scrapplier

A python package to scrape school and products information from the supplier website all across the United Kingdom. Scrapplier built on top of `undetected-chromedriver` and works well on dynamic websites. In the current version it supports more than 15 suppliers and can scrape the data from School level to Variant level

Installation

- Open your terminal and create a new environment

  ```
  $ python -m venv scrapplier-venv
  ```

- Activate your environment

  ```
  $ source scrapplier-venv/bin/activate
  ```

- Install dependencies

  ```
  (scrapplier-venv) $ pip install -r requirements.txt
  ```

- Clone scrapplier repo

  ```
  (scrapplier-venv) $ git clone git@github.com:hariesramdhani/scrapplier
  ```

- Install scrapplier

  ```
  (scrapplier-venv) $ cd scrapplier
  (scrapplier-venv) scrapplier $ pip install .
  ```

Example

```python
from scrapplier.scraper import Scraper

scraper = Scraper(username="test", password="test")

# Scrape Monkhouse
scraper.scrape(supplier="monkhouse")
```

Depending on the depth that you chose you will get the data for them

- `monkhouse_schools.csv`: School information, including school logos and school pages on the supplier website, parameter `depth="schools"`
- `monkhouse_products`: Products information, all of the products that are being sold to specific schools, parameter `depth="products"`
- `monkhouse_variants`: Products variant information, all of the product variants, parameter `depth="variants"`

Scraping logic (Lay terms)

Scraping using `undetected-chromedriver` (Selenium) works like a robot that mimics how a human would use a web browser to gather information from a website. Here's a simple breakdown of how it works, especially when scraping data from a supplier's website:

1. **Automating the Browser**:
   Selenium controls a web browser (like Chrome or Firefox) as if someone were clicking, typing, and scrolling on the website. It can open websites, click buttons, fill out forms, and even navigate through pages—just like a person would.

2. **Navigating to Supplier Websites**:
   First, Selenium goes to the supplier's website, for example, a school uniform supplier. It "loads" the website, just as you would by typing the address into your browser.

3. **Finding the Information**:
   Once the website is loaded, Selenium looks for specific elements on the page (like product names, prices, or images). It uses "selectors" (kind of like a map) to find these elements, which could be hidden in things like buttons or drop-down menus.

4. **Dealing with Dynamic Content**:
   Many modern websites are "dynamic," meaning parts of the page don't load immediately but after a few seconds or when you scroll down. Selenium waits for

this content to appear, ensuring it gathers everything that's loaded, unlike simpler methods that might miss this information.

5. **Collecting the Data**:
   Once Selenium finds the information (like product lists, prices, and school details), it "scrapes" or copies that data into a format you can use, such as a CSV file or a database.

6. **Handling Complex Interactions**:
   Some suppliers might require special actions, like logging in or clicking on specific school links. Selenium can handle these by filling in login forms or registering students, so you can access products specific to a school.

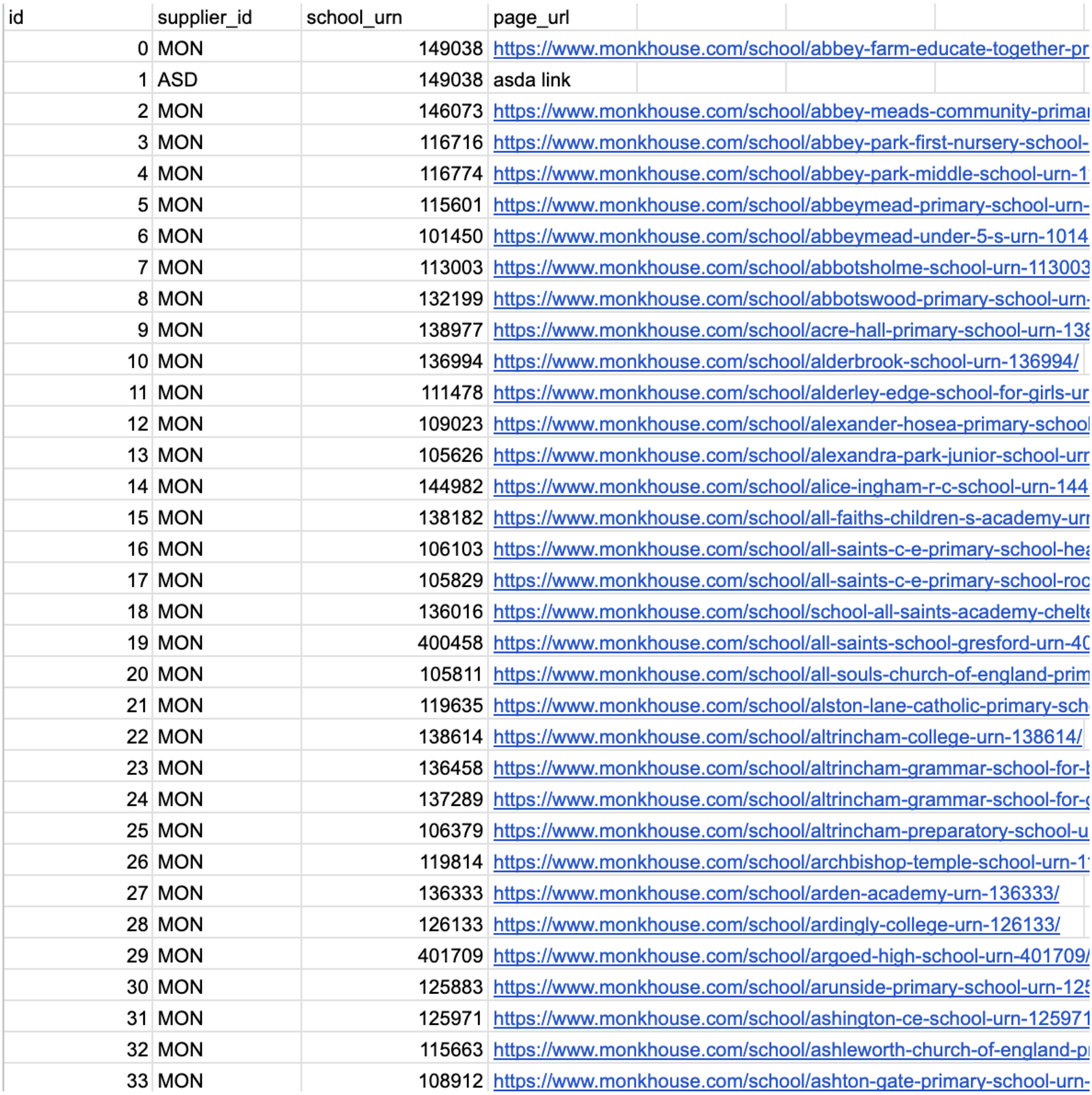
In short, Selenium acts like a human web browser user, navigating the supplier's website, gathering product details, and making that data available for analysis or comparison.

Directory information

- `scrapplier`: The python package
- `data`: All of the data collected and generated during the study
  - `raw`: Raw data (output from Scrapplier)
  - `processed`: Python processed data to match the requirements (e.g. for database or for Matt's analysis)
    - `flat_file`: Contains example of flat file as requested by Matt
    - `database`: Database for the website

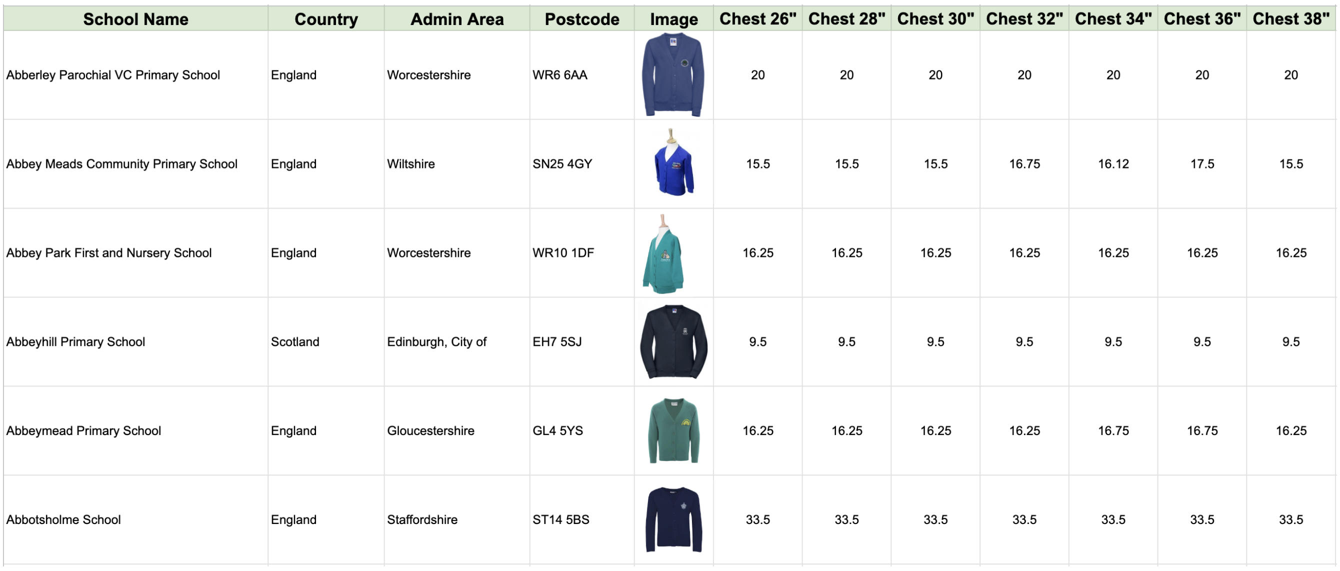
Data Example

- Raw data



id	supplier_id	school_urn	page_url
0	MON	149038	https://www.monkhouse.com/school/abbey-farm-educate-together-pr
1	ASD	149038	asda link
2	MON	146073	https://www.monkhouse.com/school/abbey-meads-community-primar
3	MON	116716	https://www.monkhouse.com/school/abbey-park-first-nursery-school-
4	MON	116774	https://www.monkhouse.com/school/abbey-park-middle-school-urn-1
5	MON	115601	https://www.monkhouse.com/school/abbeymead-primary-school-urn-
6	MON	101450	https://www.monkhouse.com/school/abbeymead-under-5-s-urn-1014
7	MON	113003	https://www.monkhouse.com/school/abbotsholme-school-urn-113003
8	MON	132199	https://www.monkhouse.com/school/abbotswood-primary-school-urn-
9	MON	138977	https://www.monkhouse.com/school/acre-hall-primary-school-urn-138
10	MON	136994	https://www.monkhouse.com/school/alderbrook-school-urn-136994/
11	MON	111478	https://www.monkhouse.com/school/alderley-edge-school-for-girls-ur
12	MON	109023	https://www.monkhouse.com/school/alexander-hosea-primary-school
13	MON	105626	https://www.monkhouse.com/school/alexandra-park-junior-school-urn
14	MON	144982	https://www.monkhouse.com/school/alice-ingham-r-c-school-urn-144
15	MON	138182	https://www.monkhouse.com/school/all-faiths-children-s-academy-urn
16	MON	106103	https://www.monkhouse.com/school/all-saints-c-e-primary-school-hea
17	MON	105829	https://www.monkhouse.com/school/all-saints-c-e-primary-school-roc
18	MON	136016	https://www.monkhouse.com/school/school-all-saints-academy-chelte
19	MON	400458	https://www.monkhouse.com/school/all-saints-school-gresford-urn-40
20	MON	105811	https://www.monkhouse.com/school/all-souls-church-of-england-prim
21	MON	119635	https://www.monkhouse.com/school/alston-lane-catholic-primary-sch
22	MON	138614	https://www.monkhouse.com/school/altrincham-college-urn-138614/
23	MON	136458	https://www.monkhouse.com/school/altrincham-grammar-school-for-b
24	MON	137289	https://www.monkhouse.com/school/altrincham-grammar-school-for-g
25	MON	106379	https://www.monkhouse.com/school/altrincham-preparatory-school-u
26	MON	119814	https://www.monkhouse.com/school/archbishop-temple-school-urn-1
27	MON	136333	https://www.monkhouse.com/school/arden-academy-urn-136333/
28	MON	126133	https://www.monkhouse.com/school/ardingly-college-urn-126133/
29	MON	401709	https://www.monkhouse.com/school/argoed-high-school-urn-401709/
30	MON	125883	https://www.monkhouse.com/school/arunside-primary-school-urn-125
31	MON	125971	https://www.monkhouse.com/school/ashington-ce-school-urn-125971
32	MON	115663	https://www.monkhouse.com/school/ashleworth-church-of-england-p
33	MON	108912	https://www.monkhouse.com/school/ashton-gate-primary-school-urn-

- Flat file



School Name	Country	Admin Area	Postcode	Image	Chest 26"	Chest 28"	Chest 30"	Chest 32"	Chest 34"	Chest 36"	Chest 38"
Abberley Parochial VC Primary School	England	Worcestershire	WR6 6AA		20	20	20	20	20	20	20
Abbey Meads Community Primary School	England	Wiltshire	SN25 4GY		15.5	15.5	15.5	16.75	16.12	17.5	15.5
Abbey Park First and Nursery School	England	Worcestershire	WR10 1DF		16.25	16.25	16.25	16.25	16.25	16.25	16.25
Abbeyhill Primary School	Scotland	Edinburgh, City of	EH7 5SJ		9.5	9.5	9.5	9.5	9.5	9.5	9.5
Abbeymead Primary School	England	Gloucestershire	GL4 5YS		16.25	16.25	16.25	16.25	16.75	16.75	16.25
Abbotsholme School	England	Staffordshire	ST14 5BS		33.5	33.5	33.5	33.5	33.5	33.5	33.5

Supplier support

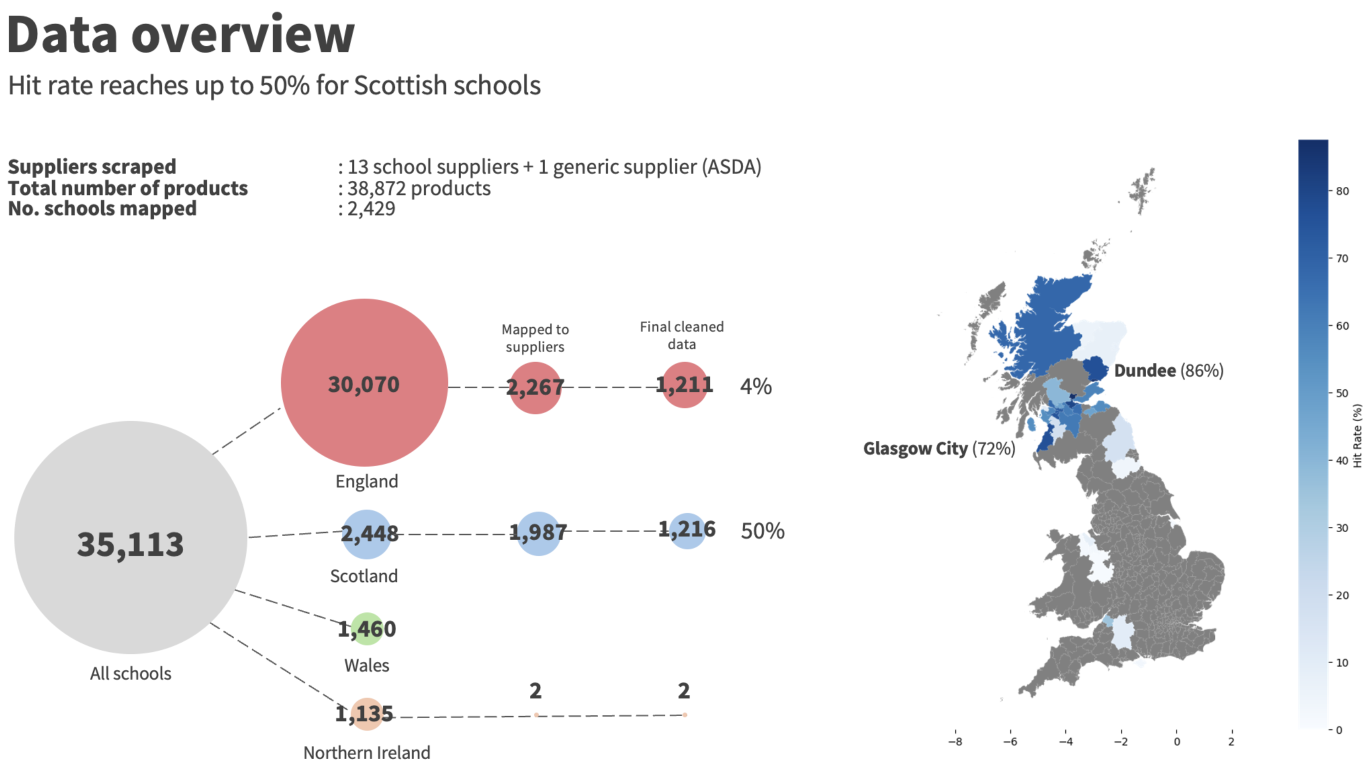
In the current version of `scrapplier` it supports the scraping of 18 suppliers, the details about the depth and counts of products and schools can be found below (as of June 2024)

id	supplier_code	supplier_name	supplier_website	website_template	school_cnt	product_cnt	product_variant_cnt
0	MON	Monkhouse	https://monkhouse.com		1,035	20,936	188,424
1	BSW	Blossom Schoolwear	https://www.blossomsschoolwear.com/	Shopify	261	4,807	30,284
2	SME	Schoolwear Made Easy	https://www.schoolwearmadeeasy.com	Shopify	679	4,703	29,628
3	SCS	Scotcrest Schools	https://www.scotcrestschools.co.uk/		192	2,402	15,132
4	MGS	MacGregor Schoolwear	https://macgregorschoolwear.co.uk	Woocommerce	159	488	3,074
5	MYC	MyClothing	https://myclothing.com/		7,728		
6	SUS	School Uniform Scotland	https://schooluniformscotland.com/	Woocommerce	13	710	4,473
7	AAG	Aspire Academy Glasgow	https://aspireacademyglasgow.com/		40	766	4,825
8	AAS	Alan Santry Schoolwear	https://www.alansantryschoolwear.co.uk/		38	260	1,638
9	BOE	Border Embroideries	https://www.border-embroideries.co.uk/		1,374	10,611	
10	DIS	Direct Schoolwear	https://directschoolwear.co.uk/		76		
11	STE	Stevensons	https://www.stevensons.co.uk/		699		
12	TRU	Trutex	https://www.trutex.com/				

id	supplier_code	supplier_name	supplier_website	website_template	school_cnt	product_cnt	product_variant_cnt
13	BAN	Banner	https://www.banner.co.uk				
14	DAL	David Luke	https://www.davidluke.com/				
15	UND	Uniform Direct	https://www.uniform-direct.com/		358		
16	TFS	Top Form Schoolwear	https://www.top-form.co.uk/		58	846	5,329
17	SMS	Smart Schoolwear	https://www.smartschoolwear.co.uk/		91		
18	PIS	Pinder Schoolwear	https://pindersschoolwear.com/		263		

Initial findings

Overview of the cleaned data that can be used for the analysis



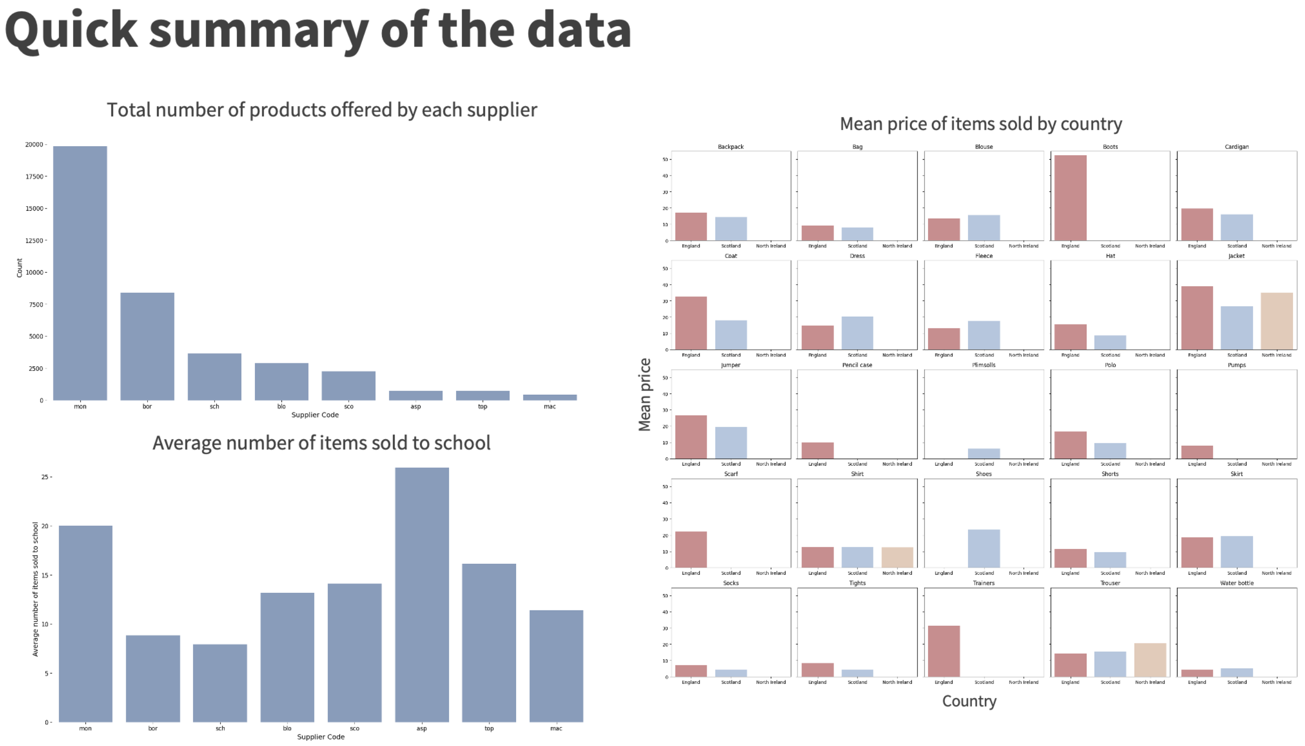
Data overview

Hit rate reaches up to 50% for Scottish schools

Suppliers scraped : 13 school suppliers + 1 generic supplier (ASDA)
Total number of products : 38,872 products
No. schools mapped : 2,429

Overview of what can be done with the data

Quick example of descriptive stastics of the data



What have been done so far

- ☑ Scrape more than 10+ suppliers data
- ☑ Try to connect the scraped data to the school database
- ☑ Create the website where we can show the comparison between the generic and supplier price
- ☑ Give an example of what kind of analyses can be done with the data

Limitations of the current approach

Here's a rephrased version of your points:

- Some suppliers do not clearly indicate whether a product is mandatory or optional for purchase.
- Accessing certain suppliers' products can be challenging; for example, with Stevensons, you must register your child with a specific school before viewing their products.
- Supplier information may be outdated, meaning some suppliers may stop providing uniforms for certain schools or begin supplying them.
- Prices may fluctuate, although the frequency of these changes is uncertain.
- Some schools, particularly in England, are not covered by the current suppliers, so the pool of suppliers we scrape from needs to be expanded.
- Some data may be incomplete or inaccurate and will require additional cleaning and validation.
- Product categorization in the flat file is overly simplistic, relying on basic string matching that may need refinement.
- Legality of web-scraping isn't very clear (Grey area).

Potential analyses that can be done with the current data

- **Supplier Coverage**: Analyze which suppliers cover the most schools or regions and identify gaps in uniform availability.
- **Cost Difference**: Compare the cost difference between generic uniforms and supplier-branded uniforms for specific schools.
- **Price Trends**: Track and compare price changes for uniforms across different suppliers over time.

Future Directions and Areas to Explore

1. **Expand Supplier Pool**
   Continue adding new suppliers to the scraping process, especially those covering regions or schools currently not included. This will ensure comprehensive coverage, particularly in underrepresented areas like England. Additionally, expanding beyond 18 suppliers could improve the variety of products and pricing options available for analysis.

2. **Enhance Data Validation and Cleaning**
   Implement more advanced data cleaning and validation techniques to handle incomplete or inaccurate data more efficiently. This could involve automated data quality checks, missing value imputation, and refining the categorization process to move beyond basic string matching.

3. **Improve Product Categorization**
   Develop more sophisticated product categorization methods, such as natural language processing (NLP) or machine learning techniques, to better group products based on attributes like school type, uniform type, or seasonality. This would provide more accurate analysis and reporting.

4. **Handle Dynamic and Restricted Access Websites**
   Explore alternative ways to scrape websites like Stevensons that require user registration or dynamic interaction. Possible solutions could include developing

more advanced scraping algorithms, investigating API access, or collaborating directly with suppliers for data sharing.

5. **Monitor Price Fluctuations and Trends**
   Implement a system to track and monitor price changes over time for both generic and supplier-branded uniforms. This would allow for detailed price trend analysis and insights into how frequently prices change and which suppliers tend to have more volatile pricing.