**Chemical in Cosmetics Dataset:**

These data reflect information that has been reported to the California Safe Cosmetics Program (CSCP) in the California Department of Public Health (CDPH). The primary purpose of the CSCP is to collect information on hazardous and potentially hazardous ingredients in cosmetic products sold in California and to make this information available to the public.

For all cosmetic products sold in California, the California Safe Cosmetics Act (“the Act”) requires the manufacturer, packer, and/or distributor named on the product label to provide to the CSCP a list of all cosmetic products that contain any ingredients known or suspected to cause cancer, birth defects, or other developmental or reproductive harm.

The data table consists of label names of cosmetic/personal care products, company/manufacturer names, product brand names, product categories, Chemical Abstracts Service registry numbers (CAS#) of the reported chemical ingredients, names of reported chemical ingredients, the number of reported chemicals for each product, and dates of reporting, product discontinuation or reformulation if applicable.

**Source:** [**https://data.ca.gov/dataset/chemicals-in-cosmetics**](https://data.ca.gov/dataset/chemicals-in-cosmetics)

**Assignment:**

**Understanding the Data**

* + How would you initially assess the dataset provided? What are the key aspects you look for before processing?
  + Example: Explain the significance of columns like ProductName, ChemicalId, and DiscontinuedDate.

**Preprocessing and Cleaning**

* + What preprocessing steps would you take to clean this dataset before building any models?

**Exploratory Data Analysis (EDA)**

* + Describe the basic exploratory data analysis steps you would perform on this dataset.
  + Example: What visualizations or statistical summaries would you use to understand the distribution of ChemicalId across different ProductNames?

**Key Insights and Analysis**

* + What key insights would you aim to extract from this dataset? How would these insights inform business decisions or further analysis?
  + Example: How would you use clustering techniques to group similar products based on their chemical composition?

**Note**: Feel free to use any clustering, regression-based, machine learning algorithms, etc., to derive insights and enhance your analysis. The examples are not required.