Mini Project 1 – What should I cook tonight?

Md Sajjad Majumdar

1. Introduction

The goal of this project is to analyze recipes from a popular recipe website, Skinnytaste.com, with the aim of deciding what to cook for dinner. The aim should be to make informed choices for dinner from analyzing a recipe's calories, tags and ratings. It will undoubtedly assist those who prefer to make dietary choices based on the recipe information.

2. Data Collection

Selenium, a web automation tool was primarily used for data collection or scrapping. For the data scrapping, Skinnytaste.com was targeted as required by the assignment instructions. The aim was to collect key data such as recipe names, images, calorie, personal points, summaries, and keys. To achieve that aim, the recipe-index page is accessed first where it collects recipe links from first 5 pages. 50 random links are then selected and saved in a csv file.

Now, the scraper was programmed to extract relevant elements dynamically by inspecting the website's structure for each individual links. The scrapped entries are then loaded into a dataframe and was cleaned, validated, and stored in a CSV file for ease of use in the subsequent data analysis steps.

Challenges:

Headless Browsing: Selenium had compatibility issues with certain browser versions and was stuck multiple times while processing, especially when running in headless mode. Latest Chrome version update and adjusting settings helped resolve this.

Timeout Errors: Some pages took longer to load or failed to even connect. Multiple attempts were needed to solve this. Later this was addressed by increasing wait times and implementing retries.

Inconsistent HTML Structure: Variations were found across different pages when scraping for data. The HTML structure across different were observed and then the use of multiple CSS selectors and fallback methods helped with overcoming this problem.

Fewer Entries in CSV: After the CSV was created, there were discrepancies with the number of entries. For example, after scrapping 50 pages, only 41 entries were saved. Some pages failed to load or were missing critical elements, leading to fewer than 50 complete entries. To capture as much data as possible, other criteria were introduced.

Storing Data:

The collected data was saved in the skinnytaste_recipes.csv file for further use. This can be used to import data for analysis with the help of tools such as pandas. CSV file can also be used to view and filter recipes based on various preferences.

3. Data Analysis

The data was analyzed using various visualization techniques to explore patterns in calorie content, Personal points distribution, and recipe tags. The following visualizations highlight key findings from the analysis.

Key Visualizations:

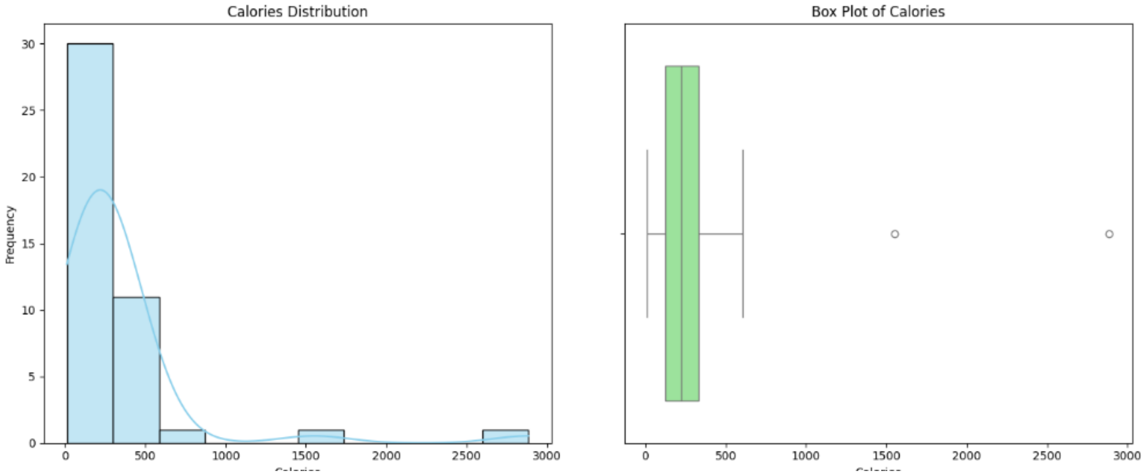

Figure 1: Calories Distribution

Histogram & Box Plot shows most of the recipes have fewer than 500 calories per serving. Only a few has more than 1000 calories visualizing the overall healthiness of the recipes.

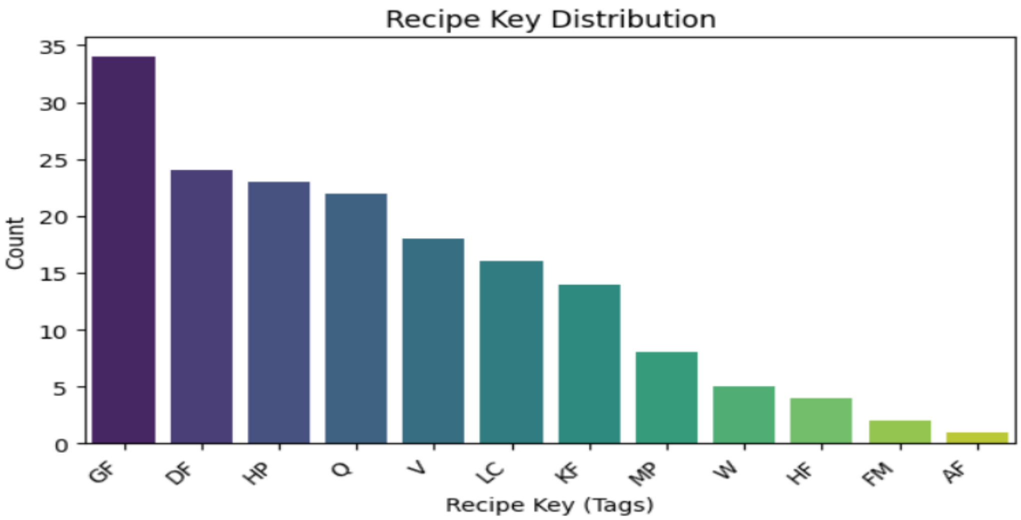

Figure 2: Recipe Key Distribution

This bar chart visualizes the most common recipe tags. The most frequent tags are **Gluten-Free (GF)** and **Dairy-Free (DF)**, indicating a trend towards dietary restrictions. Tags like **Healthy** and **Kid-Friendly** also appear frequently, showing the popularity of recipes targeting specific dietary needs.

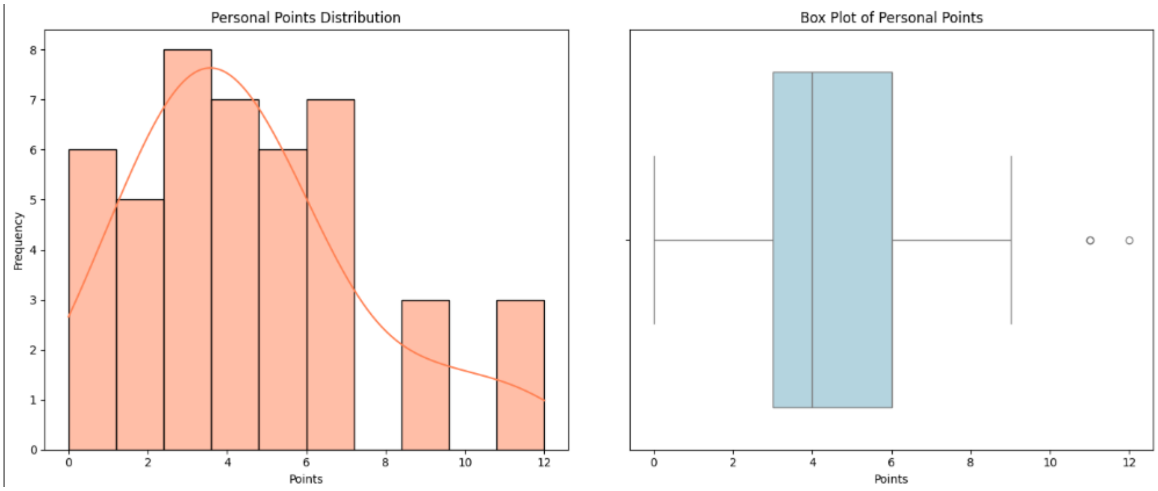

Figure 3: Personal Points Distribution

The personal points (Points) distribution shows that most recipes are within the range of **2 to 6 points**, with very few exceeding that range. This makes these recipes suitable for individuals following personal points.

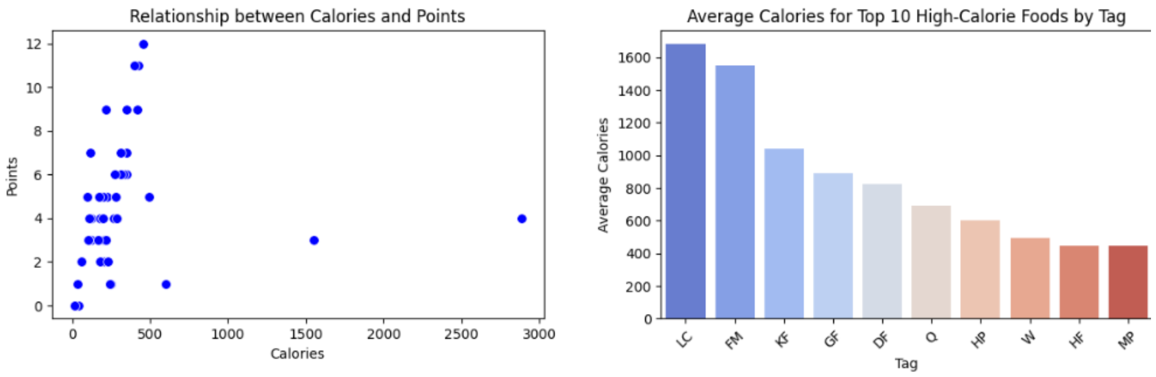

Figure 4: Relationship between Calories and Points

This scatter plot shows that, although dishes with more calories generally have more points, many recipes with comparable calorie counts have different points probably because of ingredients or portion sizes.

The bar plot shows the tags associated with the top 10 highest-calorie foods. **Low-Carb (LC)** and **Family Meal (FM)** have the highest average calorie counts, whereas recipes tagged with **Gluten-Free (GF)** and **Dairy-Free (DF)** tend to have lower calorie values.

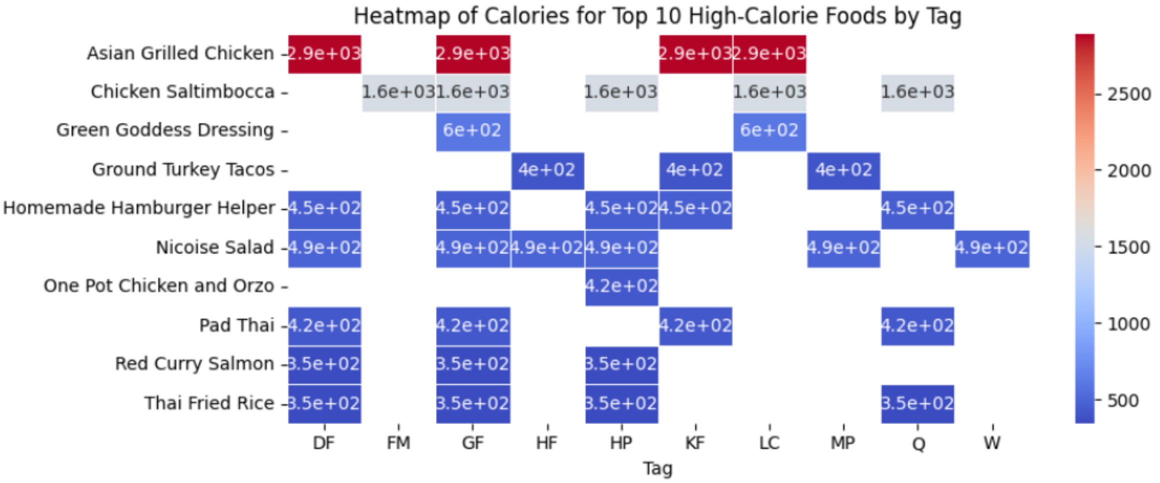

Figure 5: Heatmap of Calories for Top 10 High-Calorie Foods by Tag

This heatmap shows how the highest-calorie recipes are distributed across different recipe keys. For instance, Asian Grilled Chicken and Chicken Saltimbocca have higher calorie values, especially those tagged as LC (Low-Carb).

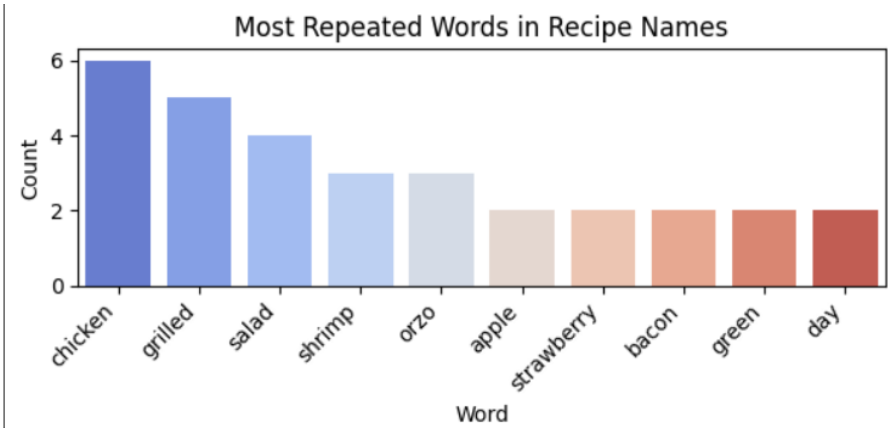

Figure 6: Most Repeated Words in Recipe Names

The most used words in recipe names are **Chicken**, **Grilled**, and **Salad** which appears frequently. This indicates popular recipe categories that the website uses to create recipes.

Observations:

From all the analysis performed, it can be said that the website skinnytaste.com mostly gravitate toward low calorie foods which are definitely preferred more by the health-conscious individuals.

4. Conclusion

Most of challenges were faced while scrapping the data from the website. As mentioned before, some of the challenges were headless browsing errors (was solved by updating the browser), timeouts (implemented retries and increasing time), inconsistent page structure (used alternative search techniques), Discrepancies in saving the entries (introduced fallout techniques), while scraping the data.

To conclude, after days of hard work, user finally found the perfect recipe to cook for dinner tonight!