# Food Places Recommendation System Based on Emotions for College Student

## Team Members

Dominic Ferrando
dferran2@kent.edu

Tram Le
*hle12@kent.edu*

Priyank Shah
*pshah14@kent.edu*

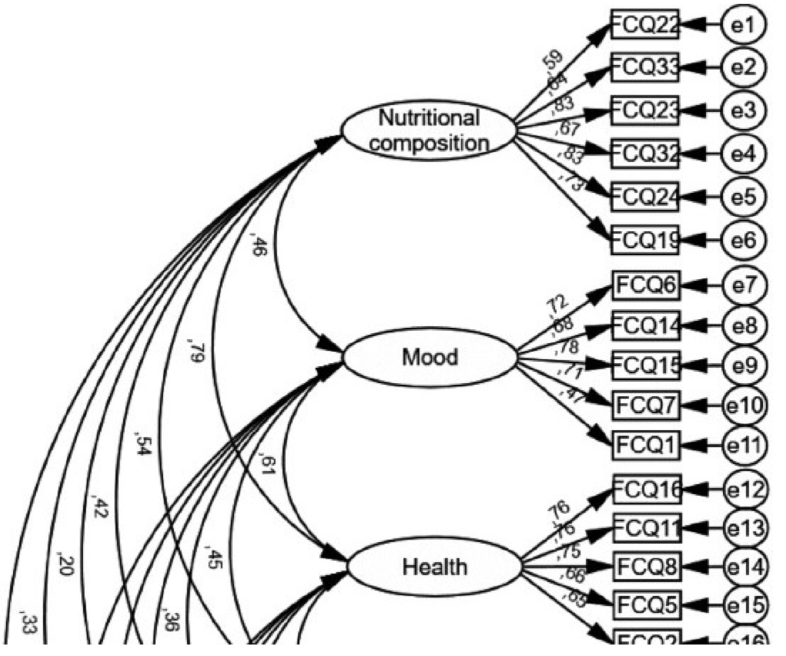
Chad Losey
*closey1@kent.edu*

## Introduction

### Motivations for this project

Nowadays, there are many college students face issues like depression due to having bad eating habits or lack of healthy eating. According to Fobes, there are various reasons related to their depression in their diet, such as being  too busy or unable to afford good food. In 2018, a Nutrients study surveyed 692 undergraduate and graduate students and showed that food insecurity significantly affected GPA. Food-insecure students had an average GPA of 3.33 out of 4.0, while food-secure students had an average GPA of 3.51.

Recognizing the importance of a high connection of nutritional composition, mood and health, especially for college students, our team came up with the idea of linking college student emotions to the food options by using recommendation system.




*Connection between nutritional composition-mood-health*

## Real Applications

We aim to provide a solution to college student emotions, such as depression, stress, laziness, and happiness by building recommendation systems to help boost their moods, and improve their food experiences. Additionally, we recommend restaurant that are connected with those food options in our system, which are nearby campus for individual college students.

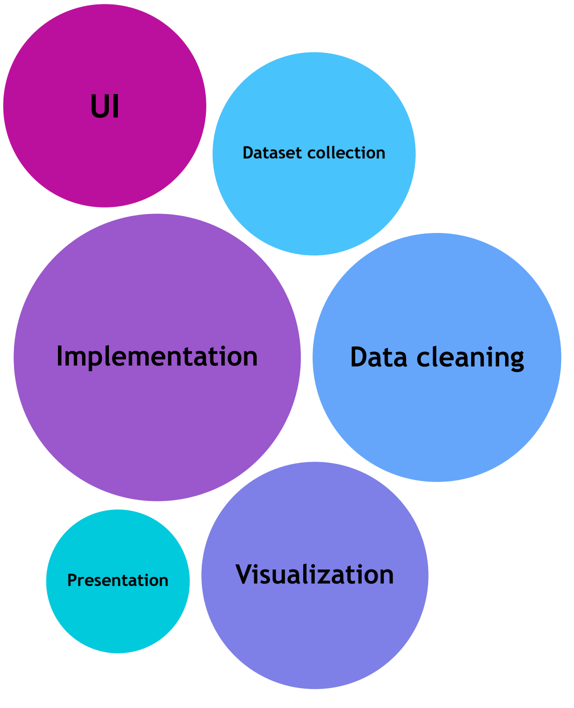
# General Project Description

**Description** We will develop a web based interface along with our recommendation system that allows college students to select a set of emotions(icon) such as happy, stress, bored, tired, etc.., based on their current mood. Then the system will provide suggestions for restaurant that fit food options near campus. We will utilize data mining techniques into the recommendation system, it will help us to select which campus in USA we should analyze on then optimize for the restaurant nearby that campus as top picks for students.

**Challenges** There will be several challenges that our project will be dealing with. We would have to decide which is the most suitable algorithm to implement and integrate into our recommendation system. Based on our dataset, we have limitations on food options that involve emotions. Moreover, it's challenging to find restaurant near by campus using Yelp dataset. Specifically, those are very large dataset, so we will need to learn how to implement those techniques in code in an efficient manner.

**Workload Distribution**

- *Writing report paper, presentation, and video demo*
- *Find and collect the dataset*
- *Data cleaning and processing*
- *Visualizing: plot, graph, chart,...*
- *Implement recommendation system algorithm for project using techniques from class and google*
- *UI web application for better illustration of what the real application would look like*

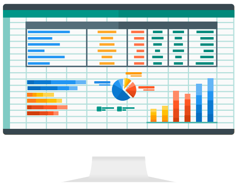


# Background

### Software Tools

Python libraries/framework such as scikit-learn, numpy, pandas to read the dataset, matplotlib to plotting the dataset, HTML+CSS, Python Flask framework for interface, and Github to deploy our application

### Hardware



Personal laptops and computers that can store and download large dataset

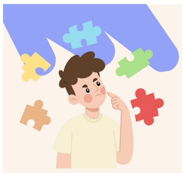
### Skills

Python, Jupyter Notebook, TensorFlow, Data Mining and Algorithm Knowledge, Web Programming

# Problem Definition

## Formal definition

The problem can be defined as



- **Let F = {f1, f2, f3,…,fn} be the set of food items mentioned in survey dataset**
- **Let E = {e1, e2, e3,…en} be the set of emotions like 'happy', 'stress', etc.**
- **Let R = {r1, r2, r3,…rn} be the set of location of restaurants nearby campus**
- **Let M = {m1, m2, m3,…mn} be the set of location of supermarket nearby campus restaurant**
- **Use the given map: E ->P(F) where each emotions in E connect to top 5 of foods in F that already connected in dataset**
- **Learn how to map: F ->P(R), F-> P(M) that maps foods to relevant restaurant/supermarket**
- **Optimize to only close by campus and average of high rating score as threshold**
- **Our goal is to optimize these connections to provide the best food recommendations for a given emotion, and related restaurant suggestions.**

Demo example:

**Given dataset:**

E1 = {sadness, stress, cold weather, not feeling well}

F1 ={chocolate, pasta, soup, chips, popcorn}

E2 = {stress, anger, sadness}

F2 = {mac and cheese, chocolate, pasta}

R1 = {cuisine(cheeseburger, pasta, soda, chocolate), location(longitude, latitude), rating()}

M1 = {food_item(veggie, mac and cheese, pasta, soup, chocolate), location(longitude, latitude)}

**Food outcome:**

Mood(stress) = {chocolate, pasta}

Mood(sadness) = {mac and cheese, soup}

**Connect food and places outcome:**

Food({chocolate, pasta}) = {R1, M1}

Food({mac and cheese, soup}) = {M1}

**Places with location nearby campus and high rating outcome:**

Location({R1, M1}) = {(school_longtitude, school_latitude), (average_rating)}

## Challenges of tackling the problem

**Dataset problems**

- The ‘food_choices’ dataset we have is a survey conducted in college, so it may not cover all possible foods-emotions combinations. Most of the foods in the survey are snacks and considered unhealthy, so it becomes quite challenging to identify good foods that are associated with cuisine of a restaurant and supermarkets.
- Cleaning and processing the datasets is also complicated because we only need necessary attributes (‘comfort_food_reasons','comfort_food') and change labels to Emotion and Food Options.
- Reading a very large dataset from yelp, specially *‘yelp_business.json’* that is over 4gb, is also challenging. We will manage to convert the dataset into .cvs file

**Algorithm selection**

- We must decide on the algorithms that are most suitable to our application. The algorithm must be able to handle the complexity of emotions and the variety of food.
- The biggest challenge in our project is how to implement location-based recommendation. This involves filtering only locations nearby Kent campus and user ratings, and dealing with a potentially large dataset of locations.

## General Solution

There are few techniques, considering our project idea, you could explore sentiment analysis on the comfort food reasons. We could consider NLTK Stopword, Lemmatizer along with BERTopic techniques to analysis to understand the emotions associated with specific food choices.

For restaurant location, we could narrow down our experiment data in small area around campus, by calculate the distance between two data points and set a threshold close to the campus

# The Proposed Techniques

## Cleaning the Data for Food Dataset

The initial step involves meticulously cleaning the data within the food dataset. This process ensures that the dataset is free from inconsistencies and inaccuracies, allowing more reliable analysis and insights. This dataset was provided by Yelp.

| | GPA | Gender | breakfast | calories_chicken | calories_day | calories_scone | coffee | comfort_food | comfort_food_reasons | comfort_food_reasons_coded | ... | soup | sports | thai_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2.4 | 2 | 1 | 430 | NaN | 315.0 | 1 | none | we dont have comfort | 9.0 | ... | 1.0 | 1.0 | |
| 1 | 3.654 | 1 | 1 | 610 | 3.0 | 420.0 | 2 | chocolate, chips, ice cream | Stress, bored, anger | 1.0 | ... | 1.0 | 1.0 | |
| 2 | 3.3 | 1 | 1 | 720 | 4.0 | 420.0 | 2 | frozen yogurt, pizza, fast food | stress, sadness | 1.0 | ... | 1.0 | 2.0 | |
| 3 | 3.2 | 1 | 1 | 430 | 3.0 | 420.0 | 2 | Pizza, Mac and cheese, ice cream | Boredom | 2.0 | ... | 1.0 | 2.0 | |
| 4 | 3.5 | 1 | 1 | 720 | 2.0 | 420.0 | 2 | Ice cream, chocolate, chips | Stress, boredom, cravings | 1.0 | ... | 1.0 | 1.0 | |

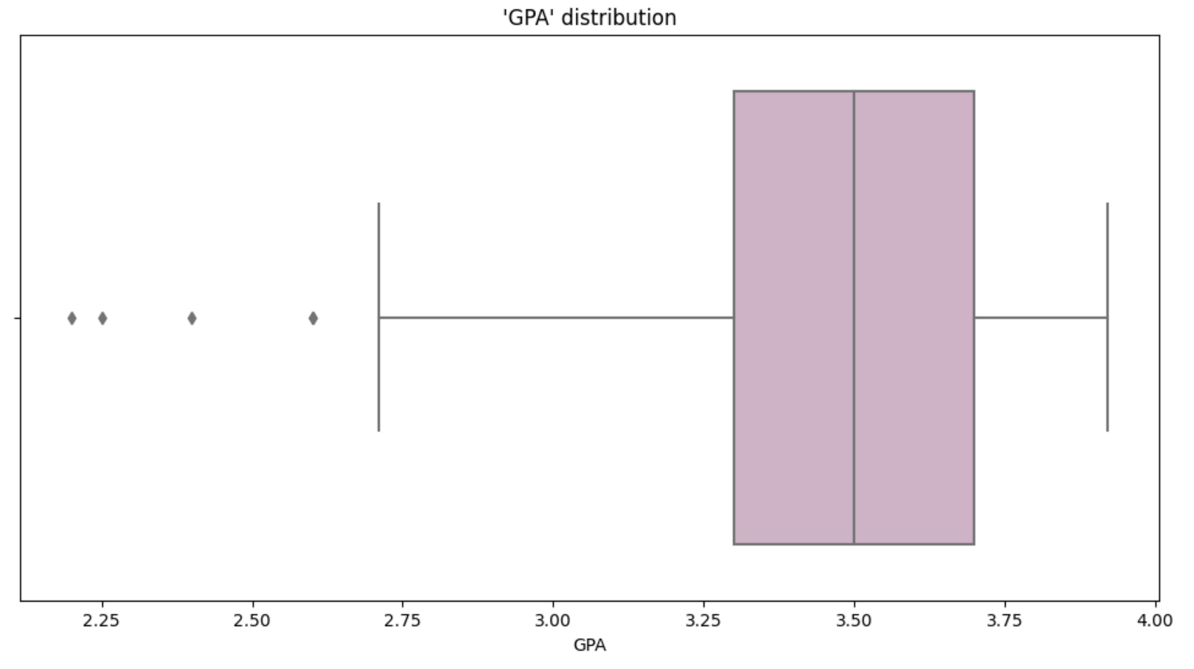
Student Overview Using GPA and Weight by Gender: We will analyze the data by considering the students' GPA and their weight, categorized by gender. This analysis aims to derive a comprehensive overview of the student demographic participating in this survey.

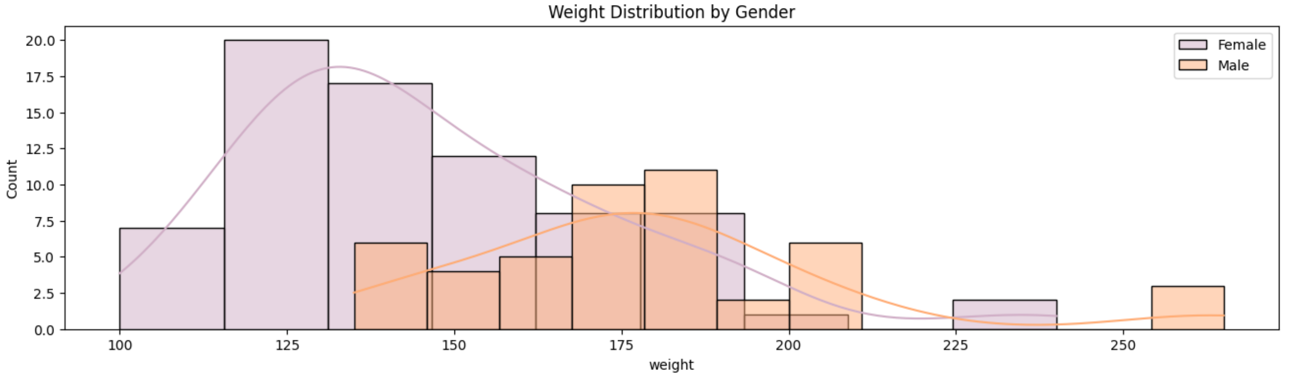
## Result Interpretation

The outcome reveals a predominance of students with high GPAs engaging in the survey. Additionally, it provides an average weight of the participating students. These results offer insights into the academic inclination and physical health metrics of the surveyed group.

## Learnings from the Results

This analysis allows us to understand the characteristics of the student population better, particularly their academic performance and health trends.




## Project Focus on Emotion and Food Choice

The project will concentrate exclusively on the attributes of emotion and the comfort foods associated with the emotion. These attributes will be used to develop a recommendation system, enhancing the relevance and personalization of food suggestions.

## Using BERT to Identify Food and Emotion Topics

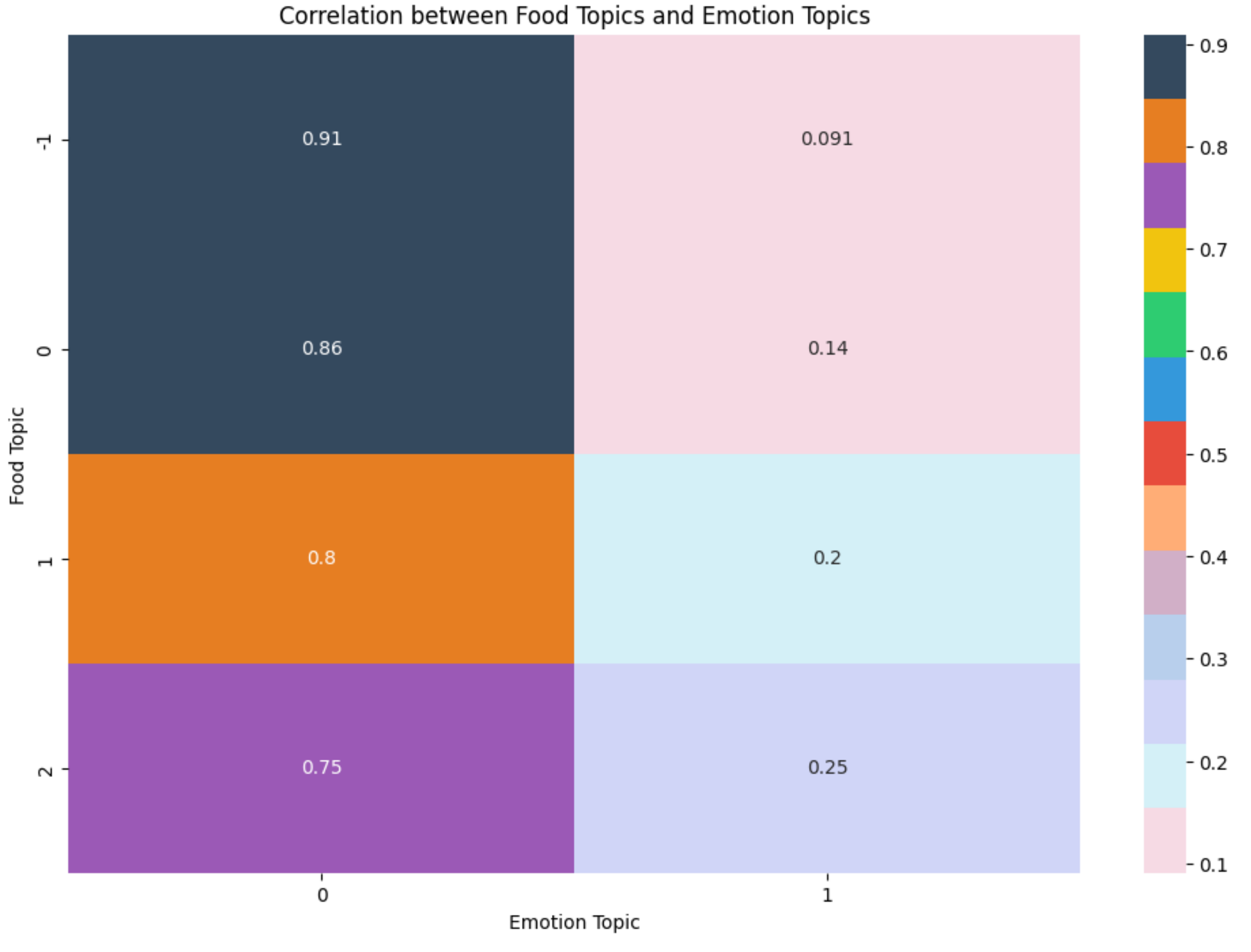
We will employ BERT, a leading language model, to discern and categorize topics related to food and emotions.

## Embedding and Topic Correlation Visualization

Post identification, these topics will be embedded and cross-referenced to understand their interrelations better. Visualizations will be utilized to effectively communicate these correlations.

## Result Demonstration

For instance, we might find that 'Food Topic 1' has a strong correlation with 'Emotion Topic 0', illustrating the interplay between specific food types and emotions.




## Supervised Learning for Accuracy Assessment

To evaluate the predictive accuracy, we apply supervised learning techniques such as a decision tree algorithm and regression analysis.

## Comparative Results

For example, we might observe that regression analysis yields an 89% accuracy rate, whereas decision trees achieve 78%. These findings will require further exploration and explanation to understand the efficacy of each method.

```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91        21
           1       0.00      0.00      0.00         4

    accuracy                           0.84        25
   macro avg       0.42      0.50      0.46        25
weighted avg       0.71      0.84      0.77        25

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91        21
           1       0.00      0.00      0.00         4

    accuracy                           0.84        25
   macro avg       0.42      0.50      0.46        25
weighted avg       0.71      0.84      0.77        25
```

## Utilizing NLTK

Stopword Removal and Lemmatization: We will also use NLTK, a powerful tool for natural language processing, focusing on the removal of stopwords and lemmatization. This process aids in refining the dataset, ensuring more precise and meaningful analysis.

## Result Application in Recommendation

The outcomes of this step will be directly applied to enhance the quality and relevance of our recommendation system, ensuring that suggestions are more tailored and accurate. Specifically, we filter out common words and punctuation.

# Visual Applications

**Choosing the right technologies:** A web based application was chosen for the ease of access it provides for users. To do this, we decided on technologies to simplify our development.

**Front end:** HTMX, CSS, HTML

**Back end:** Python, Flask

**Developing the application:** We connected the front-end and back-end with HTMX and flask. HTMX allows developers to embed HTTP requests and target elements in HTML tags as attributes.

**Designing the UI:** The application consists of various buttons on the screen as well as output views. The buttons correspond to an emotion felt by the user. When an emotion button is selected, HTMX sends a get request to python using flask. The returned HTML response is then put into the output views.

**Processing requests:** Based on the selected emotion, our python logic processes and returns food and restaurant recommendations. We use the K-Means algorithm to derive the highest rated nearby restaurants likely to contain the recommended food.

# Experimental Evaluation

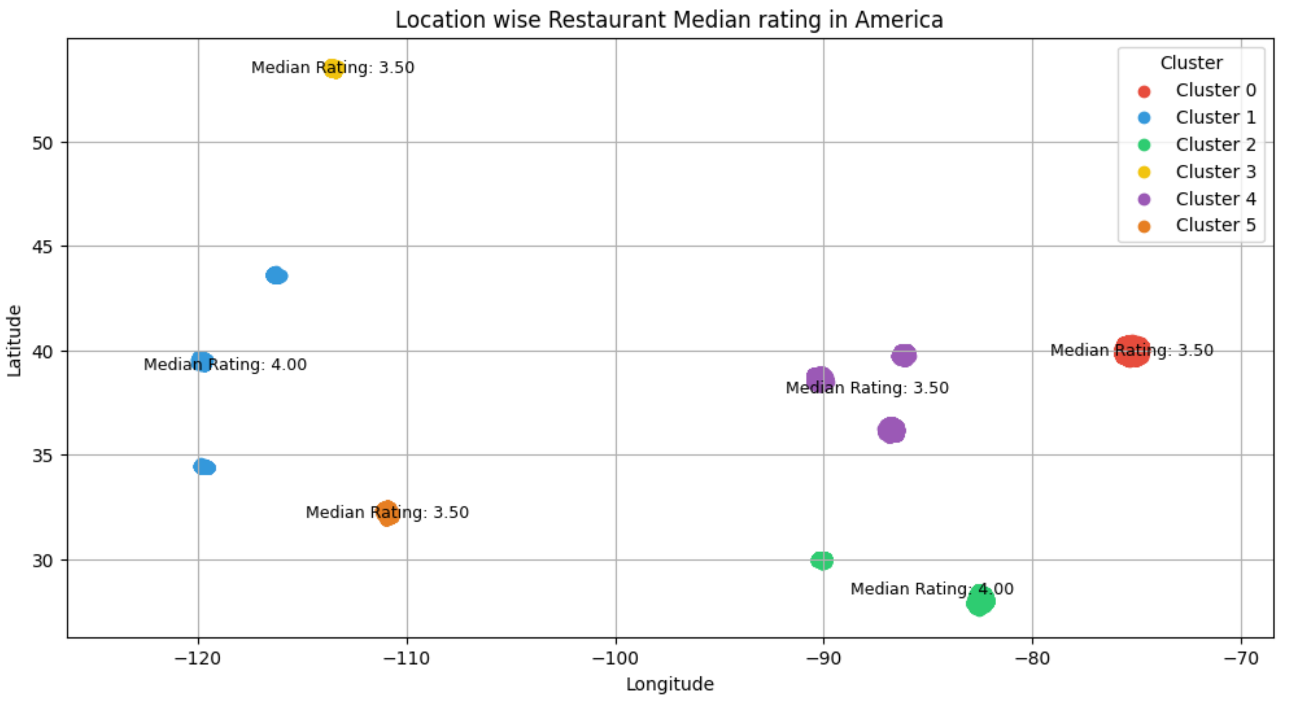
**Data Preparation for Restaurant Dataset:** The process begins with the cleaning and preparation of the restaurant data. This includes handling NaN values and negative entries.

**Results from Data Cleaning:** The cleaned dataset consists of 150,346 businesses. Among these, 103 businesses lack category specifications. After removing all entries with NaN values, 150,243 businesses remain in the dataset.

**Categorization of Business Types:** The dataset is then sorted to categorize different types of businesses.

**Results of Categorization:** The categorization reveals that there are 52,268 restaurants, 27,781 food-related businesses, and 24,395 shopping outlets.

**Filtering Process for Restaurant/Food Businesses:** The dataset is further refined by filtering out specific businesses that do not include specific categories, namely 'restaurants' and 'foods'. This allowed us to narrow down to restaurants only.

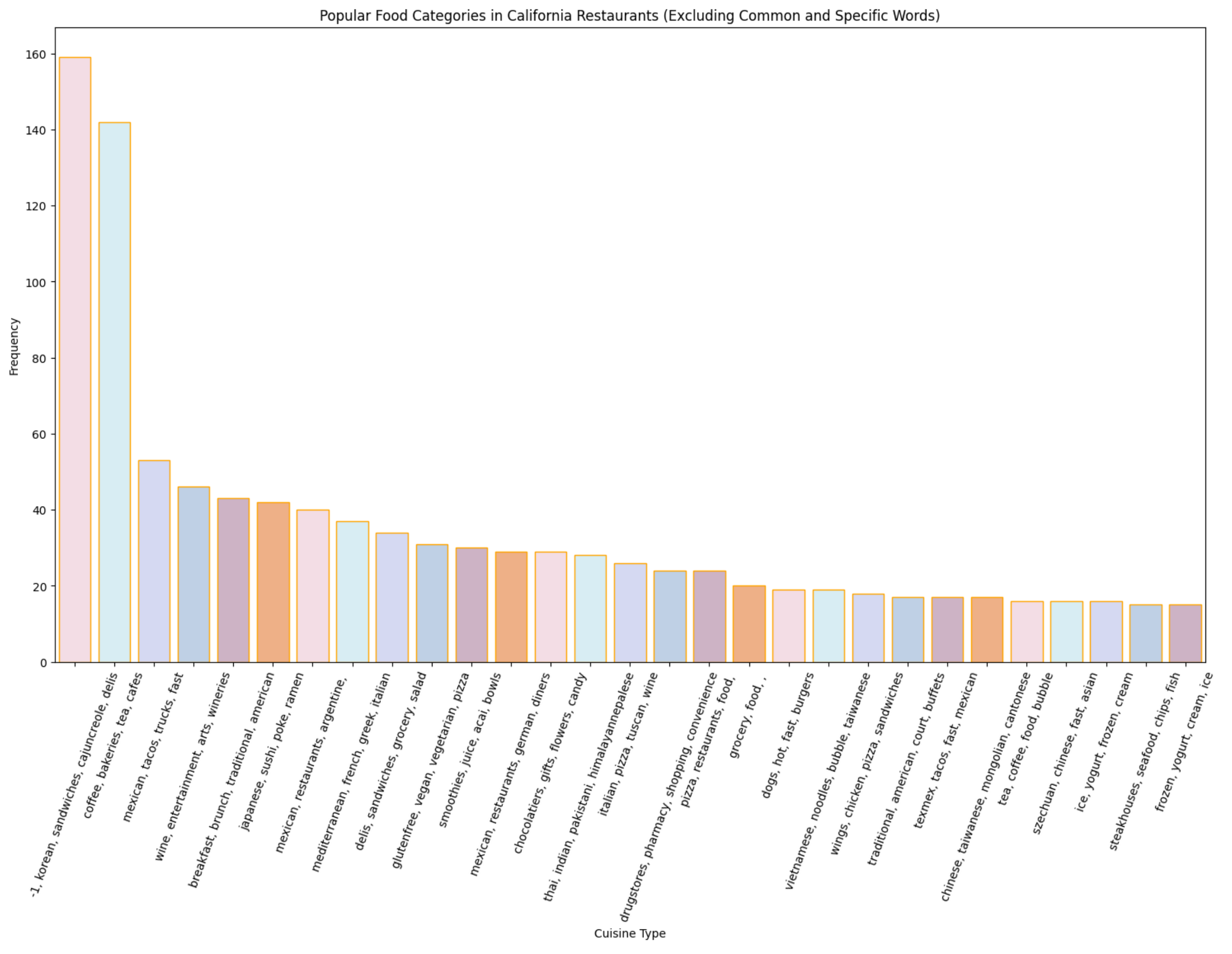



**Application of Kmeans Clustering for Popular Restaurant Locations:** Kmeans clustering is employed to identify popular restaurant locations across the USA.

- **Methodology:** The clustering process considers the median of ratings and the standard deviation to ensure the stability and quality of clusters. This helps determine the most optimal locations for restaurant recommendations.
- **Visualization and Algorithm Specifications:** A visual representation is created to illustrate the clustering. The Kmeans algorithm is configured with a parameter k=6.

**Focus on Restaurants/Food in California (CA):** The dataset is further narrowed down to focus on restaurants and food businesses in California.

- **Results and Decision Making:** Following an analysis of popular restaurant locations in the USA, cluster 1 (blue), which includes the UCLA campus area in California, is selected for restaurant recommendations. The dataset shows 64,629 restaurants/food businesses in total, with 1,596 of these located in California.




**BERT Analysis for Popular Food Categories in CA:** BERT is utilized once more to delve into popular food categories within California.

**Proximity Analysis to UCLA Campus:** The geographic proximity of restaurants in California to the UCLA campus is assessed.

- **Methodology:** The geospatial library is used to calculate the distances between the restaurants and the campus.
- **Setting Distance Threshold:** A threshold of 105 km is established to filter restaurants.
- **Results of Proximity Analysis:** A total of 27 restaurants are identified near the UCLA campus.

**Final Step: Recommending Restaurants Near UCLA Campus:** The last phase involves recommending restaurants near the UCLA campus, based on the food-emotion type analysis from the initial dataset.

- **Statistical Analysis of Local Restaurants:** The average ratings of the 27 restaurants near UCLA are calculated to gauge their overall quality.
- **Variance in Ratings:** The variance in ratings (1.07) is quite high on a scale of 1 to 5. This indicates a wide range of customer opinions, signifying that while some restaurants are rated significantly higher or lower than the average, the overall experience tends to be positive.
- **Result Interpretation:** The average rating suggests that the restaurants near UCLA are slightly above average, which can be used as a benchmark for recommendations. The high variance highlights the diversity in customer satisfaction levels.