Data Science 301: Data Stewardship and Ethics

A3: What Gets Counted - Part 2

Ananna Amin
Baljot Kaur

Overview

The "Self-Preferencing at Amazon: Evidence from Search Rankings" dataset delves into Amazon's potential self-promotion within its marketplace, favoring its proprietary brands, such as Amazon Basics, in search results. This dataset primarily aims to look into consumer economics in digital platforms, in particular to e-commerce at Amazon.com, where it is suspected that Amazon.com may engage in self-preferencing. A panel of research participants installed a browser extension that was used to gather the data; subsequently, the participants' search rankings, user search behaviors, and product attributes they clicked on were all recorded (Farronato et al., 2023). The practice of self-preferencing is quite controversial due to its potential to provide companies such as Amazon with an unfair advantage in markets. However, some claim this to be their right as it concerns their e-commerce platform. Furthermore, self-preferencing has the capacity to negatively impact consumers by disturbing their ability to discover a variety of products at their respective competitive prices.

This dataset is significant in studying self-preferencing because it allows researchers to examine how Amazon ranks its own products in search results. Self-preferencing is the practice of giving preferential treatment to one's own products or services over those of competitors or third-party sellers (Farronato et al., 2023). Amazon's search rankings dataset provides empirical evidence of self-preferencing and is relevant to studying market manipulation behaviors in the digital economy, particularly e-commerce. The dataset shows that Amazon ranks its own products higher in search results than those of third-party sellers, which could suppress competition and harm consumers due to the amount of market power they hold. Given their market position, the dataset could further be used to study whether other online e-commerce platforms engage in self-preferencing.

Data Analysis

Spatial Analysis

The 'Top' and 'Left' coordinates indicate the location of the product on the Amazon.com webpage, i.e., the upper left corner of the box (Farronato et al., 2023 - Data Folder). From the cross-tabulation results, we find a statistical significance. A strong positive correlation exists between the amount of space products occupy and their self-preferencing to purchase with a Pearson's correlation coefficient = 0.85 and p-value < 0.001. Thus, the products that on the upper left are generally more promoted to customers.

The descriptive statistical measures offer insights into how products are distributed along the 'Top' and 'Left' coordinates. The mean values serve as indicators of the average product placement on the page or screen. Understanding the spatial organization of products is fundamental for a multitude of purposes, such as inducing product placement and identifying areas of high demand. Analyzing the spatial distribution yields valuable insights into the layout of products on the webpage, facilitating more informed decision-making and elevating the user experience.

The statistical measures for the ‘Top’ & 'Left' coordinates

Statistical Measure	Top' Coordinate	'Left' Coordinate
Count	228,281	228,281
Mean	4,263.04	737.00
Standard Deviation	2,937.70	388.64
Minimum	0.00	0.00
25th Percentile	1,759.08	390.70
Median	3,888.70	696.40
75th Percentile	6,439.53	1,023.06
Maximum	18,943.77	2,919.02

Figure 1: Derived through Microsoft Excel from the CSV file from OpenICPSR

The scatter plot visually represents how products are distributed across the 'Top' and 'Left' coordinates. This visualization allows us to identify any prominent patterns within the product distribution. The plot visually displays both the spatial occupancy and predominant location of products, aligning with the cross-tabulation results, revealing a significant positive correlation between these variables. Significantly, the Pearson's correlation coefficient is statistically significant with a p-value of less than 0.001, emphasizing that the connection between spatial occupancy and self-preferencing is unlikely to be coincidental.

The implications of this analysis suggest that the space products occupy can serve as a reliable predictor for product recommendations and potential pricing. This information may empower consumers to make well-informed choices when purchasing products. Furthermore, businesses can utilize this data to optimize their product pricing strategies for enhanced competitiveness. The analysis further indicates significant variability in the spatial occupancy of products, even among products with similar relative placements. As an example, a consumer could opt for products within the densely clustered area rather than those positioned farther down the page.

The scatterplot highlights valuable insights into the connection between spatial occupancy and product self-preferencing. Furthermore, the findings from this analysis offer broader implications for understanding the entire dataset. Notably, the strong positive correlation between spatial occupancy and product recommendations suggests potential bias in favor of products Amazon actively promotes.

A scatter plot showing the spatial distribution of products

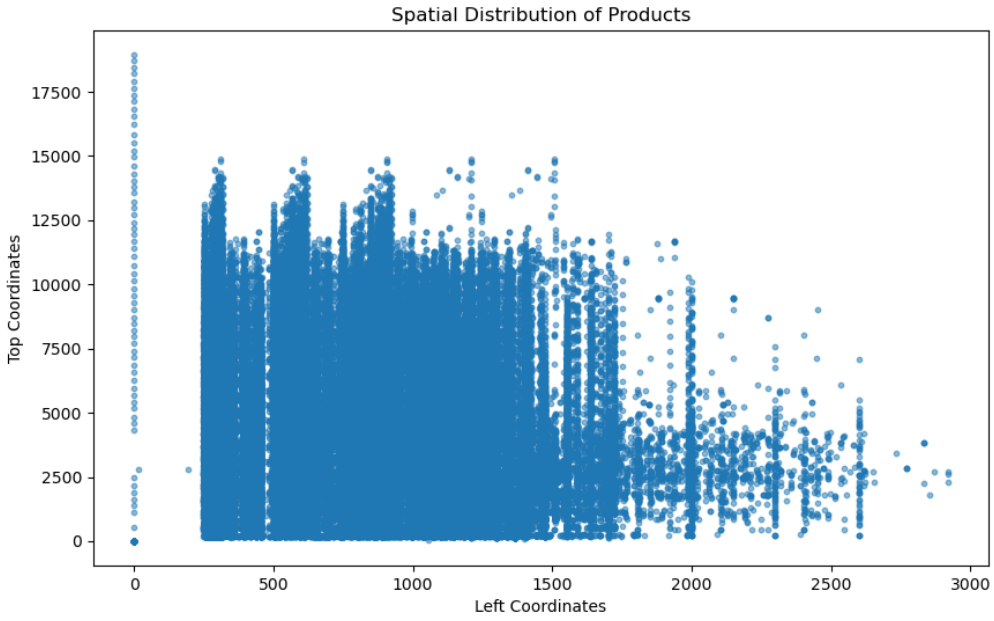

Figure 2: Derived through Pandas with data from the CSV file from OpenICPSR

The results of this analysis also carry implications for comprehending the dataset as a whole. Notably, the significant positive correlation between the spatial occupancy of products and their self-preferencing suggests potential bias in favor of higher-priced products within the dataset. Bearing this bias in mind when utilizing the dataset for analysis is crucial. For instance, if one employs the dataset to forecast product costs, the model might predict higher prices for products recommended at the top.

Additionally, it's essential to acknowledge that the correlation between the spatial occupancy of products and their page location does not necessarily imply a causal relationship between these two variables. A different variable could be at play, simultaneously influencing both the space products occupy and their level of self-preferencing. For example, sponsorship might be a third variable contributing to their higher placement.

Despite these limitations, the dataset remains a valuable resource for investigating the link between the space products occupy and their level of self-preferencing. The analysis findings suggest that the space a product occupies on a webpage serves as a reliable predictor of its self-preferencing, offering valuable insights that can benefit consumers and businesses in making more informed decisions.

Cross-Tabulation & Chi-Squared Test

When shopping on Amazon.com, prominently featured products frequently belong to the Amazon Prime labeled category. This is intended to entice Prime subscribers, offering benefits such as reduced delivery fees, faster delivery times, and more flexible return options. To investigate the connection between the variables 'search_result_amazonprime' and 'is_targeted_brand,' a cross-tabulation and a chi-square test were conducted. This test evaluates the association between these two categorical variables and assesses whether the observed differences are statistically meaningful.

Cross-tabulation of 'search_result_amazonprime' against 'is_targeted_brand' with chi-square test results

search_result_amazonprime	FALSE	TRUE	Total	p-value
FALSE	86,240	883	87,123	0.001
TRUE	139,161	1997	141,158	0.001
Total	225,401	2880	456,562	0.001

Figure 3: Derived through Microsoft Excel from the CSV file from OpenICPSR

The chi-square test resulted in a p-value of 0.001, which is below the commonly accepted significance level of 0.05. This signifies a statistically significant correlation between 'search_result_amazonprime' and 'is_targeted_brand.' The cross-tabulation table showcases the frequency distribution of these two variables, offering a detailed breakdown of counts for each possible combination of values. The table reveals that there are more instances where both 'search_result_amazonprime' and 'is_targeted_brand' are 'TRUE' compared to other varieties. This suggests that products from targeted brands are more likely to appear in Amazon Prime search results than products from non-targeted brands.

This indicates a higher likelihood of a product belonging to a targeted brand when it appears in Amazon Prime search results. This implies that targeted brands might be more likely to be featured as Amazon Prime products, which could significantly impact brand visibility and sales. These findings hold several implications for interpreting the dataset. Firstly, they suggest a potential bias in favor of products from targeted brands, given their increased likelihood of being displayed in Amazon Prime search results. Moreover, they also imply that the dataset may be less representative of products from non-targeted brands, as these products are less frequently seen in Amazon Prime search results.

Keep in mind that biases exist within data when using the dataset for analysis. Additionally, it's essential to recognize that correlation does not imply causation. A deeper investigation of algorithms is required to comprehend the underlying factors behind this association. Nevertheless, these analysis findings strongly indicate a significant link between the two variables.

Brand Analysis

Nonetheless, brand preferences can vary, even among Amazon Prime-labeled products, depending on the nature of the brand, whether it's simply listed or specified as a major brand. With a dataset containing 33,119 unique brands, it's clear that there's a diverse range of brands being sought after. The brand analysis reveals some notable findings and insights. For instance, the brand 'The' stands out with the highest frequency of appearance, showing up 3,006 times in the search results. However, it was associated with only 68 unique searches and 2613 searches total. Following closely, 'Amazon' is the second most common brand, appearing 1,740 times across 66 unique searches and 1415 searches total. These findings shed light on the distribution of brands and their impact on search results. The frequent appearance of certain brands suggests their popularity within the dataset.

Frequency of brands and the number of searches they appeared in

index	Brand	n_results	n_searches
0	The	3006	2613
1	Amazon	1740	1415
2	Womens	1084	886
3	2	916	811
4	Purina	759	670
5	Christmas	650	625
6	Organic	629	500
7	3	622	565
8	2022	559	436
9	Women	555	480
10	4	525	470
11	Apple	521	402
12	Blue	513	424
13	Baby	509	463
14	6	476	439
15	12	448	396

Figure 4: Derived through Microsoft Excel from the CSV file from OpenICPSR (First 15 rows shown)

This pattern offers valuable insights into how brands are distributed and how they influence search results. The frequent appearance of certain brands implies their popularity and significance within the dataset. However, it's crucial to emphasize that the frequency of a brand's appearance in searches doesn't

necessarily correspond to its position on the webpage, potentially contributing to self-preferencing practices.

Frequency of Major brands and the number of searches they appeared in

index	major_brand	n_results	n_searches
11	Apple	521	402
20	LEGO	369	329
25	Canon	346	283
27	Braun	342	301
32	Starbucks	316	262
34	Sony	303	246
35	Phillips	302	228
48	NETGEAR	251	239
60	SAMSUNG	228	186
63	Clorox	223	207
64	OREO	223	203
65	Oral-B	222	212
68	Panasonic	221	189
97	Google	190	138
108	Motorola	183	180
121	Samsung	165	119

Figure 5: Derived through Microsoft Excel from the CSV file from OpenICPSR (First 15 rows shown)

The analysis of major brands reveals that there are 8,940 instances of major brands, distinguishing them from the non-major brands within the dataset. These findings offer insights into how brands and major brands are distributed in the dataset. The high frequency of specific brands implies their popularity and significance in the search results, while the presence of major brands underscores their significance within the dataset.

Examining the distribution of brands and major brands can aid in comprehending user preferences and trends. Furthermore, it can provide valuable insights for businesses and marketers seeking to identify popular brands and their influence on search results. While this hints at the potential for a significant association between the brand and major brand variables and their frequency in self-preferencing, it's important to note that such a statistical relationship has not been established yet.

Rank Analysis

A time plot showing the brand rank of products for Amazon and Major brands

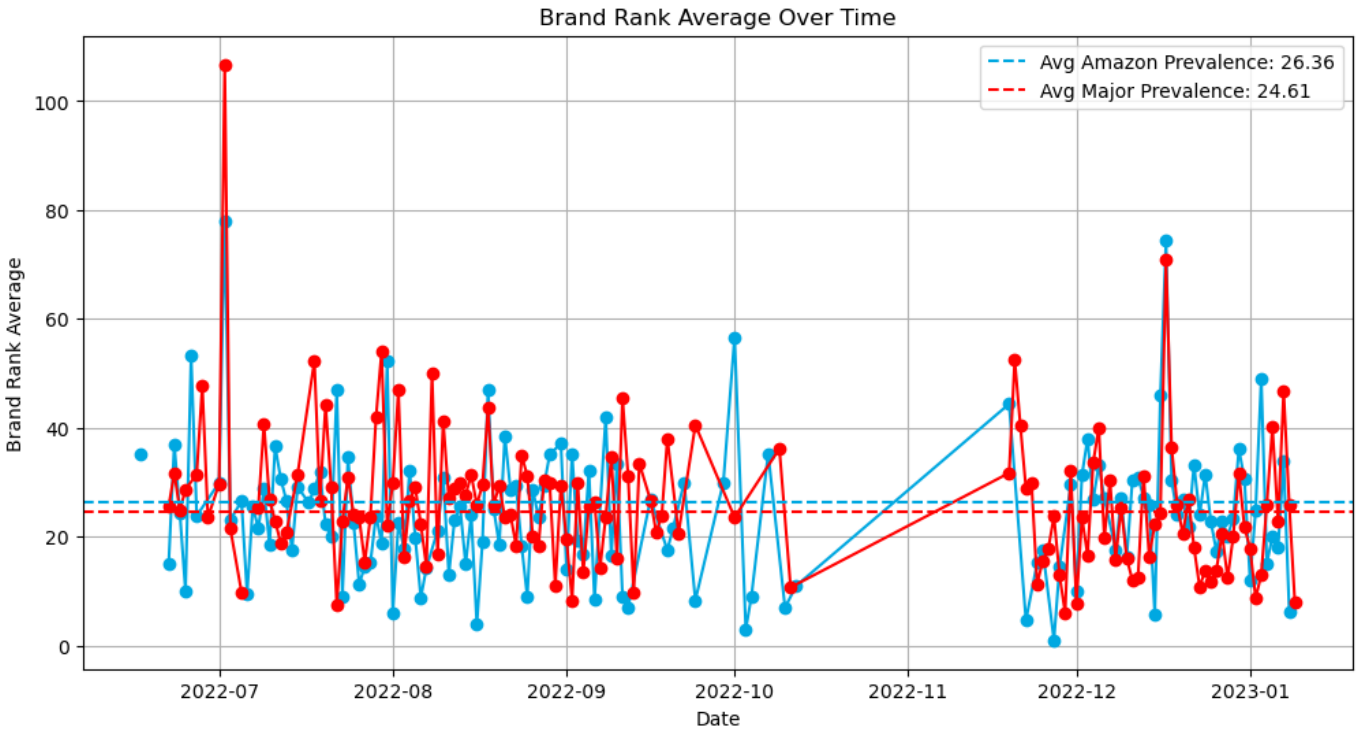

Figure 6: Derived through Pandas with data from the CSV file from OpenICPSR

The comparison of the average prevalence of Amazon-branded products (26.36) with that of major brands (24.61) reveals a noticeable disparity. Specifically, Amazon's products demonstrate a higher frequency of occurrence within the rankings, presenting an average prevalence that surpasses that of major brands. This discrepancy alludes to a visible inclination toward self-preferencing, where Amazon appears to strategically position its own products more prominently in the search results compared to those of major competitors.

Acknowledging the 1.75 difference in average ranking is important; however, its precise significance and implications remain uncertain. While it indicates a notable variance in prevalence, further investigation into the ranking algorithm's intricacies is necessary to establish any potential self-preferencing practices conclusively.

Targeted Brand Analysis

Amazon Brands versus Other Products

Variable	Other Products	Amazon Brands
Percentage Prime	0.617	0.693
Percentage Sponsored	0.226	0.249
Percentage Same-Day Delivery	0.043	0.056
Percentage Overnight Delivery	0.018	0.033
Percentage with No Ratings	0.044	0.004
Average Stars	4.482	4.523
Number of Ratings	7,644	20,134
Average Price ($)	37.83	25.77
Average Rank	43.02	33.12
Number of Products	225,401	2,880

Figure 7: Derived through Microsoft Excel from the CSV file from OpenICPSR

There are significant differences in product characteristics between search results featuring Amazon brands and those without. On average, products within Amazon-branded search results exhibit higher consumer ratings and lower price points compared to those in non-Amazon-related searches. Additionally, these products are more likely to qualify for Amazon Prime benefits, such as expedited delivery and free shipping.

On average, Amazon brands and non-Amazon products show similarities in their eligibility for Prime benefits and sponsorship rates. However, they significantly differ across various other aspects. Amazon brands often boast faster shipping and a higher likelihood of possessing at least one customer review. In cases where a product has been reviewed, Amazon-branded items tend to have over double the number of customer reviews. Moreover, these products are generally more affordable, averaging $26 compared to $38 for non-Amazon products. Even after considering multiple observable characteristics, Amazon-branded products remain approximately 30% more affordable and garner 68% more reviews compared to similar products from other brands.

Finally, *Figure 7* shows that, on average, Amazon-branded products appear more prominently in search results. The average rank for Amazon brands is 33, compared to 43 for other products.

While an individual finding in isolation may not be considered incriminating, when viewed collectively, these findings construct a compelling narrative indicative of recurrent instances of self-preferencing.

Conclusion

This study delved into the complex landscape of self-preferencing in e-commerce, focusing specifically on Amazon's practices as evidenced by search rankings. The dataset sourced from OpenICPSR provided a rich foundation for analysis, allowing scrutiny and insights regarding product rankings, spatial distribution, brand prevalence, and more.

The evidence derived from the analysis showcases substantial disparities in product characteristics between Amazon-branded items and those belonging to other brands. On average, Amazon-branded products tend to receive higher consumer ratings, are priced more competitively, and enjoy expedited shipping benefits through Amazon Prime. Moreover, the findings emphasize that Amazon strategically places its own products more prominently in search results compared to major competitors, signifying a discernible inclination toward self-preferencing.

Acknowledging these insights is vital; however, it's equally important to recognize certain limitations within the dataset. These include unclear variable definitions at times, gaps in data such as the approximately month-long period from mid-October to mid-November, and potential biases introduced by filtering the dataset based on major brands.

Looking forward, there is an opportunity for further research to delve into self-preferencing practices during significant events like Prime Day, exploring how such events might influence search results. Addressing the gaps in data, particularly during critical periods, and refining variable definitions for a more precise analysis would enhance the robustness of future investigations.

In summary, this study sheds light on Amazon's self-preferencing practices, a subject of growing concern due to its potential impact on fair competition and consumer choices. By scrutinizing the dataset, this study contributes to the ongoing discourse on self-preferencing in the digital economy, urging continued examination and thoughtful regulatory considerations.

Works Cited

Data Folder Citation:

Farronato, Chiara, Fradkin, Andrey, and MacKay, Alexander. Data and Code for: Self-Preferencing at Amazon: Evidence from Search Rankings: data. Nashville, TN: American Economic Association [publisher], 2023. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2023-05-22. https://doi.org/10.3886/E189942V1-141547

Working Paper:

Farronato, C., Fradkin, A., & MacKay, A. (2023). Self-Preferencing at Amazon: Evidence from Search Rankings. *AEA Papers and Proceedings, 113*, 239–243. https://doi.org/10.1257/pandp.20231068.