Interactive Meal Recommendation System: Using Web Scraping and User-Driven Filters for Recipe Selection

Mete Harun Akcay

I. Introduction

Deciding what to cook can be a challenging task in today's world, especially for individuals aiming to balance taste with nutritional goals. With countless online recipes available, looking through them to find meals that fit specific calorie or dietary preferences is time-consuming. Therefore, in this project, I implemented a program to speed-up the meal selection process by allowing users to input their calorie and WeightWatchers range preferences and receive the list of recipes fitting their preferences. The system is built using web scraping techniques to collect data from Skinnytaste.com, and data analysis is used to provide insights into recipe key, calorie and WW point distributions. A user-friendly interface allowing for interactive filtering of the collected recipes based on user-defined criteria was designed using Google Colaboratory.

II. Data Collection

A. Methodology

Data collection for this project was performed using web scraping techniques, specifically Selenium. The data was collected from the popular recipe website Skinnytaste. The first 50 pages of the website were scraped to gather information such as the name of the food, its calorie content, WW points, summary, image link, and list of recipe keys. Since each information has the same HTML class name in all of the recipes, they were scraped using the method CSS_SELECTOR. In total, data from 50 pages, each having 20 recipes were scraped, then using pandas library, the scraped data was converted to a data frame having 1000 rows and 6 columns.

B. Challenges

The web scraping process encountered several challenges that required adjustments to ensure successful data collection. Initially, the scraping process took approximately two hours to complete, but it was disrupted multiple times when Google Colaboratory became unresponsive during execution. After experiencing these interruptions several times, I came up with the idea that I should save the progress into an Excel file each time the process was stuck. When I ran the program, it stopped running while scraping the 680th recipe. I stopped the program, saved the information in an excel file, and restart the program from the 680th recipe. At the end of the process, I merged the two resulting Excel files into a single dataset.

Another challenge was that certain instances on the website lacked the required information, such as calorie counts or personal points. They were mostly meal plans, or suggestions for cooking; therefore, it would be unnecessary to keep them in the data frame. To address this issue, I applied error-handling mechanisms using try-except blocks to ensure that the program continued to run smoothly even when encountering incomplete data. After collecting the data, I removed rows with missing values to maintain the integrity of the dataset.

Despite these obstacles, the website's overall design was consistent, and the HTML structure remained stable across pages. Therefore, it was not that challenging to collect the data.

III. Data Analysis

After removing instances with missing values, I conducted exploratory data analysis to gain insights into the distribution of key variables such as calories, WW points, and recipe keys. Additionally, I did further analysis to reveal the relationship between the variables.

A. Calories Distribution

The calories of the recipes range between 8 (Taco Seasoning) and 608 (Sheet Pan Thanksgiving Dinner), indicating a wide range of options. Visualizations were used to better understand the calorie distribution. As it can be seen in the box plot in Figure 1, the first quartile value was 147 calories, the median was 225.5 calories, and the third quartile (Q3) value was 308 calories. These values indicate that most recipes are within the moderate calorie range, aligning with the website's goal of offering healthy meal options, while also having few outliers observed above 500 calories.

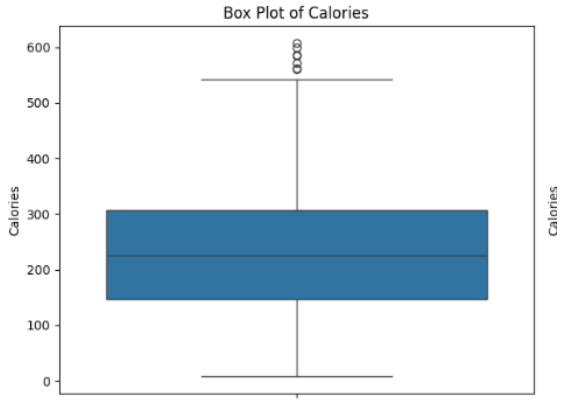

Fig. 1. Box Plot of Calories

The histogram below exhibits a right-skewed pattern, meaning that while the majority of the recipes have calories clustered around the lower to mid-range (150–300 calories), a smaller subset of recipes exhibits much higher calorie counts. These higher-calorie recipes, although less frequent, contribute to the overall skewness, again suggesting that the dataset mostly offers lower to moderate-calorie options with a few exceptions.

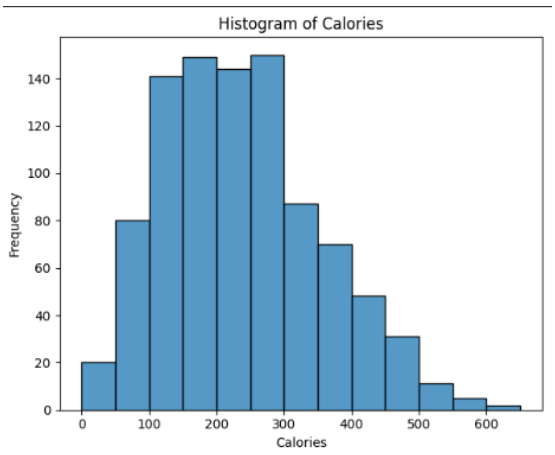

Fig. 2. Histogram of Calories

B. WW Points Distribution

WW points are a scoring system designed to simplify healthy eating by assigning foods a point value based on their nutritional content, including factors such as calories, saturated fat, sugar, and protein. While snacks, fast foods etc. can have WW points more than 20, in our dataset the foods with the maximum points are Korean Beef and Lamb Keema with Peas with 13 points. On the other hand of the spectrum, the presence of zero-point recipes (Figure 3) indicates that there are some highly nutritious, low-calorie foods that can be consumed freely without contributing to a user's daily points allowance. The histogram also reveals that the majority of recipes fall within the 3 to 7-point range, with a visible peak around 5 points.

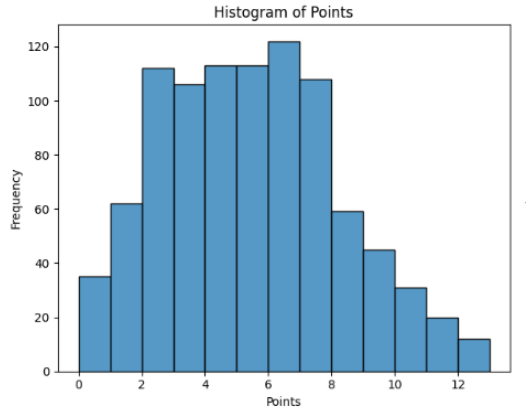

Fig. 3. Histogram of WW Points

Examining the quartile values provides further context to the distribution. The first quartile is at 3 points, the median is at 5 points, and the third quartile is at 7 points. This suggests that half of the recipes fall between 3 and 7 points, meaning most recipes are moderate in points, offering a balance between low and high-point meals.

The violin plot (Figure 4) shows a higher density of recipes around the median, suggesting that users are most likely to encounter recipes with moderate point values, around 5-6 points. The narrowing at both ends of the violin plot shows that very low or and especially very high-point recipes are rare. To sum up, distribution of WW points shows a similar trend to the distribution of calories by having right-skewed histograms, indicating that the most of the recipes have low to moderate calories and WW points while the number of high-calorie and high-point recipes decreases as the values increase.

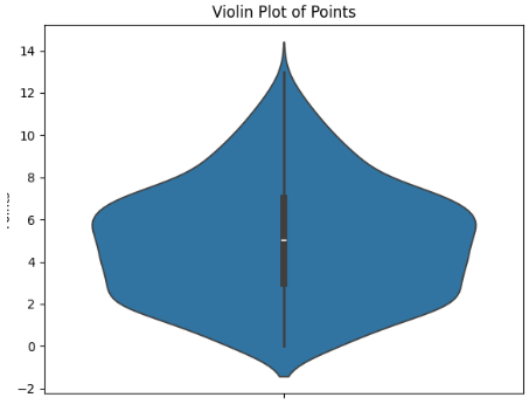

Fig. 4. Violing Plot of WW Points

C. Recipe Key Distribution

The distribution of recipe keys offers valuable insights into the variety of dietary categories available in the dataset. A pie chart showing ratio of keys and the dictionary of keys are given below. As shown in Figure 5, gluten-free (GF) recipes dominate the dataset with the highest frequency, representing almost a fifth of the dataset. This indicates a strong emphasis on gluten-free options, catering to individuals with gluten sensitivities or those following gluten-free diets. Kid friendly foods, recipes that can be cooked under 30 minutes, vegan and dairy free recipes also seem to be popular, each making up more than 10% of the total recipe collection. On the other hand, Pressure and Slow Cooker recipes make up less than 1% of the dataset, which makes them the least popular keys in the dataset. Likewise, Air Fryer recipes, freezer meals and Whole30 recipes seem to be less popular keys compared to the others.

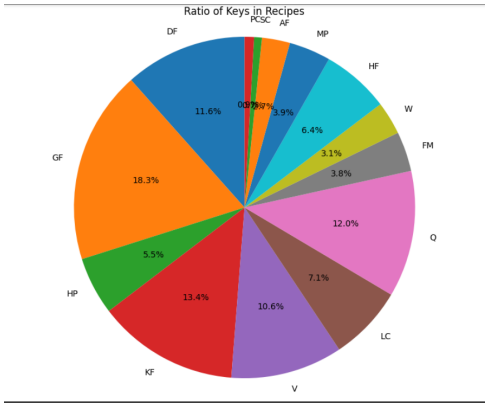

Fig. 5. Ratios of Keys in Recipes

Fig. 6. Recipe Key Dictionary

D. Further Analysis

After analysing each variable, I did further analysis, focusing on examining the relationship between recipe keys, calories, and WW points. By analyzing the average calories and points per recipe key, as well as the correlation between calories and points, we can gain deeper insights into the characteristics of different recipe categories.

The bar charts in Figure 7 and Figure 8 illustrate the average calories and WW points associated with each recipe key. The High Protein (HP) and high-fiber (HF) categories exhibit the highest average calorie content, both exceeding 300 calories per recipe. This is followed by categories like Meal Prep (MP) and Slow Cooker (SC), which also have relatively high average calories. On the lower end, recipes under keys such as vegan (V) and low-carb (LC) show much lower average calorie values, aligning with their nutritional goals.

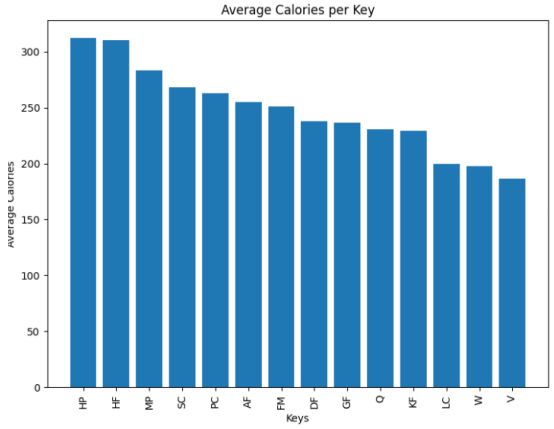

Fig. 7. Average Calories Per Key

In terms of WW points, a similar trend can be observed. High-fiber (HF), Meal Prep (MP), and High Protein (HP) recipes have the highest average points, indicating that these categories often include higher-calorie or more indulgent meals. On the contrary, categories like vegetarian (V), low-carb (LC), and Whole30 (W) recipes have lower average points, reflecting their healthier, lower-calorie focus.

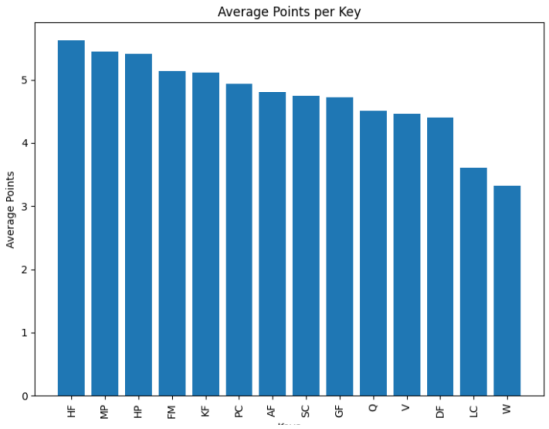

Fig. 8. Average Points Per Key

This comparison suggests that while high protein and high-fiber recipes are rich in nutrients and calories, categories like vegan and low-carb cater to users who prioritize lighter meals.

Since both figures have similar orders, I wanted to see the relationship between calories and WW points, which is given in Figure 9 below. Unsurprisingly, there is a clear positive correlation between the two variables. Recipes with higher WW points generally have more calories, with the curve steadily rising as points increase. Recipes with 10–11 points, for instance, average over 350 calories, while those with 0–2 points are closer to 100–200 calories. Although there is a decline in the average calorie value for recipes with 13 points, I think this could be ignored since there are only two recipes with 13 points.

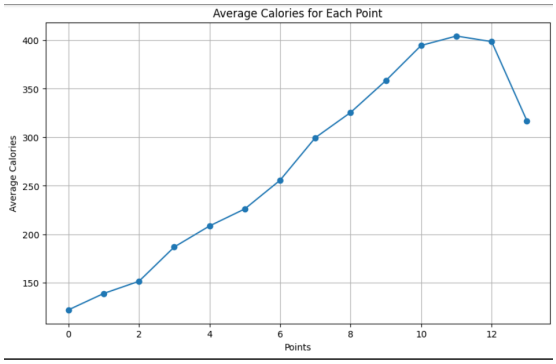

Fig. 9. Average Calories for Each Point

The analysis shows a clear relationship between recipe keys, calorie content, and WW points. Categories like high-fiber and high-protein recipes tend to have higher calories and points, whereas vegetarian and low-carb recipes maintain lower averages. Additionally, the positive correlation between calories and WW points suggests that higher-point recipes are generally more calorie-dense, although exceptions exist where non-caloric factors influence the points.

IV. Conclusion

In conclusion, the project successfully implemented a system to scrape, clean, and analyze recipe data from the Skinnytaste website. The exploratory data analysis provided valuable insights into the distribution of calories, WW points, and recipe keys. Most recipes fell into low to moderate calorie and point ranges, with right-skewed distributions indicating fewer high-calorie and high-point recipes. Popular categories such as kid-friendly, quick meals, and vegan options were well-represented, while rare categories like slow cooker and air fryer recipes were less common. Further analysis revealed strong relationships between calories and WW points, the keys with highest and lowest average calories and WW points. Finally, an interactive program, where users can filter the recipes based on the specified calories and WW points range was implemented, that can be seen in Google Colab notebook of this project.

The project encountered several bottlenecks throughout the process, particularly during the data collection phase and while handling missing or incomplete data. The main challenge was the lengthy scraping process, which took about two hours. As it was explained in Challenges in the Methodology section, this problem was solved by implementing a system to save the progress at various points by writing the scraped data to an Excel file. The pieces of dataset were then combined and the further work was done using the dataset. The other issue, recipes not having calorie or WW points, was solved with the usage of error-handling techniques.