**Project Title:**

**Exploring Sephora: Data Analysis and Visualization with Python.**

**Project Description:** This project delves into the Sephora dataset, a renowned name in cosmetics and beauty retailing. The main goal is to extract valuable insights from the dataset, exploring customer preferences, product prices and correlations between various features.

**Tools and Techniques:**

* Python (Pandas, Matplotlib, Seaborn)
* Exploratory Data Analysis (EDA)
* Data Visualization
* Correlation Analysis
* Word Clouds

**Data Acquisition**: The initial dataset for this analysis was downloaded from Kaggle.com. It comprises information on over 8,000 beauty products from Sephora's online store and was collected using a Python web scraper in March 2023. The dataset includes details like product and brand names, prices, ingredients, rating and more.

**Data Preparation:** The dataset was carefully prepared to address data issues. Columns with high percentages of missing data were removed and missing values in other columns were handled appropriately. Additionally, the "size" column, due to inconsistencies and missing values, was excluded from the analysis.

**Summary Statistics:** Python methods were employed to derive summary statistics.

**Univariate Analysis**: During this phase, individual variables such as "likes\_count", "reviews" and "price\_usd" were examined, revealing significant outliers, boxplots were generated to visualize the data, both with and without these outliers.

In terms of ratings, most products fell within the range of 4 to 4.75, with the highest count occurring between 4.25 and 4.5.

**Categorical Analysis**: Among the product categories, Skincare and Makeup were the most dominant. In terms of brands, SEPHORA COLLECTION took the lead with 352 products.

**Correlation Analysis:** A notable positive correlation of 0.69 exists between "likes count" and "reviews", but external factors may influence these features, requiring further investigation.

No significant correlation was observed between “price\_usd” and other features.

**Price Analysis**: A Boxplot visualization was created to illustrate the price distribution within each category. It reveals that the "Gift" category contains only 4 products, all priced at 50 USD each.

When considering brand prices, ELUMINAGE stands out with the highest average price.

Additionally, the analysis delved into boolean variables associated with prices, uncovering that SEPHORA EXCLUSIVE products are generally less expensive, whereas LIMITED EDITION, ONLINE ONLY and NEW products tend to be priced higher.

**Likes Count Analysis**: Average number of “likes count” were employed to reduce the impact of outliers, highlighting the Makeup category as the most popular. Regarding brands, OLAPLEX stands out with a notably higher average number of "likes count".

**Reviews Analysis**: As with "likes count", the average number of “reviews” was employed to mitigate the influence of outliers. Analyzing “reviews” revealed that the Makeup and Mini-size categories had the highest average number of “reviews” and BUXOM emerged as the brand with the significantly higher average number of "reviews" compared to other brands.

**Ingredient Analysis**: A list of the 20 most popular ingredients and an Ingredients Word Cloud were created, providing insights into the common ingredients in the dataset.

**Conclusion:**

While the analysis was not exhaustive due to data limitations, it provided valuable information about customer preferences, pricing and correlations between features. These findings provide a robust foundation for further exploration and decision-making in the cosmetics and beauty retail industry.