Data Report: Food Habit Analysis of North Americans

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 USA and Canada. The report provides an overview of the dataset used for this project and explores potential patterns in regional food preferences.

Chapter 2: Data Sources

The dataset includes key information such as restaurant names, the cuisines they serve, restaurant ratings, city, state, and five additional features. This data enables a state-wise analysis of cuisine ratings and helps identify patterns in food habits across different regions of the USA and Canada. For instance, the analysis can reveal if certain regions prefer specific cuisines more than others.

The dataset is in tabular format, stored as a CSV file, and was downloaded from Kaggle. The data quality is good, with only one missing value in the cuisine column.

The dataset is licensed under the Database Contents License (DbCL) v1.0, which allows its use. According to the dataset's source webpage, proper citation is required when using this data and this requirement will be adhered to in this project.

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

Chapter 3: Data Pipeline

To create the data pipeline for this project, I utilized Python [2] as the primary programming language. The process involved several steps, which are outlined below:

Chapter 3.1: Downloading the Data

The dataset was hosted on a web link and required downloading. To achieve this, I used the httpx [3] library, specifically its **get** method, which fetches data from URLs.

Chapter 3.2: Extracting the Data

Since the dataset was downloaded as a ZIP file, the next step involved extracting its contents. For this, I employed the zipfile [4] library, using its **open** method to unzip the file. This ensured that the data was accessible in its required format for subsequent processing.

Chapter 3.3: Reading the CSV File

Once the data was extracted, I used the **read_csv** method from the pandas [5] library to load the CSV file into a DataFrame.

Chapter 3.4: Saving Initial Metadata

After reading and preparing the dataset, I focused on saving the initial metadata for further use. For this purpose, I utilized the sqlalchemy [6] library to store the data in a SQLite database. By saving the data as a SQLite file, I ensured that the data was easily retrievable and could be used for structured queries and analysis later in the project.

Chapter 4: Result and Limitations

The dataset contains a total of 1500 rows. During the data exploration process, I identified a missing value in the cuisine column. The cuisine column itself posed an additional challenge, as it contains multiple cuisine names separated by commas. After splitting by commas, I noticed that the total count of cuisines was 21,521. However, during manual inspection, I discovered that this unusually high number was due to inconsistencies in the data. Many values were essentially duplicates caused by slight spelling variations or formatting differences. For instance:

- "burgers" and "burger"
- "Coffee and Tea" and "Coffee & Tea"

These inconsistencies created redundancy and inflated the total count. Solution to Normalize Cuisine Names. To address this issue, I leveraged the Gensim [6] library. Specifically, I utilized the **glove2word2vec** model to create a vector space representation of the cuisine names. This allowed me to measure the semantic similarity between different terms. Using these embeddings, I clustered similar terms into 10 groups, ensuring that small variations in spelling or phrasing were treated as the same entity. For example:

- "burgers" and "burger" were grouped into a single cluster and standardized as "burger."
- "Coffee and Tea" and "Coffee & Tea" were grouped and standardized as "Coffee & Tea."

After clustering and standardizing the terms, I replaced all variations with their respective normalized values. This heuristic significantly reduced the total count of unique cuisine names from 21,521 to 572, making the data more consistent and manageable for analysis. Although the cuisine count is significantly reduced, the heuristic is not perfect. So, the result will not be 100% accurate.

Another challenge I think I will face is the distribution of restaurant ratings; as shown in **Figure 1**, with most ratings concentrated on the higher end of the scale. This right-skewed

distribution makes it difficult to determine which states prefer particular cuisines, as the ratings across many restaurants reduces the variability needed to draw clear distinctions.

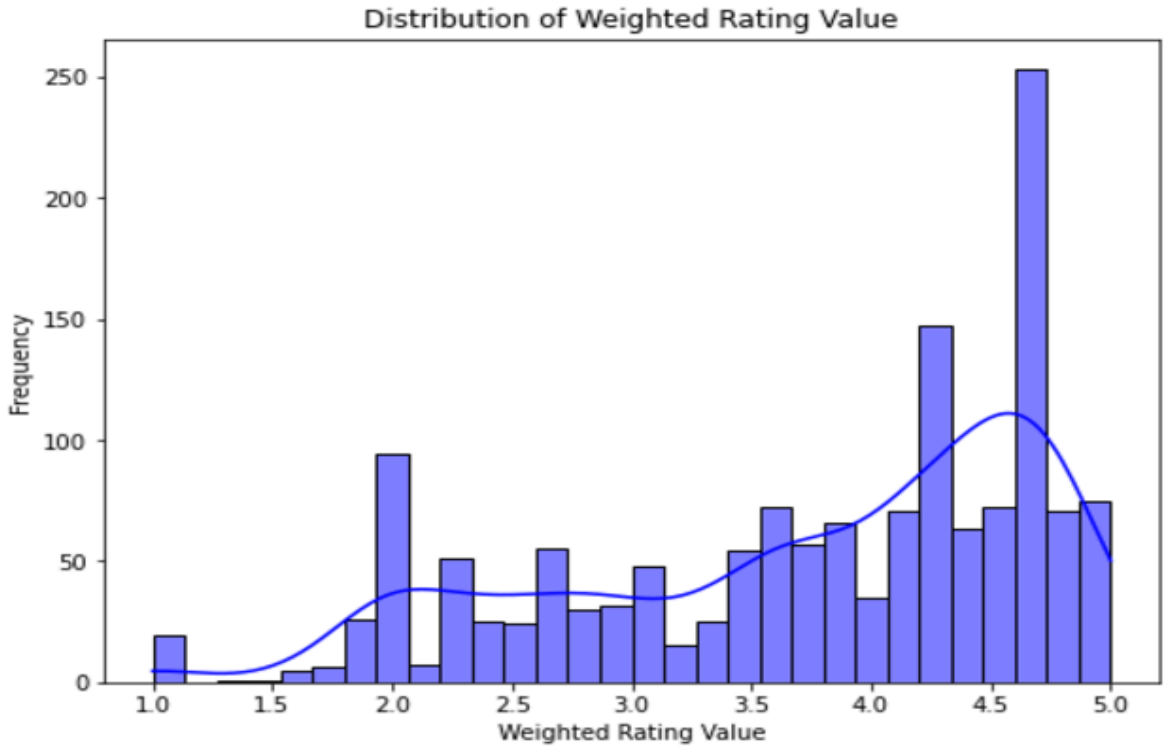

Figure 1: Rating Distribution Count of Restaurant

For example, if the majority of restaurants in several states receive similarly high ratings, it becomes challenging to identify unique preferences or regional favorites based solely on the rating data.

Reference

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

[2] Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

[3] HTTPX. (n.d.). https://www.python-httpx.org/

[4] zipfile — Work with ZIP archives. (n.d.). Python Documentation. https://docs.python.org/3/library/zipfile.html

[5] pandas - Python Data Analysis Library. (n.d.). https://pandas.pydata.org/

[6] SQLAlchemy. (n.d.). https://www.sqlalchemy.org/