**ALEXANDRIA LOW: REPORT ON CAPSTONE PROJECT**

**PROBLEM STATEMENT**:

**Identifying opportunities for growth. Such as Increasing market share. Opportunity can be evidenced.**

As per external market research, the spices and savoury segment is registering immense growth within the food flavours market. This trend is largely driven by consumers' growing interest in multicultural cuisines and adventurous eating experiences.

(Fortune Business Insights, <https://www.fortunebusinessinsights.com/food-flavors-market-102745>).

The Asia Pacific Food Flavours Market Size is expected to have a steady increase up till 2028. Currently the market size stands at more than USD 5m.

(6WResearcg, <https://www.6wresearch.com/industry-report/malaysia-food-flavors-market>).

'Increasing application of food flavours across the packaged food industry... will accelerate market growth...’

(GMI Research, <https://www.gmiresearch.com/report/food-flavors-market>[/](https://www.gmiresearch.com/report/food-flavors-market/)).

'Various multinational giants are mainly focusing on the development of exotic and unique delicious food flavours, to attract a large number of consumers towards food flavours. Strong demand for convenient food products will create growth opportunities for food flavouring industry.’

(GMI Research, <https://www.gmiresearch.com/report/food-flavors-market>[/](https://www.gmiresearch.com/report/food-flavors-market/)).

1**. An executive summary:**

**What is your goal?**

Predicting food flavour trends for product development and innovation for market share growth whilst maintaining competitiveness in a growing and fast-moving industry. This exercise is to ultimately provide some suggestions within our Company on which direction to take in terms of product development and innovation to create new products that will lead to increasing market share and maximizing profits.

**Where did you get your data?**

Data has been obtained from Kaggle.

**What are your metrics?**

Metrics used are Silhouette score.

Flavours are represented by ingredients, and cuisine represents the influence of the flavours.

The highest average rating was used to determine which Cluster to use.

**What were your findings?**

Findings were the 10 most popular ingredients and cuisines in Cluster 3. This is based on its average rating of 5.0 .

**What risks/limitations/assumptions affect these findings?**

The risk affecting the findings would be the clustering algorithm, parameters and number of clusters used

2. **Summarize your statistical analysis, including:**

**Implementation**

Clustering Kmeans was implemented on our data to identify groups of recipes with similar ingredient compositions, allowing for the identification of common flavour combinations or culinary themes. 3 clusters were made.

**Evaluation**

The results from our method resulted in the identification of popular ingredients and cuisines based on highest average ratings.

**Inference**

As there are many ingredients identified from recipes, these different ingredients can be combined with each other to produce a plethora of different flavours, with the influence of different cuisine styles.

3.**Clearly document and label each section of your notebook(s)**

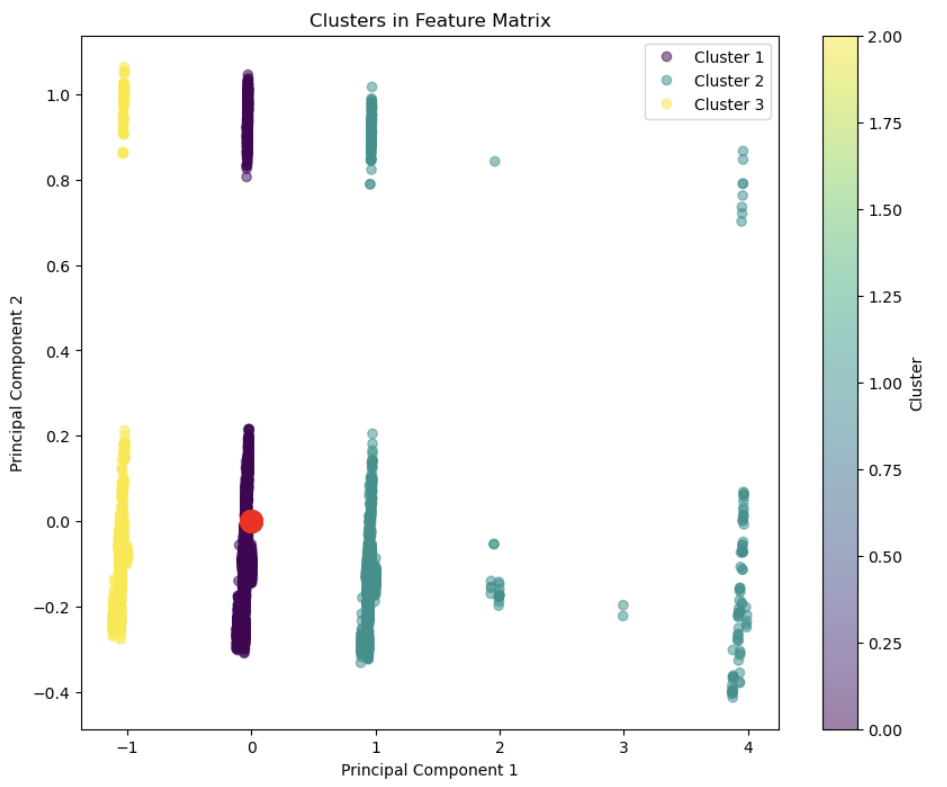
Logically organize your information in a persuasive, informative manner.

Include notebook headers and subheaders, as well as clearly formatted

markdown for all written components.

The notebook shall contain headers, subheaders, and comments following a logical flow from the start until the end of the project.

Include graphs/plots/visualizations with clear labels.

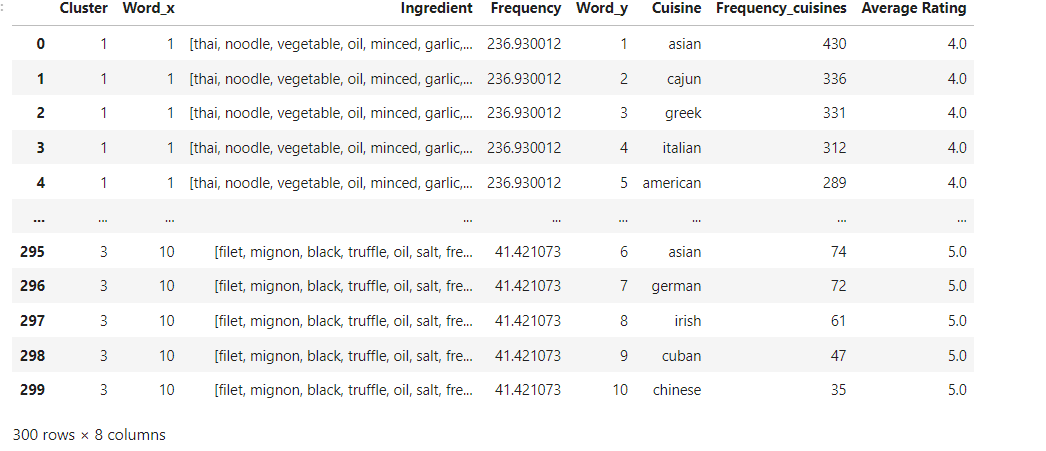


*Fig 1: Cluster Plot.*

PC1:

Captures it is the direction along which there is the greatest variance in the data, the most significant pattern in the dataset, accounting for the largest part of the data’s total variation.

PC2: Captures the next level of variation that was not captured by PC1.



*Fig2: Top 10 ingredients and cuisines with the highest frequency.*

***Findings sorted by Cluster***

|  |  |  |
| --- | --- | --- |
| ***Cluster 1 Most popular ingredients*** | |  |
| *No.* | ***Ingredients*** | ***Frequency*** |
| *1* | *'thai', 'noodle', 'vegetable', 'oil', 'minced', 'garlic', 'sweet', 'paprika', 'fish', 'sauce', 'white', 'vinegar', 'sugar', 'oil', 'leek', 'tofu', 'egg', 'scallion', 'beansprouts', 'peanut', 'lime'* | *237* |
| *2* | *'red', 'bell', 'pepper', 'olive', 'oil', 'mushroom', 'ginger', 'garlic', 'lemon', 'grass', 'sambal', 'ulek', 'chicken', 'broth', 'coconut', 'milk', 'sugar', 'soy', 'sauce'* | *199* |
| *3* | *''chicken', 'liver', 'dark', 'soy', 'sauce', 'mirin', 'sake', 'brown', 'sugar', 'garlic', 'clove', 'fresh', 'ginger', 'scallion'* | *197* |
| *4* | *'fresh', 'corn', 'pepper', 'butter', 'salt'* | *191* |
| *5* | *'sushi', 'rice', 'avocado', 'crabmeat', 'seed', 'nori', 'toasted', 'sesame', 'seed', 'wasabi', 'soy', 'sauce'* | *182* |
| *6* | *'water', 'white', 'sugar', 'grated', 'lemon', 'zest', 'lemon', 'juice', 'strawberry', 'carbonated', 'water', 'ice', 'mint'* | *175* |
| *7* | *'olive', 'oil', 'egg', 'rice', 'dark', 'sesame', 'oil', 'fresh', 'spinach', 'leaf', 'avocado', 'toasted', 'sesame', 'seed'* | *159* |
| *8* | *'chicken', 'breast', 'kielbasa', 'diced', 'tomato', 'onion', 'celery', 'green', 'bell', 'pepper', 'garlic', 'cajun', 'seasoning', 'cayenne', 'pepper', 'broth', 'dried', 'oregano', 'dried', 'parsley', 'dried', 'thyme', 'cooked', 'rice'* | *157* |
| *9* | *'red', 'bliss', 'potato', 'olive', 'oil', 'sea', 'salt', 'freshly', 'ground', 'pepper', 'fresh', 'rosemary'* | *153* |
| *10* | *'water', 'sugar', 'lime', 'rind', 'fresh', 'lime', 'juice', 'ice', 'water'* | *143* |

***Table 1. Popular ingredients Cluster***

|  |  |  |
| --- | --- | --- |
| ***Cluster 1 Most popular cuisine*** | | |
| ***No.*** | ***Cuisine*** | ***Frequency*** |
| *1* | *American* | *222* |
| *2* | *French* | *88* |
| *3* | *Italian* | *80* |
| *4* | *Cajun* | *77* |
| *5* | *Greek* | *74* |
| *6* | *Asian* | *74* |
| *7* | *German* | *72* |
| *8* | *Irish* | *61* |
| *9* | *Cuban* | *47* |
| *10* | *Chinese* | *35* |

***Table 2. Popular cuisines Cluster 1***

|  |  |  |
| --- | --- | --- |
| **Cluster 2 Most popular ingredients** | |  |
| **No.** | **Ingredients** | **Frequency** |
| **1** | **'thai', 'noodle', 'vegetable', 'oil', 'minced', 'garlic', 'sweet', 'paprika', 'fish', 'sauce', 'white', 'vinegar', 'sugar', 'oil', 'leek', 'tofu', 'egg', 'scallion', 'beansprouts', 'peanut', 'lime'** | **71** |
| **2** | **'red', 'bell', 'pepper', 'olive', 'oil', 'mushroom', 'ginger', 'garlic', 'lemon', 'grass', 'sambal', 'ulek', 'chicken', 'broth', 'coconut', 'milk', 'sugar', 'soy', 'sauce'** | **64** |
| **3** | **'chicken', 'liver', 'dark', 'soy', 'sauce', 'mirin', 'sake', 'brown', 'sugar', 'garlic', 'clove', 'fresh', 'ginger', 'scallion'** | **60** |
| **4** | **'sushi', 'rice', 'avocado', 'crabmeat', 'seed', 'nori', 'toasted', 'sesame', 'seed', 'wasabi', 'soy', 'sauce'** | **54** |
| **5** | **'olive', 'oil', 'egg', 'rice', 'dark', 'sesame', 'oil', 'fresh', 'spinach', 'leaf', 'avocado', 'toasted', 'sesame', 'seed'** | **53** |
| **6** | **'water', 'white', 'sugar', 'grated', 'lemon', 'zest', 'lemon', 'juice', 'strawberry', 'carbonated', 'water', 'ice', 'mint'** | **51** |
| **7** | **'fresh', 'corn', 'pepper', 'butter', 'salt'** | **49** |
| **8** | **'red', 'bliss', 'potato', 'olive', 'oil', 'sea', 'salt', 'freshly', 'ground', 'pepper', 'fresh', 'rosemary'** | **46** |
| **9** | **'water', 'sugar', 'lime', 'rind', 'fresh', 'lime', 'juice', 'ice', 'water'** | **45** |
| **10** | **'chicken', 'breast', 'kielbasa', 'diced', 'tomato', 'onion', 'celery', 'green', 'bell', 'pepper', 'garlic', 'cajun', 'seasoning', 'cayenne', 'pepper', 'broth', 'dried', 'oregano', 'dried', 'parsley', 'dried', 'thyme', 'cooked', 'rice'** | **41** |

***Table 3. Popular ingredients Cluster 2***

|  |  |  |
| --- | --- | --- |
| **Cluster 2 Most popular cuisine** | | |
| **No.** | **Cuisine** | **Frequency** |
| 1 | Hungarian | 140 |
| 2 | Asian | 115 |
| 3 | Cajun | 90 |
| 4 | Indian | 84 |
| 5 | Greek | 83 |
| 6 | Italian | 71 |
| 7 | French | 65 |
| 8 | Chinese | 58 |
| 9 | American | 55 |
| 10 | Cuban | 54 |

***Table 4. Popular cuisines Cluster 2***

|  |  |  |
| --- | --- | --- |
| **Cluster 3 Most popular ingredients** | |  |
| **No.** | **Ingredients** | **Frequency** |
| **1** | **chicken', 'liver', 'dark', 'soy', 'sauce', 'mirin', 'sake', 'brown', 'sugar', 'garlic', 'clove', 'fresh', 'ginger', 'scallion'** | **63** |
| **2** | **'chicken', 'breast', 'kielbasa', 'diced', 'tomato', 'onion', 'celery', 'green', 'bell', 'pepper', 'garlic', 'cajun', 'seasoning', 'cayenne', 'pepper', 'broth', 'dried', 'oregano', 'dried', 'parsley', 'dried', 'thyme', 'cooked', 'rice'** | **62** |
| **3** | **'thai', 'noodle', 'vegetable', 'oil', 'minced', 'garlic', 'sweet', 'paprika', 'fish', 'sauce', 'white', 'vinegar', 'sugar', 'oil', 'leek', 'tofu', 'egg', 'scallion', 'beansprouts', 'peanut', 'lime'** | **59** |
| **4** | **'red', 'bell', 'pepper', 'olive', 'oil', 'mushroom', 'ginger', 'garlic', 'lemon', 'grass', 'sambal', 'ulek', 'chicken', 'broth', 'coconut', 'milk', 'sugar', 'soy', 'sauce'** | **52** |
| **5** | **'extra-virgin', 'olive', 'oil', 'sweet', 'onion', 'balsamic', 'vinegar', 'chicken', 'breast', 'garlic', 'salt', 'pepper', 'chicken', 'broth', 'honey'** | **51** |
| **6** | **'olive', 'oil', 'egg', 'rice', 'dark', 'sesame', 'oil', 'fresh', 'spinach', 'leaf', 'avocado', 'toasted', 'sesame', 'seed'** | **47** |
| **7** | **'water', 'sugar', 'fresh', 'ginger', 'fresh', 'lime', 'juice', 'ginger', 'ale'** | **45** |
| **8** | **'red', 'bliss', 'potato', 'olive', 'oil', 'sea', 'salt', 'freshly', 'ground', 'pepper', 'fresh', 'rosemary'** | **43** |
| **9** | **'vegetable', 'oil', 'boneless', 'skinless', 'chicken', 'breast', 'andouille', 'sausage', 'shrimp', 'creole', 'seasoning', 'dri', 'leav', 'thyme', 'cayenne', 'pepper', 'onion', 'garlic', 'green', 'bell', 'pepper', 'celery', 'long', 'grain', 'white', 'rice', 'canned', 'chopped', 'tomato', 'chicken', 'stock', 'fresh', 'parsley'** | **41** |
| **10** | **'filet', 'mignon', 'black', 'truffle', 'oil', 'salt', 'freshly', 'ground', 'pepper', 'smoked', 'bacon'** | **41** |

***Table 5. Popular ingredients Cluster 3***

|  |  |  |
| --- | --- | --- |
| **Cluster 3 Most popular cuisine** | | |
| **No.** | **Cuisine** | **Frequency** |
| 1 | American | 222 |
| 2 | French | 88 |
| 3 | Italian | 80 |
| 4 | Cajun | 77 |
| 5 | Greek | 74 |
| 6 | Asian | 74 |
| 7 | German | 72 |
| 8 | Irish | 61 |
| 9 | Cuban | 47 |
| 10 | Chinese | 35 |

***Table 6. Popular cuisines Cluster 3***

**Comment and explain the purpose of each major section/subsection of your code.**

**Import the CSV into a pandas DataFrame.**

**Data cleaning and pre-processing:**

Tokenization

Lowercasing

Removing punctuations

Removing stopwords

Lemmatization

**Clustering and grouping of ingredients:**

Vectorization Using TD-IDF to transform document into an array (sparse matrix)

And converted into a dense array

**Convert ‘rating’ into an array:**

To be used later for clustering

**Cleaning of ‘cuisine’ column in recipe dataframe:**

To be used later for clustering after performing one-hot array.

Concatenation of the 3 pre-processed arrays into a single feature matrix.

Apply clustering algorithm to our feature matrix

Elbow method:

To determine to optimal number of clusters to use for clustering

Apply Kmeans clustering, compute centroid, perform PCA and create a dictionary to map cluster labels to map ‘legend’ to and plot the cluster plot.

Calculate the highest average rating for each plot.

Calculate the top 10 cuisine frequency

Calculate the top 10 ingredients frequency

Merge the dataframes

**Model Evaluation:**

Silhouette score was calculated to evaluate the performance of the clustering algorithm

**Document your code for your future self, as if another person needed to replicate your approach**

Refer to Appendix at the back of this Report.

4. **Clearly document all of your decision points in the relevant sections**

**How did you acquire your data?**

Data was acquired from Kaggle. A recipe dataset.

**How did you transform or engineer your data? Why?**

I have transformed the columns that I will need to work with and dropped columns which I am not using.

Tokenization, lower-casing, removing stopwords and punctuations, lemmatization have been performed and stored in a dictionary.

The dictionary was then converted into a list of documents.

Vectorization was performed on the documents to convert it into a dense array of numerical values.

‘cuisine’ column was pre-processed and One-hot encoded to converting into an array.

‘rating’ column was converted into an array.

These arrays were combined into a feature\_matrix for clustering and assigned cluster labels.

**How did you select your model?**

My model was selected based on the kind of project that was chosen. As it falls under unsupervised learning and a classification project, KMeans clustering seemed suitable. Data used is discrete and not continuous. The model is expected to demonstrate versatility especially for datasets with non-linear relationships.

Also the number of clusters get to be determined before clustering, by using the Elbow Method.

**How did you optimize hyperparameters?**

The number of clusters were determined by using the Elbow Method to get the optimal number of clusters, which is 3 clusters.

The KMeans random state was set at 42, in order to produce the same output each time.

**5. Host your notebook and any other materials in your own public Github**

**Repository.**

- Your repo should have README file that guides us through the repository and

links to important files.

- Include links and explanations to any outside libraries or source code used.

- Host a copy of your dataset or include a link to a remotely hosted version.

<https://github.com/alexandrialow/First-repository>

APPENDIX











