Sephora Data Analysis

Preliminary Work – Getting Started

The Sephora Product Dataset from Kaggle (Sephora, 2025) was selected for the Module 4 project to analyze how product attributes influence customer engagement and satisfaction within the beauty industry. The dataset includes over 10,000 products and 20+ attributes, such as brand, category, price, average rating, number of reviews, and whether a product is part of a limited edition or exclusive line.
Sephora Products and Skincare Dataset – Kaggle

This dataset offers an opportunity to explore consumer behaviors in luxury retail—particularly how product type, pricing, and brand perception shape popularity and satisfaction. The analysis will be conducted using R, employing packages like tidyverse for data manipulation, janitor for cleaning, and ggplot2 for visualization. Planned preprocessing steps include handling missing values, transforming variables into factors, and creating derived metrics such as average rating per brand and price category segmentation.

Data Questions and Planned Approach

1. **How do product categories differ in customer satisfaction and popularity?**
   Product types (skincare, makeup, fragrance, etc.) will be grouped and analyzed based on average ratings and number of "loves."
   Planned Visualization: Bar chart comparing average rating and popularity (loves count) across product categories.

2. **Does product price correlate with customer satisfaction?**
   A scatter plot will be created between price and average rating to determine whether higher prices correspond to better satisfaction levels or if affordable products perform equally well.
   Planned Visualization: Scatter plot with trendline showing correlation between price and rating.

3. **How does brand exclusivity influence customer engagement?**
   By comparing exclusive vs. limited-edition vs. regular brands, the analysis will reveal if exclusivity drives higher ratings or popularity.
   Planned Visualization: Boxplot comparing customer engagement (loves count) across exclusivity levels.

Part I – Exploring

4. Review any written description of your dataset. This is often referenced as the data dictionary.
5. Clean your data. Cleaning involves any task that prepares the dataset for analysis. This might include the following tasks:
    a. Renaming columns
    b. Managing NAs
    c. Correcting data types
    d. Removing columns or rows
    e. Manipulating Strings
6. Determine descriptive statistics for interesting variables.
7. Produce visualizations from the raw data that identify and highlight interesting aspects. These can include bar charts, histograms, line graphs, scatter plots, etc. Be sure the chosen graph best represents the information.

Part II – Expanding

1. Create new variables that better capture the data you want to report. For example, if the data shows yearly sales by month, you might calculate the month-to-month increase or decrease in sales as a new column.
2. Group, summarize, rank, arrange, count, or perform any other useful operations to create new data frames that provide access to different views of the data.
3. Extract the most interesting data results and produce visualizations that best communicate these results.

Part III – Communicating

1. Report what you have learned about your data. Identify 3–5 observations or follow-up questions that you could explore in the future.
2. Complete all data management tasks in R.