Zalando Pre-owned: "thrift it like macklemore"

Documentation Zalando Pre-owned Web scraping & dataset (Team 2)

1. MOTIVATION

1.1 • For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?

Background information

Over the past few years, the topic of sustainability has gained a lot of attention. It is well known that in order to preserve the planet for future generations, sustainability needs to be implemented in everyday life. Therefore, a lot of companies have started to implement sustainability into their organizations. This can even be beneficial for companies, as when they are able to implement good sustainability practices and adequately diffuse them towards their stakeholders, it can in the medium term become a source of competitive advantage and as a consequence lead to value creation (Sicoli, Bronzetti & Baldini, 2019).

The industry with the most need for sustainable practices is the fashion industry. The overall fashion industry, when it comes to the production of clothes, has displayed one of the highest levels of negligence concerning exploitation of the workforce, social well-being and drainage of the worlds natural resources in the past (Henninger et al., 2019). Luckily, new trends have emerged in the fashion industry which could decrease its negative environmental impact. One of such trends is the utilization of second-hand clothing. Second-hand clothing leads to environmental and financial advantages as it reduces the material, water usage, production costs, and landfill spaces needed to create new clothes (King & Wheeler, 2016). As a consequence, the trade of second-hand clothing has highly increased over the past few years through consumer-to-consumer platforms such as Vinted, Facebook Marketplace, and Marktplaats, or through business-to-consumer platforms such as Sellpy and Thredup. What all these platforms have in common is that they are specifically dedicated to the sale of second-hand items.

In more recent years, a new trend has started to emerge, namely existing clothing brands re-selling their own brand's clothes previously owned by consumers. Examples of this are *Zalando's "Pre-owned"*, *NA-KDs "Tweedehands"* and *Xan Woman's "Pre-loved"*.

The dataset

For this dataset, it was decided to focus on the website Zalando in The Netherlands, more specifically their "Pre-owned" section specified to men's clothing. Zalando started to offer this feature on their website in The Netherlands in October 2020 (Duurzaam Ondernemen, 2020). Compared to the previously mentioned websites, Zalando is an interesting choice as it offers clothes from more than 4500 brands, regular offerings and second-hand offerings, whereas a lot of other stores only offer their own brand as second-hand.

The dataset will contain data from Zalando's Pre-owned men's clothing section and Zalando's Regular men's clothing section. This dataset was created to enable research within the topic second-hand clothing. With this dataset it could for example be investigated which brands are mostly offered for second-hand clothing and which sizes and types of products are most popular in this segment. Moreover, this dataset enables researchers to determine similarities and differences between Pre-owned and Regular men's clothing offerings. Also, when comparing this data over a period of time, new trends can be analyzed and it can be determined whether it is beneficial for a clothing store to offer their own clothes as second-hand to consumers. Furthermore, as Zalando offers its products in more than 14 countries, the dataset of The Netherlands could be compared to other countries to see if there are country-specific differences.

1.2 • Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was developed by a project group of the course Online Data Collection and Management as part of the Master Marketing Analytics at Tilburg University.

1.3 • Who funded the created of the dataset? If there is an associated grant, please provide the name of the grantor and the grand name and number.

There was no funding or grant for the development of this dataset, as this dataset is a result of the web scraper which the project group built.

Additional comments

Web scraper vs. API

For this data collection project, it was chosen to use a web scraper to gather the data. This decision was made because of several reasons.

Firstly, it was decided to use web scraping for this project as the use of web scraping provides more flexibility in collecting data compared to the use of an API, since the Zalando API was a fixed format to collect data and also limited. In other words, web scraping provides more possibilities in the retrieval of specific information, and thus is more customizable in comparison to an API. Also, the Zalando API was a bit complex to get a good understanding of how to work with it.

Another important detail to take into account is the fact that the product offering of Zalando is dynamic, meaning that the product offering changes over time (e.g. per day, per hour, etc.). Web scraping is, in this case, more convenient as this method enables the researcher to access the data available on a particular moment. As a conclusion, web scraping makes it possible to gather data in form of momentarily observations, which will enable the researcher to compare product offering over time.

2. COMPOSITION

2.1 • What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)?

The instances within the dataset(s) are men's clothing items extracted from the Zalando website in The Netherlands. Within the instances a distinction is made between two types of men's clothing items, "Regular" men's clothing items and "Pre-owned" men's clothing items, as this dataset should enable researchers to conduct research on possible differences and similarities between them.

For the extracting of instance the product overview pages* are used as so called seeds for data extraction:

- Zalando Pre-owned Men's: https://www.zalando.nl/pre-owned-kleding-heren/?order=activation_date
- Zalando Regular Men's: https://www.zalando.nl/herenkleding/

From the product overview pages product URLs are extracted to derive product information for specific products (read: clothing items). When scraping Zalando "Pre-owned" Men's product overview pages the sorting option "order=activation_date" was used. When scraping Zalando "Regular" Men's product overview pages the sorting option "populair" was used. In section 3.3 more detail is given on this as it is incorporated in the sampling strategy used for the data collection.

* Product overview pages are the pages containing the total offer of products on a website.

2.2 • How many instances are there in total (of each type, if appropriate)?

Zalando supplies their products within Europe via different websites adapted to the language of the specific countries. The supply of products Zalando offers remains the same for each of the countries. As the instances in this dataset are spilt between "Regular" men's clothing items and "Pre-owned" men's clothing items, an estimation had to be made on how many instances each of the clothing sections contain in total. Table below shows how many instances each clothing section contained in total on March 4th 2022. Keep in mind that this is just a momentarily observation as the number of offerings will differ over time (e.g. per day, per hour, etc.).

Type of instance	Total number of instances
Regular men's clothing items	117,582 items divided over 15 different clothing categories
Pre-owned men's clothing items	68,442 items divided over 6 different clothing categories

(Zalando, 2022-1)(Zalando, 2022-2)

2.3 • Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated / verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances from Zalando's “Regular” men’s clothing items and Zalando's “Pre-owned” men’s clothing items.

The maximum number of product overview pages shown on the Zalando website is 428. The decision on the sample size was based on the computation of the technically feasible sample size. In the calculations the total number of product overview pages, the number of products and scraping time are taken into account. It was decided to scrape 50% of all product overview pages, meaning that data was collected from 214 product overview pages for both the Zalando's “Regular” men’s clothing section and the Zalando's “Pre-owned” men’s clothing section. Each product overview page consist of 85 products, which results in a total sample of 18,190 products per clothing section. In total data is collected from 36,380 products. It takes 35 seconds to scrape all product URLs on a product overview page. From the calculations became clear that it will take 2 hours to scrape all products from 214 product overview pages. This has to be done for both clothing sections, which leads to a total of 4 hours of scraping. This was deemed technically feasible.

However, after scraping, the team ended up with a sample size of 17,976 for Zalando's "Pre-owned" men's clothing section and 8,090 for Zalando's "Regular" men's clothing section, which corresponds to a sample of 49,41% and 22.23% of the population respectively. For the "Pre-owned" clothing section this is still very close to the initially intended 50%, but for the "Regular" clothing section this is quite a big decrease. As the web scraper did work for the "Pre-owned" section, this probably has something to do with the code on the Zalando website. The precise cause is not investigated as this was complicated and beyond the scope of this course.

2.4 • What data does each instance consist of? “Raw” data (e.g., unprocessed text or image) or features?

For each instance, a “Pre-owned” men’s clothing item or "Regular" men's clothing item, the following data is collected:

Variable	Description	Type of data
Pre-owned	Whether the item is pre-owned or not	Unprocessed text
Product type	The type of clothing item (e.g., t-shirt, sweater, pants)	Unprocessed text
Brand name	The name of the brand which made the clothing item	Unprocessed text
Size	The size of the clothing item	Unprocessed text
Price	The price at which the clothing item is offered	Unprocessed text
Color	The color of the clothing item	Unprocessed text
Delivery time	The indicated time it takes to deliver the product to the customer	Unprocessed text

As mentioned in section 2.2, the product offering on the Zalando website will differ over time because products are sold and new products are uploaded. This is one reason to include the date of scraping into the dataset as well. Another reason to include the date of scraping is that it enables researchers to investigate trends over time.

2.5 • Is there a label or target associated with each instance?

There is made a distinction between whether or not an clothing item is pre-owned. As clarified in section 2.1, the data is collected for men's clothing with the distinction between "Regular" men's clothing items and "Pre-owned" men's clothing items. To make this distinction as clear as possible in the web scraper the variables collected for the "Pre-owned" men's clothing section end with "po" as an indication to the "Pre-owned" label. The variables for the "Regular" men's clothing section do not have a special denotation. For both "Regular" men's clothing and "Pre-owned" men's clothing an individual dataset was constructed, so further labeling within the datasets was not necessary.

2.6 • Is any information missing from individual instances?

There is no missing information from individual instances observed. However, one important observation to address is that for the variable "delivery_time" 30 instances with the dataset for Zalando "Pre-owned" have the value 'no delivery time'. This value means that these products are sold out and therefore the delivery time is unknown.

2.7 • Are relationships between individual instances made explicit (e.g., users movie ratings, social network links)?

None explicit relationships between individual instances is experienced, as the instances collected are only products with their own product information. There could be a possible relationship based on brand name, for example when products are from the same brand. However, it is not considered as a relationship on itself.

2.8 • Are there recommended data splits (e.g., training, development / validation, testing)?

The data is already split into two separate datasets: (1) Zalando "Pre-owned" men's clothing and (2) Zalando "Regular" men's clothing. Due to the data collection method, it is already possible to use the data on the different men's clothing sections separately. This data collection method provides researchers with more possibilities for different research purposes.

2.9 • Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant over time; b) are there official archival versions of the complete dataset (i.e., including external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained and does not rely on any external resources. The data within the dataset is exclusively collected from the Zalando website (zalando.nl).

2.10 • Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The data within this dataset is not considered confidential, as there is no user-related information extracted during the collection of the data. All data in the dataset is publicly available on the Zalando website.

2.11 • Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

Not applicable. The dataset does not contain any offensive, insulting, threatening or anxiety causing information.

2.12 • Does the dataset relate to people?

As mentioned in section 2.10, there is no user-related information extracted during the collection of the data.

Additional comments

Intentionally removed information

As mentioned before data is collected for both the "Pre-owned" and "Regular" men's clothing section on Zalando. In section 2.4 the variables which would be collected are addressed. The original plan was to collect data on the following variables: product url, brand name, product type, price, size, delivery time and color. However, during the programming process the team discovered some problems collecting the sizes of the products within the Regular men's clothing section on Zalando. The data collection for the variable size was made complicated by a dropdown menu, as the sizes for each particular product in this section are displayed in this dropdown menu format. To be able to collect this information, more advanced web scraping techniques are required. Unfortunately, these techniques are beyond the scope of this course. Therefore, the assumption was made that for the regular offering of men's clothing, all sizes are available.

Data insights

For each dataset, for both Zalando's "Pre-owned" men's clothing and "Regular" men's clothing, basic summary statistics have been executed.

"Pre-owned" men's clothing section

In total 17,976 observations have been gathered. There is data collected for the following variables: product url, brand name, product type, price, size, delivery time and color. The total dataset for Zalando's "Pre-owned" consists of 17,976 observations and 9 variables.

The mean price is for all products is €23.89. In the dataset 625 different brands occur. The dataset contains in total 73 different categories, types of clothing. Within the different categories it is possible to combine certain categories as some are in line with others, for example "Sweater", "Trui", "Sweater met rits" and "Hoodie". Furthermore, 247 colors or color combinations occur in the dataset, of which 'Blue' occurs the most. Moreover, 247 different sizes are included in the dataset. Finally, for delivery time products mostly have a delivery time of 1-4 or 3-6 working days and for some products the delivery time is unknown.

Regular" men's clothing section

In total 8,090 observations have been gathered. There is data collected for the following variables: product url, brand name, product type, price, delivery time and color. The total dataset for Zalando's "Regular" consists of 8,090 observations and 8 variables.

The mean price is for all products is €76.86. In the dataset 533 different brands occur. The dataset contains in total 93 different categories, types of clothing. Within the different categories it is possible to combine certain categories as some are in line with others. Furthermore, 2198 colors or color combinations occur in the dataset, of which 'Black' is the most occurring color. Finally, for delivery time products mostly have a delivery time of 1-2 or 3-6 working days.

3. COLLECTION PROCESS

3.1 • How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified?

The data is acquired by scraping both Zalando’s "Pre-owned" men's clothing pages and Zalando’s "Regular" men's clothing pages. For both Zalando’s "Pre-owned" and "Regular" the following two steps are used in the scraping process:
1. Product overview pages are scraped to gather product URLs.
2. The product URLs are used to gather product information of each product. The data from the product description was directly observable as product type, brand name, size, price, color and delivery time are all displayed on the website for consumers.

3.2 • What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The mechanism used to collect the data from the Zalando website is a manually programmed web scraper. The web scraper is programmed in a Jupyter Notebook.

During the programming of the web scraper a combination of *Selenium* and *BeautifulSoup* is used. Selenium is used to scrape the different product URLs of the product overview pages as the products on these web pages load when scrolling down the page. When web scraping, it is not possible to manually scroll down the pages. Selenium allows automating web browsers, which was very useful in this case. BeautifulSoup, on its turn, is used to scrape the product pages and extract product information from them.

3.3 • If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The data in this dataset is part of a larger dataset, namely the dataset Zalando maintains their product offerings with. As mentioned in section 2.2 and 2.3, the number of product offerings on the website of Zalando (both men's "Pre-owned" and men's "Regular") is dynamic. Meaning that the total number of instances will change over time, maybe per day or even per hour. As a conclusion, a fixed number of instances can never be set. For this matter it was chosen to base the sampling strategy on a momentarily observation of the number of instances. As the number of instances in this momentarily observation is still very large, it was chosen to see how many instances could be scraped within a reasonable timeframe of 2 hours per clothing section. As explained in section 2.3, it was chosen to use a sample of all product overview pages available, for both Zalando "Pre-Owned" and "Regular", to maintain a feasible scraping time. For both clothing sections 214 product overview pages are scraped after which the product URLs retrieved from these product overview pages are scraped to gather the product information.

Sorting options on Zalando

When viewing the Zalando website, there are several different ways in which you can sort the clothing items. Sorting can be done based on the highest popularity, the newest products, the lowest or highest price, or based on whether they are in the sale or not. The chosen sorting option will determine the sample of the data.

Zalando "Pre-owned"

For this data collection project, it was chosen to select items based on the newest Pre-owned products first. This sorting option is indicated in the URL by "order=activation_date". It was decided to work with this sorting option as it is the default option when viewing the product overview page. This will result in a dataset containing the most recent product offering on Zalando "Pre-owned".

Zalando "Regular"

It was chosen to select items based on the most popular products/brands first. This sorting option is the default option when viewing the product overview page. It was decided to use this sorting option as this provides us with information on the most popular products/brands on Zalando "Regular".

Conclusion

By using this combination of sorting options, researchers will be able to research for example if popular brands/products from Zalando "Regular" also appear on Zalando "Pre-owned" and to what extent.

3.4 • Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

People involved in the data collection process were five students from the Master Marketing Analytics of whom none of they were compensated in terms of money. However, these students did get compensated with gaining a lot of new knowledge during this project, which they all can use in future projects and jobs.

3.5 • Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset related to the Zalando "Pre-owned" men's clothing section is constructed on March 19th 2022. The dataset related to the Zalando "Regular" men's clothing section is constructed on March 22nd 2022.

3.6 • Were any ethical review processes conducted (e.g., by an institutional review board)?

When selecting a source to collect web data from, among others, the ethical risks should be evaluated. Firstly, it was evaluated whether Zalando prohibits web scraping by observing the robots.txt for Zalando. This does not state anything against web scraping. Moreover, information on the website and terms and conditions do not mention anything specific about web scraping the website. However, Zalando does have a public API and takes part in Open Source Development in which they share their code and processes with the rest of the world as they want technology to benefit as many people as possible (Zalando, 2022-3). Based on this it was concluded that Zalando would allow web scraping of their website. Moreover, since user data was web scraped or the data collected is not used for commercial purposes, this project was considered to be ethical.

3.7 • Does the dataset relate to people?

As mentioned in section 2.10 and 2.11, there is no user-related information extracted during the collection of the data.

4. PRE-PROCESSING, CLEANING, LABELING

4.1 • Was any pre-processing/cleaning/labelling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Two separate datasets - one for Zalando's "Pre-owned" men's clothing section and the other for Zalando's "Regular" men's clothing section are obtained during the data collection for this project. When inspecting the data, a few issues were discovered:

● In both datasets the variable "delivery-time" was displayed as a variable with the class 'character'. This variable should be categorical as the variable has different levels, such as '1-4' and '3-6' days. Therefore, this variable is transformed into a factor.

● Also the variable "date" did not have the right class. Therefore, an extra variable ("date_ymd") was added as a variable with the class 'date'.

● The variables containing the metadata have been relocated to the first columns in the datasets.

● In the dataset for Zalando's "Regular" men's clothing items, de variable "price" had to be adjusted. For some instances it stated the price preceded with 'vanaf'. This is for clothing items for which the larger sizes are more expensive. This word was removed, leaving only the actual price.

4.2 • Was the “raw” data saved in addition to the pre-processed/cleaned/labelled data (e.g., to support unanticipated future uses)?

Yes, all raw data is saved in CSV files. In total two raw dataset are created:
● product_description_pre_owned.csv: contains the data of the scraper for the Zalando’s "Pre-owned" men's clothing page(s).
● product_description_herenkleding.csv: contains the data of the scraper for the Zalando’s "Regular" men's clothing page(s).

Both CSV files can be found on the repository of this project: https://github.com/jjacobs123/thrift-it-like-macklemore. Also the cleaned datasets can be found on this repository by the names: "product_description_po_cleaned.csv" and "product_description_reg_cleaned.csv".

4.3 • Is the software used to pre-process/clean/label the instances available?

As mentioned earlier in section 3.2, the web scraper is manually programmed in a Jupyter Notebook. The Jupyter Notebook which contains the source code can be found on: the following link: https://github.com/jjacobs123/thrift-it-like-macklemore.

5. USES

5.1 • Has the dataset been used for any tasks already?

This dataset has not been used for any tasks before, as this dataset is specially developed as a project for the course Online Data Collection and Management as part of the Master Marketing Analytics at Tilburg University.

5.2 • Is there a repository that links to any or all papers or systems that use the dataset?

There is a repository that gives access to all documents that are related to the dataset or needed to work with and get an understanding of it. The repository is publically available and can be found via the following link: https://github.com/jjacobs123/thrift-it-like-macklemore.

5.3 • What (other) tasks could the dataset be used for?

This dataset could be used for several purposes. This dataset can, for example, be used to conduct research to second-hand clothing in general, using other datasets as well to get an broader view of the general picture. Besides that, this dataset can also be used to gather insights in the differences between "Regular" clothing items and "Pre-owned" clothing items with a specific focus on men's clothing. But can also be combined with other datasets that contain for example data of women's clothing. These are just simple examples of what can be done with the data. This data, however, can serve as a bases of many other research directions.

5.4 • Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labelled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks). Is there anything a future user could do to mitigate these undesirable harms?

As explained earlier, the Zalando website is a dynamic website. This means that the website regularly gets updated, with new clothing items being added and out of stock or out of season clothing items being removed. Also, class names might be changed due to updates on the Zalando website. The dataset generated in this project is from one specific moment in time. Therefore, the dataset might not be up to date anymore when used in the future. Nonetheless, the code used to generate the dataset will still work in the future, it will just generate another dataset. There is no way to mitigate this "problem", but that is also not necessary as it will not impact future uses. In contrast, it can actually be beneficial as it enables research to analyze how the dataset changes over time.

5.5 • Are there tasks for which the dataset should not be used?

This data is collected to analyse, gather insights, and finally draw conclusions based on the findings of the analyses. The data in this dataset should only be used for research purposes. The data in this dataset cannot and should not be used to commit plagiarism or anything closely related to it.

6. REFERENCES

- Duurzaam Ondernemen. (2020). *Zalando Lanceert "Pre-Owned" in Nederland*. Duurzaam Ondernemen. https://www.duurzaam-ondernemen.nl/zalando-lanceert-pre-owned-in-nederland/
- Henninger, C. E., Alevizou, P. J., Goworek, H., & Ryding, D. (Eds.). (2017). *Sustainability in fashion: A cradle to upcycle approach*. Springer.
- King, J., & Wheeler, A. (2016). *Setting the record straight*. Recyclingwasteworld. https://www.recyclingwasteworld.co.uk/opinion/setting-the-record-straight/147367/
- Sicoli, G., Bronzetti, G., & Baldini, M. (2019). The importance of sustainability in the fashion sector: Adidas case study. *International Business Research*, 12(6), 41-51.
- Zalando. (2022-1). *Zalando Pre-owned kleding heren*. Zalando.nl: https://www.zalando.nl/pre-owned-kleding-heren/?order=activation_date
- Zalando. (2022-2). *Zalando Herenkleding*. Zalando.nl: https://www.zalando.nl/herenkleding/?order=activation_date
- Zalando. (2022-3). *Zalando Open Source.* https://opensource.zalando.com