Webscraping using Selenium

Nurul Amin

Introduction: Provided to scrape the first fifty pages of a website where the food recipe category is the target output. Every pages of this category have a vast amount of food item. The information about these food recipes is the name of the food, image of the food, calories, personal points, summary and the receipt key needed to scrape.

The extracted data will be analyzed using exploratory data analysis techniques and proper visualization. Calories distribution, Recipe key distribution and Points distribution will be shown using histograms. Then a user interaction platform will be created where calorie range and point range would be given as an input and the first 10 foods sorted based on calories, include their image and their summary as an output.

Data Collection: For collecting the data I will use the selenium library(python), as the ease of use and can be seen the connection between the website and code. Due to using the Google Colab platform, it is not been seen the interaction between the website and code but it is in the Jupyter Notebook. There are some challenges in collecting the data:

1. Install the Selenium, webdriver and other packages.
2. The website has many pages and among them 50 pages(approximately 1000 recipe) needs to scrape. For the ease of collection, first collect all the links of every recipes from every pages by doing the pagination procedure.
3. Then from every recipe's URL, collect the filtered information(Name of the food, image of the food, calories, points, recipe key).
4. Do the cleaning task for removing duplicate and unnecessary data.
5. Do the visualization task for required criteria for data analysis.
6. Do the user interaction task

After collecting the data, CSV format would be best for storing the data from pandas dataframe. There's a sample of the datasets:

	A	B	C	D	E	F
1	**Name of the food**	**The recipe Key**	**Calories**	**Personal Points**	**Image of the food**	**Summary**
2	Crustless Pumpkin Pie	DF,GF,HF,V	124.5	2	https://www.skinnytaste.com/wp-	Skip the crust and make this easy gluten-free, dairy-free crustless pumpkin pie this fall for a cozy night when you crave a pumpkin tre
3	Red Curry Salmon	DF,GF,HP	349	6	https://www.skinnytaste.com/wp-	Thai-inspired Red Curry Salmon is simmered in an incredibly tasty coconut red curry sauce with bell peppers, garlic and onions.
4	Apple Butter	DF,GF,KF,V	34	1	https://www.skinnytaste.com/wp-	This easy apple butter recipe simmers on the stove, made with apples, cinnamon, nutmeg, and allspice, it smells and tastes like fall!
5	Roasted Delicata Squash	DF,GF,V	177	4	https://www.skinnytaste.com/wp-	Roasted Delicata Squash with toasted almonds, sweetened with maple syrup and seasoned with sage and paprika, makes a perfect f
6	Apple Bread	V	124	4	https://www.skinnytaste.com/wp-	Moist cinnamon apple bread recipe made with applesauce, small chunks of fresh apples and walnuts in every bite. Itâ€™s so moist a
7	Autumn Salad with Pears and C	GF,LC,Q,V	175	5	https://www.skinnytaste.com/wp-	I love a good salad with lots of texture and flavors, and this Autumn Salad nails it. Sweet pears, honey Dijon dressing, crunchy pecans
8	Sesame Chicken	DF,FM,GF,HP	513	9	https://www.skinnytaste.com/wp-	This lighter Sesame Chicken recipe features chicken breast bites in a sweet, savory, tangy, and slightly spicy sauce topped with sesam
9	Bacon in the Oven	DF,GF,HP,KF,LC,Q,W	60	2	https://www.skinnytaste.com/wp-	Whether youâ€™re a fan of tender, crisp, or extra crispy bacon, this simple method of cooking bacon in the oven is easy. Thereâ€™s
10	Pumpkin Spice Latte	DF,GF,V	115	7	https://www.skinnytaste.com/wp-	Embrace all the cozy fall vibes with a homemade Pumpkin Spice Latte! This easy recipe tastes just like the original, but for a fraction
11	Homemade Hamburger Helper	DF,GF,HP,KF,Q	453	12	https://www.skinnytaste.com/wp-	This one-skillet, creamy Homemade Hamburger Helper is made with ground beef, macaroni and cheese â€“ real ingredients you can
12	Fried Brown Rice	DF,GF,HF,Q,V	276	6	https://www.skinnytaste.com/wp-	Fried Brown Rice is a healthy twist on classic fried rice with some extra hidden veggies to bulk it up. Just add your favorite protein to
13	Asian Grilled Chicken	DF,GF,KF,LC	288.5	4	https://www.skinnytaste.com/wp-	This Asian Grilled Chicken recipe is the perfect excuse to grill or use your indoor grill pan, an easy high-protein dinner idea!
14	Sweet Potato Salad	DF,GF,HF,MP,V,W	290	8	https://www.skinnytaste.com/wp-	This healthy Sweet Potato Salad with avocado combines sweet, creamy, warm, and cold elements and can be enjoyed warm or cold
15	Ratatouille	DF,GF,MP,V	115	2	https://www.skinnytaste.com/wp-	Ratatouille is the perfect summer side dish for your favorite roasted chicken, grilled meats, or fish! It adds a burst of color and flavo
16	Pretzel Crusted Chicken Tende	AF,DF,FM,GF,HP,KF	267	4	https://www.skinnytaste.com/wp-	Craving something crispy, kid-friendly, and absolutely delicious? Make these easy Pretzel Crusted Chicken Tenders with honey musta
17	Lemon Vinaigrette	DF,GF,LC,MP,Q,V	123	5	https://www.skinnytaste.com/wp-	This is my go-to Lemon Vinaigrette recipe. I love it over any salad, from simple green salads, Cobb salad and even over roasted veggi
18	Slow Cooker Beef Stew	DF,GF,HP,KF,MP,SC	356	7	https://www.skinnytaste.com/wp-	Slow Cooker Beef Stew is the ultimate comfort food! Itâ€™s perfect for those crisp fall evenings or chilly winter days when you crav
19	Deviled Egg Salad	AF,GF,HP,LC,MP,V	215	3	https://www.skinnytaste.com/wp-	This lazy Deviled Egg Salad takes a classic appetizer and turns it into a quick lunch! It has the same ingredients but no piping or filling.
20	Shrimp and Rice (Arroz Con Ca	DF,GF,HP	433	9	data:image/svg+xml,%3Csvg%20xn	Costa Rican Inspired Shrimp and Rice (Arroz Con Camarones) is a traditional Costa Rican recipe thatâ€™s so easy and flavorful!
21	Bruschetta Pasta Salad	MP,Q,V	364	10	https://www.skinnytaste.com/wp-	Serve this delicious Bruschetta Pasta Salad at your next BBQ or summer party. Itâ€™s the perfect light summer dish loaded with in-s
22	Watermelon Feta Salad	GF,Q,V	85	2	https://www.skinnytaste.com/wp-	This easy Watermelon Feta Salad, made with only four ingredients, is a summer staple at your next summer BBQ!
23	Coconut Popsicles	GF,KF	106	5	https://www.skinnytaste.com/wp-	These homemade coconut popsicles are rich, creamy, and super simple to make! Theyâ€™re the perfect summertime refresher.
24	Sheet Pan Shrimp with Corn an	DF,GF,HP,Q	295	3	data:image/svg+xml,%3Csvg%20xn	This Sheet Pan Shrimp with Corn and Tomatoes is the perfect one-pan summer dinner, with easy cleanup!
25	Strawberry Yogurt Bark	DF,GF,KF,V	60	2	https://www.skinnytaste.com/wp-	If youâ€™re looking for a healthy snack with some added protein for the kids this summer, youâ€™ll love this strawberry yogurt bar
26	Broccoli Cauliflower Salad	DF,GF,LC,MP,Q,V	97	2	https://www.skinnytaste.com/wp-	This Broccoli Cauliflower Salad recipe is quick and easy, ideal for warmer weather or if youâ€™re looking for a different way to prep
27	Juicy Grilled Pork Chops	DF,GF,HP,KF,LC	272	5	[<selenium.webdriver.remote.web	Juicy Grilled Pork Chops are brined in a saltwater solution and then seasoned with a flavorful dry rub to create the most delicious po
28	Grilled Eggplant with Feta	GF,LC,Q,V	106	3	https://www.skinnytaste.com/wp-	If you need an easy side dish this summer, this grilled eggplant with feta cheese is great with anything youâ€™re grilling.
29	Korean Chicken	DF,GF,HP,Q	180	2	https://www.skinnytaste.com/wp-	These grilled Korean Chicken Breasts are juicy, sweet, and spicy! I grill them all year, using my indoor grill pan when itâ€™s cold.

skinnytaste_recipe_info (1)

Fig 1: Extracted Datasets sample

Data Analysis: For proper visualization, matplotlib's subpackage pyplot will be a good option. The distribution purposes, histogram plots are better to showing the data. A histogram is a visual representation of data presented in the form of groupings. It is a precise approach for displaying numerical data distribution graphically. It's a type of bar plot in which the X-axis shows bin ranges and the Y-axis represents frequency (wiki). I will observe the calories distribution, point distribution and recipe key distribution from the total datasets of recipes. From this visualization, can demonstrates which food item will be best for health purposes and any customer can take their food from the exact matches like what calorie ranges he/she needed.

Some visualization sample:

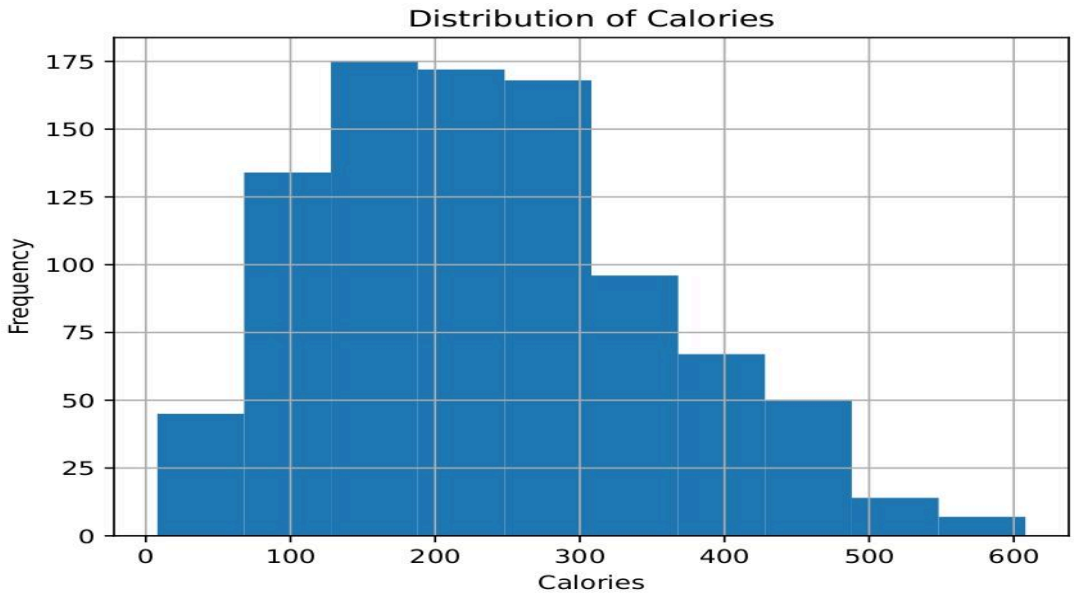

Fig 2: Calories distribution of the datasets

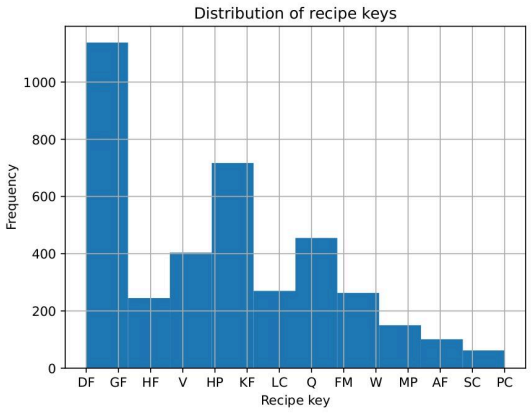

Fig 3: Recipe key distribution

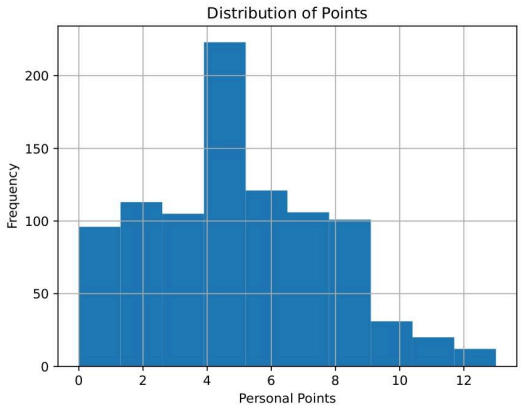

Fig 4: Points distribution

Conclusion: A website “https://www.skinnytaste.com/” , from this site I collected the desired data of some food items(approximately 1000 items). During the pagination steps, the next button didn’t work properly and then I change my policy. I use the URL directly and do the pagination by increasing the page number till 50 page. I had become bored and tied when I want to collect the image of every recipies. The image source had different criteria for different recipe and needed a vast amount of condition, but afterall some images of some recipies I couldn’t collect correctly. During the visualization task, ‘recipe key’ column had a problem like it is not a numerical data. And in my collected data every food item had a multiple recipe keys. So, the visualization task didn’t work correctly for recipe key distribution. Then I separated the column and also separated the recipe keys to different rows and lastly it could be a proper visualization. However, During the user interaction task, took the calories range and point range as an input and shows the first 10 sorted data (descending order) based on calories also filtered it by the point range.
Here’s a sample of user interaction task:

User interaction

```
cal1 = float(input("calorie range: "))
print(cal1)
point = float(input('Points range: '))
output = df[(df["Calories"]<cal2) & (df["Personal Points"]<point)]
sorted_df = output.sort_values(by=['Calories'], ascending=False)
sorted_df.head(10)
```

```
calorie range: 50
50.0
Points range: 3
```

	Name of the food	The recipe Key	Calories	Personal Points	Image of the food	Summary
967	Air Fryer Hard Boiled Eggs	AF,DF,GF,KF,LC,Q,V,W	78.0	0.0	https://www.skinnytaste.com/wp-content/uploads...	Making hard boiled eggs in the air fryer is so...
983	Hard Boiled Eggs	DF,GF,KF,LC,Q,W	77.0	0.0	[<selenium.webdriver.remote.webelement.WebElem...	If you want to know how to hard boil eggs, thi...
283	How To Grill Corn on The Cob	DF,GF,KF,V	77.0	0.0	NaN	Grilled corn on the cob is an easy, classic su...
517	Blueberry Frozen Yogurt	GF,KF,V	76.0	2.0	https://www.skinnytaste.com/wp-content/uploads...	Blueberry Frozen Yogurt, made with only five i...
253	Waffled Zucchini Fritters	DF,GF,KF,LC,V	75.0	2.0	NaN	Dust off that waffle iron and make these savor...
474	Raspberry Coconut Chia Pudding Popsicles	DF,FM,GF,LC,V	74.0	2.0	https://www.skinnytaste.com/wp-content/uploads...	These easy chia pudding popsicles with coconut...
298	Artichoke Dip	GF,KF,LC,Q,V	73.7	2.0	https://www.skinnytaste.com/wp-content/uploads...	This creamy, delicious artichoke dip is hot an...
259	Coleslaw Recipe	DF,GF,LC,Q,V,W	73.0	2.0	NaN	This easy coleslaw recipe is the perfect side ...
732	Sauteed Collard Greens with Bacon	DF,GF,LC,Q	73.0	2.0	https://www.skinnytaste.com/wp-content/uploads...	Thinly sliced collard greens sautéed with baco...
55	Grape Tomato Sauce	DF,GF,KF,LC,Q,V	73.0	1.0	[<selenium.webdriver.remote.webelement.WebElem...	When grape tomatoes are overflowing in your ga...

Next steps: Generate code with sorted_df | View recommended plots | New interactive sheet

Fig 5: User interaction task