Milestone 5: Data Warehousing

Team Name: Team6

Team Members:

- Wei Cai
- Yunyi Chi
- Jian Han
- Qifeng Huang
- Quancheng Li
- Chenyu Mao
- Xinyi Wu

Project Name: ChineseFlavors

1. External Data Sources Integration

For our project "ChineseFlavors", we have integrated external data sources to enrich our existing dataset and provide insightful analysis. Below are the external data sources we have used:

1.1 Nutrition Information

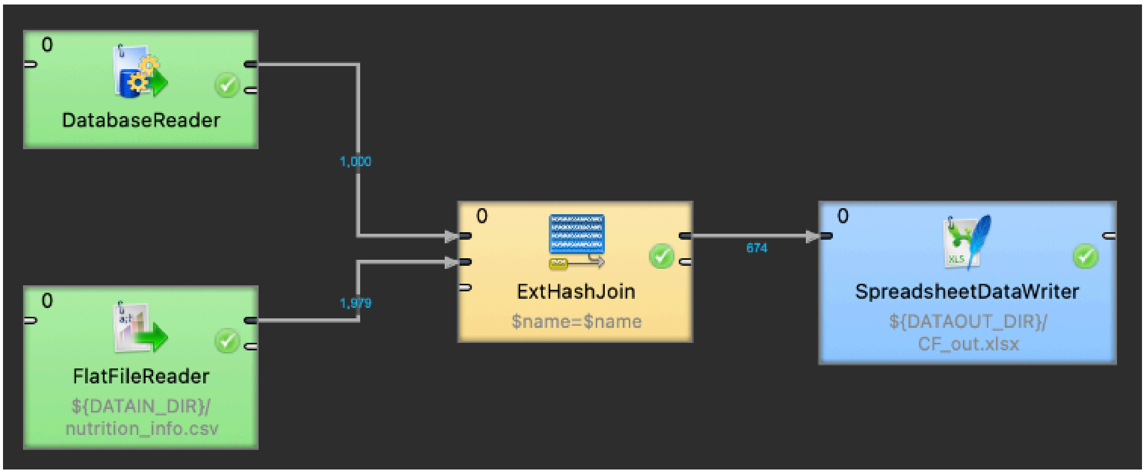
- Description: A crucial dataset, `nutrition_info.csv`, which contains the nutritional profile of food items, was integrated into the existing dataset to enhance the dishes with valuable health-related information. This dataset contains detailed nutritional information on various ingredients commonly used in Chinese cooking. It includes data on calories, fat content, vitamins, minerals, etc.
- Source: http://www.calorieninjas.com
- Utilization: We plan to use this data to analyze the nutritional content of popular Chinese dishes, aiming to offer healthier options to our users by adjusting ingredients.

1.2 Chinese Food Computing

- Description: By integrating `Chinese.csv`, the team aimed to understand and respond to consumer preferences, optimizing the menu to cater to popular tastes.
- Source: http://123.57.42.89/FoodComputing__Dataset.html
- Utilization: This dataset will guide us in identifying and forecasting trends in Chinese cuisine, allowing us to innovate our menu effectively to cater to popular demand.

2. ETL Workflows

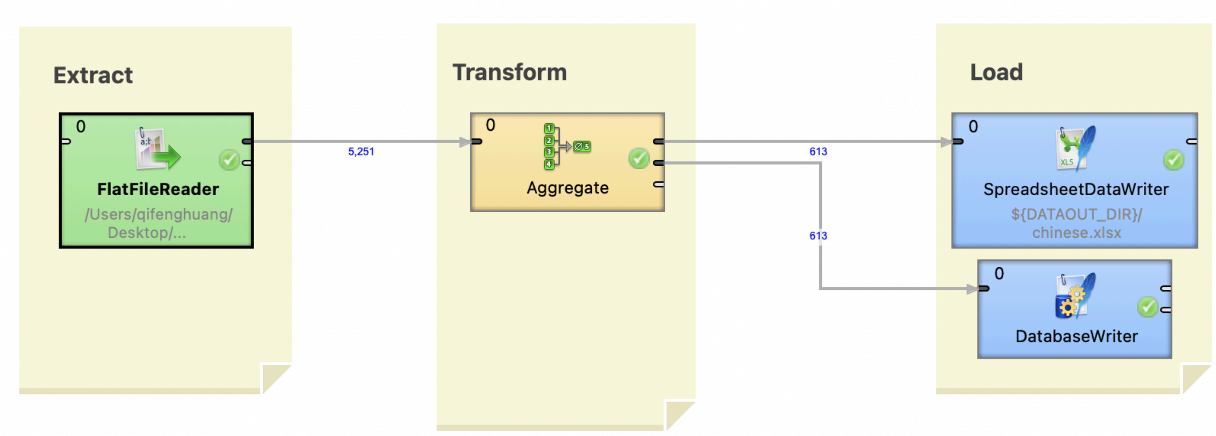
2.1 Nutrition Information



- Components Used:
  - DatabaseReader: It is configured to connect to the database where 5200_dataset - dataset_part_0.csv is stored or its corresponding database table, using the necessary credentials and connection strings to ensure secure access.
  - FlatFileReader: This component is crucial as the first step in the ETL (Extract, Transform, Load) process. It reads the data from a flat file source, such as CSV or text files. For our project, the FlatFileReader is configured to parse the nutrition_info.csv file, which contains various nutritional metrics for food ingredients used in Chinese cuisine.

        - ExtHashJoin: This component merges the nutrition data with the main dataset of our ChineseFlavors project, matching ingredients with their corresponding nutritional information.
        - SpreadsheetDataWriter: The transformed data is also exported to a spreadsheet to facilitate the creation of visual nutrition profiles for each dish.
    - Description:
        - The Nutrition Information ETL workflow systematically incorporates external nutritional data to enrich our existing ChineseFlavors dataset. Starting with the extraction of detailed nutrition information for various food items, the workflow processes this data through normalization and joins it with our internal dataset based on matching ingredients. By applying transformation logic, we can derive comprehensive nutrition profiles for our complete recipe dataset. The resulting enriched data is then stored in our Data Warehouse and outputted to a spreadsheet. The analysis derived from this data informs us about the healthiness of our dishes, which can be used to guide menu planning and marketing strategies focused on health-conscious consumers.
        - By adding these details, the role and impact of each ETL component in the Nutrition Information workflow are made clear, demonstrating how they collectively enhance our dataset for more profound health and nutrition insights.

2.2 Chinese Food Computing



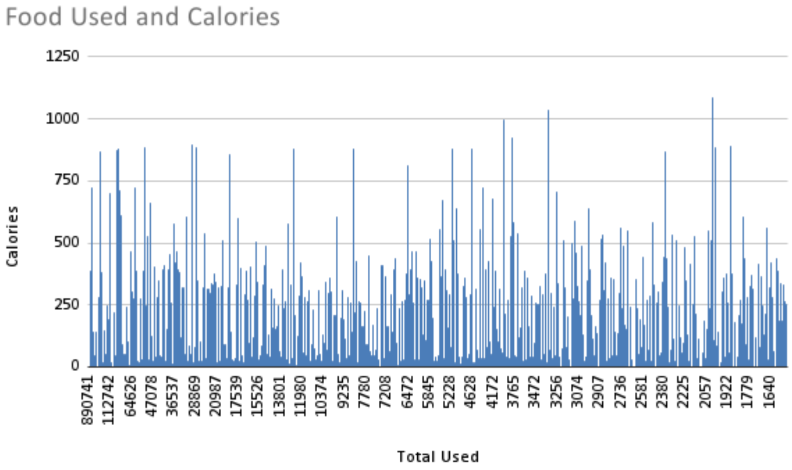
- Components Used:
    - FlatFileReader: This component is responsible for reading the data from the Chinese.csv file, which contains comprehensive information about the popularity and preferences of various Chinese dishes across different regions.
    - Aggregator: This component aggregates data points to provide a summary of trends and preferences, giving us insights into the most popular dishes and consumer inclinations toward certain flavors or ingredients.
    - SpreadsheetDataWriter: Utilized for outputting data into a spreadsheet format for further analysis and chart creation in tools like Excel or Google Sheets.
    - DatabaseWriter: It is employed to persist the transformed and aggregated data into our Data Warehouse, allowing for more complex queries and historical trend analysis.
- Description:
    - The Chinese Food Computing ETL workflow is designed to ingest consumer preference data from an external dataset. After initial extraction, the data goes through a series of transformations where it is cleaned, standardized, and aggregated to highlight trends and preferences in Chinese cuisine consumption. The outcome of this workflow is twofold; one stream outputs to a Data Warehouse, enriching our analytical capabilities, and the second stream outputs to a spreadsheet for immediate visualization and reporting purposes. The final aim is to utilize these insights to optimize our menu offerings, align with current consumer trends, and better forecast future demand.

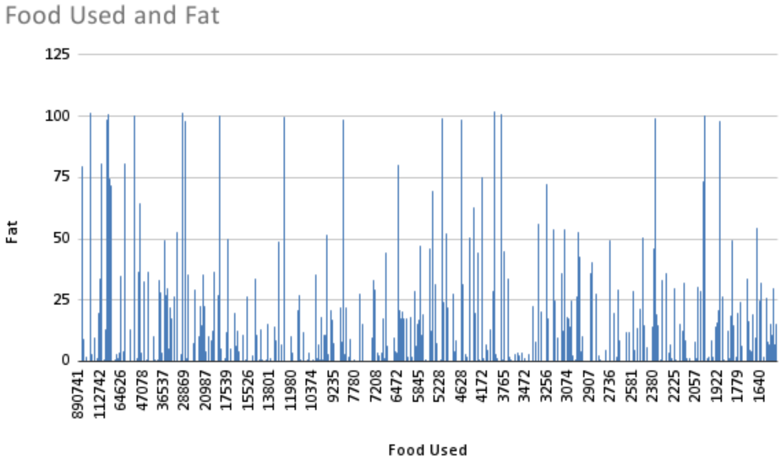
3. Data Visualization and Analysis

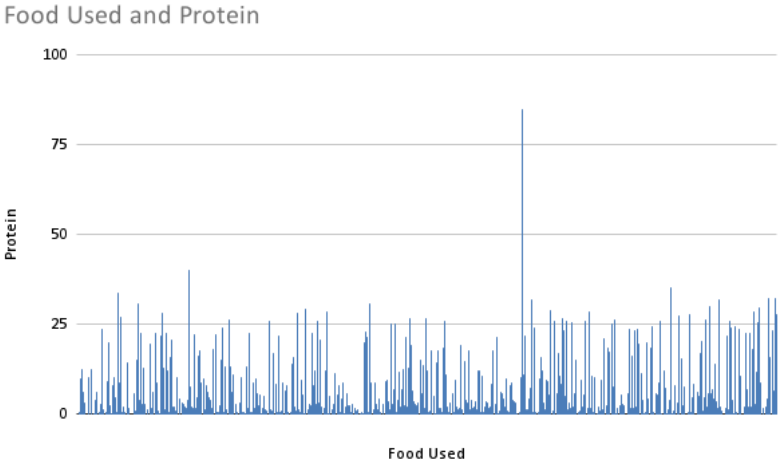
3.1 Hypothesis

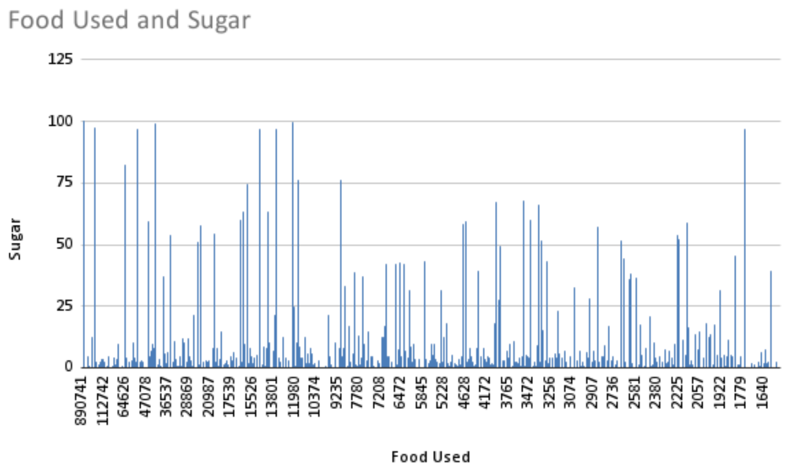
3.1.1 Nutritional Hypothesis

- Our hypothesis posited that the quantity of ingredients in recipes would directly correlate with their nutritional values, particularly calories, fats, and sugars. We anticipated that dishes with a higher count of ingredients would generally contain more calories and fats.







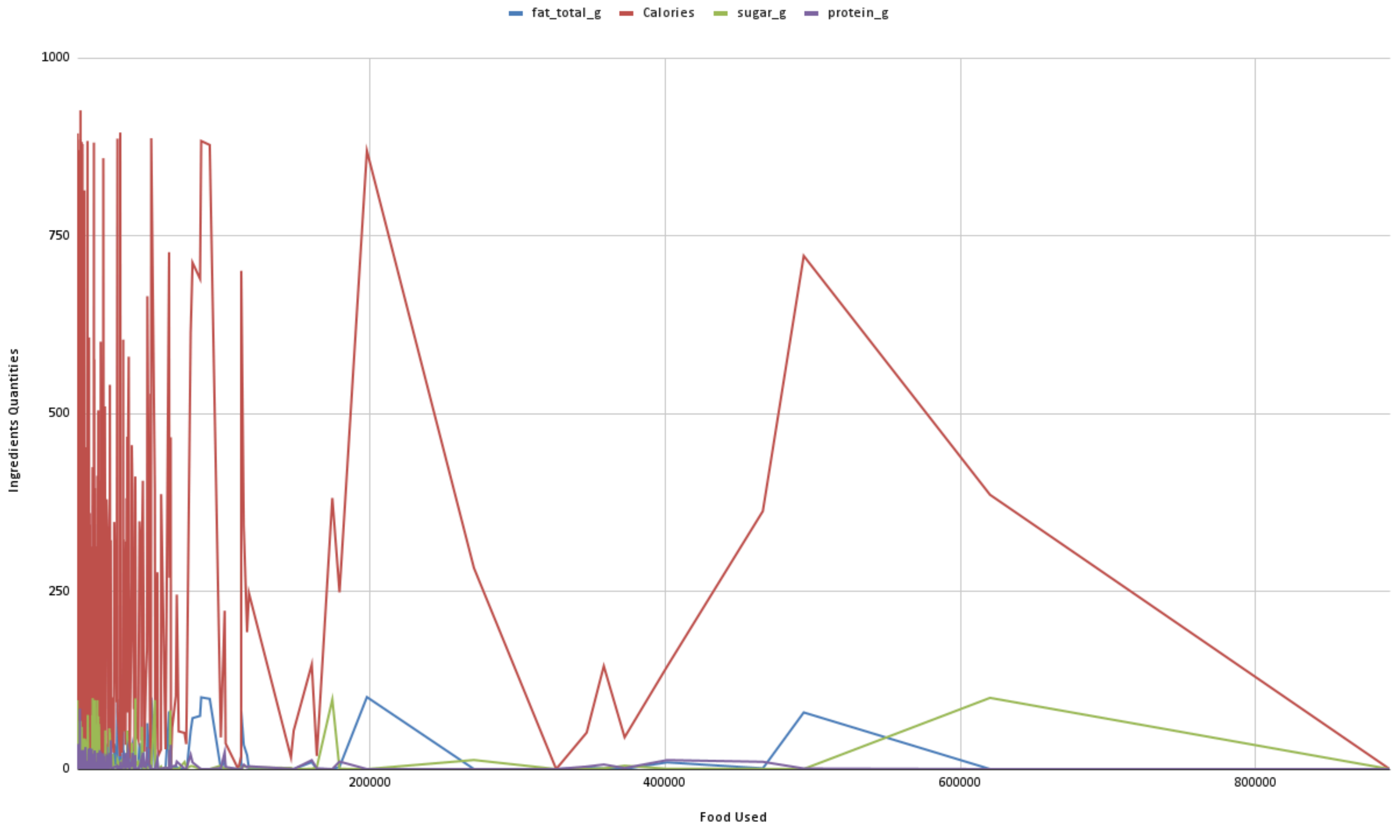


3.1.2 Chinese Food Computing Hypothesis

- The second hypothesis centered around consumer preferences, predicting that an analysis of data from Chinese.csv would allow us to pinpoint trending dishes and regional favorites, thus enabling us to tailor our menu offerings to align with current tastes and demands.

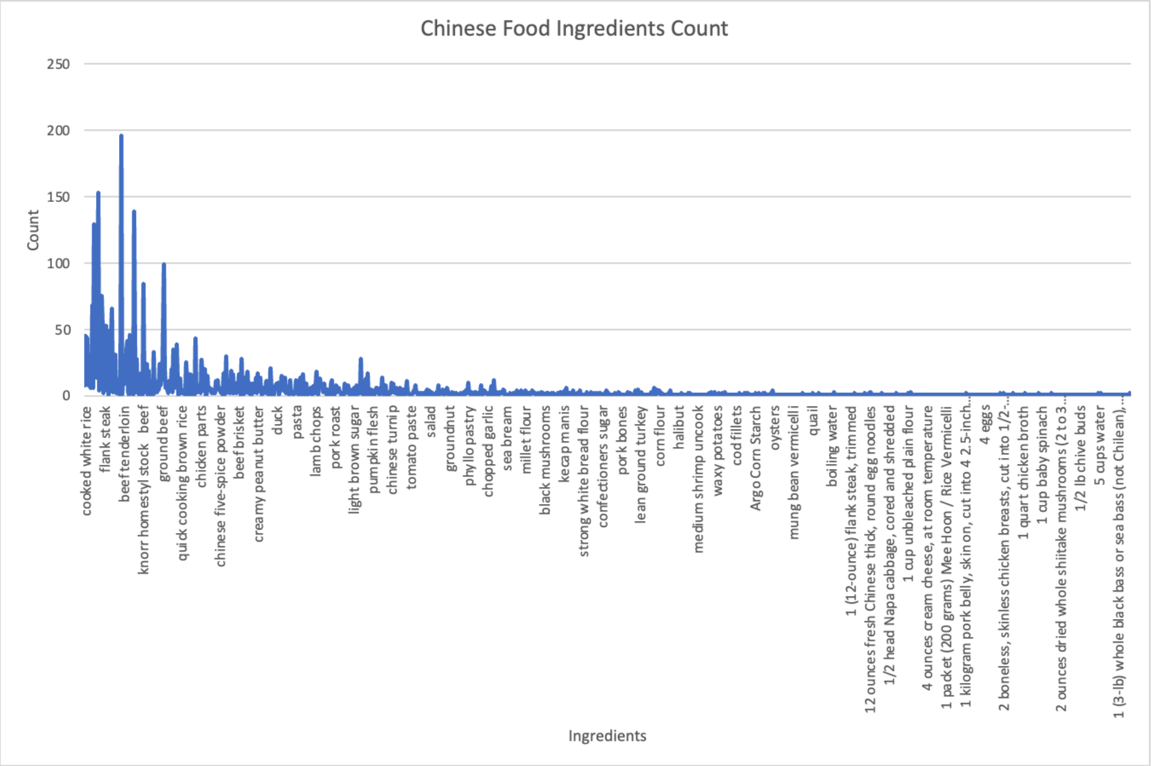
3.2 The results of combining the data, and if it validates or invalidates your hypothesis.

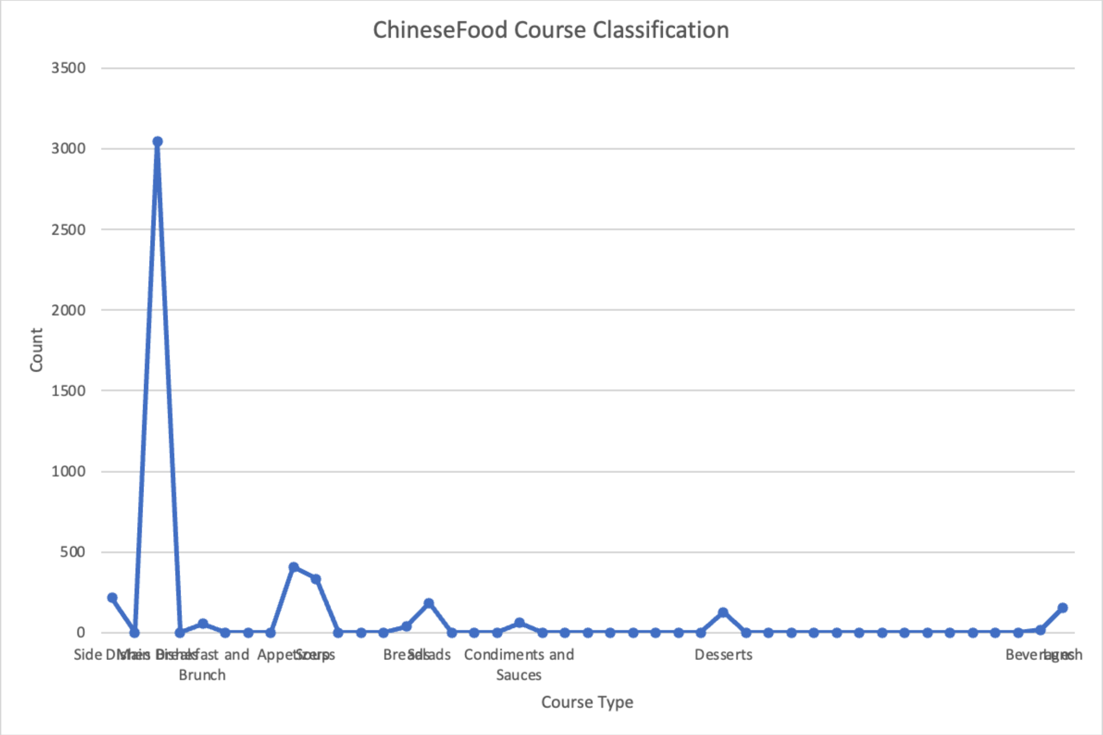
3.2.1 Nutritional Analysis



- Ingredient Nutritional Impact: Contrary to expectations, the charts revealed a weak correlation between the quantity of ingredients used in recipes and their calorie and fat content. However, a direct relationship between ingredient quantity and sugar content was noted. This suggests that while certain ingredients might be used liberally, they do not necessarily contribute to higher calorie and fat levels in Chinese dishes, indicating a nuanced approach to recipe composition.

3.2.2 Trend Analysis





- Course Classification: The classification chart clearly depicts that certain courses, like mains, are far more prevalent than others, like beverages or desserts, indicating potential areas for menu expansion or focus.

3.3 Chart's significance for your application and further actions

3.3.1 Nutritional Insightfulness

- Chart Meaning: The nutritional charts emphasize the potential for health-conscious menu innovation, identifying key areas for nutritional improvement without compromising on taste.
- Proposed Initiatives: Inspired by this analysis, we aim to refine our culinary approach, embarking on a menu overhaul to introduce dishes that embody both flavor and well-being.

3.3.2 Preference-Oriented Menu Development

- Chart Meaning: Classification charts reveal our current menu's structural bias towards certain dish types, hinting at unexplored avenues for menu expansion.

- Proposed Initiatives: Motivated by this revelation, we intend to broaden our culinary horizons, specifically targeting the enhancement of our dessert and beverage line-up to offer a more rounded dining experience.