Data science project 1 – report

In the 21st century, a growing number of people experience problems with weight gain or weight management. People go on diets every day, and a lot of them end up either quitting early, or dieting in an unsustainable way for a period of time and afterwards gaining some or all of the lost weight back. The key to sustainable weight management is eating foods that taste great, yet are low enough in calories so that you can eat enough to feel full and satisfied. In today's world there is an abundance of information available online. This project investigates the website www.skinnytaste.com to help retrieve data that can assist in finding suitable weight management recipes.

The Skinnytaste website consists of pages with blog-style posts that include recipes, and sometimes posts related to weight management in some other way. In this project I am interested in the recipes only. Each recipe has some associated data: They list the calories of the food, a brief summary on what the food is, Weight Watchers points (each recipe has a blue, green and a purple score), tags that give some information on what type of food is in question (e.g. keto, vegan, kid-friendly and more), and an image of the finished product. For this project I use the Beautifulsoup Python library to collect the name, summary, calories, WW points and image of the recipe. The Python script then writes all data to a csv file, to be used for analysis and user interaction.

To help understand the data, I use the Seaborn library to visualize information. In this project, chose to plot the calorie distribution to a histogram. This helps the user know whether Skinnytaste may be suitable for his or her own weight management purposes, since different people have different calorie needs, so knowing how many recipes there are for each calorie range is important. The Python script also allows the user to input a calorie and WW point range, and finds recipes for the user within the provided constraints.

This project came with some challenges. The first one was to find structure in the data source. Since there were both recipes and other blog posts with the same structure, it was challenging to find what makes them different in order to filter out the unwanted blog posts. This was solved when I noticed that the unwanted blog posts did not have icons for WW points or calories, and I could use that information to collect only the posts with the data I needed.

Another challenge was the fact that Beautifulsoup library is not particularly fast. This was seen when testing the script; I usually had to wait for quite a while for the script to provide me with an output to determine if the script works as desired, which slowed down the development process. However since each of the 30 pages I scraped from Skinnytaste were more or less

identical, I solved this by simply scraping the first page only for my test runs, and then expanding to the rest of the pages later when I needed to validate that my script worked for the entire data set.