京东商品数据爬取与分析说明介绍

[1、课程设计目的 1](#_Toc91183999)

[2、课程设计内容 1](#_Toc91184000)

[3、背景知识 2](#_Toc91184001)

[4、设计步骤与方法 3](#_Toc91184002)

[4.1实验环境 3](#_Toc91184003)

[4.2步骤1： 3](#_Toc91184004)

[4.3步骤2: 7](#_Toc91184005)

[5、设计结果及分析 10](#_Toc91184006)

[6、问题及心得体会 12](#_Toc91184007)

[附录 12](#_Toc91184008)

# 1、课程设计目的

手机已经成为人们日常的必备工具，没有手机可以说是寸步难行，所以手机对人们格外重要，恰逢这次课程设计，于是想通过爬虫爬取京东上面手机的重要信息，并对其进行数据统计处理后进行数据可视化来进一步了解手机的特性

# 2、课程设计内容

本次课程主要由三部分组成：

1. 首先通过京东网站进行数据爬取，爬取手机的销量，价格，发货地区，手机的属性等一系列有关信息，然后将数据保存到text文件里面,由于京东只能展示100页，所以一共爬取100页的信息内容，一共有4591条数据，保存文件1.69M的信息
2. 然后使用pandas库对这些数据进行分析出来，分别按每一列的列名对数据进行划分，并且用replace函数和strip以及split这三个数据清洗的常用函数进行清洗，然后将数据传入列表里面
3. 然后用pyecharts和seaborn，matplotlib这三个常用的画图库进行数据可视化处理

# 3、背景知识

京东（股票代码：JD），中国自营式电商企业，创始人刘强东担任京东集团董事局主席兼首席执行官 。旗下设有京东商城、京东金融、拍拍网、京东智能、O2O及海外事业部等。2013年正式获得虚拟运营商牌照。2014年5月在美国纳斯达克证券交易所正式挂牌上市。2015年7月，京东入选纳斯达克100指数和纳斯达克100平均加权指数。 2016年6月与沃尔玛达成深度战略合作，1号店并入京东。  
2017年1月4日，中国银联宣布京东金融旗下支付公司正式成为银联收单成员机构。2017年4月25日，京东集团宣布正式组建京东物流子集团。2017年8月3日，2017年“中国互联网企业100强”榜单发布，京东排名第四位。 2019年7月，发布2019《财富》世界500强：位列139位。2021年《财富》世界500强排名第59位 。  
2018年3月15日，京东成立了“客户卓越体验部”，整体负责京东集团层面客户体验项目的推进。 2018年《财富》世界500强排行榜第181名。2018年7月24日，京东增资安联财险中国的方案获得了银保监会的批准。 9月4日，京东集团与如意控股集团签署战略合作协议。 2019年8月22日，进入2019中国民营企业500强前十名 ；2019中国民营企业服务业100强发布，京东集团排名第4 。2019年9月7日，中国商业联合会、中华全国商业信息中心发布2018年度中国零售百强名单，京东排名第2位。 2019年10月，在福布斯全球数字经济100强榜排第44位。  
2020年4月，京东确认将赴港上市 。12月18日，京东集团就京东金融广告道歉：严格问责，全集团进行反省教育。

手机，全称为移动电话或无线电话，通常称为手机，原本只是一种通讯工具，早期又有“大哥大”的俗称 ，是可以在较广范围内使用的便携式电话终端，最早是由美国贝尔实验室在1940年制造的战地移动电话机发展而来。  
1958年，苏联工程师列昂尼德.库普里扬诺维奇发明了ЛК-1型移动电话，1973年，美国摩托罗拉工程师马丁·库帕发明了世界上第一部商业化手机。历经2G时代、3G时代，迄今为止已发展至4G时代了，而5G时代也紧随其后，国内已经出现5G的商用。

经过一个学期的python课程的学习，我对python语言也有了一定程度的掌握，特别是python丰富的第三方扩展库，可以使我更加容易完成对手机信息的收集和分析

# 4、设计步骤与方法

## 4.1实验环境

window10

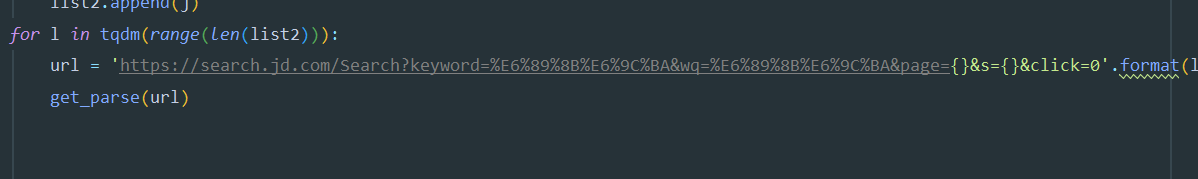
pycharm 2021.2

python3.7

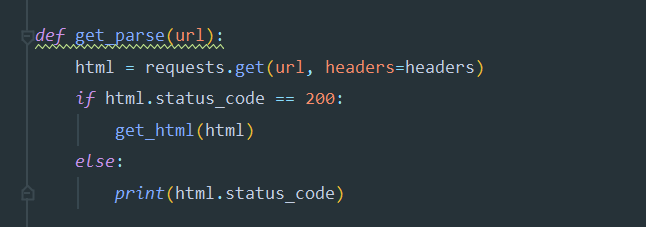
## 4.2步骤1：

数据爬取 requests，lxml，re

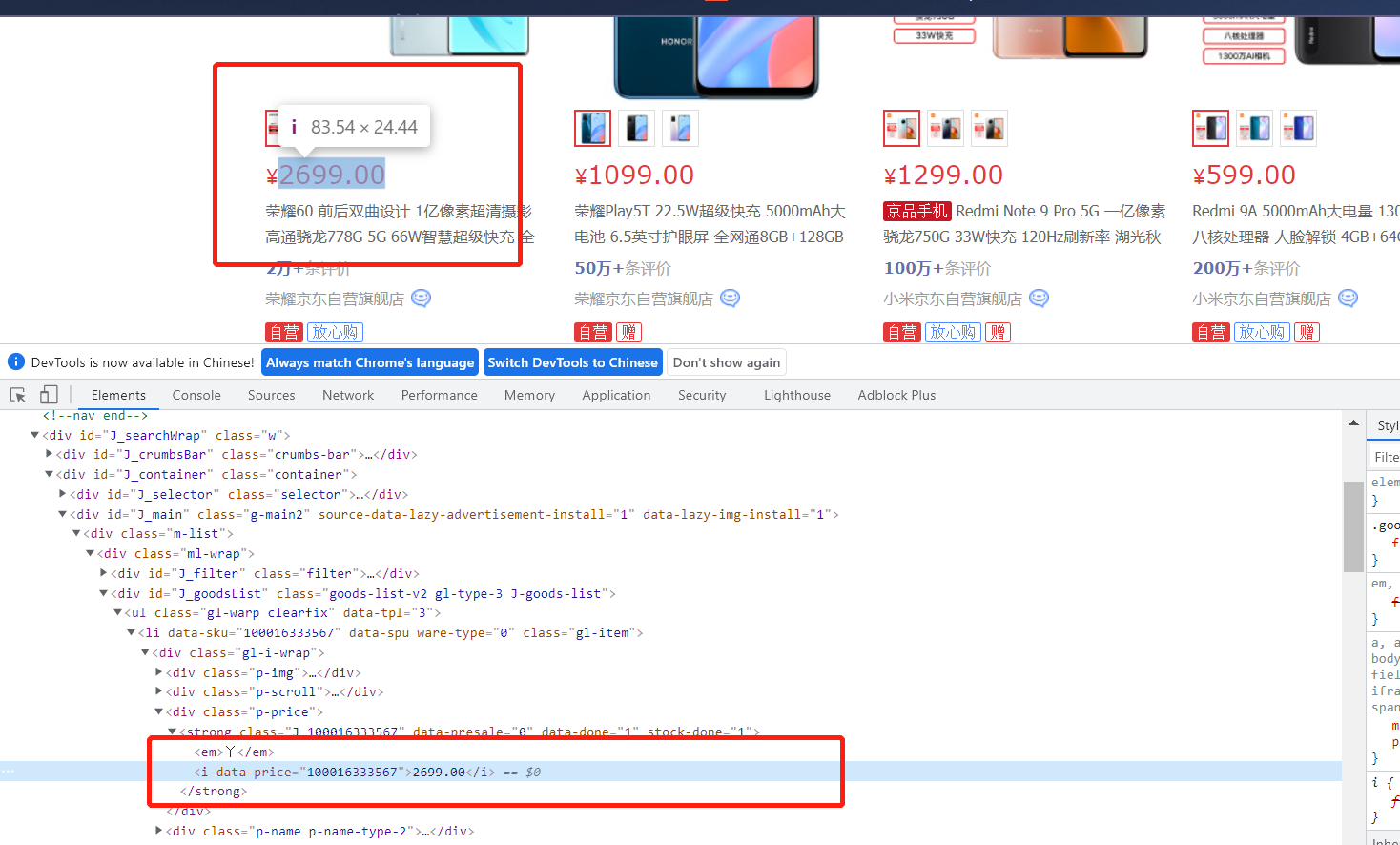
首先是先去根据京东的网址去寻找规律，它每跳转一页，对应的page和s就会发生变化，s是每隔30变一次，page是每隔1变成一次，所以我们要提前伪造好对应的URL



然后再去判断这个网页是否能正常打开，如果能正常打开就是去获取页面的源码，如果不能则返回页面的状态码

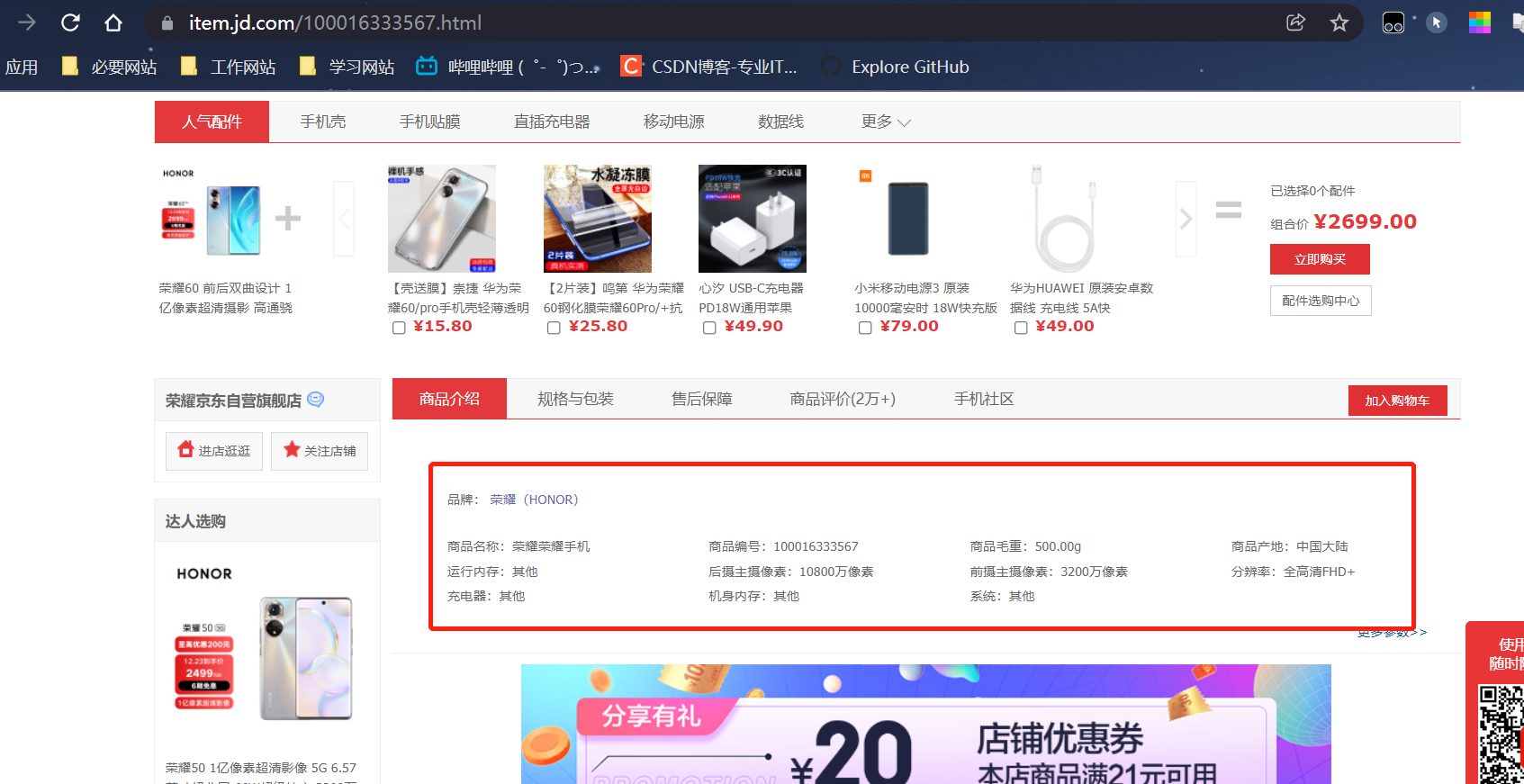


然后因为京东网站的数据都是静态页面的，所以我们直接用lxml去定位改属性的位置即可

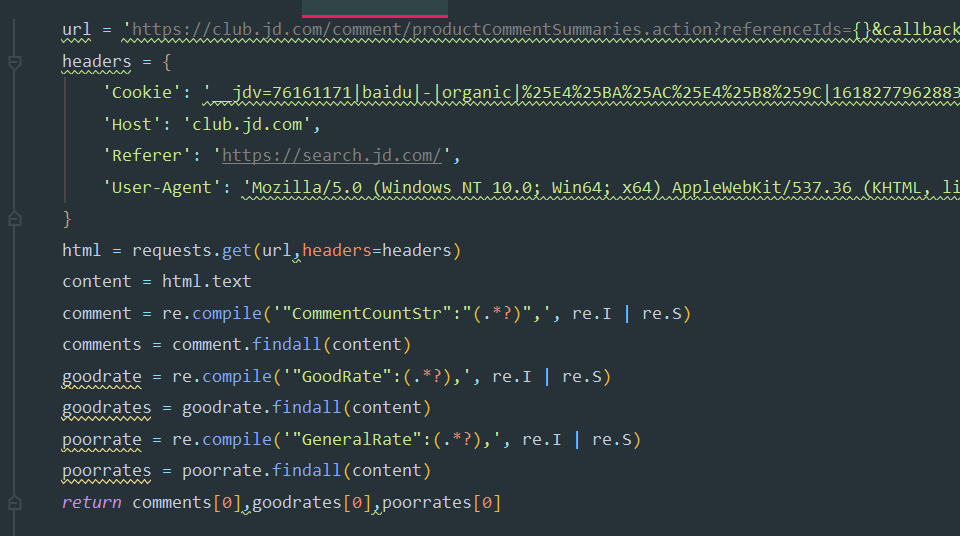




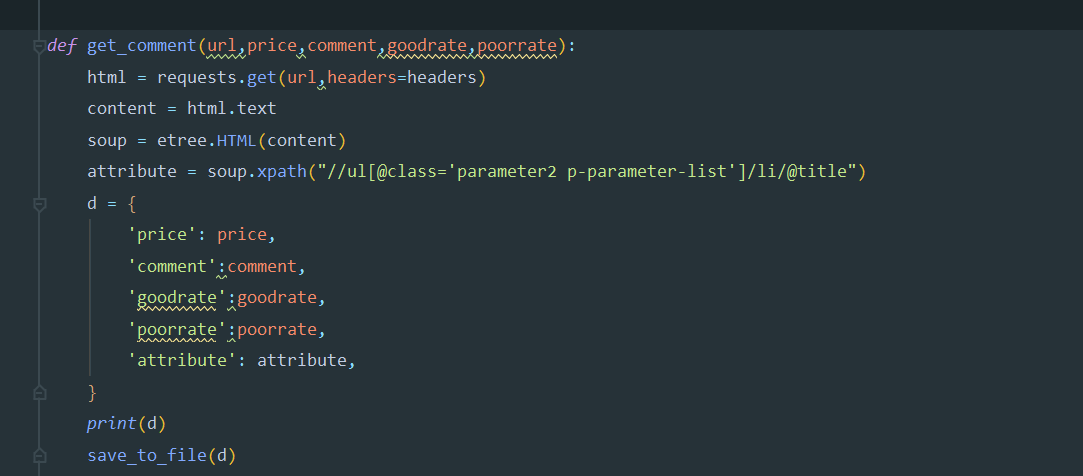
然后我们根据上面获取到每个商品的详细链接之后，我们再去获取商品详细页面里面的数据



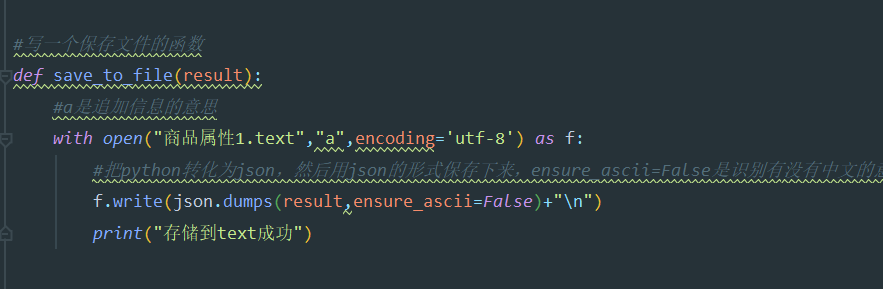
但是这时的xpath语法就失效了，因为每个页面的规则都不同，我们又要获取全部数据，这时我们就采用正则表达式去获取全部数据



然后我们把上面所有获取到的内容全部保存到text文件里面，直接使用函数进行传参，这样的目的是为了分工明确，且便于修改



最后把这些数据保存下来





## 4.3步骤2:

在把全部数据获取到之后我们开始进行一步去分析

首先分析价格和评论之间的关系，先读取价格列和评论列，然后用replace把后面的小数点去掉这里是价格的清洗步骤



再来到评论先用strip把前面的空格去掉，然后再用replace把多余的字符去掉，最后把这个转化成整数类型



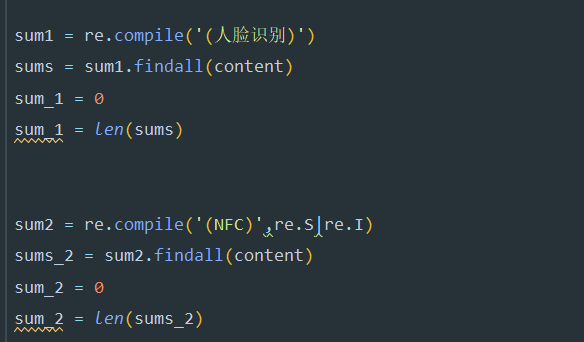
再逐个进行对比，找出0-1000,1000-2000,2000-3000,3000-4000,4000-5000,5000以上对应各个价位的手机评论数据



然后再用pyechart把图画出来



接着根据商品的属性去算出每个手机包含热点的数量，并且进行统计



然后用列表把数据收集起来，并且用pyechart把图像进行绘制



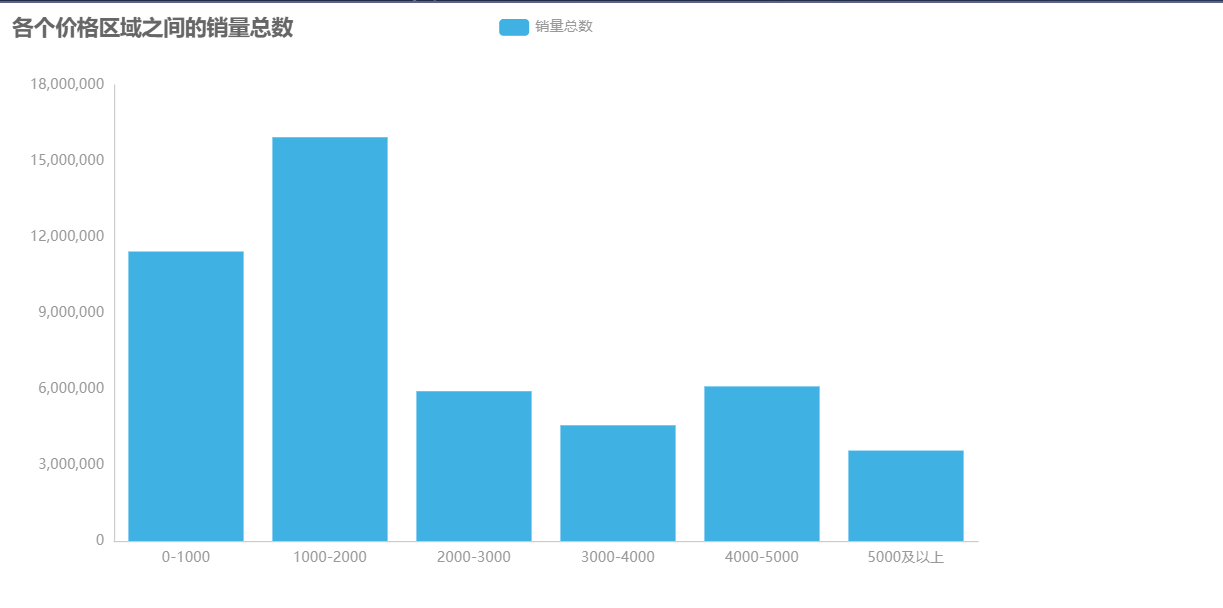
最后去计算价格与评论的线性关系

根据之前对它们已经清洗好的数据，用datafram重新包装清洗好的数据，再用seaborn 去绘制它们的线性关系是正比还是反比

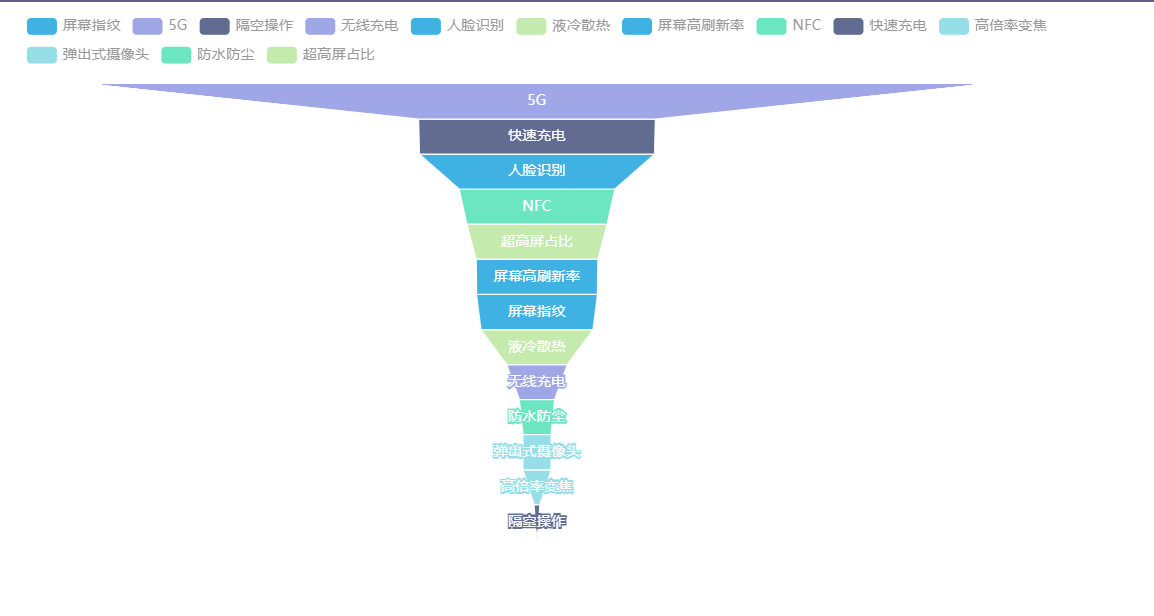


# 5、设计结果及分析

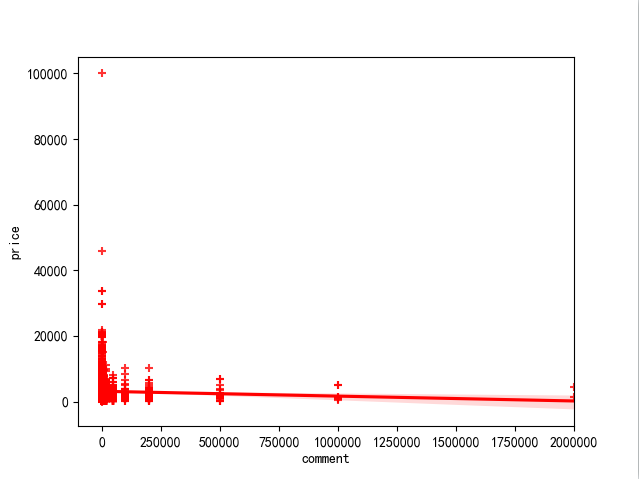
通过价格和评论的柱状图，我们可以清楚的得知销量最高的1000-2000之间的手机，这和我国的人均GDP有关，目前我国仍处于发展阶段，人均收入不高，这时手机的价格成为人们的重要选择，1000-2000这个价位区间的手机性价比最高，自然销量最好，其次就是0-1000的了



关于手机的热卖点，目前主流的热卖点是5G，其次就是快速充电，再者就是人脸识别这也说明目前人们买手机首先就是看这3点，是否是5G手机，是否支持快充，是否能进行人脸识别，如果没有符合上述的功能，那么手机的销量可能就要大打折扣



最后来看线性回归图，线性回归图和我们刚刚柱状图有关，果然正如上面说的一样，价格和评论呈负相关，价格越高评论数量越低，这符合上面的叙述



# 6、问题及心得体会

主要的问题还是爬虫的方面，最开始爬取数据的时候一直都是采用xpath语法，但是后面去爬取详细页面就失效了，果然不能一招吃遍天下鲜，我们还有多多学习，用知识来补充我们的短板，让我们越走越远

# 附录

代码。。。。。