Maria Cardillo
E-Commerce Dataset - Data Journal
July 15, 2025

Dataset: https://www.kaggle.com/datasets/gabrielramos87/an-online-shop-business/data

July 15, 2025

Dataset

In this project, I will be using a public dataset from Kaggle on E-Commerce sales data for one year. This dataset was collected from a confidential e-commerce business in London and features sales from European countries. This dataset caught my eye as it is based off of real-world data and contains around 500,000 rows of transaction data within one year. Before diving into the data, I believe I will want to convert this data into USD to make the findings more relevant to my own understanding as well as my audience.

One limitation of this data will be that there is only one year, meaning that seasonal trends will be determined based off of only one season rather than comparing year over year.

Business Scenario

To set the stage for this project, I will be using the following scenario:

A new e-commerce business is looking to enter the market for home goods and gift related items. This business contains a few product designers who will be making customized designs, but due to limited resources, the company wants to ensure they are dedicating their resources to popular items with high turnover for their initial launch. The dataset we will be using for this project includes gifts and home goods for all ages and countries all over the world.

Our primary business question will be: Which products should our new e-commerce business focus on designing for their initial launch?

To answer this question, we will be breaking our dataset down into further questions for analysis, including:

- What products sell the highest quantity?
- What products generate the highest revenue?
- Are there any trends in specific designs? (i.e. retro, flowers, dogs, etc.)
- Do certain products sell better during specific periods?
- How do holiday specific items affect sales?

As we know, our business has limited resources and is looking for a high turnover, so time specific data will be relevant when determining when our business will start and with which items. This is crucial for our business to see success from the start to allow them to continue to expand their product portfolio.

We may additionally want to consider how our customers segment to specific products. If we see trends in products being purchased together, we may want to ensure this information is provided to our stakeholders as it could impact our sales and lead to overall higher success for our business. If a product trends to be sold with another specific product, we may not achieve the same results in our business if we only offer one of these items as the customer may decide to purchase from a competitor where they can purchase both items in one transaction.

Data Tools

For this project, I will begin my analysis using Python in Jupyter Notebooks. Jupyter Notebooks is a valuable asset for a visual workflow and will quickly allow the testing of different hypotheses with the ability for a quick reference to previous queries. I will be integrating libraries including Pandas, Matplotlib, and NumPy. As I will be using JupyterLite for Jupyter Notebooks, I will be somewhat limited on visualizations for presentational purposes, so I intend to consider my tools for visualizations once I have a better understanding of the data.

Data Cleaning

This Kaggle dataset has already been cleaned, however it does contain a few null values in the customer column. The column view in Kaggle allows an overview to ensure there are no out-of-place values that would require further cleaning, however we still use .describe() and .info() to validate this.

Data Analysis

To start my analysis, I will import pandas, Matplotlib, and Numpy. I will additionally use pandas to create my dataframe. Next, I will explore my data by using .describe() and .info(), check for null values with .isnull().sum(), and use .valuecounts() to explore some of the most popular data in my columns.

Now, taking a look at our data, I notice that we have a “Price” column, and a “Quantity” column, but not a column for our total sale. To ensure that the Price column is only one unit, checking back on the original dataset, it clarifies “the price of each product per unit in pound sterling (£).” Time to make a new column! This would be a good time to convert our original price column to USD, too.

We’ll start with the conversion. Since I’m using JupyterLite, I need to write a manual conversion. To consider this limitation, it would be best to adjust my current “Price” column to “Price_GBP” and add a column for “Price_USD.” This ensures we could always go back and update our “Price_USD” column from the original “Price_GBP” column if the conversion rate were to alter in such a way it affected our analysis.

Now that we have our converted totals, we can start to check out a bit more about our products. We're going to import Counter for this and combine all of our words together to query for the words with the most results. My hypothesis is that this will yield insights for both products and designs that we can dive in deeper to soon.

Great! This provided us with products and designs that we can look a bit deeper into. We can see that the results here are a lot different than our .valuecounts(). There's a couple of words to ignore like "set," "of," and "and," and we can also see this yielded a lot more results for designs than products.

July 16, 2025

Today we're going to start off where we left off with the counter. Now that I have a list of the most popular words, I'm going to filter out some of the noise words that are unnecessary like "set," "of," and "and," and add a rule that the length of the word must be more than two letters. I'm also going to assign my top 20 products to a variable. With this, I can see that one of the products, paper craft little birdie, has the third most amount of profit but 0 quantity. Upon further inspection, it looks like this item had a partial refund so it looks like 0 quantity, but it still has the third highest profit.

Next, I'm going to plot a few visualizations. I started with a bar chart for most popular keywords, but I think a wordcloud is a bit more fitting for this situation. This does a great job at showing what words are frequently used, but doesn't really help me with sales. Let's fix that.

Now we have a bar chart that clearly shows us the revenue for each keyword. This is definitely helpful and we will save this visualization for presentation. I really like the use of the word cloud as well, so we'll also save this one.

The next thing I want to do is create a few visualizations to break down what ProductNames are contained in some of these keywords. I'm going to create a nested graph to show which products in each keyword sell best.

Next, I'm going to start showing trends over time to ensure that our business begins in the best time frame.

Now that I have my visuals, I'm going to do a quick overview of my business objectives to make sure I've answered the business questions from my original objective before compiling my graphs into a presentation. To recap:

> *Our primary business question will be: Which products should our new e-commerce business focus on designing for their initial launch?*

To answer this question, we will be breaking our dataset down into further questions for analysis, including:

- *What products sell the highest quantity?*
- *What products generate the highest revenue?*
- *Are there any trends in specific designs? (i.e. retro, flowers, dogs, etc.)*
- *Do certain products sell better during specific periods?*
- *How do holiday specific items affect sales?*

Right off the bat, I realized I got a bit carried away by our keywords and missed the chart for top products. Let's add that now. This graph was very important as the top selling item, popcorn holder, didn't appear in any of our previous graphs. Let's include a graph with these items displayed over time as well. I'm going to need to exclude the paper birdie for this as it skews our results.

With these last few graphs, and a quick snippet to get our total revenue, it's time to put together our report!