White Paper

- Missing Deadline Team

Chan Tai To
Li Haotian
Wei Rongyan
Xue Kaiwen

Executive Summary

The category our team chose is electronic devices which contain the laptop, smartphone, and tablet. The HKTVMall and the Suning Hong Kong are be selected for data crawling. HKTVMall was formally launched on 2nd February 2015. Regarding the ranking of Alexa about the top site in Hong Kong, HKTVmall lies in 110th in September 2017. The Suning entered the Hong Kong market at the end of 2009. Until 2017, there are 30 stores operated in Hong Kong. The customer consumption behaviours that happened in these two retails stores can be represented the rest of consumers. Most of the observations are based on the data we crawled while the characteristics of the operating system also have an impact on the customer's decision.

In general, it can be divided into three parts of the whole process, web crawling, data processing, and data analysis.

Methodology

Web Crawling

A Web crawler sometimes called a spider and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing. In our project, we wrote crawlers to retrieve product data and information that is related to them. Nowadays, many websites are using JavaScript to render contents, that makes us more difficult to request the source codes directly, so we used Pyppeteer, which is the Python version Puppeteer, it provides a high-level API to control Chrome or Chromium over the DevTools Protocol and it is easy for cross-platform setup and execution.

Crawl Suning

While crawling Suning (refer to crawl_HKSuning.ipynb), we found that we cannot determine the current page number by examining the URLs, which means we cannot tell the ending page to the crawler. Then we found out that, we can know if a page is the last page by seeking the “next page” button. If the next page button does not exist on the current page, that means the current page is the last page.

The strategy of crawling products on Hong Kong Suning is, first we visit the category page (e.g. phone page) and acquired all URLs of the products, then we traversed all URLs about that category, visit the detail page about that product and retrieved detail information about that information. We stored product data and comment data into two CSV files separately, for better organization.

In order to better presentation, we crawled the top nine products’ pictures of each category (refer to top-nine-imgs.ipynb) and made them into a 9 x 9 grid pictures.

Crawl HKTVMall

Refer to the Jupyter Notebook for crawling HKTVMall. HKTVMall has more data than that of Suning. The most significant challenge is Pyppeteer’s results are not definitely reproducible. For example, when sending a request to HKTVMall’s browser, the returned HTML files may have not been rendered by Javascript. In this case, we cannot find any useful information by examining HTML. We write an exception handler to repeatedly request the web page (using headless google chrome browser) until the returned HTML is correctly rendered by Javascript.

We prepared two separate tables to store our data from HTML. The design of these tables is of the same principle as database schemas. The first table is about products. The information of brand, category, name, price, sales, etc. are put in this table. The second one is about the comments of all products. Two tables are cross-referenced by the index of each product.

The URLs of HKTVMall are tagged by page numbers. We, therefore, iterated through all URLs of different categories for different information about products. Then, all information is available and can be examined by BeautifulSoup, a Python package of extracting HTML content.

Data Analysis

Price Analysis

- During the price analyzing procedure, we desired to plot the distribution of different prices of different items for each platform. After that, we calculate the mean value and median value for each category: tablet, laptop, and smartphone for references.
- For Suning HK, each transaction has its comment records. We analyze the base of the most-bought product on the transaction records.

Variety Analysis

- On HKTVMall, different sellers may sell the same products. Thus, we need to calculate the number of product variety. To achieve this, cosine similarity was applied to the names of each item.
- Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any angle in the interval (0,π] radians. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° relative to each other have a similarity of 0. This is analogous to the cosine, which is unity (maximum value) when the segments subtend a zero angle and zero (uncorrelated) when the segments are perpendicular.
- The similarity threshold was determined by a tuple of the item, whose name slightly diverse but in fact, are the same item. We set this value as the benchmark and compare other tuples.
- In each iteration, we take the current item as a new group, if we find a new item with similarity higher than the threshold, then group them together.

Sentiment Analysis

When analyzing customer sentiment towards each platform, we found that due to different shopping flow, the comments and rates from users varies. For customers on Suning, they are required to comment on the item every time they buy. Thus, for customers in Suning HK, we removed comments which are system generated because they lack sentiment. However, when we calculate the volume of transaction, this information is meaningful.

Procedures are as follow:

1. Extract user comments from each item,
2. Remove stop words (meaningless words in grammar, e.g. ‘and’, ‘the‘, 的)
3. Compute occurrence and generate word clouds (see appendix)

Data Sources

Data Description

There are eight columns for storing products data:

1. Product Index: Each product is represented by a unique number in Suning’s system.
2. Product Type: Determine the product is a phone, a laptop or a tablet.
3. Product Name: The full name of the product that is shown on the website.
4. Product Brand: The brand of the product
5. Price: The price of that product. The price may be different on different websites or different sellers.
6. Total Comment: Numbers of customers’ comments.
7. Seller: The sellers. In Suning website, all sellers are Hong Kong Suning.
8. Website: The URL of the product.

There are five columns for storing products data:

1. Product Index: The unique number of the product, as a foreign key that links to products that saved in another CSV file.
2. Username: The name of the person that commented on the product. The username is protected. In Suning, we cannot get the full username. In HKTVMall, we can only get the salutation of a user.
3. User Rate: Users’ ratings for the product.
4. Rate Date: The date that users rated.
5. Buyer Comment: Comments from the customer.

In Suning, users must rate or comment after they bought a product, but it is not necessary on HKTVMall, so the comment data are a lot less than what we got from Suning.

Data Preprocessing

Problems

1. Misclassification: wrong label by the suppliers

Restriction

1. Customer ID is protected
2. No unique ID assigned for customer name

Findings

Product Variety

For convenient, we divided the products into three categories, which is the laptop, tablet, and smartphone.

HKTVMall

There are 62 kinds of tablets, 174 kinds of laptop, and 281 kinds of smartphone observed from the HKTVMall products dataset. Inside the data, there are 9.7%(17) of tablets, 50.6%(88) of laptops, and 44.8%(126) of smartphones repeated. We can find that the more expensive products, the more available alternatives have.

Suning

There are 77 kinds of tablets, 66 kinds of laptop, and 194 kinds of smartphone observed from the Suning products dataset. Compare to the HKTVMall, Suning has a large variety of products but the available options are less.

Price

For the HKTVMall, the merchandise is not directly managed or supplied by HKTVMall. After searching on the website, it would show several options with the same products but different vendors and prices. The role of HKTVMall acts during the transaction tends to be an intermedia, to some extent, like Taobao. Due to that, the customer's bargaining power obviously increased which is beneficial for them. However, it may lead to customer service and service guarantee problems. If the goods damaged, the customer may contact the wrong person and can not find the right way to solve problems, which certainly influences the overall satisfaction of shopping experiences. Comparing to the HKTVMall, the merchandise shown on the website is all offered by the Suning itself so that the purchasing process and services guarantee will be protected by the website, the customer loss can be definitely reduced. Moreover, the bargaining power is relatively weak here, the higher price may be one significant factor that affects their purchasing decisions.

Uniqueness

- Observed Consumption Patterns

Since the Suning' s customer information is under security protection and the customer name of HKTVMall provided does not assign a unique ID as well. Both customer datasets are not appropriate for predicting the consumption patterns straightly. The purpose of finding consumption patterns is to provide some promotion packages with lower prices to attract customers, which is an effective way to increase profitability.

- System Operation

For information collection, it focuses on not only the details of variable functions but also an evaluation of services provided company. In this aspect, the Suning is doing better than HKTVmall. The comment is an indispensable reference for customers when making a purchasing decision. It compulsively requires the buyer to leave comments otherwise the system will automatically take the full mark. For the HKTVMall, the customer may depend on their needs to decide whether to write or not. Without comments, it is really hard for other buyers to elevate the quality of goods or services provided by the websites.

To combine the advantages of these two websites, we observed that a better shopping experience that can be provided almost relies on more information, no matter the price, the products guarantee, services supporting or other buyers' elevations. According to sufficient information, the gap between their expectations and the real situation would be as small as possible. In other words, the satisfaction of shopping experience would increase.

Customer Sentiment

From the observation from the comments from both websites, the customers are mainly concern on three aspects, which includes product quality or function, customer services, and price discount.

Conclusion

To do further analysis, HKTVMall may collect more customer information about the evaluation of shopping experiences. For Suning, if it can open more hidden data for public, the more deeply investigation can be found.

Reference

1.https://ir.hktv.com.hk/eng/ir/presentations/pre170919.pdf
2.http://w2.cedars.hku.hk/careersfair/2017/resources/files/Hong%20Kong%20Suning%20Commerce%20Co.,%20Limited/Hongkong%20Suning_job.pdf
3.https://en.wikipedia.org/wiki/Cosine_similarity

Appendix

HKTVMall

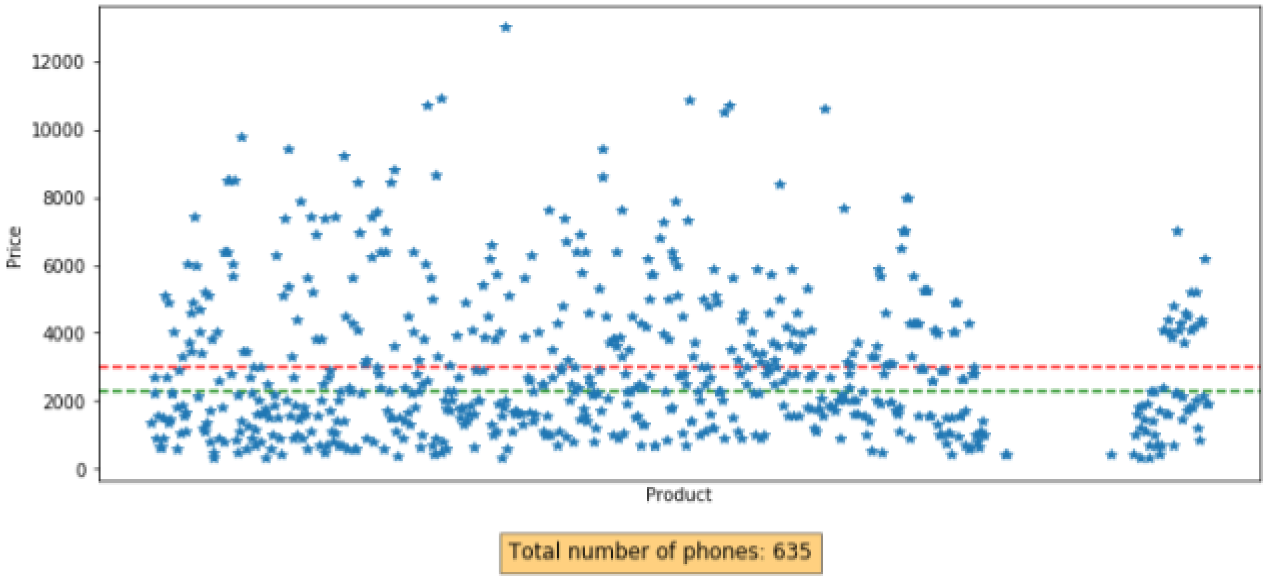

Figure 1

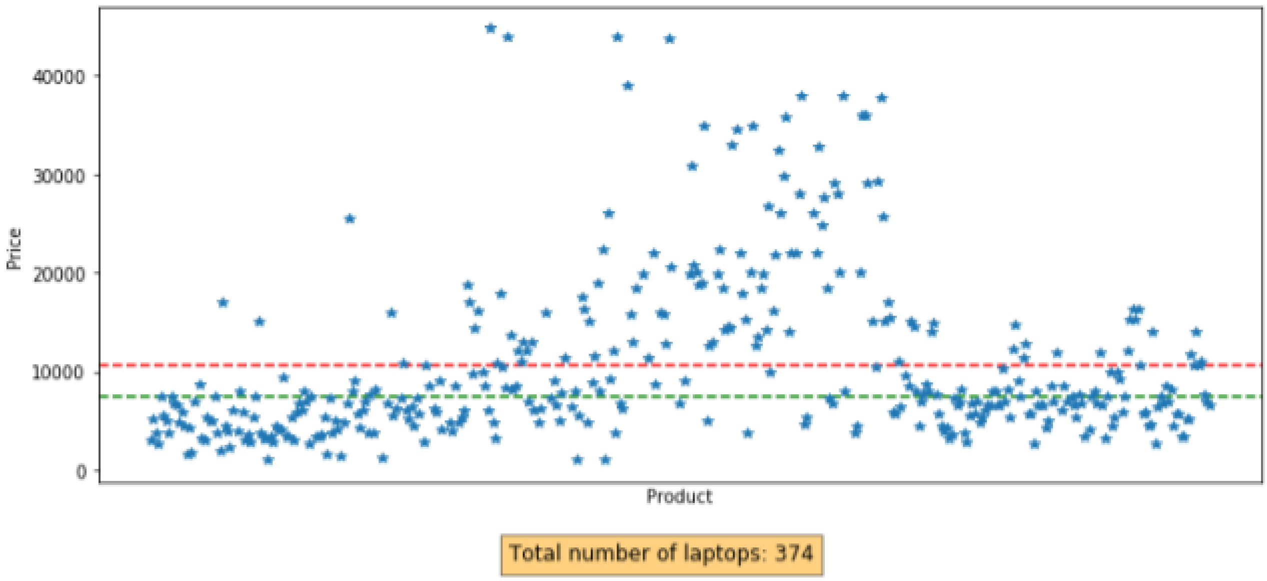

Figure 2

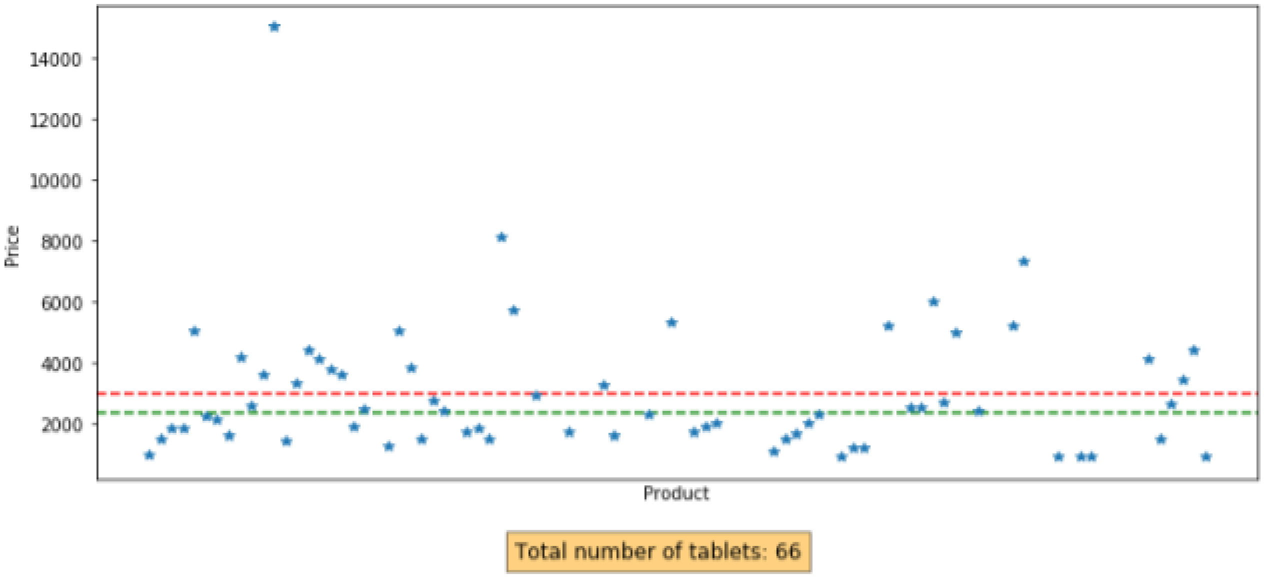
Price
14000
12000
10000
8000
6000
4000
2000
Product
Total number of tablets: 66

Figure 3

Suning

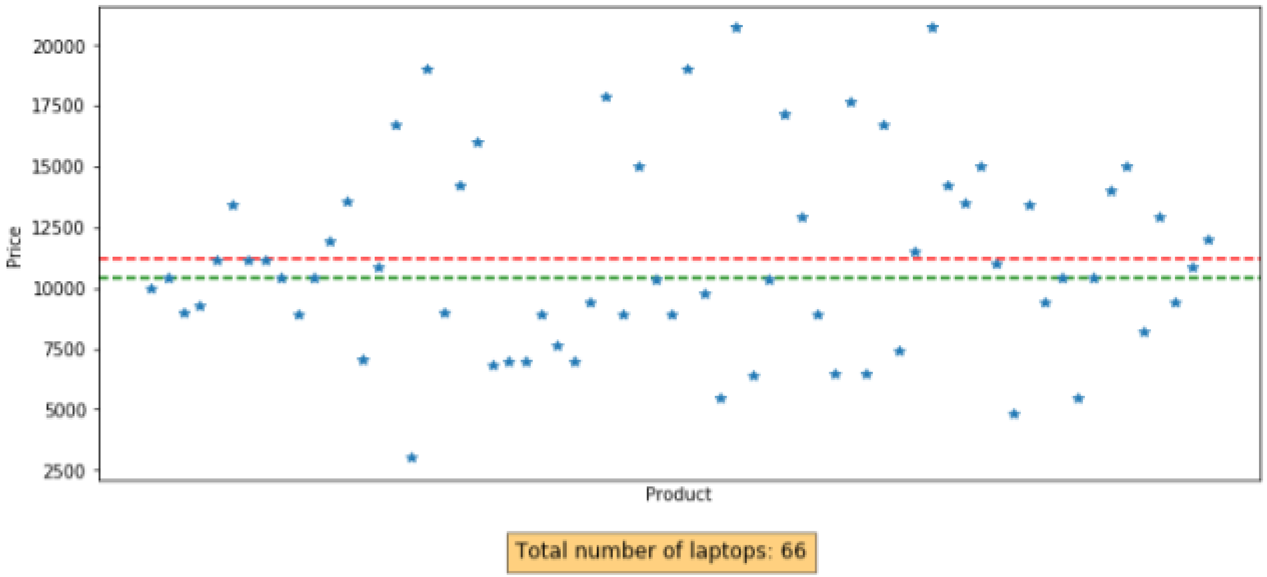

Figure 4

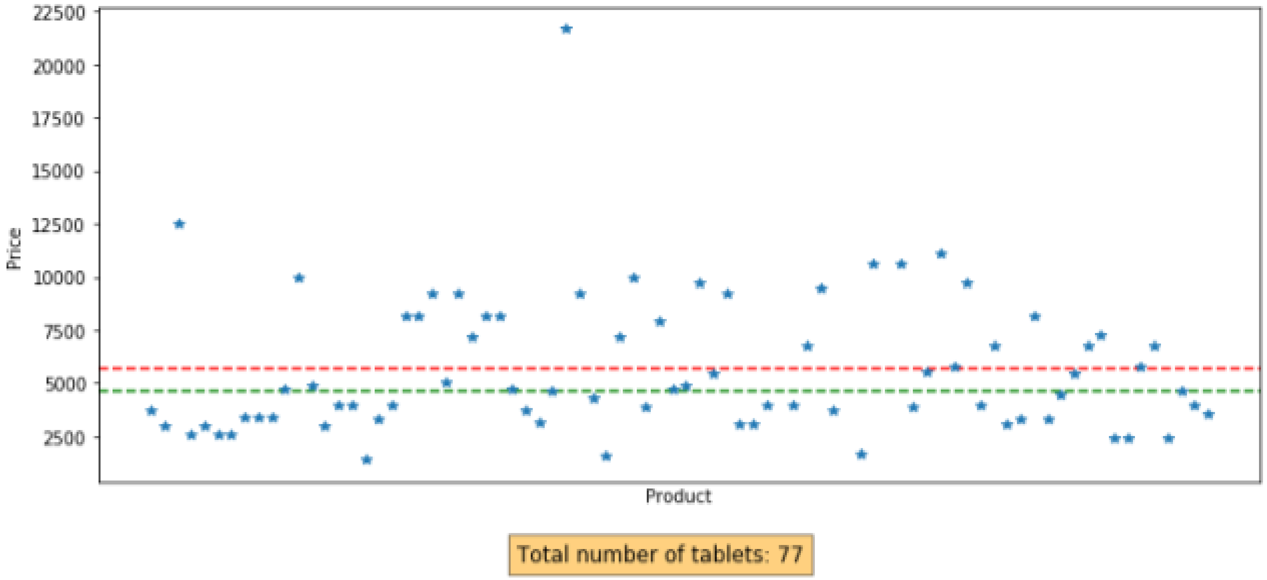

Figure 5

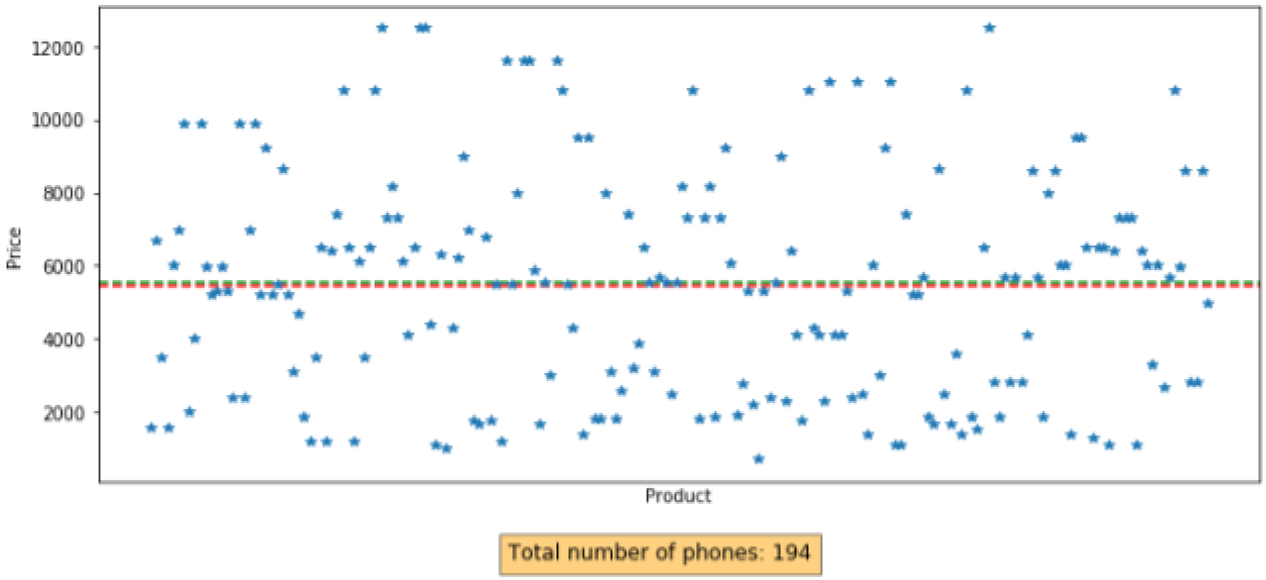
12000
10000
8000
6000
4000
2000
Price
Product
Total number of phones: 194

Figure 6

Wordcloud

比什麼爛蘋果性價比高
物流快快快快
好好用
很好
买家没有填写评价内容
买家没有填写评价内容 买家没有填写评价内容
买家没有填写评价内容 买家没有填写评价内容
买家没有填写评价内容 买家没有填写评价内容
買家没有填写评价内容 买家没有填写评价内容
價錢便宜售後服務優良
影相靚
买家没有填写评价内容 屏幕靓
此用户没有填写评价内容
买家没有填写评价内容 买家没有填写评价内容
买家没有填写评价内容 买家没有填写评价内容
买家没有填写评价内容
此用户没有填写评价内容
非常好
买家没有填写评价内容
VeryGood好
手机很好很强大
此用户没有填写评价内容

Figure 7

濾網沒有包裝袋
包裝很好
是兩腳插頭
provided
準時發貨
hour
其他幾好
Good
連線容易
adaptor
Late
half
delivery
有極大安全風險
一開機手機會提示此apps
兼且電力消耗快
manager
左右
螢光幕會不斷彈出停用通知
手機裏預載兩個有問題的apps
正常使用下午四時電量只剩20
若果啟用此apps
其中一個是file
電池很易熱滾燙
若果停用此App
good
great
far
still
word
though
app
given
MS
efficient
instead
good
days
Works
almost
ordering
C302
快靚正
google
esp
store
install
price
Delovered
play
會再買
used
nice
done
doc
剛買了但不能charge
麻煩跟進
價值合理
價格實惠
good
laptop
不知是
quality
欠了火牛和線
thin
還是火牛的問題
fast
沒法使用用
充電線有反應
跟機放火牛和線的位置是空的
用作一般文書處理0k
light
貨品狀態比預期好
試了一天
Iphone
但性能
雖然有些問題
超值
外觀100但商家好快換貨解決了
新機一樣
其他與新機無異
抵讚
好快送到
一流後備機
發貨準時
整體不錯
原來是日版
抵用
很快收到貨
除柵背有一小點花痕外
designed
still
易攜帶
Will
好用
輕便易用
sale
Excellent
輕巧
well
two
GOOD
buy
設計簡潔俐落
網購咁耐第一次見到咁離譜大譜商家
很好
虛假信息誤導消費者
咁樣明產品具體信息型號
Day
極度差評
根本頭尾支架就套唔實個電話
唔錯
中伏
本人覺得打斜扣還好用
夾好像太大
已經塞合調節過好幾次都係不中用
請問呢個插頭可以點用
service
Poor
reply
sales
但是兩腳插頭
follow
剛收到貨
用唔到
請問要怎麼樣購買新的Filter
頭先開咗盒發現個插頭唔係香港嗰種插頭

Figure 8