I’ve got two Python projects under my belt. The first looked at crime in Los Angeles and the second analyzed my cycling performance. Both have been progressive steps up. I didn’t know Python before starting the first project, and for the second, I was accessing my first API. With a solid foundation, I wanted to move another step further – web scraping.

If you’ve made it this far in my portfolio I think it would be fair to say I like cycling. For this project, I decided to scrape customer reviews from my favorite cycling apparel companies, OrNot. They’re based in California, most of their products are made in the US, no massive logos adorn their clothes, and their jerseys are downright super comfortable. So here we go!

I’m going to be using a library I’ve never used before – Selenium. This will allow me to interact with the webpage like a normal user and close popups, scroll down, click on the next page, etc. I began by collecting all the URLs of the products I would need to scrape. This was a really simple loop once I figured out there were six pages of products.

INSERT CODE

With all of the product URLs collected, I now needed to construct another loop to iterate over each URL. For each URL I wanted to collect the following information:  
  
 user – person who submitted review

review\_post\_date – date the review was submitted

location\_of\_review – user’s location

star\_ratings – user’s grade of the product on a scale of 1-5

review\_headers – title of the review

review\_body – actual review

fit\_rating – user’s grade of the product’s fit on a scale of 1-5

products – the name of the product the user was reviewing

size\_ordered

reviewers\_height\_and\_weight

I was able to inspect the webpage and tell Selenium how to find each element. If it couldn’t find an element (someone didn’t provide their height or weight for example), it would store N/A instead. Once the elements were found, the program clicked to find the next page of reviews for that product, and if there were no more pages of reviews it moved to the next URL. After completing this loop for each URL, I saved the results to a .csv.

INSERT CODE

This proved to be a rather finicky process. Occasionally the program would break because a pop-up was activated and needed to be clicked. Other times I think Selenium had some difficulty knowing where to scroll to. It also collected, for a reason I’m still unsure, hundreds of rows of completely null data. Ultimately though, I was able to collect every review for every item listed in the men’s collection on the OrNot website.

So began the cleaning process. The first and most obvious step was to remove any rows that were completely null. I also wanted to make sure I didn’t collect any duplicate entries so I dropped those as well. From there I went through the star\_ratings, review\_headers, review\_body, and products, to find any nulls, and if there were any, dropping them. The main goal of this project is to analyze each of these categories, and if there are any nulls, they’re essentially worthless. All told, after the collection and cleaning process, I had 5, 285 reviews to work with.

INSERT CODE

I began with some really simple data exploration.

INSERT CODE/GRAPH

Over all reviews, OrNot averaged 4.86 stars out of 5, which is really impressive. I’m glad I’m not alone in my love for their products.

The fit of their clothing was just as spot on too.

INSERT CODE/GRAPH

The average fit was 2.90, with 3 being a perfect fit. So if anything, their clothes fit a bit snug but not by much.

I also wanted to see where most of these reviews were coming from.

INSERT CODE/GRAPH

Basically all were from the US at 94.98%.

We should also check out how many reviews they’ve gotten each year.

INSERT CODE/GRAPH

It’s interesting to see their explosion in reviews in 2020 followed by another big amount in 2021 before they start to back off in 2022. I know cycling-related sales really took off across the whole industry in 2020 thanks to the pandemic. I remember it was really hard getting certain mechanical parts, but it’s cool to see that OrNot probably benefited from the influx of new cyclists as well. I’d love to know how closely the amount of reviews correlates to their actual sales figures.

I’m also curious to see which of their products are the most reviewed, but we’ve run into a bit of an issue because the two most reviewed products are House Bib Shorts – Black and House Bib Shorts – Stone Blue which are the same product, just different color. While it’s obviously good information to know that they’ve received more reviews for the black than stone blue, I want a more macro look. Which product categories (jerseys, bibs, jackets, shirts, etc.) get the most reviews?

INSERT CODE/GRAPH

After grouping their products into broader categories we find that jerseys get the most reviews followed by bibs/tights and then jackets/vests. I was honestly a bit surprised to find that the Other category which includes things like tools, lights, and bar bags had more reviews than socks/caps/hat/gloves. Maybe this is because OrNot gives away free mini bar bags on first orders over $99? Something to investigate in the future perhaps.

Let’s see the average star rating for each of these product categories.

INSERT CODE/GRAPH

And all categories have a rating of 4.81 or higher. No real surprise there given that all reviews for the men’s collection averaged 4.86.

Lastly, I’m curious about the sizing for products. I’ll need to do some more grouping as the sizes for the jerseys is different from the shorts which is different from the pants.

INSERT CODE/GRAPH  
  
It seems most reviewers are around a medium.

With all of the exploratory analysis done, I can move on to working with the actual reviews. I’m saving that for another write-up though because this is already long and took lots of work. Part 2 coming soon!

In this project I was able to scrape the OrNot company’s website to get a collection of all their customer reviews and performed some cursory analysis to get a feel for customer satisfaction. It turns out I’m not alone in really loving OrNot. No matter the product category, their quality and fit all consistently rate really highly. It’ll be a lot of fun to dig into greater detail into the actual reviews!