IRWA-2025-part-1
Github repository: https://github.com/laiaatomas/IRWA_FinalProject
Laia Tomàs U198723 NIA252294
Quim Ribas U198742 NIA251754

FINAL PROJECT: REPORT PART 1

PART 1: DATA PREPARATION

For simplicity, we ingest the data (in json format) and use the predefined pandas function to read it as a conventional dataframe.

1. Preprocessing

We create a function *build_terms* that takes as input a column and applies all the pre-processing steps specified: removing stop words, tokenization, removing punctuation marks and stemming. This function returns the pre-processed column. We apply this function to the text fields of our dataset (*title* and *description*) and save them into two new variables (*title_clean* and *description_clean*).

As a bonus, we added two extra pre-processing steps. First, we ensure that each entry of these columns is a string, if not, we replace them with an empty list.

For the second extra step, we only keep words with a character length above 2, because we believe that 3 is the minimum number of characters for a word to have a considerable meaning in the sentence. For instance, words present in the dataset such as wg or re are removed as we believe they are not significant in describing the product.

2. Output validation

We check that the required output columns are present in our pre-processed dataframe. As explained in the previous section, to preserve the original information on the text fields, the pre-processed text fields are stored in two new columns called *title_clean* and *description_clean,* so that we will be able to retrieve the original titles and descriptions when needed.

It is worth mentioning that this approach of not storing the pre-processed variables in place is consistent along the whole project.

3. Handling text fields

After careful analysis, we concluded that the best approach is to keep these fields separated in the index rather than merging them into a single text field. Each of these attributes carries distinct semantic information describing different product characteristics, and preserving this separation allows for a more accurate and interpretable retrieval process.

Maintaining these fields separately enables field-specific weighting during retrieval, meaning that certain fields can be given more importance depending on their relevance to the query. For example, matches in the brand field may be considered more significant than those in the seller field, as users often include brand names in their searches.

However, it is also important to consider the alternative approach of merging these fields into a single text field. The main advantage of this method is simplicity: it reduces indexing complexity and can make query processing faster, since all metadata are stored together. It can also be beneficial when queries are very short or ambiguous, as merging all descriptive information may increase the likelihood of a partial match. Nevertheless, this approach has notable drawbacks. Merging all fields reduces the semantic distinctiveness between them, which may lead to less accurate ranking.

In addition, following hint 1, we considered the evaluation context provided by validation_labels.csv, which contains relevance labels for two sample queries. This guided our decision to keep fields separate, as it emphasizes the need to distinguish between different product attributes for more effective retrieval. Preserving the distinct fields allows us to apply field-specific weights and query strategies aligned with the types of queries expected in the evaluation phase, ultimately improving retrieval precision.

In summary, the index is structured with each key product attribute as a separate field, so each term records both the field it belongs to and the products in which it appears. This enables flexible weighting, precise ranking, and interpretable results, while still maintaining manageable indexing complexity.

4. Handling numerical fields

For this section, we consider the numerical fields. For the *discount* variable, we check whether the format is text and then we keep only the numerical value as integer type. This pre-processing is done in the function *clean_discount*, which takes as input the text and returns the extracted discount value. We save this result into a new column in the dataframe *discount_clean*.

The *out_of_stock* field is boolean so we created a new field *out_of_stock_int* to store it as binary integers (0 for False and 1 for True).

For the last fields (*selling_price*, *actual_price* and *average_rating*) we transform them from object format to numeric format. In this case, since we are only ensuring the typos we modify them in place.

Regarding the question if they should be used for indexing, all of these numeric fields carry quantitative rather than semantic meaning. Consequently, these fields should not be indexed as textual terms. Instead, they should be stored as structured numeric or Boolean fields to support filtering, sorting, and ranking operations in the search process.

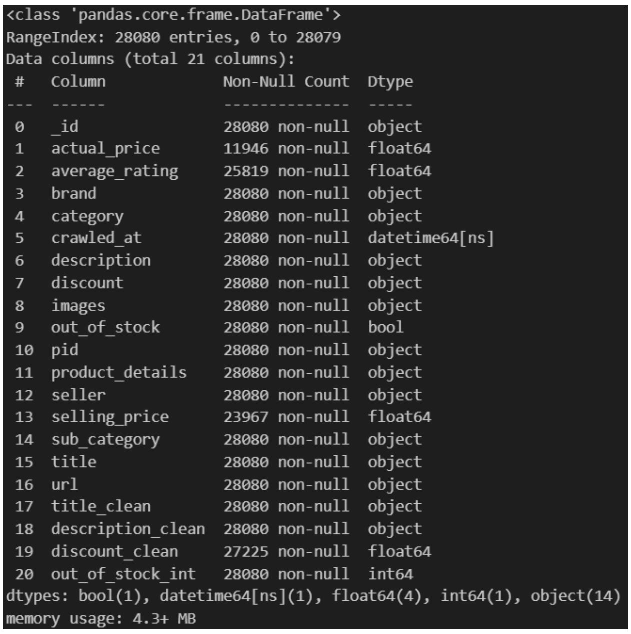
Indexing these fields as textual terms would affect the model negatively in two different ways. On one hand, it would add unnecessary noise by increasing the number of unique tokens (vocabulary size) with values that don't help ranking, which makes the index larger and slows down searches to the index. On the other hand, it would reduce efficiency of the retrieval model, as numerical values do not contribute meaningfully to textual similarity metrics (TF-IDF, for example).

PART 2: EXPLORATORY DATA ANALYSIS

In this second part of the project, we perform an exploratory analysis of our pre-processed dataframe. To avoid repetitiveness, we have not included in the report how we achieve all of these plots and statistics, meaning that we show only the results and the interpretation of them rather than the code behind it (which can be found in the notebook in GitHub).

Dataset Overview

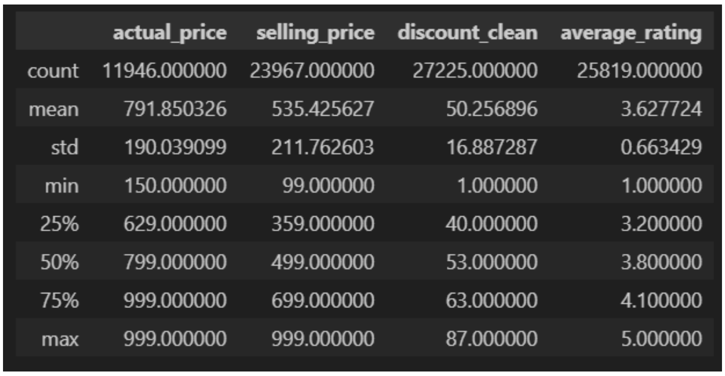
First of all, we start with an analysis on the characteristics of our dataset. We call the functions info() and describe() to get a general view of our dataset. We observe the columns with the number of non-null rows and type of each column. As we can see, the columns actual_price, average_rating, selling_price and discount_clean are type float64 as we converted them.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28080 entries, 0 to 28079
Data columns (total 21 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   _id                28080 non-null  object
 1   actual_price       11946 non-null  float64
 2   average_rating     25819 non-null  float64
 3   brand              28080 non-null  object
 4   category           28080 non-null  object
 5   crawled_at         28080 non-null  datetime64[ns]
 6   description        28080 non-null  object
 7   discount           28080 non-null  object
 8   images             28080 non-null  object
 9   out_of_stock       28080 non-null  bool
 10  pid                28080 non-null  object
 11  product_details    28080 non-null  object
 12  seller             28080 non-null  object
 13  selling_price      23967 non-null  float64
 14  sub_category       28080 non-null  object
 15  title              28080 non-null  object
 16  url                28080 non-null  object
 17  title_clean        28080 non-null  object
 18  description_clean  28080 non-null  object
 19  discount_clean     27225 non-null  float64
 20  out_of_stock_int   28080 non-null  int64
dtypes: bool(1), datetime64[ns](1), float64(4), int64(1), object(14)
memory usage: 4.3+ MB
```

Figure 1. Output of the function info().

The function *describe*() shows us some basic statistics about the numeric fields of our dataset. For example, the *average_rating* column ranges from the values 1 to 5, with an average of 3.62. We can also observe that the mean of the actual prices are higher than the mean of selling prices, and values range from 150 to 999 and 99 to 999, respectively. Regarding the discount field, the mean of discount is around 50% off.



	actual_price	selling_price	discount_clean	average_rating
count	11946.000000	23967.000000	27225.000000	25819.000000
mean	791.850326	535.425627	50.256896	3.627724
std	190.039099	211.762603	16.887287	0.663429
min	150.000000	99.000000	1.000000	1.000000
25%	629.000000	359.000000	40.000000	3.200000
50%	799.000000	499.000000	53.000000	3.800000
75%	999.000000	699.000000	63.000000	4.100000
max	999.000000	999.000000	87.000000	5.000000

Figure 2. Output of the function describe().

After checking that there are no duplicate rows in our dataset, we can perform an analysis on the missing data in our dataset. We visualize how many missing values there are per field and the proportion of values they represent in each field.

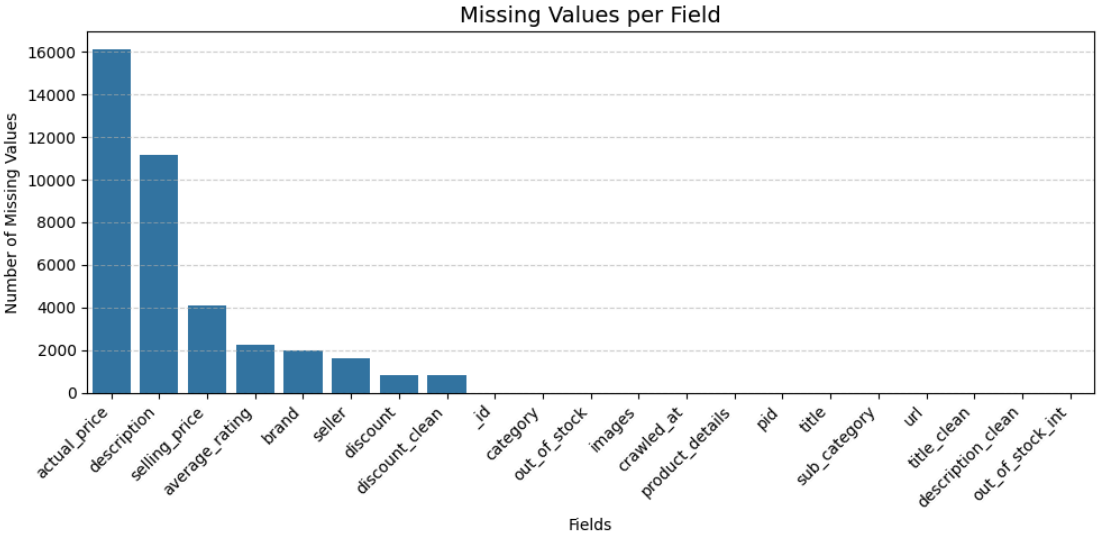


Figure 3. Histogram of missing data per field.

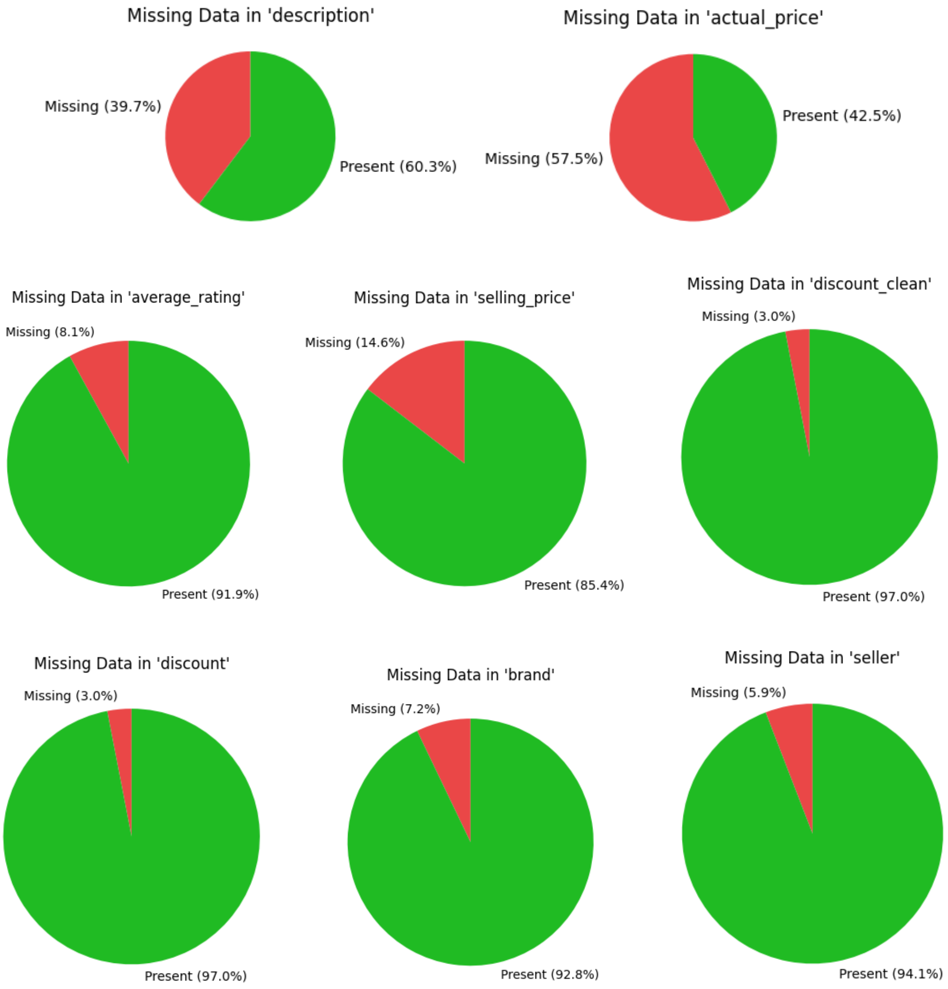


Figure 4. Pie charts of proportion of missing data.

As we can see in figure 3, the only fields with missing values are *actual_price*, *description*, *selling_price*, *average_rating, brand, seller, discount* and *discount_clean*. Therefore, we plot the percentage of missing data for these fields, in figure 4.

The two fields with the highest percentage of missing data are *actual_price*, with 57,5% of missing data and around 16000 values, followed by *description*, with 39.7%. The rest of the fields have a percentage of missing values lower than 15%.

Statistical analysis on numeric fields

We now consider the numerical fields (*selling_price, actual_price, discount_clean, average_rating*). We use three plots to visualize their characteristics. We use histograms to plot the distribution of each field, boxplot to detect any possible outliers and a correlation heatmap to analyze the relationship between all of these fields.

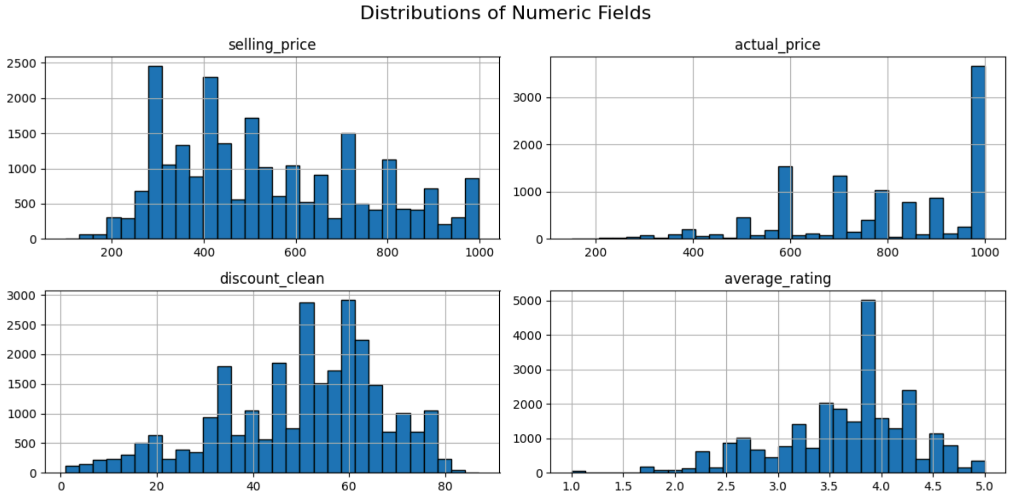


Figure 5. Histograms of numeric fields.

From figure 5 we can conclude that *selling_price* is evenly distributed across its value range, *actual_price* is right-skewed, indicating that many products have been listed before applying the discount, *discount_clean* is centered around 50-60%, meaning discounts are quite substantial, and *average_rating* values are centered around 3.5-4, meaning that most products have a great rating, and few below 3.

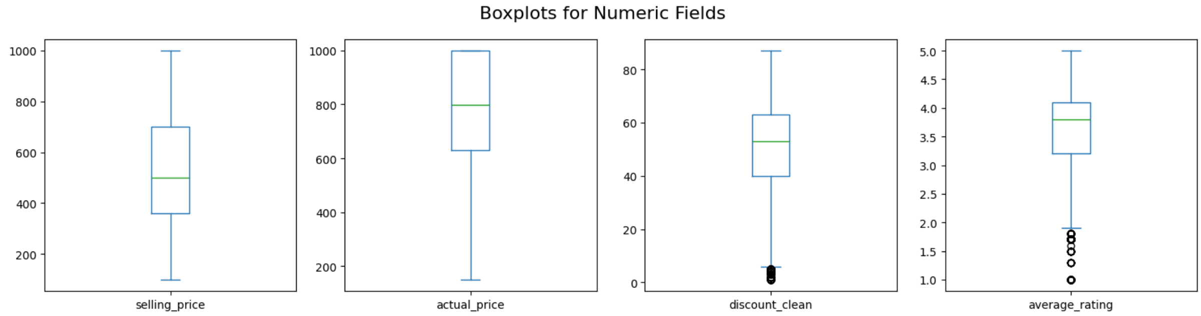


Figure 6. Boxplots of numeric fields.

From figure 6, we observe that there are a lot of outliers in the fields *discount_clean* and *average_rating*, and none on the *selling_price* and *actual_price*.

Next, we quantified the relationships between these fields by computing a heatmap. The heatmap can be visualized in figure 7.

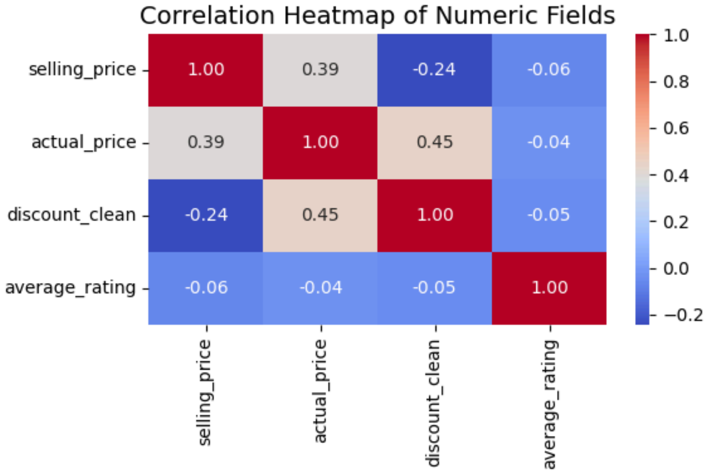


Figure 7. Heatmap of numeric fields.

Regarding figure 7, we extract these conclusions:

- Selling price vs actual price (r = 0.39): this is expected, as higher-priced products before discount generally remain higher-priced after discount.

- Actual price vs discount (r = 0.45): this suggests that products with higher original prices often have higher discounts applied, possibly to make them more competitive or appealing to buyers.

- Selling price vs discount (r = -0.24): this indicates that as discounts increase, the selling price tends to decrease. This makes intuitive sense since discounts directly reduce the final selling price.

- Average rating vs price/discount: the correlations between *average_rating* and the price-related fields are very weak (close to 0). This implies that customers' ratings are not strongly influenced by the price level or the size of the discount, they appear to depend on other factors (product quality, brand reputation, etc.)

We also performed three scatter plots for the most related variables (in absolute value).

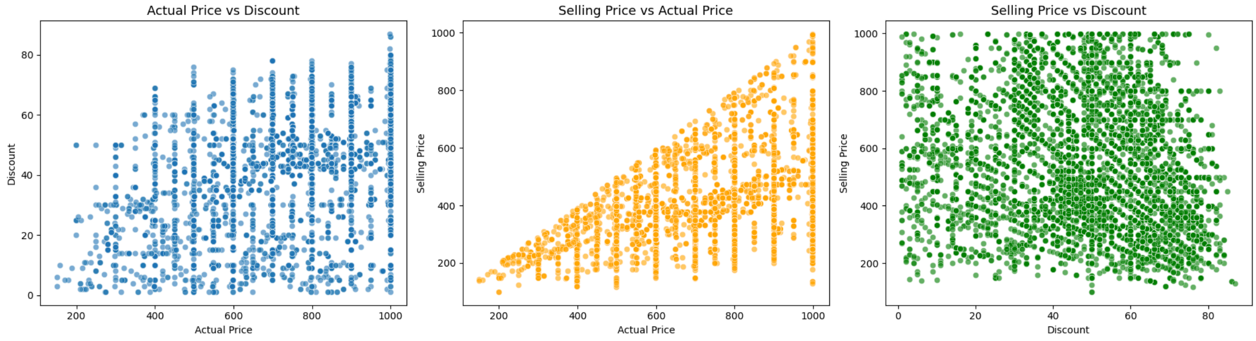


Figure 8. Scatter plot for actual price, selling price and discount.

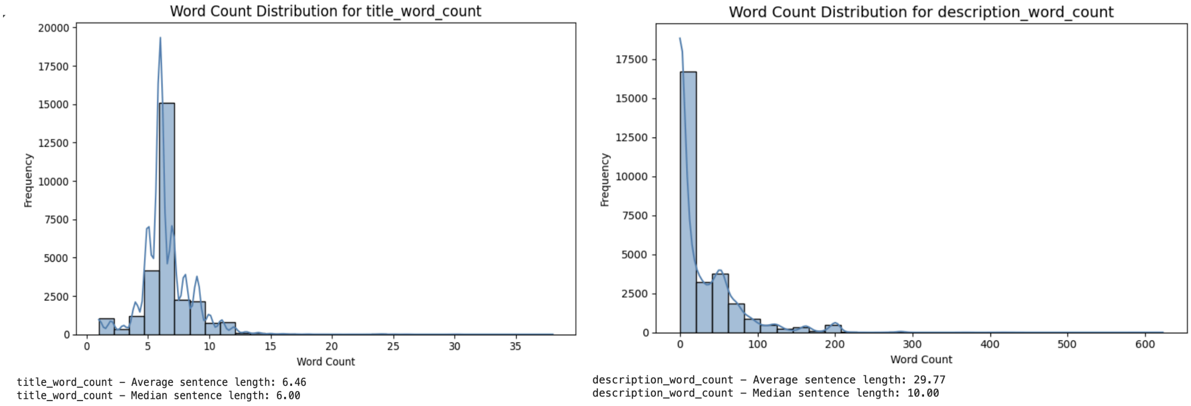
For the first plot on figure 8, comparing the discount against the actual price, we see that the points appear quite scattered. However, we can intuitively see a positive trend. This trend is not perfectly linear, but one can detect a slight upward pattern, meaning that expensive items may be discounted more aggressively.

Regarding the second plot, the selling price versus the actual price follows a clear positive linear relationship, with points only on the lower diagonal. This trend is expected, since the selling price can't be above the actual price, it can only be reduced (by applying a discount).

For the third plot on figure 8, comparing the selling price against the discount, it intuitively follows a negative trend, as higher discount being applied implies that the selling price is reduced.

Text Fields Analysis

For this section, we analyzed the text variables (title and description). First we examine their text length characteristics. We visualize the distribution of their text length by plotting two histograms.



```
title_word_count - Average sentence length: 6.46
title_word_count - Median sentence length: 6.00
```

```
description_word_count - Average sentence length: 29.77
description_word_count - Median sentence length: 10.00
```

Figure 9. Histograms of word count for title and description.

We also compute the average and median of this sentence length for both fields. The results make sense since as one would expect that the title of a product includes less words than the description of it. The results appear in figure 9.

We performed two Word Clouds for the text variables to visualize the most common terms.

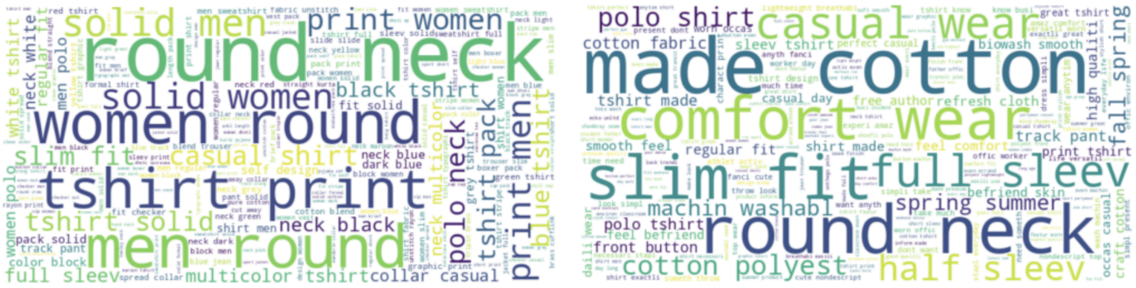
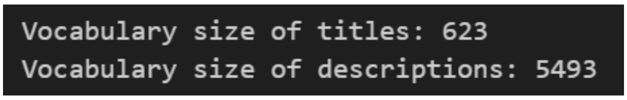


Figure 10. Word clouds for title and description.

The left-side plot on figure 10 is associated with the *title_clean* field and the right-side one is associated with the *description_clean* field.

The vocabulary size found for the fields title and description was 623 and 5493, respectively. To retrieve these unique words we used the cleaned title and description columns, which keep only the preprocessed words, so we need to consider this vocabulary could be an underestimation. The vocabulary sizes of these two fields are shown in figure 11.



```
Vocabulary size of titles: 623
Vocabulary size of descriptions: 5493
```

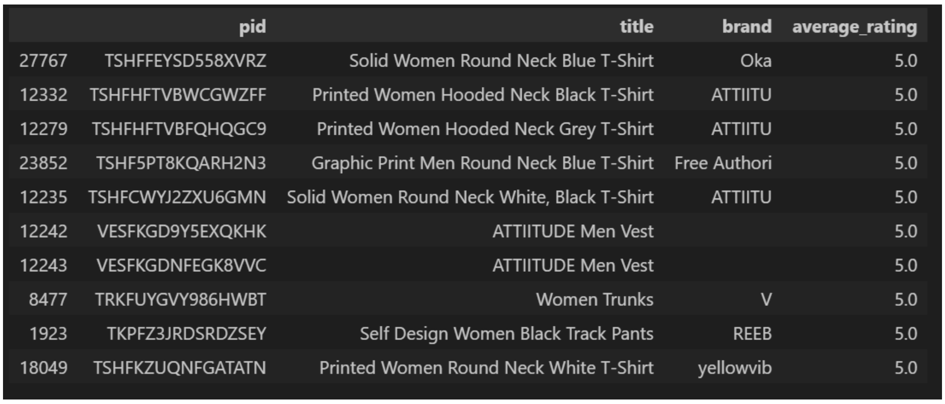
Figure 11. Vocabulary sizes for title and description.

Ranking Products Analysis

We analyzed the ranking of products according to rating, price and discount to identify trends in our dataset. The goal of this examination is to understand which product attributes might influence ranking and retrieval quality.

For the numeric fields, we displayed the top 10 rated products by ordering from highest to lowest each field. The following screenshots show the results.

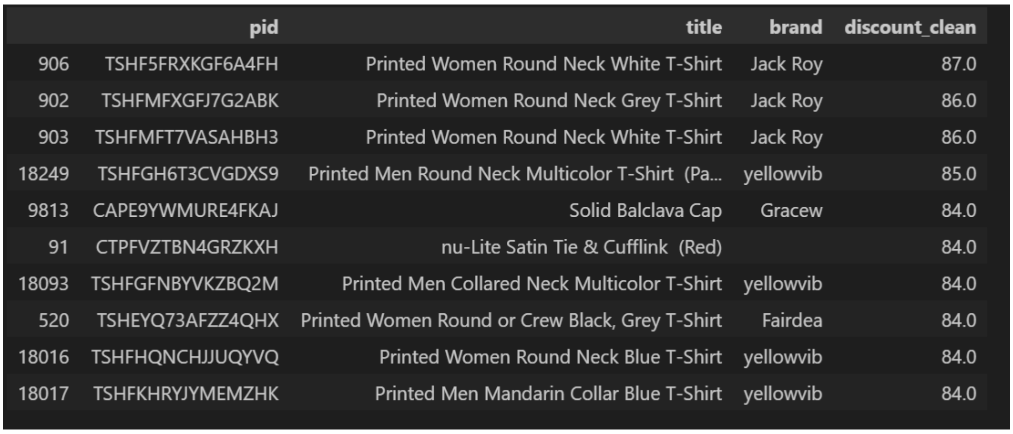
Figure 12 shows the top products ordered by average rating.



	pid	title	brand	average_rating
27767	TSHFFEYSD558XVRZ	Solid Women Round Neck Blue T-Shirt	Oka	5.0
12332	TSHFHFTVBWCGWZFF	Printed Women Hooded Neck Black T-Shirt	ATTIITU	5.0
12279	TSHFHFTVBFQHQGC9	Printed Women Hooded Neck Grey T-Shirt	ATTIITU	5.0
23852	TSHF5PT8KQARH2N3	Graphic Print Men Round Neck Blue T-Shirt	Free Authori	5.0
12235	TSHFCWYJ2ZXU6GMN	Solid Women Round Neck White, Black T-Shirt	ATTIITU	5.0
12242	VESFKGD9Y5EXQKHK	ATTIITUDE Men Vest		5.0
12243	VESFKGDNFEGK8VVC	ATTIITUDE Men Vest		5.0
8477	TRKFUYGVY986HWBT	Women Trunks	V	5.0
1923	TKPFZ3JRDSRDZSEY	Self Design Women Black Track Pants	REEB	5.0
18049	TSHFKZUQNFGATATN	Printed Women Round Neck White T-Shirt	yellowvib	5.0

Figure 12. Product ranking by highest average rating.

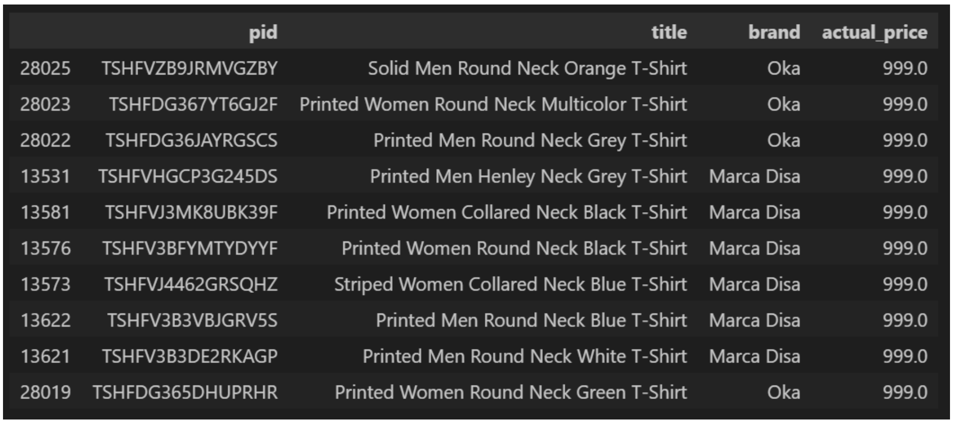
Figure 13 shows the top products ordered by discount.



	pid	title	brand	discount_clean
906	TSHF5FRXKGF6A4FH	Printed Women Round Neck White T-Shirt	Jack Roy	87.0
902	TSHFMFXGFJ7G2ABK	Printed Women Round Neck Grey T-Shirt	Jack Roy	86.0
903	TSHFMFT7VASAHBH3	Printed Women Round Neck White T-Shirt	Jack Roy	86.0
18249	TSHFGH6T3CVGDXS9	Printed Men Round Neck Multicolor T-Shirt (Pa...	yellowvib	85.0
9813	CAPE9YWMURE4FKAJ	Solid Balclava Cap	Gracew	84.0
91	CTPFVZTBN4GRZKXH	nu-Lite Satin Tie & Cufflink (Red)		84.0
18093	TSHFGFNBYVKZBQ2M	Printed Men Collared Neck Multicolor T-Shirt	yellowvib	84.0
520	TSHEYQ73AFZZ4QHX	Printed Women Round or Crew Black, Grey T-Shirt	Fairdea	84.0
18016	TSHFHQNCHJJUQYVQ	Printed Women Round Neck Blue T-Shirt	yellowvib	84.0
18017	TSHFKHRYJYMEMZHK	Printed Men Mandarin Collar Blue T-Shirt	yellowvib	84.0

Figure 13. Product ranking by highest discount.

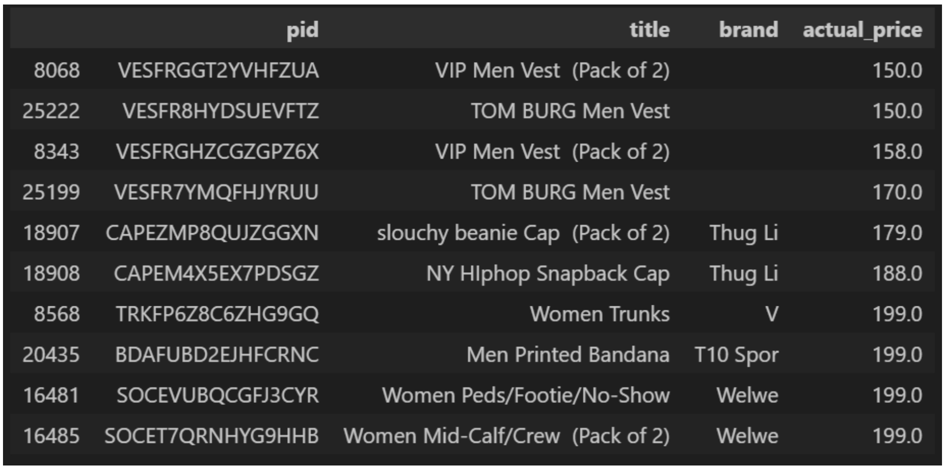
Figure 14 shows the top 10 products ordered by actual price.



	pid	title	brand	actual_price
28025	TSHFVZB9JRMVGZBY	Solid Men Round Neck Orange T-Shirt	Oka	999.0
28023	TSHFDG367YT6GJ2F	Printed Women Round Neck Multicolor T-Shirt	Oka	999.0
28022	TSHFDG36JAYRGSCS	Printed Men Round Neck Grey T-Shirt	Oka	999.0
13531	TSHFVHGCP3G245DS	Printed Men Henley Neck Grey T-Shirt	Marca Disa	999.0
13581	TSHFVJ3MK8UBK39F	Printed Women Collared Neck Black T-Shirt	Marca Disa	999.0
13576	TSHFV3BFYMTYDYYF	Printed Women Round Neck Black T-Shirt	Marca Disa	999.0
13573	TSHFVJ4462GRSQHZ	Striped Women Collared Neck Blue T-Shirt	Marca Disa	999.0
13622	TSHFV3B3VBJGRV5S	Printed Men Round Neck Blue T-Shirt	Marca Disa	999.0
13621	TSHFV3B3DE2RKAGP	Printed Men Round Neck White T-Shirt	Marca Disa	999.0
28019	TSHFDG365DHUPRHR	Printed Women Round Neck Green T-Shirt	Oka	999.0

Figure 14. Product ranking by highest actual price.

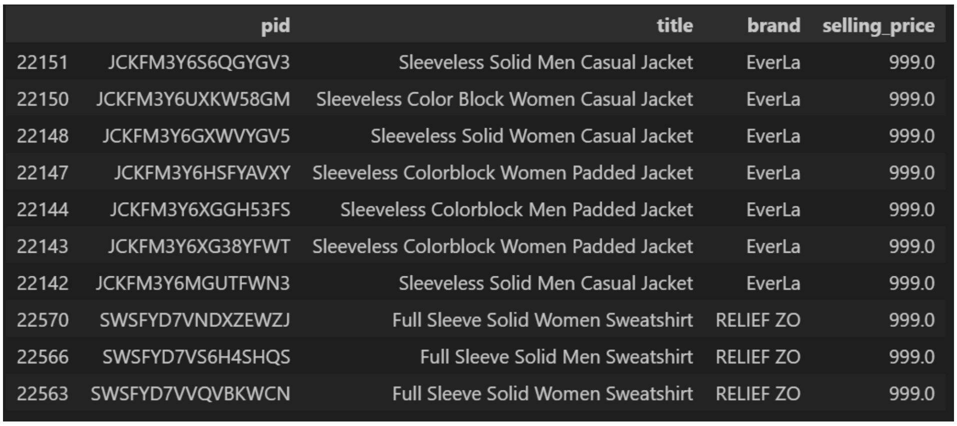
Figure 15 shows the bottom 10 products according to their actual price (the lowest valued products).



	pid	title	brand	actual_price
8068	VESFRGGT2YVHFZUA	VIP Men Vest (Pack of 2)		150.0
25222	VESFR8HYDSUEVFTZ	TOM BURG Men Vest		150.0
8343	VESFRGHZCGZGPZ6X	VIP Men Vest (Pack of 2)		158.0
25199	VESFR7YMQFHJYRUU	TOM BURG Men Vest		170.0
18907	CAPEZMP8QUJZGGXN	slouchy beanie Cap (Pack of 2)	Thug Li	179.0
18908	CAPEM4X5EX7PDSGZ	NY HIphop Snapback Cap	Thug Li	188.0
8568	TRKFP6Z8C6ZHG9GQ	Women Trunks	V	199.0
20435	BDAFUBD2EJHFCRNC	Men Printed Bandana	T10 Spor	199.0
16481	SOCEVUBQCGFJ3CYR	Women Peds/Footie/No-Show	Welwe	199.0
16485	SOCET7QRNHYG9HHB	Women Mid-Calf/Crew (Pack of 2)	Welwe	199.0

Figure 15. Product ranking by lowest actual price.

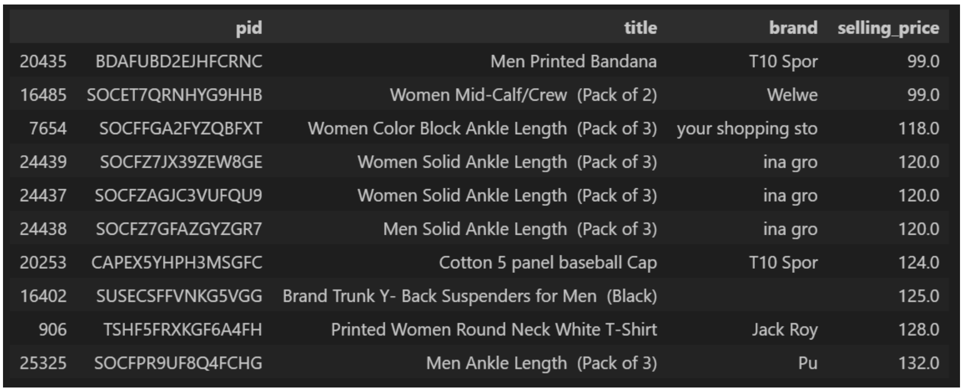
Figure 16 shows the top 10 products ordered by the selling price.



	pid	title	brand	selling_price
22151	JCKFM3Y6S6QGYGV3	Sleeveless Solid Men Casual Jacket	EverLa	999.0
22150	JCKFM3Y6UXKW58GM	Sleeveless Color Block Women Casual Jacket	EverLa	999.0
22148	JCKFM3Y6GXWVYGV5	Sleeveless Solid Women Casual Jacket	EverLa	999.0
22147	JCKFM3Y6HSFYAVXY	Sleeveless Colorblock Women Padded Jacket	EverLa	999.0
22144	JCKFM3Y6XGGH53FS	Sleeveless Colorblock Men Padded Jacket	EverLa	999.0
22143	JCKFM3Y6XG38YFWT	Sleeveless Colorblock Women Padded Jacket	EverLa	999.0
22142	JCKFM3Y6MGUTFWN3	Sleeveless Solid Men Casual Jacket	EverLa	999.0
22570	SWSFYD7VNDXZEWZJ	Full Sleeve Solid Women Sweatshirt	RELIEF ZO	999.0
22566	SWSFYD7VS6H4SHQS	Full Sleeve Solid Men Sweatshirt	RELIEF ZO	999.0
22563	SWSFYD7VVQVBKWCN	Full Sleeve Solid Women Sweatshirt	RELIEF ZO	999.0

Figure 16. Product ranking by highest selling price.

Figure 17 shows the bottom 10 products according to their selling price.



	pid	title	brand	selling_price
20435	BDAFUBD2EJHFCRNC	Men Printed Bandana	T10 Spor	99.0
16485	SOCET7QRNHYG9HHB	Women Mid-Calf/Crew (Pack of 2)	Welwe	99.0
7654	SOCFFGA2FYZQBFXT	Women Color Block Ankle Length (Pack of 3)	your shopping sto	118.0
24439	SOCFZ7JX39ZEW8GE	Women Solid Ankle Length (Pack of 3)	ina gro	120.0
24437	SOCFZAGJC3VUFQU9	Women Solid Ankle Length (Pack of 3)	ina gro	120.0
24438	SOCFZ7GFAZGYZGR7	Men Solid Ankle Length (Pack of 3)	ina gro	120.0
20253	CAPEX5YHPH3MSGFC	Cotton 5 panel baseball Cap	T10 Spor	124.0
16402	SUSECSFFVNKG5VGG	Brand Trunk Y- Back Suspenders for Men (Black)		125.0
906	TSHF5FRXKGF6A4FH	Printed Women Round Neck White T-Shirt	Jack Roy	128.0
25325	SOCFPR9UF8Q4FCHG	Men Ankle Length (Pack of 3)	Pu	132.0

Figure 17. Product ranking by lowest selling price.

As we can observe from figures 14 to 17, there is a mismatch between actual and selling price rankings. This reflects the effect of discount strategies and price elasticity in the dataset.

We see how premium products often appear in the top 10 when ranked by actual price, but may drop in the ranking based on selling price if they have significant discounts.

Similarly, differences in the bottom 10 products also indicate the influence of discounts on perceived affordability. Items with large markdowns may fall among the lowest selling prices despite having a higher original value, while consistently low-priced items remain at the bottom across both measures.

Therefore, the top 10 by selling price are likely premium items still selling at a high price even after discounts, while the bottom 10 show the most affordable deals or heavily discounted products.

We next examine the most frequent sellers and brands.

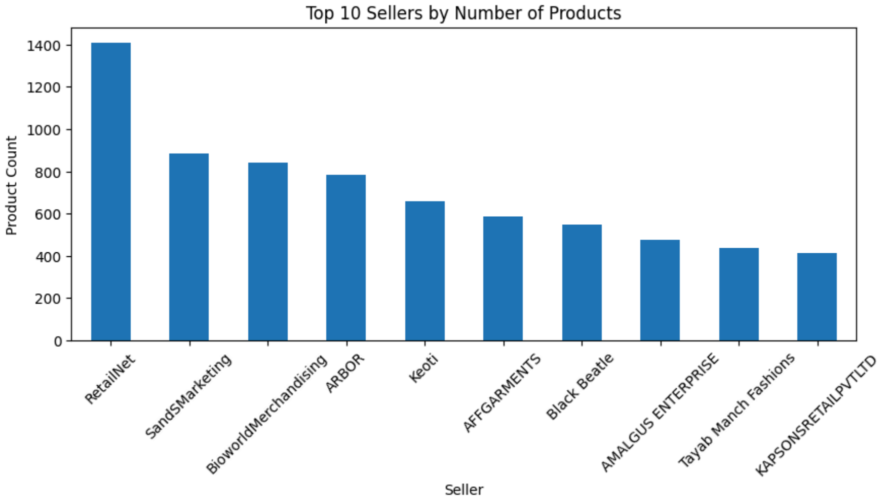


Figure 18. Top sellers by number of products.

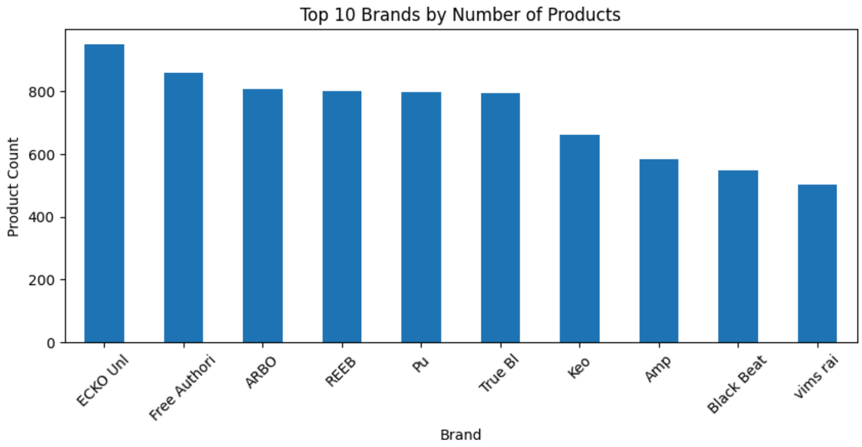
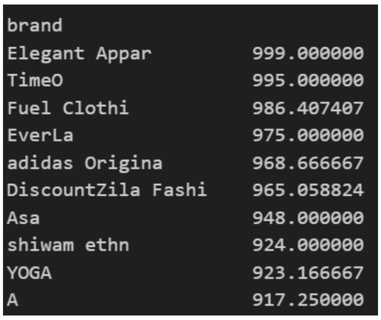


Figure 19. Top brands by number of products.

Figures 18 and 19 highlight the dataset's dominance by a few key vendors, which may bias retrieval results toward these brands. To ensure a fair evaluation, we excluded products not associated with any registered brand, as their inclusion would have made "no-brand" the top seller.

Moreover, we found the top 10 brands with the highest average product selling prices (figure 21)



```
brand
Elegant Appar          999.000000
TimeO                  995.000000
Fuel Clothi            986.407407
EverLa                 975.000000
adidas Origina         968.666667
DiscountZila Fashi     965.058824
Asa                    948.000000
shiwam ethn            924.000000
YOGA                   923.166667
A                      917.250000
```

Figure 20. Top brand by highest selling price

From figures 19 and 20, we can conclude that brands producing premium items do not dominate the market in terms of product volume. None of the top brands according to their market presence seem to have a high average selling price.

We also visualized in a pie chart (figure 21) the proportion of out of stock products, compared to the available ones.

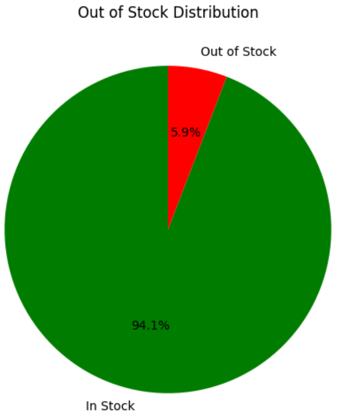


Figure 21. Pie chart of out of stock

Entity Recognition Analysis

We computed two plots to analyze the entities on the fields title and description.

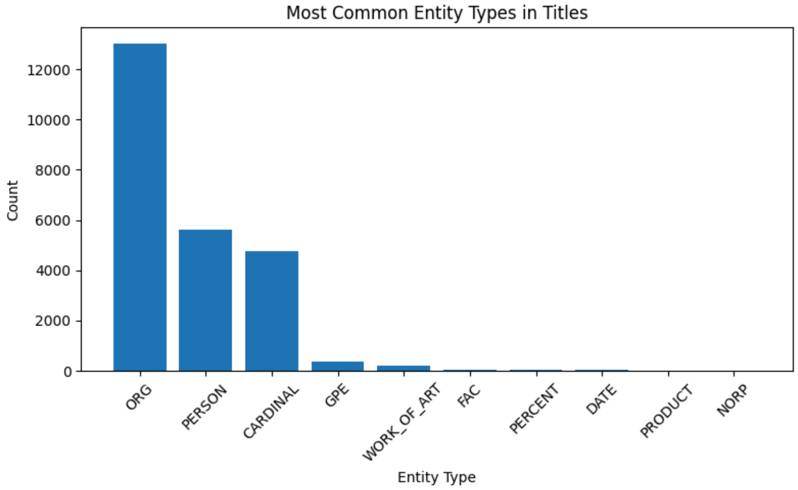


Figure 22. Entity types histogram in title.

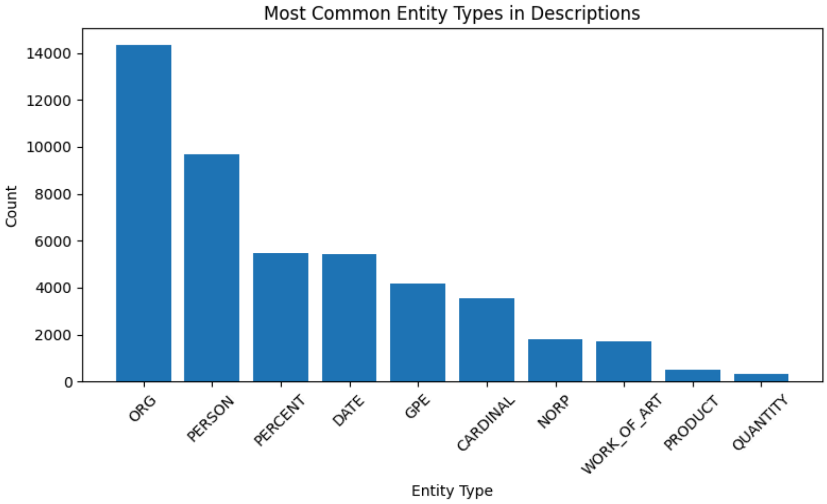


Figure 23. Entity types histogram in description

From figures 22 and 23, we can extract several conclusions. In the title field, the top entities were organizations, persons , and cardinals, indicating that product titles frequently mention brand names, designer names, or numerical identifiers such as model numbers or sizes. In the description field, the most frequent entities were also organizations and persons, but percents rather than cardinals, reflecting that descriptions often include brands, designer collaborations, and percentage information, such as material compositions (e.g., “100% cotton”) or discounts.

From all of this exploratory data analysis we have gathered a lot of valuable insights that could later enhance the retrieval process and improve search relevance in our project.