Mini Project-1

Scraping Data from Website

'What should I cook tonight?'

Introduction:

The goal of this project is to scrape recipe data from Skinnytaste website, process the data, extract relevant information, and perform exploratory Data Analysis (EDA) along with visualizations. Also, the project provides a user interface where users can filter recipes according to their preferred calories.

Data Collection:

The data has been scraped from the first 50 pages of Skinnytaste. Each recipe page has- Recipe name, Summary, Personal Points, Calorie, Image, Recipe key (Categorization based on dietary preferences like High Fiber (HF), High Protein (HP), Vegetarian (V), etc. These data were specifically extracted from the website. Once the relevant and target data were selected, scraper like Selenium was used to collect the data. Using Pandas (Python Library) to save the collected data into a Data Frame. And to ensure the further use of the data, the Data Frame has been saved as CSV file.

Data Preprocessing:

After generating the dataset, using python libraries like Pandas, the dataset was preprocessed for further analysis. As, there were multiple recipe keys for each recipe, the keys were encoded into categorical variables for better handling during analysis.

Exploratory Data Analysis (EDA):

Exploratory Data Analysis is a process that can investigate data and find out the pattern among them often through visualizations. Here the EDA involved analyzing the distribution of calories, points, and recipe keys. This analysis aimed to identify the relationship between these data and visualize them. Using Matpotlib, Seaborn (Python libraries) to create different plots and visualization graphs.

Calories Distribution:

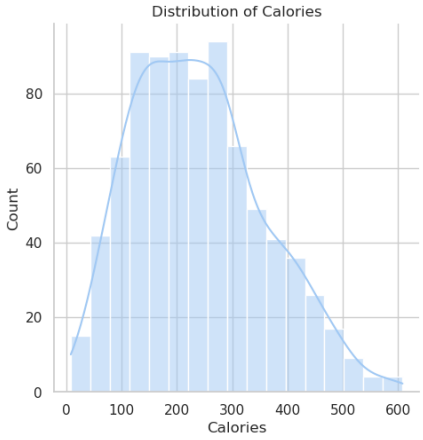


Figure 1: Calories Distribution

Mini Project-1

Scraping Data from Website

'What should I cook tonight?'

The distribution of calories (Figure 1) shows that most of the recipes fell within the range of 100 to 300 calories, reflecting the healthy focus of the website, also the data showed a skewed distribution.

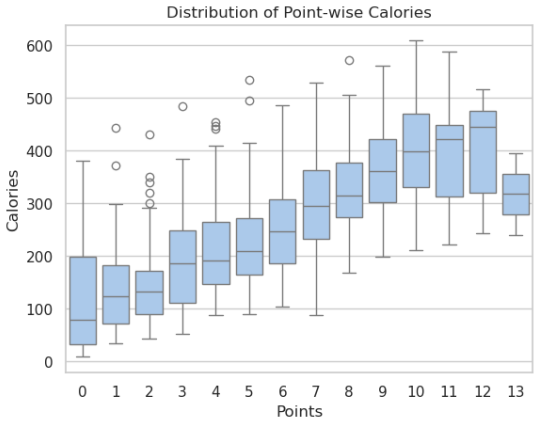


Figure 2: Point-wise Calories Distribution

Also, the data showed that (Figure 2), most of the points are given to the recipe of higher calories.

Recipe Key Distribution:

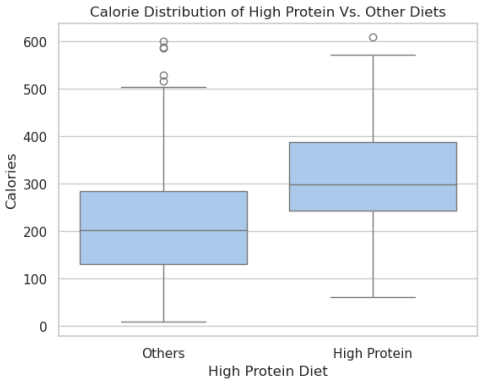


Figure 3: Calorie Distribution of High Protein Vs. Other Diets

As the recipe keys are preprocessed, it is shown (Figure 3) that based on calories that the difference of distribution of Higher Protein (HP, recipe key) and other recipes.

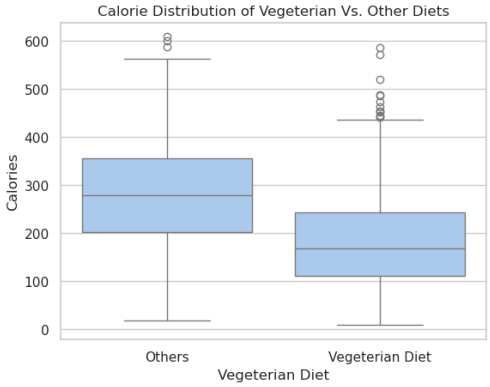


Figure 4: Calorie Distribution of Vegetarian Vs. Other Diets

The same analysis in Figure 4, shows how calories differ in various recipe keys like Vegetarian (V) recipe.

Points Distribution:

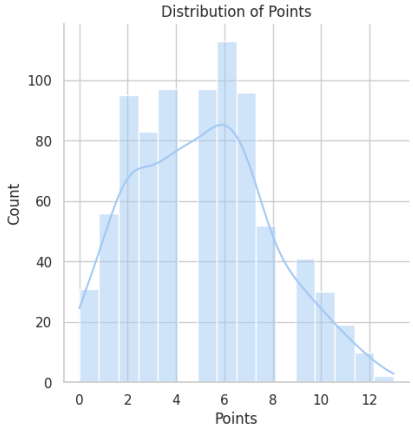


Figure 5: Points Distribution

Mini Project-1

Scraping Data from Website

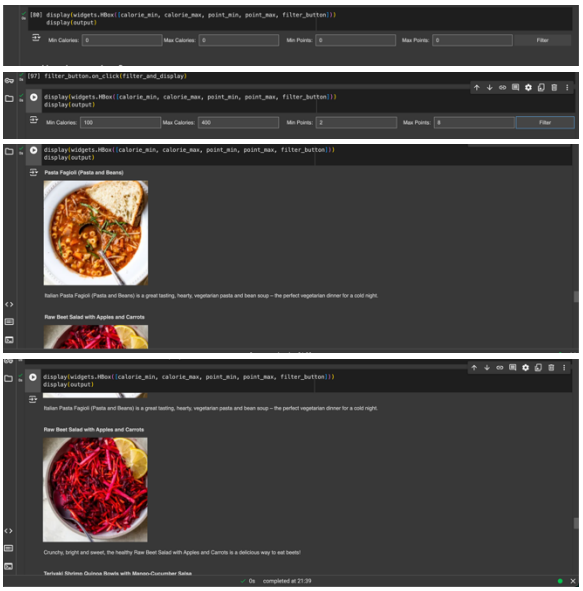
‘What should I cook tonight?’

The points distribution, which is crucial for people who are following diet programs, was visualized in Figure 5. Most recipes fell within the range of 2 to 7, which indicates the website’s focus on diet-friendly recipes.

User Interaction:

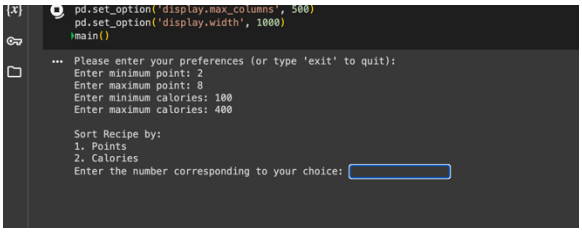
The final step involved creating a user interface that allows users to filter recipes based on their personal points and calories. The interface accepts a range of points and calories and returns 10 recipes along with recipe name, summary and images. Two interfaces were created.

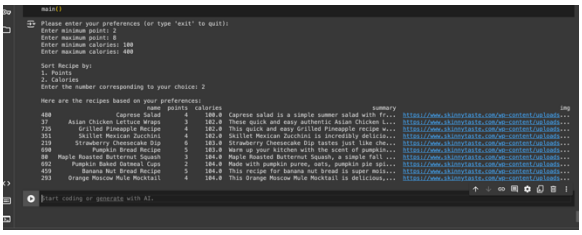
Interface-1:



Here are the outputs after filtering recipes with users preferred points and calories.

Interface-2:





In the second interface, the user can provide points and calories as input and is able to view the recipe not only sorted by calories but also by points. And here user should visit the link to check images of the desired recipe.

Conclusion:

In the project, the achieved goals successfully demonstrate how to scrape data, preprocess them, and perform proper analysis to visualize. While the project was successful, a few bottlenecks were encountered:

- **Website Structure:** The website structure was not very flexible at the beginning to find all the selected data easily like recipe keys, personal points. But, after going through it several

times it was possible to select all the data and get into it.

- **Missing Data:** On the website, many images are not accessible while scraped the data. But, later it was possible to get the image URLs, though the URLs don't work properly though.
- **User Interaction:** While building the first Interface (Interface-1), the URLs didn't work properly. At the first few executions, no images were visible.
  But after executing with an if-else condition, the Error got fixed, all the 10 images of each recipe were visible. Though before showing the images, it still shows something unwanted, but not Error.

Hopefully, further work can be done by fixing all issues, filtering more options and cleaning the data more accurately.