Analysis of Sephora Skin Care Product Popularity and the Ingredients Within.

Vrena Ranallo

Western Governors University

TABLE OF CONTENTS

Project Overview ***3***

A. Summarization of the Elements of Project **3**

Project Plan ***3***

B. Summarization of Project Execution **3**

Methodology ***4***

C. Data Selection, Collection and Limitations **4**

D. Data Extraction and Preparation **5**

E. Data Analysis Process Report **6**

1. Description of Analytical Methods 6
2. Advantages and Limitations of Analytical Methods 7
3. Application of Analytical Methods 7

Results ***12***

F. Evaluate Success of Analysis **12**

1. Statistical Significance 12
2. Practical Significance 25
3. Overall Success 29

G. Key Takeaways **29**

1. Summarization of Conclusions 29
2. Explanation of Tools and Graphical Representations 29
3. Recommendations Based on Findings 30

H. Panopto Presentation **31**

Appendices ***31***

I. Evidence of Project Completion **31**

J. Sources **31**

K. Ingredient Glossary **32**

PROJECT OVERVIEW

A. SUMMARIZATION OF THE ELEMENTS OF PROJECT

Skin care is a quickly growing market where the customers are very knowledgeable about ingredients and have incredibly high expectation for quality. In this project, we will explore methods of constructing an informed strategy to expand Makeup+ into the skin care segment by researching a relationship between the popularity of a skin care product and the ingredients used.

The scope of this project is to use insights gathered from competitors' products to create a strategy for two potential skincare lines. Both recommended lines will include 3 products, with target price ranges, and beneficial ingredients for each formulation. This will be done by using statistical regression analysis on products currently on Sephora's website. The strategies will be delivered at the end of the report, along with a video presentation, and a one-page summary report. Each of these deliverables are meant to advise Makeup+ stakeholders and allow them to make a data-driven decision for their product launches.

This is a strictly informative project to create a product strategy. It will not include any product development.

PROJECT PLAN

B. SUMMARIZATION OF PROJECT EXECUTION

The goal of this project is to create a data-driven strategy for Makeup+ to confidently expand its portfolio into skincare. This was done by analyzing the different types of skin care products and how they rank with customers. Determine prominent and beneficial ingredients for each skin care type. Then perform a regression test to evaluate any potential linear relationship between the top beneficial ingredients and the customer rank. The results of this analysis will be presented in the Results section of this report, a 1-page summary with my recommendations, and a video presentation.

This project will be considered a success when the analysis of the Sephora data has been performed, and the results have been correctly interpreted into a strategy and communicated in the three deliverables mentioned above. The project's success is not predicated on the existence of a correlation between the rank and the top ingredients.

We used agile project management to plan and execute this project. It was important to use a method that could evolve based on the insights gained and allowed for a flexible iterative approach. The project milestones are listed below with the projected vs actual start and end dates. As you can see the project started two days earlier than projected and was completed 4 days before the projected end date.

Milestone	Projected Start	Actual Start	Projected End	Actual End
Define client requirements	1/16/2023	1/14/2023	1/17/2023	1/14/2023
Define system requirements	1/16/2023	1/14/2023	1/17/2023	1/14/2023
Design specs based on requirements	1/16/2023	1/14/2023	1/17/2023	1/14/2023
Gather Data	1/17/2023	1/14/2023	1/18/2023	1/14/2023
Setup Workspace	1/17/2023	1/14/2023	1/18/2023	1/14/2023
EDA	1/17/2023	1/15/2023	1/19/2023	1/15/2023
Data Cleaning	1/17/2023	1/15/2023	1/19/2023	1/15/2023
Access products and their popularity	1/19/2023	1/15/2023	1/20/2023	1/15/2023
Analysis on Ingredients	1/19/2023	1/15/2023	1/20/2023	1/16/2023
Regression Test	1/19/2023	1/16/2023	1/20/2023	1/16/2023
Write Documentation on Findings	1/19/2023	1/17/2023	1/20/2023	1/19/2023
Record Video Presentation	1/22/2023	1/18/2023	1/23/2023	1/19/2023

METHODOLOGY

C. DATA SELECTION, COLLECTION AND LIMITATIONS

The data was found on GitHub as a downloaded CVS file that I was able to upload it to my environment in Jupyter Notebooks easily and according to the original project plan. The original data set comprised 11 columns and 1,472 product entries. After removing unnecessary columns and entries with missing data, I had 6 columns and 1,329 complete entries.

The obstacles in the data were all related to data quality, specifically, the completion of the entries. This was addressed by removing affected entries. First, I identified products with missing ingredient lists, this was done by finding entries with placeholders of either “No Info”, or “Visit the (brand) boutique”. The ingredient list is one of the most important elements in this analysis and is entirely unique in every product, therefore all the 143 products with a placeholder was removed. Next, I assessed the rank data and verified no outliers in the upper bounds of the data, but there were 17 entries that had a 0 ranking. Sephora ranks products on a scale from 1 to 5, which shows the 17 entries were missing and not just poorly rated. Since this was a very small proportion of the data, I was comfortable removing the entries.

There were several limitations discovered while assessing product types. After we removed missing values, the dataset contained 1,329 entries, which was then broken down into 6 product types, each having between 150 and 265 products. The data was still appropriate and viable for analysis however the amount in each category was limited. Part of this exercise was to investigate competitive products, this included product pricing. However, a major constraint was not having subcategory data. The most drastic example of this was found in the pricing of face masks which ranged from ~$2 - $250. Masks are typically sold in jars, tubes, and single-use sheet masks, each of these subcategory’s could have vastly different price ranges, but we don’t have sufficient data to make a full analysis or recommendation.

Even with the obstacles and limitations of using this dataset, it was still the most appropriate choice because Sephora is one of the world’s most recognizable names in the beauty industry. It’s in 35 countries worldwide and hosts hundreds of trusted brands and thousands of high-quality products (Pereira, 2022). In using data collected from Sephora we are ensuring the product information is relevant to the current trends and that the ingredients used in the products are high quality and can be used in the products formulated for Makeup+.

D. Data Extraction and Preparation

The data used in this project was originally scraped from the Sephora website and uploaded as a public dataset to Github. The scaped data had been used by the original author and was therefore clean and mostly tidy. I was able to easily download it as a CVS file and uploaded it as a pandas data frame using Jupyter Notebook without issues.

Once my data was extracted and uploaded into my environment, I went through the process of preparing the data for various analyses, tests, and visualizations. I started by evaluating duplicate or missing data, then removing the affected entries.

Next, I worked on the variables used in the regression analysis, rank (dependent) and ingredients (independent). The rank column in the dataset represents the average customer rating for each product with 1 decimal point. I multiplied this column by 10 to remove the decimal for an easier interpretation of the data. Then, I tidied up the ingredient column which that consisted of the entire ingredient list in a single column with each ingredient separated by a comma. I started by stripping any spaces and periods from the list then making them all lowercase. This was very helpful in ensuring uniform formatted for analysis and regression tests. Once uniform, I extracted each ingredient into a dictionary with the product category as the key. This allowed for loops to cycle through every ingredient, in each category, to remove ingredients that are required for the formulas but don't have skin benefits. Then finally count each instance of the ingredients by product type.

After the top 10 ingredients for each product category were identified, I created 6 data frames to hold entries for each category. To run a multivariate analysis, I added a new column for with binary values. After each category had a data frame with the appropriate encoded columns, the data reparation was complete, and the tests could be built.

E. Data Analysis Process Report

1. Description of Analytical Methods

This project includes two methods for analyzing data. First, we used descriptive statistics to summarize the features of the dataset. We looked at the frequency of products in categories, range of prices for each category, and the central tendency of the category ranking. These descriptive techniques allow us to gain beneficial insights about the skin care segment that could be the foundation of our recommended strategy. We also calculated the frequency of ingredients for use in the second analytical method, regression analysis. Ingredients are listed in descending order on the ingredient list leaving some very beneficial ingredients at the bottom. By using the frequency of the ingredients, we can see the most

persistent rather than most prominent ingredients. After determining the top 10 recurring ingredients, we will use each of them as independent variables in a linear regression.

2. Advantages and Limitations of Analytical Methods

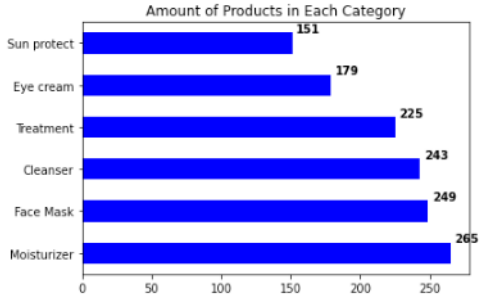
Descriptive statistics is a very common approach to data analysis because it is simple to use, easy to understand, and can be used in conjunction with visualizations to see patterns in data. However, because it is so simple it only provides a summary of data and without additional modeling, they cannot be used to make inferences from the conclusions drawn. It also only gives a one-dimensional view of the data and doesn't identify relationships between variables.

By combining the descriptive methods with regression modeling we can determine the relationship of a single dependent variable (rank) and multiple independent variables (ingredients). Depending on the relationship it could allow us to make predictions about how the product rank can be influenced by certain ingredients. This of course assumes that there is a linear relationship between rank and ingredients, without a linear relationship the model won't be accurate.

3. Application of Analytical Methods

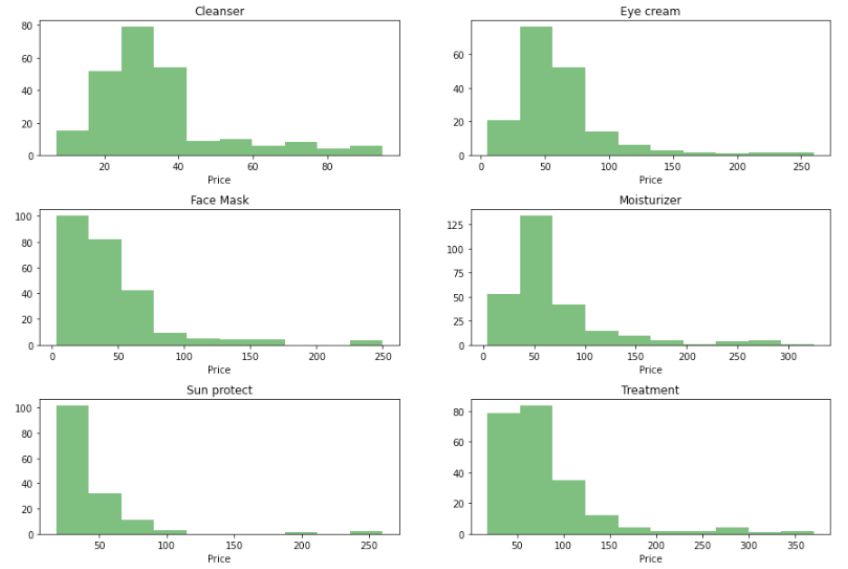
I started my analysis by creating a simple bar chart showing the count of products in each category.

```
In [19]: ax = cleaned_df.Label.value_counts().plot(kind='barh', color='blue')
         ax.set_title('Amount of Products in Each Category')
         for i, v in enumerate(cleaned_df.Label.value_counts()):
             ax.text(v + 3, i + .25, str(v), color='black', fontweight='bold')
```

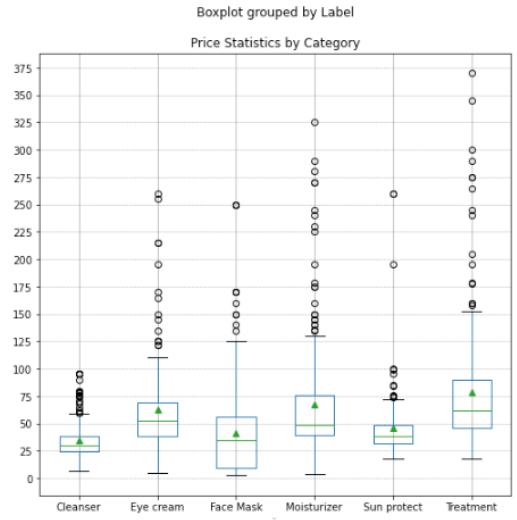


Since each category is very different, I continued my analysis by analyzing the range of prices first with histograms to verify normal distributions. The variation in price for each product type was similar enough to be able to put them into side-by-side box plots, however you can see that there are many possible outliers that make a larger scale.

```
grouped = cleaned_df.groupby('Label')

fig, axs = plt.subplots(nrows=3, ncols=2, figsize=(15, 10))
axs = axs.flatten()

# Loop through the groups and plot a histogram for each group
for ax, (name, group) in zip(axs, grouped):
    ax.hist(group['price'], color = 'green', alpha=0.5, label=name)
    ax.set_title(name)
    ax.set_xlabel('Price')

plt.subplots_adjust(wspace=0.2, hspace=0.4)
plt.show()
```

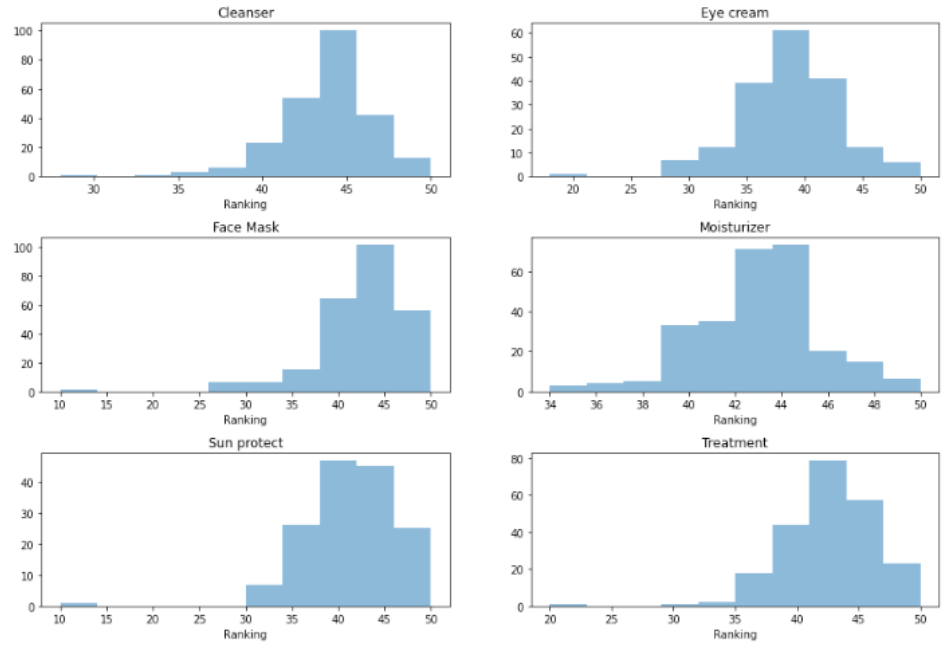


In [21]:
```
ax = cleaned_df.boxplot(column=['price'],by=['Label'],figsize=(8,8), showmeans=True)
plt.title('Price Statistics by Category', fontsize =12)
plt.xlabel('Category', fontsize=0)
ax.grid(visible=True, which='both', axis='y', linestyle='--', linewidth=0.5, color='gray', alpha=0.7)
ax.yaxis.set_ticks(np.arange(0, cleaned_df['price'].max()+25, 25));
```



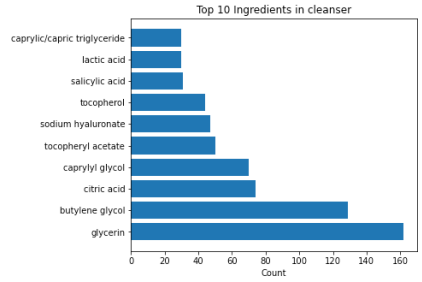
Then I created histograms to verify normal distribution of product ranks and investigate the spread.

In [22]:
```
fig, axs = plt.subplots(nrows=3, ncols=2, figsize=(15, 10))
axs = axs.flatten()

# Loop through the groups and plot a histogram for each group
for ax, (name, group) in zip(axs, grouped):
    ax.hist(group['rank'], alpha=0.5, label=name)
    ax.set_title(name)
    ax.set_xlabel('Ranking')

plt.subplots_adjust(wspace=0.2, hspace=0.4)
plt.show()
```



Understanding the frequency of popular ingredients required looping through each ingredient list for every entry, counting the instances, and outputting a bar chart that representing the top ten ingredient for every category.

```
In [30]: # Loops through the counted ingredients and outputs a bar chart for each product type
         # with their top 10 ingredients
         for product_name, ingredient_counts in products_ingredient_counts.items():
             # Get the top 10 ingredients by count
             top_ingredients = ingredient_counts.most_common(10)
             # Extract the ingredient names and counts
             ingredient_names = [i[0] for i in top_ingredients]
             ingredient_counts = [i[1] for i in top_ingredients]
             # Create a bar chart
             plt.figure(figsize=(6,5))
             plt.barh(ingredient_names, ingredient_counts)
             plt.title(f"Top 10 Ingredients in {product_name}")
             plt.xlabel("Count")
             plt.show()
```



Once this was completed, I moved on to preparing the data for the regression modeling. This included creating a dataframe with for each category and adding an encoded column for the top ingredients in the specific category. Then fitting the model and outputting the OLS summary report.

In [38]:
```
# Call model to fit, then predict the model. It will also print out the summary results.
cleanser_model = fit_mlr(cleanser_new, cleanser_ing)
cleanser_predict = predict_mlr(cleanser_model, cleanser_new)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   rank   R-squared:                       0.041
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.9807
Date:                Wed, 18 Jan 2023   Prob (F-statistic):              0.461
Time:                        14:20:50   Log-Likelihood:                -591.88
No. Observations:                 243   AIC:                             1206.
Df Residuals:                     232   BIC:                             1244.
Df Model:                          10
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       43.4083      0.408    106.291      0.000      42.604      44.213
glycerin                         0.6290      0.423      1.488      0.138      -0.204       1.462
butylene glycol                 -0.0142      0.382     -0.037      0.970      -0.768       0.739
citric acid                     -0.3892      0.414     -0.940      0.348      -1.205       0.427
caprylyl glycol                 -0.4813      0.420     -1.147      0.253      -1.308       0.346
tocopheryl acetate              -0.1772      0.470     -0.377      0.707      -1.104       0.750
sodium hyaluronate               0.2952      0.481      0.614      0.540      -0.652       1.243
tocopherol                       0.6643      0.487      1.364      0.174      -0.296       1.624
salicylic acid                  -0.1862      0.570     -0.326      0.744      -1.310       0.938
lactic acid                      0.4003      0.563      0.712      0.477      -0.708       1.509
caprylic/capric triglyceride     0.6908      0.572      1.208      0.228      -0.436       1.817
==============================================================================
Omnibus:                       71.093   Durbin-Watson:                   2.110
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              240.407
Skew:                          -1.203   Prob(JB):                     6.26e-53
Kurtosis:                       7.237   Cond. No.                         5.24
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

A potential issue when using a multivariate regression model occurs when independent variables (ingredients) are related to each other. Therefore, after each model was created, I ensured no multicollinearity issues by creating a heat map charting the correlation coefficient between each ingredient.

In [36]:
```
def create_correlation_plot(df, independent_vars, name):
    '''
    Function to create a correlation heat map that will help determine multicollinearity
    in our model
    INPUT:
     - Data Frame
     - Independent Variables (ingredients)
    OUTPUT:
     - Correlation heatmap
    '''
    correlation = df[independent_vars].corr()
    f, ax = plt.subplots(figsize=(12, 10))
    mask = np.triu(np.ones_like(correlation, dtype=bool))
    cmap = sb.diverging_palette(230, 20, as_cmap=True)
    sb.heatmap(correlation, annot=True, mask = mask, cmap=cmap)
    plt.title(f'Correlation Heatmap for {name}')
    plt.show()
```

Finally, I put the actual and predicted values calculated from the model to create a scatterplot with the results to better determine the strength of the relationships.

```
In [35]: def scatterplot_mlr_results(df, y_pred, name):
             '''
             Function to plot the results of the regression model
             INPUT:
              - Target Variable - in this case, the rank
              - Predicted Model
              - Dataframe name
             OUTPUT:
              - Scatter Plot with the predicted vs actual values
             '''
             y = df['rank']
             plt.scatter(y, y_pred)
             plt.plot(np.unique(y), np.polyld(np.polyfit(y, y_pred, 1))(np.unique(y)))
             plt.xlabel('Actual Values')
             plt.ylabel('Predicted Values')
             plt.title('Predicted vs Actual Values using {} Data'.format(name))
             plt.show()
```

RESULTS

F. EVALUATE SUCCESS OF ANALYSIS

1. STATISTICAL SIGNIFICANCE

Statistical Significance for the regression models will be evaluated with 5 numbers:

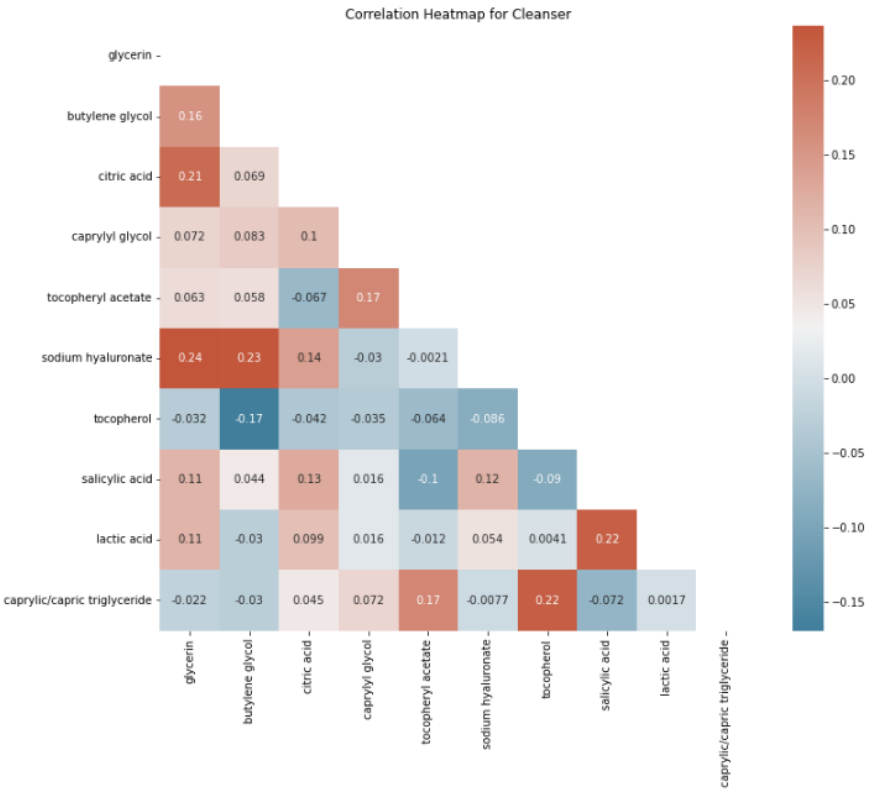
- **R-squared greater than 0.50**. This number represents the percentage of variation in the dependent variable (rank) than can be explained by the independent variables (ingredients). A score higher than .50 would tell us that the majority of the variation in the customer ranking can be explained by the presence of specific ingredients.
- **Adjusted R-squared, close to or higher than R-squared.** A score close or higher than R-squared will let us the model is a good fit. A lower number can represent issues in the model such as overfitting or multicollinearity.
- **Model p-value, less than 0.05.** A value less than 0.05 will verify that the data is a good fit for the model and show that ingredients are statistically significant in predicting the customer rank.
- **Independent Variables (ingredients) coeffects, greater than .7.** The coefficient will show the relationship between the individual ingredient and the overall rank.
- **Independent Variables (ingredients) p-values, less than 0.05.** An ingredients p-value less than 0.05 will explain statistical significance relating to the rank.

Cleanser Regression Results:

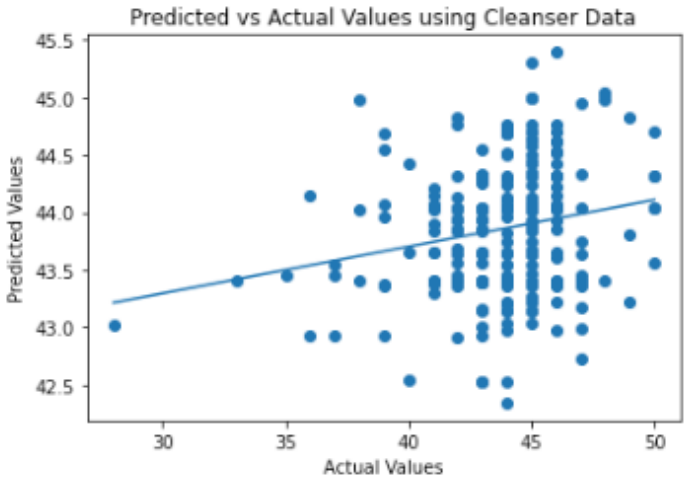
R-Squared	***Adjusted R-Squared***	***Model P-Value***
0.041	-0.001	0.461
Independent Variable	***Coefficient***	***P-value***
Glycerin	0.6290	0.138
Butylene glycol	-0.0142	0.970
Citric acid	-0.3892	0.348
Caprylyl glycol	-0.4813	0.253
Tocopheryl acetate	-0.1772	0.707
Sodium hyaluronate	0.2952	0.540
Tocopherol	0.6643	0.174
Salicylic acid	-0.1862	0.744
Lactic acid	0.4003	0.477
Caprylic/capric triglyceride	0.6908	0.228

The R-Squared number indicates that only 4.1% of the variation in the dependent variable can be explained by the independent variables. An adjusted r-squared score of -0.001 along with the model's p-value of 0.461 indicated that this model may not great fit for the data.

The p-values don't show any statistical significance in determining the outcome of the rank. We can also see that the ingredient coefficient has both positive and negative values. A potential explanation for the model's fit could be multicollinearity. If the ingredients have a strong relationship with each other (over 0.7), this would be a signal that we have multicollinearity in our model. The graph below is a correlation heat map that shows the relationships between each ingredient. We can see that the strongest correlation is only 0.24 between sodium hyaluronate and glycerin.



Since multicollinearity doesn't seem to be an issue, we want to plot the results of our model. The best way to do this is with a scatterplot with the line of best fit that was created from our model using the least square method.



The scatterplot shows that there is not a strong linear relationship between the customer rank and the ingredients.

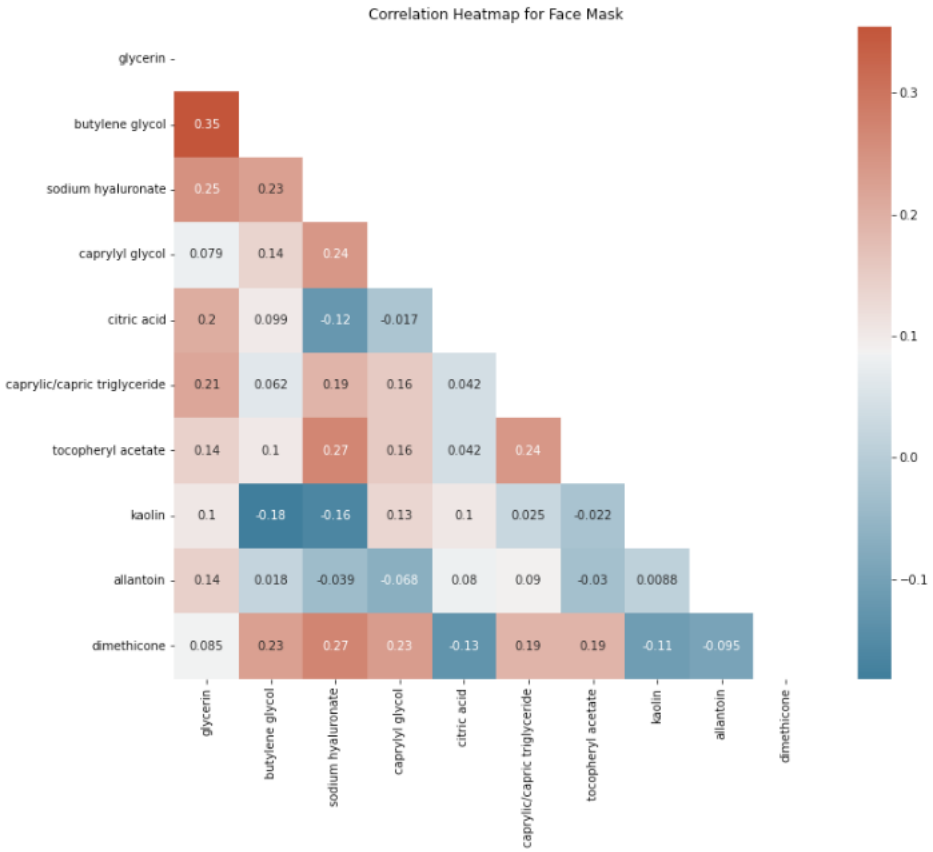
Based on this analysis, we fail to reject the null hypothesis for cleanser ingredients affecting customer ranking.

Face Mask Regression Results:

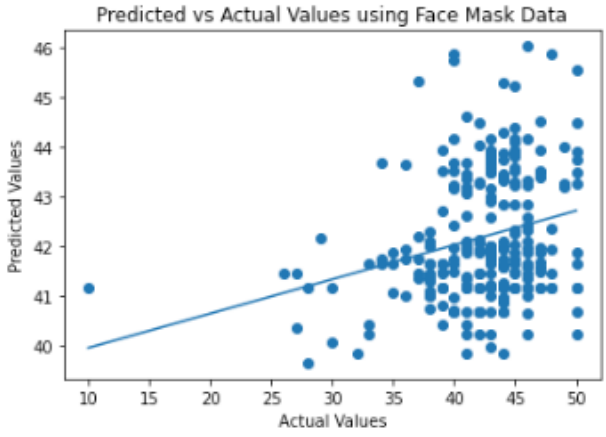
R-Squared	***Adjusted R-Squared***	***Model P-Value***
0.069	0.030	0.0691
Independent Variable	***Coefficient***	***P-value***
Glycerin	0.7517	0.413
Butylene glycol	-0.2916	0.689
Sodium hyaluronate	-0.1967	0.788
Caprylyl glycol	-0.0565	0.940
Citric Acid	0.2390	0.746
Caprylic/capric triglyceride	1.8526	**0.021**
Tocopheryl acetate	-0.4021	0.614
Kaolin	1.9919	**0.015**
Allantoin	-1.2028	0.131
Dimethicone	0.3744	0.640

The R-Squared number indicates that only 6.9% of the variation in the dependent variable can be explained by the independent variables. An adjusted r-squared score of 0.030 along with the model's p-value of 0.06 indicated that there is not a strong correlation between the ingredients and the customer rank.

The p-values for Caprylic/capric triglyceride and Kaolin could represent significance in predicting the overall customer result but first, we need to investigate the coefficients further. Just like in the cleanser, a potential explanation for the model's poor fit could be multicollinearity. The graph below is a correlation heat map for the mask ingredients. We can see that the strongest correlation is only 0.35 between butylene glycol and glycerin.



Since multicollinearity doesn't seem to be an issue, we want to plot the results of our model. The best way to do this is with a scatterplot with the line of best fit that was created from our model using the least square method.



The scatterplot shows that there is not a strong linear relationship between the customer rank and the ingredients.

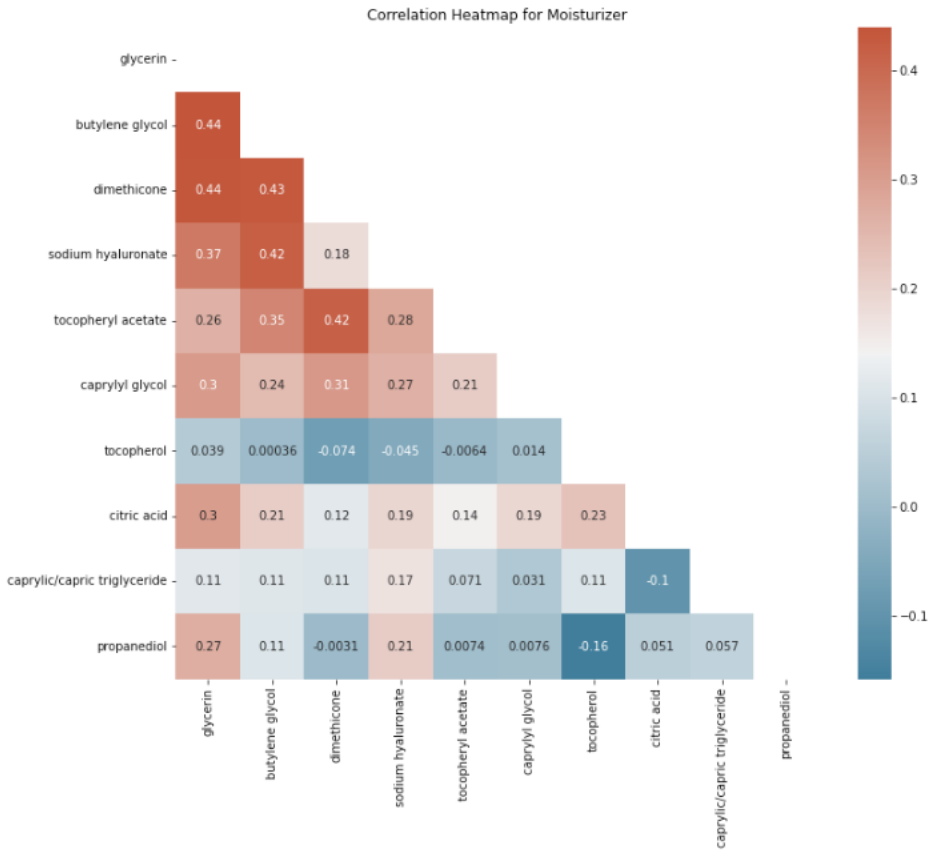
Based on this analysis, we fail to reject the null hypothesis for mask ingredients affecting customer ranking.

Moisturizer Regression Results:

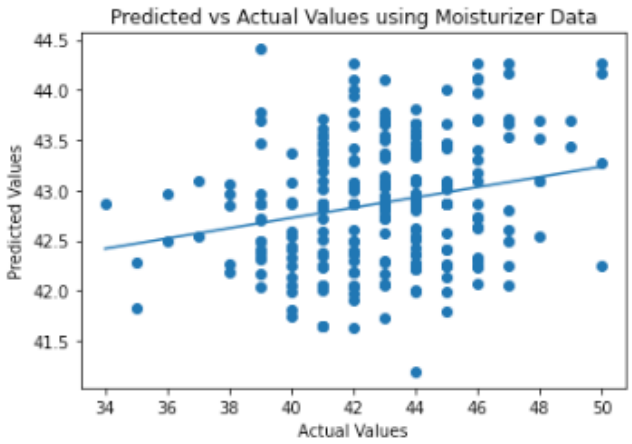
R-Squared	***Adjusted R-Squared***	***Model P-Value***
0.051	0.014	0.197
Independent Variable	***Coefficient***	***P-value***
Glycerin	0.6101	0.233
Butylene glycol	-1.2049	**0.004**
Dimethicone	0.2296	0.587
Sodium hyaluronate	-0.1034	0.798
Tocopheryl acetate	0.0154	0.968
Caprylyl glycol	-0.1622	0.670
Tocopheryl	0.2312	0.534
Citric Acid	-0.1964	0.635
Caprylic/capric triglyceride	0.5636	0.144
Propanediol	0.0005	0.999

The R-Squared number indicates that only 5.1% of the variation in the dependent variable can be explained by the independent variables. An adjusted r-squared score of 0.014 along with the model's p-value of 0.197 indicated that there is not a strong correlation between the ingredients and the customer rank.

The p-value for Butylene Glycol could represent significance in predicting the overall customer result but first, we need to investigate the coefficients further. Just like in the previous product categories, a potential explanation for the model's poor fit could be multicollinearity. The graph below is a correlation heat map for the moisturizer ingredients. We can see that the strongest correlation is only 0.44 between butylene glycol and glycerin as well as dimethicone and glycerin.



Since multicollinearity doesn't seem to be an issue, we want to plot the results of our model. The best way to do this is with a scatterplot with the line of best fit that was created from our model using the least square method.



The scatterplot shows that there is not a strong linear relationship between the customer rank and the ingredients.

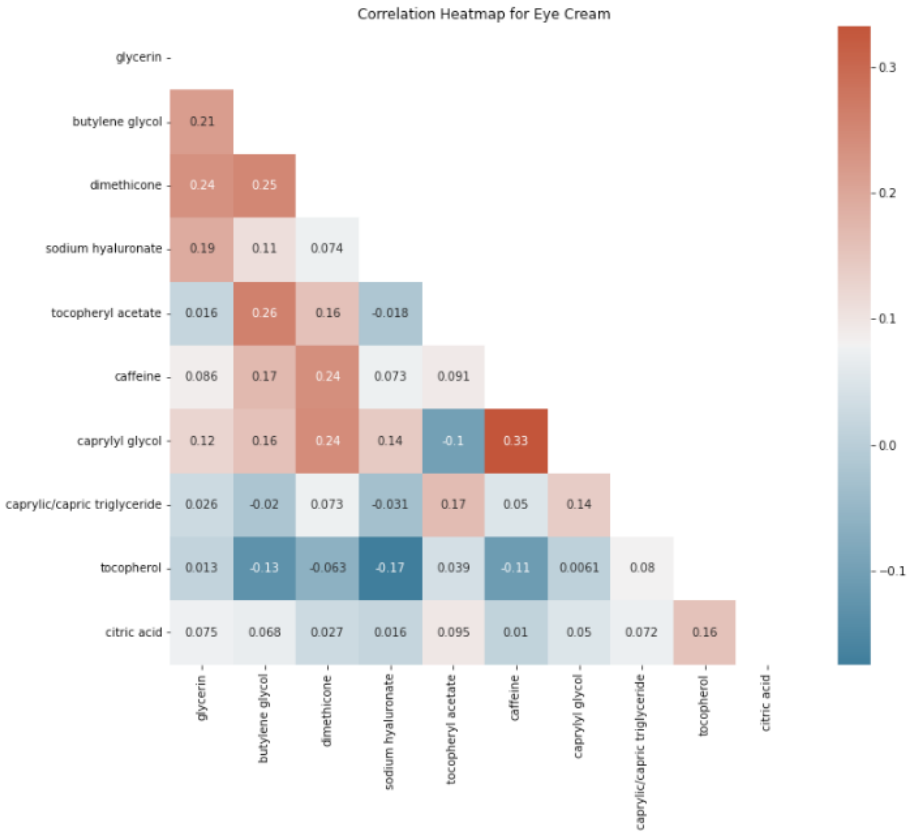
Based on this analysis, we fail to reject the null hypothesis for moisturizer ingredients affecting customer ranking.

Eye Cream Regression Results:

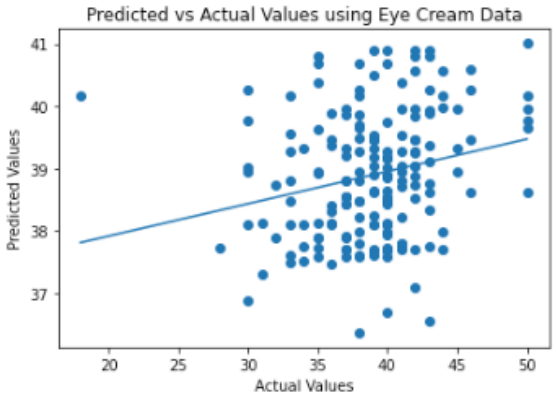
R-Squared	***Adjusted R-Squared***	***Model P-Value***
0.052	-0.005	0.517
Independent Variable	***Coefficient***	***P-value***
Glycerin	0.8435	0.540
Butylene glycol	1.3473	0.107
Dimethicone	-1.2107	0.104
Sodium hyaluronate	-0.5220	0.457
Tocopheryl acetate	0.1006	0.892
Caffeine	-0.3031	0.688
Caprylyl glycol	-1.0535	0.180
Caprylic/capric triglyceride	0.2070	0.776
Tocopheryl	0.1222	0.870
Citric Acid	-0.2033	0.792

The R-Squared number indicates that only 5.2% of the variation in the dependent variable can be explained by the independent variables. An adjusted r-squared score of -0.005 along with the model's p-value of 0.517 indicated that this model may not great fit for the data.

The p-values don't show any statistical significance in determining the outcome of the rank. We can also see that the ingredient coefficient has both positive and negative values. Just like in the previous product categories, a potential explanation for the model's poor fit could be multicollinearity. The graph below is a correlation heat map for the eye cream ingredients. We can see that the strongest correlation is only 0.33 between caprylyl glycol and caffeine.



Since multicollinearity doesn't seem to be an issue, we want to plot the results of our model. The best way to do this is with a scatterplot with the line of best fit that was created from our model using the least square method.



The scatterplot shows that there is not a strong linear relationship between the customer rank and the ingredients.

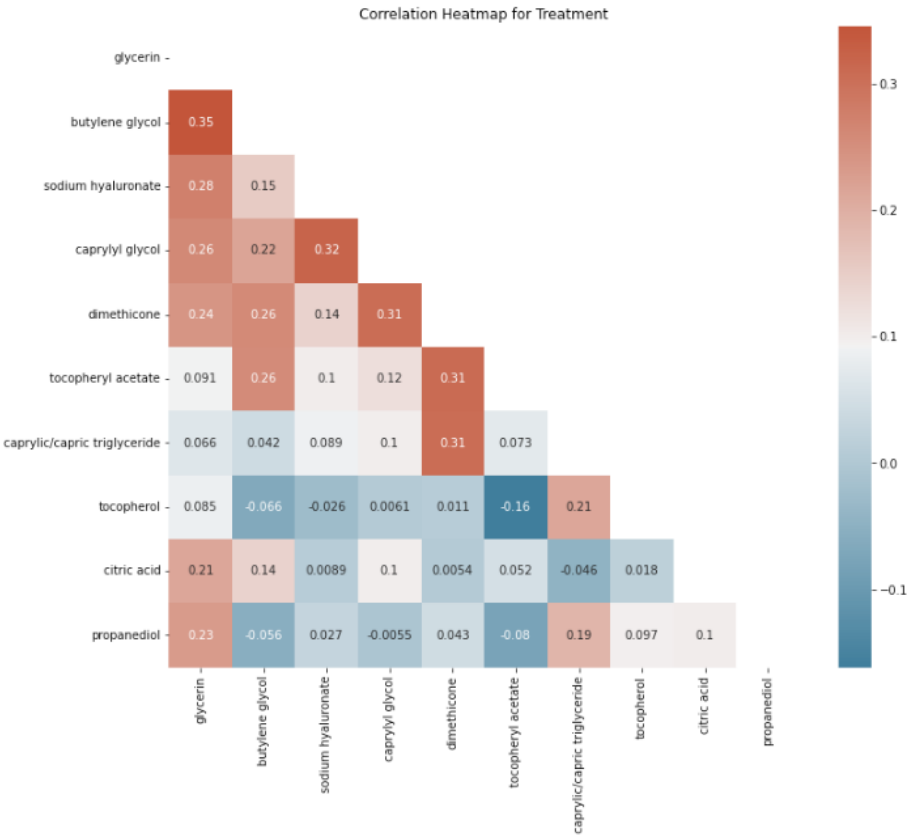
Based on this analysis, we fail to reject the null hypothesis for eye cream ingredients affecting customer ranking.

Treatment Regression Results:

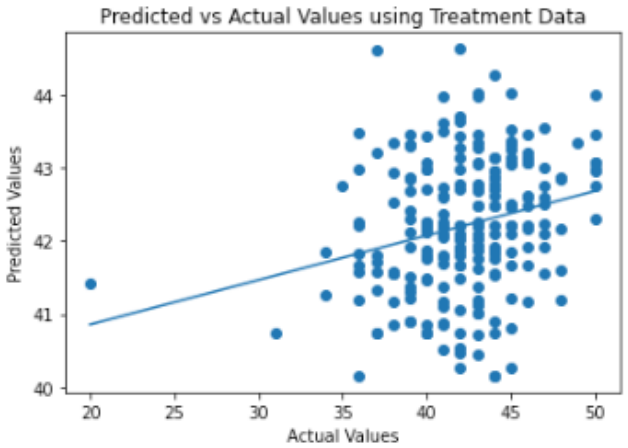
R-Squared	***Adjusted R-Squared***	***Model P-Value***
0.061	0.017	0.190
Independent Variable	***Coefficient***	***P-value***
Glycerin	0.2606	0.721
Butylene glycol	-0.5529	0.334
Sodium hyaluronate	0.5922	0.277
Caprylyl glycol	-0.7432	0.190
Dimethicone	-0.9121	0.121
Tocopheryl acetate	-0.1067	0.852
Caprylic/capric triglyceride	1.2728	**0.039**
Tocopheryl	-0.3431	0.549
Citric Acid	0.9488	0.101
Propanediol	0.2694	0.653

The R-Squared number indicates that only 6.1% of the variation in the dependent variable can be explained by the independent variables. An adjusted r-squared score of 0.017 along with the model's p-value of 0.190 indicated that there is not a strong correlation between the ingredients and the customer rank.

The p-value for Caprylic/capric triglyceride could represent significance in predicting the overall customer result but first, we need to investigate the coefficients further. Just like in the previous product categories, a potential explanation for the model's poor fit could be multicollinearity. The graph below is a correlation heat map for the moisturizer ingredients. We can see that the strongest correlation is only 0.35 between butylene glycol and glycerin.



Since multicollinearity doesn't seem to be an issue, we want to plot the results of our model. The best way to do this is with a scatterplot with the line of best fit that was created from our model using the least square method.



The scatterplot shows that there is not a strong linear relationship between the customer rank and the ingredients.

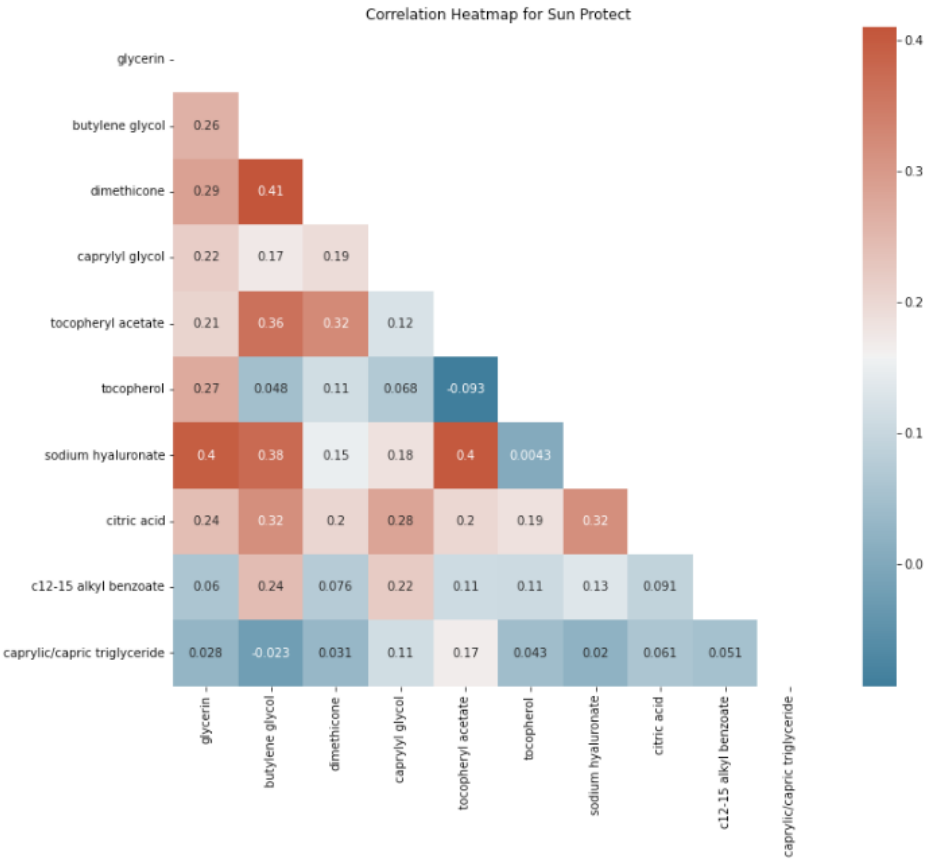
Based on this analysis, we fail to reject the null hypothesis for treatment ingredients affecting customer ranking.

Sun Protect Regression Results:

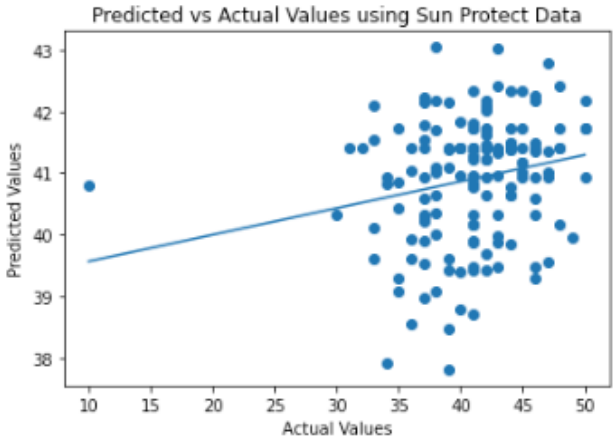
R-Squared	***Adjusted R-Squared***	***Model P-Value***
0.043	-0.025	0.784
Independent Variable	***Coefficient***	***P-value***
Glycerin	0.3164	0.755
Butylene glycol	0.1318	0.896
Dimethicone	0.6173	0.520
Caprylyl glycol	-0.6462	0.474
Tocopheryl acetate	-0.6071	0.536
Tocopheryl	-1.5398	0.099
Sodium hyaluronate	-0.9015	0.413
Citric Acid	1.1741	0.315
c12-15 alkyl benzoate	-0.8670	0.396
Caprylic/capric triglyceride	0.6844	0.519

The R-Squared number indicates that only 4.3% of the variation in the dependent variable can be explained by the independent variables. An adjusted r-squared score of -0.025 along with the model's p-value of 0.784 indicated that this model may not great fit for the data.

The p-values don't show any statistical significance in determining the outcome of the rank. We can also see that the ingredient coefficient has both positive and negative values. Just like in the previous product categories, a potential explanation for the model's poor fit could be multicollinearity. The graph below is a correlation heat map for the eye cream ingredients. We can see that the strongest correlation is only 0.41 between dimethicone and butylene glycol.



Since multicollinearity doesn't seem to be an issue, we want to plot the results of our model. The best way to do this is with a scatterplot with the line of best fit that was created from our model using the least square method.



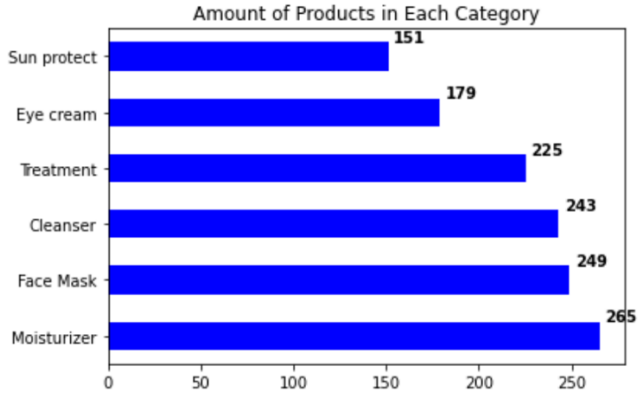
The scatterplot shows that there is not a strong linear relationship between the customer rank and the ingredients.

Based on this analysis, we fail to reject the null hypothesis for treatment ingredients affecting customer ranking.

2. Practical Significance

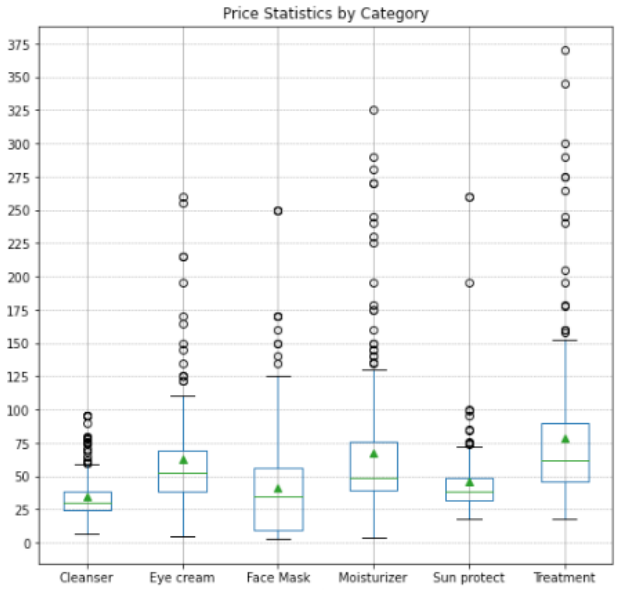
Even though we didn't find a strong linear relationship between customer rank and the ingredients we have gained a lot of insights into the skin care segment that led us to create a data-driven strategy.

We learned that Moisturizers, Face masks and Cleansers have the most products in the category and that Sun Protect has the fewest products. This can be significant in knowing which products have the biggest market share, as well as potential gaps that Makeup+ could fill.

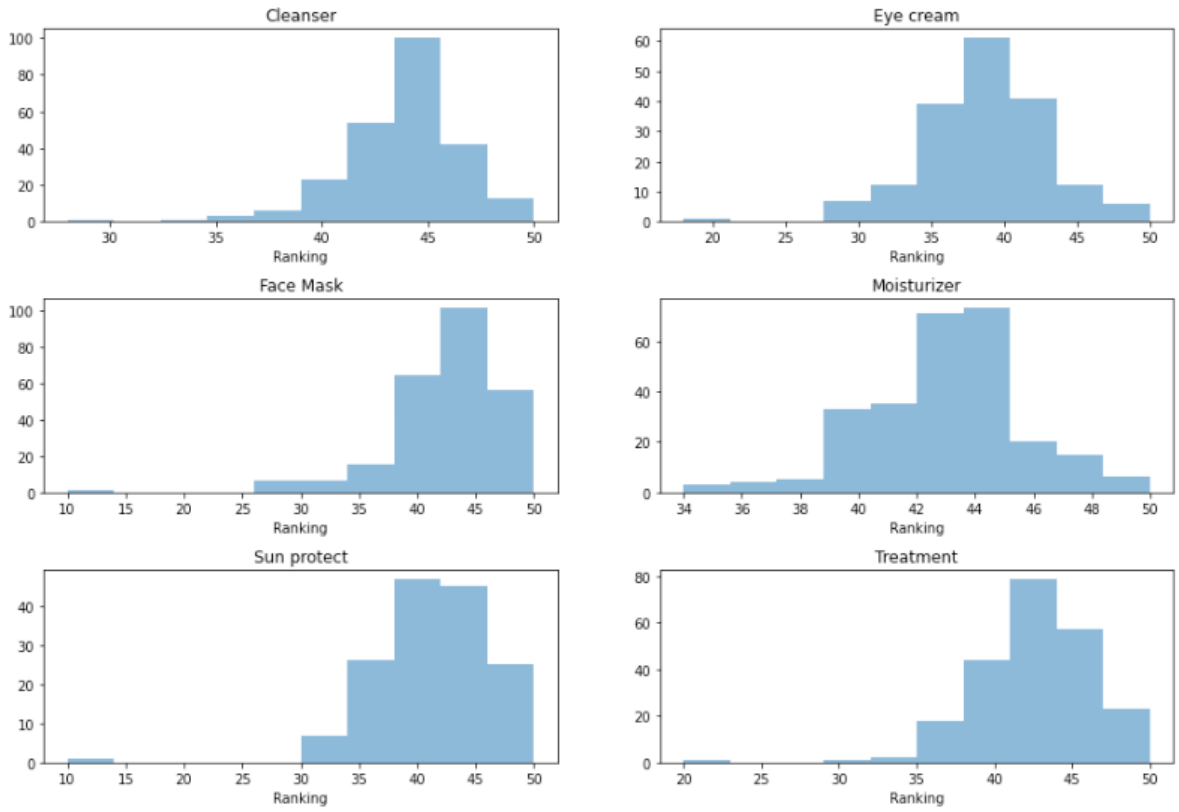


There were limitations called out in section C regarding the pricing for each product, but it is important to understand a basic pricing structure for each product category before formulating and packaging the product to stay within the company's profit margins. For general pricing guidelines, we can use the inner quartile ranges.

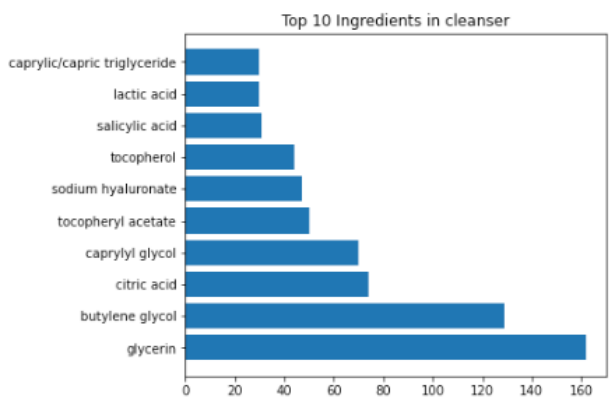
- Cleansers: $25 and $40.
- Eye Creams: $40 and $70.
- Face masks: $10 and $55.
- Moisturizers: $40 and $75.
- Sun Protect: $30 and $50.
- Treatment: $45 and $85.

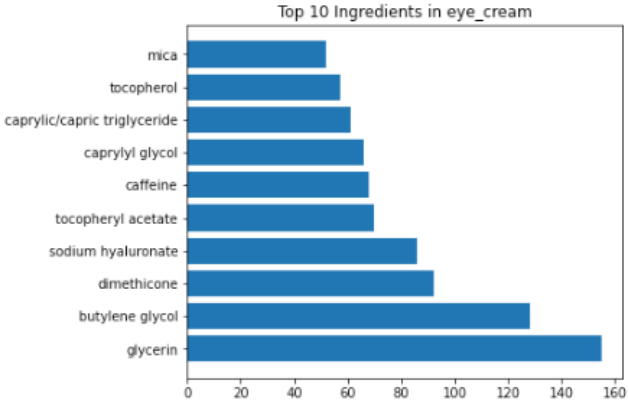


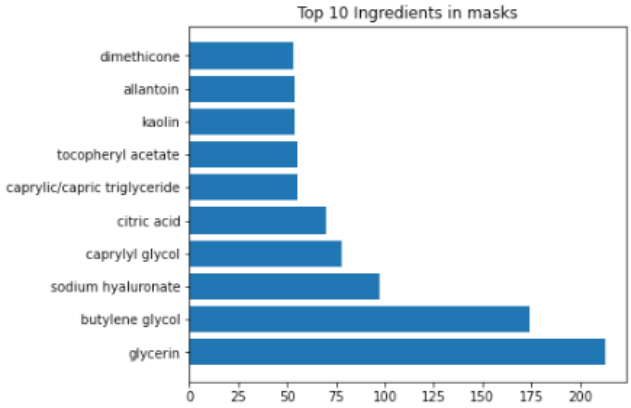
While the analysis on rank was in the pursuit to find a correlation with ingredients, we learned that skin care products are generally ranked very well, 3.5 on average (with a transformed value of 35). We can also see that not only do cleansers and moisturizers have the most products, but they are also ranked the highest.

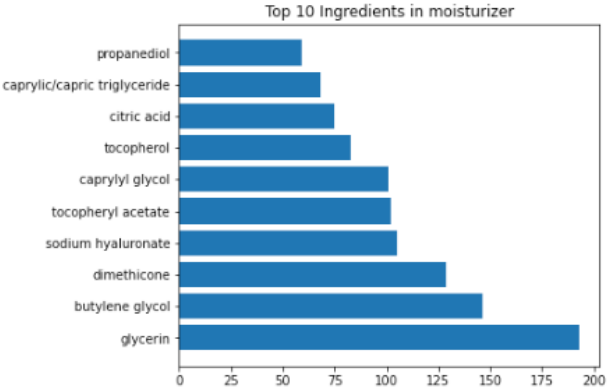


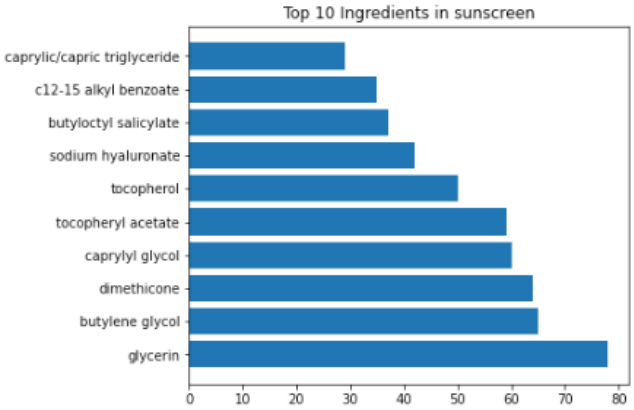
Though there isn't a correlation with the overall product rank, the research conducted for the project proposal shows that customers are very knowledgeable and interested in skin care ingredients. Therefore, our last major insight from this project is revealing the top ten most frequently used ingredients.

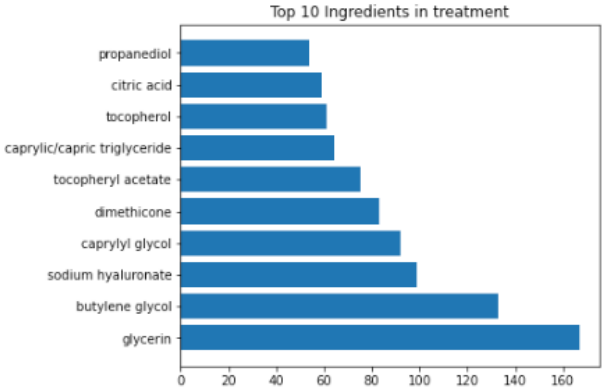




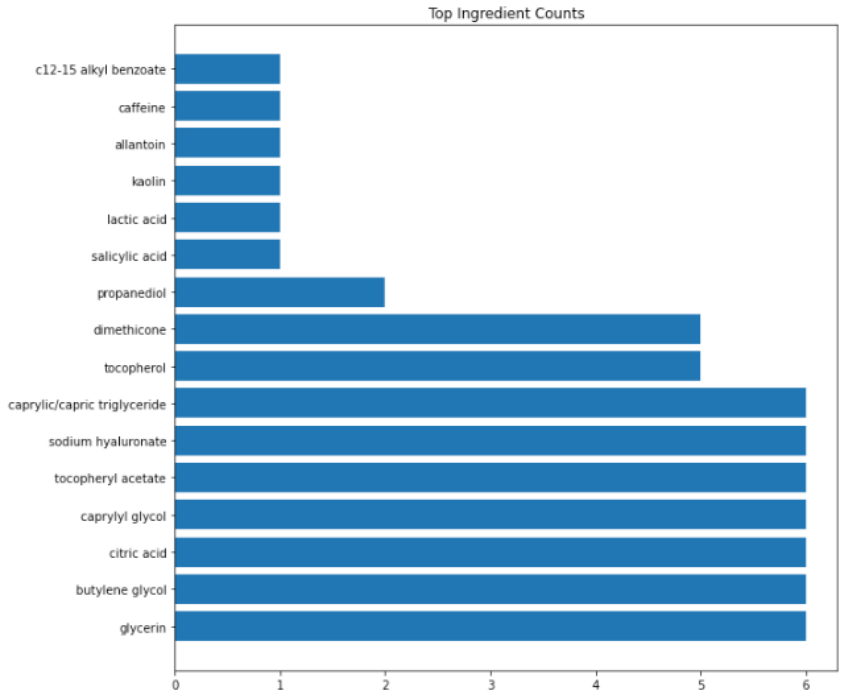








As we are reviewing the most frequently used ingredients, we also see many recurring ingredients in every category. It is interesting to note that in every category there are 10 recurring ingredients that are meant to promote hydration in skin.



- **Glycerin**: Moisturizer naturally found in skin. Works by pulling water into the skin, keeping it hydrated, it helps to restore protective barriers in skin.
- **Butylene Glycol**: Moisturizing compound that works by attracting water, quickly absorbing it, then pulling it deeper into the skin.
- **Caprylyl Glycol**: Moisturizing ingredient like Butylene Glycol that works as a humectant that attracts and retains moisture.
- **Tocopheryl Acetate**: Also known as Vitamin E. Multi-beneficial ingredient that works as a superior antioxidant, promotes regeneration of healthy cells, and protects skin from ultraviolet rays.
- **Sodium Hyaluronate**: Also known as Hyaluronic Acid. One of most hydrating ingredients that it naturally occurring in skin.
- **Tocopherol**: Is the natural form of Tocopheryl Acetate. This more natural form of Vitamin E has the same benefits and it easier for the skin to absorb.
- **Caprylic/Capric Triglyceride**: Derives from coconut oil, this ingredient works for all skin types and creates a protective barrier over the skin to decrees the loss of moisture.
- **Dimethicone**: Common moisturizing ingredient that forms a protective barrier and fills fine lines and wrinkles.
- **Citric Acid**: Gentle exfoliator that gentle lifts off dead skin cells from the top layer of the skin.
- **Propanediol**: Natural humectant that pulls moisture into skin, encourages water retention, and helps to prevent water loss.

* Information about ingredients was pulled from INCIDecoder.com.
A more comprehensive list of ingredients used in this project can be found in the appendix of this report.

3. Overall Success

The success of the project, as defined by the project plan, was to analyze Sephora data to create a strategy for Makeup+ to launch into the skin care segment. We completed each of the objectives by using descriptive statistics to inspect features of the dataset and successfully complete regression testing on the product rank and the ingredients. These analyses have been completed and synthesized into a recommended strategy found in section G3 of this report. Therefore, we can conclude that this project was a success.

G. Key Takeaways

1. Summarization of Conclusions

This project has produced many useable insights about skin care categories and ingredients. We now understand that treatments are generally the most expensive product category and should be investigated with a dataset with subcategories.

Out of a ranking scale of 1 – 5 most products, on average, rank above 3.5. Cleansers and Moisturizers have the most favorable rankings.

We learned that even though ingredients are fundamental to customers there isn't a clear linear relationship between popular ingredients and the product rank.

Out of the top ingredients found across all product categories ones that promote skin hydration seem to be the most prevalent.

2. Explanation of Tools and Graphical Representations

The visualizations used were created to make the data more understandable. Being able to understand and interpret the data helped to create an informed recommendation with the context of the insights gained. The histograms and bar were created to show the distribution of the data which made it easier to compare the product categories to one another. The boxplots and the scatterplots help to convey the relationship between multiple features of the dataset, which made it easier to see trends in the data.

3. Recommendations Based on Findings

Recommendation 1 - Moisturizing Regimen

The first product strategy it to launch Cleanser, Night Cream (moisturizer) and Treatment that gives the user ultra-hydrated skin. This strategy leverages all the hydrating ingredients found in the above analysis. A full pricing analysis should be performed before a final price of the product is set however, using the current data, we can use the following price ranges as starting point:

- Cleanser: $25 - $40 with an average price of $30.
- Night Cream: $40 - $75 with an average price of $50.
- Treatment: $50 - $80 with an average price of $60

Moisturizing ingredients should include:

- Glycerin
- Butylene Glycol
- Caprylyl Glycol
- Sodium Hyaluronate
- Caprylic/Capric Triglyceride
- Propanediol
- Squalene

Each ingredient in this recommendation is naturally occurring and used together achieves a deep moisturization with protective layers to hold in water for maximum benefits.

Recommendation 2 - Multi-Benefit Regimen

The second recommended strategy is for a Cleanser, Day-time Moisturizer, and Mask, with each product boasting multiple beneficial ingredients. This recommendation takes the three most popular categories then combines frequently used ingredients in every category to build a full regimen with many benefits.

The cleanser will have moisturizing, gently exfoliating, problem skin benefits. It should be priced between $25 and $40, and ingredients should include:

- Sodium Hyaluronate
- Citric Acid
- Salicylic Acid

A day-time moisturizer should of course have a hydrating quality as well as sun protection, and antioxidants. It should be priced between $40 and $75, and ingredients should include:

- Sodium Hyaluronate
- Tocopheryl Acetate
- C12-15 Alkyl Benzoate
- Caprylic/Capric Triglyceride

Multi use clay mask that is gently exfoliating, promotes healing, and is hydrating. It should be priced between $10 and $55, and ingredients should include:

- Kaolin Clay
- Allantoin
- Caprylic/Capric Triglyceride
- Citric Acid
- Sodium Hyaluronate

While this recommendation could not be positioned as 'natural' it provides multiple benefits for skin health that can be used as standalone products, or a full regimen.

H. Panopto Presentation

A link to the Panopto presentation can be found here: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=02b469b0-670f-48d0-8c2c-af90012b2f97&edit=true

APPENDICES

I. Evidence of Project Completion

You will find the evidence of project completion uploaded with this written report which includes the following.

1) PDF Version of the code completed in Jupyter Notebooks
2) CVS Copy of the data used to complete this project
3) One-page summary of the recommended product launches.
4) The summary presentation can also be found at the link in section H.

J. Sources

Pereira D. (2022) Sephora Business Model Retrieved from: https://businessmodelanalyst.com/sephora-business-model

INCI Decoder (n.d) Retrieved from: https://incidecoder.com/

K. INGREDIENT GLOSSARY

These are ingredient definition came from INCIDecoder.com and are either beneficial for skin or used in formulas.

Skin Beneficial Ingredients:

Glycerin: Moisturizer naturally found in skin. Works by pulling water into the skin to keep it hydrated and help to restore protective barriers in skin.

Butylene Glycol: Moisturizing compound that works by attracting water and quickly absorbing it and pulling it deep into the skin.

Citric Acid: Gentle exfoliator that gentle lifts off dead skin cells from the top layer of the skin.

Caprylyl Glycol: Moisturizing ingredient like Butylene Glycol that works as a humectant that attracts and retains moisture.

Tocopheryl Acetate: Also known as Vitamin E. Multi-beneficial ingredient that works as a superior

Sodium Hyaluronate: Also known as Hyaluronic Acid. One of most hydrating ingredients that it naturally occurring in skin.

Tocopherol: Is the natural form of Tocopheryl Acetate. This more natural form of Vitamin E has the same benefits and it easier for the skin to absorb.

Salicylic Acid: Popular ingredient for treating blackheads and acne. Works as an anti-inflammatory exfoliator for skin surface and in pores.

Kaolin: Naturally occurring clay that gently absorbs excess oil from skin.

Allantoin: Used as a soothing agent that promotes healing while softening and protecting the skin.

Dimethicone: Common moisturizing ingredient that forms a protective barrier and fills fine lines and wrinkles.

Propanediol: Natural humectant that pulls moisture into skin, encourages water retention, and helps to prevent water loss.

Squalene: Naturally occurring oil that is lightweight and extremely moisturizing.

Caffeine: Naturally occurring stimulant with antioxidants.

Cyclopentasiloxane: Works with Dimethicone to create a breathable protective barrier in skin.

C12-15 Alkyl Benzoate: Sun Protection.

Ingredients for Formulas:

Water: Acts as a solvent for other ingredients

Phenoxyethanol: Derived from Green Tea, it is used as a preservative and stabilizer in formulas

Disodium EDTA: Common ingredient that keeps for mulas integrity by preventing other ingredients to bind to trace minerals found in water.

Sodium Benzoate: Preservative that works against fungi.

Ethylhexylglycerin: Used as a deodorant and boosts the effectiveness of other preservatives.

Potassium Sorbate: Preservative against mold and yeasts.

Sodium Chloride: Also known as table salt, used as a thickener for cleansing formulas with ionic cleansing agents.

Sodium Hydroxide: Alkalotic ingredient used in small amounts to adjust the pH of the product.

Xanthan Gum: Used as a thickening agent and emulsion stabilizer to produce a more gel-like feel.

Cocamidopropyl Betaine: Stabilizer for foaming age nts in cleaning products.

Polysorbate 20: Increased solubility of ingredients and allows water and oil to mix

Glyceryl Stearate: Fatty acid that helps water and oil to mix.

Chlorphenesin: Preservative against some fungi and yeasts.

Stearic Acid: Fatty Acid that gives body to cream type products and stabilizes emulsions

1, 2-Hexanediol: Multifunctional ingredient that boosts antimicrobial activity in preservatives and works as an emollient and solvent.

Carbomer: Converts liquid to create viscous, clear gels

Sodium Benzoate: Preservative against fungi.

Cetearyl Alcohol: Gives body to creams and lotions and stabilizes oil and water mixtures.

PEG-100 Stearate: Emulsifier that works with a wide range of pH levels.

Mica: Mineral powder that increases skin adhesion and gives it a pearly sheen.

Acrylates/c10-30 Alkyl Acrylate Crosspolymer: Work s as an emulsion stabilizer.

Silica: Keeps skin matte and acts as a thickening agent.

Lecithin: Water-binding ingredient that is used to create liposomes and stabilize products

Pentylene Glycol: Broad Spectrum antimicrobial that is a solvent and emulsion stabilizer.

BHT: Synthetic Preservative

Aluminum Hydroxide: Functions as an opacifier and emollient skin protector

Butyloctyl Salicylate: used in sunscreens to solubilize UV-Filters.

Fragrance: Artificial scent

Limonene: Naturally occurring fragrance

Linalool: Naturally occurring fragrance