

× × × ×

# Project : Influence of Ingredients on Häagen-Dazs Ice Cream Flavors and Ratings

× × × ×

Presented by : Chavaporn T.

Link : <https://github.com/Pekky28/Portfolio0-Clustering.git>



x x x x

# Outline

<b>Executive Summary</b>
<b>Introduction</b>
<b>Methodology</b>
<b>Result</b>
<b>Conclusion</b>
<b>Appendix</b>

x x x x

# x x x x Executive Summary

## Summary of Methodology

The project begins by gathering a comprehensive ice cream dataset from Kaggle website which obtained the data directly from brand website. After that We check the data and column to ensure data consistency and accuracy, this includes essential information about Häagen-Dazs ice cream flavors, such as names, descriptions, ratings and rating counts.

Data wrangling is employed to extract unique ingredients from the dataset, forming the basis for subsequent analysis. Dummy columns are generated to represent the presence of each ingredient in the dataset.

Exploratory Data Analysis (EDA) is conducted to gain a deep understanding of the collected data.

The next step involves building a clustering model. The K-Nearest Neighbors (KNN) algorithm is used to group ingredients into clusters. The elbow method is employed to determine the optimal number of clusters (K) using Within-Cluster Sum of Squares (WCSS) as a metric. Silhouette scores are calculated for different K values, and the optimal K is determined to be 3. Relationships between clusters and ice cream ratings, rating counts, and ingredients are explored using scatterplots

The project sets specific thresholds for ratings and rating counts to determine which cluster has the most influence on high ratings. Ice cream flavors that meet these criteria are identified as top performers.

# × × × × Executive Summary

## Summary of Result

The project successfully identified certain clusters of ingredients within Häagen-Dazs ice cream flavors that have a significant influence on high ratings. Some Häagen-Dazs ice cream flavors with high ratings and rating counts were identified, providing valuable insights into which ingredients contribute to the brand's better-rated ice creams.

# × × × × Problems/Objective

Determine the key ingredient(s) that contribute to high ratings in ice cream.

Cluster ingredients to identify patterns and associations between different ice cream flavors.

Explore the relationship between ingredient clusters, ice cream ratings, and rating counts to gain insights into what makes an ice cream flavor popular.



× × × ×

# Introduction

## Project Background Context

The project aims to investigate the influence of different ingredients on ice cream flavors and ratings, with a particular focus on the Häagen-Dazs brand, which offers a diverse range of 70 flavors.

The primary objective is to identify which ingredient has the most significant impact on high ratings within this brand. The dataset contains columns such as 'key,' 'icecream\_name,' 'description,' 'rating,' 'rating\_count,' and 'ingredients.'

The project involves data wrangling, clustering with K-Nearest Neighbors, and analysis to find clusters of ingredients that have the most influence on Häagen-Dazs ice cream ratings.



x x x x

# Methodology

Data Collection Methodology

Perform data wrangling

Perform EDA using visualization

Perform K-Nearest Neighbors (KNN)



# × × × × Data Collection

---

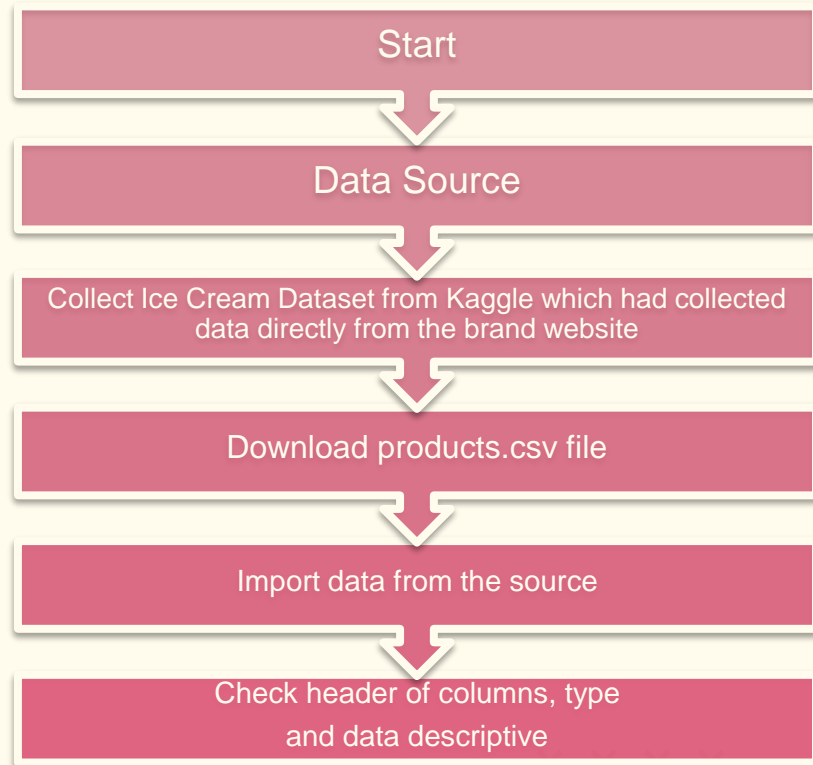
**Start:** The data collection process begins.

**Data Sources:** The data was collected from Kaggle which directly collected dataset and images from the brand website :<https://www.haagendazs.us/products>

**Data Import:** Download Häagen-Dazs's products csv file  
Then import the data into Jupyter Notebook

**Data Description:** Descriptive information about each flavor such as: the flavor name, description, average rating, and ingredients list.

Check data type, missing values, columns before perform data wrangling.





# Data Collection

**Ice Cream Dataset**

Data Card Code (11) Discussion (1)

61 New Notebook Download (55 MB)

About this directory

Data from Häagen-Dazs only

- images 70 files
- products.csv 30.51 kB
- reviews.csv 1.63 MB

Summary

- 488 files
- 82 columns

bj  
breysers  
combined  
hd  
talenti

**Data Sources:** The data was collected from [Kaggle](https://www.kaggle.com) which directly collected dataset and images from the brand website :<https://www.haagendazs.us/products>

**Ice Cream Dataset**

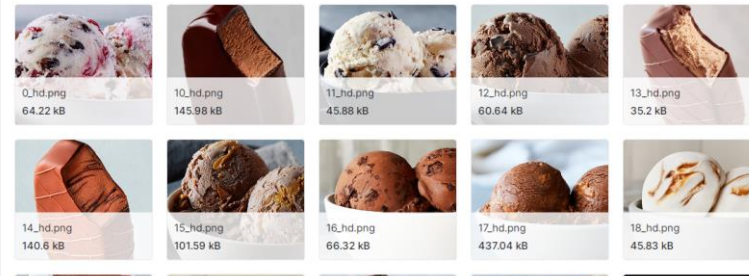
Data Card Code (11) Discussion (1)

61 New Notebook

images (70 files)

About this directory

Product images from Häagen-Dazs



0_hd.png 64.22 kB	10_hd.png 145.98 kB	11_hd.png 45.88 kB	12_hd.png 60.64 kB	13_hd.png 35.2 kB
14_hd.png 140.6 kB	15_hd.png 101.59 kB	16_hd.png 66.32 kB	17_hd.png 437.04 kB	18_hd.png 45.83 kB



x x x x

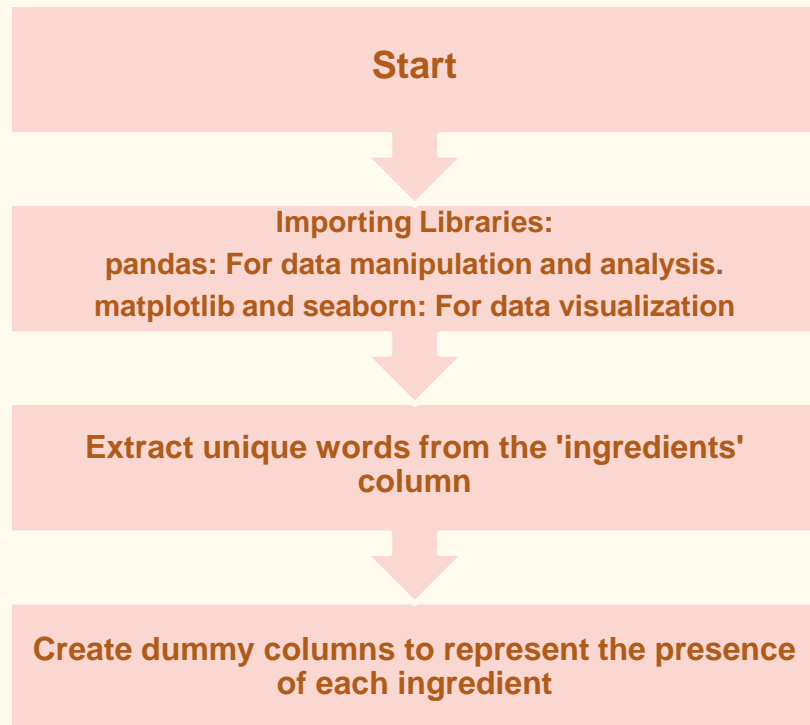
# Data Wrangling

---

The process begins by importing the necessary libraries for data manipulation and analysis.

Next, data wrangling is employed to extract unique ingredients from the dataset, forming the basis for subsequent analysis.

Dummy columns are generated to represent the presence of each ingredient in the dataset, which is displayed as numerical values, 0 and 1.

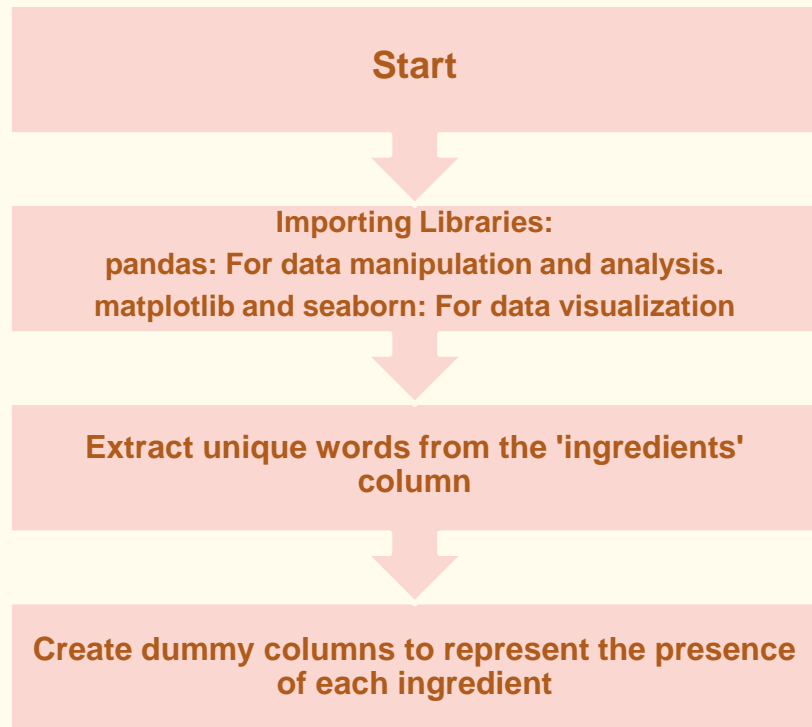


x x x x

# Data Wrangling

key	name	description	rating	rating_count	Ingredients	PUMPKIN JUICE CONCENTRATE	COCONUT EXTRACT	BUTTER OIL	EGG YOLKS	...	TOFFEE	MILK CHOCOLATE AND VEGETABLE OIL COATING
0_0_hd	White Chocolate Raspberry Truffle Ice Cream	A truly exquisite ice cream inspired by fine c...	4.9	168	CREAM, SKIM MILK, SUGAR, RASPBERRY PUREE, LACT...	0	0	0	1	...	0	0
1_1_hd	Banana Peanut Butter Chip Ice Cream	Ribbons of rich peanut butter and bits of choc...	4.7	80	CREAM, SKIM MILK, SUGAR, PEANUTS, BANANA PUREE...	0	0	1	1	...	0	0
2_2_hd	Bourbon Praline Pecan Ice Cream	Treat yourself to ice cream infused with smoot...	4.1	191	CREAM, SKIM MILK, SUGAR, BROWN SUGAR, EGG YOLK...	0	0	0	1	...	0	0

Dummy columns are generated to represent the presence of each ingredient in the dataset, which is displayed as numerical values, 0 and 1.



x x x x

# EDA with DATA VISUALIZATION

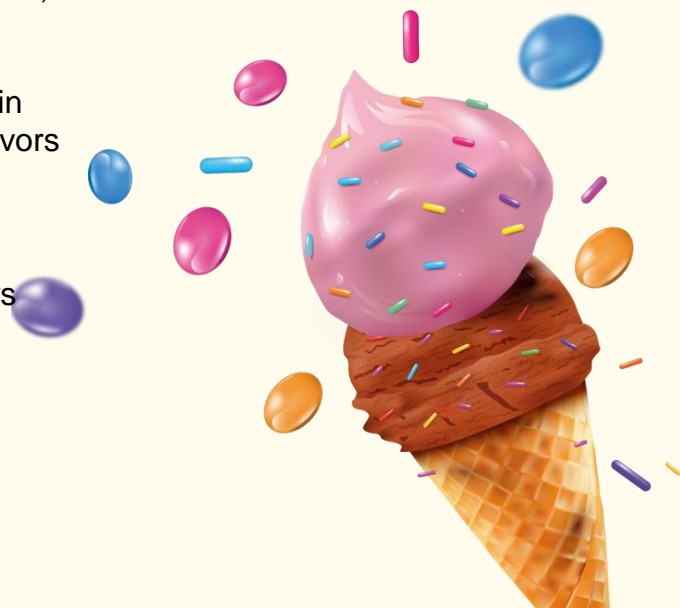
---

**Line Chart:** This chart is used to display the elbow method for finding the optimal K value, which determines the number of clusters. The line chart also visualizes the Silhouette score in relation to the number of clusters for evaluating the quality of clustering.

**Scatter Plot:** Scatter plots are used to visualize the K-means clusters of each ingredient and show the relationships between variables, including `ingredient_cluster`, `rating_count`, and `rating`. Scatter plots are effective for visualizing clusters after performing KNN.

**Box Plots:** These plots are used to visualize thresholds for ratings and rating counts in order to determine which cluster has the most influence on high ratings. Ice cream flavors that meet these criteria are identified as top performers.

**Bar Chart:** This chart is used to visualize the results of ice cream names and their ingredients' influence on ratings and rating counts. It displays the top ice cream flavors with the highest rating scores, indicating their significance.



x x x x



# K-Nearest Neighbors (KNN)

x x x x

# Summary of Clustering using K-Mean

**The K-Nearest Neighbors (KNN)** algorithm is utilized to group ice cream flavors into clusters based on the similarity of their ingredient profiles. The elbow method is employed to determine the optimal number of clusters (K) using Within-Cluster Sum of Squares (WCSS) as a metric.

Silhouette scores are calculated for different K values, and the optimal K is determined to be 3 for Häagen-Dazs flavors.

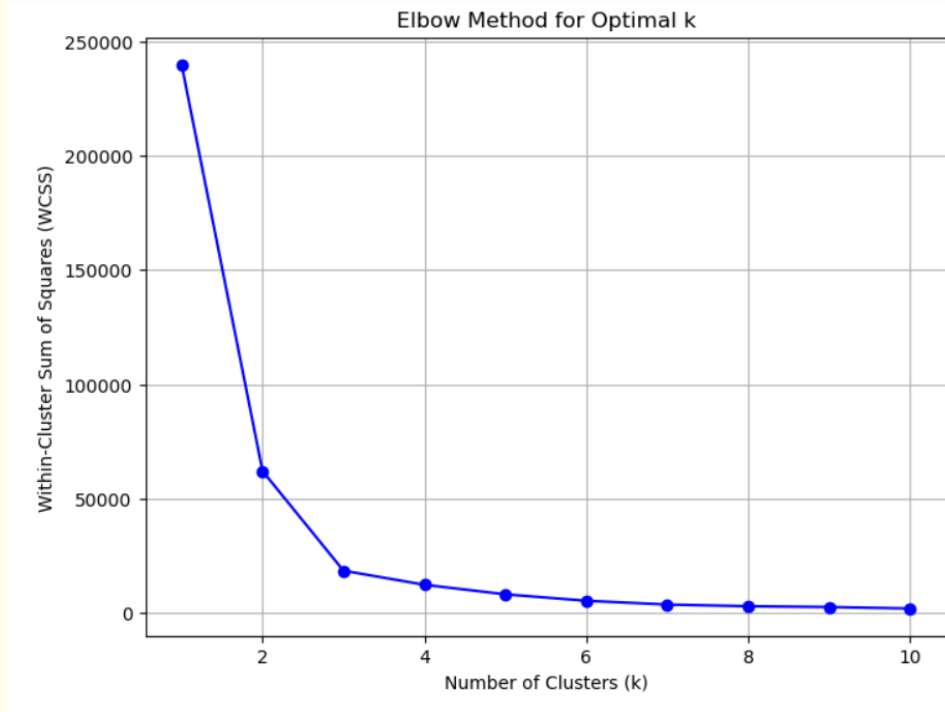
The project then analyzes the clusters created for Häagen-Dazs, where cluster 1 has 43 items, cluster 0 has 18 items, and cluster 2 has 9 items.

Cluster 1, which contains 43 Häagen-Dazs ice cream flavors, may represent a group of flavors that share common key ingredients contributing to high ratings.

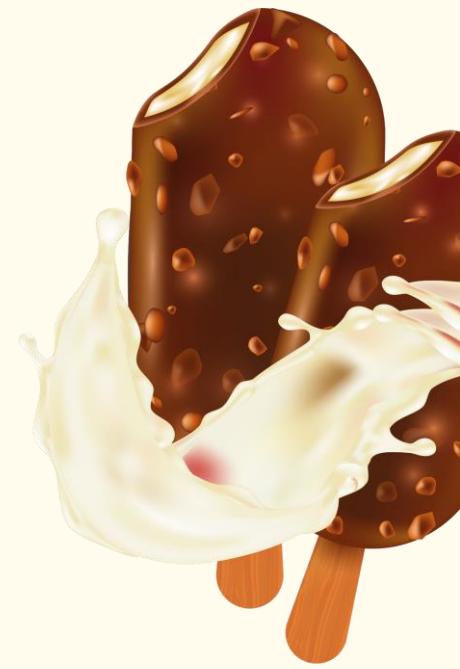
Cluster 0, with 18 flavors, and Cluster 2, with 9 flavors, likely have their own distinct ingredient profiles that also influence ratings.



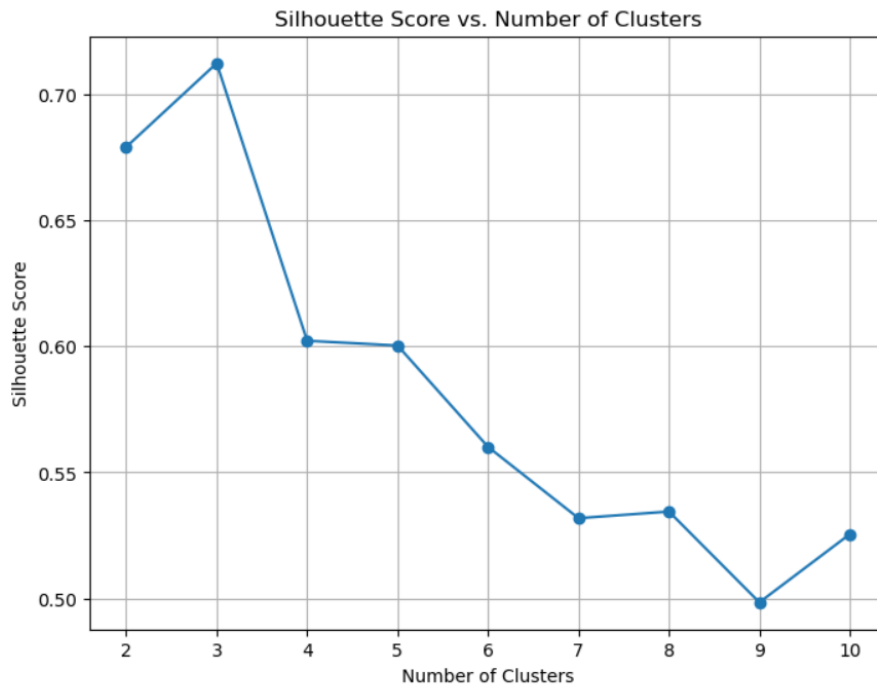
# Summary of Clustering using K-Mean



The elbow method is employed to determine the optimal number of clusters (K) using Within-Cluster Sum of Squares (WCSS) as a metric.

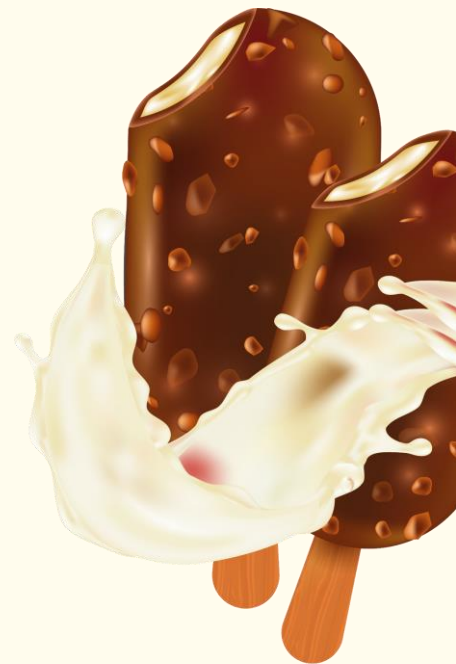


# Summary of Clustering using K-Mean



**Silhouette scores are calculated for different K values, and the optimal K is determined to be 3.**

Higher Silhouette scores indicate that the clusters are well-defined and that data points within each cluster are more similar to each other than to those in other clusters, in this case the silhouette scores are above 0.7 at number of clusters equal 3, ensuring the validity of the analysis.





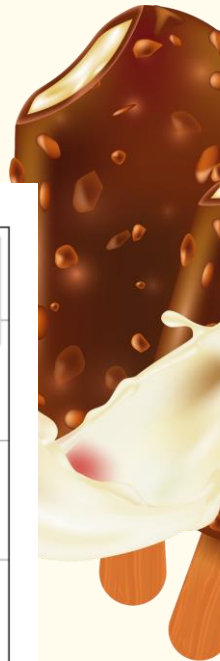
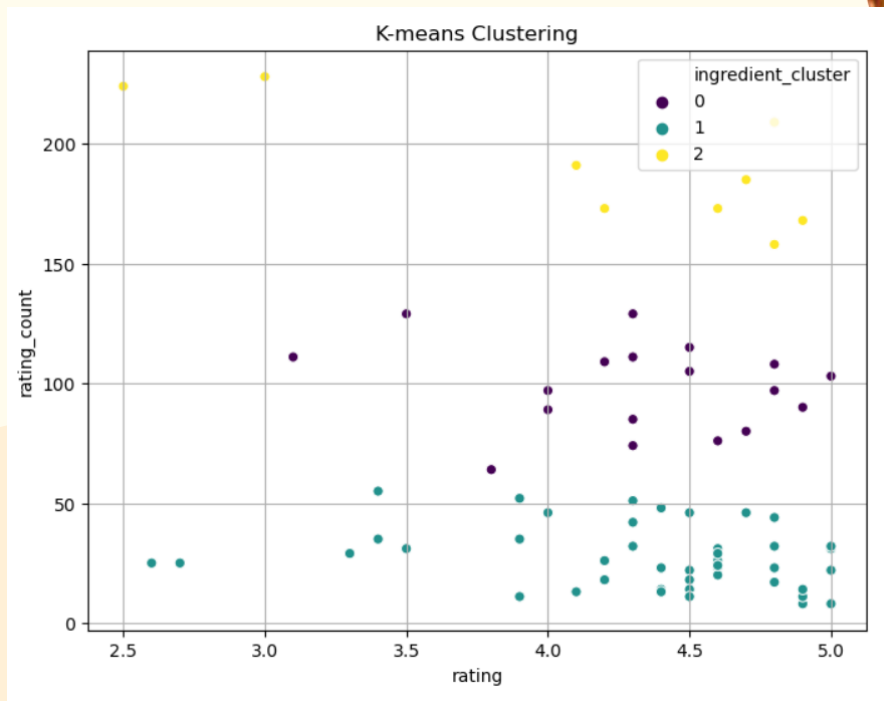
# Summary of Clustering using K-Mean

```
In [10]: cluster_counts = df['ingredient_cluster'].value_counts()
print(cluster_counts)
```

```
1    43
0    18
2     9
Name: ingredient_cluster, dtype: int64
```

The project then analyzes the clusters created, where **cluster 1 has 43 items**, **cluster 0 has 18 items**, and **cluster 2 has 9 items**.

Relationships between clusters and ice cream ratings, rating counts, and ingredients are explored using scatterplots.



x x x x



# Insight Draw From EDA

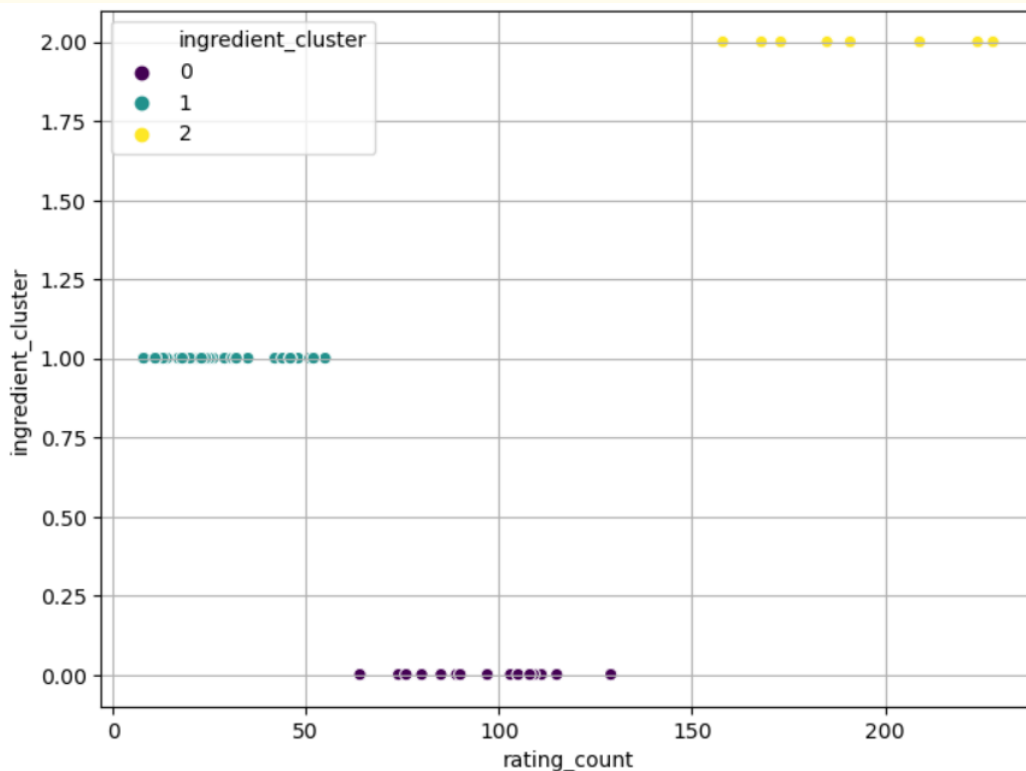
x x x x

# Insight Draw From EDA



## Ingredient\_Cluster VS Rating\_Count

We can see that ingredient\_cluster 2 has a higher rating\_count, approximately in the range of 150-200, while cluster 0 has values above 50 but does not exceed 150 in rating\_count, and cluster 1 has the lowest rating count.

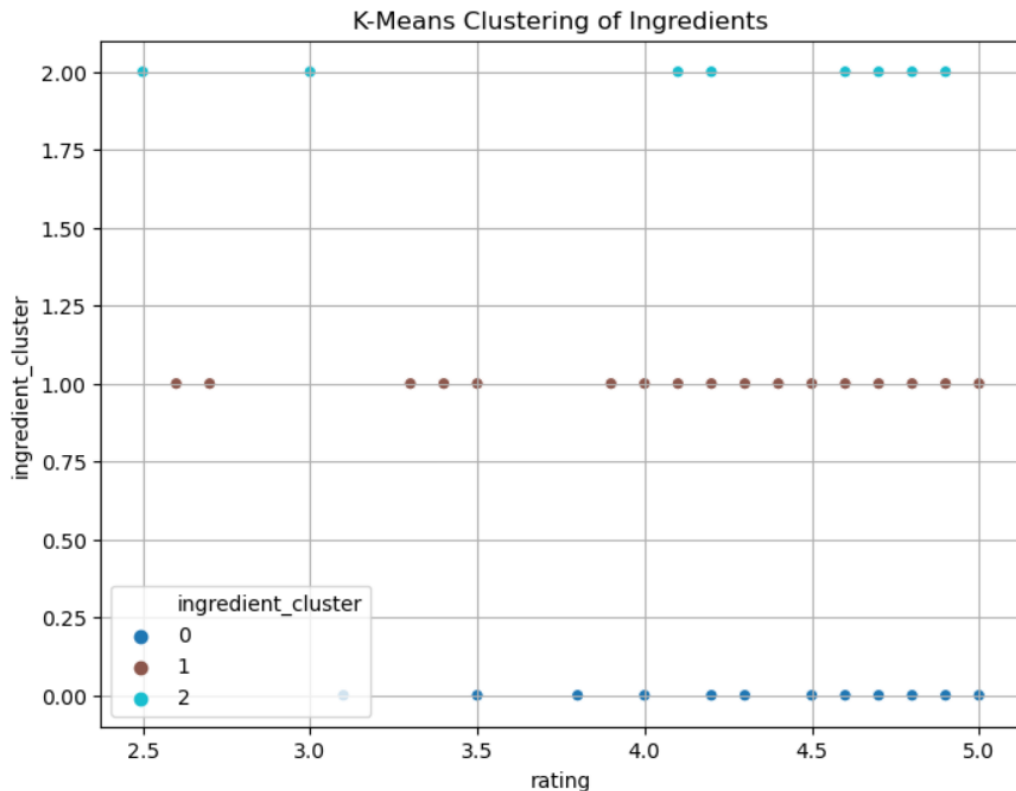


# Insight Draw From EDA



## Ingredient\_Cluster VS Rating

Most of ingredient\_cluster 2 falls within the range of ratings from 4.0 to 5.0.  
Cluster 1 falls within the range of ratings from 3.0 to 5.0, and cluster 0 falls within the range of ratings from 3.5 to 5.0.



# Influence on Rating

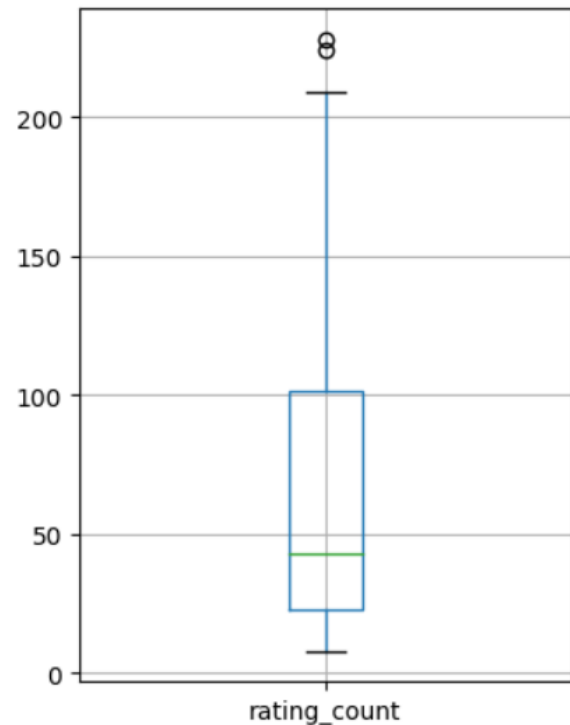
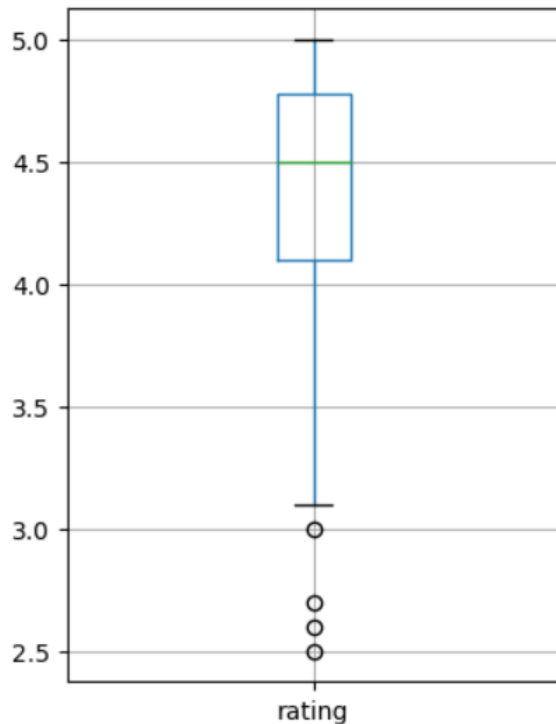


We calculate statistics to find the mean for both rating and rating\_count, and identify high ratings by calculating percentiles at 0.75.

Then, we use a box plot to display the high ratings.

**75th Percentile Rating: 4.775**

**75th Percentile Rating Count: 101.5**



# Influence on Rating



Now we can set thresholds for ratings and rating counts (4.7 and 101.5, respectively) to determine which Häagen-Dazs flavor cluster has the most influence on high ratings.

**Rows that meet the criteria**  
**rating > 4.7**  
**rating\_count >= 101.5**

are selected to identify Häagen-Dazs ice cream flavors that perform exceptionally well.

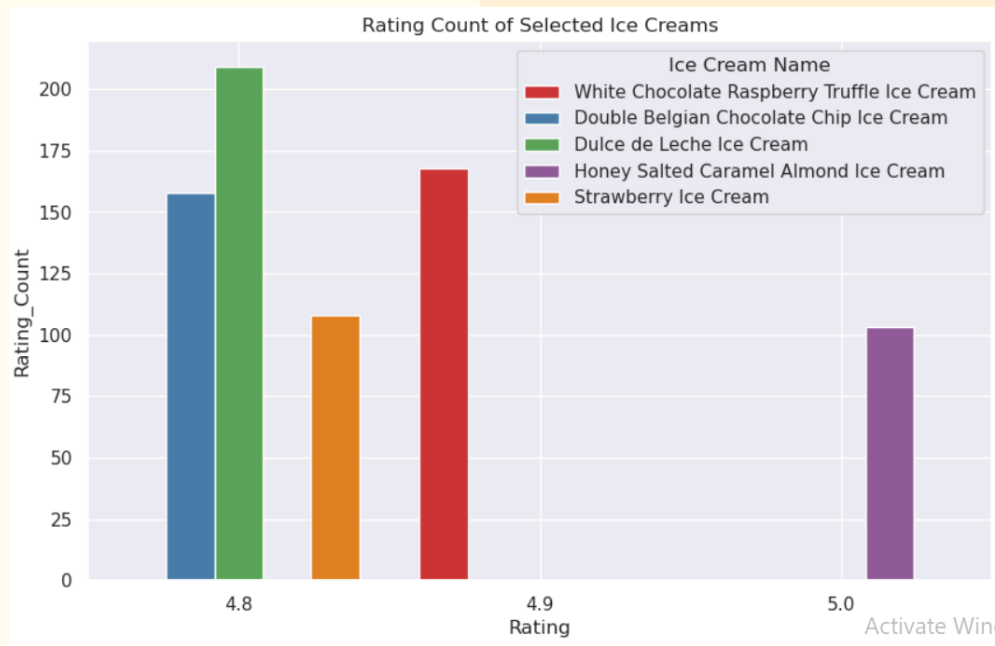
	icecream_name	rating	rating_count	description	ingredient_cluster
0	White Chocolate Raspberry Truffle Ice Cream	4.9	168	A truly exquisite ice cream inspired by fine chocolate truffles. We begin with pure white chocolate ice cream and swirl in satisfying chunks of chocolaty fudge truffles and a tangy raspberry ribbon.	2
27	Double Belgian Chocolate Chip Ice Cream	4.8	158	Your favorite Belgian chocolate ice cream, renamed. Our Belgian chocolate combines rich, velvety chocolate ice cream with finely shaved Belgian chocolate for a uniquely textured experience.	2
29	Dulce de Leche Ice Cream	4.8	209	Inspired by Latin America's treasured dessert, our dulce de leche ice cream is a delicious combination of caramel and sweet cream, swirled with ribbons of golden caramel.	2
31	Honey Salted Caramel Almond Ice Cream	5.0	103	This decadent blend of honey ice cream, swirls of salted caramel, and crunchy toasted almonds was lovingly crafted to raise awareness for the honeybees and other pollinators who bring us so many of our favorite ingredients.	0
55	Strawberry Ice Cream	4.8	108	We introduce sweet summer strawberries to pure cream and other natural ingredients. Because it's brimming with real fruit, the true flavor of our strawberries comes shining through.	0





x x x x

# Summary of Result



**Cluster 0** includes flavors like "Honey Salted Caramel Almond" and "Strawberry Ice Cream," The "Honey Salted Caramel Almond Ice Cream" stands out with a perfect rating of 5.0.

**Cluster 2** includes flavors like "White Chocolate Raspberry Truffle," "Double Belgian Chocolate Chip," and "Dulce de Leche," which also receive high ratings and are likely characterized by their unique ingredient combinations.

x x x x



× × × ×

# Summary of Result

Based on the project's analysis, the following Häagen-Dazs ice cream flavors and their respective clusters, ratings, and descriptions have been identified

## Cluster 0 (Ingredients Profile)

### Honey Salted Caramel Almond Ice Cream

Rating: 5.0

Rating Count: 103

**Description:** This flavor is a decadent blend of honey ice cream, swirls of salted caramel, and crunchy toasted almonds, created to raise awareness for honeybees and other pollinators.



### Strawberry Ice Cream

Rating: 4.8

Rating Count: 108

**Description:** This flavor features sweet summer strawberries mixed with pure cream and other natural ingredients, offering the true flavor of strawberries.



× × × ×





× × × ×

# Summary of Result

## Cluster 2 (Ingredients Profile)

### **White Chocolate Raspberry Truffle Ice Cream**

**Rating:** 4.9

**Rating Count:** 168

**Description:** This flavor is described as an exquisite ice cream inspired by fine chocolate truffles. It features pure white chocolate ice cream, chunks of chocolaty fudge truffles, and a tangy raspberry ribbon.



### **Double Belgian Chocolate Chip Ice Cream**

**Rating:** 4.8

**Rating Count:** 158

**Description:** This is described as a favorite Belgian chocolate ice cream, combining rich, velvety chocolate ice cream with finely shaved Belgian chocolate for a unique textured experience.



× × × ×



x x x x

# Summary of Result

## Cluster 2 (Ingredients Profile)

### Dulce de Leche Ice Cream

**Rating:** 4.8

**Rating Count:** 209

**Description:** This flavor is inspired by Latin America's treasured dessert, featuring a delicious combination of caramel and sweet cream, swirled with ribbons of golden caramel.



In summary, the project successfully addressed all three objectives by identifying clusters of ingredients that influence ratings, revealing patterns and associations between flavors, and exploring the relationships between ingredient clusters, ratings, and rating counts to understand what makes a Häagen-Dazs ice cream flavor popular.

x x x x



# Conclusion

## Model Interpretation:

The clustering model effectively grouped Häagen-Dazs ice cream flavors into two clusters, highlighting distinct characteristics within these clusters.

**Cluster 0** seems to be characterized by flavors that have a niche appeal, resulting in high ratings but lower rating counts.

**Cluster 2** includes flavors that appeal to a broader audience, as they have high ratings and significantly higher rating counts.

## Business Interpretation:

Häagen-Dazs can leverage these insights for product management and marketing strategies:

**Cluster 0:** Häagen-Dazs may consider maintaining and promoting these flavors to cater to a niche, discerning audience.

These flavors may appeal to those seeking unique and distinct ice cream experiences.

**Cluster 2:** Flavors in this cluster appear to have broader appeal and receive high ratings. These flavors could serve as flagship products and be promoted more extensively to target a wider customer base.



# Problems/Solution



Determine the key ingredient(s) that contribute to high ratings in ice cream.

The project analyzed ice cream flavors in different clusters based on their ingredient profiles. **Cluster 0 includes flavors like "Honey Salted Caramel Almond" and "Strawberry Ice Cream"**. These flavors are characterized by specific ingredients that likely contribute to their high ratings.

Cluster ingredients to identify patterns and associations between different ice cream flavors.

The project successfully used the **K-Nearest Neighbors (KNN) algorithm** to cluster ice cream flavors based on their ingredient profiles. **Cluster 0 and Cluster 2** were identified as distinct groups with varying levels of popularity and high ratings. The clustering analysis revealed patterns and associations between different ice cream flavors. Häagen-Dazs can categorize its products based on ingredient similarity.



# Problems/Solution



Explore the relationship between ingredient clusters, ice cream ratings, and rating counts to gain insights into what makes an ice cream flavor popular.

It identified that **Cluster 0** contains flavors with high ratings but lower rating counts, whereas **Cluster 2** includes flavors with consistently high ratings and significantly higher rating counts.

This exploration provided insights into what makes an ice cream flavor popular: **Cluster 0** is characterized by niche appeal and unique ingredient combinations that appeal to a specific audience.

**Cluster 2** includes flavors with broader market appeal, contributing to higher ratings and more substantial rating counts.

Additionally, Häagen-Dazs could explore creating new flavors with ingredient profiles similar to those in Cluster 2.



x x x x



# Recommendation

Häagen-Dazs can focus on promoting and further developing flavors within Cluster 0, as these already have high ratings and could have broad appeal. It can be marketed as a premium product.

- **Differentiation:** By understanding the unique appeal of Cluster 0 flavors, the brand can stand out in the market by offering distinctive, high-quality ice cream flavors.
- **Broad Appeal:** Cluster 2 flavors can serve as core offerings with broad market appeal, attracting a larger customer base.

**Innovation:** The brand can consider developing new flavors inspired by the ingredient profiles found in Cluster 2 to cater to the preferences of a larger customer base. These innovations may also contribute to maintaining high ratings.





# Appendix

**Source of Data** The dataset used for this project was sourced from Kaggle and is publicly available at the following link:

[Ice Cream Dataset on Kaggle](#)

This dataset played a pivotal role in the project, enabling data-driven analysis and findings related to the influence of ingredients on Häagen-Dazs ice cream flavors and ratings.

**Code Repositories:** GitHub repository with the Python scripts for data preprocessing and modeling

**Python Libraries:** Various Python libraries were employed for data manipulation, analysis, and visualization, including but not limited to Pandas, Matplotlib, Seaborn, and Scikit-Learn.

**K-Nearest Neighbors (KNN) Clustering:** The KNN algorithm was used for clustering analysis, enabling the grouping of ice cream flavors based on ingredient similarity. Code for KNN implementation, including elbow method and Silhouette score calculations, was a crucial project asset.

**Data Visualization Tools:** Data visualization assets included code for generating line charts, box plots, bar charts, and scatter plots to visualize the project's findings effectively.







**THANKS!**