Exploratory Data Analysis

**Detected Key Columns**

|  |  |
| --- | --- |
| **Role** | **Column** |
| Category | category |
| Price | price\_pkr |
| Gender/Target | gender\_target |
| Retailer | retailer |
| Size (explicit) | retailer |

**Quick Scan (complete)**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Rows | 6,430 |
| Columns | 13 |
| Duplicate rows | 0 |
| Columns with any missing values | 9 |
| Numeric columns | 2 |
| Non-numeric columns | 11 |

**Columns — Missingness & Uniqueness (complete; top 30)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **column** | **dtype** | **missing\_count** | **missing\_pct** | **unique\_count** | **likely\_identifier** |
| size\_value | float64 | 3,917 | 60.92 | 119 | 0 |
| size\_unit | object | 3,917 | 60.92 | 3 | 0 |
| size\_raw | object | 2,002 | 31.14 | 329 | 0 |
| product\_url | object | 689 | 10.72 | 5,741 | 0 |
| observed\_at | object | 689 | 10.72 | 2 | 0 |
| price\_pkr | float64 | 324 | 5.04 | 1,013 | 0 |
| title | object | 2 | 0.03 | 6,190 | 1 |
| brand | object | 2 | 0.03 | 2,899 | 0 |
| image\_url | object | 2 | 0.03 | 2,597 | 0 |
| category | object | 0 | 0.00 | 36 | 0 |
| retailer | object | 0 | 0.00 | 4 | 0 |
| source\_file | object | 0 | 0.00 | 4 | 0 |
| gender\_target | object | 0 | 0.00 | 3 | 0 |

**Top Retailer Values**

|  |  |  |
| --- | --- | --- |
| **retailer** | **count** | **percent** |
| daraz | 3,811 | 59.27 |
| alfatah | 1,930 | 30.02 |
| metro | 367 | 5.71 |
| imtiaz | 322 | 5.01 |

## Issues:

## Overlapping Categories

There are many names for the same categories

**Price (PKR) — Stats by Category**

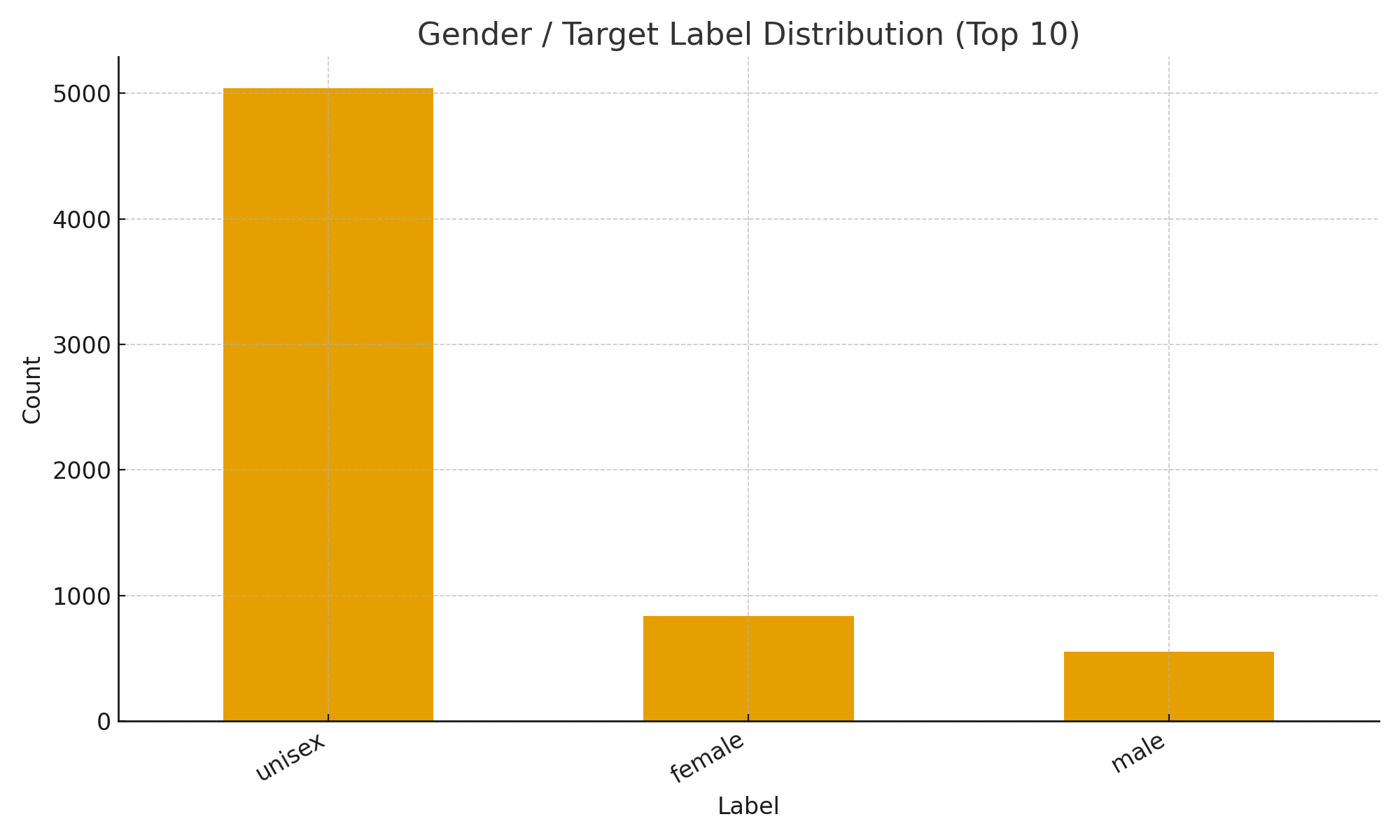
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **category** | **count** | **mean** | **median** | **p25** | **p75** | **min** | **max** | **std** |
| anti-perspirant | 315 | 506.83 | 549.00 | 219.95 | 699.00 | 100.00 | 999.00 | 260.68 |
| anti\_perspirants | 22 | 95.45 | 50.00 | 50.00 | 150.00 | 50.00 | 150.00 | 50.96 |
| body-spray | 369 | 531.68 | 499.00 | 299.00 | 799.00 | 100.00 | 999.00 | 290.96 |
| body\_spray | 41 | 405.66 | 340.00 | 150.00 | 595.00 | 15.00 | 2,000.00 | 385.93 |
| conditioner | 411 | 442.83 | 365.00 | 169.90 | 613.50 | 101.00 | 6,000.00 | 448.56 |
| cream | 469 | 344.12 | 299.00 | 139.50 | 509.00 | 0.10 | 999.00 | 265.28 |
| creams | 39 | 465.03 | 75.00 | 50.00 | 175.00 | 17.00 | 4,800.00 | 1,270.53 |
| deodorant | 403 | 366.35 | 395.00 | 0.74 | 650.00 | 0.10 | 999.00 | 323.67 |
| deodrant\_stick\_rollon | 142 | 0.54 | 0.68 | 0.13 | 0.84 | 0.10 | 0.99 | 0.31 |
| face wash | 479 | 0.43 | 0.40 | 0.18 | 0.65 | 0.10 | 0.99 | 0.25 |
| facewash | 376 | 458.06 | 445.00 | 299.00 | 589.25 | 100.00 | 999.00 | 211.69 |
| facewashes | 33 | 377.12 | 100.00 | 50.00 | 100.00 | 40.00 | 3,975.00 | 957.50 |
| handwash | 388 | 341.52 | 298.50 | 199.00 | 434.00 | 100.00 | 999.00 | 198.20 |
| lotion | 906 | 104.45 | 0.50 | 0.20 | 103.68 | 0.10 | 955.00 | 211.31 |
| razor | 455 | 307.93 | 238.00 | 149.00 | 399.00 | 0.11 | 999.00 | 256.55 |
| shampoo | 483 | 277.71 | 160.00 | 0.70 | 450.00 | 0.10 | 999.00 | 290.12 |
| shampoos | 168 | 489.20 | 360.00 | 180.00 | 400.00 | 10.00 | 6,000.00 | 840.80 |
| soap | 607 | 253.00 | 180.00 | 0.40 | 399.00 | 0.10 | 999.00 | 265.14 |

**Top Category Values**

|  |  |  |
| --- | --- | --- |
| **category** | **count** | **percent** |
| lotion | 919 | 14.29 |
| soap | 607 | 9.44 |
| shampoo | 483 | 7.51 |
| face wash | 479 | 7.45 |
| cream | 469 | 7.29 |
| razor | 455 | 7.08 |
| conditioner | 411 | 6.39 |
| deodorant | 403 | 6.27 |
| handwash | 388 | 6.03 |
| facewash | 376 | 5.85 |
| body-spray | 369 | 5.74 |
| anti-perspirant | 315 | 4.90 |
| shampoos | 168 | 2.61 |
| deodrant\_stick\_rollon | 142 | 2.21 |
| body\_spray | 49 | 0.76 |
| creams | 46 | 0.72 |
| facewashes | 33 | 0.51 |
| shampoo2 | 24 | 0.37 |
| anti\_perspirants | 24 | 0.37 |
| soaps6 | 24 | 0.37 |

## Uneven Gender Distribution

Unisex share appears high at approximately 78.41% of all rows.

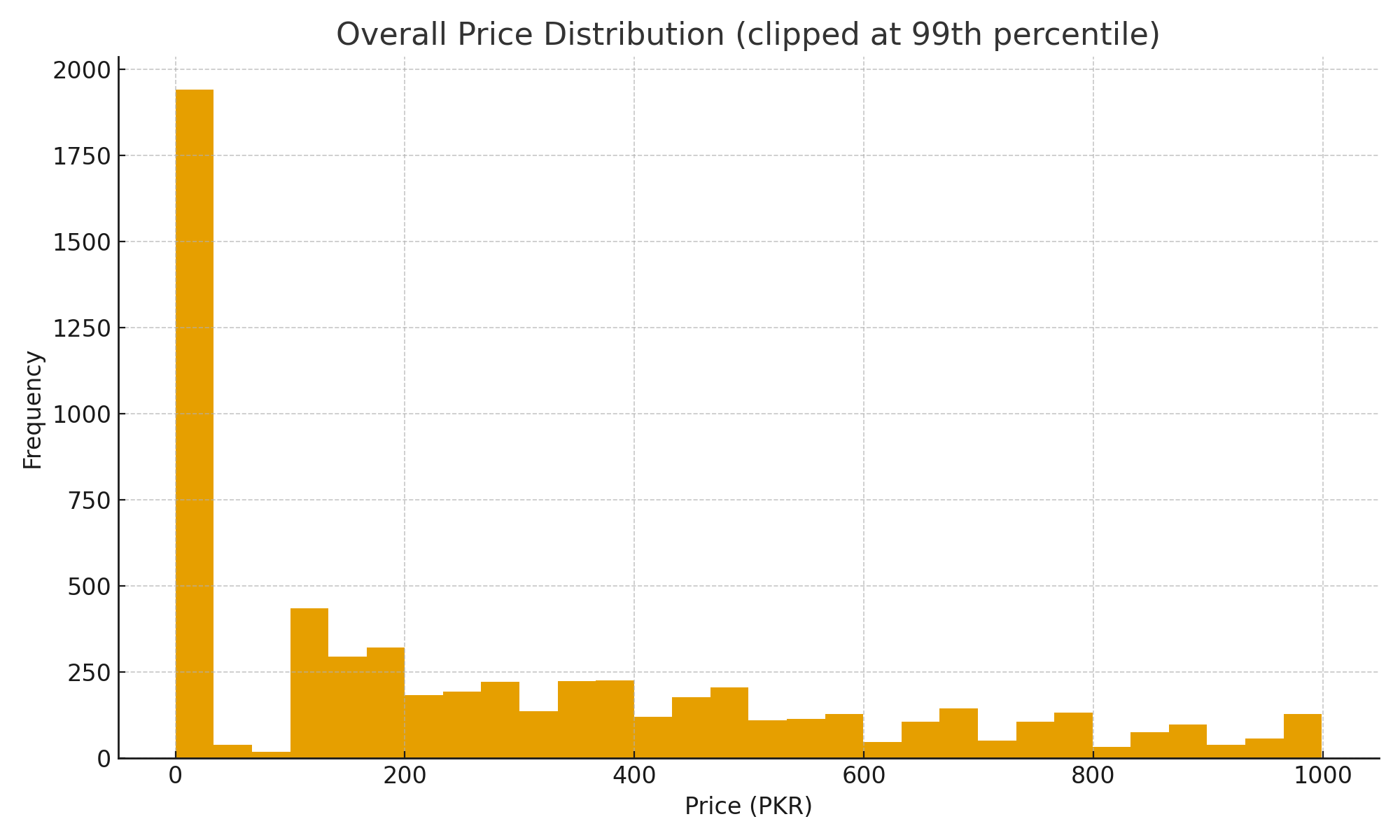


**Top Gender/Target Values**

|  |  |  |
| --- | --- | --- |
| **label** | **count** | **percent** |
| unisex | 5,042 | 78.41 |
| female | 837 | 13.02 |
| male | 551 | 8.57 |

## Overall Price Distribution

The prices are not standardized hence clustered at 0.



**Data Cleaning**

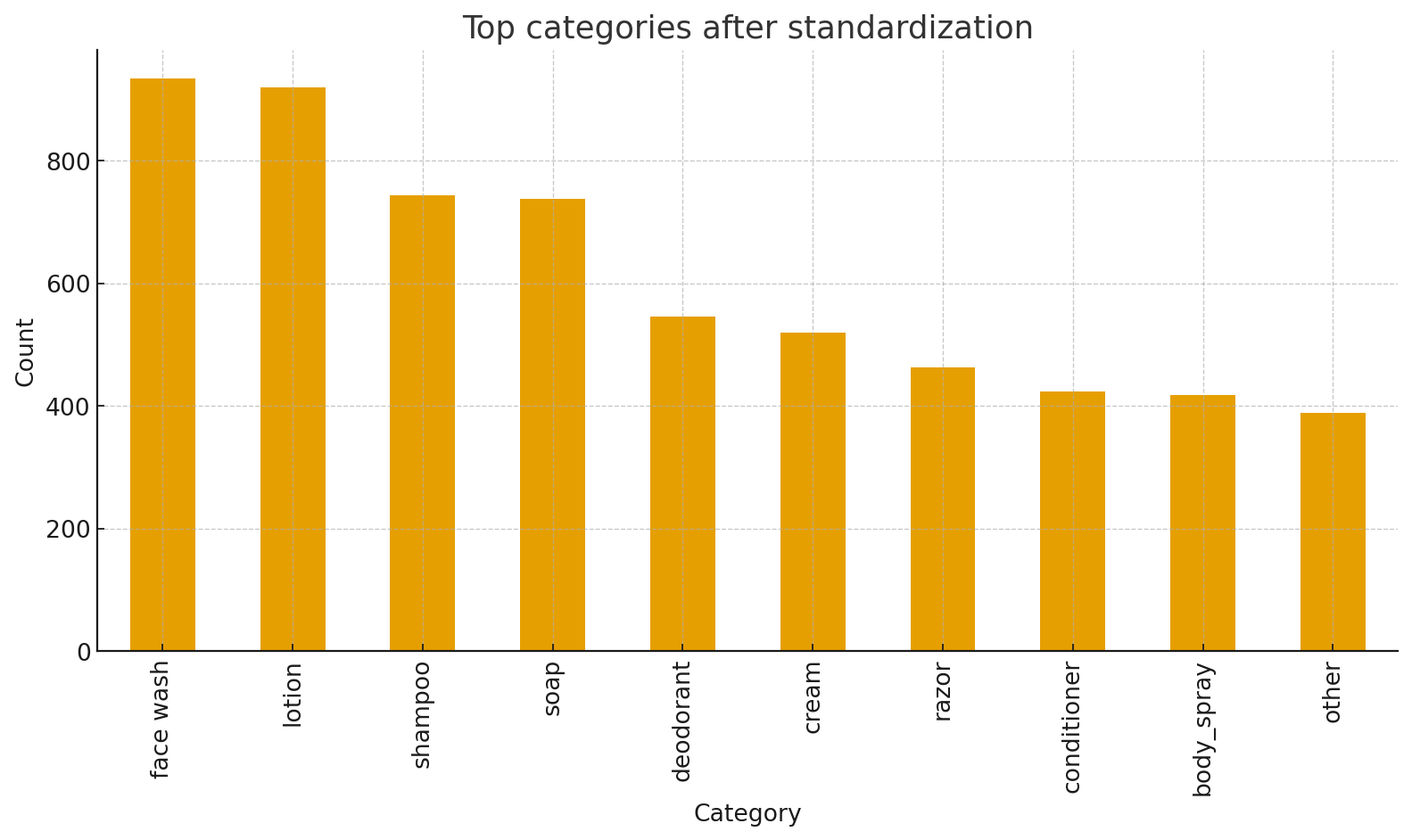
**Unique Categories:**

## BEFORE mapping (from dataset)

|  |  |
| --- | --- |
| category | count |
| lotion | 919 |
| soap | 607 |
| shampoo | 483 |
| face wash | 479 |
| cream | 469 |
| razor | 455 |
| conditioner | 411 |
| deodorant | 403 |
| handwash | 388 |
| facewash | 376 |
| body-spray | 369 |
| anti-perspirant | 315 |
| shampoos | 168 |
| deodrant\_stick\_rollon | 142 |
| body\_spray | 49 |
| creams | 46 |
| facewashes | 33 |
| facewashes3 | 24 |
| soaps1 | 24 |
| shampoo1 | 24 |
| shampoo3 | 24 |
| anti\_perspirants | 24 |
| soaps3 | 24 |
| soaps5 | 24 |
| soaps6 | 24 |
| shampoo2 | 24 |
| shampoo4 | 20 |
| soaps2 | 17 |
| soaps7 | 16 |
| conditioners | 13 |
| facewashes1 | 12 |
| facewashes2 | 8 |
| razors1 | 7 |
| creams2 | 5 |
| soaps4 | 2 |
| facewashes4 | 2 |

## AFTER mapping (canonical categories)

|  |  |
| --- | --- |
| category | count |
| soap | 1461 |
| lotion | 919 |
| fash wash | 855 |
| shampoo | 575 |
| deodorant | 545 |
| cream | 469 |
| razor | 462 |
| body\_spray | 418 |
| conditioner | 411 |
| anti\_perspirants | 315 |



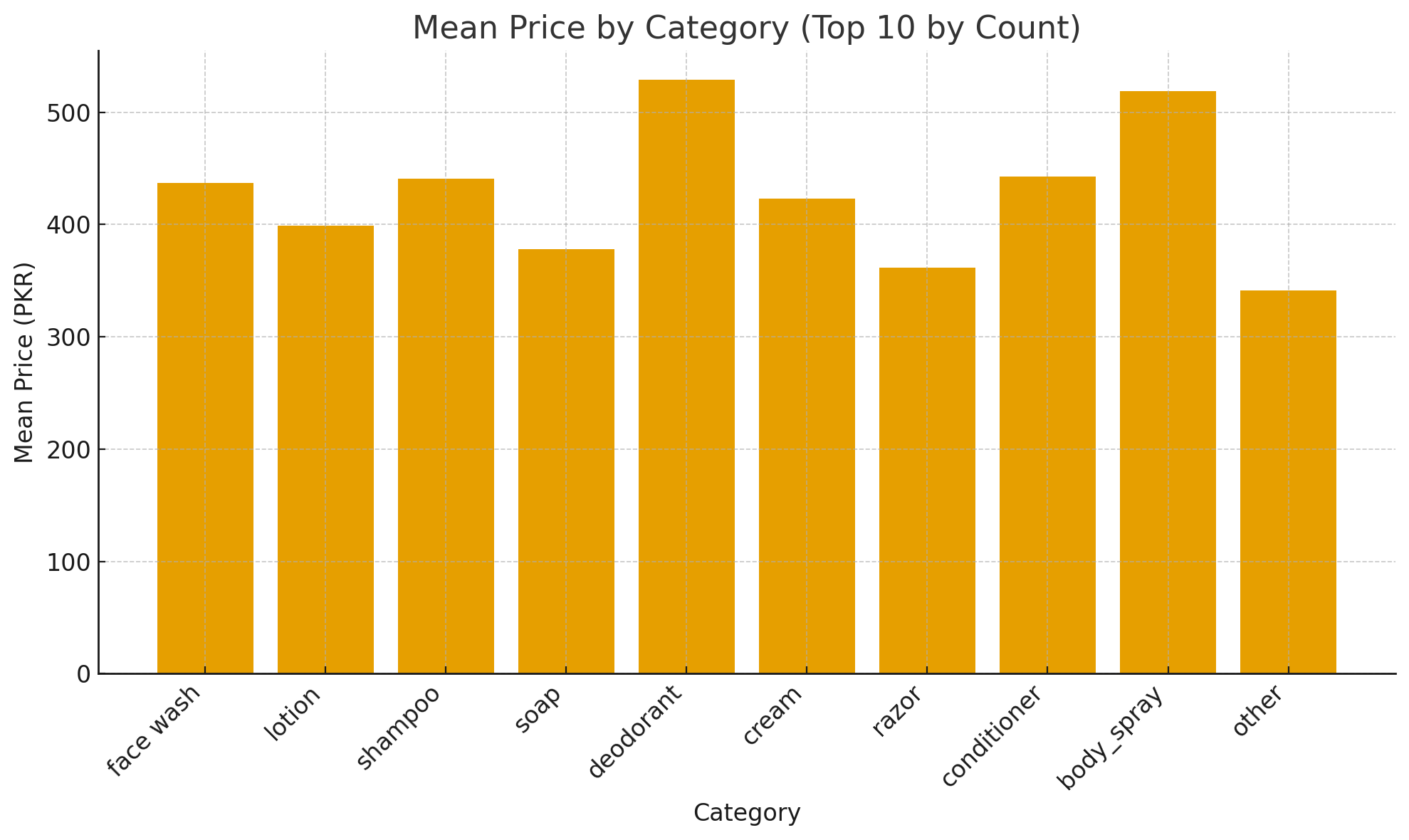
**Standardize Prices:**

We standardized product prices recorded in PKR by addressing entries captured as PKR/1000. Specifically, within the “**price\_pkr”** column, any value beginning with 0. (e.g., 0.695) was interpreted as “per-thousand” and multiplied by 1000 to convert it to plain PKR (e.g., 0.695 → 695). All other numeric values were left unchanged. This procedure affected **1,930** values (**31.61%** of non-missing records). After standardization, price distribution shifted upward as expected: the median increased from **199.9** to **375.0**, and the mean from **295.75** to **429.58**. No additional transformations were applied. The final dataset preserves original structure, with prices now expressed consistently in PKR.

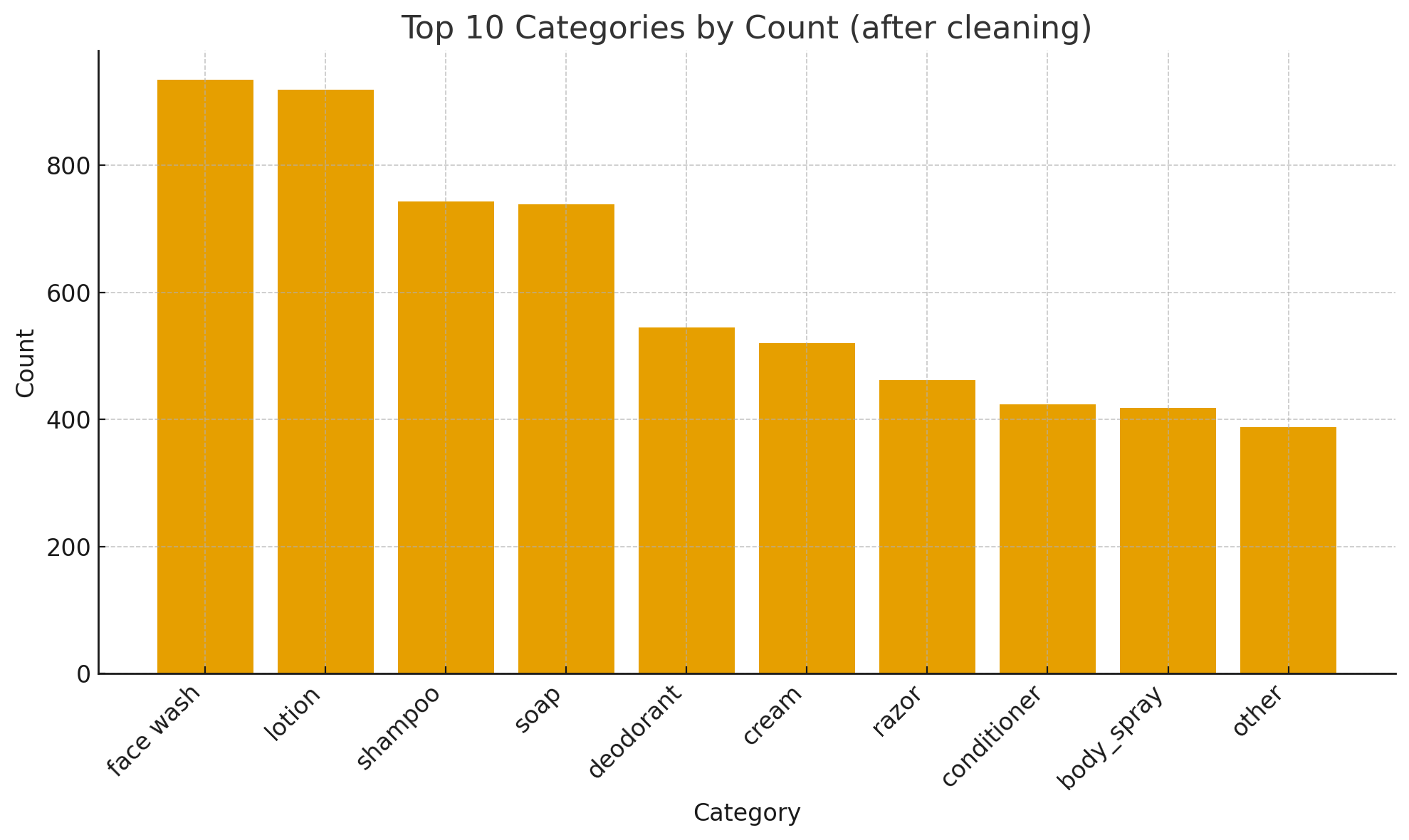
## Table : Adjustment Examples

|  |  |  |  |
| --- | --- | --- | --- |
| price\_column | row\_index | price\_before\_raw | price\_after\_numeric |
| price\_pkr | 0 | 0.42 | 425.00 |
| price\_pkr | 1 | 0.16 | 159.50 |
| price\_pkr | 2 | 0.12 | 119.50 |
| price\_pkr | 3 | 0.13 | 129.50 |
| price\_pkr | 4 | 0.12 | 120.00 |
| price\_pkr | 5 | 0.12 | 119.50 |
| price\_pkr | 6 | 0.16 | 158.00 |
| price\_pkr | 7 | 0.13 | 129.50 |
| price\_pkr | 8 | 0.26 | 264.50 |
| price\_pkr | 9 | 0.23 | 229.50 |

## Mean Price (PKR) by Category



## Top Categories by Count

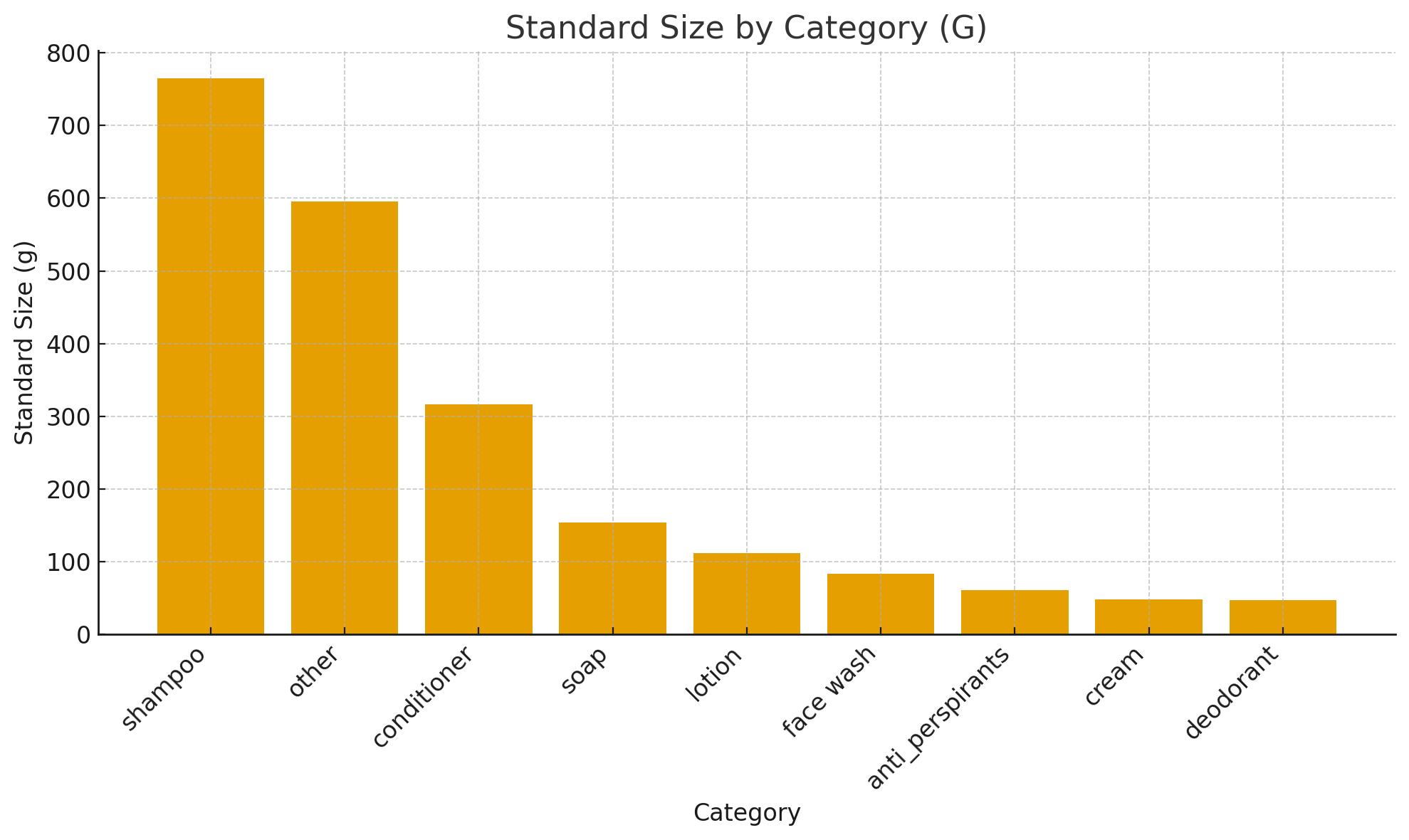


**Standardize size:**

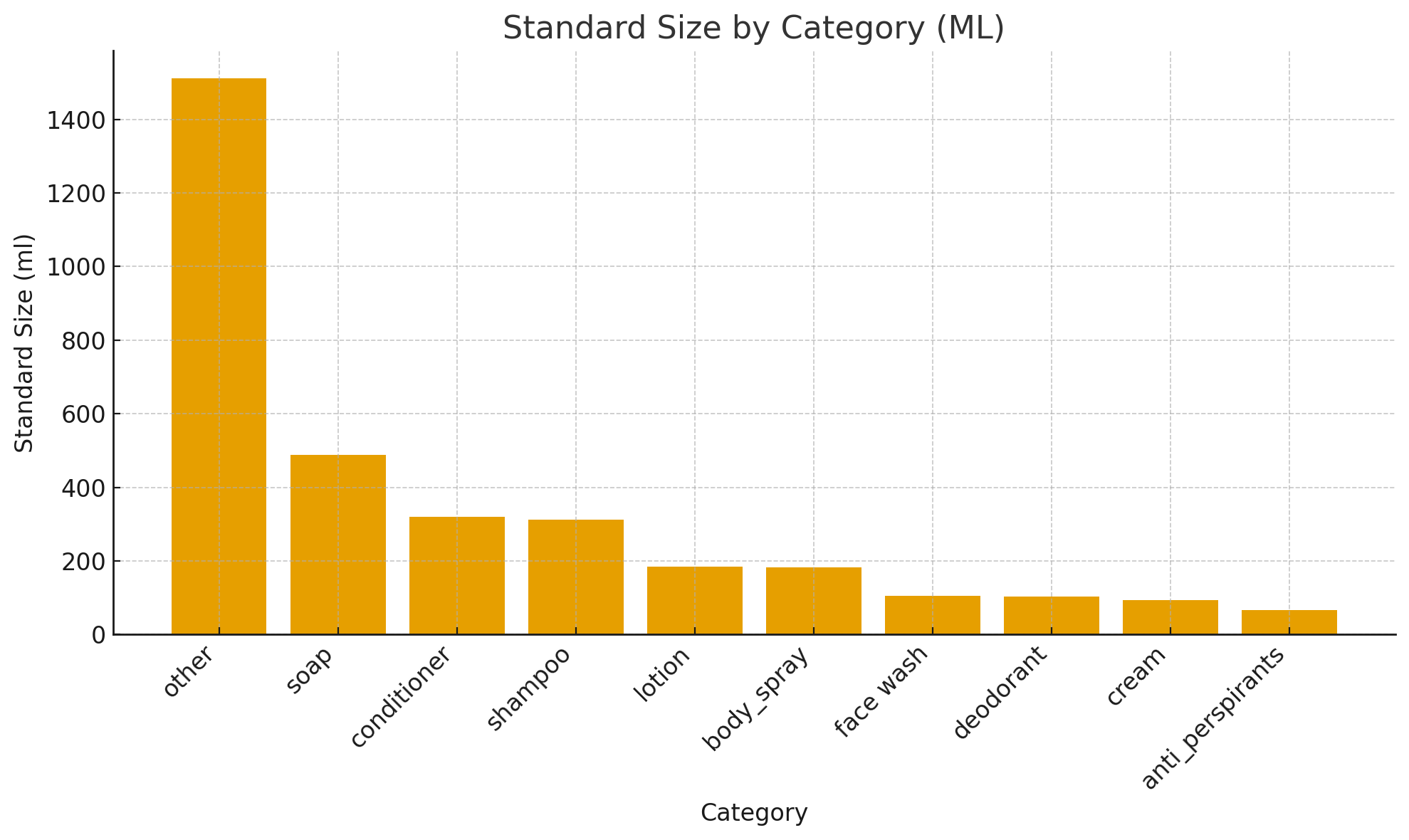
## Non-Standard Prices per Category

Multiple distinct pack sizes within categories indicate non-standardized unit pricing across items. No reliable volume-based sizes detected. No reliable weight-based sizes detected. So we find average size per category and keep that as the standard size. We then find standard prices for that standard size so we can easily compare price based gender differences across categories.  **Average Size per Category**

1. **Average size (g) per category**

****

1. **Average size (ml) per category**

****

1. **Standard size per category:**

## Table 1: Category and Standard Size (ML)

|  |  |
| --- | --- |
| Category | Standard Size (ml) |
| other | 1512 |
| soap | 489 |
| conditioner | 321 |
| shampoo | 313 |
| lotion | 184 |
| body\_spray | 183 |
| face wash | 105 |
| deodorant | 104 |
| cream | 94 |
| anti\_perspirants | 67 |

## Table 2: Category and Standard Size (G)

|  |  |
| --- | --- |
| Category | Standard Size (g) |
| shampoo | 765 |
| other | 595 |
| conditioner | 317 |
| soap | 154 |
| lotion | 112 |
| face wash | 84 |
| anti\_perspirants | 61 |
| cream | 48 |
| deodorant | 47 |

We currently see some categories recorded in both grams and milliliters (for example, shampoo appears as 765 g and 313 ml). To ensure consistency, each category must have a single standard unit and a single standard size. We will determine the standard unit by comparing counts in ml and g within the category and selecting the unit with the higher count. We will then convert all other entries in that category to the chosen unit, compute the average, and round it to define the category’s standard size.

Step 1  
Identify the size unit for every product as either milliliters or grams and ensure the size value is numeric.

Step 2  
Within each category, count how many products are in milliliters and how many are in grams.

Step 3  
Select the standard unit for that category as the unit with the higher count. If the counts are equal, use milliliters.

Step 4  
Convert all size values in the minority unit to the chosen standard unit. Use a one to one conversion between milliliters and grams unless a category specific density is provided.

Step 5  
After conversion, compute the average size for the category in the standard unit.

Step 6  
Round the average to a whole number. This rounded value is the standard size for the category.

Step 7  
Record the final output for each category as category name, standard unit, and standard size. Include a companion table that lists the counts in milliliters and in grams used to determine the standard unit.

## Table 3: Counts by Category (ml vs g)

|  |  |  |
| --- | --- | --- |
| Category | Count (ml) | Count (g) |
| body\_spray | 328 | 0 |
| conditioner | 321 | 12 |
| shampoo | 284 | 15 |
| other | 265 | 4 |
| anti\_perspirants | 218 | 15 |
| face wash | 215 | 37 |
| deodorant | 198 | 13 |
| lotion | 130 | 14 |
| cream | 87 | 85 |
| soap | 53 | 214 |

## Table 4: Chosen Standard Size by Category

|  |  |  |
| --- | --- | --- |
| Category | Chosen Unit | Standard Size |
| soap | g | 154 |
| anti\_perspirants | ml | 67 |
| cream | ml | 94 |
| deodorant | ml | 104 |
| face wash | ml | 105 |
| body\_spray | ml | 183 |
| lotion | ml | 184 |
| shampoo | ml | 313 |
| conditioner | ml | 321 |
| other | ml | 1512 |