TILBURG UNIVERSITY

Online Data Collection and Management (328061-M3)

Tilburg School of Economics and Management (TiSEM)

Tilburg University

Team Project

Web Scraping AliExpress

Unveiling Market Trends in Consumer Electronics

Author: T.J. van der Schaaff M.Sc.

Student ID: 2062861

A research project report following the requirements for Tilburg University Master of Science in Marketing Management.

March 25, 2024

CONTENTS

LIST OF FIGURES

Chapter 1

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. The dataset was created to analyse market trends and consumer preferences in the consumer electronics category on AliExpress, one of the world's largest e-commerce platforms. The primary objective was to collect comprehensive data on various consumer electronics products, including their titles, prices, customer reviews, and seller information. By scraping and analysing this data, we aimed to gain valuable insights into the demand, pricing strategies, and customer sentiment surrounding different types of electronic devices sold on AliExpress (Insights, 2024).

The specific task in mind was to develop a robust web scraping methodology to extract relevant data from AliExpress product listings and transform it into a structured dataset suitable for further analysis. This dataset would enable researchers, businesses, and market analysts to study the consumer electronics market on AliExpress, identify popular product categories, compare prices across sellers, and understand customer opinions based on product reviews (Research & Ltd., 2023).

At present, there is a lack of publicly available datasets that provide detailed information on consumer electronics products sold on e-commerce platforms like AliExpress. While AliExpress itself may have internal data on product sales and customer behaviour, this information is not readily accessible to external researchers and analysts. By creating this dataset through web scraping, we aimed to fill this gap and provide a valuable resource for anyone interested in studying the consumer electronics market on one of the largest global e-commerce platforms (Grand View Research, 2023).

Moreover, the dataset may well serve as a foundation for various other AliExpress web scraping research projects and commercial applications. For instance, the dataset may be used to:

1. Identify trending products and emerging technologies in the consumer electronics space (Rita & Ramos, 2022).

2. Analyse pricing strategies employed by different sellers and brands (Verghese, 2023).
3. Study customer sentiment and consumer preferences based on product reviews (Cirqueira et al., 2020).
4. Compare product offerings and prices across different regions or countries.
5. Develop machine learning models for product recommendation or demand forecasting (Speckmann, 2021).

In closing, the primary motivation behind creating this dataset was to provide a comprehensive and accessible resource for studying the consumer electronics market on AliExpress. By filling the gap in publicly available data and enabling various research and business applications, this dataset aims to contribute to a better understanding of consumer preferences, market trends, and e-commerce dynamics in the electronics industry (DeVito, Richards & Inglesby, 2020).

Chapter 2

COMPOSITION

What do the instances that comprise the dataset represent?
The instances in the dataset represent consumer electronics products listed on the AliExpress e-commerce platform. Each instance corresponds to a unique product listing, containing various attributes such as `product id`, `URL`, `type`, `title`, `price`, `currency`, `units sold`, `thumbnail image URL`, `store URL`, `store name`, and `store IDs`.

Are there multiple types of instances? Please provide a description.
The dataset primarily contains one type of instance, which is product listings for consumer electronics items. However, there is some variation in the specific types of electronics products, such as relays, earphones, microphones, game consoles, and projectors, as evident from the `title` data column.

How many instances are there in total?
The dataset consists of 59 instances in total, each representing a distinct consumer electronics product listing on AliExpress.

Is the dataset a sample or does it contain all possible instances?
The dataset appears to be a sample of consumer electronics product listings from AliExpress, rather than an exhaustive collection of all possible listings in this category. The relatively small size of the dataset (i.e., 59 instances) suggests it is a curated subset rather than a comprehensive catalogue.

What data does each instance consist of?
Each instance in the dataset consists of the following attributes:

- `id`: A unique numeric identifier for the product listing.
- `url`: The URL of the product listing page on AliExpress.
- `type`: The type of the listing (e.g., *'natural'* or *'ad'*).

- `title`: The title or name of the product.
- `price`: The listed price of the product.
- `currency`: The currency in which the price is denominated (e.g., USD).
- `trade`: The number of units sold (with some missing values).
- `thumbnail`: The URL of the product thumbnail image.
- `store_url`: The URL of the store selling the product.
- `store_name`: The name of the store selling the product.
- `store_id`: A numeric identifier for the store.
- `store_ali_id`: Another numeric identifier for the store.

Most of the data is in string format, with some numeric columns like `id`, `price`, `store_id`, and `store_ali_id`. The `trade` column contains the sales volume but has some missing values.

Is any information missing from individual instances?
Yes, the `trade` column which indicates the sales volume has 4 missing values out of the 59 total instances. Apart from this, the dataset appears to be largely complete, with no other missing values in the remaining columns.

Are there any errors, sources of noise, or redundancies in the dataset?
No obvious errors, noise, or redundancies are apparent from the provided summary of the web scraped product dataset. However, more in-depth data validation and cleaning would likely benefit the integrity and consistency of the insights obtained through the analyses.

Are relationships between individual instances made explicit?
Relationships between instances are not made explicit in the data as provided. Each row represents an individual product listing without overt connections to other listings. That said, products could be related by virtue of being sold by the same store (i.e., as indicated by the `store_url`, `store_name`, and `store_id` data columns) or by being of a similar product type as indicated in the titles.

In brief, the dataset comprises a sample of 59 consumer electronics product listings from the AliExpress platform, with each instance consisting of 12 attributes that describe key details about the product and the selling store. While largely complete, the dataset does have some missing values in the `trade` data column. Relationships between products are not made explicit but could potentially be inferred from shared store details or similar product types.

Does the dataset contain data that might be considered confidential?
No, the dataset does not contain any confidential data. All the information has been scraped from publicly available AliExpress product listings.

Does the dataset contain data that might be offensive, insulting, threatening, or might otherwise cause anxiety?
The dataset itself is unlikely to contain any offensive or insulting content, as it captures objective information about products. However, it is possible that certain products, if viewed directly on the AliExpress site, may have marketing content that could be considered offensive to some individuals.

Does the dataset identify subpopulations? Is it possible to identify individuals from the dataset?
No, the dataset does not identify any subpopulations or individuals. The data is solely focused on consumer electronic product information and does not capture any personal details about sellers or customers.

Does the dataset contain data that might be considered sensitive?
No, the dataset does not contain any sensitive data. The captured information is limited to publicly available product details and metadata.

Any other comments?
The dataset provides a snapshot of consumer electronics products available on AliExpress at the time of scraping. Moreover, the dataset can be used to analyse product categories, price ranges, seller details, and other related information. However, as the data only represents a single crawl, it does not capture any temporal trends or changes in the product listings on the site over time.

Chapter 3

COLLECTION PROCESS

How was the data associated with each instance acquired?
The data for each consumer electronics product instance on AliExpress was acquired through direct observation. Specifically, the raw text data describing each product's attributes, including `title`, `price`, `number sold`, `thumbnail image URL`, `store name`, etc. was programmatically scraped from the consumer electronics subpage comprising the AliExpress webpage. No data were reported by subjects or indirectly inferred. The directly scraped observable data were considered validated and verified since the data came straight from the authoritative source of the AliExpress webpage itself.

What mechanisms or procedures were used to collect the data?
The data collection process was facilitated by a software program developed in Python, utilising the `BeautifulSoup` and `Selenium` libraries. Furthermore, `BeautifulSoup` was used for HTML content parsing, while `Selenium` was used for automating web page interactions. Together, these libraries allowed for navigation through dynamically loaded content within the context of the AliExpress website.

If the dataset is a sample from a larger set, what was the sampling strategy?

A *"Stratified Random Sampling"* strategy was employed to aggregate the product data from the *"Consumer Electronics"* product category from the AliExpress webpage, into a single dataset. The data collection procedure focused on scraping product data from products that are both popular and highly rated, with the aim of accurately reflecting prevalent consumer preferences and technological trends. This stratified random sampling strategy was used to scrape the product listings sequentially from the first few result pages until a sample size of around 119 products was reached. No further sampling probabilities were employed in the data collection procedure.

Who was involved in the data collection process, and how were they compen-

sated?

The data collection process was carried out solely by the graduate researcher, named T.J. van der Schaaff, as part of an academic project for an online data collection and management course. No other parties were involved, and no compensation was provided.

Over what timeframe was the data collected?
Between the March 20th, 2024, and March 25th, 2024, the data collection process was carried out. The collected data represent a recent snapshot of consumer electronics listings on AliExpress at that time. The timeframe mentioned above, matches when the scraped product data and cases were present on AliExpress' online retail platform.

Were any ethical review processes conducted?
No formal ethical review or institutional oversight was necessary, as the dataset does not involve human subjects or sensitive data. The data was scraped from public webpages without logging in or circumventing restrictions, and scraping was done at a reasonable rate to avoid burdening AliExpress servers.

Does the dataset relate to people?
The dataset comprises only product information originating from the consumer electronics category of AliExpress' webpage, and does not identify individuals or people. To exemplify, the resulting dataset includes only superficially related data on AliExpress stores and AliExpress sellers by including store names and store URLs. However, no personal details about the sellers themselves were stored in the resulting dataset due to respect for the privacy of the people and individuals on AliExpress' online retail platform.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?
The data were directly extracted and aggregated from the AliExpress webpage, ensuring an unaltered and accurate representation of the information as presented on the platform.

Were the individuals in question notified about the data collection? Did the individuals in question consent to the collection and use of their data?
As the data collection and management research project exclusively targets product infor-

mation without collecting personal data, notification to and consent from individuals were not applicable or required.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
Given the absence of any sort of personal data collection process within this AliExpress web scraping research project, issues regarding consent revocation and impact analysis on data subjects do not apply.

Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?
No formal data protection impact analysis has been conducted, as the dataset does not contain personal data on individuals. The scraped product information is publicly available on the AliExpress website.

Any other comments?
The custom *R* code script used for web scraping and the resulting dataset will be made publicly available in the web scraping project's data repository on GitHub to ensure transparency and reproducibility. To respect AliExpress's web servers and avoid overloading their systems, the script was designed to scrape data at a reasonable rate with appropriate delays between requests. The script can be easily modified and re-run to collect a larger or updated dataset if needed for future research, while still adhering to web scraping best practices and ethical considerations.

Chapter 4

PREPROCESSING/CLEANING/LABELLING

Was any preprocessing/cleaning/labelling of the data done (e.g., discretisation or bucketing, tokenisation, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, several preprocessing, cleaning, and labelling steps were performed on the AliExpress consumer electronics dataset:

- ➢ Missing values were identified for each column using the `is.na()` function in *R*. The output shows the number of missing values in each column.
- ➢ The `price` variable was converted to numeric by removing non-numeric characters using `gsub("[^0-9\\.]", "", product_data$price)`. This allows for mathematical operations on the price data.
- ➢ Similarly, the `trade` variable was converted to numeric and missing values were filled with 0 using `as.numeric(gsub("[^0-9\\.]", "", product_data$trade))` and `product_data$trade_numeric[is.na(product_data$trade_numeric)] <- 0`.
- ➢ A new `price_usd` data column was created by multiplying the `price_numeric` with a conversion rate (i.e., set to the value of 1 in this code script). This allowed for standardised financial analysis across different currencies.
- ➢ A new `thumbnail_url` column was created by concatenating a base URL with the `thumbnail` data column. This enriches the data with complete URLs for product thumbnails.

The specific code script used to preprocess and clean the product data that were web scraped from the AliExpress consumer electronics webpage is included in the `R_Code_Script_Data_Preprocessing_and_Analysis.R` *R* code script I provided together with the web

scraping research project documentation.

Were the 'raw' data saved in addition to the preprocessed/cleaned/labelled data (e.g., to support unanticipated future uses)? If so, please provide a link or access point to the 'raw' data.

The code script does not explicitly save the raw product data. Rather, the script loads data from the `AliExpress_Consumer_Electronics_Product_Data.csv` data file, which contains the raw product data that were web scraped from items that are part of the *"Consumer Electronics"* product category on the AliExpress webpage, and then proceeds with preprocessing and analysis. The cleaned dataset was saved at the end of the *R* code script, using the `write.csv(product_data, "Cleaned_AliExpress_Consumer_Electronics_Data.csv", row.names = FALSE)` command, thereby bypassing the original AliExpress consumer electronics data file.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, the *R* code script used for data preprocessing, cleaning, and labelling is available in the `R_Code_Script_Data_Preprocessing_and_Analysis.R` *R* code file provided together with the research project documentation. This script loads the raw product data, applies the cleaning steps described above, and then proceeds with data analysis and visualisation.

Any other comments?

The data preprocessing and cleaning steps in this script are focused on handling missing values, converting data types (i.e., particularly for price and trade volumes), and enriching the data with additional information (i.e., thumbnail URLs and USD price conversions). The cleaned data is then used for a series of exploratory data analyses and visualisations, including examining product type distribution, price distribution, the correlation between trade volume and price, and identifying top stores on AliExpress' retail platform by their respective product counts. As such, the resulting code script provides a solid foundation for understanding key aspects of the AliExpress Consumer Electronics dataset, but could be extended with additional preprocessing or analysis steps depending on the specific research questions being pursued.

Chapter 5

USES

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

Has the dataset been used for any tasks already? If so, please provide a description.
This dataset has not yet been used for any specific tasks. As it was collected as part of a research project on web scraping and online data collection, its primary purpose so far has been for educational and skill development purposes.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
Due to the fact that the AliExpress consumer electronics dataset has not been published yet, no repository exists to link the data to papers or systems at the present time.

What tasks could the dataset be used for? Is there anything about the composition of the dataset or the way it was collected and preprocessed, cleaned or labelled that might impact future uses?

This dataset of consumer electronics products from AliExpress could potentially be used for a variety of market research and analysis tasks, including:

- ➢ Examining pricing trends and strategies in the consumer electronics space.
- ➢ Analysing product titles and descriptions to understand common features and selling points.
- ➢ Comparing products between different sellers and brands.

- ➢ Exploring the breadth of the consumer electronics category and identifying potential opportunities or gaps in the market.

Since the AliExpress consumer electronics dataset was collected through web scraping and only contains free and readily available product information, there are minimal risks around unfair treatment of individuals or groups. Nevertheless, users should be aware that the data collected for this research project only represents a snapshot in time of the total number of products available on AliExpress' retail platform. Moreover, users should note that the pricing, availability, and other details regarding products sold through AliExpress' retail platform are subject to change over time.

Also, users should remember that the data comes from a single e-commerce platform and may not be representative of the entire consumer electronics market. Avoiding over-generalisation and combining this data with other sources could help mitigate the risk of biased or skewed insights.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset only contains product information and metadata, and does not include any details about individuals, transactions, or consumer behaviour. As such, it would not be appropriate for tasks that require that type of data, including analysing purchasing patterns, conducting user segmentation, or building recommender systems. The data is not verified or guaranteed to be fully accurate, so it should not be used for critical business decisions without additional validation from other sources.

Any other comments?

Potential users of this dataset should be aware of its limitations as web scraped data from a single source captured at one point in time. However, for general market research and analysis of the consumer electronics space on AliExpress, it provides a solid foundational overview of the types of products available and their key characteristics. Users can build upon this dataset with additional information to draw more robust insights as needed for their particular use case.

BIBLIOGRAPHY

Cirqueira, Douglas et al. (Jan. 2020): "Customer Purchase Behavior Prediction in E-commerce: A Conceptual Framework and Research Agenda". In: *Lecture Notes in Computer Science*, pp. 119–136. DOI: 10.1007/978-3-030-48861-1_8. URL: https://doi.org/10.1007/978-3-030-48861-1_8.

DeVito, Nicholas J, Georgia C Richards & Peter Inglesby (Sept. 2020): "How we learnt to stop worrying and love web scraping". In: *Nature* 585.7826, pp. 621–622. DOI: 10.1038/d41586-020-02558-0. URL: https://www.nature.com/articles/d41586-020-02558-0.

Grand View Research, Inc. (Oct. 2023): *Consumer Electronics Market Size, Share Trends Analysis Report by product (Smartphones, tablets, desktops, laptops, digital cameras, hard disk drives, e-readers), by sales channel (Offline, online), by region, and segment Forecasts, 2023 - 2030*. URL: https://www.grandviewresearch.com/industry-analysis/personal-consumer-electronics-market.

Insights, Fortune Business (Mar. 2024): *Consumer Electronics Market Size, Share Industry Analysis, By Product Type (Electronic Devices (Television, Computer, Digital Camera Camcorder, and Others) and Home Appliances (Refrigerator, Washing Machine, Air Conditioner, and Others)), Distribution Channel (Offline and Online), and Regional Forecast, 2024-2032*. URL: https://www.fortunebusinessinsights.com/consumer-electronics-market-104693.

Research & Markets Ltd. (2023): *Global Consumer Electronics Market - Research and Markets*. URL: https://www.researchandmarkets.com/report/consumer-electronics.

Rita, Paulo & Ricardo F. Ramos (Aug. 2022): "Global Research Trends in Consumer Behavior and Sustainability in E-Commerce: A Bibliometric Analysis of the Knowledge Structure". In: *Sustainability* 14.15, p. 9455. DOI: 10.3390/su14159455. URL: https://www.mdpi.com/2071-1050/14/15/9455.

Speckmann, Felix (Dec. 2021): "Web scraping". In: *Zeitschrift Fur Psychologie-journal of Psychology* 229.4, pp. 241–244. DOI: 10.1027/2151-2604/a000470. URL: https://doi.org/10.1027/2151-2604/a000470.

Verghese, S. (Nov. 2023): *Consumer Electronics market*. URL: https://www.futuremarketinsights.com/reports/consumer-electronics-market.