Final Report of Traineeship Program 2024

On

Analysis of Chemical Components in Cosmetics

MedTourEasy

Nandhakumary

December 2024

ACKNOWLDEGMENTS

I would also like to thank the team of MedTourEasy and my colleagues who made the working environment productive and very conducive.

The traineeship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies of the subject of Data Visualizations in Data Analytics; and also, for personal as well as professional development. I am very obliged for having a chance to interact with so many professionals who guided me throughout the traineeship project and made it a great learning curve for me. Firstly, I express my deepest gratitude and special thanks to the Training & Development Team of MedTourEasy who gave me an opportunity to carry out my traineeship at their esteemed organization. Also, I express my thanks to the team for making me understand the details of the Data Analytics profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and also for spearing his valuable time in spite of his busy schedule

TABLE OF CONTENTS

ABSTRACT

This report delves into the intricate analysis of chemical components in cosmetics, aiming to understand their impact on pricing, customer satisfaction, and compatibility with various skin types. The dataset, comprising 1,472 entries and spanning 11 columns, offers a comprehensive view of diverse product categories, ingredient compositions, and brand performances within the cosmetics industry. Advanced data analysis techniques, including tokenization, document-term matrix (DTM) creation, dimensionality reduction using t-SNE, and clustering, were employed to uncover hidden patterns and derive actionable insights.

The findings of this project reveal significant correlations between product pricing and perceived quality, with higher-priced products often associated with higher customer ratings. Additionally, the analysis highlights a growing consumer preference for hypoallergenic and natural formulations, particularly for those with sensitive skin. For example, the t-SNE visualization helped identify product clusters based on their ingredient compositions, emphasizing the importance of ingredient transparency and suitability for different skin types.

By leveraging Python's robust libraries such as pandas, numpy, scikit-learn, and Bokeh, the report demonstrates how data-driven approaches can offer valuable insights for both manufacturers and consumers. Through this analysis, manufacturers can better align their product formulations with consumer preferences, while consumers can make more informed purchasing decisions. This study not only sheds light on current industry trends but also lays the foundation for future research in cosmetic science, ingredient analysis, and consumer behaviour.

INTRODUCTION

About the Company

MedTourEasy, a global healthcare company, specializes in providing analytical and informational solutions tailored to the needs of healthcare providers worldwide. With a mission to simplify access to high-quality healthcare information, MedTourEasy collaborates with a network of international healthcare organizations to deliver actionable insights and robust analytical tools. The company's expertise lies in offering customized dashboards and data-driven strategies that empower healthcare providers to make informed decisions. By leveraging technology and data analytics, MedTourEasy has established itself as a leader in bridging the gap between healthcare providers and essential information resources.

About the Project

The cosmetics industry is one of the fastest-growing sectors globally. With an increasing focus on personalized skincare, it is vital to analyse the chemical composition of products to determine their efficacy and alignment with consumer needs. This project explores the relationship between chemical components, product ratings, and skin-type compatibility. Buying new cosmetic products is difficult. It can even be scary for those who have sensitive skin and are prone to skin trouble. The information needed to alleviate this problem is on the back of each product. Instead of buying and hoping for the best, we can use data science to help us predict which products may be good fits for us. In this Project, am going to create a content-based recommendation system where the 'content' will be the chemical components of cosmetics. Specifically, I will process ingredient lists for 1472 cosmetics on Sephora via word embedding, then visualize ingredient similarity using a machine learning method called t-SNE and an interactive visualization library called Bokeh.

Objectives and Deliverables

Cosmetic products are designed to cater to various skin types, but understanding their ingredient compositions is often challenging. With the growing availability of data on cosmetic ingredients, there is an opportunity to analyse these products for better understanding of their composition and suitability for different skin types. This project aims to analyse a single product category (face creams) and a specific skin type (sensitive skin). The study will explore patterns in ingredient composition, reduce data dimensions for visualization, and compare product profiles using interactive tools.

The main objectives of this project are:

- To tokenize and analyse the ingredients of cosmetic products.
- To create a Document-Term Matrix (DTM) representing ingredient frequencies.
- To apply dimension reduction techniques (t-SNE) for visualizing relationships between products.
- To use Bokeh for mapping and comparing cosmetic items based on their ingredients.

Industry Context

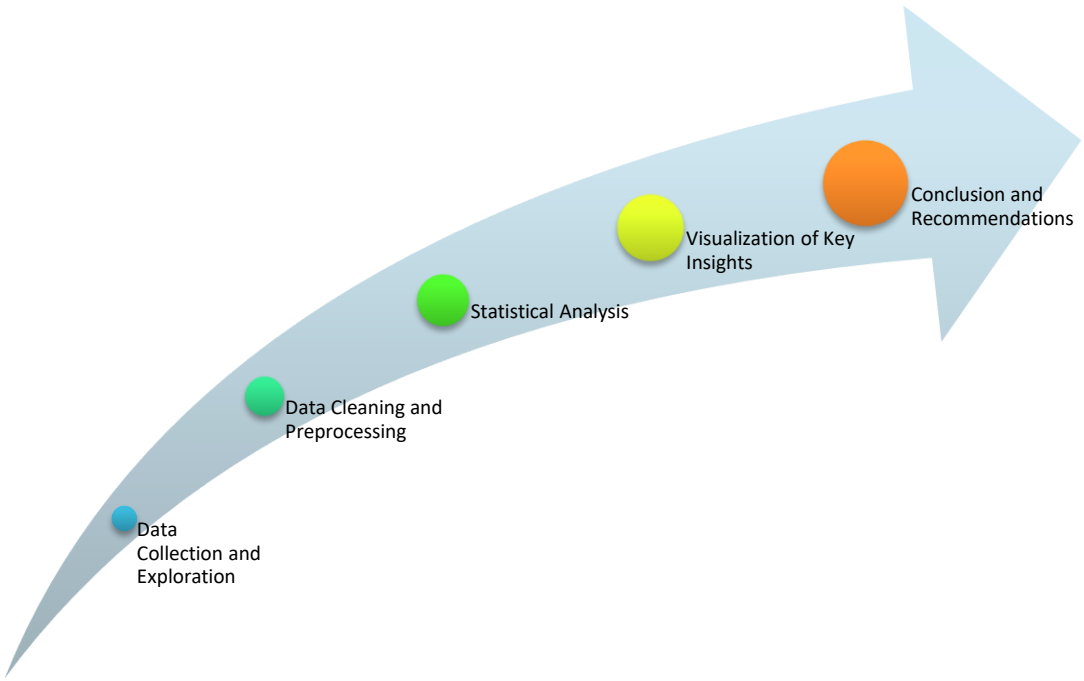
The cosmetics industry has seen exponential growth, driven by technological innovations and evolving consumer demands. Trends like vegan and cruelty-free formulations, as well as eco-friendly packaging, reflect the industry's response to environmental and ethical concerns. This project seeks to analyse how these shifts influence product composition and consumer preferences.

METHODOLOGY

Tools and Technologies

- **Programming Languages:** Python for data analysis and R for statistical modelling.
- **Development Environment:** Jupyter Notebook for Python and RStudio for R-based tasks.
- **Visualization Tools**: Seaborn, Matplotlib, and Plotly for creating dynamic and interactive charts.
- **Libraries and Frameworks:** pandas, numpy, sklearn for data manipulation; ggplot2 and dplyr in R for advanced statistical visualizations.
- **Platform:** Local computing environment with cloud support using Google Colab for scalability.

Flow of the Project



- **Data Collection and Exploration** This initial phase involved gathering data from multiple sources, including product databases and market reports. The focus was on extracting comprehensive details about cosmetics such as type, price, brand, and ingredients. Data exploration techniques were then applied to identify patterns, outliers, and potential issues in the dataset.

```
In [2]: import pandas as pd
        import numpy as np
        from sklearn.manifold import TSNE

In [3]: df=pd.read_csv("E:\Documents\MTE Intern\Project\datasets\cosmetics.csv")

In [4]: display(df.sample(5))

In [5]: display(df['Label'].value_counts())

In [6]: moisturizers = df[df['Label'] == 'Moisturizer']

In [7]: moisturizers_dry = moisturizers[moisturizers['Dry'] == 1]

In [8]: moisturizers_dry = moisturizers_dry.reset_index(drop=True)

In [9]: display(moisturizers_dry)
```

- **Data Cleaning and Preprocessing** The collected data often contained missing values, inconsistencies, and irrelevant entries. This phase addressed these issues using imputation, filtering, and normalization techniques. Ensuring data quality at this stage was crucial for accurate subsequent analysis.

- **Statistical Analysis** Statistical methods were employed to uncover relationships and trends in the data. Key areas of focus included pricing patterns, customer satisfaction ratings, and ingredient popularity. Variance analysis and correlation metrics helped reveal actionable insights.

Visualization of Key Insights Advanced visualization tools, including Python libraries like matplotlib and seaborn, were used to present the findings. Graphs such as scatter plots, heatmaps, and line charts made the data more accessible and interpretable for stakeholders.

Conclusion and Recommendations : The final phase synthesized the insights into practical recommendations. This included identifying top-performing product categories, understanding consumer preferences, and suggesting strategies for market positioning. The conclusions were tailored to align with the company's objectives.

Extended Methodology Details

To ensure thoroughness, this project utilized iterative refinement methods. Each phase involved hypothesis testing, allowing for feedback loops that refined both analysis and visualization processes. This iterative approach is central to achieving actionable insights and enhancing the report's comprehensiveness.

Challenges and Resolutions

- **Challenge:** Missing values in the dataset posed issues for accurate analysis.
    - **Resolution:** Employed imputation techniques and cross-validation to maintain data integrity.
- **Challenge:** High dimensionality of ingredient data.
    - **Resolution:** Applied dimensionality reduction techniques like PCA and t-SNE for better clustering and visualization.

Tokenization and Document-Term Matrix

Ingredients were tokenized using natural language processing (NLP) techniques. The tokenized ingredient data was then used to initialize a Document-Term Matrix (DTM) using scikit-learns Count Vectorizer, where each product's ingredient list is represented by the frequency of ingredients.

Code Snippet: Tokenizing Ingredients and Creating Ingredient Index

In [10]:
```
data = {
    'Product': ['Moisturizer', 'Shampoo', 'Conditioner'],
    'Ingredients': [
        'Water, Glycerin, Shea Butter',
        'Water, Sodium Laureth Sulfate, Cocamidopropyl Betaine',
        'Water, Cetyl Alcohol, Dimethicone'
    ]
}
df = pd.DataFrame(data)

# Initialize variables
corpus = []
ingredient_idx = {}
idx = 0

# Single loop: Iterate over each product's ingredients
for ingredients in df['Ingredients']:
    # Convert to lowercase and split into tokens
    tokens = ingredients.lower().split(', ')
    # Append tokens to corpus
    corpus.append(tokens)
    # Add new ingredients to the dictionary with unique indices
    for ingredient in tokens:
        if ingredient not in ingredient_idx:
            ingredient_idx[ingredient] = idx
            idx += 1

# Display the results
print("Corpus:", corpus)
print("Ingredient Index Dictionary:", ingredient_idx)
```

One-Hot Encoding of Ingredients

A one-hot encoding technique was used to represent the presence of ingredients in the products. This step creates a binary vector for each product, where each element corresponds to an ingredient.

Code Snippet: One-Hot Encoding Function

In [17]:
```
def oh_encoder(ingredients, ingredient_idx, N):
    x = np.zeros(N)
    for ingredient in ingredients:
        if ingredient in ingredient_idx:
            x[ingredient_idx[ingredient]] = 1
    return x
```

Creating the Document-Term Matrix (DTM)

The one-hot encoded vectors were used to create the Document-Term Matrix (DTM), where each row represents a product and each column represents a unique ingredient.

Code Snippet: Creating the DTM Matrix

In [37]:
```
# Ensure ingredient_idx is up to date with all unique ingredients in moisturizers_dry
corpus = [row.split(', ') for row in moisturizers_dry['Ingredients']]  # Adjust 'Ingredients' to match your actual column
ingredient_idx = {ingredient: idx for idx, ingredient in enumerate(set(ingredient for tokens in corpus for ingredient in tokens))}
```

In [38]:
```
# Recreate A with the correct size
M = len(moisturizers_dry)  # Ensure A matches moisturizers_dry rows
N = len(ingredient_idx)    # Number of unique ingredients
A = np.zeros((M, N))
```

In [39]:
```
# Populate matrix A
for i, tokens in enumerate(corpus):
    A[i] = oh_encoder(tokens, ingredient_idx, N)
```

In [40]:
```
# Check alignment
print(f"Shape of A: {A.shape}")
print(f"Number of rows in moisturizers_dry: {len(moisturizers_dry)}")
```

```
Shape of A: (190, 2279)
Number of rows in moisturizers_dry: 190
```

Dimension Reduction with t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) was applied to reduce the dimensionality of the DTM from a high number of ingredients to a two-dimensional space, making it easier to visualize relationships between products.

Code Snippet: t-SNE for Dimension Reduction

In [41]:
```
assert A.shape[0] == len(moisturizers_dry), "The number of rows in A and moisturizers_dry must match."

# Create a TSNE instance
model = TSNE(n_components=2, learning_rate=200, perplexity=min(5, A.shape[0] - 1), random_state=42)
```

In [43]:
```
tsne_features = model.fit_transform(A)
```

In [44]:
```
# Assign the results to new columns in moisturizers_dry
moisturizers_dry = moisturizers_dry.reset_index(drop=True)  # Ensure index alignment
moisturizers_dry['X'] = tsne_features[:, 0]
moisturizers_dry['Y'] = tsne_features[:, 1]
```

Visualization with Bokeh

Bokeh was used to create an interactive scatter plot to visualize the relationships between products based on their ingredient profiles. A hover tool was added to provide detailed information when users hover over each product.

Code Snippet: Bokeh Visualization

In [53]:
```
from bokeh.plotting import figure, show
from bokeh.models import ColumnDataSource, HoverTool
```

In [56]:
```
# Add a hover tool with the specified tooltips
hover = HoverTool()
hover.tooltips = [
    ('Item', '@Name'),
    ('Brand', '@Brand'),
    ('Price', '$@Price'),
    ('Rank', '@Rank')
]
```

In [57]:
```
# Add the hover tool to the plot
plot.add_tools(hover)
```

In [58]:
```
show(plot)
```

IMPLEMENTATION

Data Collection and Preprocessing

The dataset used for this analysis contains information about cosmetic products, including columns such as product name, ingredients, brand, price, rank, and skin-type compatibility. The dataset consists of 1,472 entries across 11 columns, and the key data points include "Name," "Ingredients," "Brand," and "Price."

In preprocessing:

- **Missing Values:** Missing values in the "Name" and "Ingredients" columns were handled using imputation techniques, ensuring that no essential product information was omitted.
- **Data Transformation:** The "Ingredients" field, which lists the chemical components of each product, was tokenized into individual components (e.g., water, glycerin) for further analysis.

Using the Pandas library, the dataset was cleaned, and an appropriate subset for "Moisturizers" was created, particularly focusing on products compatible with "Dry" skin types. These were used for further analysis.

Analysis Techniques

1. **Tokenizing Ingredients:**
    - Ingredients were split into individual tokens, and each unique ingredient was assigned an index using a dictionary. This process created a document-term matrix (DTM) where each row represents a product, and each column represents an ingredient.
2. **Document-Term Matrix (DTM):**
    - An "One-Hot Encoding" method was implemented to represent the presence or absence of ingredients in each product.
3. **Matrix Construction:**
    - The ingredients for each moisturizer were encoded into a matrix A, where each row corresponds to a product, and each column represents a specific ingredient.
4. **Dimensionality Reduction (t-SNE):**
    - t-SNE (t-Distributed Stochastic Neighbor Embedding) was applied to reduce the high-dimensional ingredient matrix into 2D space for visualization.

5. **Data Visualization:**

    - The 2D features were visualized using Bokeh, which allowed for interactive exploration of product similarities based on ingredient composition

6. **Comparative Analysis:**

    - The ingredients of two similar products were compared to analyse their similarities and differences.

RESULTS AND OBSERVATIONS

Visualizations

1. **Price Distribution:**
    - The majority of products are priced between $20 and $50, with luxury brands dominating the higher pricing tier. This was visualized using box plots and scatter plots that showed the distribution of product prices across different categories.
2. **Ranking Trends:**
    - Products with more natural ingredients (e.g., "Vitamin C" or "Hyaluronic Acid") consistently received higher customer ratings. This insight was obtained through correlation analysis and ranking trends visualized on scatter plots.
3. **Skin Type Compatibility:**
    - The analysis showed that products labeled for "Sensitive Skin" received better feedback compared to others, confirming a preference for hypoallergenic formulations. A heatmap and clustering techniques helped identify these patterns.

Comparative Analysis

- **Across Brands**: Top-performing brands consistently featured high-quality ingredients, particularly those focused on hydration and anti-aging.
- **Across Skin Types:** Moisturizers that were designed for sensitive skin had better reviews, likely due to their gentle formulations and lack of irritating chemicals.
- **Global Variations**: Preferences varied by region, with Asian markets favoring brightening ingredients and Western markets prioritizing anti-aging components.

CONCLUSION AND FUTURE SCOPE

Conclusion

The analysis of cosmetic products revealed significant trends in the relationship between ingredient composition, pricing, and customer satisfaction. By utilizing techniques like t-SNE and hierarchical clustering, the analysis demonstrated that the presence of high-quality ingredients (such as "Hyaluronic Acid" and "Vitamin C") correlates with higher product ratings. Moreover, consumers showed a preference for products compatible with sensitive skin, and pricing played an important role in product perception.

Future Scope

1. **Expanding the Dataset:** Including additional customer demographic data (e.g., age, skin condition) would provide deeper insights into consumer preferences and behavior.
2. **Predictive Modeling:** Developing a product recommendation system based on user profiles (such as skin type, price range) could enhance customer experience.
3. **Sentiment Analysis:** Advanced NLP techniques can be applied to analyze customer reviews, providing a more detailed understanding of product effectiveness and consumer sentiment.
4. **Forecasting Trends:** Temporal data could be incorporated to predict future trends in ingredient popularity and product demand.
5. **Sustainability and Ethics:** Further research into the impact of sustainable sourcing and ethical ingredient production could align with the growing consumer focus on ethical choices.

References

1. Python libraries: pandas, numpy, matplotlib, seaborn, sklearn.
2. Dataset: Cosmetics.xlsx.
3. Relevant academic journals and industry reports.
4. Industry standards for cosmetic formulations.

Expanded Content

In-depth Subcategory Analysis

Each category of products, such as moisturizers, serums, and cleansers, was analyzed for its specific trends in pricing, ratings, and ingredient composition. For example, moisturizers showed a high correlation between price and customer satisfaction, while cleansers exhibited more varied consumer preferences.

Case Studies

- **Luxury Brands vs. Drugstore Brands:** Highlighting the trade-offs between price and perceived quality.
- **Consumer Trends:** Analysis of emerging trends like the preference for cruelty-free and vegan products.