Analysis of tablet data sets from Zhongguancun Online

GUO Shengming 22465758

Wu Menglei 22457968

Long Zhaonan 22435328

Zhang Yuwei 22417915

1.Abstract

Tablet PC, also called portable PC, is a small, portable personal computer with a touch screen as the basic input device. Its biggest feature is that the touch screen and handwriting recognition input function, as well as powerful pen input recognition, voice recognition, gesture recognition, and mobility. With the advancement of technology and information technology, tablet PCs are becoming more and more common in our daily lives, and their functions and roles are getting closer to those of computers and cell phones, which can improve our productivity and work efficiency. The purpose of this report is to help us quickly understand the development of the tablet PC and its various parameters, to understand what factors influence the price of a tablet PC, and to provide ideas and references for novice buyers of tablet PCs.

2.Introduction

2.1 Introduction of studied problem

The topic of this report is to crawl tablet PC data information from the offer website of Zhongguancun Online by writing a python program and analyze the data visually.

Our main work includes writing a crawler program, analyzing the composition of the URL data page, thinking about how to write incremental crawler logic, using request requests to get html, and then using BeautifulSoup to parse the html to get field by field, and finally successfully crawling the various data parameters of the tablet PC and saving the csv file to local. Then, after having the dataset csv file, we need to use pandas to read and process it, and then use the matplotlib library to plot and analyze the data and graphs.

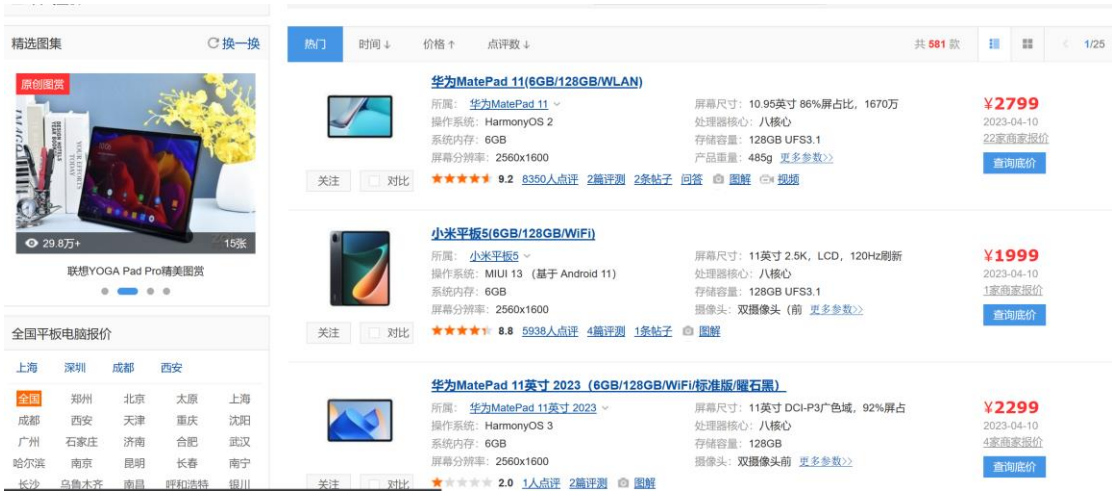
2.2 The motivation of studying this topic and its associated value

The purpose of this report is to help us quickly understand the development of the tablet PC and its various parameters, to understand what factors influence the price of a tablet PC, and to provide ideas and references for novice buyers of tablet PCs.

3. Methodology

3.1 Data source

The tablet PC data in this analysis report comes from the Zhongguancun Online website: https://detail.zol.com.cn/tablepc/



3.2 Data collection method

1. Import modules

At the beginning of the code, the required modules, including requests for sending network requests, re for regular expression matching, traceback for obtaining exception information, and BeautifulSoup for parsing HTML, are imported using the import statement.

```
import requests
import re
import traceback
from bs4 import BeautifulSoup
import csv
```

2. Encapsulate functions

The code encapsulates multiple functions to get webpage HTML, parse HTML documents, and store data to a CSV file.

Parse HTML documents

```
def GetHtmlforProUrl(url, bn):
    try:
        html = GetHtml(url)
        # 先判定是否有数据
        if len(re.findall('对不起，暂时没有', html)) != 0:
            print(f'{url}没有数据')
            return False
        soup = BeautifulSoup(html, 'html.parser')
        divlis = soup.find_all('div', 'pro-intro')
        for div in divlis:
            a = div.find('a', 'more')
            if a is None:
                continue
            prodetailurl = a.attrs['href']
            GetHtmlAndParserPro(prodetailurl, bn)
    except Exception as e:
        print(e)
        traceback.print_exc(limit=2, file=open('tt.log', 'a+'))
        print(f'GetHtmlforProUrl: {url}发生错误')
    return True
```

3.Main function

The Main() function is the main entry point of the program. It defines the URLs to be crawled and iterates through these URLs, calling other functions to extract data from the web page and save it to a CSV file.

```
def Main():
    baseUrl = 'https://detail.zol.com.cn/tablepc/'
    brandname = ['苹果', '华为', '小米', '三星', '荣耀', '中柏', '酷比魔方', '台电', '微软', '联想', 'E人E本', '诺基亚',
                 '海信', '索立信', 'vivo', 'OPPO', '松下', 'realme', '谷歌', '海尔', '神基', 'VOYO', '纽麦', '清华同方',
                 'a豆', 'ROG', '睿刚', '集思宝']
    brannamelist = ['apple', 'huawei', 'xiaomi', 'samsung','honor','jumper','kubi','teclast','microsoft','lenovo','eben',
                    'nokia','hisense','soulixin','vivo','oppo','panasonic','realme','google','haier','getac','voyo',
                    'nerlmiay','tongfang','adol','rog','ruggon','jisibao']
    for i in range(0, len(brannamelist)):
        for j in range(1, 25):
            url = baseUrl + brannamelist[i] +'/'+ str(j) +'.html'
            isco = GetHtmlforProUrl(url, brandname[i])
            if not isco:
                break
```

4.CSV file operations

The program uses the open() function to create a CSV file and open a file object. It then creates a csv file writing object through the csv.writer() function and writes the table headers. Next, in the SaveContentToCsv() function, the program stores the data to be written to the CSV file in a list and then calls the csv_writer.writerow() function to write the data in the list to the CSV file.

```
# 1. 创建文件对象（指定文件名，模式，编码方式）
f = open("pbdninfo6.csv", "w", encoding="gbk", newline="")
# 2. 基于文件对象构建 csv写入对象
csv_writer = csv.writer(f)
# 3. 构建列表头
csv_writer.writerow(['公司名称','产品名称','上市年份','上市月份','价格(元)','运行内存(GB)','存储容量(GB)','操作系统','处理器型号','屏幕尺寸(英寸)','屏幕分辨率','重量(g)','评论人数','评分','电池类型','电池容量', ])
def SaveContentToCsv(infolist):
    global csv_writer
    csv_writer.writerow(infolist)# 4. 写入csv文件内容
```

5. Website structure analysis

By analyzing the structure of the target website and the tags in the HTML pages, the program uses

the BeautifulSoup library to parse the HTML pages and extract the required data.

```
def GetHtmlforProUrl(url, bn):
    try:
        html = GetHtml(url)
        # 先判定是否有数据
        if len(re.findall('对不起，暂时没有', html)) != 0:
            print(f'{url}没有数据')
            return False
        soup = BeautifulSoup(html, 'html.parser')
        divlis = soup.find_all('div', 'pro-intro')
        for div in divlis:
            a = div.find('a', 'more')
            if a is None:
                continue
            prodetailurl = a.attrs['href']
            GetHtmlAndParserPro(prodetailurl, bn)
    except Exception as e:
        print(e)
        traceback.print_exc(limit=2, file=open('tt.log', 'a+'))
        print(f'GetHtmlforProUrl: {url}发生错误')
    return True
```

```
def GetSpanString(detaildiv, sname):
    getstr = ''
    try:
        getstr = detaildiv.find('span', string=sname).find_parent('tr').find('td').find('span').string
        # print(getstr)
    except Exception as e:
        # print('没有', sname)
        it = ''
    return getstr
```

6.Exception handling

The original program returns the Boolean value 'False' if an exception occurs during web page acquisition or parsing. This method is relatively straightforward and requires improvement.

```
def GetHtmlforProUrl(url, bn):
    try:
        html = GetHtml(url)
        # 先判定是否有数据
        if len(re.findall('对不起，暂时没有', html)) != 0:
            print(f'{url}没有数据')
            return False
        soup = BeautifulSoup(html, 'html.parser')
        divlis = soup.find_all('div', 'pro-intro')
        for div in divlis:
            a = div.find('a', 'more')
            if a is None:
                continue
            prodetailurl = a.attrs['href']
            GetHtmlAndParserPro(prodetailurl, bn)
    except Exception as e:
        print(e)
        traceback.print_exc(limit=2, file=open('tt.log', 'a+'))
        print(f'GetHtmlforProUrl: {url}发生错误')
    return True
```

7.Performance optimization

Due to the large amount of data to be processed, program performance bottlenecks may occur. To further improve program execution efficiency, multithreading or coroutine mode can be used to process data, which can greatly improve program execution efficiency.

8.Other improvements

In addition, when crawling data, it is necessary to prevent the crawler from being blocked by the website. This can be mitigated by adding User-Agent or setting access intervals. When writing code, attention should be paid to code readability and maintainability, making the program easy to understand and modify.

When using Matplotlib to make a diagram, we set the font to SimHei, but the runtime prompt did not find this font. At first, we thought it was a system font problem, but after installing the fonts on the system, it was still not solved. After checking, it is because there is no relevant Chinese font in Matplotlib.
So, there is a messy code in some legend of pie charts. To solve this problem of missing Chinese fonts, we tried many ways, such as inserting codes with a translation function. But finally, we solved this question by typing the labels manually with the English names and abbreviations of the different brands in this line: labels=['Apple', 'Koobee', 'Huawei', 'Teclast', 'Samsung', 'Lenove', 'Honor', 'Mi', 'Jp', 'Mc'].

3.3 Data analysis and exploration

1.Get csv. file and upload it to Github

```
url = "https://raw.githubusercontent.com/Leizisayhelloworld/Jour7280-Final-project/main/pad_Info.csv"
pad_info = pd.read_csv(url)
pad_info
```

Scraping data takes a lot of time, and the online file was produced and stored in the temporary stored file of colab. The author downloaded the file immediately after finishing the scraping and upload the csv. file to Github, so that we get an online document, through which the authors are easy to request the dataframe by using the function read_csv() during the group cooporation.

2.Data cleaning

When extracting data, missing values may occur. To ensure data accuracy and integrity, data cleaning is required.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 453 entries, 0 to 452
Data columns (total 16 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   公司名称       453 non-null    object
 1   产品名称       453 non-null    object
 2   上市年份       411 non-null    float64
 3   上市月份       350 non-null    float64
 4   价格(元)      453 non-null    int64
 5   运行内存(GB)   354 non-null    float64
 6   存储容量(GB)   421 non-null    float64
 7   操作系统       433 non-null    object
 8   处理器型号      413 non-null    object
 9   屏幕尺寸(英寸)   453 non-null    float64
 10  屏幕分辨率      453 non-null    object
 11  重量(g)      395 non-null    float64
 12  评论人数       453 non-null    int64
 13  评分         453 non-null    float64
 14  电池类型       396 non-null    object
 15  电池容量       215 non-null    object
dtypes: float64(7), int64(2), object(7)
memory usage: 56.8+ KB
```

By analyzing the function .info(), we can analyze our dataframe ------ which was assigned to pad_info. The dataframe shows that the column “上市月份”, “运行内存(GB)” and “电池容量” has too many null values, all of which do not have sufficient data. The authors find that the null values of column “上市月份” accounts for 22% of the total value in the column; the null values of column “电池容量” accounts for 52.5% of the total value in the column; the null values of column “运行内存(GB)” accounts for 21.9% of the total value in the column. Because so many missing value will affect the exactitude of our analysis, the authors decided to delete these two columns by using .drop() function.

4. Deleting the useless column ---- Month

```
Useful_pad_info = pad_info.drop(['上市月份'], axis = 1)
Useful_pad_info
```

Apart from columns"上市月份", "运行内存(GB)" and "电池容量”, the pad information of the other columns is partially missing, and the missing information is scattered distribution, and the volume of the missing information is less, so we cannot delete them directly. What we want to do is dealing with those missing values according to the specific situation.

After deleting the useless columns, the authors also checked that there is not any duplicated rows and get a new dataframe, which was assigned to Useful_pad_info.

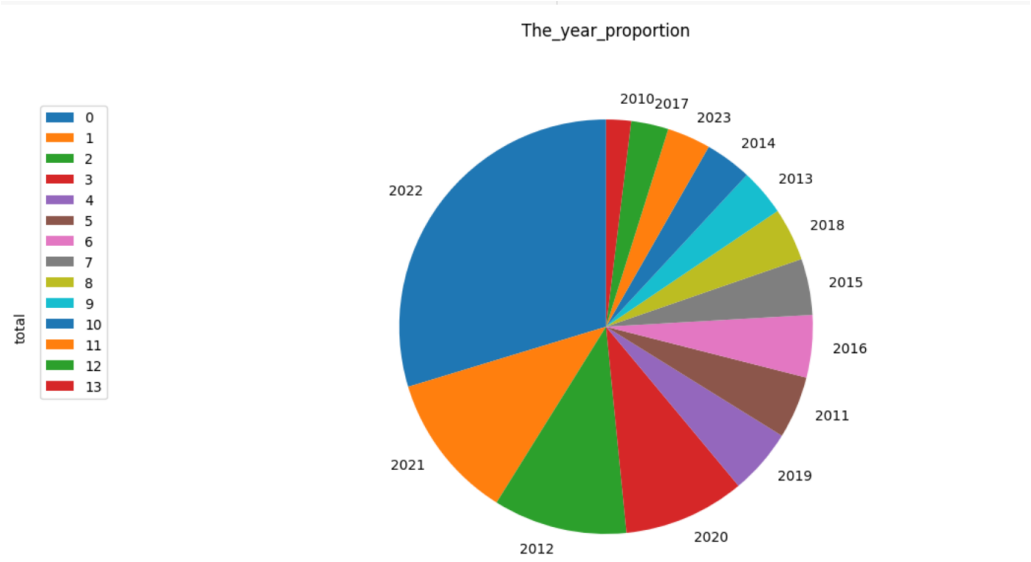
3.Data visualization

Based on the knowledge we have and the goal of our research, the authors want to know more about

the development of the tablet computer, the different types of the tablet computer on the market and to understand what factors influence the price of a tablet PC.

4.The development of tablet computer

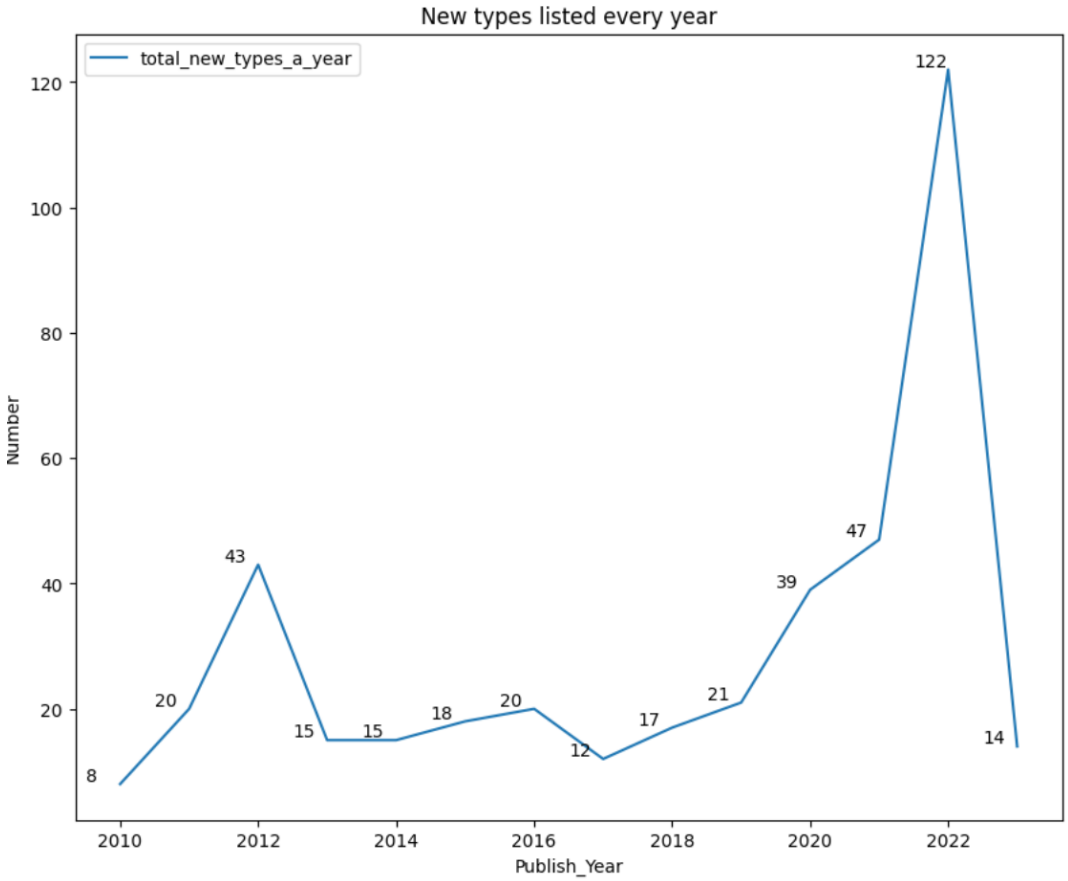
The pie chart and line chart show us the development of tablet computer. The authors decide to explain the two charts one by one.



(The Chart subplot)

After getting the pie chart, the authors want to find more about the history behind the tablet market. Although the first business tablet computer was introduced in 1989, it became popular due to the strong promotion by Microsoft in 2002. In 2010, Apple's iPad with Apple's iOS operating system, which expanded the iPad from the workplace to daily life, is used for multimedia. Therefore, the tablet computer became easily accepted by consumers. That’s the reason why the data started in 2010.

Apple, the leader in the tablet market, released its third and fourth generation iPads in 2012. Meanwhile, Microsoft launched its Surface line of tablets on June 19, and other brands such as Android continued to launch new tablet products. As a result, 2012 was known as the year of the tablet computer, and sales in this year increased greatly.



When it comes to the line chart, it is much easier to see the trend of development of tablet computer. And by adding another line of code, the authors find that the brand of tablet computer that was released in 2010 was not only the Apple, but also the Teclast.

```
year_df = pd.DataFrame(Useful_pad_info.loc[Useful_pad_info['上市年份'] == 2010])
print(year_df)
```

	公司名称	产品名称	上市年份	价格(元)	运行内存(GB)	存储容量(GB)
101	苹果	苹果iPad（16GB/WiFi版）	2010.0	3200	NaN	16.0
110	苹果	苹果iPad（32GB/WiFi版）	2010.0	3450	NaN	32.0
111	苹果	苹果iPad（64GB/WiFi版）	2010.0	3650	NaN	64.0
161	苹果	苹果iPad（16GB/WiFi+3G版）	2010.0	3400	NaN	16.0
162	苹果	苹果iPad（32GB/WiFi+3G版）	2010.0	3650	NaN	32.0
163	苹果	苹果iPad（64GB/WiFi+3G版）	2010.0	3800	NaN	64.0
391	台电	台电T720-3GE（8GB）	2010.0	1580	NaN	NaN
406	台电	台电T720-WiFi（8GB）	2010.0	930	NaN	NaN

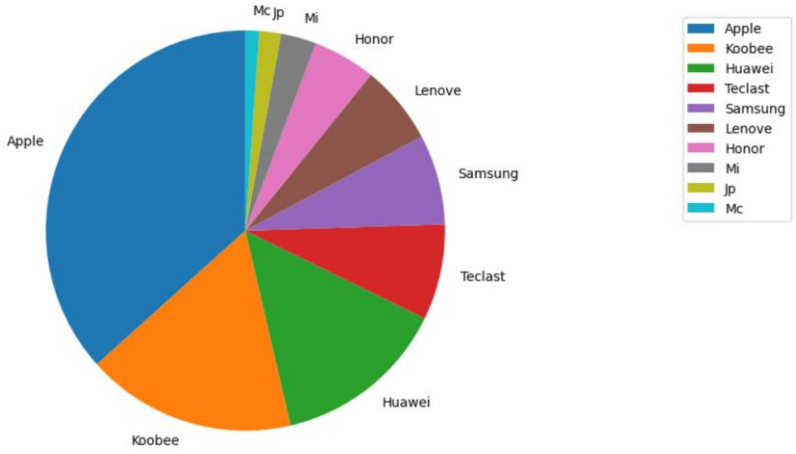
The later brand is not a famous brand, but its product is quiet cheap, whose lowest price of a 8GB-storage-capacity is only 299 RMB, much lower than Apple. As for Apple, the price of its 8GB storage capacity is 1350 RMB, which is about three times more expensive than the latter.

In the year 2012, the first type of “Ipad Mini” was released by Apple, and in this year many new types were shown up in the market. After that, the yearly released new types of tablet computer went

down again, until the 2020.

The new types of tablet in 2023 drops sharply, but this does not mean something bad happen. The reason for this is that this report's writing time is at April, and the new product launches of many brand do not start yet.

5. Comparison of the number of tablets by company



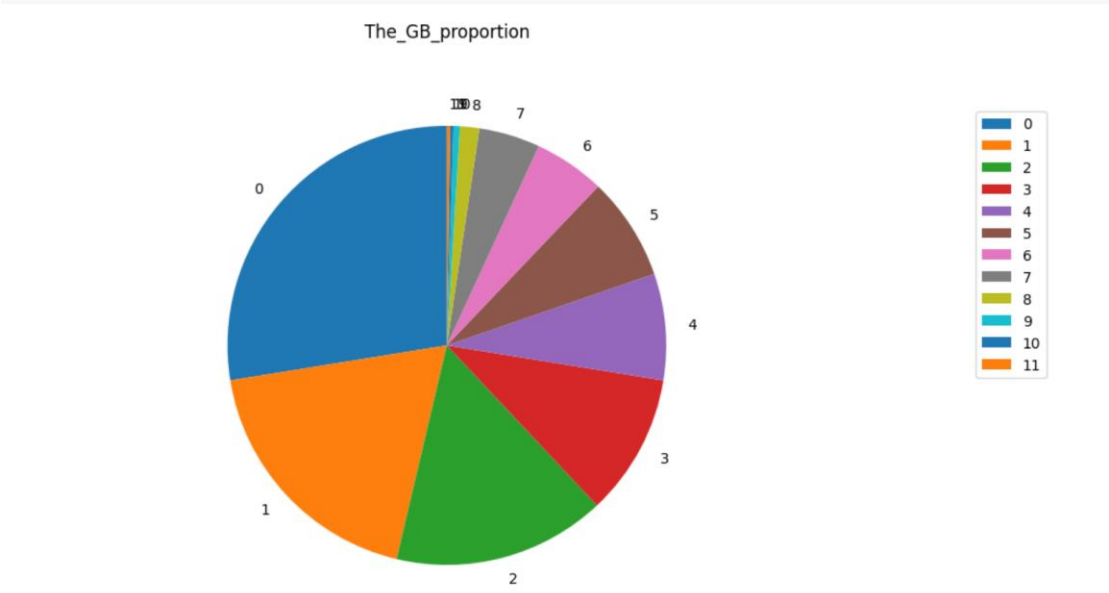
First of all, the authors find that Apple leads in the number of tablets, which means that Apple is the leader in the tablet market. When it comes to iPad, a line of tablet computers released by Apple since 2010, which is also an electronic product between Apple's smartphone iPhone and laptop products. Tablet computers had been around for more than a decade before iPad, but never attracted the attention of mainstream users. Since the launch of the first iPad, Apple remains the top-selling tablet company, which means that the Apple's iPad has had an impact on existing tablet PC sales.

Secondly, in terms of the number of products, Apple is followed by Koobee, Huawei, Teclast, Samsung, Lenove and Honor, all of which have a similar number of products, indicating that brands in the tablet market are diverse.

Except for Samsung, the Chinese brands account for almost 50% among all of the tablet computer types. There are three plausible reasons. First, the website we scraped is a Chinese website. Second, China is good at producing electronic product. Third, some of the Chinese brands, such as Koobee and Teclast are tend to produce many different but cheap types to scrape the market.

In this case, on the one hand, different brands can offer consumers a choice of different products and prices. On the other hand, different brands also make it difficult for consumers to find a brand with good quality. To find a proper type of tablet computer, the authors' suggestion is that ------ choose the product which has famous brand.

6.The numbers of types of tablet with different storage capacities



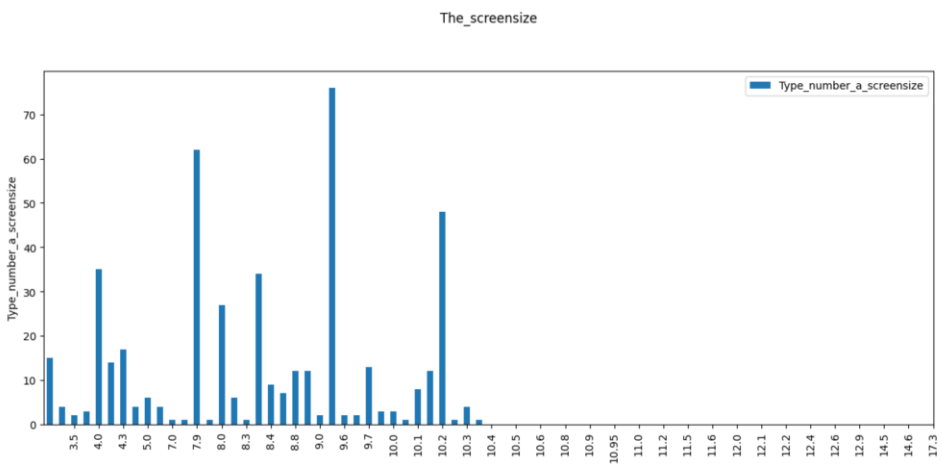
The top is the 128GB tablet with a total of 116 units. The next was the 64GB tablet with a total of 79 units. Then came the 256GB tablets with a total of 66 units. And these three memory capacity of the tablet has accounted for about two-thirds of the entire market.

If we want to know how much memory is appropriate to buy a tablet computer, in addition to consider the use of scenarios, but also must be distinguished from the running memory and body memory.

If you just look at the video with little memory capacity to cache, do not need to add a lot of APP, light video and web APP, this type of use of the tablet can be with 6 + 64G or 128G combination.

If the user is a game player, the CPU performance requirements are very high, the CPU and GPU specifications must also be on a certain rate, this time the running memory and body memory have to buy large not small, there are conditions directly with the 8 + 256G combination, will play very cool.

7. The numbers of types of tablets with different screen sizes



The largest number is the 11.0 inch tablet, with 76 types. This means that this size pad has the most types in the market, with the main model being the Apple Ipad pro. The second most numerous is the 9.7-inch tablet, with 62 units. The third most is the 12.9-inch tablet, with 48 units.

The size of the tablet is one of the most important factors to consider when buying a tablet. Generally speaking, the size of the tablet can be divided into three intervals: below 10 inches, 10~11 inches and above 11 inches. Different sizes are suitable for different usage scenarios and tasks.

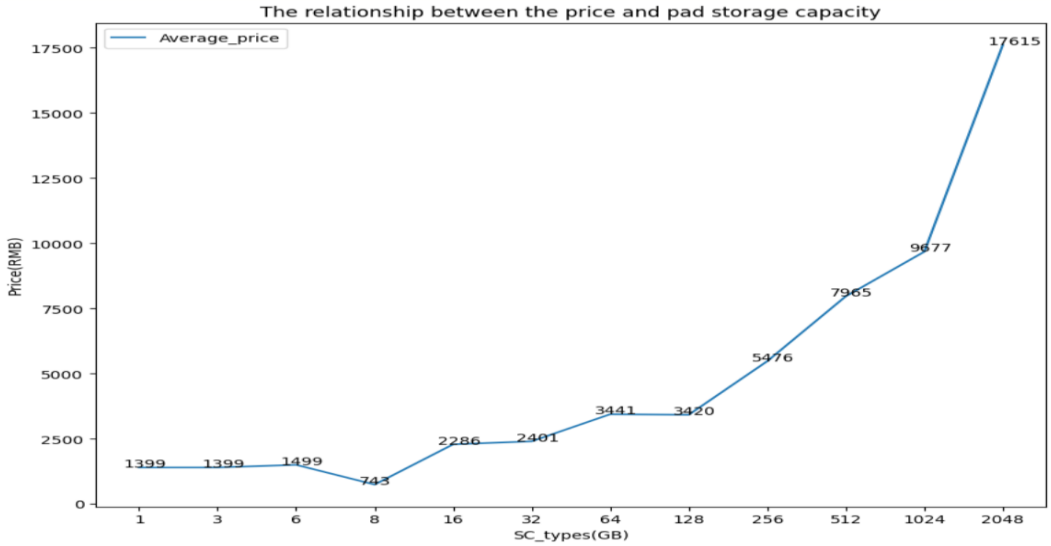
Tablet PC below 10 inches is a relatively small and lightweight product, suitable for use on the go or in the usual use of subway, bus and other transportation. This size tablet is easy to carry and can be easily used to watch videos, play games, brush the web and other basic operations, and is also ideal for users who like to use it in bed or on the sofa.

The 10 to 11-inch tablet is one of the more popular sizes in recent years. This size is more suitable for everyday tasks such as watching online classes, watching movies and taking notes. It is larger than tablets under 10 inches and can display more content, but is still relatively lightweight and easy to carry.

There is also a gradual increase in tablets above 11 inches, with the Apple iPad Pro reaching a maximum of 12.9 inches and the Samsung Tab S8 Ultra reaching 14.6 inches. Tablets of this size are often used for more complex tasks and are more productive, such as drawing and note-taking. The larger screen can better display details and more content, but also easier to operate.

The authors suggested, when choosing the size of the tablet, the consumer need to consider the use scenario and tasks. If the consumer needs a lightweight and portable tablet, a product under 10 inches may be more appropriate. If the consumer needs a larger screen to complete more tasks, 10 to 11-inch tablets may be more suitable. If the consumer needs to perform more complex tasks, such as drawing and note-taking, then a tablet of 11 inches or more may be more appropriate.

8. The relationship between the price and pad storage capacity



Through the line chart, we can observe the relationship between the price and the pad storage capacity shows positive correlation. The range of storage capacity is between 1GB to 2048GB. The cheapest type is not the type whose storage capacity only has 1GB, but the type whose storage capacity has 8GB.

The cause of this exceptional is that brand Teclast has the extremely cheap price. For instance, its lowest and highest price of an 8GB-storage-capacity product is 299 RMB, 988 RMB separately.

To find a proper type of tablet computer, except for staring at the famous brand, it is also important to find a type which has enough storage capacity but also has a reasonable price. The average salary of Chinese people is 2000-3000 RMB. So for most Chinese consumers, it is suitable to purchase a 128GB-storage-capacity product.

3.4 Work distribution of group members

GUO Shengming: Data scratching, report and PPT
WU Menglei: Data analysis, report and PPT
LONG Zhaonan: Data analysis, report and PPT
ZHANG YuWei: Data analysis, report and PPT

4.Reflection and Summary

Due to the large amount of data to be processed, program performance bottlenecks may occur. To further improve program execution efficiency, multithreading or coroutine mode can be used to process data, which can greatly improve program execution efficiency.

In addition, when crawling data, it is necessary to prevent the crawler from being blocked by the website. This can be mitigated by adding User-Agent or setting access intervals. When writing code, attention should be paid to code readability and maintainability, making the program easy to understand and modify.

When using Matplotlib to make a diagram, we set the font to SimHei, but the runtime prompt did not find this font. At first, we thought it was a system font problem, but after installing the fonts on the system, it was still not solved. After checking, it is because there is no relevant Chinese font in Matplotlib.
So, there is a messy code in some legend of pie charts. To solve this problem of missing Chinese fonts, we tried many ways, such as inserting codes with a translation function. But finally, we solved this question by typing the labels manually with the English names and abbreviations of the different brands in this line: labels=['Apple', 'Koobee', 'Huawei', 'Teclast', 'Samsung', 'Lenove', 'Honor', 'Mi', 'Jp', 'Mc'].