Task 6.1
Negar Nikeghbali
21.05.2024

6.1: Sourcing Open Data

The e-commerce cosmetics dataset available on Kaggle provides comprehensive information about cosmetic products sold online. It includes various attributes such as product names, brands, prices, ratings, and customer reviews. The dataset is structured in a tabular format, likely in CSV or Excel format, containing rows for each product and columns for different attributes. The data can be accessed via the following Kaggle link.

This dataset likely originated from an e-commerce platform or retailer specializing in cosmetics. Each entry represents a unique cosmetic product available for purchase online. The dataset might have been collected over a period to capture a diverse range of products and customer feedback.

Data Choice:

This dataset was chosen due to its relevance to the project objectives, which may involve analyzing trends, customer preferences, and market dynamics within the cosmetics industry. E-commerce datasets offer valuable insights into consumer behavior, product performance, and market trends, making them ideal for various analytical tasks such as market segmentation, pricing optimization, and customer sentiment analysis.

Furthermore, the availability of detailed attributes such as product ratings and total customer reviews allows for in-depth analysis of customer satisfaction and product quality. By analyzing this dataset, we aim to gain actionable insights that can inform business strategies, marketing campaigns, and product development initiatives within the cosmetics industry.

Data Profile

- **website:** Categorical distribution of different e-commerce platforms.
- **country:** Frequency distribution of product listings by country.
- **category:** Distribution of products across different categories.
- **subcategory:** Further granularity within the main categories.
- **price:** Range, mean, median, and standard deviation of product prices.
- **brand:** Distribution of products across different brands.
- **form:** Types of product forms (e.g., liquid, powder).
- **type**: Specific product types within the form (e.g., foundation, lipstick).
- **rating:** Distribution of product ratings, mean and median ratings.
- **count of ratings:** Distribution of the number of ratings received, t

Descriptive Analysis

	price	rating	count of ratings
count	12615.000000	12615.000000	12615.000000
mean	2281.180935	4.257113	994.659770
std	3118.747543	0.556765	6570.809313
min	1.700000	1.000000	0.000000
25%	499.000000	4.060000	6.000000
50%	1390.000000	4.300000	26.000000
75%	2799.500000	4.700000	266.500000
max	94099.000000	5.000000	220040.000000

Column Name	Data Type	Quantitative/Qualitative	Discrete/Continuous/Nominal/Ordinal	Time Invariant/Variant
website	String	Qualitative	Nominal	Time Invariant
country	String	Qualitative	Nominal	Time Invariant
category	String	Qualitative	Nominal	Time Invariant
subcategory	String	Qualitative	Nominal	Time Invariant
price	Float	Quantitative	Continuous	Time Invariant
brand	String	Qualitative	Nominal	Time Invariant
form	String	Qualitative	Nominal	Time Invariant
type	String	Qualitative	Nominal	Time Invariant
rating	Float	Quantitative	Continuous	Time Invariant
count of ratings	Integer	Quantitative	Discrete	Time Invariant

Data Cleaning/Renaming/reformatting

Column	Action	Count
type	NaN replaced with mode	2681
color	Removed: it has many false entered data such as NaN, #Error, 9#, 4#, long text,...	
size	Removed: The same unit for the size is not mentioned and could not be also estimated, therefore should this column be removed	

rating	2067 NaN's replaced with mean,
	57 Values > 5 replaced with 5;
	replace(',', '.');
	replace(' out of 5 stars', '')
	From str into float
	replace(',' into'.');
	2 replace(' out of 5 stars' into'')
	From str into float
		2126

noofratings	733 replace 'New to Amazon', 'No reviews', 'Write A Review' with 0 459 replace NaN's with 0 replace(','into '') 374 replace('ratings' into '') Renamed 'noofratings) into 'count of rating' From str into int 459 replace NaN's with 0 replace(','into '') 374 replace('ratings' into '') Renamed 'noofratings) into 'count of rating'	1566
price	NaN replaced with mean	317

Limitations and Ethics

- **Limitations**: The dataset has many missing and inconsistent data like in the 'color' and 'size' columns. In addition, only India and the U.S. have been investigated. The dataset has no/column to represent the time of gathering data(year, months,…) so this data set might not accurately represent the entire market.
- **Ethical Considerations**: The dataset does not contain personal information, mitigating privacy concerns. However, the source and data collection method should be considered to ensure the data is used ethically.

Questions to Explore

In this analysis, the following questions will be explored:

1. Which brands are the most famous?

2. Which category is the most famous?

3. Which subcategory is the most famous?

4. What is the distribution of product prices across different categories and subcategories?

5. How do product ratings vary by product category and subcategory?

6. What is the relationship between the number of ratings and the average rating of products?

7. Which brands have the highest number of high-rated products?

8. Which country represents the better price?