Food Habit Analysis of the People of USA and Canada

Name: Md Rashel Uzzaman
Student ID: 23099317
M.Sc. in Artificial Intelligence
Friedrich-Alexander-Universität Erlangen-Nürnberg

Chapter 1: Introduction

This research aims to analyze the food habits of people in the United States (US) and Canada. This report provides the motivation behind the research, an overview of the dataset with licensing and pipeline and the methodology employed to analyze and interpret the findings. This research seeks to identify regional food preferences and trends which can serve as valuable information for those who are planning to enter the market or establish new restaurants.

The primary motivation of this research is to investigate whether there are discernible patterns in the food preferences among people in the USA and Canada. For example, in some regions burgers are popular while in others steak are more popular. Understanding these preferences will provide valuable information for market entry strategies.

Chapter 2: Data

The dataset used in this research is sourced from kaggle and licensed under Database Contents License v1.0 which permits its usage provided proper citation is maintained.

Source	https://www.kaggle.com/datasets/saketk511/1500-north-american-resturants
License Page	https://opendatacommons.org/licenses/dbcl/1-0/

Chapter 2.1: Dataset Details

The dataset[1] contains 1,500 rows and includes: names, cuisines served, ratings, city, state and additional features. During the data exploration phase, several challenges were identified.

Chapter 2.2: Data Cleaning and Normalization

The "cuisine" column contained multiple cuisine names separated by commas, resulting in a total count of 21,521 unique cuisine entries. However; upon manual inspection, it was discovered that many entries were duplicated due to slight spelling variations. For example:

"burgers" and **"burger"**. To address this, a systematic approach to normalize cuisine names was implemented.

The Gensim[2] library was utilized, specifically the glove2word2vec model, to create vector space representations of cuisine names. This enabled the measurement of semantic similarity between terms. Using these embeddings, similar terms were clustered into 10 groups, ensuring that variations in spelling or phrasing were treated as equivalent. For example: **"burgers"** and **"burger"** were standardized as **"burger"**. This normalization process significantly reduced the total count of unique cuisine names from 21,521 to 572. While the approach enhanced data consistency, it is acknowledged that the clustering heuristic may not achieve perfect accuracy, leaving room for minor discrepancies.

Chapter 3: Analysis

In this section, I will answer my methodology, results, findings and interpretations. For description purposes, both states of the USA and provinces of Canada will be referred to as states.

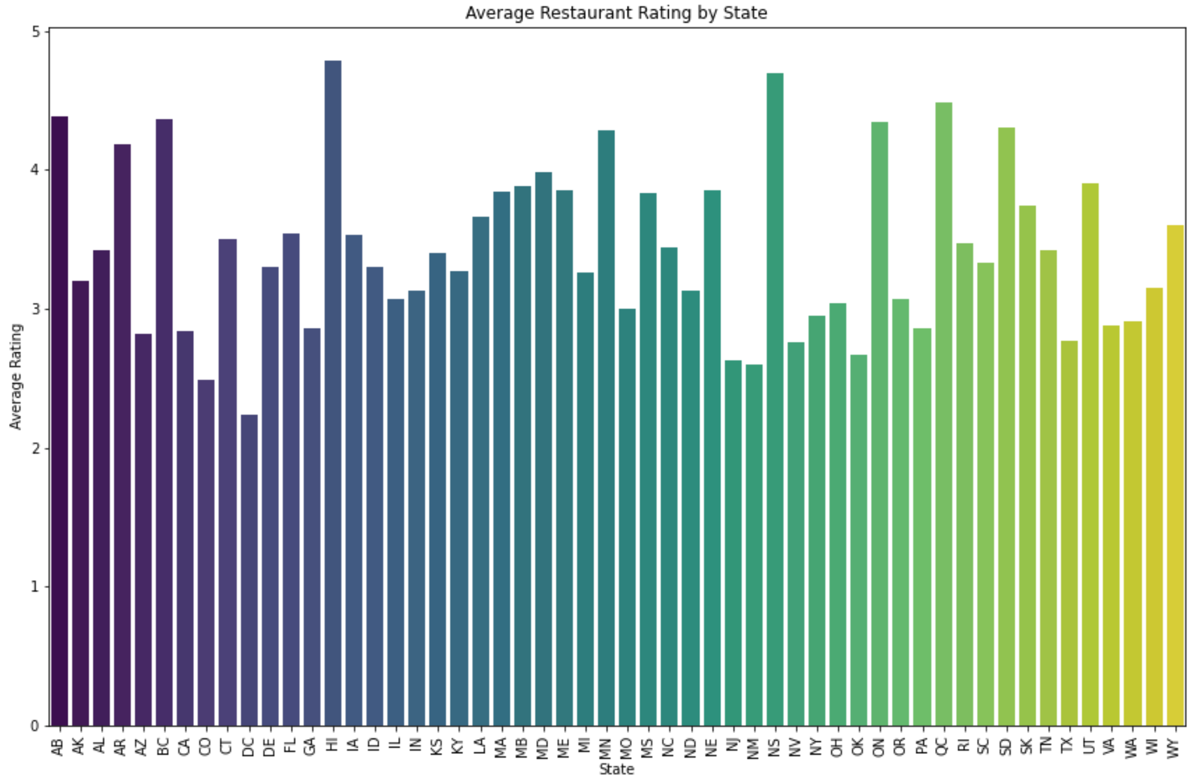

Figure 1: Average Restaurant Rating by State

First, **Figure 1** shows the distribution of average resultant ratings in different states. Here it can be clearly observed that states such as Connecticut (CT), California (CA) and New Jersey (NJ) have comparatively lower ratings than others. On the other hand, states such as Hawaii (HI) and Quebec (QC) have higher ratings for the restaurants. This shows that those states have good quality restaurants compared to others. For calculating this bar chart, I took the sum of the ratings of all restaurants in a particular state and divided it by the total number of restaurants in that state.

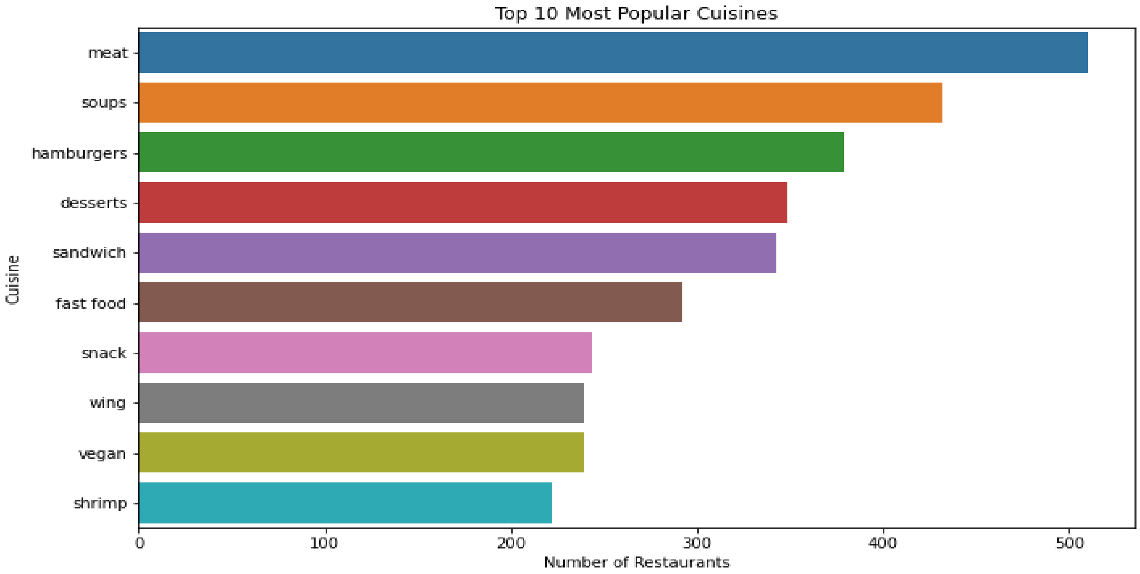

Figure 2: Top 10 Most Popular Cuisines

Furthermore, the bar chart in **Figure 2** highlights the most popular cuisines based on restaurant frequency. "Meat" leads as the most common cuisine, followed by "soups" and "hamburgers." Other popular categories include "desserts," "sandwich," and "fast food." Interestingly, categories like "wing" and "vegan" are also prominent, reflecting evolving dietary preferences.

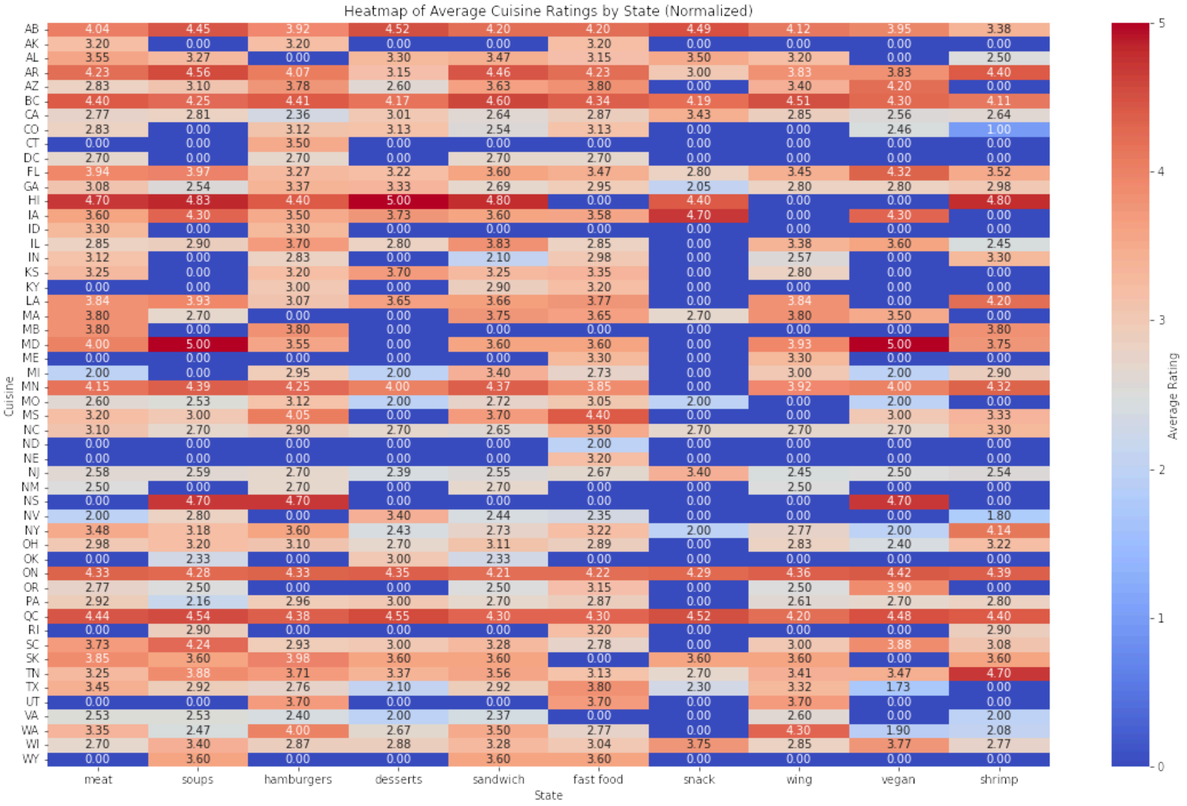

Figure 3: Heatmap of Average Cuisine Ratings by State

Based on the heuristic described in **Chapter 2**, here I count the cuisines served in restaurants. For example, the value of the soups is 432; that means out of 1500 restaurants in the dataset, soup is served in 432.

Lastly, **Figure 3** is the most important part of this research. The heatmap shows the pattern of the food habits of the people of a particular state. Also, it gives us an overview of people's diverse food choices based on the ratings in different states. For example, people of Maryland (MD) rated restaurants very highly that serve Soup or Vegan. On the other hand, people of Georgia (GA) rated restaurants very poorly that serve Soup or Vegan food. It shows the diversity of the food choice. If we look into Michigan (MI), we can see that the ratings are comparatively low across all the 10 cuisines. But in **Figure 1**, the average rating for Michigan (MI) was not that much lower. It means that, for Michigan (MI) people, their favorite cuisines are not in the top 10 cuisines from their country. Another finding is that Hamburgers and Fast Food items are rated well throughout the country compared to other cuisines. Lastly, another finding is that there are 0 values. That means according to the dataset, there are no restaurants of the cuisine there. Although the dataset has only 1500 restaurants, it is highly unlikely this would be the case. Now let me move on to how I calculated this heatmap. After doing the normalization from **Chapter 2**, the dataset was grouped by state and cuisine, and the average rating was calculated for each combination. This ensured that the heatmap accurately reflected the average preference for each cuisine within a state. Because of that, we can see the 0 values in the heatmap.

Chapter 4: Conclusion

The primary question posed in this research was to analyze the food habits of people in North America, specifically identifying state-wise preferences for cuisines. The findings provide significant insights into regional preferences and the overall dining trends in the USA and Canada. The analysis uncovered intricate regional patterns, such as Maryland's high ratings for "Soup" and "Vegan" cuisines, and Michigan's apparent deviation from the top 10 cuisines. This demonstrates the diversity of food habits across states. While the research successfully answered the primary question, some limitations affected the scope and depth of the conclusions. The dataset contained 1,500 restaurants, which may not fully represent the diverse and expansive food industry in North America. The reduction of cuisine names from 21,521 to 572, while essential for analysis, might have oversimplified nuanced distinctions, potentially affecting accuracy. Some states showed zero values for certain cuisines, likely due to data sparsity rather than an actual absence of those cuisines. Despite these limitations, the research provides a robust foundation for understanding food habits in the USA and Canada. The findings have practical applications for businesses in the food and hospitality industries, enabling better regional targeting and menu optimization.

Reference

[1] Saket Kumar. (2024). 1500 North American Restaurants [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/7615817

[2] Gensim - Python NLP Library. (n.d.). https://pypi.org/project/gensim/