迎的语言。Python 在网络爬虫开发中具有独特的优势,其丰富的开发库和大量优秀的代码使得 python成为爬虫的助理开发语言

网络爬虫技术发展到现在,已经有多如牛毛的开发库和现成的技术供人们使用，比如多线程和 requests库就是其中的几种,这篇论文就是基于它们实现的。在这么多的爬虫开发程序员眼中，公认的难题已经不是开发周期这个问题。在网络迅速发展的当下,网页的更新速度极为地迅速。在这特殊的发展背景下,网络爬虫是否能够检测到网页的更新并且采取相应的对策变成判断这个程序是否合格的指标。其次，版权意识越来越深入人心，网页作为知识的一个相当重要的载体,在某种程度上，网页里的信息也在知识版权的范围，应该受到法律的保护。所以有很多的网页不希望里面的知识被爬虫程序抓取下来,于是在程序员设计网页的时候,会特别关注这方面的设计,以防止爬取。当然纳,有相当一部分的爬虫程序不是出于赚钱的目的而开发出来的,其开发出来的意义是为了让人们免费的使用。所以开发人员就需要在法律和不违法之间做一个平衡,这样就可以让爬虫程序更好的为人民服务。

近年来,通过各学者的分析研究，已解决了基于技术的网络爬虫关键问题,实现了对动态网页信息的抓取。基本的实现方法是模拟用户动作,触发JavaScript事件，解析网页中的脚本，通过网页的树来获取信息，虽然，这些方法去了一些成果，但是仍然存在很多问题，比如，怎样提升网络爬虫的性能等，因此，基于技术的网络爬虫还需要进一步的研究。

# 数据来源介绍与数据分析

### 2.1数据来源介绍

本次作业的数据主要来源与京东商城（www.jd.com），商品类型方面主要选取了上一年度即2021年发布的搭载了“Nvidia GeForce RTX 3060 Laptop”显卡的主流游戏本（京东自营），选取该类商品的主要原因为本人对该方面产品较为了解，以及最近有换新机的想法，能更好的对数据进行分析和理解。

### 2.1.1原始数据展示

以下为本次作业爬取的商品具体型号及url地址，涉及不同层次的品牌及型号，一线品牌主要包含联想、华硕……,以及二线及一些三四线品牌：红米、神舟、七彩虹……

r9000p = 'https://item.jd.com/100014794825.html'  
 r7000p = 'https://item.jd.com/100019736872.html#crumb-wrap'  
 tx2 = 'https://item.jd.com/100010367173.html#crumb-wrap'  
 g15 = 'https://item.jd.com/100012622295.html'  
 redmiG = 'https://item.jd.com/100026497910.html#crumb-wrap'  
 z8\_ta5ns = 'https://item.jd.com/100021707404.html'  
 x15 = 'https://item.jd.com/10032218944060.html'  
 TR911zero = 'https://item.jd.com/100012544831.html'  
 ttPlus = 'https://item.jd.com/100021647694.html'



图-1 联想r9000p京东商城页面

### 2.1.2数据处理结果展示

本次爬取的数据主要为Json类型,经过python处理为DataFrame数据，最后通过.to\_csv()方法将数据存入csv文件中，方便后续处理。

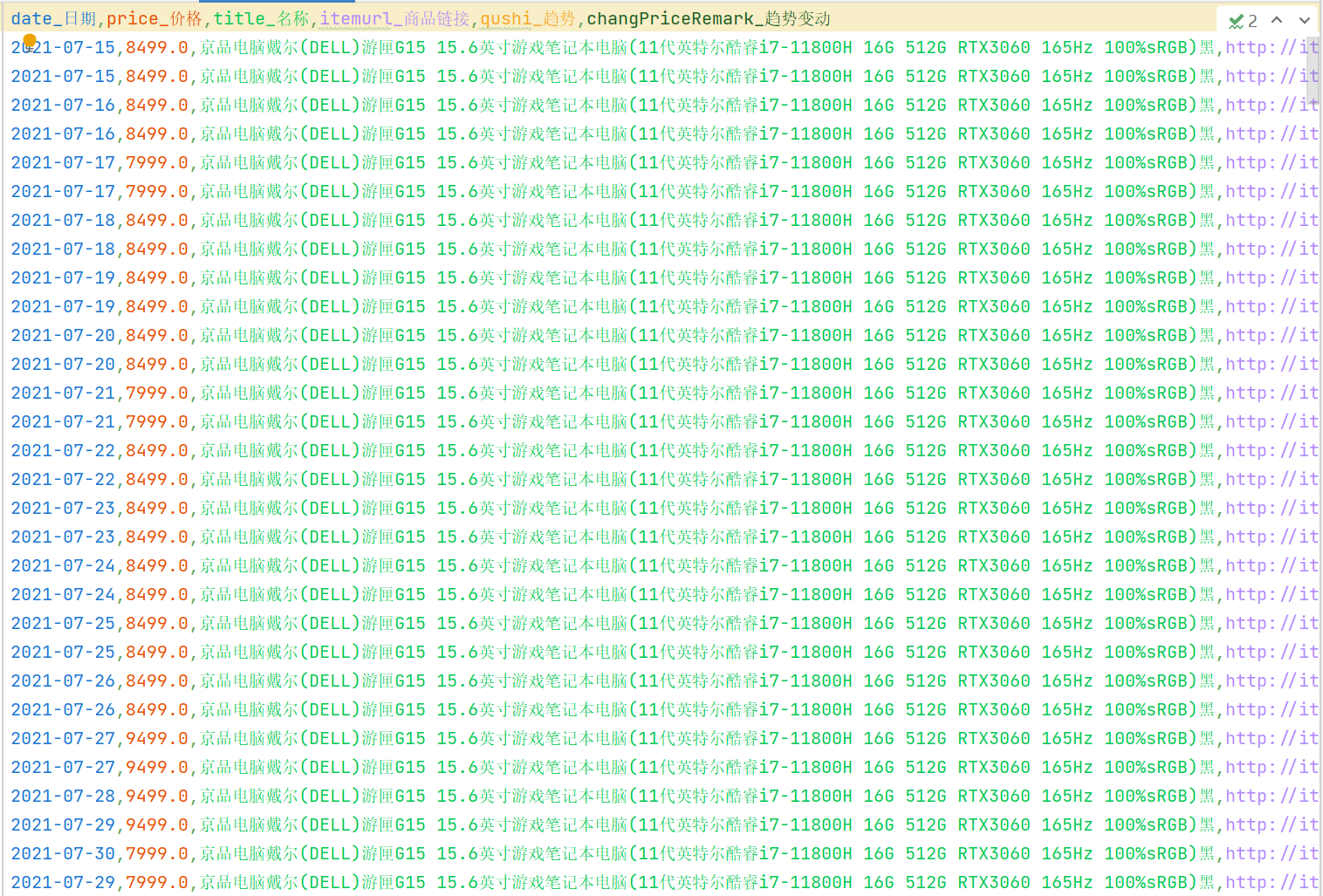


图-2 戴尔g15 5511 csv文件内容

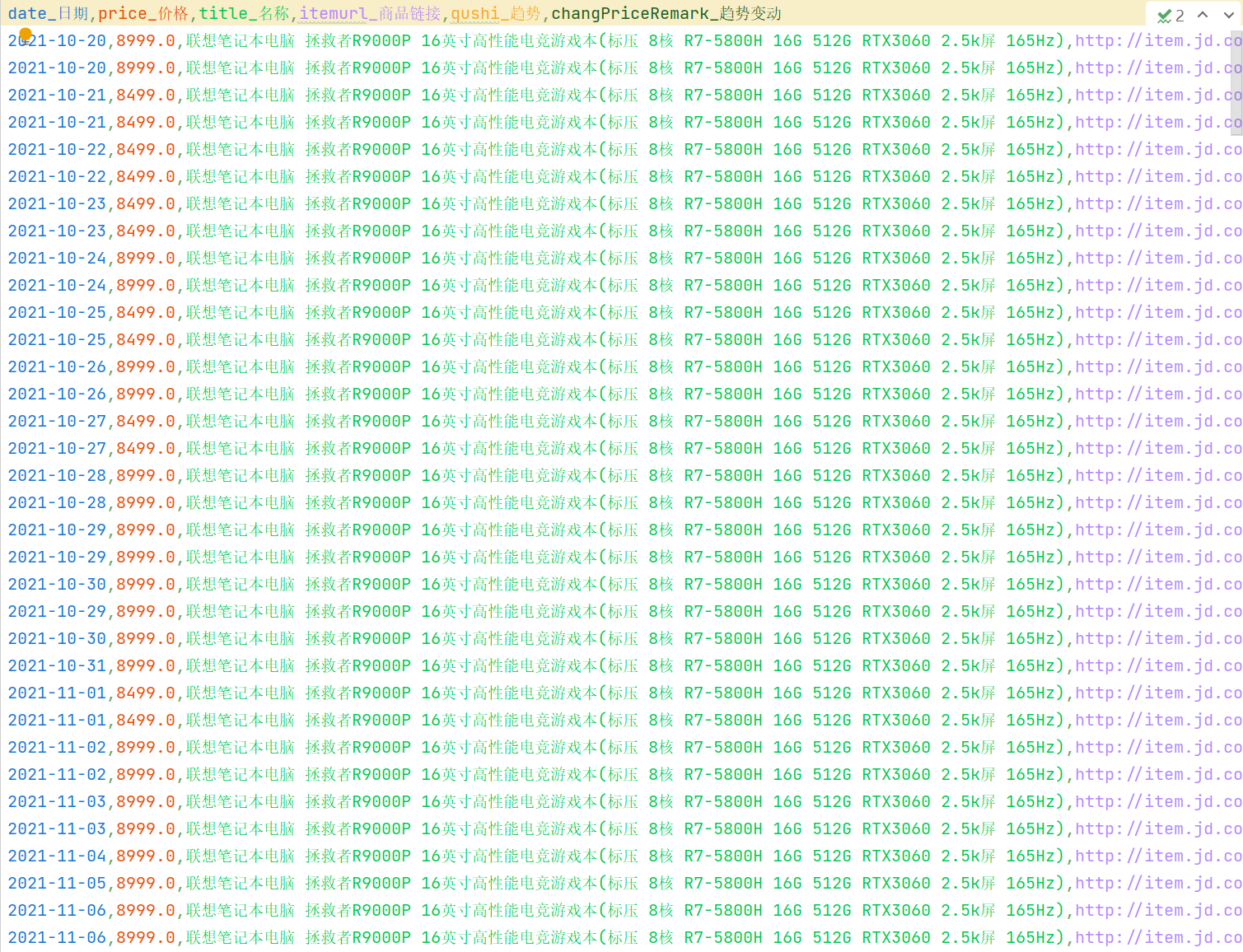


图-3 联想r9000p csv文件内容

# 第三章 涉及方法介绍

### 网页爬虫技术爬取数据

### **网页爬虫**

在涉及网页数据时主要使用的爬虫技术用于下载，找到网页特定的数据所在段，经过网页数据分析，利用python的request模块；或可将网页json数据直接下载保存到本地json格式，便于后续调用。

### 3.1.2 商品信息网页爬虫

由于京东网页上并无以往价格信息，所以借助第三方比价网站（[慢慢买——购物比价网 (manmanbuy.com)](http://manmanbuy.com/)）来获取商品的历史价格和其他信息，该网站请求方式为post请求，爬取过程较为简单，主要是将url字符串拼接然后请求，之后就是数据处理了。



### 3.1.3 商品信息数据清洗

Request请求到的为一段很长的json数据。

部分实例：【{'ok': 1, 'msg': '', 'single': {'id': '0', 'title': '京品电脑机械革命钛钽PLUS 17.3英寸165Hz电竞游戏笔记本电脑(11代英特尔酷睿i7-11800H 16G 512G RTX3060 2.5K)', 'smallpic': 'http://img14.360buyimg.com/n7/jfs/t1/36967/37/16701/133777/626d17faEd9a583f6/3a7c7f171518712e.jpg', 'bigpic': 'http://img14.360buyimg.com/n7/jfs/t1/36967/37/16701/133777/626d17faEd9a583f6/3a7c7f171518712e.jpg', 'jiagequshi': '[Date.UTC(2021,4,11),9999.00,""],[Date.UTC(2021,4,12),9999.00,""],[Date.UTC(2021,4,13),9999.00,""],[Date.UTC(2021,4,14),9999.00,""],[Date.UTC(2021,4,15),9999.00,""],[Date.UTC(2021,4,16),9999.00,""],[Date.UTC(2021,4,17),9999.00,""],[Date.UTC(2021,4,18),9999.00,""]......】

格式为[时间，价格]以及开头和结尾的一些其他数据，在数据清洗中使了.split .replace 切片等基本操作，最后利用pandas将数据写为了dataframe形式，再利用to\_csv、to\_excel将数据写入了csv文件和excel表格中方便后续使用。



图片-4 csv、excel文件

# 第四章 数据分析

### 数据分析

数据分析时主要使用了pyplot中的plt.plot（）进行了图标的绘制，使数据更为清晰明了，以及excel中的一些统计函数对数据进行了统计。



图-5 折线图绘制代码部分

### 4.1.1数据展示

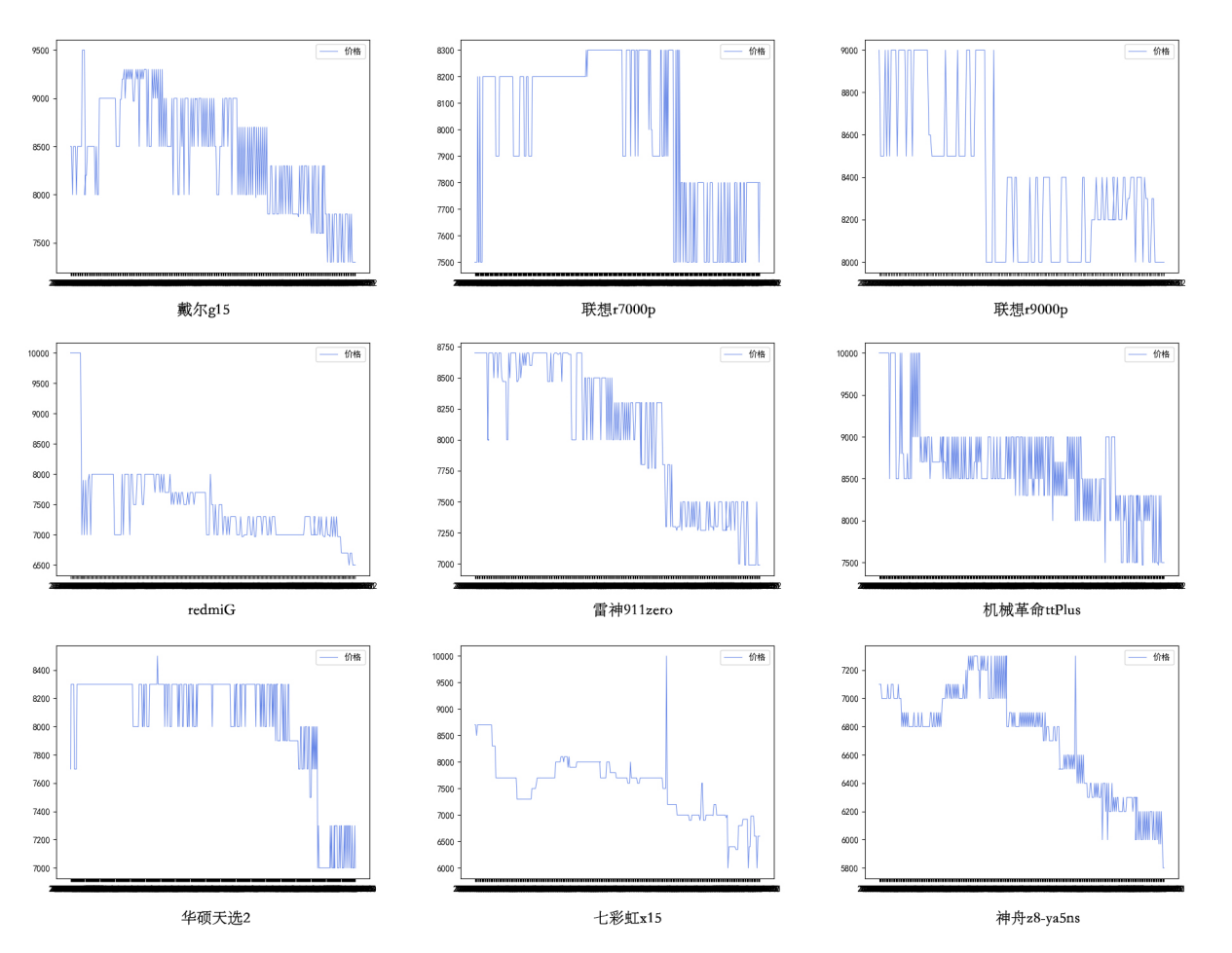


图-6价格走势曲线汇总图

由图-5价格走势曲线汇总图可知，大多数定位较高的品牌如联想、华硕、戴尔等，在一段时间区间内价格总是：保持高价->促销降价->然后再保持高价,再两个价格上反复横跳这样的销售策略。这样的策略使得消费者不能在任何时候都可以以最低价买到，但也使得消费者在将其产品加入购物车几天后被提示降价\*\*\*元，也起到了一定的促使购物的效果。而像七彩虹(图8)这样的品牌定位较低的厂商价格策略与其他家截然不同，它采取了长期保持低价的价格策略，使得消费者随时都能买到最低价的产品。各种定价策略相信都是商家市场部人员商议的结果，各有利弊，无好坏之分。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| price\_价格 | g15 | r7000p | r9000p | redmiG | TR911zero | ttPlus | tx2 | x15 | z8 |  |
| count | 529 | 735 | 323 | 417 | 515 | 673 | 813 | 649 | 635 | 5289 |
| mean | 8443.378072 | 8015.598639 | 8389.712074 | 7454.251799 | 8077.23301 | 8620.872214 | 8041.558426 | 7534.608629 | 6706.433071 |  |
| std | 573.0190918 | 287.613256 | 361.639367 | 637.2356137 | 576.7187003 | 583.6684085 | 407.4543682 | 571.8935607 | 374.2759907 |  |
| min | 7299 | 7499 | 7999 | 6499 | 6989 | 7469 | 6999 | 5998 | 5799 |  |
| 25% | 7999 | 7799 | 7999 | 6999 | 7499 | 8299 | 7999 | 6999 | 6399 |  |
| 50% | 8499 | 8199 | 8399 | 7299 | 8299 | 8499 | 8299 | 7699 | 6799 |  |
| 75% | 8999 | 8199 | 8499 | 7699 | 8699 | 8999 | 8299 | 7899 | 6999 |  |
| max | 9499 | 8299 | 8999 | 6999 | 8699 | 9999 | 8499 | 9999 | 7299 |  |
| changPriceRemark | 23% | 10% | 11% | 7% | 20% | 25% | 18% | 40% | 21% |  |
|  | 戴尔 | 联想 | 联想 | 红米 | 雷神 | 机械革命 | 华硕 | 七彩虹 | 神舟 |  |

表-1 汇总数据分析统计表

上表数据由pandas中的.describe()方法计算得出，共统计了5289条数据，max一般为商品的首发价，min一般为至今位置商品的最低价，changPriceRemark一栏为最高价和最低价之间的价格变化幅度百分比。由上表易知降价幅度较低的品牌一般为品牌自身定位较高的品牌，如联想的拯救者系列，该产品可以称之为行业标杆，各方面素质均十分出众，自然价格也十分坚挺，不会有跳水行为背刺之前购买的消费者的行为，但相应的，其产品性价比就不如其他。

# 第五章 讨论和心得

人工智能在当代社会中应用越来越广泛，对其的学习是十分有必要的，在本次作业的完成过程中也对人工智能及相关方面有了更深刻的了解和掌握，对今后的学习有很大裨益。

本次作业为人工智能课程的第一次作业，由于本人水平有限及期末考试提前时间不足等等因素，本次项目的成果可能不尽如人意，但作业的目的就是为了提高自己，在完成本次作业的过程中我学习到了许多之前不曾了解知识，也巩固了课上所学，让我自己得到了许多提升。