Report 1: Data Wrangling and Preparation

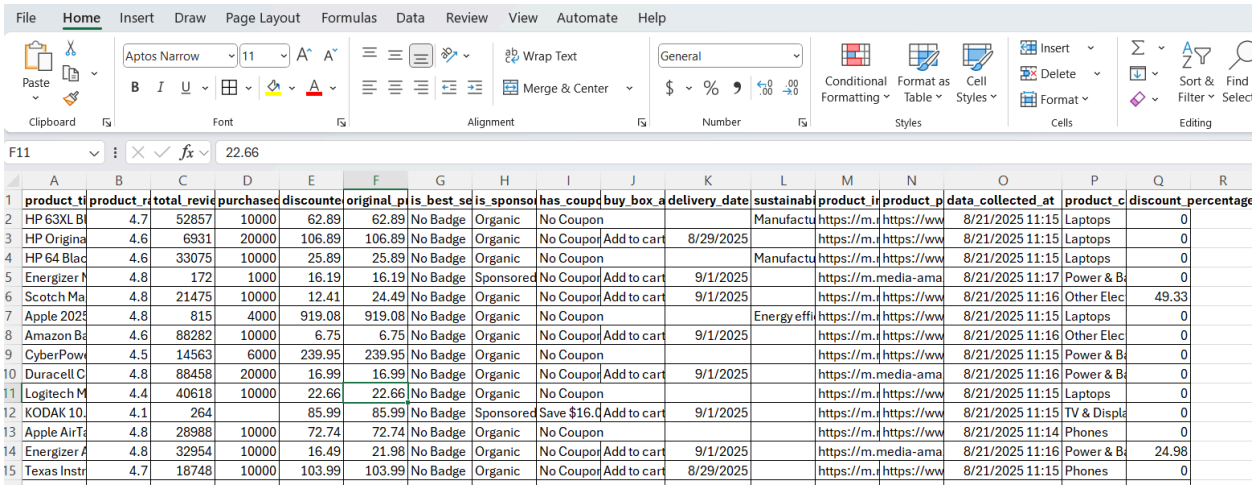
Amazon Product Sales Data

BUS 211A-2: Foundations of Data Analytics
Brandeis University – Fall 2025
Team Members: **Asef Haque, Bora Chaush, Emiraldo Prifti, Hanyu Jia, Mason Marathias**

Date: September 12, 2025

1. Business Context & Problem Definition

As Amazon's retail data consultants, our job is to assist the company to get more out of its huge sales data by organizing it in a useful way. Amazon sales data for its products can tell us a lot about how people shop, how they set prices, and how well the products work. For a company like Amazon, looking at this data can really help them find sales trends and see how well promotions are really working, which will help them make decisions about inventory and marketing. Amazon's own success is a great example of how powerful a data-driven strategy can be. For example, its recommendation engine, which uses data from over 150 million consumer purchases, makes up about 35% of sales. This shows why it is very important to have good, well-prepared data when making important decisions.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	product_ti	product_r	total_revie	purchased	discounte	original_p	is_best_se	is_sponso	has_coupo	buy_box_a	delivery_date	sustainabi	product_ir	product_p	data_collected_at	product_c	discount_percentage	
2	HP 63XL Bl	4.7	52857	10000	62.89	62.89	No Badge	Organic	No Coupon			Manufactu	https://m.r	https://ww	8/21/2025 11:15	Laptops	0	
3	HP Origina	4.6	6931	20000	106.89	106.89	No Badge	Organic	No Coupon	Add to cart	8/29/2025		https://m.r	https://ww	8/21/2025 11:15	Laptops	0	
4	HP 64 Blac	4.6	33075	10000	25.89	25.89	No Badge	Organic	No Coupon			Manufactu	https://m.r	https://ww	8/21/2025 11:15	Laptops	0	
5	Energizer N	4.8	172	1000	16.19	16.19	No Badge	Sponsored	No Coupon	Add to cart	9/1/2025		https://m.media-ama		8/21/2025 11:17	Power & Ba	0	
6	Scotch Ma	4.8	21475	10000	12.41	24.49	No Badge	Organic	No Coupon	Add to cart	9/1/2025		https://m.r	https://ww	8/21/2025 11:16	Other Elec	49.33	
7	Apple 2025	4.8	815	4000	919.08	919.08	No Badge	Organic	No Coupon			Energy effi	https://m.r	https://ww	8/21/2025 11:15	Laptops	0	
8	Amazon Ba	4.6	88282	10000	6.75	6.75	No Badge	Organic	No Coupon	Add to cart	9/1/2025		https://m.r	https://ww	8/21/2025 11:16	Other Elec	0	
9	CyberPow	4.5	14563	6000	239.95	239.95	No Badge	Organic	No Coupon				https://m.r	https://ww	8/21/2025 11:15	Power & Ba	0	
10	Duracell C	4.8	88458	20000	16.99	16.99	No Badge	Organic	No Coupon	Add to cart	9/1/2025		https://m.media-ama		8/21/2025 11:16	Power & Ba	0	
11	Logitech M	4.4	40618	10000	22.66	22.66	No Badge	Organic	No Coupon				https://m.r	https://ww	8/21/2025 11:15	Laptops	0	
12	KODAK 10.	4.1	264		85.99	85.99	No Badge	Sponsored	Save $16.0	Add to cart	9/1/2025		https://m.r	https://ww	8/21/2025 11:15	TV & Displa	0	
13	Apple AirTa	4.8	28988	10000	72.74	72.74	No Badge	Organic	No Coupon				https://m.r	https://ww	8/21/2025 11:14	Phones	0	
14	Energizer A	4.8	32954	10000	16.49	21.98	No Badge	Organic	No Coupon	Add to cart	9/1/2025		https://m.media-ama		8/21/2025 11:16	Power & Ba	24.98	
15	Texas Instr	4.7	18748	10000	103.99	103.99	No Badge	Organic	No Coupon	Add to cart	8/29/2025		https://m.r	https://ww	8/21/2025 11:15	Phones	0	

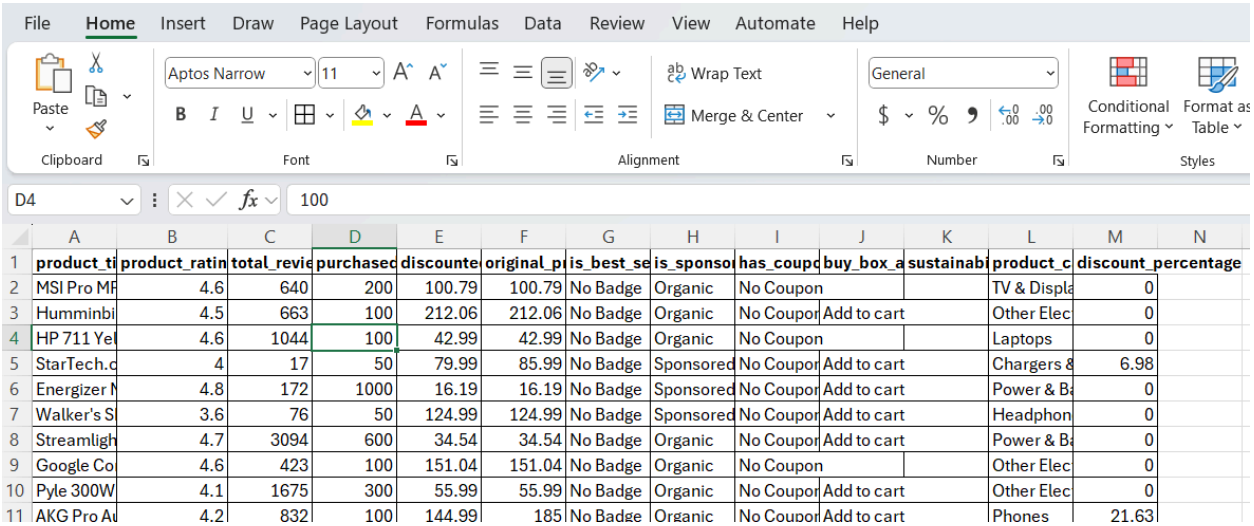
Figure 1: Original Columns

The issue that we tackle is that raw sales data frequently arrives in substantial and disordered formats, obstructing straightforward analysis. We chose a dataset from Kaggle that had a month's worth of Amazon product sales data. It had 42,676 rows and 17 columns, although some of the fields were not useful or were missing information. It is hard to work with a dataset this big and complicated in a classroom or demo context. We had to set a clear scope and use data cleansing and reduction to make the data easier to work with and more useful for our analysis. The purpose of our consultation is to show how competent data wrangling can make data better and easier to use, which will help businesses make better decisions. We will show you how we have changed the raw Amazon sales data into a smaller, clearer, and more useful form that can be used for analysis and to get more useful information.

2. Dataset Description

The dataset originates from Kaggle and is called "Amazon Product Sales for a Month." It shows a snapshot of Amazon's product sales listings that were scraped from the Amazon website, with data broken down by its product category. Before any cleanup, the original dataset had 42,676 rows and 17 columns. After we cleaned it up, we kept 13 columns that were most important to our analysis. The

dataset is seen as reliable and very well organized because the columns are clearly labeled and prepared in the same way. To make sure everything was authentic, our team checked a sample of product names, pricing, and other details against current Amazon listings to make sure they matched. Also, the Kaggle community's positive feedback and conversations about the dataset showed that it was reliable for use in education and analysis.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	product_ti	product_ratin	total_revie	purchased	discounte	original_pi	is_best_se	is_sponso	has_coupo	buy_box_a	sustainabi	product_c	discount_percentage	
2	MSI Pro MF	4.6	640	200	100.79	100.79	No Badge	Organic	No Coupon			TV & Displa	0	
3	Humminbi	4.5	663	100	212.06	212.06	No Badge	Organic	No Coupor	Add to cart		Other Elec	0	
4	HP 711 Yel	4.6	1044	100	42.99	42.99	No Badge	Organic	No Coupon			Laptops	0	
5	StarTech.c	4	17	50	79.99	85.99	No Badge	Sponsored	No Coupor	Add to cart		Chargers &	6.98	
6	Energizer M	4.8	172	1000	16.19	16.19	No Badge	Sponsored	No Coupor	Add to cart		Power & Ba	0	
7	Walker's Sl	3.6	76	50	124.99	124.99	No Badge	Sponsored	No Coupor	Add to cart		Headphon	0	
8	Streamligh	4.7	3094	600	34.54	34.54	No Badge	Organic	No Coupor	Add to cart		Power & Ba	0	
9	Google Co	4.6	423	100	151.04	151.04	No Badge	Organic	No Coupon			Other Elec	0	
10	Pyle 300W	4.1	1675	300	55.99	55.99	No Badge	Organic	No Coupor	Add to cart		Other Elec	0	
11	AKG Pro Au	4.2	832	100	144.99	185	No Badge	Organic	No Coupor	Add to cart		Phones	21.63	

Figure 2: Retained Columns

Retained Columns (13): We concentrated on the fields that give us useful information about how well a product works and how much it costs. For analysis and SQL practice, the dataset kept the following important columns:

- **product_title** – The name of the item
- **product_rating** – The average rating from customers.
- **total_reviews** – The number of reviews that customers have written about the product.
- **purchased_last_month** – The number of items sold in the past month
- **discounted_price** – The price at which the item is currently for sale, minus any discounts.
- **original_price** – The price that was on the list before any discounts
- **is_best_seller** – A flag that shows whether the item is listed as an Amazon "Best Seller."
- **is_sponsored** – A flag that shows if the listing is a sponsored (advertisement) product.
- **has_coupon** – A flag that tells you if the product has a coupon
- **buy_box_availability** – Tells you if the product is in stock and ready to buy right away (in other words if it has the Buy Box).

- **sustainability_tags** – Tags that have to do with sustainability, including eco-friendly labels or certificates on the products.
- **product_category** – This is the type of product, like electronics, home, books, etc.
- **discount_percentage** – The percentage off, which is the difference between the original price and the discounted price
- These columns that we decided to keep include the most important things you need to look at to figure out how well sales are doing: Product identification, customer sentiment (ratings and reviews), sales volume, pricing and discounts, and special marketing flags (such best-seller or sponsored status).

Columns that were taken away: During the first cleansing, we got rid of a few columns because they were not very useful for our study or had private information in them. Taking them away made the dataset easier to work with without sacrificing any important information. The following columns were removed from the dataset:

- **image_url** – Which is the URL of the product image (not relevant for SQL queries or data analysis).
- **page_url** – The URL of the product page, but we don't need it for our needs.
- **date_of_data_collected** – The day the data was scared. This information is not important for looking at how well a product did in a month.
- We got rid of some noise and privacy risk in the data by eliminating these columns. For example, image_url and page_url don't help you figure out sales trends.
- After we cut down on the number of columns, the dataset contained 13 distinct columns that our team comprehended. The data looked consistent because the ratings were on a scale of 0-5, the prices were numbers in a common currency, and the entries in categorical variables like product_category made sense. These methods lay strong groundwork for more data cleaning and minimization.

3. Data Reduction Approach

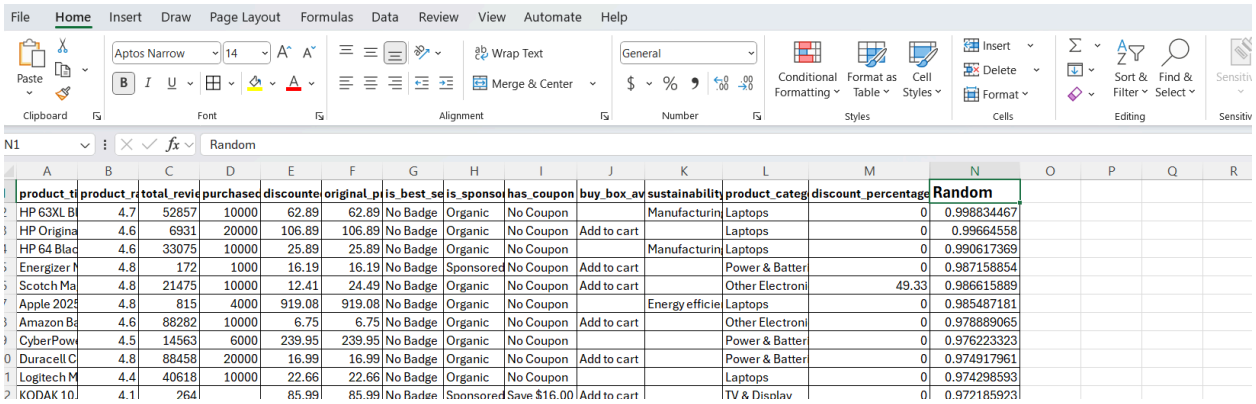
The dataset was too large to work with. We had more than 42,000 rows, and that's way too much for practical use. It can slow down tools like Excel and SQL, which we need to work on it. The process of making it so we can determine whether it's reliable or not drastically falls in efficiency. It takes way too much time to analyze and draw conclusions, as there are too many rows to check and go through. Even if we made some kind of mistake, it would take too much time to find and fix it. So, what we decided to do was to reduce it as much as we could, so we could work more accurately on it. It was reduced to up to 5,000 rows. We considered multiple sampling strategies for this reduction:

- **Filter by Product Category:** We decided to reduce the dataset by selecting categories. For example, we removed the small and niche categories and decided to keep only the big ones. We

did this by counting the rows in each category, then setting a threshold to only keep categories with at least a certain number of products or the top 20 categories by size. We believe that this was useful because a smaller and cleaner dataset means faster queries and easier analysis. Also, by focusing on major product groups, it is better for practicing SQL joins and group-by operations because they have plenty of data.

- **Filter by Purchase Volume/Amount:** For example, items with only 1–2 purchases last month or products that generate very low revenue. That's useful because it focuses on top-selling, high-value products and makes it easier to identify key revenue drivers and important products. The downside is that many lower-cost or niche products would be completely removed, and another thing is that the dataset might only reflect blockbuster products, ignoring smaller steady sellers that still matter to the business. Even so, it's worth the risk as it's easier to process, clean, and analyze without wasting time on products that barely matter to the overall revenue.
- **Random Sampling:** So, this method is about taking a random sample of about 10–15% of all rows. That way, we are keeping a smaller dataset without losing the characteristics of the original. This way, every row has a chance to be included, and it also works well when you want a dataset that represents the whole without favoring any category or group. The risk to it, though, is that the smaller categories might disappear because if a category has very few products, there's a chance it won't appear at all in the sample. However, the risk is still worth it, as random sampling becomes quick and easy to implement, which saves time compared to more complex methods like stratified sampling. It works well when you need to reduce a huge dataset fast without overthinking every single detail.

Chosen Method – Stratified Sampling: The method we chose was Stratified Sampling: With that method we managed to reduce the data to 5,000 rows while keeping the overall structure of the original dataset. Each major product category was included in the sample in roughly the same proportion as in the full dataset. We started by organizing the dataset according to the product_category. From there, we took a random selection of entries from each group. By doing this, we made sure that even the less common categories had some representation in the final dataset, while the bigger categories still held enough data to work effectively. With that the dataset balanced and avoided letting just a few popular categories take over the whole analysis.



product_ti	product_ra	total_revie	purchased	discounte	original_pi	is_best_se	is_sponso	has_coupon	buy_box_av	sustainabilit	product_categ	discount_percentage	Random
HP 63XL Bl	4.7	52857	10000	62.89	62.89	No Badge	Organic	No Coupon		Manufacturin	Laptops	0	0.998834467
HP Origina	4.6	6931	20000	106.89	106.89	No Badge	Organic	No Coupon	Add to cart		Laptops	0	0.99664558
HP 64 Blac	4.6	33075	10000	25.89	25.89	No Badge	Organic	No Coupon		Manufacturin	Laptops	0	0.990617369
Energizer M	4.8	172	1000	16.19	16.19	No Badge	Sponsored	No Coupon	Add to cart		Power & Batter	0	0.987158854
Scotch Ma	4.8	21475	10000	12.41	24.49	No Badge	Organic	No Coupon	Add to cart		Other Electroni	49.33	0.986615889
Apple 2025	4.8	815	4000	919.08	919.08	No Badge	Organic	No Coupon		Energy efficie	Laptops	0	0.985487181
Amazon Ba	4.6	88282	10000	6.75	6.75	No Badge	Organic	No Coupon	Add to cart		Other Electroni	0	0.978889065
CyberPowe	4.5	14563	6000	239.95	239.95	No Badge	Organic	No Coupon			Power & Batter	0	0.976223323
Duracell C	4.8	88458	20000	16.99	16.99	No Badge	Organic	No Coupon	Add to cart		Power & Batter	0	0.974917961
Logitech M	4.4	40618	10000	22.66	22.66	No Badge	Organic	No Coupon			Laptops	0	0.974298593
KODAK 10	4.1	264		85.99	85.99	No Badge	Sponsored	Save $16.00	Add to cart		TV & Display	0	0.972185923

Figure 3: Random Sampling

We decided to go with the stratified sampling because it keeps the variety of Amazon's product range intact, which in return helps with the dataset making it more balanced and representative. For instance, if we decided to go with filtering only by category size or sales volume, the final dataset might have been heavily skewed toward a few dominant categories or only top-selling items. The use of stratification ensured that the mix of categories stayed similar to the original dataset. The original data had 5% of items in Books and 10% in Electronics, so in our smaller sample, we pretty much kept it the same way to be as thorough as possible. This makes it easier to run SQL queries and still see trends for each category, letting us compare them in a meaningful way. Filtering by category alone would have excluded potentially valuable data from niche markets. Random sampling was also considered, but it had the risk of leaving out smaller categories entirely. Stratified sampling solved this by combining the randomness of sampling with guaranteed coverage of every product group. The result was a 5,000-row dataset that is far easier to handle while still offering a wide variety of products, categories, and sales data that reflect Amazon's overall marketplace.

4. Data Cleaning Steps

We performed data cleaning to be sure that the dataset was accurate, consistent, and ready for analysis. The data cleaning multi step workflow:

1. **Remove Irrelevant or Personal Data Columns:** First, we had to drop the columns identified as irrelevant (image URL, page URL, etc.). The process was done in Excel by simply deleting those columns from the sheet. Removing these fields early on helped us to focus on more important variables and save any private information. After this step, we worked strictly with the 13 retained columns containing product and sales information.

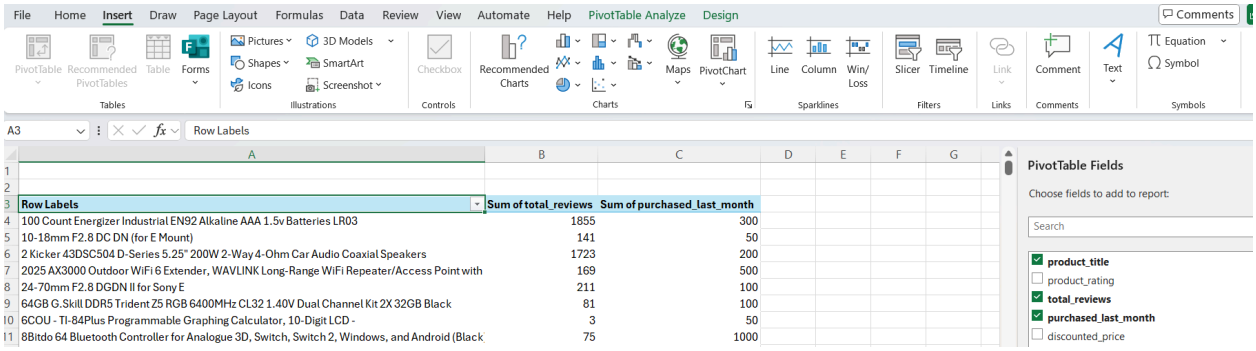
2. **Handle Missing Values:** We checked the dataset for null entries. Generally, the missing data was less than 2% of records and we still addressed it to avoid skewing results. For numerical fields like product_rating or discounted_price that had an occasionally missing value, we imputed reasonable estimates. Our strategy was to use category-level averages for filling missing numeric values. So, if purchased_last_month was missing for a particular product we tried to fill it with the average number of purchases last month for other products in the same category. A missing

product_rating could be replaced with the average rating of products in that category. This approach helped us to keep the context of the product type – if products in similar categories might have similar performance and prevents a missing value from completely removing a data point from analysis.

3. **Standardize Data Formats and Types:** We took care that all columns have a consistent and appropriate data type. For example, prices which were read as text due to a dollar sign or comma were converted to numeric values representing USD. We removed currency symbols or $ format or commas to store prices as numbers. Ratings were on a 0 to 5 scale and with some stored as text like "4.5 out of 5" and were formatted as numeric decimal values. We also standardized text fields and category names, for instance, product_category values were consistently capitalized or formatted, and trimmed any extra spaces. Standardizing units and formats are an important part of cleaning because it eliminates inconsistencies that could cause errors in analysis or when loading into a SQL database. After this step, all numeric columns (total_reviews, purchased_last_month, all price fields, ratings, discount_percentage) were truly numeric and free of anomalies like text characters, which made them ready for calculations and comparisons.

4. **Remove Duplicates and Invalid Entries:** We checked for duplicate rows and any obviously invalid data points. Using Excel's remove duplicates feature on a combination of product_title and other identifiers, we found only a handful of duplicate lines, which we dropped to avoid double-counting. We also looked for logically impossible values, for example, negative numbers in purchased_last_month or a discount_percentage over 100%. None of the products had ratings above 5 or below 0 (which is good, indicating data consistency). We encountered a few cases where discounted_price was higher than original_price which would imply a negative discount; those were treated as data errors. For such cases, it was small, we assumed maybe the prices were swapped and corrected them or if the data was too inconsistent, we dropped that record to avoid distorting the analysis. So, we removed a small number of records under 1% of the samples that were too problematic.

5. **Verify Category Labels and Value Ranges:** Lastly, we looked after checks on categorical and numeric fields. We reviewed the product_category entries to be sure that they were valid and consistent, like, no obvious typos or mix-ups like "Electronics" vs "Electronic" – if found, we would standardize them to one. We also verified that the price related fields made sense, discount_percentage was correctly calculated as original_price - discounted_price / original_price * 100% and fell in the range of 0%–100% for all records after our corrections. These checks are going to help to know that the cleaned data is consistent and credible for analysis.



Row Labels	Sum of total_reviews	Sum of purchased_last_month
100 Count Energizer Industrial EN92 Alkaline AAA 1.5v Batteries LR03	1855	300
10-18mm F2.8 DC DN (for E Mount)	141	50
2 Kicker 43DSC504 D-Series 5.25" 200W 2-Way 4-Ohm Car Audio Coaxial Speakers	1723	200
2025 AX3000 Outdoor WiFi 6 Extender, WAVLINK Long-Range WiFi Repeater/Access Point with	169	500
24-70mm F2.8 DGDN II for Sony E	211	100
64GB G.Skill DDR5 Trident Z5 RGB 6400MHz CL32 1.40V Dual Channel Kit 2X 32GB Black	81	100
6COU - TI-84Plus Programmable Graphing Calculator, 10-Digit LCD -	3	50
8Bitdo 64 Bluetooth Controller for Analogue 3D, Switch, Switch 2, Windows, and Android (Black	75	1000

Figure 4: Pivot Table

After these cleaning steps the feedback was a clean dataset of 5,000 rows and 13 columns. No missing or duplicate records. All data types are consistent and now the dataset is ready to be uploaded in a MySQL database or be analyzed with analytical tools.

To make analysis easier in steps, our team decided to create binned categories for some numeric fields. This was not explicitly required for cleaning, but we planned it as the next step to simplify certain queries and comparisons:

- **Price Ranges:** We created price groups by labeling each product as Low, Medium, or High priced. Products under $20 were tagged at a low price, 20 to $50 as medium and above $50 as high price. This categorization will help in analyzing trends and see if high-priced items have better average ratings. Do low-priced items get more reviews?
- **Rating Categories:** We segmented product ratings into broad categories poor with a rating between 0 and 2.5, average 2.6 to 4.0, and excellent 4.1 to 5.0. This allows us to group products by customer satisfaction rather than deal with a spectrum of decimal values and we can quickly count how many product frequencies fall into each quality.
- **Discount Levels:** We also bucketed the discount percentages into a few groups: none 0% discount, moderate 1–20% discount and high 21% or more discount. This'll help in assessing how common large discounts are and whether high-discount items sell more or have different attributes compared to non-discounted products.

By grouping these variables into categories, we simplify the comparative analysis across different product segments. These will make it easier to perform aggregations. Instead of saying that product A has a 37% discount and product B has a 15% discount we can say that product A falls in the group of high discount categories while product B is in the moderate group of discount categories.

5. Business Benefits, Limitations, and Trade-offs

Performing the above data reduction and cleaning yields several benefits, but it also involves trade-offs. It's important to reflect on how these choices impact the potential insights:

- **Benefits:** The streamlined dataset offers a marked increase in manageability and analytical performance. A smaller, cleaner dataset is easier to load into tools like Excel or a local SQL database, and queries or calculations run much faster. This efficiency is crucial in both academic and professional environments, but it's also relevant in business – cleaner data leads to quicker analysis and faster insights. By eliminating irrelevant information (e.g., image URLs, personal data) and focusing on core metrics, we reduce distraction and can concentrate on the variables that matter for business decisions.

  Additionally, ensuring data quality (no missing or inconsistent values) improves the accuracy of any analysis we do; decisions based on this data will be more reliable because they're grounded in well-vetted information. In short, the prepared dataset supports robust analysis and decision-making processes, enabling Amazon to derive insights such as recognizing high-demand product categories, evaluating discounts, or refining pricing approaches, without the interference of noisy or poorly structured data.

- **Limitations:** Reducing rows and columns comes at the cost of granularity. Removing certain columns like seller location or user information limits opportunities for certain analyses, such as geographic sales performance or customer segmentation.

  Similarly, by sampling down to 5,000 rows, we might have left out some rare products or outlier cases that could hold valuable insights. There is a risk that in making the data set smaller, some subtle patterns (perhaps present only in the full data) might be missed. For instance, a very niche category or a product with unusual sales spike might not appear in our sample, making those trends invisible. Moreover, a reduced data set might not capture seasonal or temporal variations if any (though in this case it's just one month of data). Lastly, although privacy measures (such as excluding sensitive fields) support compliance with standards like GDPR or CCPA, they also constrain the range of possible explorations.

  In essence, removing data reduces richness – we must accept that we traded off a bit of completeness for simplicity.

- **Trade-offs:** Our decisions balanced academic needs with data completeness. Since the purpose is to demonstrate data wrangling and to practice SQL/analysis techniques (rather than to perform a specific real-world business analysis on all 42k products), the trade-off is justified. We opted for a diverse but downsized dataset that still maintains integrity for analysis.

  In a real business scenario, one might keep the full dataset and use more powerful tools to handle it, rather than dropping data. However, the approach we took is a practical compromise: it prioritizes clarity, diversity, and manageability of data over having every possible record. We preserved as much representativeness as possible through stratified sampling, mitigating some concerns about bias. It's also worth noting that all cleaning operations (like dropping columns for privacy) align with good data governance and focus on the analysis objectives.

  Every change was intentional: we improved data quality and focused at the expense of some granularity. This is a common trade-off in analytics – sometimes a slightly smaller but cleaner dataset can yield better insights than a large, messy one. As consultants, we communicated these

trade-offs to Amazon's team, ensuring they understand what was removed and why, so they can interpret the forthcoming analysis with that context in mind.

The benefit of our data wrangling is a dataset that is accurate, relevant, and ready to drive insights quickly, which ultimately supports faster and more confident decision-making. The limitation is reduced scope and detail, but we mitigated that by careful sampling and by focusing on the core questions at hand. The choices made were deliberate, aiming to maximize learning and insight while minimizing noise and complexity.

6. Tools and Techniques

Tool Used for Wrangling: Our idea was to use Microsoft Excel. We decided that excel was convenient for this stage because it allowed us to quickly visual scan the data, also was easier for manual filtering, and straightforward operations like column removal and find-and-replace for cleaning values. We used Excel functions and features such as sorting (for spotting outliers or duplicates), filters (identifying missing values), and pivot tables (to verify aggregated statistics like average ratings by category). As we were already familiar with how Excel works it was easier to get an overview of the dataset and perform one-off cleaning actions. For example, using Excel we could rapidly fill down category averages or use formulas to calculate discount_percentage to double-check values.

Summary

Overall, the Amazon product sales data set taught us how business challenges can be addressed using analytical tools. Initially, we checked Kaggle's usability score, which was 10/10. Notably, it was sourced properly and there were a lot of user interactions. After our own observations, we deleted the unnecessary columns and used the random function to create averages on rows that needed imputation.

Our report was organized using Excel; however, we soon noticed for further cleaning and reading such large data, tools like SQL and Tableau would help find specific info to properly consult Amazon. By simply deleting unnecessary columns, it helped organize what information would be useful to observe. This experience showed us how data preparation tools can improve quality and streamline workflows.

One of the most significant findings was how much cleaner data leads to better decision-making. Clean data provides confident findings and results, and less of the garbage in garbage out. Excel was effective for the initial steps, but automation from other applications would make the process more seamless. With cleaner, consistent and structured data, implementing business decisions and observing patterns to optimize sales and advertising becomes clearer.

On top of that, using Tableau would also be helpful to visualize and interpret the data even for a lay person. Hypothetically, creating histograms or a timeline chart for sponsored/discounted products could help observe sales patterns. Tableau offers great tools to visualize larger datasets, which would help

present our findings and conclusions to Amazon. Providing a thorough report and having quantitative data backing up our conclusions make us a more reliable business partner.

Additionally, we found that data preparation supports collaboration in meaningful ways. Because Excel is widely used and accessible, it allows us to work together effectively even with different levels of technical skill. In real organizations, this accessibility helps teams across departments stay aligned. People from multiple backgrounds and departments can engage with data to help make crucial business decisions.

In conclusion, we simplified the amazon dataset into something more actionable by us as data consultants. We reduced the dataset from 17 columns down to 13 meaningful columns, and we sampled the rows to a manageable size while preserving the dataset. The resulting dataset is free of clutter and errors, making it ready for analysis in MySQL or other analytical tools.