What should I cook tonight?

Mahira Ibnath Joytu

Introduction:

Deciding what to cook every day is a cumbersome task, be it for students, professionals or any individual who have a lot on their plate throughout the week. More and more people are relying on the internet to find suitable and easy recipes for everyday meals. It is even more challenging to find a good recipe if someone must keep dietary restrictions, calorie intake and other relevant factors in mind. On top of that, there is a sea of recipes available online making it more difficult to land on an apt recipe for an individual. Someone who would've gone online to save time by searching up for some recipes can easily end up wasting more time looking for one which they could have used to prepare the meal itself. Many people struggle to sift through the options, which often leads to frustration and poor food choices.

Thus, personalized recipe suggestions can make it easier to stay on track with dietary goals and simplify healthy eating. By making the decision process smoother, we can promote better eating habits and improve overall quality of life.

The goal of this project is to build an easy-to-use tool that uses data analysis, web scraping, and visualization to offer personalized recipe recommendations. By collecting and analyzing data from a popular recipe website, the system can suggest meals that match the user's nutritional goals. The model will extract information like recipe names, images, calorie counts, and summaries and then filter the data based on user preferences. Ultimately, the project aims to create a user-friendly interface that makes meal planning much simpler and encourages healthier eating ultimately saving time without going through the hassle of filtering, browsing and sorting through websites.

Data Collection:

The recipes and its data have been collected from the SkinnyTaste website, a popular site for searching up recipes. The first 50 pages of the site have been scraped to extract all the recipes present within this range. Both quantitative and categorical data have been extracted for each recipe. The python library Beautiful Soup has been used to scrape the data.

The following categorical features have been extracted:

- **Name of the Recipe:** The recipe title
- **Image of the Recipe:** The URL of the source for the picture of the dish
- **Summary:** An overview of the recipe
- **Recipe Key:** A unique identifier tag for each recipe with a total of 14 variations such as AF, Q, V, and more

The following quantitative features have been extracted:

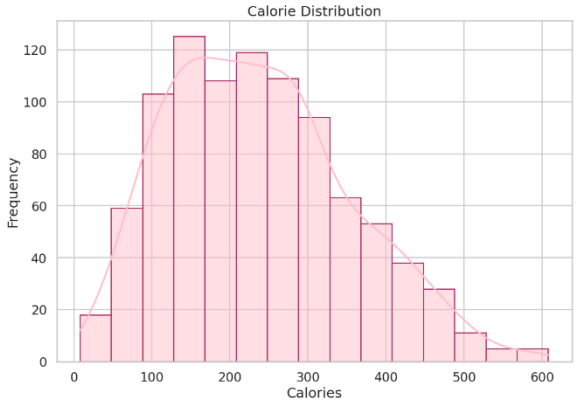
- **Calories:** The recipe's total calorie count
- **Personal Points:** A rating system within the range of 1-16 used to evaluate recipes

Data Analysis:

An exploratory data analysis has been done with the data gathered from scraping the website to find key insights and understand the data better.

a. Calorie Distribution:

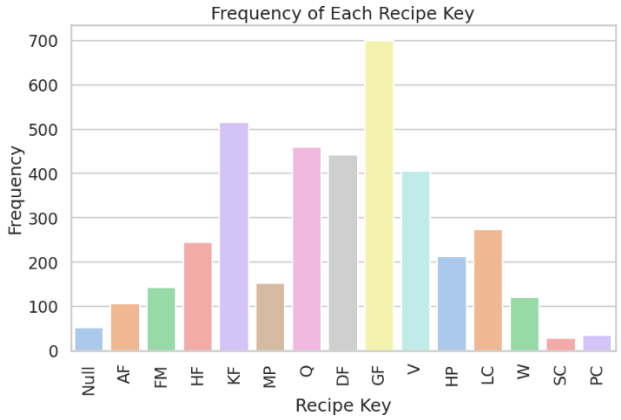
The following histogram shows the distribution of calories of all the recipes found on the website's first 50 pages. Quite visibly, the recipes are mostly distributed within the caloric range of 50 to 600. The maximum frequency of the recipes is found to be in the 150-300 range. Thus, it is evident that the website focuses on healthy and calorie-conscious recipes.



The distribution is right skewed since there are few dishes with high calorie count. This pattern suggests that the dataset offers a variety of recipes with moderate calorie counts, which is beneficial for users seeking balanced meals. The higher-calorie outliers, while less common, may provide options for those with higher energy needs or preferences for more indulgent meals. Overall, the distribution indicates a healthy spread of recipes, with a focus on moderate-calorie options.

b. Recipe Key Frequency:

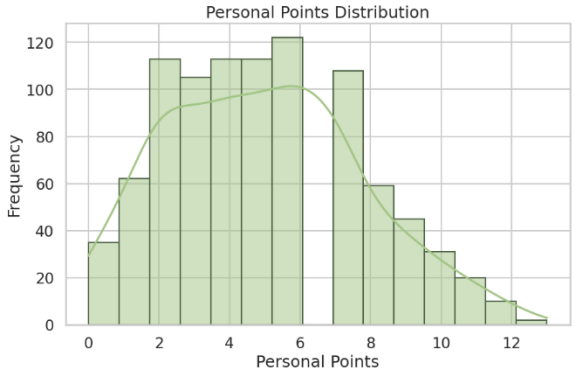
Observing the frequency plot of the recipe keys found on the website we can see that the GF (Gluten-Free) key is the most prevalent one in the dataset with over 700 occurrences. Q (Quick Meals) and KF (Kid-Friendly) are also common, with approximately 500 and 400 occurrences, respectively. These categories seem to be popular, indicating that many recipes are either quick to prepare or suitable for children.



DF (Dairy-Free), V (Vegetarian), and HP (High-Protein) categories also have relatively high frequencies, with around 400-500 occurrences each. These suggest a good number of options for users following specific dietary preferences, such as avoiding dairy or focusing on high-protein meals. LC (Low-Carb) recipes are moderately represented, with around 300 occurrences. SC (Slow Cooker) and PC (Pressure Cooker) recipes have the lowest frequencies, both under 100, suggesting that there are relatively fewer recipes designed for these cooking methods. A small number of entries were found to be Null, indicating unspecified recipe keys. Overall, the website emphasizes convenience, family meals, and specialized diets, while providing fewer options for slow or pressure-cooked meals.

c. Personal Points Distribution:

The Personal Points Distribution chart shows that most recipes in the dataset fall within the 2 to 7 points range, with a peak around 5 to 6 points, where over 120 recipes are present. Recipes with fewer than 2 points are less common, as are those with points above 10. The distribution is slightly right skewed, meaning there are more recipes with lower to mid-range personal points, while recipes with higher personal points (above 7) are less frequent. The highest frequency of recipes is concentrated in the 5 to 6 point range, suggesting that most recipes are rated around the middle of the scale.



The insights from this distribution indicate that the majority of the recipes in the dataset are rated with moderate personal points, which likely reflect a balance between healthiness and indulgence. Recipes with higher personal points (above 8) are much less common, suggesting that fewer recipes are richer or more indulgent based on whatever criteria the points measure. The spread of personal points indicates that users have a wide range of moderately rated recipes to choose from, with only a few higher-point recipes representing more indulgent or calorie-dense options.

As a whole, it seems the SkinnyTaste website provides a balanced selection of recipes, with a focus on moderate-calorie options and a strong emphasis on dietary preferences such as gluten-free, quick meals, and kid-friendly choices. It

caters to various dietary needs, including dairy-free and vegetarian options, while the personal points distribution shows that most recipes offer a healthy balance between indulgence and nutrition. Overall, the website serves as a versatile tool for meal planning, offering diverse recipes that accommodate a wide range of preferences and health goals.

Conclusion:

The task was fun, and I learned many new things related to web scraping, but multiple bottlenecks had to be overcome to reach the conclusion. The first issue I encountered was scraping the images of each recipe. I tried to extract the images from the individual recipe pages but was unable to do so due to complicated HTML tags, so I ended up extracting them from the recipe index page which was the general catalog for recipes. The source links of each image had to be extracted. The source links differed in terms of 'src' and 'srec' attributes on both pages and it was easier to extract them using the 'src' attribute from the main catalog. I figured this out after trying to extract the images multiple times from the individual recipe page and failing every time as for output I kept getting an empty array for the image link field. Then, I looked at the html code for the catalog page and attempted to extract the source link from there and was able to do so in one go so I continued with that approach. Apart from this issue, I faced difficulty in scraping the recipe keys of each recipe as they were accumulated as strings. So, I had to split them individually using the space character and then run a counter to gather the frequency of each key.

To conclude, I have a much better understanding of web scraping and its techniques after this project. I can now implement Beautiful Soup easily to extract data from websites and it was very exciting to see how I am able to gather data from websites. Moreover, I have gained some insight into the process of data accumulation and handling as a part of this process and can't help but wonder how difficult it can get to manage and store even higher volumes of data from larger website.