# California Safe Cosmetics Program Data - Exploratory Analysis

Aakriti Poudel

2026-02-01

## Table of contents

## 1 Introduction

This analysis explores the California Safe Cosmetics Program (CSCP) Data, which contains information about cosmetic products sold in California state that contain chemicals known or suspected to cause cancer, birth defects, and other reproductive harm.

## 2 Load necessary packages

```
# Load libraries
library(janitor)
library(scales)
library(viridis)
library(lubridate)
library(tidyverse)
```

## 3 Read the dataset

```
# Read the CSV file
cscp_data <- read_csv(here::here("data", "cscpopendata.csv")) %>%
  clean_names()
```

```
# Basic summary statistics
cat("Dataset Dimensions:", nrow(cscp_data), "rows x", ncol(cscp_data), "columns\n")
cat("\nColumn Names:\n")
print(names(cscp_data))
```

## 4 Data cleaning & wrangling

```
# Clean and wrangle the data
cscp_clean <- cscp_data %>%
  # Convert date columns to date format
  mutate(initial_date_reported = mdy(initial_date_reported),
         most_recent_date_reported = mdy(most_recent_date_reported),
         discontinued_date = mdy(discontinued_date),
         chemical_created_at = mdy(chemical_created_at),
         chemical_updated_at = mdy(chemical_updated_at),
         chemical_date_removed = mdy(chemical_date_removed)) %>%
  # Extract year and month from `initial_date_reported` for time based analysis
  mutate(report_year = year(initial_date_reported),
         report_month = month(initial_date_reported, label = TRUE)) %>%
  # Clean up space in text columns
  mutate(primary_category = str_trim(primary_category),
         sub_category = str_trim(sub_category),
         chemical_name = str_trim(chemical_name),
         company_name = str_trim(company_name),
         brand_name = str_trim(brand_name)) %>%
  # Remove rows with missing information
  filter(!is.na(primary_category) & !is.na(chemical_name))
```

```
# Summary of unique values
cat("Number of products:", n_distinct(cscp_clean$cdph_id), "\n")
cat("Number of companies:", n_distinct(cscp_clean$company_name), "\n")
cat("Number of brands:", n_distinct(cscp_clean$brand_name), "\n")
cat("Number of chemicals:", n_distinct(cscp_clean$chemical_name), "\n")
cat("Number of primary categories:", n_distinct(cscp_clean$primary_category), "\n")
cat("Number of subcategories:", n_distinct(cscp_clean$sub_category), "\n")
cat("Date range:", as.character(min(cscp_clean$initial_date_reported, na.rm = TRUE)),
    "to", as.character(max(cscp_clean$initial_date_reported, na.rm = TRUE)), "\n")
```

```
# Check for missing values in columns
missing_summary <- cscp_clean %>%
  summarise(across(everything(), ~sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "columns", values_to = "missing_counts") %>%
  filter(missing_counts > 0) %>%
  arrange(desc(missing_counts))
```

# 5 Exploratory data visualizations

## 5.1 Figure 1: Top 10 most reported chemicals

This visualization explores which chemicals are most commonly reported in cosmetic products.

```
# Create a subset for count of products by chemical
chemical_counts <- cscp_clean %>%
  group_by(chemical_name) %>%
  summarise(product_count = n_distinct(cdph_id),
            total_reports = n()) %>%
  arrange(desc(product_count)) %>%
  slice_head(n = 10) %>%
  mutate(chemical_name_short = case_when(
    chemical_name == "Silica, crystalline (airborne particles of respirable size)" ~ "Crystal
    chemical_name == "Retinol/retinyl esters, when in daily dosages in excess of 10,000 IU, o
    TRUE ~ chemical_name))

# # Plot top 10 most reported chemicals in cosmetic products
ggplot(chemical_counts, aes(x = reorder(chemical_name_short, product_count), y = product_coun
  geom_segment(aes(xend = chemical_name_short, y = 0, yend = product_count),
               linewidth = 0.8, color = "#820c70") +
  geom_point(size = 3, color = "#820c70") +
  geom_text(aes(label = comma(product_count)), hjust = -0.3, size = 3.5) +
  coord_flip() +
  scale_y_continuous(labels = comma, expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Top 10 most commonly reported chemicals in cosmetic products",
       subtitle = "California Safe Cosmetics Program Data",
       x = "Chemical Name",
       y = "Number of products containing chemical",
       caption = "Source: CSCP Open Data") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 12),
        axis.text.y = element_text(size = 10))
```

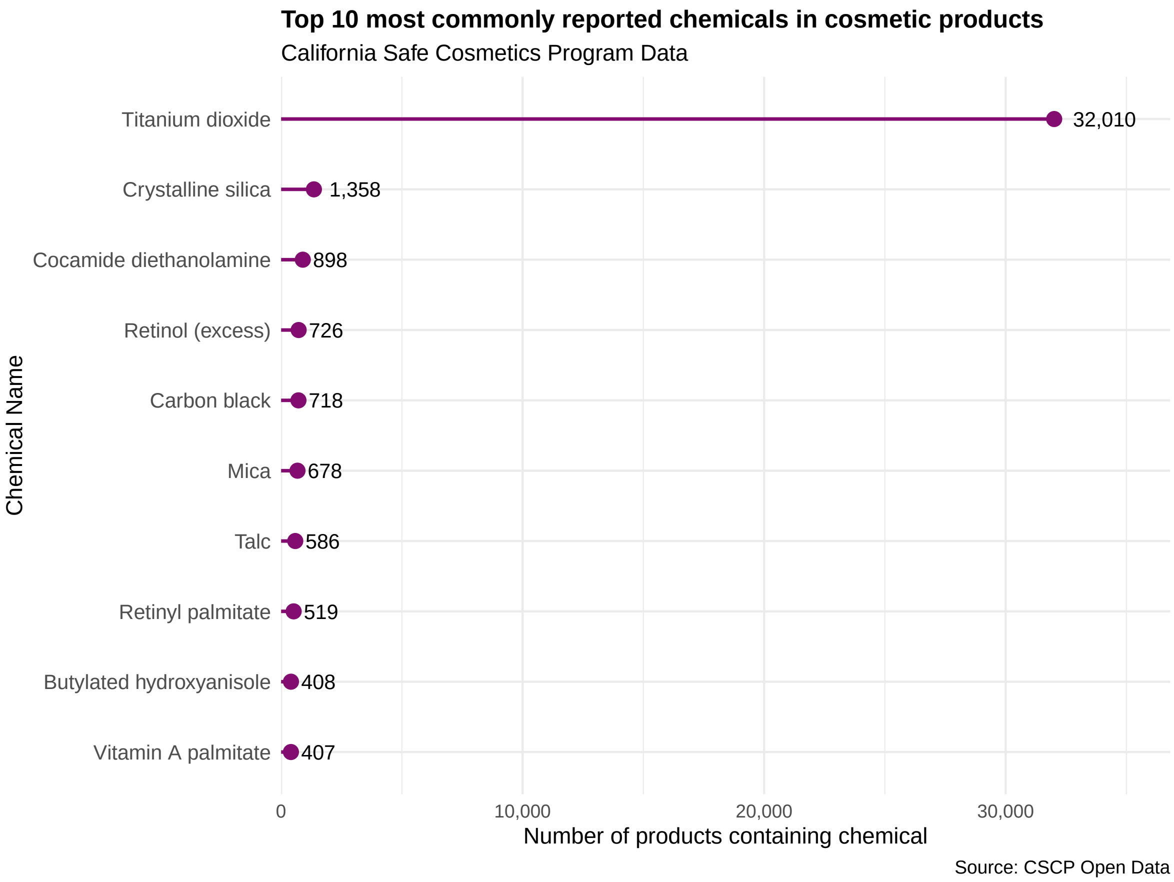


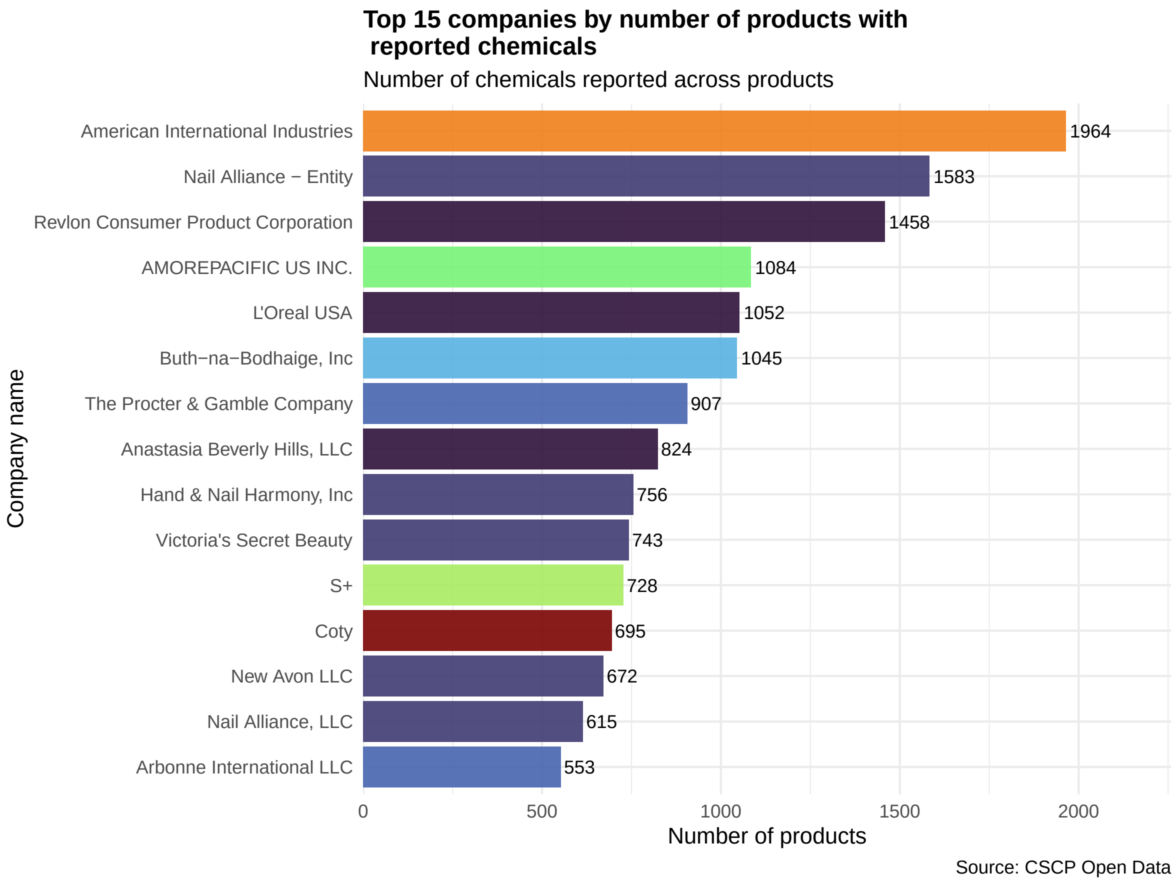
Figure 1: This visualization explores which chemicals are most commonly reported in cosmetic products. Titanium dioxide is the most frequently reported chemical across all product categories.

## 5.2 Figure 2: Top 10 companies with most products with reported chemicals.

This visualization identifies which companies have the most products with reported chemicals.

```r
# Create subset for products by company
company_products <- cscp_clean %>%
  group_by(company_name) %>%
  summarise(unique_products = n_distinct(cdph_id),
            unique_chemicals = n_distinct(chemical_name),
            unique_brands = n_distinct(brand_name)) %>%
  arrange(desc(unique_products)) %>%
  slice_head(n = 15)

# Plot a horizontal bar chart of top companies
ggplot(company_products, aes(x = reorder(company_name, unique_products), y = unique_products)
  geom_col(aes(fill = unique_chemicals), alpha = 0.9) +
  geom_text(aes(label = unique_products), hjust = -0.1, size = 3) +
  coord_flip() +
  scale_fill_viridis_c(option = "turbo") +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Top 15 companies by number of products with\n reported chemicals",
       subtitle = "Number of chemicals reported across products",
       x = "Company name",
       y = "Number of products",
       caption = "Source: CSCP Open Data") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 12, hjust = 0),
        plot.subtitle = element_text(hjust = 0),
        axis.text.y = element_text(size = 9),
        legend.position = "none")
```



Top 15 companies by number of products with reported chemicals

Number of chemicals reported across products

Source: CSCP Open Data

Figure 2: This visualization identifies which companies have the most products with reported chemicals.

### 5.3 Figure 3: Trends over time by primary category

This visualization shows trends in chemical reporting across different product categories over time.

```
# Create a subset from reports by year and category
time_category <- cscp_clean %>%
  filter(!is.na(report_year) & report_year >= 2009 & report_year <= 2020) %>%
  group_by(report_year, primary_category) %>%
  summarise(report_count = n(), .groups = "drop")

# Get top 6 categories by total reports
top_categories <- time_category %>%
```

```r
  group_by(primary_category) %>%
  summarise(total = sum(report_count)) %>%
  arrange(desc(total)) %>%
  slice_head(n = 6) %>%
  pull(primary_category)

# Filter for top categories
time_category_top <- time_category %>%
  filter(primary_category %in% top_categories)

# Plot a line chart of reports over time by category
ggplot(time_category_top, aes(x = report_year, y = report_count, color = primary_category))
  geom_line(linewidth = 1.2) +
  geom_point(size = 2) +
  facet_wrap(~primary_category, scales = "free_y", ncol = 2) +
  scale_color_manual(values = c(
    "Bath Products" = "#ce15b5",
    "Hair Coloring Products" = "#6aa60c",
    "Makeup Products (non-permanent)" = "#dcc50e",
    "Nail Products" = "#d95f09",
    "Skin Care Products" = "#0ee096",
    "Sun-Related Products" = "#115ce0")) +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(breaks = c(2009, 2012, 2015, 2018, 2020)) +
  labs(title = "Trends in chemical reporting by product category over time",
       subtitle = "Top 6 product categories over time",
       x = "Year",
       y = "Number of reports",
       caption = "Source: CSCP Open Data") +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", size = 14),
        legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))
```

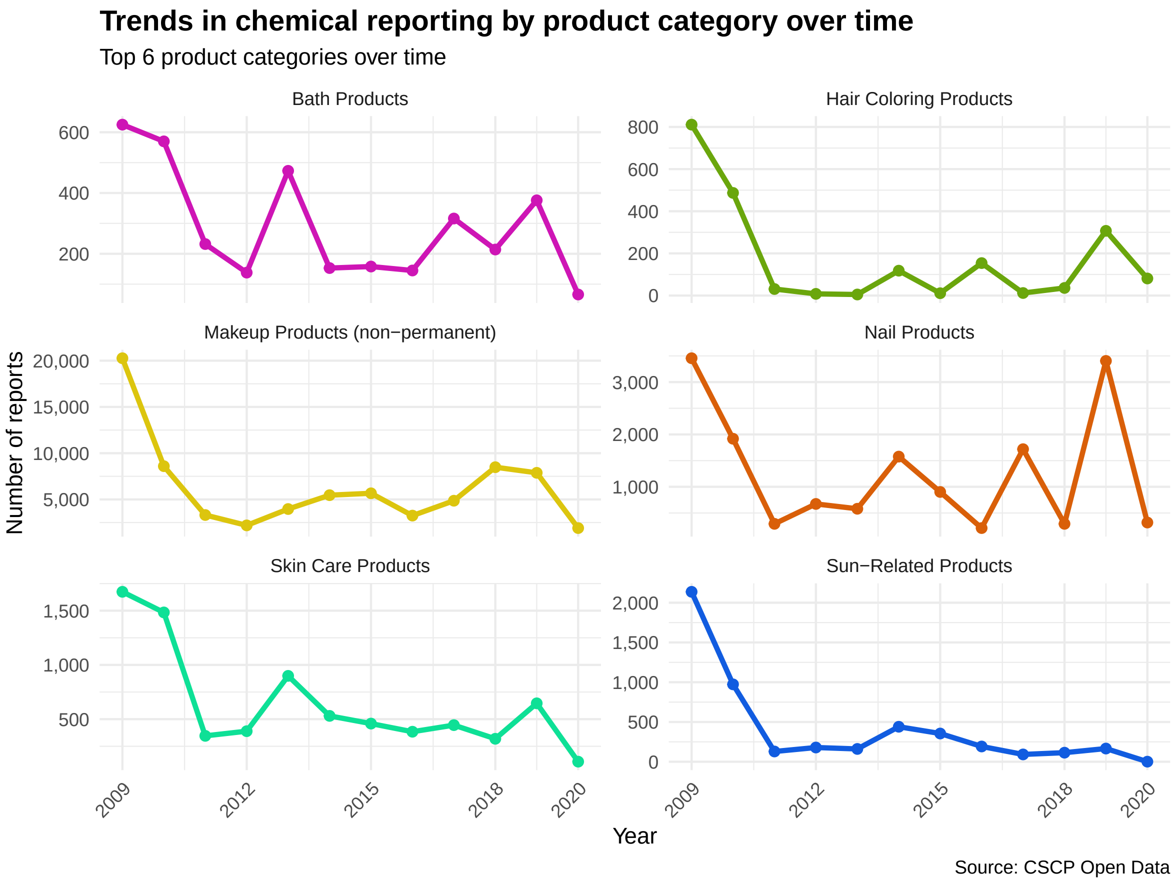


Figure 3: Trends in chemical reporting across the top six cosmetic product categories from 2009 to 2020.

## 6 Answer questions

**Q1. What have you learned about your data? Have any potentially interesting patterns emerged? Point to specific visualizations that you created as you describe your findings.**

**Answer:** My data set is large, containing 114,635 rows and 24 columns. There are 606 companies, 2,573 brands, 36,972 products, and 123 chemicals. The products span 13 primary categories and 89 subcategories in first version of cleaned data set. My data would require significant cleaning and wrangling even if I chose to focus on one brand or one company.

The first visualization explores which chemicals are most commonly reported in cosmetic products. It was found that Titanium Dioxide is the most commonly reported chemical in cosmetic products, followed by Crystalline Silica and Cocamide Diethanolamine.

The second visualization identifies which companies have the most products with reported chemicals. It was found that American International Industries has the most products with reported chemicals, followed by Nail Alliance Entity and Revlon Consumer Products Corporation.

The third visualization examines trends in chemical reporting across different product categories over time. For this, I selected six categories. Bath products, Hair coloring products, Makeup products, Nail products, Skin care products, and Sun-related products. The trend shows that Makeup products has the highest chemical reporting over time.

**Q2. In FPM #1, you outlined some questions that you wanted to answer using these data. Have you made any strides towards answering those questions? If yes, how so? If no, what next steps do you need to take (e.g. I need to create X plot type, I still need to track down Y data, I need to restructure existing data so that you can visualize it in Z ways, etc.)? Have any new questions emerged?**

**Answer:** The visualization answered one of my questions what are the most common harmful chemicals in cosmetics. With my data set, the other two questions can be answered, but I would have to do a lot of data cleaning and wrangling. As of now, I do not intend to change my preliminary questions. I want to explore the data more, and if I feel like I could do something more, then I will consider changing the questions.

**Q3. What challenges do you foresee encountering with your data? These can be data wrangling and / or visualization challenges.**

**Answer:** As I mentioned, the data set is huge. There is so much information, and so much can be done with it. The biggest challenge for me will be data cleaning and wrangling, as far as I can see now.