Online Shopping Platform Product Consumption Comparison

DSCI 510

Team Members:
Yiming Xiong (xiongyim@usc.edu 1057091668)
Haosen Guo (haosengu@usc.edu 7183115772)

1. Introduction

Online shopping platforms such as Newegg often list the same product from multiple sellers, with noticeable differences in price, buyer ratings, and sales volume. This makes it difficult for consumers to make a quick and informed choice. We developed a Python-based web crawler and data analysis tool targeting popular electronic products like the RTX 5090 graphics card and 2TB hard drive. This tool visualizes seller information from Newegg, providing users with a more intuitive and transparent purchasing guide.

2. Data

This project uses two product categories (NVIDIA RTX 5090 Graphics Cards and 2TB Solid State Drives) from Newegg.com, an online e-commerce platform specialized in computer hardware. All data was collected through web scraping.

NVIDIA RTX 5090 Graphics Cards: GeForce RTX 5090 GPU / Video Graphics Cards | Newegg

2TB Solid State Drives (SSD): 2tb ssd | Newegg.com

For these two products, we crawled the search results pages and extracted structured information such as product title, brand, price, customer rating, and number of reviews; some even included shipping information. For NVIDIA RTX 5090 Graphics Cards, we crawled eight pages (288 data entries) of product websites because we found that much of the data was empty during the crawling process. Therefore, we increased the amount of data crawled to ensure that we had enough valid data for analysis and visualization. However, for the 2TB SSD data, there was a large amount of high-quality valid data, so crawling only two pages (89 data entries) was sufficient.

3.1. Data Cleaning

The data cleaning part is performed mainly using Clean.py and Classify.py, which processes raw scraped data from Newegg.com for two product categories: RTX 5090 graphics cards and 2TB SSDs.

Actually, part of cleaning and filtering are included in Fetch.py. Because the Fetch.py will generate raw html (and json if want. The saving of JSON has been commented out, so it will not run.) and csv files. Here come explanations following the code running order.

After using *beautifulsoup4* to obtain html pages, the built-in functions (such as get; select_one; search) will be used to extract the product titles, URLs, prices, ratings, review counts, shipping fees, and brand information. Among them, the titles and brands are very easy to obtain. As for the price, shipping cost, rating and the number of reviews, since the labels stored in HTML are uncertain and most of the numbers are with decimal points, this project has built separate functions to filter multiple labels of different name formats, and uses regular expressions to obtain the accurate numbers. Finally, the data will be stored in a raw CSV file and then further data cleaning will be carried out.

In Clean.py, GPU and SSD data will be removed rows with any missing values applying *dropna().* The GPU data also needs to be filtered based on the condition that includes the keyword "Graphics Card". Because the data obtained from the graphics card may include models like the 5090 laptops, but the specific data we need is just for the 5090 graphics card.

Classify.py employs a hierarchical keyword-matching approach with four primary categories: Water Cooled Flagship, Air Cooled Flagship, Game-enhanced, Basic, Uncategorized. However, due to the different naming rules of various manufacturers, actually classifying them is quite difficult. We create different lists based on the naming conventions of different manufacturers or the special abbreviations they use to match the titles.

3.2. Data Analysis

The results obtained from analysis.py only cover the most important parts. For detailed information, please refer to the txt file named "analysis_results.txt" generated by output. This report provides analytical conclusions based on the original data, serving as a quantitative basis for subsequent market insights and business decisions.

Descriptive statistical analysis: Quantitatively describe the basic characteristics of the RTX 5090 GPU and 2TB SSD market. This includes calculating the total number of products, the number of brands, and core price indicators (average, standard deviation, range, etc.).

Brand Competition Analysis: Evaluate the market performance and price positioning of each brand. The prices of GPUs from different manufacturers are relatively close to each other, while for SSDs, Samsung holds a dominant position in terms of both price and market share.

Statistical hypothesis testing: A one-way analysis of variance (ANOVA) was used to verify whether the price differences among different GPU categories were statistically significant. The results showed that the F-statistic was 0.6762, and the p-value = 0.5778 > 0.05, indicating that at the 95% confidence level, the price differences among different categories were not significant.

Outlier detection for price analysis: The interquartile range (IQR) method was used to identify outliers. The results showed that there were 8 price-abnormal products in the SSD market. Meanwhile, the prices of GPU products were more concentrated and stable.

3.3. Visualization

We have also visualized the average prices of the products (5090 graphics cards by category, and SSDs by brand) for easier observation. We have also visualized the number of products in each 5090 category and the market share of the two product brands. Due to page limitations, we won't show all the visualizations here, but you can find the results in the `data\images` folder.

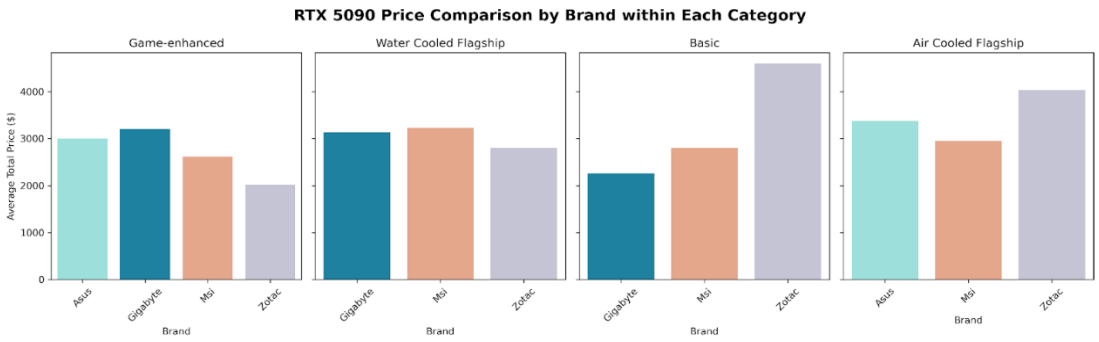


Figure 1. RTX 5090 Price Comparison Visualization

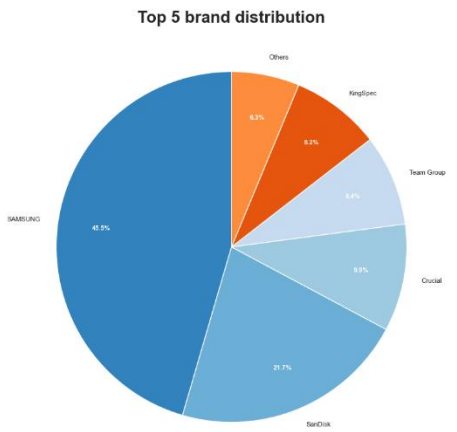
Figure 1 above shows a price comparison for each brand within the same product category. We can clearly see that Gigabyte's higher-end products are priced higher than other brands, while Zotac's mid-range and lower-end products are also relatively expensive.

Figure 2 . Price and Sales by Brand line graphs

Then we created line graphs of the price and sales volume for each product, as shown in Figure 2, to see how sales changed as the price increased. We expected that sales would be higher at lower and medium price points, so the curve should generally show a downward trend. However, in our visualization results, the sales volume for each price point of the Gigabyte and Asus brands was the same. We then checked the Newegg website and found the same thing: regardless of the model of the 5090 Gigabyte, the sales volume was always 120. The same was true for Asus. Therefore, we suspect that they set all the sales data to be the same.

Figure 3. Price and Sales by Category line graphs

We also created line graphs showing the price and sales volume for each category of GPUs, as shown in Figure 3. It can be seen that for almost every category, the price range slightly above 3000 yuan has the highest sales volume.



Regarding the pie chart analyzing market share, because we suspect there are errors (the sales volume for each Gigabyte product was 120) in the sales figures for the 5090 product, we will only analyze the SSD 2TB data here. So for the 2TB SSD, we visualized the sales pie of the top 5 brands and considered the others as one, as shown in figure 4. The result is exactly as obtained from the above analysis.

Figure 4. 2TB SSD brand distribution

3.4. Conclusion

By combining analysis and visualization, it can be concluded that the prices of various GPU products have stabilized, and the market is relatively balanced, excluding anomalies. While for SSDs, Samsung holds the majority share and there is a significant price difference. Furthermore, based on the ANOVA statistical test, it can be concluded that there is no significant correlation between brand classification and price.

4. Changes from Original Proposal

Compared to the original proposal, we have added one more product. At the same time, we strengthened the original data cleaning process, then selected the data that needed to be further classified in a more detailed manner, and finally conducted data analysis and visualization.

The original data contains a great deal of information that is not necessary for the analysis. There are various issues such as differences in product classification and naming for different goods. Moreover, the specific solutions and their details have all been mentioned in the previous text.

5. Future Work

Time series analysis: The current analysis merely provides a snapshot of the market conditions at the time of data collection. Given the significant fluctuations in the prices of graphics cards and storage devices, establishing an automated scheduling script to collect longitudinal data would be of great value. Through time series analysis, we can track the changing trends of prices over time and predict future price movements.

Multi-product Integration: Combining the analysis of different products, such as GPUs, CPUs, monitors, and other accessories, and providing price position analysis, thereby helping users select appropriate product combinations.