**ACKNOWLEDGEMENT**

The traineeship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies of the subject of 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 also for spearing his valuable time in spite of his busy schedule.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| S. No. | Topic | Page No |
|  | Introduction |  |
|  | About The Company |  |
|  | About the Project |  |
|  | Objectives and Deliverables |  |
|  | Methodology |  |
|  | Flow Of Project |  |
|  | Languages, Libraries and Platform Used |  |
|  | Implementation |  |
|  | Dataset and Libraries Importing |  |
|  | Data Exploration and Cleaning |  |
|  | t-SNE Model Implementation |  |
|  | Power BI Dashboard |  |
|  | Sample Screenshots and Observations |  |
|  | Conclusion and Future Scope |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**ABSTRACT**

Selecting new cosmetic products can be a daunting task, especially for individuals with sensitive skin or those prone to skin issues. The ingredient lists on cosmetic packaging often contain complex chemical names that are difficult to interpret without a background in chemistry. This complexity creates a barrier to making informed purchasing decisions.

To address this challenge, this project proposes the development of a content-based recommendation system for cosmetics, leveraging data science techniques. The system will process ingredient lists from 1,472 cosmetic products available on Sephora. By employing word embedding techniques, the chemical components of these products will be transformed into a numerical format suitable for analysis.

The project will then utilize t-SNE (t-Distributed Stochastic Neighbour Embedding), a machine learning method, to visualize the similarity between ingredients. This visualization will be enhanced using Bokeh, an interactive visualization library, allowing users to explore the relationships between different cosmetic products based on their ingredient lists.

The goal of this project is to provide a data-driven approach to cosmetic selection, helping users identify products that are more likely to suit their individual skin needs. By leveraging advanced data processing and visualization techniques, this recommendation system aims to simplify the decision-making process and reduce the risk of adverse reactions to new cosmetic products.

**INTRODUCTION**

* 1. **About the Company**

MedTourEasy, a global healthcare company, provides you the informational resources needed to

evaluate your global options. It helps you find the right healthcare solution based on specific health needs, affordable care while meeting the quality standards that you expect to have in healthcare.

MedTourEasy improves access to healthcare for people everywhere. It is an easy to use platform

and service that helps patients to get medical second opinions and to schedule affordable, high

quality medical treatment abroad.

* 1. **About the Project**

Buying new cosmetic products is difficult. The ingredient lists on product labels contain crucial information, but interpreting these lists can be challenging. Without a background in chemistry, understanding the implications of various ingredients can be complex and confusing. 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, you are going to create a content-based recommendation system tat could help the individuals in buying better products, which are low in cost and better for the skin based on their skin type.

* 1. **Objective and Deliverables**

Working on the project, I aim to create a recommendation system for the customers, where all the suited ingredients for individual skin type would be recommended using some techniques:

1. So, the first aim would be using an Machine Learning Model like t-SNE that would help in building the recommendation system for the individuals based on their skin types.
2. Second would be a better visualization for customers that would help them understand why certain products are recommended to them using Power BI Dashboard.

For the recommendation system part, we will process ingredient lists for 1472 cosmetics on Sephora via word embedding, then visualize ingredient similarity using a machine learning method called t-SNE.

While for the visualization part, it would be done on Power BI, due to certain advantages like, Its user-friendly interface allows for easy creation of interactive reports and dashboards, even for users without extensive technical expertise. Power BI integrates seamlessly with a wide range of data sources, including cloud-based and on-premises systems, enabling users to consolidate and analyze data from various platforms.

**METHODOLOGY**

* 1. **Flow Of Project**
  2. **Languages, Libraries and Platforms Used**
     1. Language: Python

Python is a versatile and widely-used programming language known for its simplicity and readability. It supports multiple programming paradigms, including procedural, object-oriented, and functional programming, making it suitable for a broad range of applications. Python's extensive standard library and robust community support contribute to its popularity among developers for web development, data analysis, machine learning, automation, and more.

i. Python's clear and straightforward syntax enhances code readability and maintainability.

ii. Supports various programming paradigms, including procedural, object-oriented, and functional programming.

iii. Provides a rich set of libraries and frameworks for tasks ranging from web development (e.g., Django, Flask) to data analysis (e.g., Pandas, NumPy).

iv. Runs on various operating systems, including Windows, macOS, and Linux.

v. A large and active community offers extensive documentation, tutorials, and third-party tools.

vi. Easily integrates with other languages and technologies, making it suitable for diverse development environments.

This combination of features makes Python a powerful and flexible tool for both novice and experienced developers.

* + 1. Library: Scikit-learn

Scikit-learn is a popular and widely-used machine learning library in Python, designed for easy implementation of a variety of machine learning algorithms. It provides simple and efficient tools for data mining and data analysis, making it accessible to both beginners and experienced practitioners. Scikit-learn is built on top of other foundational Python libraries such as NumPy, SciPy, and matplotlib, and it supports various supervised and unsupervised learning techniques.

i. Includes a comprehensive collection of machine learning algorithms for classification, regression, clustering, and dimensionality reduction.

ii. Provides a consistent and easy-to-use interface for implementing and evaluating machine learning models.

iii. Works seamlessly with other scientific computing libraries such as NumPy, SciPy, and pandas, allowing for efficient data manipulation and analysis.

iv. tools for model selection, hyperparameter tuning, and performance evaluation, facilitating the development of robust models.

v. Offers various data preprocessing techniques for scaling, encoding, and transforming data to improve model performance.

vi. Well-documented with extensive tutorials and supported by an active community, which helps in troubleshooting and learning.

These features make scikit-learn a valuable library for developing and deploying machine learning models with ease and efficiency.

* + 1. Platform: Power BI

Power BI is a powerful business analytics tool from Microsoft designed to help users visualize and share insights from their data. It offers a range of features that simplify data analysis and reporting, making it accessible to users with varying levels of technical expertise. Power BI enables the creation of interactive and visually appealing reports and dashboards that facilitate data-driven decision-making.

i. Allows users to create dynamic and interactive dashboards that provide real-time insights into their data.

ii. Seamlessly connects to a wide variety of data sources, including cloud-based services, databases, and Excel spreadsheets.

iii. Offers a wide range of built-in visualizations and supports custom visuals to tailor reports to specific needs.

iv. Provides robust data modeling capabilities for creating complex calculations, relationships, and aggregations.

v. Supports real-time data updates and refreshes, ensuring that users always have access to the latest information.

vi. Facilitates easy sharing and collaboration through Power BI Service, allowing teams to access and work with reports from anywhere.

vii. Integrates well with other Microsoft products such as Excel, Azure, and Office 365, enhancing productivity and data accessibility.

These features make Power BI a versatile and user-friendly tool for transforming raw data into actionable insights and driving informed business decisions.

**IMPLEMENTATION**

* 1. **Dataset and Libraries Importing**

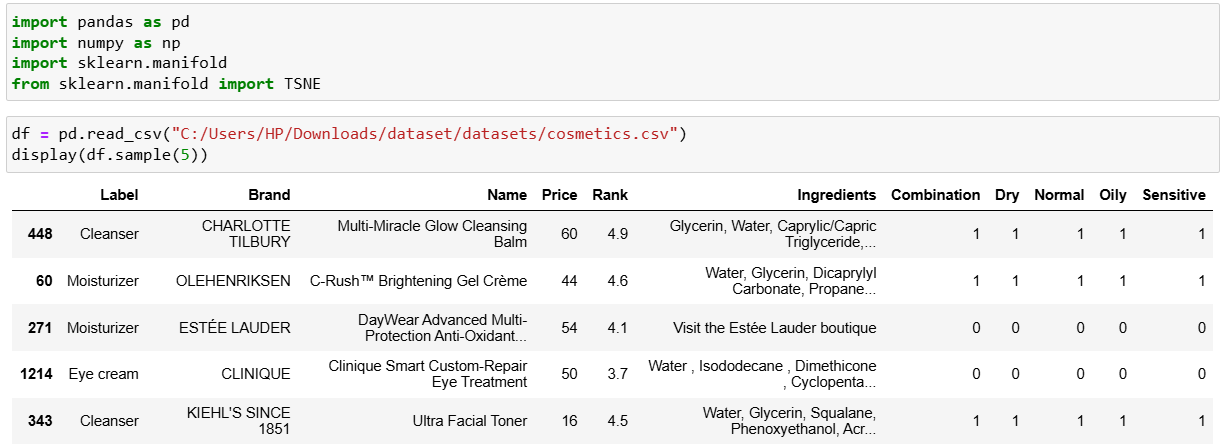
The dataset has been provided through the MedTourEasy Company, as a part of the project through a google drive link:

<https://drive.google.com/uc?export=download&id=1-OBlj1fTfFwciscLKIZIvWbIeHwBUzif>

Data importing is referred to as uploading the required data into the coding environment from internal sources (computer) or external sources (online websites and data repositories). This data can then be manipulated, aggregated, filtered according to the requirements and needs of the project.

We used the Python Pandas package for importing the dataset using the read\_csv() function in the Pandas Library.

Sample Code:



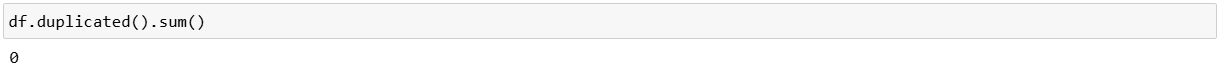
* 1. **Data Exploration and Cleaning**

Data cleaning is essential in data analytics as it ensures the accuracy, consistency, and reliability of data before analysis. By identifying and correcting issues within the dataset, data cleaning enhances its quality and usability. It ensures that analysis produces reliable results, maintains consistency by standardizing formats, improves efficiency by eliminating redundant information, and supports better decision-making. Additionally, clean data helps meet regulatory standards and data governance policies, making it crucial for compliance and overall data integrity.

There are certain steps involved in data cleaning:

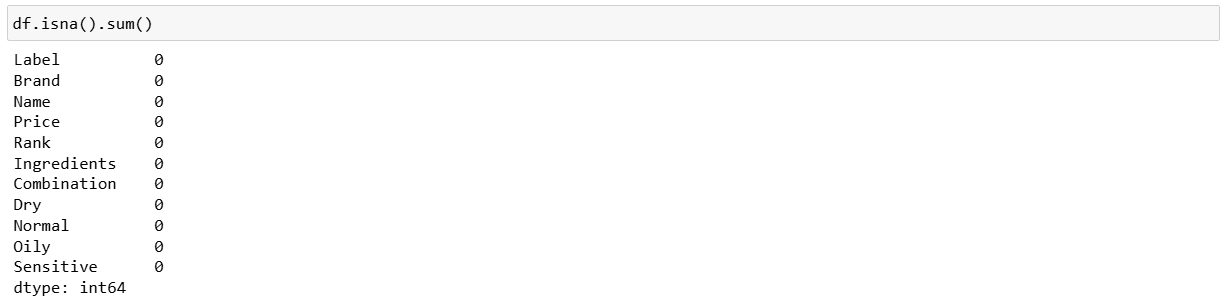
1. **Remove Duplicates**: Identify and remove duplicate records to avoid redundancy and ensure each data entry is unique.

Sample Code:



1. **Handle Missing Values**: Address missing or null values by:
   * Imputing them with statistical measures (mean, median, mode).
   * Using algorithms to predict missing values.
   * Removing records with missing values if appropriate.

Sample Code:



1. **Correct Errors**: Identify and correct data entry errors, such as typos, incorrect values, and out-of-range data, to ensure data accuracy.
2. **Standardize Data**: Convert data to a consistent format (e.g., dates, units) and standardize categorical values to ensure uniformity across the dataset.
3. **Normalize Data**: Adjust data to a common scale or range, especially for numerical data, to ensure fair comparison and analysis.
4. **Filter Outliers**: Identify and address outliers that may skew the analysis, either by investigating their cause or by removing them if they are deemed erroneous.

While there are so many steps involved in data cleaning of data analytics, we have only used two of them. One is, clecking for duplicate values and the other is checking for missing values where we found that there are no missing values so we didn’t need to replace it with any mean, median or mode values for the project.

After cleaning the dataset, since we need to work with the dataset, it’s a good practise to explore the dataset about it’s values which would further help in the project implementation.

1. After certain exploration there are certain inputs we get about the dataset, it has a total of 1472 rows and 11 columns.
2. The columns are labelled as - ‘Label', 'Brand', 'Name', 'Price', 'Rank', 'Ingredients', 'Combination', 'Dry', 'Normal', 'Oily', 'Sensitive'.
3. The label column talks about the type of product used by the customer, which involves 6 unique categories - 'Moisturizer', 'Cleanser', 'Treatment', 'Face Mask', 'Eye cream' and 'Sun protect'.
4. The Brand talks about the Company Brand that the product belongs to which are a total of 116 brands.
5. The Name column talks about the name of the product which belongs to a certain company and a certain type of product. Many products can belong to the same brand.
6. The Price and Rank column talk about the price of the product and their standing in the global market as the rank.
7. The ingredients columns talks about all the ingredients used in the product making as a array of strings.
8. While the last 5 columns which involve - 'Combination', 'Dry', 'Normal', 'Oily' and 'Sensitive', all these column contain values in 0s and 1s as in True or False, indicating if the person has a dry or normal or sensitive or a oily skin, while if the customer has a combination of two or three, the required columns are set as 1s along with the combination column set as 1 also.
   1. **t-SNE Implementation**

t-SNE (t-Distributed Stochastic Neighbour Embedding) is a dimensionality reduction technique used to visualize high-dimensional data in a lower-dimensional space, typically two or three dimensions. It works by preserving the local structure of the data, maintaining the similarities between nearby points while mapping distant points apart. One of its key advantages is its ability to reveal complex patterns and clusters in the data that may not be apparent in higher dimensions. t-SNE is particularly effective for exploring and visualizing large datasets, making it a popular choice in machine learning and data analysis for understanding the underlying structure and relationships within the data.

Since we need to implement a recommendation system that has many parameters to consider and t-SNE Model helps in reducing these parameters into a lower dimensional space, it would a good choice to use this for our project that would recommend the customers certain products based on the best ranking and prices while also considering the chemicals used in the product which is the ultimate goal of the project.

1. There are six categories of product in our data (moisturizers, cleansers, face masks, eye creams and sun protection) and there are five different skin types (combination, dry, normal, oily and sensitive). Because individuals have different product needs as well as different skin types, let's set up our workflow so its outputs (a t-SNE model and a visualization of that model) can be customized. For the example, let's focus in on moisturizers for those with dry skin by filtering the data accordingly.

Sample Code:



1. To get to our end goal of comparing ingredients in each product, we first need to do some preprocessing tasks and bookkeeping of the actual words in each product's ingredients list. The first step will be tokenizing the list of ingredients in Ingredients column. After splitting them into tokens, we'll make a binary bag of words. Then we will create a dictionary with the tokens, ingredient\_idx, which will have the following format:{"ingredient" index value, … }.

Sample Code:



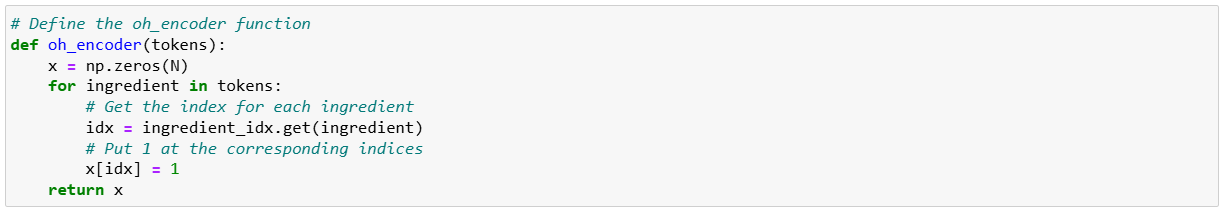
1. The next step is making a document-term matrix (DTM). Here each cosmetic product will correspond to a document, and each chemical composition will correspond to a term. This means we can think of the matrix as a <em>“cosmetic-ingredient”</em> matrix. To create this matrix, we'll first make an empty matrix filled with zeros. The length of the matrix is the total number of cosmetic products in the data. The width of the matrix is the total number of ingredients.

Sample Code:



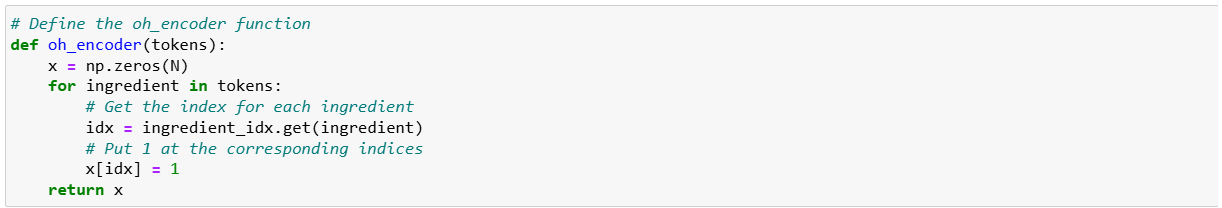
1. Before we can fill the matrix, let's create a function to count the tokens (i.e., an ingredients list) for each row. Our end goal is to fill the matrix with 1 or 0: if an ingredient is in a cosmetic, the value is 1. If not, it remains 0. The name of this function, oh\_encoder will become clear next.

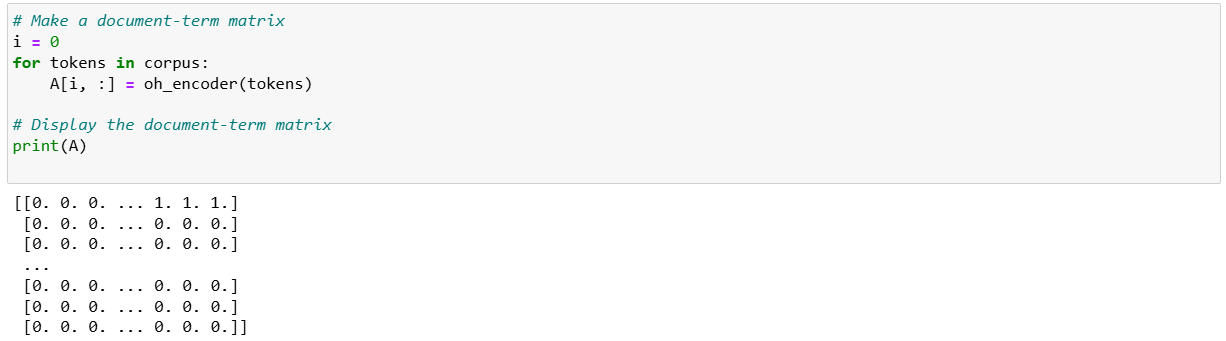
Sample Code:

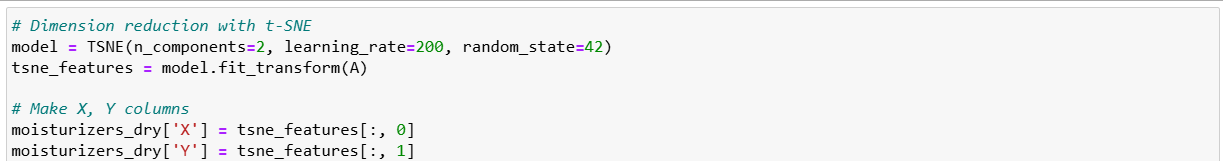


1. Now we'll apply the oh\_encoder() functon to the tokens in corpus and set the values at each row of this matrix. So the result will tell us what ingredients each item is composed of. For example, if a cosmetic item contains water, niacin, decyl aleate and sh-polypeptide-1, the outcome of this item will be as follows. This is what we called one-hot encoding. By encoding each ingredient in the items, the Cosmetic-Ingredient matrix will be filled with binary values.

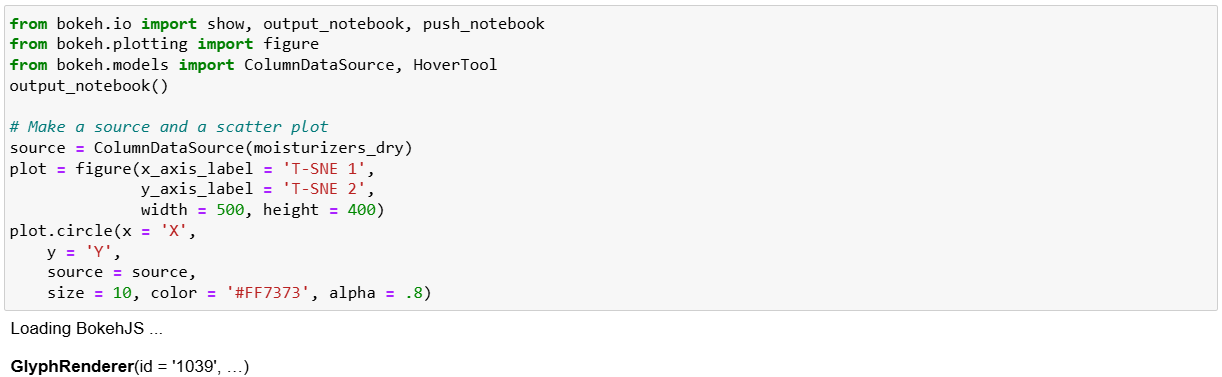
Sample Code:





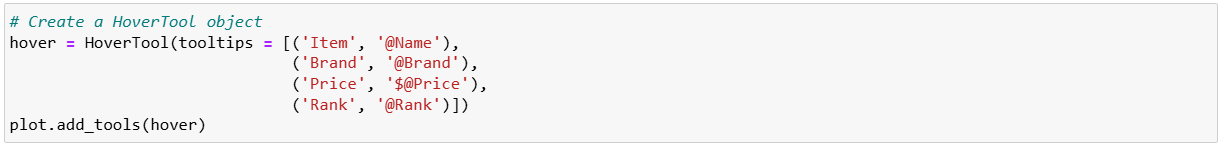


1. We are now ready to start creating our plot. With the t-SNE values, we can plot all our items on the coordinate plane. And the coolest part here is that it will also show us the name, the brand, the price and the rank of each item. Let's make a scatter plot using Bokeh and add a hover tool to show that information. Note that we won't display the plot yet as we will make some more additions to it.

Sample Code:  


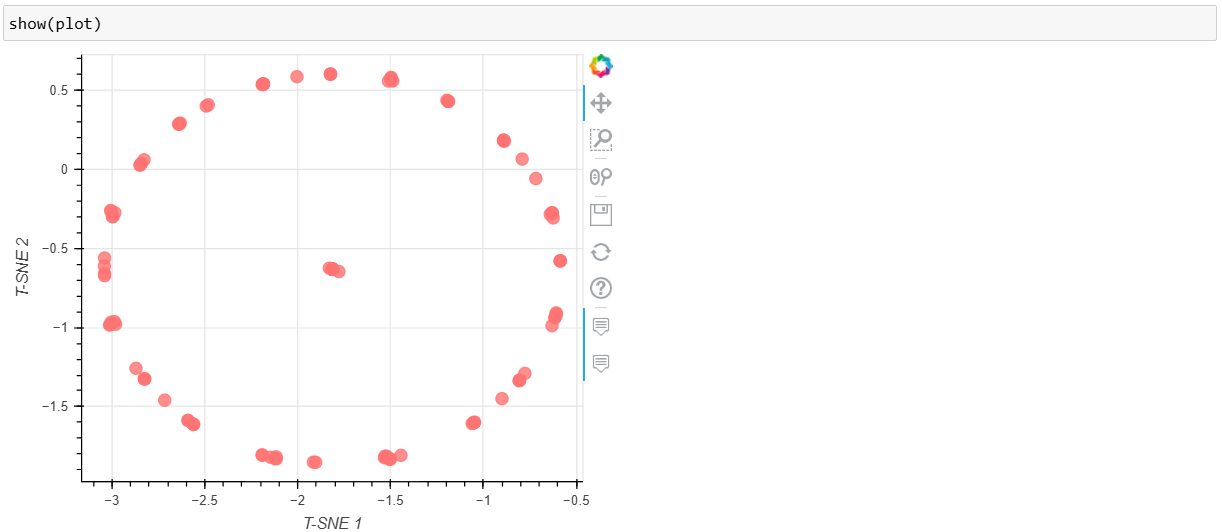
1. Why don't we add a hover tool? Adding a hover tool allows us to check the information of each item whenever the cursor is directly over a glyph. We'll add tooltips with each product's name, brand, price, and rank (i.e., rating).

Sample Code:



1. Finally, it's show time! Let's see how the map we've made looks like. Each point on the plot corresponds to the cosmetic items. Then what do the axes mean here? The axes of a t-SNE plot aren't easily interpretable in terms of the original data. Like mentioned above, t-SNE is a visualizing technique to plot high-dimensional data in a low-dimensional space. Therefore, it's not desirable to interpret a t-SNE plot quantitatively. Instead, what we can get from this map is the distance between the points (which items are close and which are far apart). The closer the distance between the two items is, the more similar the composition they have. Therefore this enables us to compare the items without having any chemistry background.

Sample Code:



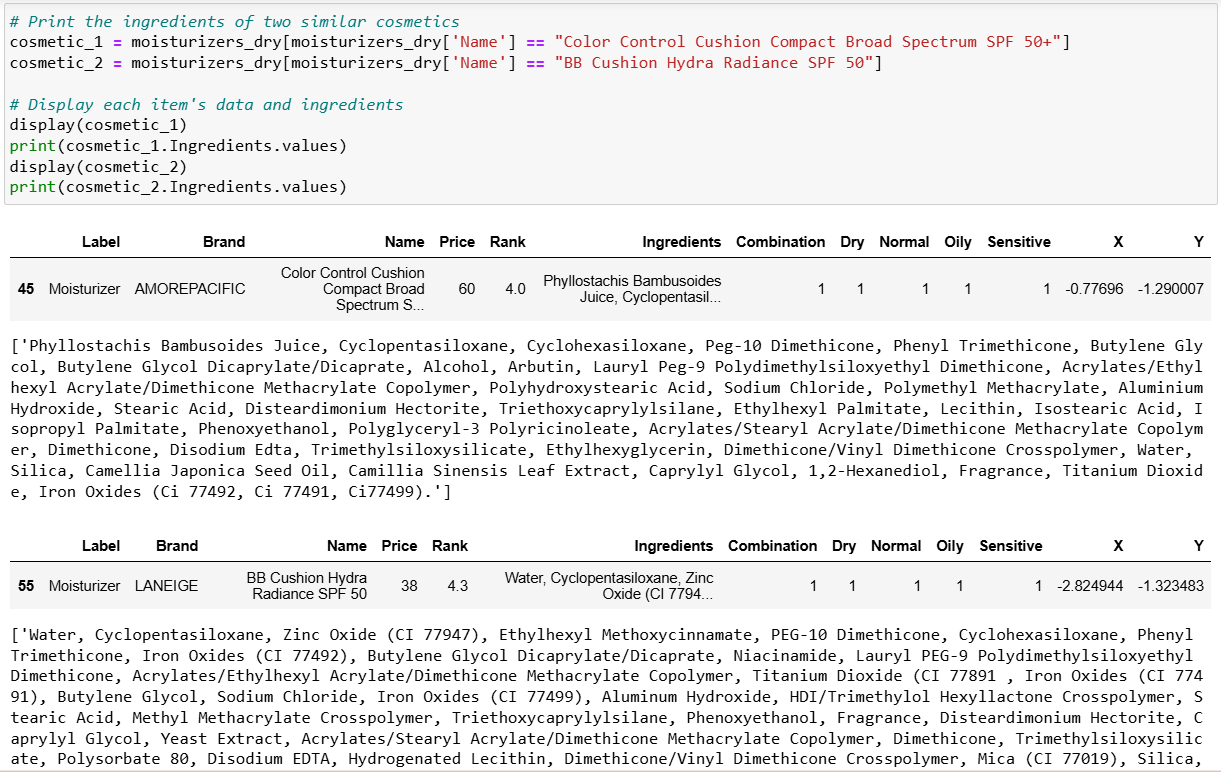
After we have created our recommendation system suppose we need to test it, Say we enjoyed a specific product, there's an increased chance we'd enjoy another product that is similar in chemical composition. Say we enjoyed AmorePacific's.

Color Control Cushion Compact Broad Spectrum SPF 50+</a>. We could find this product on the plot and see if a similar product(s) exist. And it turns out it does! If we look at the points furthest left on the plot, we see LANEIGE's.

BB Cushion Hydra Radiance SPF 50 essentially overlaps with the AmorePacific product. By looking at the ingredients, we can visually confirm the compositions of the products are similar (though it is difficult to do, which is why we did this analysis in the first place!), plus LANEIGE's version is $22 cheaper and actually has higher ratings.

It's not perfect, but it's useful. In real life, we can actually use our little ingredient-based recommendation engine help us make educated cosmetic purchase choices.

Sample Code:



* 1. **Power BI Dashboard**

Power BI is a robust business analytics tool developed by Microsoft designed to help users turn data into actionable insights. It offers a comprehensive platform for visualizing and sharing data through interactive reports and dashboards. Power BI enables users to connect to a wide range of data sources, including databases, cloud services, and spreadsheets, integrating them into a unified view. This integration allows for in-depth analysis and the generation of meaningful business intelligence. With its intuitive interface and powerful features, Power BI facilitates data exploration, enabling users to uncover trends, monitor key metrics, and make informed decisions based on real-time information. It is widely used for its ability to simplify complex data and provide clear, actionable insights for effective business planning and strategy.

Some key advantages of Power BI in the field of data analytics is:

i. User-Friendly Interface: Simplifies the creation of interactive and visually appealing reports.

ii. Robust Data Integration: Seamlessly connects to numerous data sources for comprehensive analysis.

iii. Real-Time Data Refresh: Keeps insights up-to-date with the latest information.

iv. Advanced Data Modeling: Supports complex data modeling for in-depth analysis.

v. Extensive Customization Options: Offers various customization features to tailor reports and dashboards to specific needs.

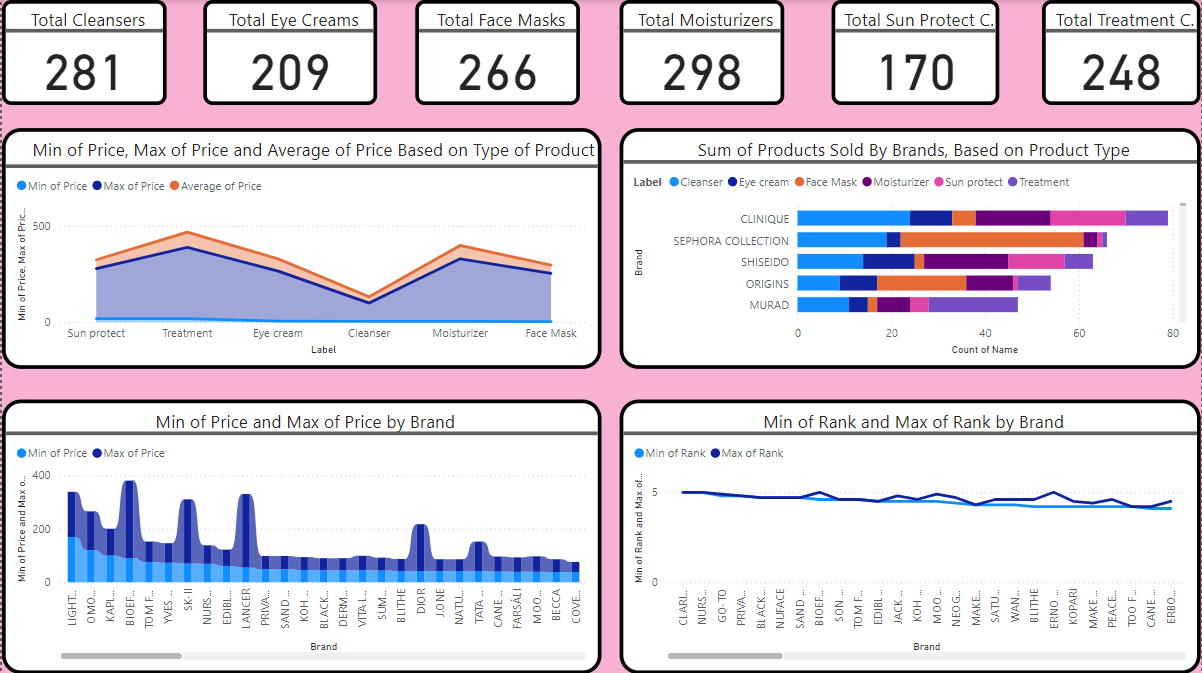
vi. Powerful Decision-Making Tool: Facilitates data-driven decision-making and strategic planning through actionable insights.

Since we are aiming to ease the conflict of the customers towards buying a suitable product. We could do so through data visualisation via Dashboards made using Power BI. Because the dataset that any specific E-Commerce Company or any other Industrial Companies have is large. Gaining valuable insights from that data can be a bit hard to interpret for the company’s growth which is why the go for creating dashboards that help in gaining these insights through interactive visualisations in Power BI. Similarly coming to our problem statement, gaining trust over any recommendation system could be hard for them at first that’s where these dashboards can have a positive impact, by helping them understand the patterns in the data and hence interpret why a certain product is recommended to him/her.

For this project we plan to make a Power BI Dashboard having certain crucial information in the form of Data Cards or Graphs like:

1. First having a count of the total number of distinct product categories that the company has in the form of cosmetics.
2. A Stacked Area Chart representing the Maximum, Minimum and Average of Price based on the Type of Product.
3. A Stacked Bar Chart representing the Sum of Products sold by Brands based on the Product Type.
4. A Ribbon Chart representing the Minimum and Maximum Of Price by Brand.
5. A Line Chart representing Minimum and Maximum of Rank based on Brand.

Screenshot for the Power BI Dashboard is as follows:



**SAMPLE SCREENSHOTS AND OBSERVATIONS**

**4.1 Heading and Quick Facts**

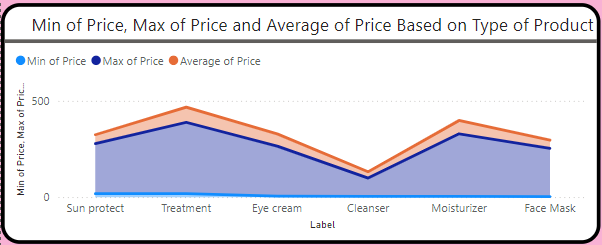
This the topmost bar of the dashboard which shows the certain main facts about the cosmetics. There are 6 facts displayed as follows: Total Cleansers, Total Eye Creams, Total Face Masks, Total Moisturizers, Total Sun Protect Creams, Total Treatment Creams.



OBSERVATION – There are a total of 281 Cleansers, 209 Eye Creams, 266 Face Masks, 298 Moisturizers, 170 Sun Protect Creams, 248 Treatment Creams in the company.

**4.2 Price Based On Type Of Product**

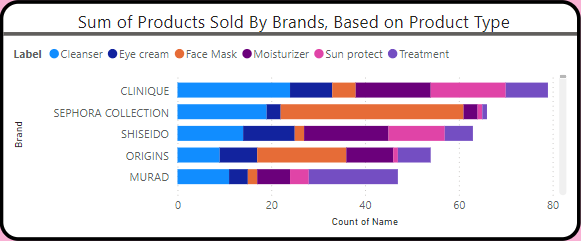
The first visual of the dashboard is a Stacked Area Chart representing the Minimum, Maximum and Average Price based on Type of Product. This visual is user interactive, since if the user clicks on any of the product type, the whole dashboard will display the visualization based on that Product Type.



OBSERVATION – The line chart shows that there is a very high difference between the minimum and maximum price of any Product Type. Also, according to the visual, the two Product Types – Treatment Creams and the Moisturizers arehaving the highest of Prices while a Cleaner has the lowest of costs.

**4.3 Sum Of Products Sold By Brands Based On Product Type**

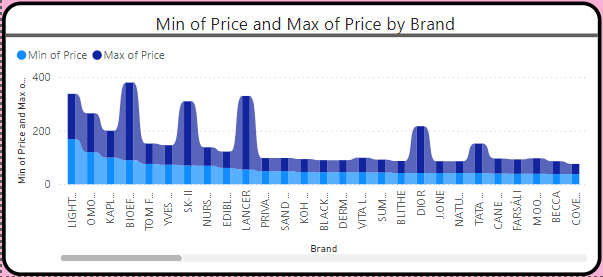
The second visual of the dashboard is a Stacked Bar Chart representing the total of the individual Product Types sold by a specific Brand. This visual is also user interactive since if the user clicks on any Product Type, only the that product type’s total product sold will be shown in each company or brand.



OBERVATION – From the top 5 distinct brands that are visible in the picture, there are certain different ratio of products sold in each company, now maybe that could be because that company has a higher edge over that product type compared to other brands, say, there is a visually high looking ratio for the Face Masks sold by the Sephora Collection Brand.

**4.4 Minimum and Maximum Price Of a Product Based on Brand**

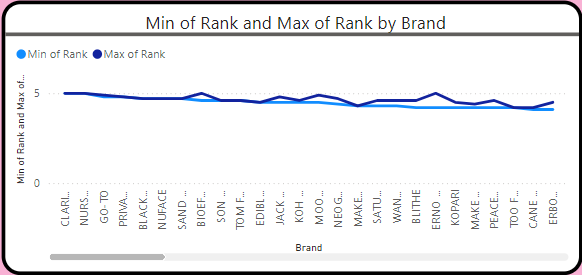
This is the third visual of the dashboard representing a Ribbon Chart that shows the minimum and maximum price of any product sold by a specific brand name. This visual is also user interactive since if the user clicks on any Brand, only the that brand’s information i.e minimum and maximum price of any product would be shown.



OBSERVATION - From the 26 brands that are visible in the above figure, there are not very high difference in the minimum and maximum price of product for most brands. But some have a significant difference in prices, say, from the figure, BIOEFFECT, SK II, LANCER, DIOR and to some extent TATA HARPER.

**4.5 Minimum and Maximum Rank By Brand**

This is the fourth visual of the dashboard representing a Double Line Chart that shows the minimum and maximum rank given to a specific brand. This visual is also user interactive since if the user clicks on any Brand, only the that brand’s information i.e minimum and maximum rank of that brand would be shown.



OBSERVATION - From the 26 brands that are visible in the above figure, there are not very high difference in the minimum and maximum rank of specific brands. But some have a a little significant difference in rank, say, from the figure, BIOEFFECT, JACK BLACK, MOON JUICE, ERNO LASZLO and PEACE OUT. Even though the difference is not very big still the customer has better options to go for in the global market based on ranking. This can imply that the specific brand has certain category of product with high ranking while some other category product having low ranking, indicating the brand is only good for a certain category of product.

**CONCLUSION AND FUTURE SCOPE**

This project aimed at analysing the current situation of cosmetics sold in the global market by creating intuitive and user interactive dashboards and drawing conclusions on the reasons for certain parameters (say, prices and rank) to have certain specific values. Currently the project is in it’s last stage of development with the dashboards been developed and submitted for review and feedback.

With regards to the future work, the firm aims at regularly updating the dashboards with the time and integrating it with their systems so as to continually draw conclusions and analyse the results. This will enable them to predict future business opportunities and provide a basis on which they can plan on increasing their market presence and capacity planning.