Yidan Sun

Sunyidan@bu.edu

Final project  
Cosmetic chemical system

MET CS 779 Final Project

Outline

Table of Contents

[1. Data Background 2](#_Toc165465029)

[1) Data Link 2](#_Toc165465030)

[2) Data Origin 2](#_Toc165465031)

[3) Data Files 2](#_Toc165465032)

[2. Goal 2](#_Toc165465033)

[3. Tool 2](#_Toc165465034)

[4. Normalization 3](#_Toc165465035)

[5. ERD 3](#_Toc165465036)

[6. Constellation and Slow Changing Dimension 4](#_Toc165465037)

[7. SQL Implementation-Snowflake 4](#_Toc165465038)

[8. Insights and Data Visualization-Tableau 5](#_Toc165465039)

[Bibliography 9](#_Toc165465040)

# Data Background

## Data Link

<https://catalog.data.gov/dataset/chemicals-in-cosmetics-2a971>

## Data Origin

This project utilizes data of chemicals in cosmetics records in California from the year of 2009 to 2020. The 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.

## Data Files

There are three datasets in the file, where one of them called ” cscpopendata.csv” is the main flat spreadsheet that covers all the raw data, 114637 data volume and 22 variables.

File ”chemicalsincosmetics-dd.csv” covers variables explanation.

Another file ”chemicalsincosmetics-dd-subcategories.csv” covers 13 primary categories and corresponding subcategories.

# Goal

The raw data has one flat worksheet with two subsidiary data, recording all California cosmetic associated features from Product name, CDPHID, Company name, Brand name, Primary category to Chemical Number and Chemical count. Project goal is to denormalize data into ERD, do normalization and build slow changing dimensions, write queries to turn data into several relational tables, clean abnormal values, and pop up some business questions for visualizing.

# Tool

Snowflake, Drawio, Excel, Tableau

# Normalization

Original dataset has one spread sheet covers all cosmetic records; the primary key is not showing. It is synthetic primary key combined with using 4 Ids, CDPHId, CSFId, SubCategoryId, ChemicalId, which will need to generate later.

The BrandId is not available and hence need to generate for primary key of brand table. Dates are normalized and create the unique Id for each different date type.

The spreadsheet is separated to dimensional tables, Cosmetic Table, Product Table, Company Table, Brand Table, Color/Scent/Flavor Table, CAS Table, Primary Category Table, Subcategory table, InitialDateReport, MostRecentDateReported, ChemicalDateRemoved, ChemicalUpdatedAt, DiscontinuedDate.

# ERD

A diagram of a computer

Description automatically generated

# Constellation and Slow Changing Dimension

A diagram of a computer

Description automatically generated with medium confidence

# SQL Implementation-Snowflake

Implementation Link: https://app.snowflake.com/lxkxbbu/el48151/w56W9p9HzmJN#query

Import cosmetic table, do data cleaning(null values check, manual mistakes),separate to different tables. See detail implementation as follows:

The final clean version is called “Cosmetic\_Cleaned.csv”

# Insights and Data Visualization-Tableau

Join cosmetic table with initialdatereport table, company, and brand table.

The overall California cosmetic chemicals have significantly decreased for 2009 year and has reached the new lowest in 2020 year. Even though there are local peaks around year of 2015 and 2019, the overall trend of using poisonous chemicals in California cosmetic industry is decreasing.

A graph with a line

Description automatically generated

During the recent year of 2019 and 2020, Charlotte Tibury Beauty Ltd and S+ company have most hazardous chemicals.

A graph of a number of people

Description automatically generated with medium confidence

Among the 13 primary categories of cosmetic products, Most Chemicals reported to be hazardous mainly falls into makeup products, nail products across years.

A graph with green and pink lines

Description automatically generated

Brands like Sephora, Charlotte Tibury, NYX, Palladio, bareMinerals ect have the most chemical reports to Public Health Department of California.

A graph of blue bars

Description automatically generated

As one of brands containing hazardous chemicals, sephora is still increasing the chemical reports in between 2019 and 2020. While most other brands have gradually decreased their chemicals usage.

A graph with lines and numbers

Description automatically generated

Among Brands with top counted hazardous chemicals, Titanium dioxide is dominant in the market with Silica, crystalline, Butylated hydroxyanisole accordingly .

A group of squares with text

Description automatically generated with medium confidence

# Bibliography

U.S. Department of Health & Human Services. (n.d.). Chemicals in cosmetics. Retrieved April 25, 2024, from https://catalog.data.gov/dataset/chemicals-in-cosmetics-2a971