PROJECT
Data Visualization

ANALYSIS ON THE DATASET

ADVERSE FOOD EVENTS

About the Dataset

- Overview:
- The dataset contains adverse event reports from the CAERS (Center for Food Safety and Applied Nutrition Adverse Event Reporting System) system.
- The dataset originally had 90,786 rows and 12 columns.

Overview about the Dataset

	count	unique	top	freq	first	last	mean	std	min	25%	50%	75%	max
Report_no	64512.0	NaN	NaN	NaN	NaT	NaT	153693.741769	41443.187708	65325.0	120158.75	163482.5	189232.25	214610.0
Created_date	64512	4020	2017-04-18 00:00:00	185	2004-01-01	2017-06-30	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Start_date	64512	5227	2017-03-22 00:00:00	121	1931-06-19	2017-06-30	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Products_role	64512	2	Suspect	57675	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Products/Brand name	64512	33749	REDACTED	6078	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Industry_code	64512.0	NaN	NaN	NaN	NaT	NaT	40.342711	17.54712	2.0	24.0	53.0	54.0	54.0
Industry_name	64512	40	Vit/Min/Prot/Unconv Diet(Human/Animal)	27830	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Age	64512.0	NaN	NaN	NaN	NaT	NaT	23.799557	29.355295	0.0	0.0	0.0	50.0	736.0
Age_unit	64512	6	Not Available	32301	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Genders	64512	3	Female	40812	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Outcomes	64512	298	NON-SERIOUS INJURIES/ ILLNESS	21240	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Symptoms	64512	33516	OVARIAN CANCER	4499	NaT	NaT	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Data Cleaning

- Before commencing our analysis, it is essential to perform an initial check on our dataset to identify any necessary modifications. Given that the column names are in medical terminology, we opt for simplifying them using the rename() function. Additionally, to enhance clarity and comprehension, we convert the values in the 'start_date' and 'created_date' columns into a datetime format.
- To ensure the integrity of our data, we employ the isnull() function to identify any null values within the dataset.
- This step is crucial as null values can hinder our analysis, and the sum() function is utilized to quantify the count of null values.

```
df.isnull().sum()
```

- Regrettably, there are around 37,000 null values in three columns of this dataset, specifically in:
- Start_Date
- Age
- Symptoms

- Given that the 'Start_Date' represents when the consumer begins to experience the harmful or adverse event, we propose a solution for handling null values. We intend to substitute the null values in the 'Start_Date' column with the corresponding values from the 'Created_Date.' This decision is based on the assumption that consumers would have experienced the adverse event at the time of reporting. The following code snippet illustrates the implementation of this approach:

```
df['Start_date'] = df['Start_date'].mask(df['Start_date'].isna(), df['Created_date'])
```

- With this we could eliminate the NULL values in our Start_Date column . We could also verify it using:

```
df['Start_date'].isnull().sum()
```

- Regarding the 'Age' column, we propose assigning a default value of 0 where information about the consumer's age is unavailable.

- For the 'Symptoms' column, which contains only 5 null values, we decide to drop these specific rows. This removal is deemed acceptable as it will have minimal impact on the overall dataset.

- The 'Age_unit' column exhibits various units, necessitating a uniform unit for consistency in calculations. To achieve this, we compare both the 'Age' and 'Age_unit' columns and convert all values into years, standardizing the age unit for individuals who reported adverse events.

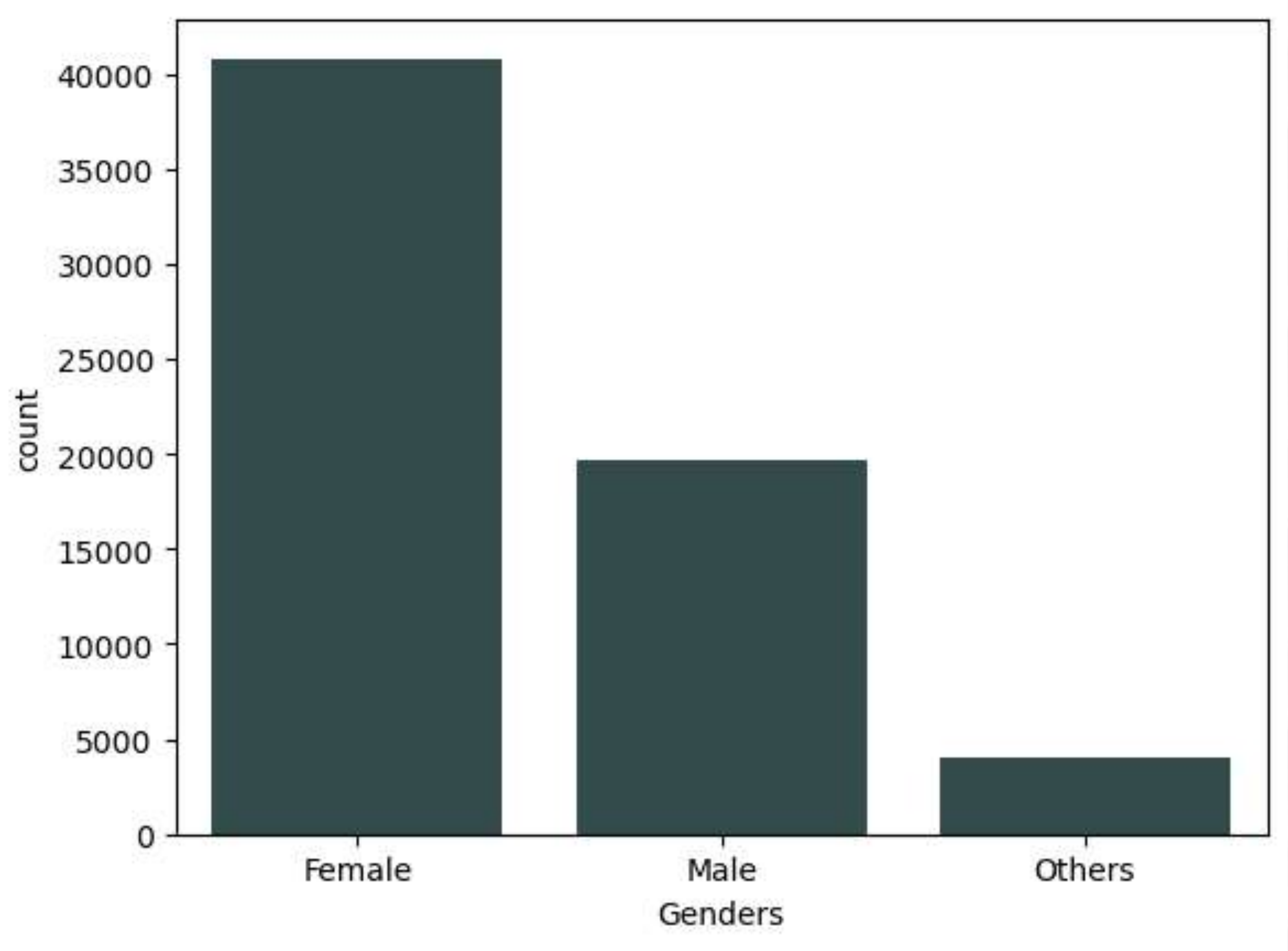
Checking on the Genders

- Upon inspecting the values in the 'Genders' column, we observed five distinct types. To enhance clarity and simplify the categorization, we intend to streamline it into three types: 'Male,' 'Female,' and 'Not Available.' This simplification will provide a clearer understanding of the gender distribution. The transformation will be applied in-place for immediate effect.

```
df['Genders'].replace({'Unknown':'Others', 'Not Reported':'Others','Not Available':'Others'}, inplace=True)
```

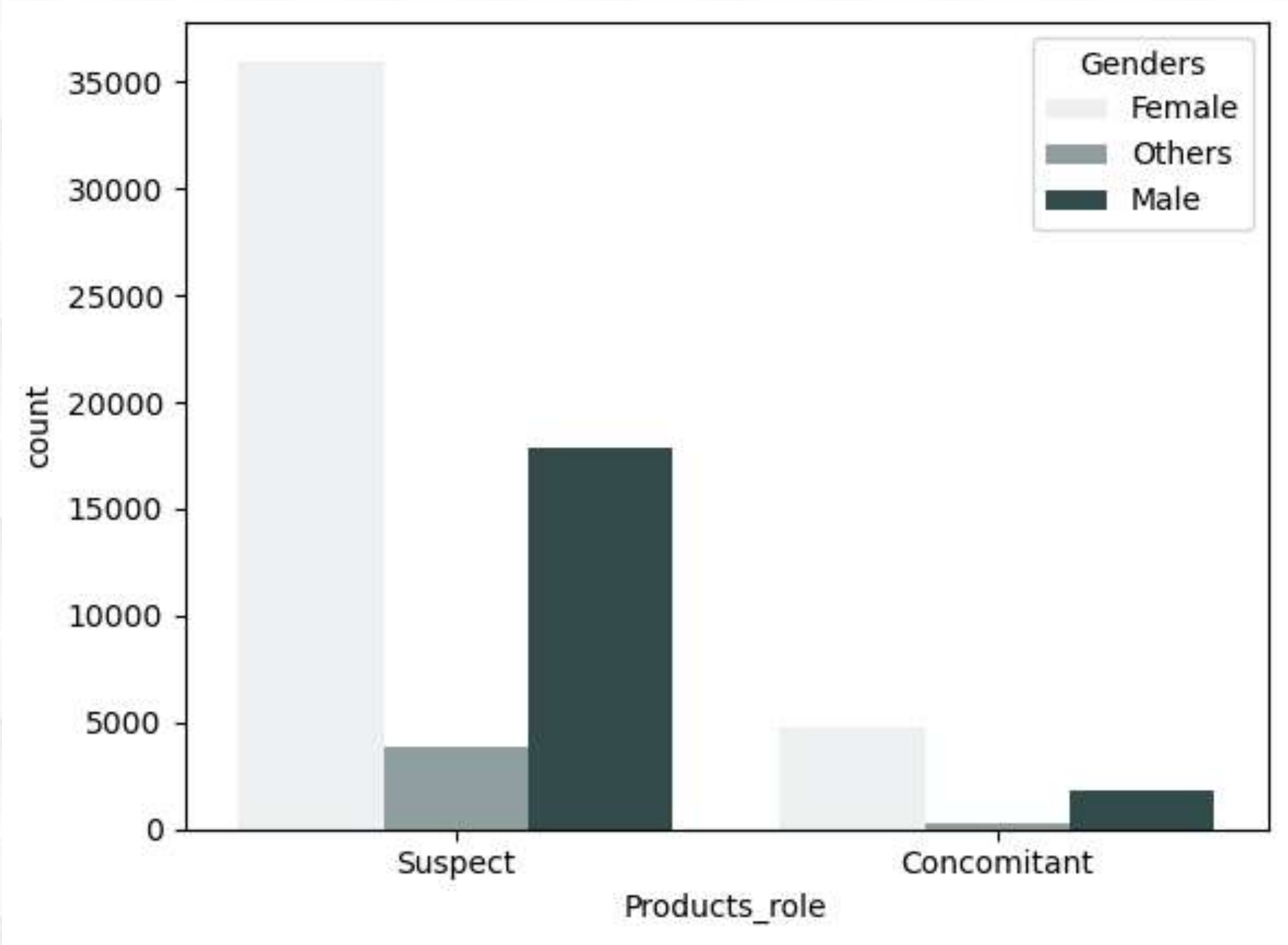
- After changing the values
- Women are more impacted than men, as the majority of reported cases originate from women.
- Approximately 5000 individuals have left this column unfilled.

Genders



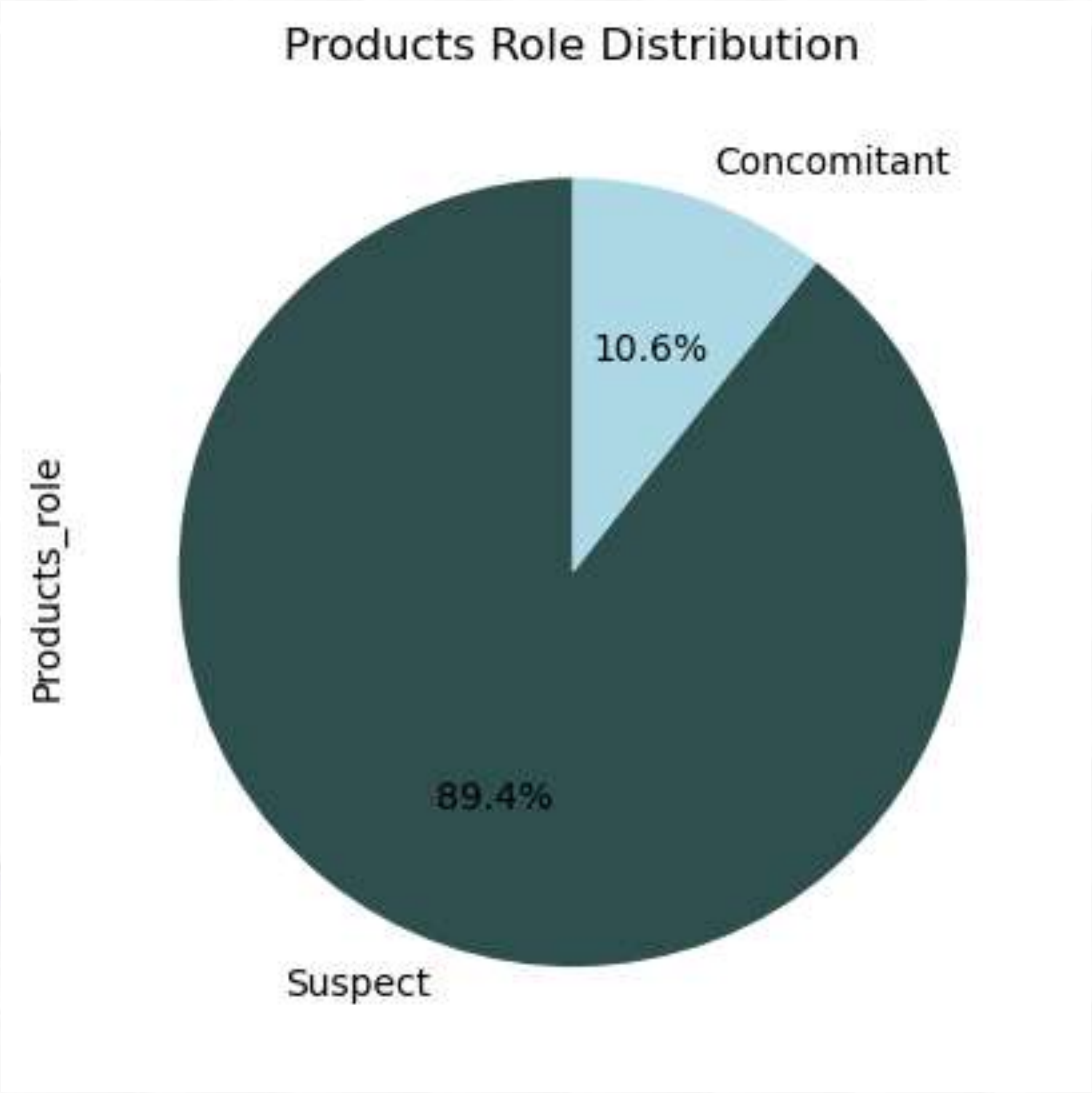
40000
35000
30000
25000
20000
15000
10000
5000
0
count
Female
Male
Others
Genders

Gender Vs Product role



Observing that over 35,000 females have consumed products marked as suspect implies that the public might not have been aware that these products were flagged as suspect.

Products Role

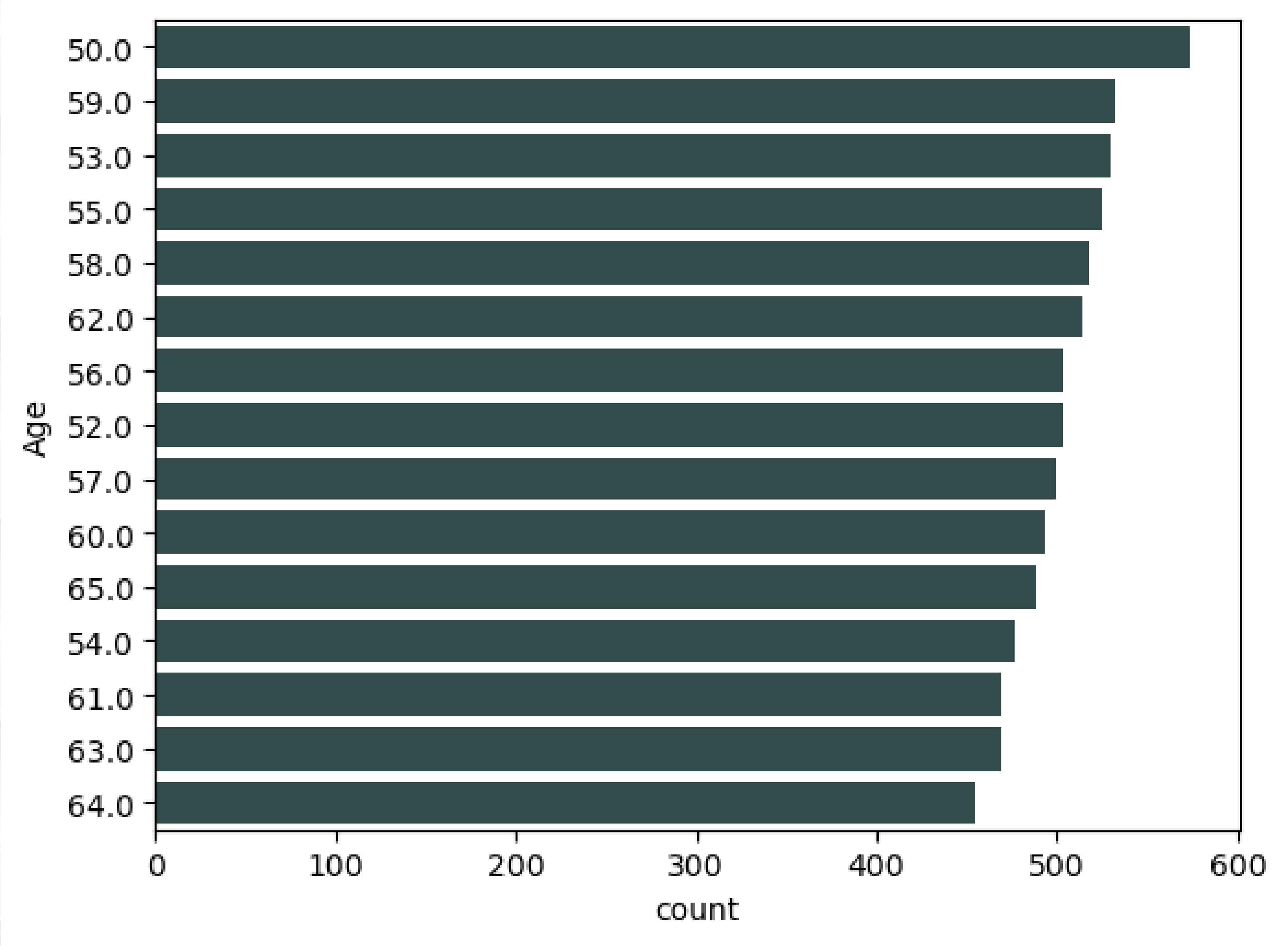


Upon examining the unique values in the 'Products Role' column, two distinct values were identified:

Suspect – Refers to the product under investigation.

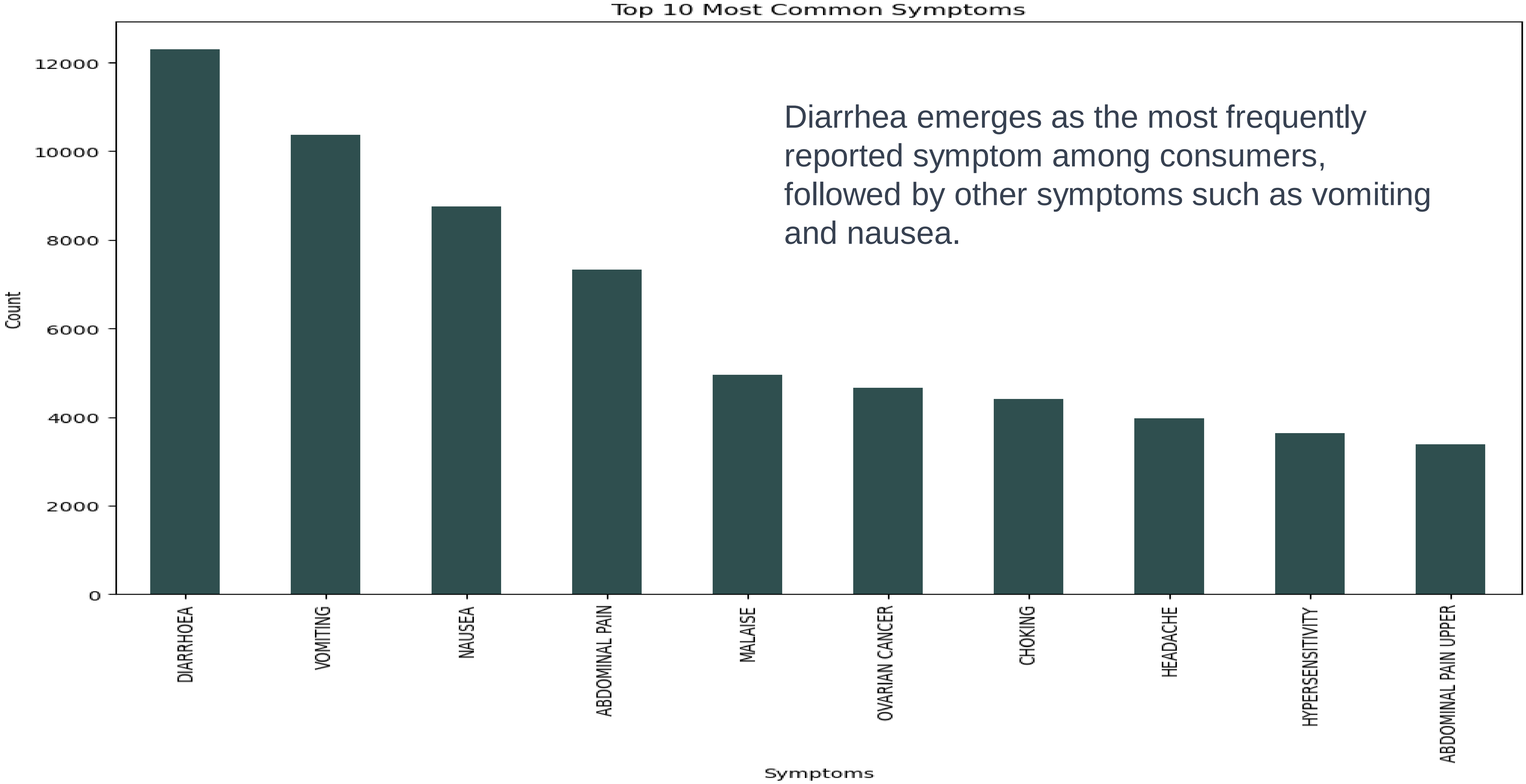
Concomitant – This term can be used to characterize factors or conditions that are simultaneous or interconnected with a specific event or situation.

AGE



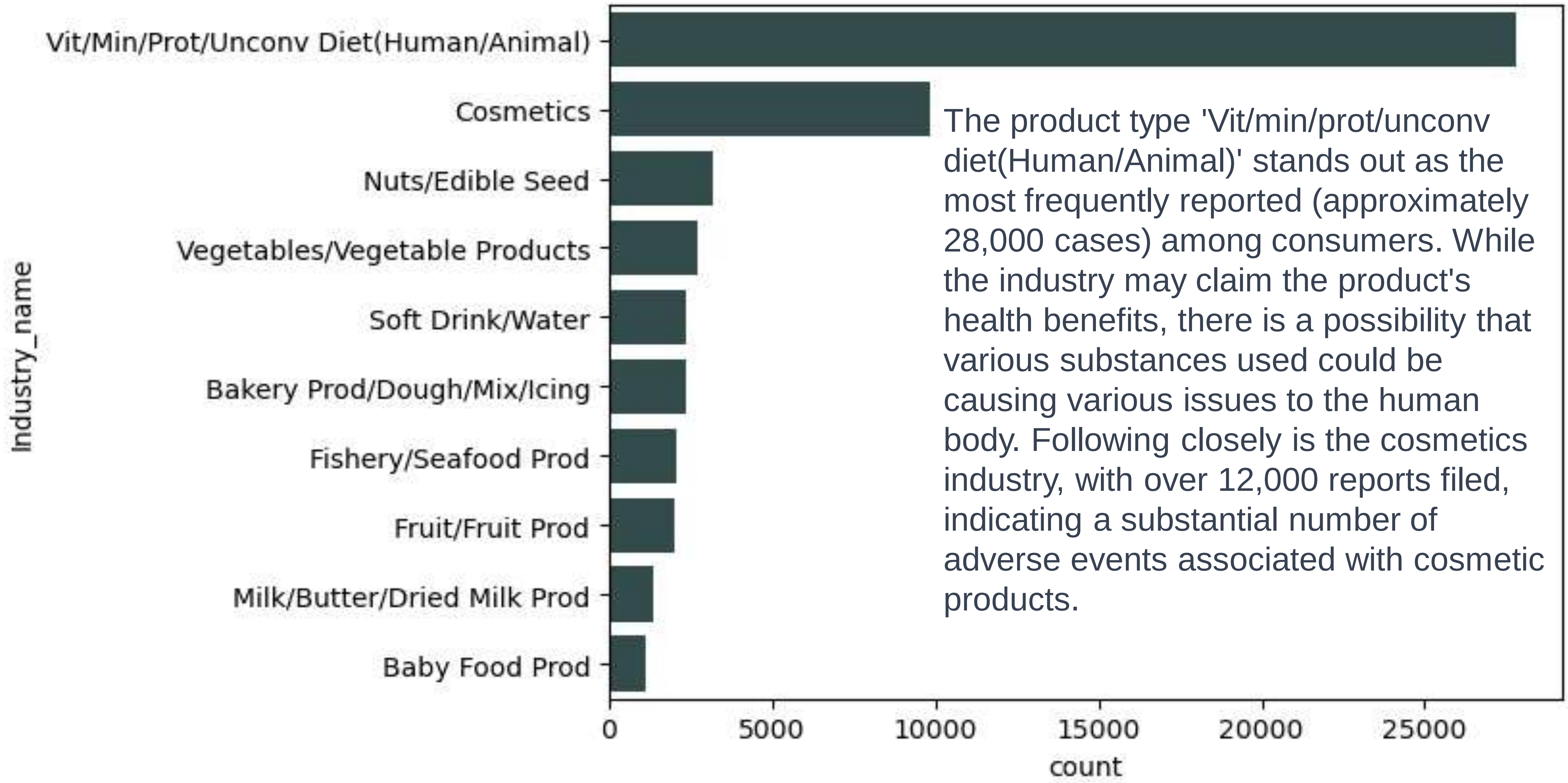
The products primarily affected individuals in their 50s and 60s. There was comparatively less impact on younger individuals, indicating that the immune systems of teenagers and young adults played a significant role in this trend.

Symptoms



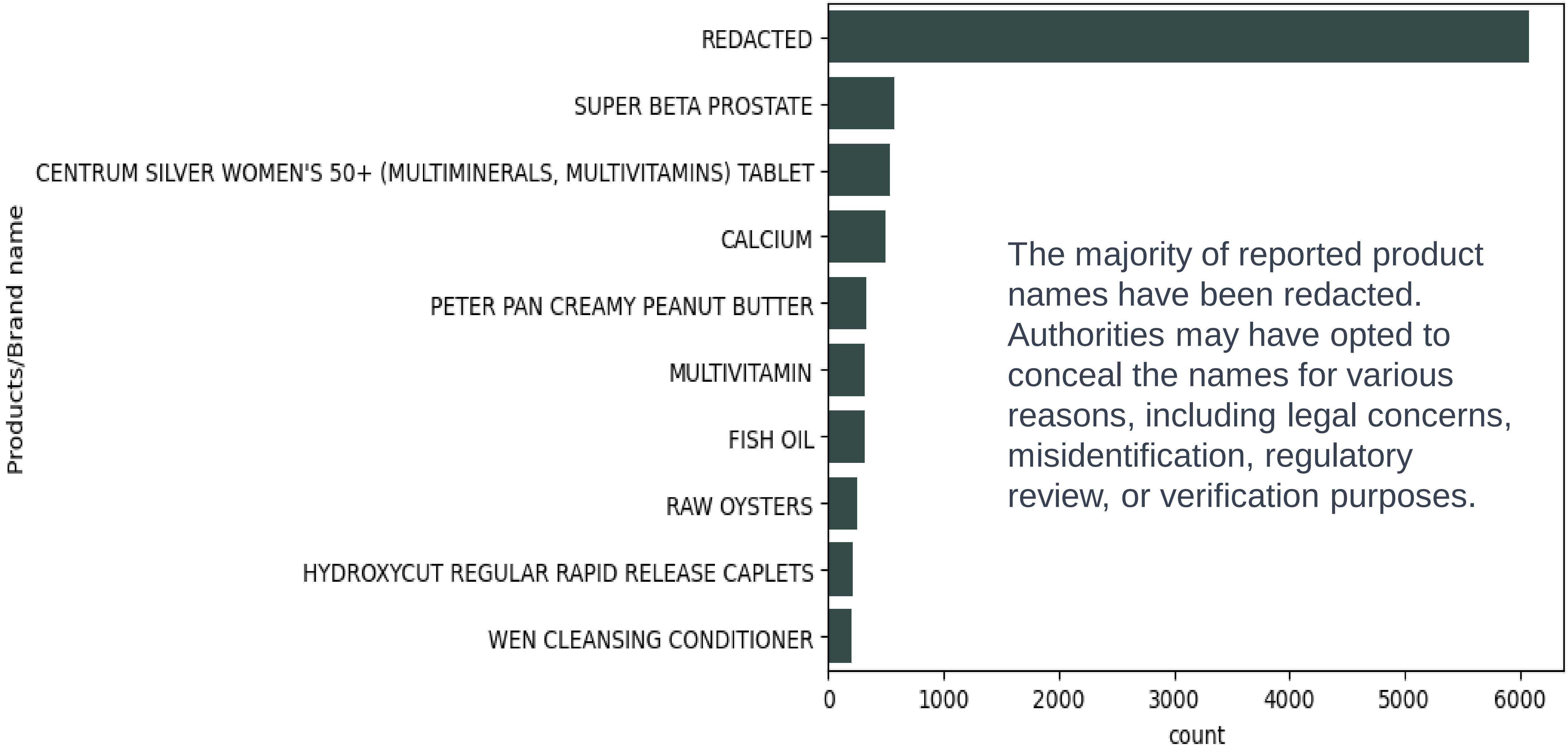
Diarrhea emerges as the most frequently reported symptom among consumers, followed by other symptoms such as vomiting and nausea.

The Most reported industry



The product type 'Vit/min/prot/unconv diet(Human/Animal)' stands out as the most frequently reported (approximately 28,000 cases) among consumers. While the industry may claim the product's health benefits, there is a possibility that various substances used could be causing various issues to the human body. Following closely is the cosmetics industry, with over 12,000 reports filed, indicating a substantial number of adverse events associated with cosmetic products.

Product/Brand name



The majority of reported product names have been redacted. Authorities may have opted to conceal the names for various reasons, including legal concerns, misidentification, regulatory review, or verification purposes.

Important details this analysis revealed

Data Cleaning and Preprocessing:

The dataset underwent initial cleaning, including renaming columns for clarity. Conversion of 'start_date' and 'created_date' values into datetime format. Identification and handling of null values in critical columns.

Gender Distribution:

Women are more commonly affected, as the majority of reported cases originate from females. Approximately 5000 cases lack information about gender.

Suspect Products and Public Awareness:

Over 35,000 females have consumed products marked as suspect, suggesting potential lack of public awareness regarding these flagged products.

Age Analysis:

Products appear to have a more significant impact on individuals in their 50s and 60s, with fewer adverse events reported among younger age groups.

Symptoms Analysis:

Diarrhea is the most frequently reported symptom among consumers, followed by vomiting and nausea.

Product Type Impact:

'Vit/min/prot/unconv diet(Human/Animal)' is the most reported product type, with approximately 28,000 cases. Cosmetics is the second most reported industry, with over 12,000 cases.

These insights provide a comprehensive understanding of adverse events reported in the dataset, highlighting gender distribution, age patterns, symptom prevalence, and the impact of specific product types and industries. Additionally, the presence of redacted product names emphasizes the importance of legal and regulatory considerations in reporting adverse events.